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

$$(1.1) \qquad P_\lambda\{X(m) \leq n\} = \alpha.$$

Since

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

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

$$(1.3) \qquad P_\lambda\{T > t\} = \alpha.$$

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

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

(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 $\lambda_*[T', X(T')] \leq \lambda_{*2}[X(m)]$ with equality if and only if $X(m) \leq r$, 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 $\lambda_1^*(m) < \lambda$.

1.3. (a) For the curtailed test of $H_0: p \leq p_0$ against $H_1: p > p_0$, show how the test can be further curtailed with *acceptance* of H_0.

(b) Discuss confidence bounds for p following a curtailed test of $H_0: p \leq p_0$ against $H_1: p > p_0$.

1.4.* Let $X(t)$, $t \geq 0$, be a Poisson process with mean λt and set $\Pi_\lambda(t, n) = P_\lambda\{X(t) \geq n\}$. For testing $H_0: \lambda = \lambda_0$ against $H_1: \lambda > \lambda_0$ based on a fixed sample of size $t = m$, the attained significance level or p-value is defined to be $\Pi_{\lambda_0}(m, X(m))$. That is, the p-value is the Type I error probability for that test with rejection region of the form $\{X(m) \geq r\}$ for some r, for which the value of $X(m)$ actually observed just barely lies in the rejection region. Small values of $\Pi_{\lambda_0}(m, X(m))$ are conventionally interpreted as providing more evidence against H_0 than large values. For a curtailed test of $H_0: \lambda \leq \lambda_0$ against $H_1: \lambda > \lambda_0$ defined by the constants m and r, suggest a definition of the attained significance of the observed value $(T', X(T'))$. For your definition, explain why data yielding a small p-value should be thought to provide strong evidence against H_0.

CHAPTER II

The Sequential Probability Ratio Test

1. Definition and Examples

We begin by recalling the Neyman–Pearson Lemma for testing a simple hypothesis against a simple alternative. Let x denote a (discrete or continuous) random variable (or vector) with probability density function f. To test H_0: $f = f_0$ against $H_1: f = f_1$, define the likelihood ratio $l(x) = f_1(x)/f_0(x)$, choose a constant $r > 0$, and

$$\text{Reject } H_0 \quad \text{if } l(x) \geq r,$$
$$\text{Accept } H_0 \quad \text{if } l(x) < r.$$

This class of tests (depending on r) is optimal from both a Bayesian and a frequentist viewpoint. In particular, any test of H_0 against H_1 which is based on observing x and has significance level no larger than $\alpha = P_0\{l(x) \geq r\}$ must have power no larger than $P_1\{l(x) \geq r\}$ (cf. Cox and Hinkley, 1974, p. 91). Here P_i denotes probability under the hypothesis H_i, $i = 0, 1$.

A sequential probability ratio test of H_0 against H_1 admits a third possibility in addition to rejecting H_0 for large $l(x)$ and accepting for small $l(x)$, namely that for intermediate values of $l(x)$ one collects more data. More precisely let $x_1, x_2, \ldots$ be a sequence of random variables with joint density functions

$$P\{x_1 \in d\xi_1, \ldots, x_n \in d\xi_n\} = f_n(\xi_1, \ldots, \xi_n)\, d\xi_1, \ldots, d\xi_n \qquad (n = 1, 2, \ldots).$$

Consider testing the simple hypotheses $H_0: f_n = f_{0n}$ for all n against $H_1: f_n = f_{1n}$ for all n. Let $l_n = l_n(x_1, \ldots, x_n) = f_{1n}(x_1, \ldots, x_n)/f_{0n}(x_1, \ldots, x_n)$. Choose constants $0 < A < B < \infty$ (usually $A < 1 < B$) and sample $x_1, x_2, \ldots$ sequentially until the random time

$$(2.1) \qquad \begin{aligned} N &= \text{first } n \geq 1 \quad \text{such that } l_n \notin (A, B) \\ &= \infty \text{ if } l_n \in (A, B) \quad \text{for all } n \geq 1 \end{aligned}$$

√Stop sampling at time N and if $N < \infty$

$$\text{Reject } H_0 \quad \text{if } l_N \geq B,$$
$$\text{Accept } H_0 \quad \text{if } l_N \leq A.$$

We defer temporarily the technical issue of whether this procedure actually terminates, i.e. whether $P_i\{N < \infty\} = 1$ for $i = 0$ and 1. Assuming that it does, we have a test of size $P_0\{l_N \geq B\}$ and power $P_1\{l_N \geq B\}$. As additional properties of the test we shall consider its sample size distribution $P_i\{N = n\}$, $n = 1, 2, \ldots$ or more simply its expected sample size $E_i(N)$ for $i = 1, 2$.

It will turn out that the sequential probability ratio test has a very strong optimality property when the x_n are independent and identically distributed: it minimizes E_i (sample size) for $i = 0$ and 1 over all (sequential) tests having the same size and power. The basic idea behind this fact is very simple, although a complete proof is circuitous and difficult (see Ferguson, 1967, p. 365).

The goals of this chapter are to study the sequential probability ratio test as a test of a simple hypothesis against a simple alternative (Section 2), to extend it to certain problems involving composite hypotheses (Section 3), to give an informal discussion of its optimality property (Section 4), and to criticize it in Section 5, in preparation for the various modifications of Chapters III and IV. An introduction to the related "cusum" procedures of statistical quality control is given in Section 6.

Before proceeding we present two simple, but especially instructive examples.

Example 2.2. Let $x_1, x_2, \ldots$ be independent and normally distributed with mean μ and unit variance. For testing $H_0: \mu = \mu_0$ against $H_1: \mu = \mu_1$ (say $\mu_0 < \mu_1$), the likelihood ratio is

$$(2.3) \qquad \begin{aligned} l_n &= \prod_{k=1}^{n} \{\phi(x_k - \mu_1)/\phi(x_k - \mu_0)\} \\ &= \exp\{(\mu_1 - \mu_0)S_n - \tfrac{1}{2}n(\mu_1^2 - \mu_0^2)\}, \end{aligned}$$

where $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ and $S_n = \sum_1^n x_k$. Hence the stopping rule (2.1) can be re-written

$$(2.4) \qquad \begin{aligned} N &= \text{first } n \geq 1 \text{ such that } S_n - \tfrac{1}{2}n(\mu_1 + \mu_0) \notin (a, b) \\ &= \infty \text{ if no such } n \text{ exists,} \end{aligned}$$

where $a = \log A/(\mu_1 - \mu_0)$, $b = \log B/(\mu_1 - \mu_0)$, and if $N < \infty$ the sequential probability ratio test rejects H_0 if and only if

$$S_N \geq b + \tfrac{1}{2}N(\mu_1 + \mu_0).$$

A simple special case is the symmetric one $\mu_1 = -\mu_0, b = -a$, for which (2.4) becomes

$$N = \text{first } n \geq 1 \text{ such that } |S_n| \geq b \qquad (2.5)$$
$$= \infty \text{ if } |S_n| < b \text{ for all } n.$$

EXAMPLE 2.6. Let $x_1, x_2, \ldots$ be independent and identically distributed with $P_p\{x_k = 1\} = p, P_p\{x_k = -1\} = q\ (p + q = 1)$. For testing $H_0: p = p_0$ against $H_1: p = p_1\ (p_0 < p_1)$ the likelihood ratio is

$$l_n = (p_1 p_0^{-1})^{(n+S_n)/2}(q_1 q_0^{-1})^{(n-S_n)/2} \qquad (2.7)$$
$$= (p_1 q_0 p_0^{-1} q_1^{-1})^{S_n/2}(p_1 p_0^{-1} q_1 q_0^{-1})^{n/2},$$

where $S_n = \Sigma x_k$. Again (2.1) can be expressed directly in terms of the random walk $\{S_n\}$, and a simple symmetric case occurs when $p_0 = q_1$ and $B = A^{-1}$, so (2.1) becomes

$$N = \text{first } n \geq 1 \text{ such that } |S_n| \geq b, \qquad (2.8)$$
$$= \infty \text{ if } |S_n| < b \text{ for all } n,$$

where $b = (\log B)/\log(q_0 p_0^{-1})$.

2. Approximations for $P_i\{l_N \geq B\}$ and $E_i(N)$

We continue to use the notation and assumption of the preceding section. In particular we continue to assume that $P_i\{N < \infty\} = 1\ (i = 0, 1)$, and consider the following problems

(a) Relate $\alpha = P_0\{l_N \geq B\}$ and $\beta = P_1\{l_N \leq A\}$ to A and B;

(b) Relate $E_i(N)$ to A, B for $i = 0, 1$.

We begin with a simple calculation related to (a). More general versions of this idea are one of the principal techniques developed in the following chapters. Let B_n denote the subset of n-dimensional space where $A < l_k(\xi_1, \ldots, \xi_k) < B$ for $k = 1, 2, \ldots, n - 1$ and $l_n(\xi_1, \ldots, \xi_n) \geq B$. Hence $\{N = n, l_n \geq B\} = \{(x_1, \ldots, x_n) \in B_n\}$. By direct calculation

$$\alpha = P_0\{l_N \geq B\} = \sum_1^\infty P_0\{N = n, l_n \geq B\} = \sum_1^\infty \int_{B_n} f_{0n}\, d\xi_1 \cdots d\xi_n$$
$$= \sum_1^\infty \int_{B_n} \frac{f_{0n}}{f_{1n}} f_{1n}\, d\xi_1 \cdots d\xi_n = \sum_1^\infty E_1[l_n^{-1}; N = n, l_n \geq B] \qquad (2.9)$$
$$= E_1[l_N^{-1}; l_N \geq B] \leq B^{-1} P_1\{l_N \geq B\} = B^{-1}(1 - \beta).$$

A similar argument with the roles of A and B interchanged leads to

$$\beta = P_1\{l_N \leq A\} \leq A P_0\{l_N \leq A\} = A(1 - \alpha). \qquad (2.10)$$

The inequalities (2.9) and (2.10) fail to be equalities only because l_n does not have to hit the boundaries exactly when it first leaves (A, B). However, if we agree to ignore this discrepancy and treat (2.9) and (2.10) as approximate equalities

$$\alpha \cong B^{-1}(1 - \beta) \quad \text{and} \quad \beta \cong A(1 - \alpha), \qquad (2.11)$$

we can solve for α and β to obtain crude but extremely simple approximations:

$$\alpha \cong \frac{1 - A}{B - A}; \qquad \beta \cong A\left(\frac{B - 1}{B - A}\right). \qquad (2.12)$$

To study the expected value of N we make the additional assumption that the observations x_n are independent and identically distributed, so $l_n = \prod_{k=1}^n \{f_1(x_k)/f_0(x_k)\}$, where f_i is the probability density function of x_1 under the simple hypothesis $H_i\ (i = 0, 1)$. In this case the log scale is convenient, and

$$\log l_n = \sum_{k=1}^n \log\{f_1(x_k)/f_0(x_k)\}$$

is a sum of independent, identically distributed random variables. Moreover,

$$N = \text{first } n \geq 1 \text{ such that } \log l_n \notin (a, b),$$
$$= \infty \text{ if } \log l_n \in (a, b) \text{ for all } n,$$

where $a = \log A$, $b = \log B$.

The basic argument for approximating $E_i(N)$ consists of two parts. By Wald's identity given below in Proposition 2.18

$$E_i\{\log l_N\} = \mu_i E_i(N), \qquad (2.13)$$

where

$$\mu_i = E_i[\log\{f_1(x_1)/f_0(x_1)\}] \qquad (i = 0, 1).$$

Moreover, the approximation we used in deriving (2.11) suggests that $\log l_N$ be regarded as a two-valued random variable taking on the values a and b, so

$$E_i\{\log l_N\} \cong a P_i\{l_N \leq A\} + b P_i\{l_N \geq B\}, \qquad (2.14)$$

where the probabilities in (2.14) are given approximately in (2.12). Putting (2.12), (2.13), and (2.14) together yields the approximations

$$E_1 N \cong \mu_1^{-1}\{aA(B - 1) + bB(1 - A)\}/(B - A) \qquad (2.15)$$

and

$$E_0 N \cong \mu_0^{-1}\{a(B - 1) + b(1 - A)\}/(B - A). \qquad (2.16)$$

Note that $\mu_0 < 0 < \mu_1$. In fact, since $\log x \leq x - 1$ with equality if and only if $x = 1$,

$A=1$

$\frac{1}{\mu_1}\ a$

$A=0, B=1: \frac{b}{\mu_1} = \frac{\log B}{\mu_1} = 0$

$$\mu_0 = E_0[\log\{f_1(x_1)/f_0(x_1)\}] = \int \log\{f_1(\xi)/f_0(\xi)\}f_0(\xi)\,d\xi$$
$$\leq \int \{f_1(\xi)/f_0(\xi) - 1\}f_0(\xi)\,d\xi = 0$$

with equality of and only if f_0 and f_1 are the same density function. A similar argument shows $-\mu_1 < 0$. Since by the strong law of large numbers

(2.17) $$P_i\{n^{-1}\log l_n \to \mu_i\} = 1 \qquad (i = 0, 1),$$

if $A \to 0$ and $B \to \infty$ we expect to find under P_i that $\log l_n$ leaves the interval (a, b) at approximately the time and place that $n\mu_i$ does, to wit b/μ_1 under H_1 and $|a|/|\mu_0|$ under H_0. It is easy to see that this is asymptotically consistent with (2.15) and (2.16).

The following two propositions justify (2.13).

Proposition 2.18 (Wald's identity). *Let $y_1, y_2, \ldots$ be independent and identically distributed with mean value $\mu = Ey_1$. Let M be any integer valued random variable such that $\{M = n\}$ is an event determined by conditions on $y_1, \ldots, y_n$ (and is independent of $y_{n+1}, \ldots$) for all $n = 1, 2, \ldots$, and assume that $EM < \infty$. Then $E(\sum_{k=1}^M y_k) = \mu EM$.*

PROOF. Suppose initially that $y_k \geq 0$. We write $\sum_{k=1}^M y_k = \sum_{k=1}^\infty I_{\{M \geq k\}} y_k$, and note that $\{M \geq k\} = (\bigcup_{j=1}^{k-1}\{M = j\})^c$ is independent of $y_k, y_{k+1}, \ldots$. Hence by monotone convergence

$$E\left(\sum_1^M y_k\right) = \sum_1^\infty E(y_k; M \geq k) = \mu \sum_1^\infty P\{M \geq k\} = \mu EM.$$

For the general case we write

$$\sum_{k=1}^M y_k = \sum_{k=1}^M y_k^+ - \sum_{k=1}^M y_k^-,$$

where $a^+ = \max(a, 0)$, $a^- = -\min(a, 0)$, and apply the case already considered to these two terms separately. $\square$

Proposition 2.19 (Stein's lemma). *Let $y_1, y_2, \ldots$ be independent and identically distributed with $P\{y_1 = 0\} < 1$. Let $-\infty < a < b < \infty$ and define*

$$M = \text{first } n \geq 1 \text{ such that } \sum_1^n y_k \notin (a, b)$$
$$= \infty \text{ if } \sum_1^n y_k \in (a, b) \text{ for all } n.$$

Then there exist constants $C > 0$ and $0 < \rho < 1$ such that $P\{M > n\} \leq C\rho^n$ $(n = 1, 2, \ldots)$. In particular $EM^k < \infty$ for all $k = 1, 2, \ldots$ and $Ee^{\lambda M} < \infty$ for $\lambda < \log \rho^{-1}$.

For a proof of Proposition 2.19 see Problem 2.6.

Remark 2.20. The expected sample size approximations, (2.15) and (2.16) can also be expressed in terms of the error probabilities α and β. From (2.11), (2.13), and (2.14) follow

(2.21) $$E_1 N \cong \mu_1^{-1}\left\{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta\log\left(\frac{\beta}{1-\alpha}\right)\right\}$$

and

(2.22) $$E_0(N) \cong \mu_0^{-1}\left\{\alpha\log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right)\right\}.$$

The calculation (2.9) turns out to be very important in a variety of cases. It is useful to present a more abstract version for future reference.

Let $z_1, z_2, \ldots$ be an arbitrary sequence of random variables. For each n, let $\mathscr{E}_n$ denote the class of random variables determined by $z_1, \ldots, z_n$, i.e. a random variable $Y \in \mathscr{E}_n$ if and only if $Y = f(z_1, \ldots, z_n)$ for some (Borel) function f of n variables. For an event A the notation $A \in \mathscr{E}_n$ means that the indicator of A, I_A, belongs to $\mathscr{E}_n$. A random variable T with values in $\{1, 2, \ldots, +\infty\}$ is called a stopping time if $\{T = n\} \in \mathscr{E}_n$ for all n. Hence an observer who knows the values of $z_1, \ldots, z_n$ knows whether $T = n$. A random variable Y is said to be prior to a stopping time T if

(2.23) $$YI_{\{T=n\}} \in \mathscr{E}_n \quad \text{for all} \quad n, \qquad Y \in \mathcal{F}_T$$

or equivalently $YI_{\{T \leq n\}} \in \mathscr{E}_n$ for all n. In particular T is prior to itself. Condition (2.23) has the interpretation that by the time T an observer who knows the values $z_1, z_2, \ldots, z_T$ also knows the value of Y.

Proposition 2.24 (Wald's likelihood ratio identity). *Let P_0 and P_1 denote two probabilities and assume that there exists a likelihood ratio l_n for $z_1, \ldots, z_n$ under P_1 relative to P_0 in the sense that $l_n \in \mathscr{E}_n$ and for each $Y_n \in \mathscr{E}_n$*

(2.25) $$E_1(Y_n) = E_0(Y_n l_n). \qquad Y \in \mathcal{F}_T$$

For any stopping time T and non-negative random variable Y prior to T

$$E_1(Y; T < \infty) = E_0(Yl_T; T < \infty\}.$$

In particular, if $Y = I_A$

$$P_1(A \cap \{T < \infty\}) = E_0(l_T; A \cap \{T < \infty\}).$$

PROOF. The proof basically repeats (2.9) with (2.23) and (2.25) used to justify the second equality:

$$E_1(Y; T < \infty) = \sum_{n=1}^\infty E_1(Y; T = n) = \sum_{n=1}^\infty E_0(Yl_n; T = n)$$
$$= E_0(Yl_T; T < \infty). \qquad \square$$

Remark. If $z_1, \ldots, z_n$ have joint density p_{in} under P_i $(i = 0, 1)$, then $l_n = p_{1n}/p_{0n}$. In this case for $Y_n = f_n(z_1, \ldots, z_n)$, (2.25) follows from

$$E_1(Y_n) = \int f_n(\xi_1,\ldots,\xi_n) p_{1n}(\xi_1,\ldots,\xi_n)\, d\xi_1,\ldots, d\xi_n$$

$$= \int f_n(\xi_1,\ldots,\xi_n) \frac{p_{1n}(\xi_1,\ldots,\xi_n)}{p_{0n}(\xi_1,\ldots,\xi_n)} p_{0n}(\xi_1,\ldots,\xi_n)\, d\xi_1,\ldots, d\xi_n$$

$$= E_0\left\{Y_n \frac{p_{1n}(z_1,\ldots,z_n)}{p_{0n}(z_1,\ldots,z_n)}\right\}.$$

✓ **Remark 2.26.** We have already used Proposition 2.24 to derive the approximations (2.12). It can also be very useful in Monte Carlo studies. Consider the problem of estimating $\alpha = P_0\{l_N \geq B\}$ by simulation. The naive estimator is

$$\tilde{\alpha} = n^{-1} \sum_{k=1}^{n} I\{l_{N_k} \geq B\}$$

based on n independent realizations of (N, l_N) under the probability P_0. Since α is typically small, and the standard deviation of $\tilde{\alpha}$ is $[\alpha(1-\alpha)/n]^{1/2}$, it is often necessary to take large values of n to provide an accurate estimate. For example, if one wants to estimate α to within 10% of its value with probability .95, for small α n must be about $400/\alpha$. $4n \cdot \frac{1-\alpha}{\alpha}$

Alternatively, by Proposition 2.24, another unbiased estimator of α is

$$\hat{\alpha} = n^{-1} \sum_{k=1}^{n} \frac{f_{0N_k}}{f_{1N_k}} I\left\{\frac{f_{1N_k}}{f_{0N_k}} \geq B\right\},$$

where now the experiment generating the observations is conducted under P_1. By (2.11) and Proposition 2.24

$$n\,\mathrm{var}(\hat{\alpha}) \leq E_1\left\{\left(\frac{f_{0N}}{f_{1N}}\right)^2; \frac{f_{1N}}{f_{0N}} \geq B\right\}$$

$$\leq B^{-1} E_1\left\{\frac{f_{0N}}{f_{1N}}; \frac{f_{1N}}{f_{0N}} \geq B\right\} = B^{-1}\alpha \cong \frac{\alpha^2}{1-\beta}.$$

For this experiment only about 400 replications are required to achieve 10% relative accuracy with probability .95 no matter how small α is.

3. Tests of Composite Hypotheses

Although the sequential probability ratio test is derived as a test of a simple hypothesis against a simple alternative, it is natural to consider the consequences of using it for composite hypotheses. For example, to test $H_0\colon \theta \leq \theta^*$ against $H_1\colon \theta > \theta^*$ in a one-parameter family of distributions one might choose surrogate simple hypotheses $\theta_0 \leq \theta^*$ and $\theta_1 > \theta^*$, and use a sequential probability ratio test of θ_0 against θ_1. Then one would want to know the entire power function and expected sample size as a function of θ, in addition to their values at θ_0 and θ_1. Before discussing the general case we consider some simple examples.

Recall Example 2.6, and assume that the symmetric case with stopping rule (2.8) is to be used for the composite hypotheses $H_0\colon p \leq \frac{1}{2}$ against $H_1\colon p > \frac{1}{2}$. There is no loss of generality in assuming that B is chosen so that $b = \log B/\log(q_0 p_0^{-1})$ is an integer. Since the random walk $\{S_n\}$ proceeds by steps of ± 1, $P_p\{|S_N| = b\} = 1$, so in this case the approximations of Section 2 are equalities. In particular by (2.11) and (2.15)–(2.16) (or Remark 2.20),

$$\alpha = \beta = P_{p_0}\{S_N = b\} = 1/(1+B) = 1/[1 + (q_0/p_0)^b]$$

and

$$E_{p_0}(N) = |q_0 - p_0|^{-1} b(1 - 2\alpha).$$

[In making the required identifications it is helpful to keep in mind that $b = \log B/\log(q_0 p_0^{-1})$, for this example is proportional but not identical to $b = \log B$ in (2.15)–(2.16).]

Now suppose $p < \frac{1}{2}$, but $p \neq p_0$. Since (2.8) does not explicitly involve p_0, it is also a sequential probability ratio test of p against $1 - p$, but with a different value of B. Put another way, (2.8) can be re-written

(2.27) $$N = \text{first } n \geq 1 \text{ such that } (q/p)^{S_n} \notin \left(\left\{\left(\frac{q}{p}\right)^b\right\}^{-1}, \left(\frac{q}{p}\right)^b\right),$$

which is of the form $l_n \notin (B_1^{-1}, B_1)$, where l_n is (2.7) with $p_0 = p$, $p_1 = q$, and $B_1 = (q/p)^b$. Hence with the proper re-interpretation the approximations of Section 2 yield $P_p\{S_N = b\}$ and $E_p(N)$ for general $p \neq \frac{1}{2}$. The results are

(2.28) $$P_p\{S_N = b\} = 1/\{1 + (q/p)^b\}$$

and

(2.29) $$E_p(N) = |q - p|^{-1} b |1 - 2P_p\{S_N = b\}|,$$

for all $p \neq \frac{1}{2}$. The corresponding results for $p = \frac{1}{2}$ are easily computed by taking the limit as $p \to \frac{1}{2}$ in (2.28) and (2.29).

A similar discussion applies to Example 2.2, but now the approximations do not turn into equalities. Consider again the symmetric case (2.5) for simplicity, and note that by symmetry (2.9) becomes $\alpha < B^{-1}(1 - \alpha)$, so the approximation (2.11) is an inequality

$$\alpha = P_{\mu_0}\{S_N \geq b\} < 1/(1+B) = 1/(1 + e^{2b|\mu_0|}).$$

In this case for arbitrary $\mu > 0$, (2.5) can be written

$$N = \text{first } n \geq 1 \text{ such that } e^{2\mu S_n} \notin (e^{-2\mu b}, e^{2\mu b}),$$

which has the form of a symmetric sequential probability ratio test of $-\mu$ against μ with $B = A^{-1} = e^{2\mu b}$ (cf. (2.3)). Hence for N defined by (2.5), for all $\mu > 0$

(2.30) $$P_{-\mu}\{S_N \geq b\} < 1/(1 + e^{2\mu b})$$

Table 2.1. Symmetric Sequential Probability Ratio Test for Normal Observations: $b = 4.91$

μ	$P_{-\mu}\{S_N \geq b\} \cong$	$E_\mu(N) \cong$
.4	.019	11.8
.3	.050	14.7
.2	.123	18.5
.1	.272	22.3
0	.500	24.1

and

$$E_\mu(N) \cong \mu^{-1} b\{(e^{2\mu b} - 1)/(e^{2\mu b} + 1)\}. \tag{2.31}$$

An approximation for $\mu = 0$ can be obtained by taking the limit as $\mu \to 0$.

Table 2.1 gives a numerical example of the approximations (2.30) and (2.31). Since (2.30) and (2.31) are only approximations, one would like to know how good they are. It may be shown (cf. Problem 2.2, also III.5, III.6, VIII.5, and X.2) that if $p(b)$ denotes the right hand side of (2.30) and $e(b)$ the right hand side of (2.31), then $p(b + .583)$ and $e(b + .583)$ give very good approximations for a wide range of values of b and μ. Thus for $\mu = .3$, in order that $P_{-.3}\{S_N \geq b\} = .05$, the right hand side of (2.30) suggests taking $b = 4.91$. However, the true value of $P_{-.3}\{S_N \geq 4.91\}$ is much closer to $p(4.91 + .583) = .036$. The approximation (2.31) is generally somewhat more satisfactory: for $\mu = .3$, $e(4.91) = 14.7$ compared to the more accurate $e(4.91 + .583) = 17.0$.

For a different example concerning the accuracy of (2.12) and (2.15)–(2.16), see Problem 2.1. A general conclusion about these approximations is the following. Usually (2.12) overestimates α and β, and the relative error is often as large as 30–70%; (2.15) and (2.16) typically underestimate the expected sample size, and except in the case of quite small samples sizes the relative error is usually on the order of 5–25%. Later we shall see that (2.12), (2.15)–(2.16), and related results can be substantially improved by approximating rather than neglecting the excess over the boundary. For example, the number .583 of the preceding paragraph is approximately the expected difference between the random variable S_N and the boundaries $\pm b$. Hence the improved approximations amount to using the no-overshoot approximations with new boundaries $\pm b'$ differing from the old ones by the average excess over the boundaries. See Chapter X for the theoretical justification.

In spite of their lack of accuracy, the remarkable generality and simplicity of the approximations of Section 2 make them quite useful.

Note that a fixed sample size test of $\mu = -.3$ against $\mu = +.3$ with significance level and Type II error probability of .036 requires 36 observations, so when $|\mu| \geq .3$ a sequential probability ratio test saves on the average more than 50% of these observations. Even when $\mu = 0$, this sequential probability ratio test has an expected sample size of only 30 observations. See Problem 2.9 for additional comparisons of fixed sample and sequential probability ratio tests.

The preceding examples show that it is sometimes possible to obtain approximations to the entire power function and expected sample size function of a sequential probability ratio test of composite hypotheses using only the theory developed for simple hypotheses. The following discussion shows that this is generally possible in the context of a one-parameter exponential family of distributions.

Consider a general sequential probability ratio test defined by (2.1) with the additional hypothesis that $x_1, x_2, \ldots$ are independent and identically distributed, so that

$$l_n = \prod_{k=1}^{n} \{f_1(x_k)/f_0(x_k)\},$$

where f_i is the density function of x_1 under H_i $(i = 0, 1)$. Let f be some third probability density function, and consider the problem of evaluating $P_f\{l_N \geq B\}$. This was feasible in the preceding examples because there was yet a fourth density function f^* such that the original test of f_0 against f_1 was equivalent to a test of f against f^* with new values of A and B. Moreover, the new A and B were the old A and B raised to a power. Hence given f, assume that there exists a probability density function f^* and a $\theta_1 \neq 0$ such that

$$\frac{f^*(x)}{f(x)} = \left[\frac{f_1(x)}{f_0(x)}\right]^{\theta_1}. \tag{2.32}$$

Then for $\theta_1 > 0$ (for example)

$$N = \text{first } n \geq 1 \text{ such that } \prod_1^n \{f_1(x_k)/f_0(x_k)\} \notin (A, B)$$

$$= \text{first } n \geq 1 \text{ such that } \prod_1^n \{f^*(x_k)/f(x_k)\} \notin (A^{\theta_1}, B^{\theta_1}),$$

and it follows from (2.12) that

$$P_f\{l_N \geq B\} \cong \frac{1 - A^{\theta_1}}{B^{\theta_1} - A^{\theta_1}}. \tag{2.33}$$

A similar calculation applies when $\theta_1 < 0$, and an approximation for $E_f(N)$ is also easily obtained (see Problem 2.16).

Note that (2.32) is satisfied if and only if

$$\int_{-\infty}^{\infty} \left[\frac{f_1(x)}{f_0(x)}\right]^{\theta_1} f(x)\,dx = 1. \tag{2.34}$$

Let $z(x) = \log\{f_1(x)/f_0(x)\}$, and define a function $\psi(\theta)$ by

$$e^{\psi(\theta)} = \int_{-\infty}^{\infty} e^{\theta z(x)} f(x)\,dx$$

whenever the integral converges. Then

$$f_\theta(x) = e^{\theta z(x) - \psi(\theta)} f(x)$$

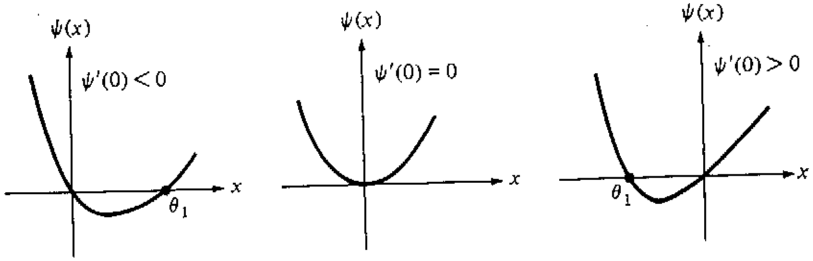

Figure 2.1.

defines an exponential family of distributions and the existence of a θ_1 satisfying (2.32) or (2.34) is equivalent to the existence of a $\theta_1 \neq 0$ such that $\psi(\theta_1) = 0$. It is easy to see by differentiation (assuming always that the integrals converge) that

$$\psi'(\theta) = \int_{-\infty}^{\infty} z(x) f_\theta(x)\, dx$$

and

$$\psi''(\theta) = \int_{-\infty}^{\infty} [z(x)]^2 f_\theta(x)\, dx - [\psi'(\theta)]^2 \geq 0,$$

so ψ is convex. Hence under some modest convergence assumptions ψ has the appearance of one of the three functions in Figure 2.1; and the desired $\theta_1 \neq 0$ satisfying (2.34) exists if $\psi'(0) = \int_{-\infty}^{\infty} z(x) f(x)\, dx = E_f z(x_1) \neq 0$, but not if $\psi'(0) = 0$. Note that the case $\psi'(0) = 0$ corresponds to cases where we obtained approximations to $P\{l_N \geq B\}$ and $E(N)$ by continuity in the two examples at the beginning of this section. That technique works quite generally, but for a different approach see Problems 2.10 and 2.11.

To see how this general discussion relates to the examples, observe that if $f_i(x) = \phi(x - (-1)^i \mu_0)$ for some $\mu_0 \neq 0$, and if f is a normal density with mean $\mu \neq 0$, then f^* is normal with mean $-\mu$. Similarly in the symmetric (about $\frac{1}{2}$) Bernoulli example, if f is Bernoulli with success probability $p \neq \frac{1}{2}$, then f^* is Bernoulli with success probability $1 - p$.

Example. Morgan *et al.* (1951) made a detailed study of the sequential probability ratio test as a method of grading raw milk prior to pasteurization. The classification process involved counting bacterial clumps in each of several fields of a film of milk in order to determine their average density. After deciding that subject to certain qualifications the number of bacterial clumps per field was reasonably approximated by a Poisson distribution, the authors proposed a sequential probability ratio test to determine the acceptability of the milk as (for example) Grade A milk in the state of Connecticut.

Let x_i denote the number of clumps of bacteria in the ith field, and assume that $x_1, x_2, \ldots$ are independent Poisson random variables with mean λ. The standard fixed sample plan for accepting milk as Grade A was equivalent to taking a sample of size $m = 30$ and testing $H_0\colon \lambda = .1808$ against $H_1\colon \lambda = .5249$ with Type I and Type II error probabilities of .05. Let $S_n = x_1 + \cdots + x_n$, $n = 1, 2, \ldots$. A simple calculation shows that a sequential probability ratio test of $H_0\colon \lambda = \lambda^{(0)}$ against $H_1\colon \lambda = \lambda^{(1)}$ $(\lambda^{(0)} < \lambda^{(1)})$ is defined by a stopping rule of the form $N =$ first $n \geq 1$ such that $S_n - \lambda^* n \notin (a, b)$, where $\lambda^* = (\lambda^{(1)} - \lambda^{(0)})/\log(\lambda^{(1)}/\lambda^{(0)})$, $b = \log B/\log(\lambda^{(1)}/\lambda^{(0)})$, and $a = \log A/\log(\lambda^{(1)}/\lambda^{(0)})$. For $\lambda^{(0)} = .1808$ and $\lambda^{(1)} = .5249$, $\lambda^* = .3229$. In order to achieve the desired .05 error probabilities the approximations (2.11) suggest the values $A = B = 19$, so $-a = b = 2.7629$.

Moreover, the arguments of this section give approximations to the power function and expected sample size, as follows. For each $\lambda_0 < \lambda^*$ there exists a $\lambda_1 > \lambda^*$ such that $(\lambda_1 - \lambda_0)/\log(\lambda_1/\lambda_0) = \lambda^*$, and conversely. One such pair is, of course, $\lambda_0 = \lambda^{(0)}$ and $\lambda_1 = \lambda^{(1)}$. The given test is also a sequential probability ratio test of λ_0 against λ_1 with error probabilities

$$p_{\lambda_0}(a, b) = P_{\lambda_0}\{S_N - \lambda^* N \geq b\} \cong \frac{1 - (\lambda_1/\lambda_0)^a}{(\lambda_1/\lambda_0)^b - (\lambda_1/\lambda_0)^a},$$

$$1 - p_{\lambda_1}(a, b) = P_{\lambda_1}\{S_N - \lambda^* N \leq a\} \cong \frac{1 - (\lambda_0/\lambda_1)^b}{(\lambda_0/\lambda_1)^a - (\lambda_0/\lambda_1)^b},$$

and expected sample size

$$E_{\lambda_i}(N) \cong (\lambda_i - \lambda^*)^{-1}[(b - a) p_{\lambda_i}(a, b) + a].$$

In particular, for $\lambda_0 = .1808$ and $\lambda_1 = .5249$, $E_{\lambda_0}(N) \cong 17.5$ and $E_{\lambda_1}(N) \cong 12.3$, both of which are considerably less than the fixed sample size of 30. Even for $\lambda = \lambda^*$, the expected sample size is only about 23.6.

In order to evaluate these approximations, Morgan *et al.* tried out the sequential probability ratio test (truncated at a maximum of 60 observations—cf. III.6, especially Table 3.6) on the data they had previously gathered to confirm the adequacy of the Poisson model. The "true" value of λ was determined by counting the clumps per field for 100 fields, and then several sequential tests were run on that group of 100 fields by starting at different points in the sequence. The authors' general conclusions were that the error probabilities were generally slightly smaller and the expected sample sizes slightly larger than the approximations suggest, and that the sequential test did indeed result in an overall savings in sampling cost.

4. Optimality of the Sequential Probability Ratio Test

For testing a simple hypothesis against a simple alternative with independent, identically distributed observations, a sequential probability ratio test is optimal in the sense of minimizing the expected sample size both under H_0 and under H_1 among all tests having no larger error probabilities. To understand this result it is helpful to consider first a degenerate version of the problem in which only one error probability and expected sample size appear.

Suppose $x_1, x_2, \ldots$ are independent observations from f, which may be either f_0 or f_1. Assume also that if f_0 is the true density, sampling costs nothing, and our preferred action is to observe $x_1, x_2, \ldots$ *ad infinitum*. On the other hand, if f_1 is the true density each observation costs a fixed amount, so in this case we want to stop sampling as soon as possible and reject the hypothesis $H_0: f = f_0$.

It may help to imagine that a new drug is being marketed under the hypothesis that its side effects are insignificant. However, physicians prescribing the drug must record and report the side effects. As long as the hypothesis of insignificant side effects ($f = f_0$) remains tenable, no action is required. If it even appears that the level of side effects is unacceptably high ($f = f_1$), this must be announced and the drug withdrawn from use.

A "test" of $H_0: f = f_0$ in the sense described above is a stopping time T. If $T < \infty$, H_0 is rejected. We seek a stopping time for which $P_0\{T < \infty\}$ is acceptably small, say no larger than some prescribed α, and for which $E_1(T)$ is a minimum. A candidate is a "one-sided" sequential probability ratio test, i.e. the stopping time

$$
\begin{aligned}
N &= \text{first } n \geq 1 \text{ such that } l_n \geq B \\
&= \infty \text{ if } l_n < B \text{ for all } n,
\end{aligned} \tag{2.35}
$$

where as usual $l_n = \prod_{k=1}^{n} \{f_1(x_k)/f_0(x_k)\}$. By letting $A \to 0$ in (2.9) and (2.15), or by a direct argument along the lines of Section 2,

$$P_0\{N < \infty\} \leq B^{-1} \tag{2.36}$$

and

$$E_1(N) \cong \log B/E_1[\log\{f_1(x_1)/f_0(x_1)\}]. \tag{2.37}$$

The following proposition gives a lower bound for $E_1(T)$ for any stopping time for which $P_0\{T < \infty\} < 1$. It implies that if (2.36) and (2.37) were actual equalities, the stopping time N of (2.35) would achieve the lower bound and hence would be optimal.

Proposition 2.38. *For any stopping time T with $P_0\{T < \infty\} < 1$,*

$$E_1(T) \geq -\log P_0\{T < \infty\}/E_1[\log\{f_1(x_1)/f_0(x_1)\}].$$

Proof. We may assume $E_1(T) < \infty$; otherwise the result is trivially true. Note that for any random variable y with mean μ, since $e^{\xi} \geq 1 + \xi$, $Ee^{(y-\mu)} \geq 1 + E(y - \mu) = 1$ and hence $Ee^{y} \geq e^{Ey}$. The following computation uses Propositions 2.24 and 2.18 (Wald's identities):

Jamaru

$$
\begin{aligned}
P_0\{T < \infty\} &= E_1 \exp\left(-\sum_{1}^{T} \log\{f_1(x_k)/f_0(x_k)\}\right) \\
&\geq \exp\left(-E_1\left[\sum_{1}^{T} \log\{f_1(x_k)/f_0(x_k)\}\right]\right) \\
&= \exp(-E_1 T\, E_1 \log\{f_1(x_1)/f_0(x_1)\}),
\end{aligned}
$$

which immediately implies the proposition, since $E_1[\log\{f_1(x_1)/f_0(x_1)\}] > 0$. □

Now consider a conventional (sequential) test of $H_0: f = f_0$ against $H_1: f = f_1$ with error probabilities $\alpha = P_0\{\text{Reject } H_0\}$ and $\beta = P_1\{\text{Accept } H_0\}$. For a sequential probability ratio test we have the approximate relations (2.21) and (2.22) between the expected sample sizes and the error probabilities. Theorem 2.39 generalizes Proposition 2.38 in asserting that these expected sample sizes are approximately minimal.

Theorem 2.39. *Let T be the stopping-time of any test of $H_0: f = f_0$ against $H_1: f = f_1$ with error probabilities α, β $(0 < \alpha < 1, 0 < \beta < 1)$. Assume $E_i(T) < \infty$ $(i = 0, 1)$. Then*

$$E_1(T) \geq \mu_1^{-1}\left\{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta\log\left(\frac{\beta}{1-\alpha}\right)\right\}$$

and

$$E_0(T) \geq \mu_0^{-1}\left\{\alpha\log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right)\right\},$$

where

$$\mu_i = E_i[\log\{f_1(x_1)/f_0(x_1)\}] \qquad (i = 0, 1).$$

Proof. Let $R = \{\text{Reject } H_0\}$, $R^c = \{\text{Accept } H_0\}$. As in the proof of Proposition 2.38, by Wald's likelihood ratio identity (2.24)

$$
\begin{aligned}
\alpha = P_0(R) &= E_1\left\{\prod_{1}^{T} \frac{f_0(x_k)}{f_1(x_k)}; R\right\} \\
&= E_1\{e^{-\log l_T} | R\} P_1(R) \geq \exp[-E_1(\log l_T | R)](1-\beta) \\
&= \exp[-E_1(\log l_T; R)/(1-\beta)](1-\beta).
\end{aligned}
$$

Taking logarithms yields

$$(1-\beta)\log\left(\frac{\alpha}{1-\beta}\right) \geq -E_1(\log l_T; R).$$

A similar calculation gives

$$\beta \log\left(\frac{1-\alpha}{\beta}\right) \geq -E_1(\log l_T; R^c),$$

so by addition and Wald's identity

$$(1-\beta)\log\left(\frac{\alpha}{1-\beta}\right) + \beta\log\left(\frac{1-\alpha}{\beta}\right) \geq -E_1(\log l_T) = -\mu_1 E_1(T),$$

which is equivalent to the first assertion of the proposition, since $\mu_1 > 0$. The second assertion is proved similarly. □

Since in general no precise meaning is attached to the approximate equalities (2.21) and (2.22), we have not really proved the optimality of the sequential probability ratio test. Indeed a complete proof is quite difficult and involves the introduction of several auxiliary concepts (see, for example, Ferguson (1967), p. 365). For the very special symmetric Bernoulli case of Example 2.6, the approximations (2.21) and (2.22) are actual equalities, so the preceding discussion contains a completely rigorous proof.

5. Criticism of the Sequential Probability Ratio Test and the Anscombe–Doeblin Theorem

Although the optimality of the sequential probability ratio test is a remarkably strong property in some respects, it applies only to simple hypotheses. For applications involving composite hypotheses the test has noteworthy deficiencies. One is the open continuation region, which leads occasionally to very large sample sizes, especially when $E\{\log[f_1(x_1)/f_0(x_1)]\} \cong 0$. Another is the difficulty associated with estimating a parameter when the data are obtained from a sequential probability ratio test.

The first of these problems can be treated in principle by truncating the stopping rule to take no more than some maximum number of observations. The analysis of such a test is more difficult, but no new statistical concepts are involved (see Chapter III).

The problem of estimation exists to some extent with all sequential tests. If one wants to stop sampling as soon as it is possible to tell in which of two subsets of the parameter space a parameter lies, there presumably are cases where the amount of data necessary to make this rather coarse distinction is inadequate for estimation. A possible solution is to enforce artificially a larger sample size when estimation is of interest. Investigation of the issue is complicated by the fact that even in very simple cases, when one would ordinarily consider unbiased estimators, sequentially stopped versions of the estimators are biased, and their sampling distributions can be quite complicated.

The problems of estimation following sequential tests are discussed in detail in Chapters III and IV. Here we prove a crude but useful result, which shows

that randomly stopped averages are asymptotically normal under quite general conditions.

Theorem 2.40 (Anscombe, 1952, Doeblin, 1938). *Let $x_1, x_2, \ldots$ be independent and identically distributed with mean μ and variance $\sigma^2 \in (0, \infty)$. Let $S_n = \sum_{k=1}^n x_k$. Suppose M_c, $c \geq 0$, are positive integer valued random variables such that for some constants $m_c \to \infty$, $M_c/m_c \xrightarrow{P} 1$. Then as $c \to \infty$*

$$P\{M_c^{-1/2}(S_{M_c} - \mu M_c) \leq x\} \to \Phi(x/\sigma),$$

where Φ denotes the standard normal distribution function.

Remark 2.41. Consider the symmetric sequential probability ratio test of (2.5) for the mean of a normal distribution. Assume $\mu > 0$. As $b \to \infty$, it follows from (2.30) that with probability close to one

$$S_{N-1} < b \leq S_N.$$

Divide by N and note that by the strong law of large numbers $N^{-1}S_{\tilde{N}} \to \mu$ *with probability one, where* $\tilde{N}$ is either N or $N-1$. Hence as $b \to \infty$

$$\mu b^{-1} N \xrightarrow{P} 1. \tag{2.42}$$

It follows from Theorem 2.40 that $N^{-1/2}(S_N - N\mu)$ is approximately normally distributed, and $N^{-1}S_N \pm 1.645N^{-1/2}$ is an approximate 90% confidence interval for μ for large b. Unfortunately this approximation is very poor for moderate values of b (see III.4, 6).

Proof of Theorem 2.40. Without loss of generality assume that $\mu = 0$ and $\sigma^2 = 1$. Also recall that if Z_n converges in law to Z, $\xi_n \xrightarrow{P} 1$, and $\eta_n \xrightarrow{P} 0$, then $\xi_n Z_n + \eta_n$ converges in law to Z (cf. Cramér, 1946, p. 254). Hence, since

$$M^{-1/2}S_M = (m/M)^{1/2}m^{-1/2}S_m + (m/M)^{1/2}m^{-1/2}(S_M - S_m),$$

and $m^{-1/2}S_m$ is asymptotically normally distributed by the central limit theorem, it suffices to show

$$m^{-1/2}(S_M - S_m) \xrightarrow{P} 0. \tag{2.43}$$

Let $\varepsilon, \delta \in (0, 1)$. Let $m_1 = m(1-\delta)$, $m_2 = m(1+\delta)$. Then

$$\{m^{-1/2}|S_M - S_m| > \varepsilon\} \subset \{|m^{-1}M - 1| > \delta\} \cup \left\{\max_{m_1 \leq n \leq m_2} |S_n - S_{m_1}| > m^{1/2}\varepsilon/2\right\}.$$

By hypothesis $P\{|m^{-1}M - 1| > \delta\} \to 0$ for all $\delta > 0$. Also by Kolmogorov's inequality

$$P\left\{\max_{m_1 \leq n \leq m_2} |S_n - S_{m_1}| > m^{1/2}\varepsilon/2\right\} \leq 4(m_2 - m_1)/m\varepsilon^2 \leq 8\delta/\varepsilon^2.$$

Hence

$$\limsup_{c\to 0} P\{m^{-1/2}|S_M - S_m| > \varepsilon\} \le 8\delta/\varepsilon^2.$$

Since δ can be made arbitrarily small, this proves (2.43) and hence the Theorem. □

Remark. The preceding proof is deceptively simple. It seems remarkably difficult by comparison to obtain any more accurate approximation, e.g. an Edgeworth type asymptotic expansion, even for fairly simple stopping rules.

6. Cusum Procedures

Imagine a process which produces a potentially infinite sequence of observations $x_1, x_2, \ldots$. Initially the process is "in control" in the sense that an observer is satisfied to record the x's without taking any action. At some unknown time ν the process changes and becomes "out of control." The observer would like to infer from the x's that this change has taken place and take appropriate action "as soon as possible" after time ν.

To give this problem a simple, precise formulation, assume that the x_i are independent random variables and that for some $\nu \ge 1$, $x_1, x_2, \ldots, x_{\nu-1}$ have the probability density function f_0, whereas $x_\nu, x_{\nu+1}, \ldots$ have the probability density function f_1.

Let P_ν denote probability when the change from f_0 to f_1 occurs at the νth observation, $\nu = 1, 2, \ldots$; and let P_0 denote probability when there is no change, i.e. $\nu = \infty$, so $x_1, x_2, \ldots$ are independent and indentically distributed with probability density function f_0. We seek a stopping rule τ which makes the P_ν distributions of $(\tau - \nu)^+$ stochastically small, $\nu \ge 1$, subject to the constraint that the P_0 distribution of τ be stochastically large. A simple formal requirement is to minimize

$$\sup_{\nu\ge 1} E_\nu(\tau - \nu + 1 | \tau \ge \nu) \tag{2.44}$$

subject to

$$E_0\tau \ge B \tag{2.45}$$

for some given (large) constant B.

An *ad hoc* proposal to solve this problem approximately is the following. Suppose that $x_1, \ldots, x_n$ have been observed. Consider for $1 \le \nu \le n$ the hypotheses H_ν that $x_1, \ldots, x_{\nu-1}$ are distributed according to f_0 and $x_\nu, \ldots, x_n$ according to f_1, and H_0 the hypothesis of no change. The log likelihood ratio for testing H_ν against H_0 is $\sum_{k=\nu}^n \log\{f_1(x_k)/f_0(x_k)\}$. For testing the composite hypothesis that at least one of the H_ν hold $(1 \le \nu \le n)$ against H_0, the log likelihood ratio statistic is

$$\max_{0\le k\le n} (\tilde S_n - \tilde S_k) = \tilde S_n - \min_{0\le k\le n} \tilde S_k, \tag{2.46}$$

where $\tilde S_n = \sum_{j=1}^n \log[f_1(x_j)/f_0(x_j)]$. An intuitively appealing stopping rule based on (2.46) is

$$\tau = \inf\left\{n: \tilde S_n - \min_{0\le j\le n} \tilde S_j \ge b\right\}. \tag{2.47}$$

Note that $\tilde S_n - \min_{0\le k\le n} \tilde S_k$ measures the current height of the random walk $\tilde S_j$, $j = 0, 1, \ldots$ above its minimum value. Whenever the random walk establishes a new minimum, i.e. $\tilde S_n = \min_{0\le k\le n} \tilde S_k$, the process forgets its past and starts over in the sense that for all $j \ge 0$, $\tilde S_{n+j} - \min_{0\le k\le n+j} \tilde S_k = \tilde S_{n+j} - \tilde S_n - \min_{0\le k\le j}(\tilde S_{n+k} - \tilde S_n)$.

This renewal property has several important consequences. First, it implies that for τ defined by (2.47), $\sup_{\nu\ge 1} E_\nu(\tau - \nu + 1|\tau \ge \nu) = E_1\tau$, because at all times after $\nu - 1$ the process (2.46) must be at least as large as if there had been a renewal at $\nu - 1$. Hence to evaluate (2.44) and (2.45) one must calculate $E_\nu(\tau)$ for $\nu = 0, 1$, and for each of these "extreme" cases $x_1, x_2, \ldots$ are identically distributed. ~~Second, it means that~~ τ can be defined in terms of a sequence of sequential probability ratio tests as follows. Let

$$N_1 = \inf\{n: \tilde S_n \notin (0, b)\}. \tag{2.48}$$

If $\tilde S_{N_1} \ge b$, then $\tau = N_1$. Otherwise $\tilde S_{N_1} = \min_{0\le k\le N_1} \tilde S_k$ and we define

$$N_2 = \inf\{n: n \ge 1, \tilde S_{N_1+n} - \tilde S_{N_1} \notin (0, b)\}.$$

If $\tilde S_{N_1+N_2} - \tilde S_{N_1} \ge b$, then $\tau = N_1 + N_2$. Otherwise $\tilde S_{N_1+N_2} \le \tilde S_{N_1}$, and $\tilde S_{N_1+N_2} = \min_{0\le k\le N_1+N_2} \tilde S_k$. In general let

$$N_k = \inf\{n: n \ge 1, \tilde S_{N_1+\cdots+N_{k-1}+n} - \tilde S_{N_1+\cdots+N_{k-1}} \notin (0, b)\}. \tag{2.49}$$

It is easy to see that

$$\tau = N_1 + \cdots + N_M, \tag{2.50}$$

where

$$M = \inf\{k: \tilde S_{N_1+\cdots+N_k} - \tilde S_{N_1+\cdots+N_{k-1}} \ge b\}. \tag{2.51}$$

(See Figure 2.2.) Hence under any probability P which makes $x_1, x_2, \ldots$ independent and identically distributed, in particular for $P = P_1$ or P_0, Wald's identity (2.18) and (2.50) yield

$$E\tau = EN_1 EM.$$

Moreover, by (2.51), M is geometrically distributed, with $E(M) = 1/P\{\tilde S_{N_1} \ge b\}$. As a consequence we obtain the ~~basic~~ identity

$$E(\tau) = E(N_1)/P(\tilde S_{N_1} \ge b), \tag{2.52}$$

which expresses $E(\tau)$ in terms of the ~~expected sample~~ size and error probability of a single sequential probability ratio test having lower (log) boundary $a = 0$.

For (2.52) the Wald approximations of Section 2 yield the expression 0/0 when $a = 0$, but it is possible to give an approximation by evaluating the ratio

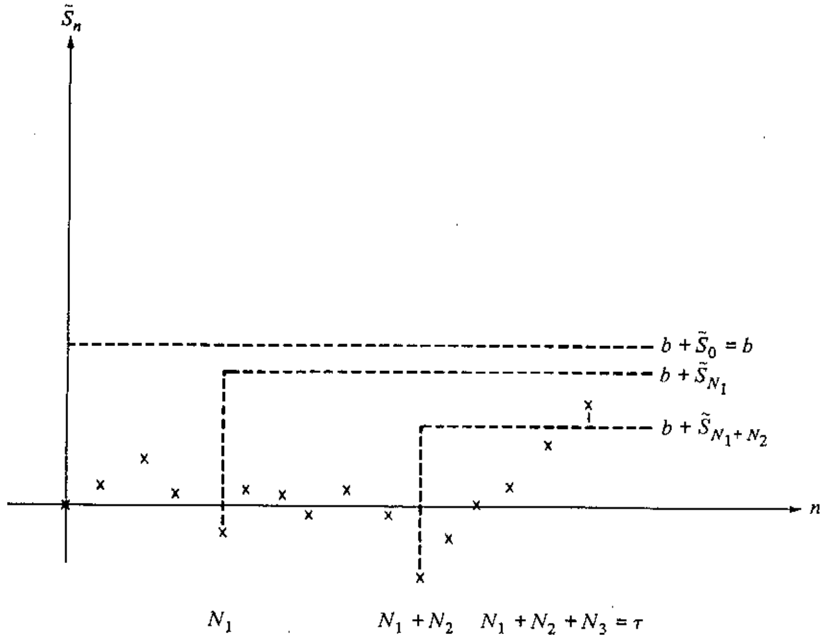

Figure 2.2.

on the right hand side for arbitrary negative a and letting $a \to 0$. The results of these calculations, ~~using (2.12), (2.15)~~ and (2.16), are

(2.53) $\boxed{E_0(\tau) = |e^b - b - 1|/|\mu_0|}$ and $E_1(\tau) = (e^{-b} + b - 1)/\mu_1$,

where

$$\mu_i = \int \{\log\{f_1(x)/f_0(x)\}\} f_i(x)\,dx \qquad (i = 0, 1).$$

If f_0 and f_1 belong to a one-parameter exponential family of distributions, as in Section 3, it is possible to obtain an analogous approximation for the expectation of τ when $x_1, x_2, \ldots$ are independent and identically distributed according to a fixed distribution in that family. Specifically, let $z(x) = \log f_1(x)/f_0(x)$ and assume that the P_θ distribution of $z_n = z(x_n)$ is of the form

(2.54) $$P_\theta\{z_n \in dx\} = dF_\theta(x) = \exp\{\theta z(x) - \psi(\theta)\}\,dF(x)$$

relative to some fixed distribution function $F(x)$. It is convenient to assume that F is standardized to satisfy $\int z(x)\,dF(x) = 0$, so $\psi'(0) = 0$. This can be accomplished by a change of origin on the θ axis. Let θ_0 and θ_1 denote *conjugate* pairs of θ values defined by $\theta_0 < 0 < \theta_1$ and $\psi(\theta_0) = \psi(\theta_1)$. See the middle graph in Figure 2.1.

Now suppose $\theta \neq 0$. If $\theta < 0$, it is denoted by θ_0 and its conjugate by θ_1.

Otherwise it is denoted by θ_1 and its conjugate by θ_0. The stopping rule (2.47) can be written

(2.55) $$\tau = \inf\left\{n: \Delta\tilde{S}_n - \min_{0 \le k \le n} \Delta\tilde{S}_k \ge b'\right\},$$

where $\Delta = \theta_1 - \theta_0$ and $b' = \Delta b$. Since by (2.54)

$$\Delta\tilde{S}_n = \sum_1^n \log\{dF_{\theta_1}(x_k)/dF_{\theta_0}(x_k)\},$$

(2.55) is again of the form (2.47), but with b' in place of b. It follows immediately from (2.53) (in different, but self-evident notation) that

(2.56) $$E_{\theta_i}(\tau) \cong |\exp[(-1)^i \Delta b] - (-1)^i \Delta b - 1|/\Delta|\psi'(\theta_i)|.$$

Taking a limit as $\theta_i \to 0$ yields

$$E_0(\tau) \cong b^2 \psi''(0).$$

Consider the special case $f_i(x) = \phi(x + (-1)^i/2)$ and $dF(x) = \phi(x)\,dx$, so $z(x) = x$ and the family of distributions (2.54) is just the normal family with mean θ and variance 1. In this case (2.56) specializes to

(2.57) $$E_\theta(\tau) \cong |\exp(-2\theta b) + 2\theta b - 1|/(2\theta^2).$$

Unfortunately these approximations are not especially accurate. For example for $b = 6$, (2.57) gives $E_{-.4}(\tau) \cong 362$ and $E_{.4}(\tau) \cong 11.9$. Van Dobben de Bruyn (1968) has calculated these quantities numerically and has obtained 940 and 14.9, respectively.

Often it is possible to improve upon (2.56), and for the normal case the improvement is very similar to that suggested in Section 3. If the right hand side of (2.57) is denoted by $e(b)$, the improved approximation for $E_\theta(\tau)$ is $e(b + 2 \cdot 0.583)$. For $b = 6$ as above, this gives $E_{-.4}(\tau) \cong 944$ and $E_{.4}(\tau) \cong 14.8$. See X.3 for justification and generalization of this approximation.

The preceding discussion is concerned with a change in distribution from F_0 to F_1. Suppose now that $F_i = F_{\theta_i}$ for $\{F_\theta\}$ the exponential family (2.54). (Note that θ_0 and θ_1 no longer have the meaning of the preceding paragraphs.) If $\theta_1 > \theta_0$ ($\theta_1 < \theta_0$), the stopping rule (2.47) can be used to signal a change from θ_0 in the direction $\theta > \theta_0$ ($\theta < \theta_0$). In order to signal a change in either direction one can splice together a pair of one-sided stopping rules as follows.

Suppose that the process is in control if $\theta = \theta_0$ and that (2.54) has been normalized so that $\theta_0 = 0 = \psi(\theta_0) = \psi'(\theta_0)$. (This is always possible by a linear transformation.) To detect a change in θ in either direction, let $\theta_1 < 0 < \theta_2$ and define

(2.58) $$\tau_i = \inf\left\{n: \theta_i S_n - n\psi(\theta_i) - \min_{0 \le k \le n} [\theta_i S_k - k\psi(\theta_i)] \ge b_i\right\},$$

where $S_n = x_1 + \cdots + x_n$. The stopping rule τ_i is of the form (2.47) for detecting a change from $\theta_0 = 0$ in the direction of θ_i, $i = 1, 2$. To detect a change in either direction let $\tau = \min(\tau_1, \tau_2)$.

In a number of special cases there is a simple relation among $E\tau$, $E\tau_1$, and $E\tau_2$. Here E denotes expectation under any probability P which makes $x_1, x_2, \ldots$ independent and identically distributed. It is shown in the somewhat technical Lemma 2.64 below that if

$$|\theta_1|^{-1}\psi(\theta_1) + \theta_2^{-1}\psi(\theta_2) \geq ||\theta_1^{-1}|b_1 - \theta_2^{-1}b_2| \tag{2.59}$$

then

$$\tau = \tau_i \Rightarrow \theta_j S_\tau - \tau\psi(\theta_j) = \min_{0\leq k\leq \tau} [\theta_j S_k - k\psi(\theta_j)] \qquad (j \neq i) \tag{2.60}$$

(see Problem 2.20). Obviously for $j \neq i$

$$\tau_i = \tau + I_{\{\tau=\tau_j\}}(\tau_i - \tau).$$

By (2.60) and the renewal property of $\tilde{S}_n - \min_{0\leq k\leq n} \tilde{S}_k$ described above, the conditional distribution of $\tau_i - \tau$ given $\tau = \tau_j$ is the same as the unconditional distribution of τ_i $(i \neq j)$. Hence for $i \neq j$,

$$E\tau_i = E\tau + E(\tau_i - \tau; \tau = \tau_j) = E\tau + P\{\tau = \tau_j\}E\tau_i.$$

Since $P\{\tau = \tau_1\} + P\{\tau = \tau_2\} = 1$, one can solve these two equations for $E\tau$ to obtain

$$(E\tau)^{-1} = (E\tau_1)^{-1} + (E\tau_2)^{-1}. \tag{2.61}$$

Summarizing this argument yields the following result.

Theorem 2.62. *Let $x_1, x_2, \ldots$ be independent and identically distributed. Let τ_i $(i = 1, 2)$ be defined by* (2.58) *and $\tau = \min(\tau_1, \tau_2)$. If* (2.59) *holds then $E\tau$, $E\tau_1$, and $E\tau_2$ satisfy the relation* (2.61).

Example 2.63. Wilson *et al.* (1979) have used cusum techniques to minitor the quality of radioimmunoassays. A laboratory makes repeated assays in the form of an average of 2 or 3 independent measurements of a concentration in a plasma. In order to maintain quality, a plasma of known concentration is occasionally submitted to be assayed. From these assays of known concentration one obtains observations $x_1, x_2, \ldots$ which are assumed to be independent and normally distributed with mean 0 and (known) variance σ_0^2— as long as the procedure remains in control. A two-sided cusum procedure can be used to detect a change in the mean in either direction from its target value of 0.

It is also important to detect a change in the variance σ^2 from σ_0^2 to some larger value. However, the target value σ_0^2 is a much fuzzier concept than the target value of 0 for the mean error. Presumably σ_0^2 is determined from prior experimentation, but it may be revised as experience accumulates. Given σ_0^2, one has associated with each x_n an independent y_n; and when the process is under control, $y_1/\sigma_0^2, y_2/\sigma_0^2, \ldots$ are independent and have a χ^2 distribution with one or two degrees of freedom, according as the original assays involve 2 or 3 measurements.

For an approximation to the average run length of a one-sided cusum procedure to detect an increase in σ^2, see Problem 2.17.

Lemma 2.64. *If* (2.59) *holds, then* (2.60) *follows.*

Proof. Suppose $\tau = \tau_2 = n$ and that l denotes the largest k, $0 \leq k \leq n$ such that

$$\theta_2 S_k - k\psi(\theta_2) = \min_{0\leq j\leq n} [\theta_2 S_j - j\psi(\theta_2)].$$

Then

$$S_n - n\psi(\theta_2)/\theta_2 - [S_l - l\psi(\theta_2)/\theta_2] \geq b_2/\theta_2 \tag{2.65}$$

but

$$S_i - i\psi(\theta_2)/\theta_2 - \min_{0\leq j\leq i} [S_j - j\psi(\theta_2)/\theta_2] < b_2/\theta_2 \tag{2.66}$$

for all $i < n$. Since $\theta_1 < 0$, (2.60) is equivalent to

$$S_n - n\psi(\theta_1)/\theta_1 = \max_{0\leq j\leq n} [S_j - j\psi(\theta_1)/\theta_1]. \tag{2.67}$$

For any $l \leq j \leq n$

$$\begin{aligned} S_n - n\psi(\theta_1)/\theta_1 - [S_j - j\psi(\theta_1)/\theta_1] &= S_n - n\psi(\theta_2)/\theta_2 - (S_j - j\psi(\theta_2)/\theta_2) \\ &\quad + (n - j)[\psi(\theta_2)/\theta_2 - \psi(\theta_1)/\theta_1] > 0, \end{aligned}$$

because by the definition of l, (2.65), and (2.66)

$$S_n - n\psi(\theta_2)/\theta_2 - S_j + j\psi(\theta_2)/\theta_2 \geq 0,$$

and by the normalization $\theta_0 = 0 = \psi(\theta_0) = \psi'(\theta_0)$ and the convexity of ψ, $\psi(\theta_i) \geq 0$ for $i = 1, 2$. Hence if (2.67) is not satisfied there exists $0 \leq j < l$ such that

$$S_n - n\psi(\theta_1)/\theta_1 < S_j - j\psi(\theta_1)/\theta_1. \tag{2.68}$$

By (2.68) and (2.65)

$$\begin{aligned} S_l - l\psi(\theta_1)/\theta_1 &- [S_j - j\psi(\theta_1)/\theta_1] \\ &< S_l - l\psi(\theta_1)/\theta_1 - [S_n - n\psi(\theta_1)/\theta_1] \\ &= S_l - l\psi(\theta_2)/\theta_2 - [S_n - n\psi(\theta_2)/\theta_2] \\ &\quad - (n - l)[\psi(\theta_2)/\theta_2 - \psi(\theta_1)/\theta_1] \\ &\leq -b_2/\theta_2 - [\psi(\theta_2)/\theta_2 + \psi(\theta_1)/|\theta_1|] \leq b_1/\theta_1, \end{aligned}$$

where the last inequality follows from (2.59). But then $\tau \leq \tau_1 \leq l$, contradicting the hypothesis that $\tau = n > l$. □

Remarks. Condition (2.59) can be interpreted as a measure of symmetry. It is trivially satisfied if $\theta_1 = -\theta_2$ and $b_1 = b_2$. Lemma 2.64 was first proved at this level of generality by van Dobben de Bruyn (1968), who also pointed out that (2.59) is necessary for (2.60).

Problems

2.1.* Let $x_1, x_2, \ldots$ be independent with probability density function $f_\theta(x) = \theta e^{-\theta x}$ for $x \geq 0$ and $= 0$ otherwise. Consider a "one-sided" sequential probability ratio test of H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$, where $\theta_1 < \theta_0$ (cf. (2.35)). Argue that by the lack of memory property of the exponential distribution, the P_{θ_1} distribution of $\log l_N - \log B$ is exponential. Hence $P_{\theta_0}\{N < \infty\} = \theta_1/B\theta_0$ and $E_{\theta_1}(N) = [\log B + \theta_0\theta_1^{-1} - 1]/[\theta_0\theta_1^{-1} - 1 - \log \theta_0\theta_1^{-1}]$. Note that for $\theta_0 \cong 1.5\theta_1$, say, there is a considerable discrepancy between this result and the Wald approximation, $P_{\theta_0}\{N < \infty\} \cong B^{-1}$.

2.2. Investigate the numerical accuracy of the modifications of (2.30) and (2.31) suggested in Section 3. One possibility is to compare results of the modified approximations with numerically computed values given by Barraclough and Page (1959) or Kemp (1958). Another is to make comparisons with simulated values using the technique of Remark 2.26 to increase the accuracy of the simulations.

2.3. For $i = 1, 2$, let $x_{i1}, x_{i2}, \ldots$ be independent Bernoulli variables with $P\{x_{ij} = 1\} = p_i$, $P\{x_{ij} = 0\} = q_i = 1 - p_i$ for $j = 1, 2, \ldots$. Assume that the x_{1j} and x_{2j} are also independent and that observations are made in pairs (x_{11}, x_{21}), (x_{12}, x_{22}), $\ldots$. Suppose $p \neq \frac{1}{2}$ and $p + q = 1$. Find a sequential probability ratio test of H_0: $p_1 = p$, $p_2 = q$ against H_1: $p_1 = q$, $p_2 = p$. Calculate the error probability and expected sample size for arbitrary p_1, p_2 (a) by using the theory developed in the text and (b) by reducing this problem to the one considered in Example 2.6.

2.4. (a) Let $x_1, x_2, \ldots$ be independent random variables with exponential distribution, $P_\lambda\{x_k \in dx\} = \lambda e^{-\lambda x}\,dx$ for $x > 0$. Let $\lambda^{(0)} < \lambda^{(1)}$. Show that a sequential probability ratio test of H_0: $\lambda = \lambda^{(0)}$ against H_1: $\lambda = \lambda^{(1)}$ is defined by a stopping rule of the form

$$N = \text{first } n \geq 1 \qquad \text{such that } n - \lambda^* \sum_1^n x_k \notin [a, b],$$

where $\lambda^* = (\lambda^{(1)} - \lambda^{(0)})/\log(\lambda^{(1)}/\lambda^{(0)})$. Find approximations to the power function and expected sample size. Denote them by $p_\lambda(a, b)$ and $e_\lambda(a, b)$. It is shown in Chapter X that improved approximations similar to those suggested in Section 3 for normal variables are given by $p_\lambda(a - 1, b + \frac{1}{3})$ and $e_\lambda(a - 1, b + \frac{1}{3})$. Compare numerically the approximations in the case $b = \infty$ with the exact results of Problem 1. (Note the difference in notation between the two problems.)

(b) Now suppose that observations are made continuously on a Poisson process with intensity λ. Find a sequential probability ratio test of $\lambda^{(0)}$ against $\lambda^{(1)}$. What is the relation between the "no excess" approximations in this case and in part (a)? Now because the process does not jump over a, improved approximations are given by $p_\lambda(a, b + \frac{1}{3})$ and $\lambda e_\lambda(a, b + \frac{1}{3})$. See X.3 or Hald (1981), p. 266, for a numerical comparison.

2.5. Assume that m items with exponentially distributed lifetimes are simultaneously put on test. Let $x_1, x_2, \ldots, x_m$ denote their failure times, so one observes sequentially $y_1 = \min(x_1, \ldots, x_m)$, y_2 = second smallest of the x_i, etc. Show how the theory of the preceding problem can be used to set up a sequential probability ratio test for this experimental situation, where now, however, there is a maximum number of m observations. (The effect of truncation on sequential probability ratio tests is discussed in III.6.) What is the relation between the expected number of failures and the expected length of the test measured in real time? Discuss also the situation where an item which fails is replaced immediately by a good item, so until the test is terminated there are always m items on test.
Hint: Show that $(m - i + 1)(y_i - y_{i-1})$, $i = 1, 2, \ldots, m$ $(y_0 = 0)$ are independent and exponentially distributed and that the likelihood function after observing $y_1, \ldots, y_k$ can be expressed in terms of $\Sigma(m - i + 1)(y_i - y_{i-1})$, $i = 1, 2, \ldots, k$.

2.6. Prove Proposition 2.19. *Hint:* Suppose $\delta > 0$ is such that $P\{y_1 \geq \delta\} \geq \delta$ and let $r > (b - a)/\delta$. Note that $P\{y_1 + \cdots + y_r > b - a\} \geq \delta^r$. Let $a \leq 0 \leq b$ and compare M with the "geometric" random variable

$$\tilde{M} = \text{first value } mr \ (m \geq 1) \qquad \text{such that} \sum_{i=(m-1)r+1}^{mr} y_i > b - a.$$

2.7. Suppose that l_n is the likelihood ratio of $z_1, \ldots, z_n$ under P_1 relative to P_0. Show that if $P_0\{l_n > 0\} = 1$, then l_n^{-1} is the likelihood ratio of $z_1, \ldots, z_n$ under P_0 relative to P_1 (cf. (2.23)). Note that it is unnecessary to assume $P_1\{l_n > 0\} = 1$. Why?

2.8. Use Proposition 2.24 and a symmetry argument to show that the right hand side of (2.31) is actually a lower bound for $E_\mu(N)$.

2.9. Show that the Wald approximation for $E_0(N)$ for the stopping rule (2.5) in the symmetric normal case is $E_0(N) \cong b^2$. Use this result but with b replaced by $b + .583$ (cf. Section 3) to show that even for error probabilities as small as .01, $E_0(N)$ is smaller than the sample size of a competing fixed sample test.

2.10.* Use Wald's identity (2.18) to derive an approximation to $P_f\{l_N \geq B\}$ when $E_f\{\log l_1\} = 0$.

2.11.† Prove Wald's identity for the second moment: Let $y_1, y_2, \ldots$ be independent and identically distributed with $Ey_1 = 0$ and $\sigma^2 = Ey_1^2 < \infty$. If T is any stopping time with $ET < \infty$, then $E[(\sum_1^T y_i)^2] = \sigma^2 ET$.
Hint: For a *bounded* stopping time T this can be proved using the representation $\sum_1^T y_i = \sum_1^\infty I_{\{T \geq i\}} y_i$ as in the proof of (2.18). To extend this to a general stopping time with $ET < \infty$, show that

$$\lim_{m<n,\, m\to\infty,\, n\to\infty} E\left[\left(\sum_{i=T\wedge m+1}^{T\wedge n} y_i\right)^2\right] = 0,$$

and use Fatou's lemma.

2.12. Use Problem 11 to obtain an approximation for $E_f(N)$ when $E_f\{\log l_1\} = 0$. Check that this answer is consistent with the special case in Problem 9.

2.13. Let $x_1, x_2, \ldots$ be independent with probability density function of the form

$$f_\theta(x) = \exp[\theta x - \psi(\theta)]f(x),$$

where f is some given density function. Assume that the parameter space is an open interval $(\underline{\theta}, \overline{\theta})$ and that $\psi(\theta) \to \infty$ as $\theta \to \overline{\theta}$ or $\underline{\theta}$. Consider a sequential probability ratio test of $H_0: \theta = \theta^{(0)}$ against $H_1: \theta = \theta^{(1)}$. Find approximations for the power function and the expected sample size function of the test. Specialize the results to the Bernoulli and Normal examples considered in the text.

Remark. Although one can apply the general theory developed in the text to solve this problem, it is probably more instructive to proceed from first principles with the ideas given in the chapter as guidelines. An important and particularly simple special case occurs when $\psi(\theta^{(0)}) = \psi(\theta^{(1)})$.

2.14.† For detecting a change of distribution, the following alternative to (2.47) has been suggested by Shiryayev (1963) and Roberts (1966):

$$T = \inf\left\{n: \sum_{k=0}^{n-1} \prod_{i=k+1}^{n} \frac{f_1(x_i)}{f_0(x_i)} \geq B\right\}.$$

Show that

$$E_\infty T = E_\infty\left\{\sum_{0}^{T-1} \prod_{k+1}^{T} \frac{f_1(x_i)}{f_0(x_i)}\right\},$$

so by neglecting the excess over the boundary $E_\infty T \cong B$. Obtaining a reasonable approximation to $E_1 T$ is difficult, although it is at least intuitively apparent that a crude approximation is given by

$$E_1 T \sim \log B/E_1[\log\{f_1(x_1)/f_0(x_1)\}] \qquad \text{as } B \to \infty.$$

2.15. Prove the following generalization of Theorem 2.40 (Anscombe, 1952). Suppose that $Y_n \xrightarrow{\mathscr{L}} Y$ and that the sequence $\{Y_n\}$ is slowly changing in the sense that for each $\varepsilon > 0$ there exists $\delta > 0$ such that for all sufficiently large m

$$P\left\{\max_{m \le n < m(1+\delta)} |Y_n - Y_m| \geq \varepsilon\right\} < \varepsilon. \tag{2.69}$$

Suppose $\{M(c), c \geq 0\}$ is a family of positive integer valued random variables such that for some constants $\rho(c) \to \infty$, $M(c)/\rho(c) \xrightarrow{P} 1$. Then $Y_{M(c)} \xrightarrow{\mathscr{L}} Y$. (See IX.2 for a condition similar to (2.69) in a different context.)

2.16.* Show that under the condition (2.32)

$$E_f(N) \cong [\log(BA^{-1})P_f\{l_N \geq B\} + \log A]/E_f\{\log[f_1(x_1)/f_0(x_1)]\}$$

(where $P_f\{l_N \geq B\}$ is given approximately by (2.33) in the case $\theta_1 > 0$ and by $(A^{\theta_1} - 1)/(A^{\theta_1} - B^{\theta_1})$ in the case $\theta_1 < 0$).

2.17. Let $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and 0 otherwise. Consider the problem of detecting a change in distribution from $\lambda = 1$ to $\lambda = \lambda^{(1)} < 1$. (This is equivalent to detecting an increase in σ^2 in Example 2.63 when two degrees of freedom are available from each assay for estimating σ^2.) Let $\lambda^* = (1 - \lambda^{(1)})/\log(1/\lambda^{(1)})$, $\theta = 1 - \lambda/\lambda^*$, and $\psi(\theta) = -(\theta + \log(1 - \theta))$. Show that the stopping rule N_1 of (2.48) equals $\inf\{n: \sum_1^n (\lambda^* x_k - 1) \notin (0, b')\}$, where $b' = b/\log(1/\lambda^{(1)})$, and that the distribution of $z_n' = \lambda^* x_n - 1$ in the form (2.54) is given by

$$\exp[\theta x - \psi(\theta)]\exp[-(x+1)]\,dx$$

for $x \geq -1$. Evaluate the approximation (2.56) for arbitrary λ. (If this approximation is denoted by $e(b')$, an improved approximation taking excess over the boundary into account is $e(b' + \frac{4}{3})$ (see X.2).)

2.18. Discuss Problem 2.17 if one seeks to detect a change from $\lambda = 1$ to $\lambda = \lambda^{(1)} > 1$. Now $\lambda^* = (\lambda^{(1)} - 1)/\log \lambda^{(1)}$, $\theta = \lambda/\lambda^* - 1$, $\psi(\theta) = \theta - \log(1 + \theta)$, $b' = b/\log \lambda^{(1)}$, and $z_n' = 1 - \lambda^* x_n$. (The correction for excess over the boundary is again of the form $e(b' + \frac{4}{3})$. For a more complete discussion of this problem which contains a different approximation to the expected run length and some numerical examples, see Lorden and Eisenberger, 1973.)

2.19.* Consider the general model of Section 1 and an arbitrary sequential test of the simple hypotheses $H_0: f_n = f_{0n}, n = 1, 2, \ldots$ against $H_1: f_n = f_{1n}, n = 1, 2, \ldots$. Let T denote the stopping time, A the acceptance region, and R the rejection region of the test. Let P denote probability and E expectation when f_n $(n = 1, 2, \ldots)$ is the true sequence of joint densities, not necessarily either f_{0n} or f_{1n}. Suppose that $P\{T < \infty\} = 1$. Show that the total error probability $\alpha + \beta$ satisfies $\alpha + \beta \geq E\{\min(f_{0T}/f_T), (f_{1T}/f_T)\}$, and that there is equality for a sequential probability ratio test provided $A < 1 < B$. (For an application of this inequality see III.7.)

2.20. For the normal distribution (i.e. (2.54) with $dF(x) = \phi(x)\,dx$, $\psi(\theta) = \theta^2/2$, $z(x) = x$), demonstrate (2.60) geometrically by an appropriate picture in the symmetric case $|\theta_1| = \theta_2$, $b_1 = b_2$.