A RETROSPECTIVE OF WALD'S SEQUENTIAL ANALYSIS—ITS RELATION
TO CHANGE-POINT DETECTION AND SEQUENTIAL CLINICAL TRIALS

BY

David Siegmund

TECHNICAL REPORT NO. 9

December 1992

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT DMS91-04432

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
A Retrospective of Wald's Sequential Analysis—Its Relation to Change-point Detection and Sequential Clinical Trials

By

David Siegmund

Technical Report No. 9

December 1992

Prepared Under the Auspices

of

National Science Foundation Grant DMS91-04432

Department of Statistics

Stanford University

Stanford, California
A Retrospective of Wald’s Sequential Analysis—Its Relation to Change-point Detection and Sequential Clinical Trials
by
David Siegmund

1. Introduction. Motivated by problems of ordnance testing and quality control during World War II, Wald created the field of Sequential Analysis. A stylistic comparison of his 1947 book, Sequential Analysis (organized into 11 Chapters about basic statistical concepts supplemented by a 50 page mathematical appendix), with some of his uncompromisingly mathematical papers on decision theory, large sample theory, etc. confirms the applied origins of this research.

In the intervening years Sequential Analysis has undergone substantial development. The purpose of this paper is to discuss Wald’s original contributions, particularly as they relate to recent research. The paper is necessarily limited in scope and reflects to a considerable extent my personal interest in two problems which have also engaged the attention of applied statisticians over an extended period of time. These are the problems of (i) sequential change-point detection and (ii) sequential clinical trials. The list of references does not purport to be comprehensive. I also do not attempt to assess the role of British contributions to the beginnings of the subject except to note the existence of the paper by Anscombe (1946), which was also addressed to problems of quality control.

One specific theme of the paper will be the evolution of the very elegant mathematical techniques introduced by Wald in his analysis of the sequential probability ratio test. A broader theme will be a description of the change in focus that has occurred as a response to criticism of the sequential probability ratio test and the stimulus of new applications.

The paper is organized as follows. Section 2 contains a brief description of Wald’s analysis of the sequential probability ratio test together with some criticisms of the test. The CUSUM test and its relation to the sequential probability ratio test are discussed in Section 3. Section 4 introduces an idea from renewal theory which has provided the basis for substantially more sophisticated theoretical analysis of a variety of sequential statistical procedures. Section 5 contains a discussion of sequential clinical trials, especially the rapid theoretical advances of the late 70’s and early 80’s. A recent result that is related to many of the ideas of Sections 2-5 is described in Section 6. Section 7 contains some miscellaneous comments.

2. The Sequential Probability Ratio Test.

Wald started working on the subject of Sequential Analysis early in 1943. The immediate stimulus was questions about ordnance testing addressed by (Naval) Captain G. L. Schuyler to W. Allen Wallis, who together with Milton Friedman brought the problem
to Wald’s attention (Wald (1947, p.2), Wallis (1980)). During the war Wald’s work was classified. It was subsequently published in Wald (1945, 1947).

Wald’s sequential probability ratio test is originally formulated as a test between a simple hypothesis and a simple alternative. Suppose that \(X_1, X_2, \cdots\) are random variables and for every positive integer \(n\), \(X_1, \cdots, X_n\) have the joint probability density function \(f_n\). We are interested in testing whether \(f_n\) is one of two completely specified sequences of joint densities: \(f_{0,n}\) or \(f_{1,n}\). Let \(L_n = f_{1,n}/f_{0,n}\) denote the likelihood ratio statistic for comparing the two hypotheses on the basis of the first \(n\) observations, where we assume for simplicity that the corresponding joint distributions are mutually absolutely continuous. Let \(A < B\) (usually \(0 < A < 1 < B\)) and define the stopping time

\[N = \inf\{n : L_n \leq A \text{ or } L_n \geq B\}.\]  

(2.1)

The test proceeds by taking observations \(X_1, X_2, \cdots\) one at a time until the stopping time \(N\) defined in (2.1). The sequence \(f_{0,n}\) is rejected if \(N < \infty\) and \(L_N \geq B\). It is accepted if \(N < \infty\) and \(L_N \leq A\). In the event that \(N = \infty\), sampling does not terminate and no conclusion is reached. (A question, naturally, is whether this third contingency actually can happen. In the important case that the observations are independent and identically distributed, a result of Stein (1946) shows that its probability is 0 unless \(f_0\) and \(f_1\) are the same density function.)

Wald’s analysis of the sequential probability ratio test was based on two identities: his likelihood ratio identity, and his identity for a randomly stopped sum. The general structure of these identities was recognized by Wald and generalized by Doob (1953) into powerful tools for the development of martingale theory.

One reasonably general form of Wald’s likelihood ratio identity is as follows. For \(i = 0, 1\) let \(P_i\) denote the joint distribution of the infinite sequence \(X_1, X_2, \cdots\) when the finite dimensional joint densities are \(f_{i,n}, n = 1, 2, \cdots\). Then for any stopping time \(T\), not necessarily of the form (2.1), and for any event \(A\) defined in terms of the random variables \(X_1, \cdots, X_T\) observed until the stopping time \(T\),

\[P_i(A \cap \{T < \infty\}) = E_0[L_T; A \cap \{T < \infty\}].\]  

(2.2)

The proof is immediate, by summing over \(n\) the obvious relation \(P_i(A \cap \{T = n\}) = E_0[L_n; A \cap \{T = n\}]\).

If we assume that the stopping time in (2.1) satisfies \(P_i\{N < \infty\} = 1\) for \(i = 0, 1\), an immediate application of (2.2) yields the following inequality between the Type I and Type II error probabilities, \(\alpha\) and \(\beta\), of a sequential probability ratio test:

\[\beta = P_1(L_N \leq A) = E_0(L_N; L_N \leq A) \leq A(1 - \alpha).\]  

(2.3)
A similar calculation with the roles of \( P_0 \) and \( P_1 \) interchanged yields

\[
\alpha = P_0 \{ L_N \geq B \} = E_1(L_N^{-1}; L_N \geq B) \leq (1 - \beta)/B. \tag{2.4}
\]

Wald suggested that since the only source of the inequalities in (2.3) and (2.4) is the inequalities in (2.1), which can be, but in general are not equalities, it would be reasonable to regard (2.3) and (2.4) as approximate equalities and solve them simultaneously to find approximations for \( \alpha \) and \( \beta \) in terms of \( A \) and \( B \). Although the resulting approximations are striking in their simplicity, they often involve a relative error on the order of 50%. Methods to improve the approximation are described below.

Wald’s identity for randomly stopped sums says that if \( Y_1, Y_2, \ldots \) are independent, identically distributed random variables with finite expectation \( \mu \), and if \( T \) is a stopping time with finite expectation, then

\[
E(Y_1 + \cdots + Y_T) = \mu E(T). \tag{2.5}
\]

In the case that the \( X_i \) above are independent and identically distributed, so \( f_{i,n} = f_i(X_1) \cdots f_i(X_n) \) is a product and \( \log(L_n) \) is a sum of independent identically distributed random variables under both \( P_0 \) and \( P_1 \), Wald combined (2.5) and the heuristic ‘no excess over the boundaries’ approximation suggested above to obtain expressions for \( E_i(N) \) in terms of \( A \) and \( B \).

Wald (1947, p. 199, especially footnote 4) used these same basic tools to conjecture an optimality property of the sequential probability ratio test: in the case of independent identically distributed observations, among all tests having no larger Type I and Type II error probabilities, the sequential probability ratio test minimizes the expected sample size under both \( P_0 \) and \( P_1 \). The proof of this result is extraordinarily ingenious and subtle. An important step in the argument is to establish a decision theoretic version of the problem and show that a sequential probability ratio test is the Bayes solution to the new problem. Two seminal papers, Wald and Wolfowitz (1948) and Arrow, Blackwell, and Girshick (1949), use the same general logic, although each seems incomplete in some part of the argument.

An indication of the extent of the influence of Wald’s analysis of the sequential probability ratio test is found in Morton (1955), where a standard for proving genetic linkage is established that persists until the present, even though its rationale is not entirely appropriate.

The sequential probability ratio test has limitations which affect its use in statistical practice. Examples are (i) its unbounded random sample size, which makes it impossible to put a definite upper bound on the cost or duration of an experiment and (ii) the unknown and difficult to study effect that the random stopping time has on parameter
estimates. One must also remember that the sequential probability ratio test was designed for testing a simple hypothesis against a simple alternative, and its optimality property refers to this artificial situation. Although the test can be used for problems involving composite hypotheses by selecting simple surrogate hypotheses to represent the composite hypotheses (indeed one of Wald’s most ingenious techniques was invented to study this issue), it nevertheless seems plausible that one can do better by developing a test tailored to the features of the problem at hand.

These limitations have stimulated subsequent research, a small part of which is discussed in Section 5.


Detecting when the quality of a production process changes is an important part of statistical process control. A classical reference is Shewhart (1931). An important contribution is the so-called CUSUM test, which originated with Page (1954) and is closely related to the sequential probability ratio test. A rather different approach which goes back to Girshick and Rubin (1952) and Shiryayev (1963) is also discussed below.

We imagine a production process which may be in control or out of control. We cannot observe the state of the process directly, but can observe independent measurements of the quality of the output of the process. If the process is in control, we would like to let the process continue without interruption. If it is out of control, we must stop the process and correct it.

A mathematical model is as follows. We can observe independent random variables \( X_1, X_2, \ldots \). For some unknown value \( j \), often called the change-point, the observations \( X_1, \ldots, X_j \) have probability density function \( f_0 \), while \( X_{j+1}, \ldots \) have density function \( f_1 \). Either or both of the \( f_i \) may involve unknown parameters. Let \( P_j(E_j) \) denote probability (expectation) when the change-point is \( j \), \( 0 \leq j \leq \infty \). We seek a stopping rule \( T \) with the properties that \( E_{\infty}(T) \) is large, say greater than a prescribed constant \( C \), while \( E_j[T - j | T \geq j] \) is small in some suitable sense. The first condition requires that the process be stopped infrequently in cases where no change has occurred; the second requires that it be stopped soon after a change occurs. Since either of these expectations may depend on unknown parameters, and the second depends on \( j \) in any case, our problem is not yet clearly specified. Often the dependence of the second expectation on \( j \) is removed by considering \( \sup_{j \geq 0} E_j[T - j | T \geq j] \) or \( \lim_{j \to -\infty} E_j[T - j | T \geq j] \) as the criterion to be minimized. For many stopping rules one wants to consider, including (3.1) defined below, the supremum is attained at \( j = 0 \). A useful example to illustrate basic concepts has \( X_1, \ldots, X_j \) normally distributed with a known mean value \( \mu_0 \) and unit variance, while \( X_{j+1}, \ldots \) are normally distributed with mean value \( \mu_1 > \mu_0 \) and unit variance. The value of \( \mu_1 \) is typically unknown, although a nominal value may be assumed for the purpose of defining a procedure. For simplicity we consider only the case that \( \mu_0 \) is known.
The procedure suggested by Page (1954) is the following. We define a "score," $Z_n$, and put $S_n = Z_1 + \cdots + Z_n$. In the artificial case that $f_0$ and $f_1$ are completely specified, an appropriate score is the log likelihood ratio, $Z_k = \log[f_1(X_k)/f_0(X_k)]$. This possibility was mentioned by Page, but he did not use it in his examples, and its importance was not generally appreciated until somewhat later (Lorden (1970), see also Moustakides (1986)). In the case that $f_1$ contains an unknown parameter, one can use an arbitrary value of the parameter to form the likelihood ratio. The stopping rule proposed by Page is

$$M = \inf\{n : S_n - \min_{0 \leq k \leq n} S_k \geq b\}. \quad (3.1)$$

In the case that the score is the log likelihood ratio indicated above, the stopping rule (3.1) amounts to using a sequence of sequential probability ratio tests. A new test is started with each new observation and is continued until either $f_0$ is accepted (with $\log A$ in (2.1) set equal to 0) or rejected (with $\log B$ in (2.1) set equal to $b$). The CUSUM test stops as soon as one of the sequential probability ratio tests stops. From this relation it follows by a simple renewal argument that for any probability $P$ under which the observations are independent and identically distributed, in particular for the two extreme cases, $P = P_\infty$ and $P = P_0$,

$$E(M) = P\{S_N \geq b\}/E(N), \quad (3.2)$$

where $N$ is defined by (2.1) with $A = 1, B = \exp(b)$.

Although Wald's approximations described in Section 2 do not apply directly to the case $A = 1$, by considering the limit of the ratio in (3.2) as $A \to 1$, one obtains approximations for the average run length of a CUSUM test for the extreme values $j = \infty$ and $j = 0$. For the simple case that $f_0$ is standard normal and $f_1$ is normal with mean $\delta$ and unit variance, the approximation is given by

$$E_{0,\mu}(M) \approx 2(2\mu - \delta)^{-2}[c(2\mu/\delta - 1) + \exp[-c(2\mu/\delta - 1)] - 1]. \quad (3.3)$$

In (3.3) the subscript 0, $\mu$ on the expectation indicates that a change of the mean from 0 to $\mu$ occurs at $j = 0$. When $\mu = 0$ there is no change, i.e., $j = \infty$. The value $\mu = \delta$ specifies that the actual change at $j = 0$ is equal to the nominal change assumed for the purpose of defining the stopping rule (3.1). Note than when $\mu < \delta/2$, in particular when $\mu = 0$, (3.3) grows exponentially as a function of $c$, whereas it grows linearly if $\mu > \delta/2$.

It was quickly recognized that the approximation (3.3) is quite poor in the particularly important case $j = \infty (\mu = 0)$, with the result that considerable effort was invested in numerical methods for calculating these and related quantities (e.g., van Dobben de Bruyn (1968), Brook and Evans (1972)). In the next section we shall see that the approximation (3.3) can be considerably improved.
An outstanding recent contribution to the theory of CUSUM tests is Moustakides' (1986) proof of an optimality property similar to the optimality property of the sequential probability ratio test.

A different approach to change-point detection, which goes back to Girshick and Rubin (1952) and Shirayev (1963), is to formulate a decision theoretic problem and find the Bayes' solution. Girshick and Rubin's formulation is conceptually somewhat different than Shirayev's although they arrive at essentially the same procedure, which stops the process as soon as the posterior probability that a change has occurred at some time before the present exceeds a threshold \( B \). For a 'flat' prior distribution on the value of the change-point, the stopping rule becomes

\[
M_1 = \inf \{ n : \sum_{k=0}^{n-1} \exp(S_n - S_k) \geq B \},
\]

(3.4)

where \( S_n = \sum_{i=1}^{n} \log[f_1(X_i)/f_0(X_i)] \).

These decision theoretic formulations, like that found in the Wald-Wolfowitz (1948) and Arrow-Blackwell-Girshick (1949) discussions of the sequential probability ratio test, assume that the probability density functions \( f_0 \) and \( f_1 \) are completely known. Hence the optimality property is of questionable relevance when these densities depend on unknown nuisance parameters.

For the problem of detecting a change in a normal mean from a known initial value when the variance is known, Roberts (1966) compares by Monte Carlo methods a number of different procedures, including a CUSUM test defined by (3.1) and the approximate Bayes procedure having the stopping rule (3.4). Pollak and Siegmund (1984) make an analytic comparison of these two procedures when testing for a change of drift in a Brownian motion process. In both papers the two procedures are found to behave similarly.


Let \( Z_1, Z_2, \cdots \) be independent and identically distributed with mean value \( \mu = EZ_1 < 0 \). Let \( S_n = Z_1 + \cdots + Z_n \), and for \( b > 0 \) define

\[
\tau = \inf \{ n : S_n \geq b \},
\]

(4.1)

where it is understood that the \( \inf \) of the empty set is \( +\infty \). The probability \( P\{\tau < +\infty\} \) is of interest in queueing theory, where it gives the stationary waiting time distribution in a single server queue, and in insurance risk theory, where it gives the probability of ruin of a risk reserve process. See Feller (1972, Chapters VI and XII.) In the special case that \( Z_k = \log[f_1(X_k)/f_0(X_k)] \), where the \( X \)'s are independent, identically distributed random variables and the \( f_i \) are probability density functions, this probability equals the
The probability of rejecting $f_0$ when using the sequential probability ratio test defined by (2.1) in the degenerate special case $A = 0, B = \exp(b)$. The inequality (2.4) becomes

$$P_0\{\tau < +\infty\} \leq B^{-1}. \quad (4.2)$$

In applications outside statistics, it will rarely be the case that $Z_k$ is given directly in the form of a log likelihood ratio. However, by the technique of exponential imbedding, which was ingeniously exploited by Wald (1947), we can assume that it is up to to a constant of proportionality which can be absorbed into the definition of $b$, whenever the $Z$'s have a finite moment generating function. Feller (1972, Chapter XII) and Siegmund (1985, Chapter 8) give thorough discussions of this technique. Then $E_1(Z_k) > 0$, so $P_1\{\tau < +\infty\} = 1$, and hence the equality between the second and third terms of (2.4) becomes

$$P_0\{\tau < +\infty\} = \exp(-b)E_1 \exp[-(S_\tau - b)]. \quad (4.3)$$

Wald's analysis was to replace the expectation on the right hand side of (4.3) by the obvious upper bound of 1 and hence to obtain (4.2).

The following argument gives a much more precise result. If the $Z$'s were positive random variables, the random variable $S_\tau - b$ appearing on the right hand side of (4.3) would be the familiar residual life time encountered in renewal theory, for which the renewal theorem would give the limiting distribution as $b \to \infty$ (e.g., Feller (1972, Chapter XI)). Since the $Z$'s have positive mean, we can define the stopping times, i.e., the successive stopping times at which the random walk attains new (strict) maxima. The random variable $S_\tau - b$ equals the residual life time at age $b$ in the renewal process with life distribution equal to the $P_1$-distribution of $S_1$, where

$$\tau_+ = \inf\{n : S_n > 0\}$$

is the first ladder time. It follows from (4.3) and the renewal theorem that as $b \to \infty$

$$P_0\{\tau < +\infty\} \sim K \exp(-b), \quad (4.4)$$

where $K$ is a complicated, but explicit constant defined in terms of the $P_1$-distribution of $S_\tau$. For the special case where the $Z$'s are proportional to $N(\mu,1)$, the constant $K$ is $v(2||\mu||)$, where $v$ is defined in terms of the standard normal distribution function $\Phi$ by the equation

$$v(x) = 2x^{-2} \exp[-2 \sum_1^{\infty} \Phi(-x n^{1/2} / 2)] \quad (x > 0). \quad (4.5)$$

A convenient approximation for small $x$ is given by

$$v(x) = \exp(-\rho x) + o(x^2) \text{ as } x \to 0, \quad (4.6)$$
where $\rho$ is a numerical constant having a value about equal to 0.583. See Siegmund (1985, Chapters 8 and 10) for a detailed discussion.

Remark. It is difficult to say who was the first to use the preceding renewal argument. It appears, for example, in Spitzer (1964, Chapter IV), but in an analytic form which makes the argument appear less promising for generalization than it is. Feller (1972, First Edition 1966) gives the argument a probabilistic flavor only slightly different from that given here. See also Siegmund (1975, 1985), where the method is applied to the sequential probability ratio test and the CUSUM test.

For a statistical example of this line of argument, consider a CUSUM test applied to detect a change in a normal mean, as described in Section 3. Assume that $\mu \to 0$ and $\delta \to 0$ in such a way that the ratio $\mu / \delta$ is fixed. Under these conditions the ratio of the two sides of (3.3) converges to one. A more precise asymptotic result is

$$E_{0,\mu}(M) = 2(2\mu - \delta)^{-2}[(c + 2\rho \delta)(2\mu / \delta - 1) + \exp[-(c + 2\rho \delta)(2\mu / \delta - 1)] - 1] + o(1), \quad (4.7)$$

where $\rho$ is the numerical constant in (4.6) and the double subscript on the expectation has the same meaning as in (3.3).

Remarks. (i) The asymptotic normalization of (4.5) is different from that of (4.4), and some additional arguments are required for the proof. The natural level of generality for a result like (4.5) is a one parameter exponential family. Siegmund (1985, Theorem 10.16) gives such a result with an error which in the present notation would be $o(\delta^{-1})$. A more careful examination of that argument indicates that the error term is $O(1)$ and can be evaluated in terms of moments of ladder height variables. Because of symmetry this term vanishes for normally distributed observations, so the error term is $o(1)$. A more general result along these lines was stated without proof by Pollak and Siegmund (1986, unpublished).

(ii) The approximation (4.7) is of the same form as (3.3), but the threshold $c$ has been replaced by $c + 2\rho \delta$ to account for the expected excess by which the process exceeds the threshold when a crossing takes place.

(iii) The approximation given by Siegmund (1975) is more directly in the spirit of (4.4) in the sense that the distribution is held fixed while $c \to \infty$. It has the disadvantages compared to (4.7) that it is complicated and requires numerical computation to evaluate, and it contains several second order terms which do not appear to have a precise mathematical justification. On the other hand, it reduces to (4.7) when $\mu$ and $\delta$ are close to 0, and experience shows it is more accurate when they are not close to 0. For example, for the special case of normal $X$'s and $\mu = 0$, the dominant term (which has a simple justification
based on (3.2) and the renewal theoretic argument given above) is given by

$$E_{0,0}(M) \sim 2 \exp(c)/[\delta \nu(\delta)]^2 \text{ as } c \to \infty. \quad (4.8)$$

Note that when $c$ is large (4.8) and (4.7) will give similar numerical results to the extent that (4.6) is an accurate approximation. To the extent that it is not, (4.8) can be more accurate. See below.

For a simple numerical comparison of (4.7) and the Wald type approximation, (3.3), suppose $c = 4, \delta = 0.8, \mu = 0$. The value of (3.3) is 155; the value of (4.7) is 415. Van Dobben de Bruyn (1968) has evaluated this quantity numerically and gives the value of 414, which he says is accurate to one unit in the last significant figure given. For an example where (4.8) is more accurate than (4.7), suppose that $c = 5, \delta = 2, \mu = 0$. Then (4.6) gives 760, (4.8) gives 723, and van Dobben de Bruyn gives 716. In this case the nonexponential terms in (4.7) make only a minor contribution to the total approximation, so it is unimportant that the analogous terms have been omitted from (4.8) (although including them would yield an even better approximation).

A result similar to (4.8), but much deeper mathematically, is Pollak’s (1987) approximation to the expectation of the stopping time in (3.4) when $j = \infty$. In the special case that $f_0$ is standard normal and $f_1$ is $N(\delta,1)$, Pollak’s result is

$$E_{0,0}(M_1) \sim B/\nu(\delta) \text{ as } B \to \infty,$$

which also provides very accurate numerical approximations.

5. Sequential Clinical Trials. Almost from the beginning of the subject, it was recognized that sequential hypothesis testing might be usefully applied to clinical trials, where data are necessarily collected sequentially and ethical considerations suggest that a trial to compare, say, two treatments should be terminated as soon as it becomes apparent that one of the treatments is better than the other. Early research used essentially the mathematical arguments developed by Wald. After writing a number of papers on sequential clinical trials during the 1950’s, Peter Armitage published the first edition of his book, Sequential Medical Trials, in 1961. Armitage, McPherson, and Rowe (1969) introduced systematic use of high speed numerical computing into the study of sequential clinical trials in order to evaluate operating characteristics of repeated significance tests, for which Wald’s methods are inadequate. These developments led to the second edition of Armitage’s book (1975).

The following simple model has become standard for discussions of basic problems of sequential clinical trials involving two treatments. We assume $X_1, X_2, \cdots$ are independent normally distributed observations with mean $\mu$ and variance 1. The observation $X_n$ gives the difference in the (immediate) responses of two paired subjects, one of whom receives
Treatment A while the other receives Treatment B. To be specific we assume that a large response is desirable, so a positive value of \( \mu \) indicates the superiority of Treatment A while a negative value indicates the superiority of Treatment B. Our first goal is to test the hypothesis of no treatment effect, i.e., \( \mu = 0 \). We would like to do so in such a way that only a small number of observations are required when \(|\mu| >> 0\). In this case one treatment is much better than the other, and the trial should terminate as quickly as possible, to avoid giving the inferior treatment to more patients than necessary.

A repeated significance test is defined as follows. Let \( S_n = X_1 + \cdots + X_n \). Given \( b > 0, m_0 > 0 \) let

\[
T = \inf\{n : n \geq m_0, |S_n| > b n^{1/2}\}.
\]  

(5.1)

For an arbitrary positive integer \( m \geq m_0 \), stop sampling at \( \min(T, m) \) and reject the hypothesis that \( \mu = 0 \) if and only if \( T \leq m \). The power function of this test is \( P_\mu \{T \leq m\} \).

The question of approximating the significance level of the test had been posed in the early 1950's by Robbins (1952) in a discussion of a still earlier paper by Feller (1940) addressed to statistical aspects of research on extra sensory perception. Robbins gave a simple upper bound for the significance level, which was computed numerically by Armitage, McPherson, and Rowe (1969).

A series of theoretical advances began in 1976. Woodroofe (1976) and Lai and Siegmund (1977), using quite different methods, gave essentially the same approximation to the significance level of a repeated significance test. Woodroofe's method is related to that of Anscombe's (1953) classical paper on sequential estimation, whereas Lai and Siegmund use Wald's likelihood ratio identity along the lines of Robbins (1970) in conjunction with a generalized renewal theorem. Siegmund (1977) studied different approximations to the significance level and used a variant of Woodroofe's method to approximate the power of a repeated significance test. Pocock (1977) suggested it would be more feasible to take data in groups and showed that even as few as two or three groups secured many of the advantages of a purely sequential test. Siegmund (1978) discussed confidence intervals for \( \mu \) following a repeated significance test, and (although anticipated by Haybittle (1971)) suggested a modification of the repeated significance test that in some important respects has more favorable operating characteristics. Woodroofe (1978), Jones and Whitehead (1979), Lalley (1983), Tsiatis (1981), Sellke and Siegmund (1983), Woodroofe and Takahashi (1982) and others studied a number of related problems. O'Brien and Fleming (1979) (anticipated by Miller (1970)) proposed a different class of sequential tests, which they studied by Monte Carlo methods. Their test can be studied theoretically via the approximations given by Siegmund (1979, 1985).

By the end of this period there existed a variety of theoretical and computational methods for evaluating sequential tests, and for addressing many questions which previously had not been adequately treated. In particular truncated sequential tests defined
by possibly curved boundaries, and estimation following sequential tests had become subjects of active investigation. Monographs summarizing these developments are Woodroffe (1982), Whitehead (1983), and Siegmund (1985).

Since the Lai-Siegmund (1977) method for approximating the error probability of a repeated significance test builds directly on Wald’s methods along the lines of the discussion of Section 4, I indicate briefly the nature of the argument. Let $P_\mu$ denote the joint distribution of $X_1, X_2, \ldots$ when the true value of their expectation is $\mu$. Introduce the measure $Q = \int P_\mu d\mu/(2\pi)^{1/2}$. The range of integration is the entire real line. Although $Q$ is not a probability measure, that fact is not important to the following calculations. Let $L_n$ denote the likelihood ratio of $X_1, \ldots, X_n$ calculated under $Q$ relative to $P_0$. It is straightforward to see that

$$L_n = \exp(S_n^2/n)/n^{1/2}.$$  \hspace{1cm} (5.2)

It follows from Wald’s likelihood ratio identity, equation (2.2), and the definition of $Q$ that

$$P_0\{T \leq m\} = \int_{\{T \leq m\}} L^{-1}_T dQ$$

$$= b\phi(b) \int E_\mu[(T/b^2)^{1/2} \exp(-(S_T^2/T - b^2)/2); T \leq m] d\mu.$$ \hspace{1cm} (5.3)

The successful application of this identity depends on two observations, one almost trivial and the second less so. Assume that $b \to \infty$ and $m \to \infty$ in such a way that $b/m^{1/2} \to \mu_1$. For simplicity assume that $m_0 = 1$. The easy observation is that for each $\mu \neq 0$, under the measure $P_\mu$ we have $T/b^2 \to \mu^{-2}$ with probability one as $b \to \infty$, and hence $P_\mu\{T \leq m\} \to 0$ or 1 according as $|\mu|$ is less than or greater than $\mu_1$. Hence the right hand side of (5.3) behaves asymptotically like

$$b\phi(b) \int_{|\mu| \geq \mu_1} E_\mu(\exp[-(S_T^2/T - b^2)/2]|\mu^{-1} d\mu.$$ \hspace{1cm} (5.4)

The second observation is that writing (5.1) (with $m_0 = 1$) in the form

$$T = \inf\{n : S_n^2/n > b^2\}$$ \hspace{1cm} (5.5)

puts the random variable $S_T^2/T - b^2$ in (5.4) into the form of an excess over the boundary or residual lifetime for the nonlinear process $S_n^2/n$. Lai and Siegmund showed that for each $\mu \neq 0$ this random variable behaves like the residual lifetime of the random walk $\mu(S_n - n\mu/2)$ and hence has a limiting distribution given by the renewal theorem, as in Section 4. The limit is somewhat complicated and will not be described here, except to say that as a consequence (5.4) asymptotically is of the form

$$2b\phi(b) \int_{\mu > \mu_1} \mu^{-1} \nu(\mu) d\mu,$$ \hspace{1cm} (5.6)
where \( \nu(\mu) \) is defined in (4.5), and given approximately in (4.6).

The approximation (5.6) is extremely accurate, even for \( m = 2 \), as is easily seen by comparing it with exact numerical calculations of Pocock (1977). For example, for \( m = 2, b = 2.178 \) Pocock (1977) has evaluated (5.3) to be 0.05; Table 4.2 of Siegmund (1985) shows that (5.6) is 0.0494. In order to deal with a larger class of related tests Siegmund (1985) suggests a slightly different approximation.

An elegant decision theoretic approach to testing composite statistical hypotheses was developed by Chernoff in a series of papers beginning in 1960 (cf. Chernoff (1972)). For the problem of clinical trials, the formulation described above was criticized by Anscombe (1963) in his review of the first edition of Armitage’s book for not being Bayesian and, more importantly, not decision theoretic. Since that time a number of authors have studied versions of the model suggested by Anscombe (e.g., Lai, Levin, Robbins, Siegmund (1980), Chernoff and Petkau (1981)). In contrast to the situation described in Section 3, where the decision theoretic formulation addressed the same problem and produced procedures which behave very much like the CUSUM test, Anscombe’s decision theoretic model seems designed to promote an essentially different formulation of the problem than the one discussed above. The result is that there has been very little interaction between these two lines of research. For a more thorough discussion see Siegmund (1985, Section 4.6). For a decision theoretic approach employing a different loss function, see Lerche (1986).

6. Detecting a change in distribution using the generalized likelihood ratio statistic. The CUSUM test is designed to detect a change from one completely specified distribution to another, although it is typically applied to problems involving composite hypotheses. For detecting a change in a normal mean Barnard (1959) proposed what amounts to applying the CUSUM idea to the generalized likelihood ratio statistic. This idea is very appealing because it suggests procedures for a wide variety of parametric problems involving nuisance parameters, and has been proposed by others (e.g., Basseville (1988)). It appears, however, that no one has given an approximation to its average run length under the hypothesis of no change in distribution, even in the simplest case.

To describe a recent result of Siegmund and Venkatraman (1992) along these lines, assume that \( X_1, X_2, \ldots \) are independent, \( N(0, 1) \) random variables, and write \( S_n = X_1 + \cdots + X_n \). Given \( b > 0 \) define

\[
T = \inf \{ n : \max_{0 \leq k < n} (n - k)^{-1/2} |S_n - S_k| > b \}. \tag{6.1}
\]

Then as \( b \to \infty \),

\[
E(T) \sim (2\pi)^{1/2} \exp(b^2/2) \int_0^b x\nu^2(x)dx. \tag{6.2}
\]
It is easily seen that the stopping rule (6.1) stands in the same relation to the stopping rule (3.1) of the simple CUSUM test (for the problem of a change in a normal mean with known variance) as the stopping rule (5.5) of a repeated significance test stands in relation to that of a sequential probability ratio test. However, the method of Lai-Siegmund (1977) for giving approximations to the null distribution of (5.5) seems of no use here. Rather one needs a method which can be adapted to processes with a multidimensional time index. One possibility is Siegmund’s (1988) extension of Woodroofe’s (1976) method. From that method one obtains the following preliminary result.

Suppose \( b \to \infty, \ m \to \infty \) in such a way that \( m = cb^2 \) for some \( c > 0 \). Then

\[
P\{T \leq m\} \sim m b \phi(b) \left( \int_{c^{-1/2}}^{\infty} x \nu^2(x) dx - c^{-1} \int_{c^{-1/2}}^{\infty} x^{-1} \nu^2(x) dx \right). \tag{6.3}
\]

A substantial additional argument shows that \( T \) is asymptotically exponentially distributed with mean value given by the right hand side of (6.2) (which appears formally by letting \( c \to \infty \) in (6.3)); and the proof is completed by a simple uniform integrability argument.

**Remark.** Since the integral in (6.2) is convergent at \( \infty \), that result would also be correct if the upper limit in the integral were put equal to \( \infty \). However, there are heuristic reasons for preferring the upper limit of \( b \), which also yields more accurate numerical approximations. See Siegmund and Venkatraman (1992) for a more complete discussion.

7. **Summary.**

I would like to thank the conference organizers for inviting me to prepare this retrospective view of Wald’s contributions to Sequential Analysis. As indicated in the introduction, I have restricted myself to two subjects which have involved an extended dialogue between theoretical and applied statisticians and where Wald’s technical contributions have had a continuing role. I have omitted all reference to large subfields of Sequential Analysis whose origins can be traced to the intellectual climate created by Wald’s work, but to which Wald did not contribute directly. Examples would be the theory of fixed width interval estimation along the lines initiated by Stein (1949) and Anscombe (1953) and problems of adaptive sampling strategy as originally discussed by Robbins (1952). An even more important example, as measured by its subsequent development, is the subject of stochastic control theory, which quickly grew into a field in its own right, but whose origins and essential ideas come directly out of the Wald-Wolfowitz (1948) and Arrow-Blackwell-Girshick (1949) papers on the optimality of the sequential probability ratio test.

Also, since high speed computers were not available to Wald, I have mentioned but have not emphasized the role they have played in subsequent research. It is interesting to note that the mathematical difficulties associated with the evaluation of operating characteristics of even very simple sequential procedures led to early systematic use of numerical
methods. From the research cited above, the interval 1968-1972 can be seen as a time of particularly rapid development of computational methods.

The subject of sequential change-point detection is presently in a period of rapid development, and is changing too quickly to permit even a superficial summary of contemporary research. This comment becomes even more true if one enlarges the subject to include fixed sample change-point problems. A particularly interesting line of recent research stems from Assaf (1988). The discussion in Section 6 was included because of its natural relation to Sections 3, 4, and 5.

By comparison the subject of sequential clinical trials is to some extent in a state of consolidation. In my view the most interesting problems do not involve more sophisticated analyses of the basic two treatment single response design, although there is clearly more to be said, especially for survival data (e.g., Sellke (1988), Gu and Lai (1991)). Other important problems are those concerned with more complex experimental situations, where problem formulation is still an important consideration. Examples are problems of multiple endpoints (e.g., Lin (1991)), repeated measurement designs (Lee and DeMets (1991)), and designs involving more than two treatments (Siegmund (1992)).

References


Page, E. S. (1954). Continuous inspection schemes, Biometrika 41 100-115.


Tsiatis, A. (1981). The asymptotic joint distribution of the efficient scores test for the
proportional hazards model calculated over time, *Biometrika* 68, 311-315.
117-186.
Wald, A. and Wolfowitz J. (1948). Optimum character of the sequential probability ratio
Assoc.* 75, 320-330.
Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials*, Ellis Hor-
wood, Chichester.
Woodrooffe, M. (1976). Frequentist properties of Bayesian sequential tests *Biometrika* 63,
101-110.
Woodrooffe, M. (1978). Large deviations of the likelihood ratio statistic with applications
to sequential testing, *Ann. Statist.* 6, 72-84.
Woodrooffe, M. (1982). *Nonlinear renewal theory in sequential analysis*, Society for In-
dustrial and Applied Mathematics, Philadelphia.
837.