METHODS OF SEQUENTIALLY TESTING COMPOSITE HYPOTHESES
WITH SPECIAL REFERENCE TO THE TWO-SAMPLE PROBLEM

By W. J. Hall
University of North Carolina

Institute of Statistics Mimeo Series No. 441

August 1965

This research was supported by the National Institutes of Health under Grant GM-10397.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.
Correction: pg. 37, line 12. "Test 1a" and "Test 1b" should be "Test 1c" and "Test 1d".

Addendum to Appendix 3:

A PRELIMINARY REPORT ON SOME CALCULATIONS OF RELATIVE
EFFICIENCY OF THE PERMUTATION TEST

\[ e_1 = \frac{\text{approx. ASN of normal theory SPRT under } H_1}{\text{approx. ASN of permutation test under } H_1} = \frac{E_1 L^e}{E_1 L^*} \]

The RHS is the ratio of expected increments in the cumulative sums defining the tests (see pp. 52-3). The hypotheses \( H_1 \) are normal theory hypotheses about the distribution of the observed difference \( Z \) when sampling in pairs (see bottom of pg. 52). The parameter \( h \) is a measure of the difference between \( H_1 \) and \( H_0 \) (middle of pg. 53).

The approximations in \( e_1 \) are due solely to the inaccuracy of the error probabilities, due to ignoring overshoot. The ratio \( e_1 \) is a measure of relative efficiency of the permutation test relative to the normal theory test when both tests are valid (\( H_1 \) true).

It is shown on page 53 that \( e_1 \) tends to 1 as \( h \) tends to 0 — in fact, \( e_1 = 1 - (4h + 0(h^2)) \) — and that the average number of additional observations required by the permutation test is asymptotically \( c_1 \) as \( h \) tends to 0; \( c_1 \) is commonly around two to six (see bottom of pg. 51). Calculations of \( e_1 \) are in progress by Peter Nemenyi.

Calculations indicate that:

1. The efficiency ranges from about 75% or less (in situations requiring very small ASN's) monotonically towards 100 percent;
2. The asymptotic theory is fairly rapidly approached;
3. The extra number of observations required by the permutation test is approximately \( c_1 \) or less—not just asymptotically, but over the range of \( h \) investigated.

<table>
<thead>
<tr>
<th>( h )</th>
<th>2.0</th>
<th>0.5</th>
<th>0.08</th>
<th>0.02</th>
<th>( H_0 ) true</th>
<th>( H_1 ) true</th>
</tr>
</thead>
<tbody>
<tr>
<td>efficiency ( e_1 )</td>
<td>0.749</td>
<td>0.903</td>
<td>0.981</td>
<td>0.995</td>
<td>( e_1 ) = 0.853</td>
<td>0.980</td>
</tr>
<tr>
<td>( 1 - \frac{1}{2} h ) (est. of ( e_1 ))</td>
<td>0.50</td>
<td>0.875</td>
<td>0.980</td>
<td>0.995</td>
<td>( e_1 ) = 0.820</td>
<td>0.980</td>
</tr>
</tbody>
</table>

\[ E_1 N = \begin{cases} \text{perm. test} & 7.1 \quad 23.5 \quad 135.1 \quad 532.4 \\ \text{normal test} & 5.3 \quad 21.2 \quad 132.5 \quad 530.0 \\ \text{difference} & 1.8 \quad 2.3 \quad 2.6 \quad 2.4 \end{cases} \]

\[ E_1 N = \begin{cases} \text{perm. test} & 12.0 \quad 39.9 \quad 229.4 \quad 904.3 \\ \text{normal test} & 9 \quad 36 \quad 225 \quad 900 \\ \text{difference} & 3.0 \quad 3.9 \quad 4.4 \quad 4.3 \end{cases} \]

W. J. Hall
Department of Statistics
Phillips Hall, U. N. C.
Chapel Hill, N. C. - 27515
METHODS OF SEQUENTIALLY TESTING COMPOSITE HYPOTHESES
WITH SPECIAL REFERENCE TO THE TWO-SAMPLE PROBLEM \(^1,2\)

By W. J. Hall

University of North Carolina

Summary: Three methods for developing sequential probability ratio
tests (SPRT's) for testing composite (parametric and non-parametric) hypotheses
are described in Part I: (1) parameter separation, (2) data reduction, and
(3) conditioning. Method (1) is an elaboration and extension of an approach
attributed to Nandi [24] and used in a special context by Girshick [12]; it
is closely related to Fraser-sufficiency [9]. It generalizes naturally in
either of two ways, leading respectively to methods (2) and (3). Method (2)
includes Cox's method [7] and the invariance method of Stein [16] and has been
widely used. Method (3) has not been described previously except in an example
of Wald's [31]. A primary contribution of this paper, which is otherwise
largely expository, is to describe this conditioning method in some generality
and to apply it to develop some new tests for the two-sample problem, namely
sequential permutation tests analogous to the non-sequential tests of Pitman
[27] as described by Lehmann and Stein [23]. All three methods are illustra-
ted by tests for the two-sample problem.

Part II of the paper consists of a listing and brief description of a
number of SPRT's for this two-sample problem: three kinds of normal tests,
some sign tests, some permutation tests, and some rank tests. They include
several standard tests, several recent tests, and several new tests. Some
numerical comparisons of these tests are planned for a subsequent paper.

---

1. Presented at the 125\textsuperscript{th} annual meeting of the American Statistical

2. Research supported by the National Institutes of Health under Grant
   GM-10397.
# Table of Contents

**Summary** ....................................................... 1

**Part I: Three Methods of Sequentially Testing Composite Hypotheses**

Introduction .................................................. 3

Method (1): Parameter Separation ................................ 7

Method (2): Data Reduction ...................................... 13

Method (3): Conditioning ........................................ 18

**Part II: Sequential Probability Ratio Tests for the Two-Sample Problem**

Summary of Part II .............................................. 32

Introduction to the Tests ..................................... 33

Normal Tests .................................................. 36

Normal mean, variance known (Tests 1a-1d) .................. 36

Normal mean, variance unknown (Tests 2a-2b) ............... 37

Normal mean in standard units—t-tests (Tests 3a-3d) .... 38

Sign Tests .................................................. 40

Sign test of differences (Test 4) ............................ 41

Double-dichotomy tests (Tests 5a-5b) ....................... 42

Permutation Tests (Test 6) .................................... 43

Rank Tests (Tests 7a-7f) ..................................... 45

**Appendix 1: ASN of the Extended Double-Dichotomy Test** 49

**Appendix 2: ASN of the Permutation Test** .................. 52

**References** .................................................. 54
PART I: THREE METHODS OF SEQUENTIALLY TESTING COMPOSITE HYPOTHESES

INTRODUCTION

Throughout we are concerned with a sequentially performed experiment. We designate the data potentially available at stage \( n \) by \( X_n \), and the accumulated data from the first \( n \) stages by \( X_{(n)} = (X_1, X_2, \ldots, X_n) \), \( n = 1, 2, \ldots \).

We shall assume the stages are independent, although it would not always be essential, but there is no need for assuming identical distributions, or even identical sample spaces for the successive stages. A family of probability models is assumed for the sequence of data, including two sub-families designated \( H_0 \) and \( H_1 \). We may alternatively describe \( H_0 \) and \( H_1 \) as composite hypotheses about the probability model of the data.

We seek methods for constructing valid sequential probability ratio tests (SPRT's) of these hypotheses. By "valid" we mean that the strength \((\alpha, \beta)\) is correctly prescribed: the probability of incorrectly rejecting \( H_0 \) is never more than \( \alpha \) and the probability of incorrectly accepting \( H_0 \) is never more than \( \beta \)--at least within the limits of approximation that are inherent in SPRT's due to the possibility of overshooting the stopping boundaries. By a SPRT we mean of course a sequential test based on some kind of a joint likelihood ratio with constant termination boundaries. Designating the joint likelihood ratio (l.r.) at stage \( n \) by \( R_n \) (the \( H_1 \) likelihood in the numerator and the \( H_0 \) likelihood in the denominator), the test is: after observing the data at stage \( n \) (\( n = 1, 2, \ldots \)),

- continue to the next stage if \( B < R_n < A \),
- terminate and accept \( H_0 \) if \( R_n \leq B \),
- terminate and reject \( H_0 \) if \( R_n \geq A \).
and \( A = (1-\beta)/\alpha \) and \( B = \beta/(1-\alpha) \). If \( R_n \) is a bonafide joint l.r., then the test has (approximate) strength \((\alpha, \beta)\) \([31]\).\(^3\)

The point about which we have so far been vague is: how can we get a bonafide l.r. for composite hypotheses? That is indeed the subject of Part I of this paper. We shall describe three kinds of situations in which joint l.r.'s, of some useful sort, may be constructed, and exemplify each. Our examples will all be related to the two-sample problem, but of course the methods are not confined to this single area of application. It will be noted that there is some overlap among the methods, and that more than one of them may be required in a single problem; nevertheless, many problems are not amenable to these methods.

In brief, the applicability of the three methods may be summarized as follows: Method (1) is applicable when the l.r., with specific values assigned to the nuisance parameters so that the two hypotheses are simple, is free of the assigned values; two sufficient conditions, each utilizing Fraser-sufficiency \([9]\), are given and they generalize respectively to the other two methods. In method (2), the available data are reduced to a sequence of statistics having probability models completely specified by the two hypotheses, and thereby a l.r. can be constructed; the Stein Theorem of invariance and sufficiency theory \([16]\) (or Cox's factorization theorem \([7]\)) may be useful. In method (3), a sequence of statistics is chosen and fixed in such a way that the hypotheses about the conditional distributions become simple hypotheses, and hence a conditional l.r. can be constructed; the resulting conditional SPRT's (CSPRT's) are readily seen to be valid unconditional tests of the composite hypotheses.

\(^3\) One can assure bounds on the error probabilities, rather than approximate bounds, by using Wald's conservative stopping boundaries \( A = 1/\alpha \) and \( B = \beta \) instead of his recommended ones given above; see paragraph headed strength on page 586 of \([16]\).
We shall not attempt a complete survey of the methodology of sequential composite-hypothesis testing. The methods introduced by Wald ([31], Chapter 4) and by Bernardo [5], in which nuisance parameters are integrated out of the likelihoods, will not be discussed at all; familiar applications of them are also covered by Cox's method or invariance and the latter will be briefly described (see also [19] and [16]). Nandi [24] described certain situations, which we understand to be special cases of methods (1) or (2), in which composite hypotheses may be tested. Girshick([12] and Section 4.2.4 of [31]) described certain kinds of two-sample problems---of the "which of two treatments is the better one" type rather than the "is one treatment better or is there no difference" type treated in the examples here---in which SPRT's of composite hypotheses are possible; this is included in method (1).

Johnson and Leone ([20], Section 16.11) give several examples in which SPRT's of composite hypotheses may be constructed; they may be viewed as examples of what we call method (2). Both Johnson [19] and Hall, Wijsman and Ghosh [16] mention many examples not mentioned here. Large-sample methods of sequentially testing composite hypotheses have recently been given by Cox [8], but are not considered here.

Of course, knowledge of a l.r. \( R_n \) facilitates any generalized sequential probability ratio test (GSPRT)\(^4\) [21], and not just a SPRT. In particular, a fixed sample size test (FSST)\(^5\) may be based on \( R_n \) for fixed \( n \), with specified significance level \( \alpha \), but not necessarily with specified \( \beta \). A critical region

\(^4\) A GSPRT is a test which may be described in an identical manner to the description of a SPRT given earlier, except that the constants \( A \) and \( B \) may be chosen in other ways and may vary with the stage index \( n \).

\(^5\) A l.r. FSST is a GSPRT with \( A_n = B_n = c_n \) for the fixed sample size \( n \), and \( A_m = \infty \), \( B_m = 0 \) for \( m \neq n \).
of the form \( R_n \geq c_n \) where \( c_n \) is chosen to achieve a specified significance level provides such a test. Such a FSST would be most powerful in some sense, being based on a l.r. We shall not consider GSPRT's here, other than SPRT's--that is, we confine attention to the case of constant termination boundaries \( A \) and \( B \). Other types of boundaries have been investigated by various authors, including Armitage [3] and Anderson [1], but largely in special contexts.

Approaches other than those based on a joint l.r. are also omitted from consideration here (e.g., Sec. 15.3 of Wilks [33], Nemenyi, Adelman and Miller [25], and Hall [13]), as are all decision-theoretic methods (except as they overlap with SPRT's).

It should be emphasized that we are not making optimality claims--we are simply seeking some test with pre-assigned strength \((\alpha, \beta)\), and the class of SPRT's seems a likely place to search. We state once more that the methods have non-sequential application as well; in fact, they have been used extensively there, and what we describe here may be viewed as sequential analogs of them.

It will frequently be convenient to use parametric notation and let \( \theta \) index the family of possible models. We can then characterize \( H_i \) as \( \gamma(\theta) = \gamma_i \) \((i=0,1)\). When \( \theta \) may be written in in the form \((\gamma, \eta)\), we may then call \( \gamma \) the parameter of interest and \( \eta \) the nuisance parameter. Thus, \( \gamma \) characterizes the hypotheses to be tested and \( \eta \) indexes each of the families \( H_0 \) and \( H_1 \).

We shall use density function notation, tacitly assuming a suitable dominating measure; in fact, we frequently use conditional densities as well. For our purposes of exposition, we shall act as if we are dealing only with the discrete case so that densities (probability mass functions) and conditional densities (probabilities) can be assumed and manipulated without concern for measure-theoretic regularities. The methods described are certainly valid in the usual continuous model problems as well; the full extent of their validity will not be pursued.
Before turning to the three methods to be surveyed and exemplified in this paper, we wish to remark that monotonicity provides a technique of extending the validity of tests of simple hypotheses to certain kinds of composite hypotheses—a technique which is well-known [31] and which is also useful in extending tests of composite hypotheses. Specifically, if the model is specified by a parameter (vector) \( \theta \), and we test \( \theta = \theta_0 \) vs. \( \theta = \theta_1 \), then, if the OC function (probability of accepting \( H_0 \) as a function of \( \theta \)) of the test is decreasing in a numerical function \( \gamma(\theta) \), the test has strength \((\alpha, \beta)\) not only for testing the hypotheses as given but also for testing the composite hypotheses \( \gamma(\theta) \leq \gamma(\theta_0) \) vs. \( \gamma(\theta) \geq \gamma(\theta_1) \). Simple (and well-known) examples include the SPRT for testing such hypotheses about the normal mean \( \theta = \mu \) when the variance \( \sigma^2 \) is known \( (\gamma(\theta) = \theta, \theta_0 < \theta_1, \) random sampling from a normal population) and the sequential t-test \( (\theta = \mu/\sigma, \gamma(\theta) = \theta \) in the one-sided test and \( \gamma(\theta) = (\mu/\sigma)^2 \) in the two-sided test). Monotonicity proofs are available when the joint l.r. is monotone in \( \gamma \) —both for the case of tests based on independent observations (Lehmann [22]) and more generally (Ghosh [11]). The latter reference covers many of the common examples of sequential tests derived by use of Cox’s theorem or invariance (method (2)), and the former applies to many examples of what we describe as methods (1) and (3) below.

**METHOD (1): PARAMETER SEPARATION**

In this section we assume a parametric formulation with \( \theta = (\gamma, \eta) \), and we assume further that the parameter space is a product space. We can confine attention to the two values \( \gamma_0 \) and \( \gamma_1 \) for \( \gamma \), specified by \( H_0 \) and \( H_1 \), but a larger range is possible too; \( \eta \) is the nuisance parameter with the same range under \( H_0 \) and \( H_1 \).
Suppose \( \eta \) is fixed at \( \eta' \) so that \( H_0 \) and \( H_1 \) become simple hypotheses. When is the l.r. free of \( \eta' \)? It will be seen that a necessary and sufficient condition is provided by the following notion of \( \gamma \) and \( \eta \) being **separable**.

**DEFINITION:** Suppose, for each \( n \) and each \( x_n \), the likelihood of \( x_n \) factors into a function of \( \gamma \) only times a function of \( \eta \) only; we then say that \( \gamma \) and \( \eta \) are separable parameters.

Suppressing \( n \) from the notation, we thus have

\[
f_{\gamma, \eta}(x) = g_{\gamma}(x) h_{\eta}(x)
\]

where \( f_{\gamma, \eta} \) is the density function of \( X_n \). Since the joint likelihood is the product of the marginal likelihoods for the respective stages, the joint likelihood would likewise factor, and the factor depending only on \( \eta \) would cancel out of the l.r. On the other hand, if the l.r. at \((\gamma_1, \eta')\) and \((\gamma_0, \eta')\) is free of \( \eta' \) for all \((\gamma_1 \text{ and } \gamma_0 \text{ under consideration})\), then we can let \( h_{\eta} = f_{\gamma_0, \eta} \) for fixed \( \gamma_0 \) and \( f_{\gamma, \eta} / f_{\gamma_0, \eta} = g_{\gamma} \), which is \( \eta \)-free by assumption, so that \( f_{\gamma, \eta} = g_{\gamma} h_{\eta} \); hence, \( \gamma \) and \( \eta \) are separable. Note also that if \( S_n = s_n(X(n)) \) is a sufficient statistic for \( \theta = (\gamma, \eta) \), then the separability test could equivalently be applied to the distribution of \( S_n \) rather than the distribution of \( X_n \) (for every \( n \)).

In practice, the choice of nuisance parameter \( \eta \) as a function of \( \theta \) is not fixed by the formulation of the hypotheses \( H_0 \) and \( H_1 \). Hence, to determine the applicability of the method one would ordinarily write down the density function of \( X_n \) (or of \( S_n \)) with parameter \( \theta \), factor out that part depending solely on \( \gamma(\theta) \), and determine if the remaining factor depends on \( \theta \) only through some \( \eta = \eta(\theta) \) for which \((\gamma, \eta)\) is one-to-one in \( \theta \) and the range of \( \eta \) does not depend on \( \gamma \). We thus speak of \( \gamma \) being "separated out of \( \theta \)", or simply of " \( \gamma \) being separable."

The consequence of \( \gamma \) being separable is that the nuisance parameter is really not a nuisance! We can act as if \( \eta \) were known, assigning it an arbitrary
value $\eta'$. The resulting SPRT will have strength $(\alpha, \beta)$, whatever the value of $\eta$. Since the test is in effect one of simple hypotheses (for $\eta'$ fixed), all of Wald's methods [31] for investigating its properties are applicable--for every fixed value of $\eta$. Thus, under conditions of simple random sampling, the test terminates with certainty, Wald's OC and ASN approximation methods are applicable, and the test has minimal ASN under $H_0$ and $H_1$ uniformly in $\eta$. Also, assuming simple random sampling, the approximate OC (Wald's method) may be seen to be free of the nuisance parameter $\eta$, but the ASN may depend on $\eta$ since it depends on the distribution of the l.r. and that distribution may depend on $\eta$ (but see below). By way of contrast, for a FSST of level $\alpha$ based on the l.r. $R_n$, the critical number $c_n$ may depend on $\eta$ so that we cannot conclude either that the test is uniformly (in $\eta$) most powerful, nor that its power (or OC) is $\eta$-free.

Trivial examples of parameter separation occur whenever $X$ is of the form $(U, V)$ where $U$ and $V$ are independent and the $U$-distribution is governed by $\gamma$ and the $V$-distribution by $\eta$. Obviously, $\eta$ (and $v$) drop out of the l.r., and might just as well be ignored from the outset. Other examples are provided by the class of problems considered by Girshick [12] (not pursued here$^6$); we will give other examples below, relevant to the two-sample problem.

---

6. In the notation of Sec. 1.8 of [12], we hold $v(\theta_1) - v(\theta_2)$ fixed, at 2c (say), let $\eta(\theta_1, \theta_2) = (v(\theta_1) + v(\theta_2), w(\theta_1) + w(\theta_2)) = (\eta_1, \eta_2)$, and let $\gamma(\theta_1, \theta_2) =$ sign of $(\theta_2 - \theta_1)$. Then the joint density of the pair $(X_1, X_2)$, with $u(x_1) = u_1, (u_1^2 + u_2^2)/2 = \bar{u}$ and $r(\bar{x}_1) = r_1$, is found to be $\exp[\frac{(u_2 - u_1)\gamma + \bar{u}}{2} (\eta_1^2 + \eta_2^2 + r_1^2 + r_2^2)]$. Hence, $\gamma$ and $\eta$ are separable. The hypotheses to be tested are $\gamma = \gamma_0 = +1$ vs. $\gamma = \gamma_1 = -1$, or equivalently $v(\theta_1) - v(\theta_2) = 2c$ or $-2c$. Girshick noted some of the properties of his test, as described above in a more general context.
Two sufficient conditions for parameter separation will now be given, both involving the concept of Fraser-sufficiency [9].

**DEFINITION:** Suppose $X$ has a probability model with parameter $(\gamma, \eta)$ (with a product-space range). Let $T = t(X)$ and $U = u(X)$. Then $T$ is Fraser-sufficient for $\gamma$ if its distribution depends only on $\gamma$ and if the conditional distribution of $X$ given $T$ depends only on $\eta$. Likewise, $U$ is Fraser-sufficient for $\eta$ if its distribution depends only on $\eta$ and if the conditional distribution of $X$ given $U$ depends only on $\gamma$.

If such a statistic $T_n = t_n(X_n)$, or $t_n(X_{(n)})$, exists which is Fraser-sufficient for $\gamma$, then $\gamma$ may certainly be separated, for, writing the density of $X_n$ (or $X_{(n)}$) as the product of the marginal density of $T_n$ and the conditional density (probability)\(^7\) of $X_n$ given $T_n$ achieves the required factorization. Then, in this case, the density of $X_{(n)}$ (and of $X_n$) is the product of the density of a statistic $T_n$ depending only on $\gamma$ and a factor free of $\gamma$. Thus, one might just as well confine attention to the statistic $T_n$ instead of the actual data, and use its l.r. for testing $H_0$ vs. $H_1$. This situation was described by Nandi [24]\(^8\). We pursue this conclusion (with or without separability) as method (2) in the next section. Since the l.r. is identical with the l.r. for the statistic $T_n$, which has a model governed solely by $\gamma$, any GSPRT would have properties depending solely on $\gamma$—specifically, the ASN is $\gamma$-free. (That such a test would have minimal ASN under $H_0$ and $H_1$ has been noted by J. K. Ghosh; see page 543 in [16].) Before exemplifying this situation, we turn to a second sufficient condition for separability.

\(^7\) In regular continuous cases, we must first replace $X$ (or $X_{(n)}$) by a one-to-one function $(T, Y)$ and the role of conditional density is played by the conditional density of $Y$ given $T$.

\(^8\) At least, this is Johnson's [19] interpretation of Nandi's exposition, but Nandi is not quite so clear.
Note that the definition of separability is symmetric in the roles of $\gamma$ and $\eta$. Hence, we conclude immediately that if there is a statistic $U_n = u_n(X_n)$, or $= u_n(X(n))$, which is Fraser-sufficient for $\eta$, then $\gamma$ and $\eta$ are separable. In this case, the density of $X(n)$ (and of $X_n$) is the product of the conditional density given $U_n$, which depends only on $\gamma$, and a factor which is $\gamma$-free. Consequently, the l.r. (at $x(n)$) is identical with the conditional l.r., each likelihood (or rather density) being conditional on a fixed value of $U_n$ (or $U_1, \ldots, U_n$). This situation has not been previously described (in a sequential context). For $U_n$ to be Fraser-sufficient for $\eta$, it is necessary that it be sufficient for $\eta$ for every fixed value of $\gamma$ (utilizing the factorization theorem of sufficiency)--in particular, it must be sufficient for $H_0$ and for $H_1$.

Further consideration of conditional l.r.'s, conditioned on such $U_n$'s, is given under method (3).

We thus conclude that the existence of a Fraser-sufficient statistic, either for the nuisance parameter or for the parameter of interest, implies the separability of the parameters. Thus, the separability of $\gamma$ and $\eta$ is seen to be a symmetric generalization of the concept of Fraser-sufficiency. We now turn to some examples.

**Example: trinomial.** (The relevance of this example to the two-sample problem will appear later.) Suppose each stage (experiment) results in one of three outcomes--label them success, failure, and tie (the labels need not have the connotated significance)--with respective probabilities $p_1$, $p_2$ and $p_3$ (positive and summing to unity). If $\gamma = p_1/p_2$ is the parameter of interest, then $\eta = p_1 + p_2$ may be taken to be the nuisance parameter ($0 < \gamma < \infty$, $0 < \eta < 1$).

It is readily seen that the joint likelihood at stage $n$ may be expressed as $\gamma^t(1+\gamma)^{-u} \eta^u(1-\eta)^{n-u}$ where $t$ is the number of successes in the first $n$ experiments and $u$ is the number of successes and failures (non-ties) in the
first n experiments. That \( \gamma \) may be separated is obvious. Hence, the joint
l.r. for testing \((\gamma_0, \eta')\) vs. \((\gamma, \eta')\) is

\[
(\gamma/\gamma_0)^t \left[ (1+\gamma_0)/(1+\gamma) \right]^{\eta}
\]

for arbitrary \( \eta' \). It is seen that the "tie" category data do not affect the l.r.,
which seems intuitively correct. The resulting SPRT was introduced by Wald
(see next section), but by a different approach. Incidentally, \( U \) may be shown
to be Fraser-sufficient for \( \eta \); a Fraser-sufficient statistic for \( \gamma \) apparently
does not exist.

If we interchange the definition of \( \eta \) and \( \gamma \), we have an example of separa-
tion where \( T \) is Fraser-sufficient for the parameter of interest.

**Example:** difference between two normal means (variance known).

Suppose each observation consists of a pair \((X, Y)\) of independent normal random
variables with means \( \lambda \) and \( \mu \), respectively, and unit variances. Let \( \gamma = \mu - \lambda \)
and \( \eta = (\mu + \lambda)/2 \). The density of a single pair \((X, Y)\) is

\[
f_{\gamma, \eta}(x, y) = (2\pi)^{-1} \exp\left[-\frac{1}{2}(x-\lambda)^2 - \frac{1}{2}(y-\mu)^2\right]
\]

\[
= (2\pi)^{-1} \exp\left[-\frac{1}{2}(y-x-\gamma)^2\right] \exp\left[-\frac{1}{4}(x+y-2\eta)^2\right]
\]

so that the separation is achieved. (In fact, letting \( X' = Y-X \) and \( Y' = X+Y \),
this problem in terms of \((X', Y')\) falls in the "trivial example" category noted
earlier.) The resulting l.r. depends only on \( y-x \) and the resulting SPRT only
on the difference between sample means.

At stage \( n \), \( S = (T, U) \) is sufficient for \((\gamma, \eta)\) where \( T = \overline{Y} - \overline{X} \) and \( U = (\overline{X} + \overline{Y})/2 \).
(\( \overline{X} \) and \( \overline{Y} \) are the sample means of the first \( n \) \( X \)'s and \( Y \)'s, respectively.) Note
that \( T \) and \( U \) are independently distributed and \( T \) is normal with mean \( \gamma \) and
variance 2/\( n \) while \( U \) is normal with mean \( \eta \) and variance 1/2\( n \). Hence \( T \) is
Fraser-sufficient for \( \gamma \) and \( U \) is Fraser-sufficient for \( \eta \).
This example may also be treated by invariance (method (2); see next section). That there is an overlap between the concepts of invariant sufficiency and Fraser-sufficiency has been noted elsewhere [16] but not extensively explored.

Example: sign test. Fraser [9] gives a non-parametric example of Fraser-sufficiency in which the "parameter of interest" is the probability, \( p \), to the right of the origin in the population (with continuous distribution function) from which one is sampling, and the "nuisance parameter" consists of the two conditional distributions, to the right and to the left of the origin.

Suppose one takes one observation from each of two populations (with continuous distribution functions) at each stage and wishes to test hypotheses about \( p \), the probability that the difference between the observations is positive. Then separation is applicable (the number of positive differences being Fraser-sufficient for \( \gamma = p \)), and the SPRT is seen to be the sequential sign test (first described by Armitage [2]), applied to the sequence of differences. (See Test 4 in Part II; also see Test 7b.)

METHOD (2): DATA REDUCTION

A standard technique of statistical inference in the presence of nuisance parameters, and in nonparametric theory, is to replace the original data by a statistic whose distribution depends solely on the parameter of interest. In non-sequential theory, one need only know the distribution of the statistic under \( H_0 \) to construct a test of level \( \alpha \). The method is of course applicable sequentially as well, although distributions under both \( H_0 \) and \( H_1 \) need to be known.

Sequentially, two variations may be conveniently distinguished. In the first, we let \( Y_n = y_n(X_n) \) be a statistic based on the stage \( n \) data \( X_n \), with
distribution governed solely by \( \gamma \) (since the stages are mutually independent, the joint distribution of the \( Y_n \)'s is determined by the marginals). In the second, we let \( Y_n = y_n(X_{(n)}) \) be based on all the accumulated data \( X_{(n)} = (X_1, \ldots, X_n) \), again with distribution governed solely by \( \gamma \)-in fact, we must be sure that the joint distribution of \( Y_{(n)} = (Y_1, \ldots, Y_n) \) be governed solely by \( \gamma \) for each \( n \). (For the first, it may be necessary for each stage to represent a group of observations for such a \( Y_n \) to exist non-trivially). In the first, a SPRT for the hypotheses \( H_0 \) vs. \( H_1 \) is readily constructed (in principle) since \( H_0 \) and \( H_1 \) reduce to the simple hypotheses \( \gamma = \gamma_0 \) vs. \( \gamma = \gamma_1 \) about the distribution of the members of the sequence \( Y_1, Y_2, \ldots \) of independent statistics. The \( \text{L.R.} \) for each \( Y_n \) is determined, and the joint \( \text{L.R.} \) \( R_n \) obtained by multiplying together the stage-wise \( \text{L.R.'s} \). The second case is quite similar, except that the \( Y_n \)'s are no longer independent. Therefore, the joint \( \text{L.R.} \) \( R_n \) requires knowledge of the joint distribution of \( Y_{(n)} \). We refer to either of these variations as data reduction methods.

Cox [7] introduced conditions under which the joint distribution of \( Y_{(n)} \) would conveniently factor; under these conditions the joint \( \text{L.R.} \) of \( Y_{(n)} \) is identical with the \( \text{L.R.} \) of \( Y_n \). A more general version of Cox's Theorem is due to Stein, and appears in [16], together with an extensive discussion of its use. To obtain this convenient factorization, an invariance structure is required, the statistic \( Y_n \) being derived as an invariantly sufficient statistic under some appropriate group of transformations leaving the problem (and the hypotheses \( H_0 \) and \( H_1 \) invariant. The distribution of such a statistic is \( \eta \)-free. For further explanation and details, see [16].

Invariance may also be applicable in the first case considered above, each \( Y_n \) being invariantly sufficient for the data from that stage. This case is also discussed in [16] (see also Tests 3c, 3d, 7c, 7d, 7e, and 7f in Part II).
We shall now give some examples of the invariance approach, with relevance in the two-sample problem.

Example: two normal means (variance known). This example, introduced in the previous section, may also be treated by invariance. The difference between the sample means is an invariantly sufficient statistic for $\gamma$ under common shifts in location: $(x,y) \rightarrow (x+\alpha, y+\alpha)[16]$, and hence reducing the data to this difference permits construction of a SPRT. In fact, the observations need not be in pairs, but can be taken one at a time in an arbitrary order (so long as the order is free of dependence on the previous observations; see Test 1 in Part II). With such a sampling scheme, $\gamma$ and $\eta$ are no longer generally separable; in fact, neither $T$ nor $U$ is Fraser-sufficient except when the number of X's, say $m$, is equal to the number of Y's, say $n$. $S = (T,U)$ is still sufficient for $(\gamma, \eta)$, and the marginal distributions still depend solely on a single parameter, but the conditional distributions depend on both parameters. Specifically, with $T$ the difference between the sample means and $U$ the average of the two sample means, $T$ given $U=u$ is normal with mean $\gamma + 2\frac{(m-n)}{m+n} (u - \eta)$ and variance $4/(m+n)$, and hence not $\eta$-free unless $m = n$; and likewise for the conditional distribution of $U$ given $T$ (and the same would hold if $U$ were replaced by the grand mean $U' = (\Sigma X + \Sigma Y)/(m+n)$).

Example: two-sample t-tests. A two-sample t-test may be analogously derived, using invariance theory. After observing $n$ X's and $n$ Y's, the data are reduced to the usual t-statistic, based on the difference between the sample means and the pooled (within samples) estimate of the common but unknown $\sigma^2$ (m+n-2 degrees of freedom). This t-statistic has a distribution depending only on the parameter $\gamma = (\mu-\lambda)/\sigma$, so that an SPRT may be constructed to test hypotheses about $\gamma$. See Tests 3a and 3b in Part II; a two-sample t-test for testing against two-sided alternatives ($\gamma=0$ vs. $|\gamma| = \gamma_1$) has been given by Hajnal [17].
Example: **rank tests.** Sequential rank tests have been described in [32], [16], and [26]. We briefly summarize; for testing whether two samples come from the same population against the alternative that one population has a distribution function which is a power of the distribution function of the other (Lehmann alternatives), one may reduce the data to the ranks of the observations in one sample from a combined ranking of the two samples. Invariance theory guarantees that the original hypotheses will be simple hypotheses about the distribution of the reduced data. Use of the Stein Theorem [16], or alternatively the notion of sequential rank vectors [26], facilitates construction of the l.r. on which an SPRT may be based. Wilcoxon, Rhodes and Bradley [32] used what we called the first variation at the beginning of this section, thereby preserving independence among the stage-wise statistics (each stage consisting of several observations from each population). Their approach has the advantage that rank sum tests can also be devised, since the distribution of the rank sum can be worked out under both $H_0$ and $H_1$ within stages. The other approaches have the advantage that they permit sampling from one population at a time and in an arbitrary order. These tests are summarized in Part II (Tests 7a to 7f).

Now the data reduction method is not confined to the method of invariance. In particular, whenever there is a statistic $T_n$ which is Fraser-sufficient for $\gamma$, method (1) leads to a data reduction to $T_n$. See, for example, the sign test example of the previous section. (Sign tests can also be derived by invariance; see [22] and [16].) We now give an example in which invariance is not applicable.

**Example:** double dichotomy test. Suppose (simple random) sampling is done in pairs from each of two Bernoulli processes. (Actually, the data may have been more elaborate, but reduced to two Bernoulli sequences by a preliminary data reduction.) Thus, each stage results in one of the four possible outcomes:
SS, SF, FS, FF, the symbols indicating success or failure in the respective Bernoulli trials. Wald considered this problem in Chapter 6 of [31]. He proposed merging the tied pairs—that is, of replacing the above double dichotomy with the trichotomy SF, FS, and (SS, FF), which we now re-label as success, failure, and tie, respectively. The probability model for the trichotomy has parameters \((p_1, p_2, p_3)\), say, where

\[ p_1 = q_1' q_2', \quad p_2 = q_1' p_2', \quad \text{and} \quad p_3 = p_1' p_2' + q_1' q_2' = 1 - p_1 - p_2 \]

and \((p_1', p_2')\) are the probabilities of success, respectively, in the two Bernoulli sequences; \(q_1' = 1 - p_1'\).

Suppose the parameter of interest is the pair \((p_1', p_2')\), a function of \(\theta = (p_1', p_2')\), but not 1-1 (actually 1-2). Then the merging of tied pairs is an example of eliminating a nuisance parameter by data reduction. Any justification—other than convenience—for such a reduction is not known; a minimal sufficient statistic in the reduced data trinomial problem (the numbers of successes and of failures) is not a function of a minimal sufficient statistic in the original double dichotomy problem (the numbers of successes in each of the Bernoulli sequences), so that inference procedures after data reduction will not be functions of the original sufficient statistic. This suggests some lack of efficiency. (Cox [8] has introduced sequential tests for this problem which are based solely on the minimal sufficient statistic at each stage, but they have only an asymptotic justification.)

If instead we are only interested in the parameter \(\gamma = p_1/p_2 = p_1' q_2' / q_1' p_2'\) (the odds ratio), a combination of the above data reduction (merging tied pairs) with the separation exemplified in the previous section (trinomial example) leads to Wald’s double dichotomy test [31] (Test 5a in Part II). This test
may be viewed as a SPRT of hypotheses about the odds ratio $\gamma$ with an arbitrary value assigned to the nuisance parameter $p_1 + p_2 = p_1'q_2' + q_1'p_2'$. This test may (and was by Wald) also be derived by method (3) below. That this test is a SPRT about simple hypotheses in the trinomial model could be exploited in developing its properties--e.g., approximating the OC or ASN functions. In this particular example, it has been found [31] just as convenient to exploit the conditioning approach from method (3) below. The simple hypothesis SPRT approach just alluded to, however, has been useful in a recent extension of Wald's double dichotomy test [15].

METHOD 3: CONDITIONING

We now describe a method which, to our knowledge, has only been used sequentially in the double dichotomy problem as treated by Wald [31]; however, it is a fairly straightforward extension of the conditional test notions that have been widely used in non-sequential theory. The most familiar example, perhaps, is Fisher's exact test for $2 \times 2$ tables which is a conditional test, conditional on fixed marginal totals; but, having the prescribed significance level whatever the marginal totals, it is a valid unconditional test (valid in that its significance level is correct). Most of the similar tests of classical Neyman and Pearson theory, and what Lehmann has called tests of Neyman structure, are all examples of conditional tests. Another important class of examples to be mimicked here, are the permutation tests introduced by Pitman [27], and derived by Lehmann and Stein [23] as tests of non-parametric hypotheses against parametric alternatives. All of these are well-known in non-sequential theory (e.g., [22] and [10]) and all frequently have some kind of optimality. We shall now describe an analogous technique in sequential theory
and construct CSPRT's--conditional sequential probability ratio tests. (No optimal properties are known for these CSPRT's.)

Let $U_n$ be a statistic based solely on the stage $n$ data $X_n$. Suppose the conditional distribution of $X_n$ given $U_n = u$ is completely determined by each of $H_0$ and $H_1$--that is, the conditional distribution depends only on $\gamma$. A CSPRT of $H_0$ vs. $H_1$, which are simple hypotheses about the conditional distribution, is constructed by letting the role of the l.r. $R_n$ be played by the conditional likelihood ratio (c.l.r.). For each $n$, we condition $X_n$ on $U_n$, and since we assume the $X_n$'s to be independent, the joint conditional density of $X_n = (X_1, \ldots, X_n)$ given $U_1, \ldots, U_n$ is the product of the conditional densities of $X_j$ given $U_j$ ($j=1, \ldots, n$). The c.l.r. is this product.

For a fixed $U$-sequence $U_1, U_2, \ldots$, an SPRT based on $R_n$ is a perfectly valid SPRT, and hence has the prescribed error probabilities $\alpha$ and $\beta$ (approximately). Since these conditional error probabilities do not depend on the condition, namely the values assumed by the $U$-sequence, the unconditional error probabilities are also $\alpha$ and $\beta$ (approximately). That is, since the conditional probability of rejecting $H_0$, when $H_0$ is true and the $U$-sequence is fixed, is approximately $\alpha$, whatever the fixed values of the $U$-sequence, so is its expectation for any $\theta$-value in $H_0$; thus the unconditional probability of rejecting $H_0$, when $\theta$ (in $H_0$) is true, is equal to $\alpha$. Likewise, the second kind of error is not only conditionally equal to $\beta$, but unconditionally equal to $\beta$ for any $\theta$-value in $H_1$. We thus see that such CSPRT's are doubly similar--similar both with respect to $H_0$ and with respect to $H_1$. Of course, the performance of the test may depend on the nuisance parameter $n$, but the error probabilities do not (except for the approximation introduced by overshoot).
It is important to note that at stage $n$ the data are being conditioned on a function $U_n$ of the stage $n$ data, and not on a function of all the data available so far (other than $U_1, \ldots, U_{n-1}$); for, having fixed $U_1, \ldots, U_{n-1}$ at the previous stage they cannot be "unfixed" at stage $n$. Thus, we are led to more comprehensive "fixing" than is done in nonsequential problems. This remark may be clearer in the context of an example; see the first example below. Incidentally, it is because of this extensive conditioning that no optimal properties of CSPRT's are immediately apparent. Even conditionally, they are not known to be optimal in Wald's sense since they are not based on a sequence of conditionally identically distributed observations--each observation $X_n$ being conditioned on a value for $U_n$, say $u_n$, so that the conditional distributions have a parameter $u_n$ varying from stage to stage. Also, because of the extensive conditioning, it may be necessary to sample in groups.

When is this conditioning approach applicable? What is required is that there exist a statistic $U_n$ for which the conditional distribution of $X_n$ given $U_n = u$ be completely specified by $\gamma = \gamma_i (i=0,1)$; of course, it is also necessary that these two conditional distributions not be identical for all values of $u$. Thus, it is sufficient that there exist a statistic $U_n$ which is (i) sufficient for each of $H_0$ and $H_1$, since, by definition of sufficiency, the conditional distributions would then be parameter-free, and (ii) that the probability be less than one that the conditional distributions given this statistic are identical. In parametric notation, a statement essentially equivalent to (i) is that $\gamma$ be partially separable--that is, that, for each $n$, the density of $X_n$ may be written in the form

$$f_\theta(x) = g_\gamma(x) f_\theta(u(x)) ;$$
for then, setting \( \gamma = \gamma_0 \) or \( \gamma_1 \), this factorization implies (by the factorization theorem of sufficiency theory) the sufficiency of \( U \) for fixed \( \gamma \). The factor \( \ell_\theta \) may always be taken to be the density of \( U \) and \( g_\gamma \) the conditional density (probability) given \( U \); hence, the condition (ii) above is simply that the above separation of the \( g_\gamma \) factor be non-trivial—that the factor should indeed vary with \( \gamma \), at least for some \( u \)-values. This sufficiency for \( H_0 \) and for \( H_1 \), or this non-trivial partial separation of \( \gamma \), are of course not necessary for the conditioning approach to be applicable.

A special case to be noted is that when a statistic \( U_n \) exists which is Fraser-sufficient for \( \eta \). It was already noted under method (1) that such a statistic is sufficient for \( H_0 \) and for \( H_1 \), and hence \( \gamma \) is then completely, not just partially, separable. Then, the CSPRT is identical with the SPRT for any fixed value of the nuisance parameter \( \eta \) since, as noted earlier, the c.l.r. is identical with the l.r. for arbitrary \( \eta \). All examples of method (1) based on a Fraser-sufficient \( U_n \) are therefore examples of method (3) as well. But method (3) is more general as seen in the examples to follow.

Before exemplifying the conditioning method, we briefly consider the properties of CSPRT's when the \( X \)'s are mutually independent and identically distributed. Let the c.l.r. of the stage \( n \) data be denoted by \( r_n \) (the ratio of conditional densities of \( X_n \) given \( U_n \) under \( H_1 \) and \( H_0 \)). Then \( \log R_n = \sum_{i=1}^{n} \log r_i \). Now the \( r_i \)'s, considered as random variables, are unconditionally independent and identically distributed; conditionally, they are independent but not identically distributed. Thus, considered unconditionally, the CSPRT is based on a cumulative sum of a random number of independent and identically distributed terms. Therefore, Wald's method [31] of proving the certainty of termination is applicable. Also, Wald's equation (namely, \( E \log R_n = E \log r_1 \cdot E N \)), on which his ASN approximation
is based, is applicable. However, \( r_1 \) is not (unconditionally) a l.r., and hence Wald's methods of investigating and approximating the OC function are not generally applicable. Thus, the OC is only known for \( \theta \) in \( H_0 \) and in \( H_1 \) where it is (approximately) \( 1-\alpha \) and \( \beta \), respectively. Therefore, the ASN is only known (approximately) in \( H_0 \) and \( H_1 \), being given by

\[
E_{\theta} N = E_{\theta} \log R_N / E_{\theta} \log r_1
\]

where \( N \) denotes the (random) number of stages required for termination, and where

\[
E_{\theta} \log R_N = (\text{approximately}) \begin{cases} 
\alpha a + (1-\alpha)b & \text{for } \theta \text{ in } H_0 \\
(1-\beta)a + \beta b & \text{for } \theta \text{ in } H_1;
\end{cases}
\]

\( a = \log (1-\beta)/\alpha \) and \( b = \log \beta/(1-\alpha) \). Of course, if (for each \( n \)) the conditioning statistic \( U_n \) is Fraser-sufficient, then the CSPRT is a valid SPRT for arbitrary \( \eta \) and Wald's methods of approximating the OC and ASN for any \( \theta \)-value are applicable.

One further property of CSPRT's may be noted. Lehmann's method ([22], pp. 101-2) of proving monotonicity of the OC is applicable for CSPRT's. His condition that \( X_i \) have a monotone l.r. in \( \gamma \) and in \( t(X_i) \) may be replaced by the condition that \( X_i \) have a monotone c.l.r. in \( \gamma \) and \( t(X_i) \) for fixed \( u_i \); identical distributions are not required. Then one can conclude that the conditional OC is monotone in \( \gamma \), whatever the values of the U-sequence, and hence unconditionally monotone as well. Examples of this will be noted below.

We now turn to examples, first reviewing the double dichotomy example of Wald, then extending it, and then introducing a permutation test.

**Example: double dichotomy test.** We return to the double dichotomy example of the previous section, with \( \gamma = p_1/p_2 = p_1'q_2'/q_1'p_q' \). Denote the data from a single stage by \((X,Y)\) where \( X \) and \( Y \) each = 1 (success) or 0 (failure) in the
respective trials. For fixed \( \gamma \), the density is

\[
p_1^x q_1^{1-x} p_2^y q_2^{1-y} = \gamma^x \cdot q_1' q_2' (p_2'/q_2')^u
\]

where \( u = x + y \). Hence, \( U \) is sufficient for \( (p_1', p_2') \) for each fixed \( \gamma \)-value, and the conditioning on \( U \) is possible. In contrast, non-sequentially, after \( n \) stages, one would condition the distribution of accumulated successes in each sequence on the total accumulated number of successes. The conditional distribution of \( (X, Y) \) given \( U = u \) is:

\[
(X, Y) = \begin{cases} 
(0,0) \text{ with probability 1 when } u = 0, \\
(1,0) \text{ or } (0,1) \text{ with probabilities } \gamma/(1+\gamma) \text{ and } 1/(1+\gamma), \\
(1,1) \text{ with probability 1 when } u = 2.
\end{cases}
\]

The c.l.r. is thus unity except when \( u = 1 \), and then is \( (\gamma_1/\gamma_0)(1+\gamma_0)/(1+\gamma_1) \) when \( (X, Y) = (1,0) \) and \( (1+\gamma_0)/(1+\gamma_1) \) when \( (X, Y) = (0,1) \). In effect, only those pairs with \( u = 1 \)--the untied pairs--contribute to the cumulative sum \( \log R_n \) on which the test is based. This test has been described this way (by Wald [31] and Armitage [3]) as being a conditional one confined to the sequence of untied pairs. This latter sequence is simply a Bernoulli sequence, and the usual SPRT for testing \( p = \gamma_0/(1+\gamma_0) \) vs. \( p = \gamma_1/(1+\gamma_1) \) is applied. Its properties are described by Wald, utilizing the fact that it is simply a test based on a Bernoulli sequence. The ASN has to be adjusted by dividing by the probability of an untied pair.

As noted above, its properties could also be derived from the fact that it is a SPRT of \( (\gamma_0', \eta') \) vs. \( (\gamma_1', \eta') \), for arbitrary \( \eta' \), after ties have been merged. Not at all apparent from the method (3) approach, but obvious from the method (1) approach, is that this test has Wald's optimal property uniformly in \( \eta \)--that is, the (unconditional) ASN is minimized at \( (\gamma_1, \eta') \) among

23
all tests of these simple hypotheses with strength \((\alpha, \beta)\), and this is true uniformly in \(\eta'\). From Wald's approach (method (3)), it is only apparent that the conditional ASN (which depends only on \(\gamma\)) is minimized at \(\gamma_1\) among all tests of the same strength which ignore tied pairs. This optimality of Wald's double dichotomy test has not been pointed out before.

**Example:** extended double dichotomy test. We consider the same model as in the double dichotomy example, namely, that of two (independent) Bernoulli sequences; but we now no longer confine attention to sampling in pairs. Instead, we permit sampling in groups of arbitrary and varying sizes, with the sole requirement that at least one observation from each Bernoulli sequence be included in each group (or otherwise the data from some groups may have to be ignored). There are of course possible advantages of pairing in that certain kinds of variation may be eliminated by matching the pairs [3]. On the other hand, there may be advantages to permitting larger group sizes; for example, a group may be constituted by all observations reported during a particular month to a center supervising clinical trials in a number of hospitals. Or it may be desirable to place a larger proportion of patients on the new treatment, in the context of comparing a new treatment with a standard one in clinical trials, than on the standard, or vice-versa.

We temporarily confine attention to a single stage of the experiment, and suppress its index \(n\) from the notation. Suppose \(n_1\) and \(n_2\) (positive integers) observations are taken from the two Bernoulli sequences, and let \(U\) denote the total number of successes among the \(n_1 + n_2\) trials. Let \(X\) and \(Y\) represent the numbers of successes from the respective Bernoulli sequences \((U = X + Y)\); \((X,Y)\) is sufficient for this stage of the experiment. The density (probability) of \((X,Y)\) may be expressed in the form
\[ f_{p_1', p_2'}(x, y) = \binom{n_1}{y} \binom{n_2}{x} \gamma^x \cdot q_1^{n_1} q_2^{n_2} \left( \frac{p_2}{q_2} \right)^u \]

where \( \gamma \) is the odds ratio \( p_1'q_2' / p_2'q_1' \). The sufficiency of \( U \) when \( \gamma \) is specified is thus apparent; that is, \( \gamma \) is partially separable. Conditioning on fixed \( U \), say \( U = u \), the distribution of \((X,Y)\) is effectively that of \( X \) (given \( U \)); this conditional density (probability) is readily found to be

\[ \binom{n_1}{x} \binom{n_2}{u-x} \gamma^x / \sum_{j=1}^{u} \binom{n_1}{j} \binom{n_2}{u-j} \gamma^j . \]

An alternative form is obtained by replacing the coefficient of \( \gamma^j \) in the denominator by \( \binom{u}{j} \binom{u'}{j'} \) where \( u' = n_1 + n_2 - u = \text{total number of failures} \), and \( j' = n_1 - j \) (failures in first sequence), and likewise in the numerator.

The c.l.r. \( r_n \) for this one stage is thus

\[ \left( \frac{\gamma_1}{\gamma_0} \right)^x \cdot \sum_{j=1}^{u} \binom{u}{j} \binom{u'}{j'} \gamma_0^j / \sum_{j=1}^{u} \binom{u}{j} \binom{u'}{j'} \gamma_1^j . \]

The CSPRT for testing \( \gamma_0 \) vs. \( \gamma_1 \) may be carried out by computing \( \log R_n = \sum_{i=1}^{n} \log r_i \) and comparing it with Wald's boundaries \( a \) and \( b \) in the usual way.

We now briefly examine this test in the case of sampling in double pairs--\( n_1 = n_2 = 2 \) at every stage. The c.l.r. for a single stage is seen to assume one of six possible values, given in the table below; the unconditional probabilities of these six types of data (indexed 1 to 6 for convenience) are also given.

25
<table>
<thead>
<tr>
<th>type</th>
<th>(x,y)</th>
<th>c.l.r. = r_n</th>
<th>unconditional probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0,0) or (2,2)</td>
<td>1</td>
<td>( q_1^2 q_2^2 + p_1^2 p_2^2 )</td>
</tr>
<tr>
<td>2</td>
<td>(0,1) or (1,2)</td>
<td>((1+\gamma_0)/(1+\gamma_1))</td>
<td>( q_1^2 p_2 q_2 + p_1 q_1^2 p_2^2 )</td>
</tr>
<tr>
<td>3</td>
<td>(1,0) or (2,1)</td>
<td>( \gamma_1 (1+\gamma_0)/\gamma_0 (1+\gamma_1) )</td>
<td>( p_1^2 q_1^2 q_2 + p_1 p_2 q_1^2 )</td>
</tr>
<tr>
<td>4</td>
<td>(0,2)</td>
<td>((1+\gamma_0+\gamma_1)/(1+\gamma_1+\gamma_1^2))</td>
<td>( p_1^2 p_2^2 )</td>
</tr>
<tr>
<td>5</td>
<td>(1,1)</td>
<td>( \gamma_1 (1+\gamma_0+\gamma_1^2)/\gamma_0 (1+\gamma_1+\gamma_1^2))</td>
<td>( p_1 q_1^2 p_2 q_2 )</td>
</tr>
<tr>
<td>6</td>
<td>(2,0)</td>
<td>( \gamma_1^2 (1+\gamma_0+\gamma_1^2)/\gamma_1^2 (1+\gamma_1+\gamma_1^2))</td>
<td>( p_1^2 q_2^2 )</td>
</tr>
</tbody>
</table>

Letting \( c_{in} \) denote the accumulated number (through stage \( n \)) of stages with data of type \( i \), and letting \( d_i \) be the logarithm of the c.l.r. for a single stage with data of type \( i \), we thus have

\[
\log R_n = \sum_{i=1}^{6} c_{in} d_i
\]

and the test may be carried out conveniently by plotting this cumulative sum until it reaches one of the horizontal termination boundaries with ordinates \( a \) and \( b \).

In the special case when \( \gamma_1 = 1/\gamma_0 \), it may be shown that this test is identical with the double dichotomy test (except that termination is permitted here only after even numbers of pairs of observations). Otherwise, it is readily seen that the extended test is not the same as Wald's test.

For, the extended test would treat two successive pairs (making up the \( n^{th} \) stage, say) with outcomes SS and FF identically with two pairs with outcomes SF and FS (each of these double pairs is of type 5), whereas Wald's test would ignore the tied pairs SS and FF but not the untied pairs SF and FS. (If \( \gamma_1 = 1/\gamma_0 \), the contribution of SF to the cumulative sum would cancel the contribution of FS, but not otherwise.) It is thus seen that the
extended test utilizes more of the data.

Now except in the case \( n_1 = n_2 = 1 \) already considered, \( U_n \) is not Fraser-sufficient. The CSPRT for the extended problem is not a SPRT. Its ASN at \( \gamma_0 \) and \( \gamma_1 \) is approximable (see Appendix I). But no approximation to its OC is available, although it is monotone in \( \gamma \) for fixed \( \eta \). The monotonicity property holds since, for fixed \( U_n \), the stage n c.i.r. is monotone in \( x_n \), and Lehmann's method of proof cited above therefore leads to the conclusion that the conditional OC for fixed \( U_1, U_2, \ldots \) is monotone in \( \gamma \); since this is so for every fixed U-sequence, monotonicity also holds unconditionally. It is apparent that the OC and the ASN do indeed depend on \( \eta \).

The ASN's of the double dichotomy test and the extended test (under \( H_0 \) and \( H_1 \)) are compared in Appendix I. (\( E_\theta \log r_1 \) is readily computed from the table above.) It is found that a slight reduction may be possible—but typically only a very slight one—by using the extended test. On the other hand, the increased group size in the extended test will, if sufficiently increased, have an adverse effect on the ASN. Some numerical studies are planned for later publication.

This extended test, for the case \( \gamma_0 = 1 \) (i.e., \( p_1 = p_2 \)), appears in Part II as Test 5b; Wald's double dichotomy test is Test 5a.

Example: two-sample permutation test. The non-sequential Pitman permutation test for the two-sample problem, as derived by Lehmann and Stein [23] (or see [10] or [22]), is a test of the nonparametric null hypothesis that the two independent samples come from identical populations (with some unspecified continuous distribution), against a parametric alternative that the two populations are normal, with a common variance, but with differing means.

Under the null hypothesis, the two samples constitute a single random sample
and the order statistic of this single sample is a sufficient statistic. Thus, under $H_0$, the conditional distribution of the observations, given the order statistic, is completely known—it is in fact an equal-probability distribution over the possible permutations of the values making up the order statistic. The parametric family of alternatives serves to define a critical region which is conditionally most powerful, whatever the order statistic; in fact, it is uniformly most powerful, among conditional tests, against all such normal alternatives with positive (say) difference in means.

We now develop an analogous sequential test by using a similar line of reasoning within each state of the sequential experiment, each state consisting of at least one observation from each population. It will be found that the order statistic based on the combined data from a single stage is not only sufficient for the null hypothesis but also sufficient for certain families of parametric alternatives, and hence a CSPRT may be constructed by conditioning on the sequence of stage-wise order statistics.

We shall now describe the simple special case in which sampling is done in pairs; each stage consists of a single observation from the X-population and a single observation from the Y-population. The general case appears in Part II as Test 6. The null hypothesis is that the X- and Y-populations are identical, both with continuous but unspecified distributions. (Actually, this common distribution could vary from stage to stage.) The alternative hypothesis, for the present, is that the X-population is normal with mean $\lambda^2$ and variance $\sigma^2$ and the Y-population is normal with mean $\lambda+\delta$ and variance $\sigma^2$, all parameters being specified. Let $U_n$ represent the unordered pair $\{X_n, Y_n\}$, $X_n$ and $Y_n$ being the respective observations at stage $n$. ($U_n$ is l-l in the order statistic.) Since the alternative hypothesis is simple, it is trivially true that $U_n$ is sufficient for the stage $n$ data—in fact, a constant is a
sufficient statistic for the stage n data under this alternative hypothesis. We shall explore presently to what extent the alternative hypothesis can be enlarged without destroying this sufficiency of $U_n$.

We now fix attention on a single stage, and omit the subscript n. The fixed unordered observations are denoted $u = \{x,y\}$. Under $H_0$, the conditional probability of $(x,y)$ is $1/2$, as is the conditional probability of $(y,x)$. Ties may be ignored since we are confining attention to continuous distributions. Let $\varphi(x,y)$ denote the density of $(X,Y)$ under the alternative hypothesis (the product of two normal densities). The conditional probability of $(x,y)$ is then seen to be $\varphi(x,y)/[\varphi(x,y) + \varphi(y,x)]$ (see [10] or [22]), and the conditional probability of $(y,x)$ is the same with $x$ and $y$ interchanged. Algebraic reduction leads to the value $1/(1 + e^{-\Delta z})$ for the conditional probability of $(x,y)$, where $z = y-x$ and $\Delta = \delta/\sigma^2$. Hence, the conditional distribution under the normal alternative depends solely on the parameter $\Delta$, and we thus see that normal alternatives with a fixed value of $\Delta$, say $\Delta'$--but otherwise arbitrary values of $\lambda$, $\delta$, and $\sigma^2$--may be substituted for the simple alternative above, and $U = \{X,Y\}$ remains sufficient (the conditional distribution being known). Further enlargement will be noted later.

Thus, for the stage n data, the c.l.r. $R_n$ is $2/(1 + e^{-\Delta' z_n})$. A CSPRT thus may be based on the cumulative sum $\log R_n = \sum_{i=1}^{n} \log \frac{1}{2}(1 + e^{-\Delta' z_i})$. Sampling is continued until this cumulative sum reaches one of the boundaries $a$ and $b$. Since, for each fixed sequence of values of the unordered pairs $\{X_n,Y_n\}$, the test is a valid SPRT with conditional strength $(\alpha,\beta)$, it is also a test with unconditional strength $(\alpha,\beta)$ (ignoring overshoot). The OC function is (approximately) constant within $H_0$ and within $H_1$, but is otherwise unknown. It is, however, monotone in $\Delta$ for fixed values of $\lambda$ and $\sigma^2$ (and $\delta = \Delta \sigma^2$) whenever the two populations are normal with common variance and
means $\lambda$ and $\lambda's$. (See earlier remarks on proof of monotonicity.) The test terminates with certainty, being based on a cumulative sum of independent and identically distributed random variables, namely the summands in the expression for $\log R_n$. (This assumes identical distributions at each stage, but this can be relaxed.) The ASN under $H_0$ and under $H_1$ may be approximated; an asymptotic form of this approximation, for small $\Delta'$, is given in Appendix 2.

It is also shown there that, asymptotically, these ASN approximations coincide with the ASN of the fully parametric SPRT assuming normality under $H_0$ as well as under $H_1$. Hence, this test, which has a broader validity than the normal theory test, is asymptotically efficient in this sense. Some numerical studies will be reported in a subsequent paper.

Two final comments about this test will now be made. First, just how far can the class of alternatives be extended? That is, for what other models will the test have power $1-\beta$? It may be shown that if the density functions of the two populations differ only in an exponential factor $e^{\rho x}$ and $e^{\rho' y}$ where $\rho'-\rho = \Delta'$, then $U_n$ remains sufficient and the above CSPRT remains valid. For example, if the $X$-population and $Y$-population both have gamma distributions with densities proportional to $e^{-\mu x} x^{\lambda-1}$ and $e^{-\mu' y} y^{\lambda-1}$, where $\mu-\mu' = \Delta'$ but $\lambda$ is arbitrary, then the above test has power $1-\beta$. Comparison of this test with other sequential tests appropriate for exponential and gamma distributions is planned for a subsequent paper.

The final comment is that, although this test is not a SPRT for any simple hypotheses included within $H_0$ and $H_1$, $U_n$ is not Fraser-sufficient--it is an SPRT for testing the simple hypothesis $H_0$ that the differences $Z_1, Z_2, \ldots$ have logistic distributions with distribution function $F(z) = 1/(1+e^{-\Delta'z})$, against the simple alternative $H_1$ that the $Z$'s have the distribution function $[F(z)]^2$--that is, that each $Z$ may be regarded as the larger of two independent
random variables each with the logistic distribution function \( F(z) \). This is a curiosity. It does permit approximation of the OC against alternatives of the form: \( Z \) has distribution function \( F^k \) for any \( k (\Delta' \text{ fixed}) \), but any real use of this fact is not apparent.

Other composite hypothesis testing problems are certainly amenable to method (3). Specifically, CSPRT's for the two-sample problem when testing against discrete parametric alternatives may be derived in complete analogy with the permutation test above; in fact, the double dichotomy tests may be viewed as tests of this type, but Poisson or negative binomial alternatives could likewise be considered. Also, CSPRT's for tests of independence against correlated normal alternatives have been derived and will be reported elsewhere.
PART II: SEQUENTIAL PROBABILITY RATIO TESTS FOR THE TWO-SAMPLE PROBLEM

SUMMARY OF PART II

A number of SPRT's for the two-sample problem are listed and briefly described. The two-sample problem considered is that of testing whether two populations, from which samples are sequentially taken, are identical or whether observations from the second population tend to be larger than those from the first. Thus, the tests would be appropriate for comparing a treatment with a control, or a new treatment with a standard treatment. Analogous tests are available, but not presented, for testing against two-sided alternatives; these would be appropriate when attempting to determine which of two treatments is superior or whether there is no real difference. Such two-sided tests can be constructed by running two one-sided tests simultaneously, as proposed by Armitage ([2] and [3]), or by use of invariance theory (one sets up a model for the two samples with the "labels lost"—ignoring which population is which); these will not be considered further here.

The tests given here all presume that each population has a continuous probability distribution (although the sign tests are more generally valid), and the only parametric tests given presume normal populations. All of the nonparametric tests given are also valid if the populations are actually normal, with the exception of the rank tests; these presume Lehmann alternatives and such alternatives imply that at least one of the populations is non-normal.

All tests are SPRT's (or CSMRT's—see Part I)—that is, sampling is continued until a joint likelihood ratio falls outside a fixed interval (B,A); actually, the normal tests with variance unknown are not quite of this type, but are sufficiently similar (and relevant) to be included. No other types of sequential tests, nor any non-sequential or two-stage tests,
are included. In every case, the hypotheses are composite; and in every case bounds on the two types of error are guaranteed (except for inaccuracies due to overshoot).

Two basic types of sequential sampling are considered: (1) sampling one at a time from the two populations in an arbitrary order, or in batches of arbitrary and varying compositions, and (2) sampling in pairs from the two populations; the stages of the sampling scheme are always (tacitly) assumed to be independent. Sometimes there is a restriction in (1) that there be at least one observation from each population at each stage. When sampling scheme (2) is used, tests can be based on the sequence of differences between the observations making up the pairs; such tests do not require the assumption of independence between the two populations, and also permit the elimination of certain factors by matching the pairs—that is, the two populations from which one samples may vary from stage to stage as long as the distributions of the observed differences remain constant. Some of the tests using sampling scheme (1) also permit some variation from stage to stage in the two populations.

Four general categories of tests are presented: normal theory tests, sign tests, permutation tests, and rank tests. Some empirical comparisons of the tests are planned for later publication (with Peter Nemenyi).

INTRODUCTION TO THE TESTS

We introduce some notation, conventions, and overall comments relevant to the tests that follow. We refer to an X-population and a Y-population. When sampling in pairs from the X- and Y-populations, the difference between the Y-member of a pair and the X-member of a pair is denoted by Z. The null hypothesis is always that the two populations are identical or that the distribution of Z is centered around zero (together with additional assumptions.
varying from test to test), and the alternative is always that $Y$-values tend
to be larger than $X$-values—at least in the same stage of sampling—or that
$Z$-values tend to be positive (again with additional assumptions in each test).
In every instance, the hypotheses could be expressed in the form: $H_0: \gamma = 0$
vs. $H_1: \gamma = \gamma' (>0)$ for some numerical parameter $\gamma$. Whenever the OC function
is known to be monotone in $\gamma$, the hypotheses could be enlarged to $H_0: \gamma \leq 0$
vs. $H_1: \gamma \geq \gamma'$.

The following sampling scheme notations will be used:

$S_1$: sampling in batches of arbitrary composition and size. Let $m'$ and
$m''$ (which may vary with $n$—if they do not, we say the scheme is regular)
denote the numbers of $X$- and $Y$-observations in stage $n$, and let $n'$ and
$n''$ denote the accumulated numbers of $X$- and $Y$-observations through
stage $n$. Both $n'$ and $n''$ are assumed to tend to infinity with $n$ (batch
size and composition may be random, but not dependent on previous
observations; see relevant discussion in [17]).

$S_1^*$: $S_1$ with $m' \geq 1$ and $m'' \geq 1$ for every $n$.

$S_1^{**}$: $S_1$ with $m' \geq 2$ or $m'' \geq 2$ (or both) for every $n$.

$S_2$: sampling in pairs (regular $S_1$ with $m' = m'' = 1$ for every $n$).

$S_2^*$: sampling in groups of pairs ($S_1$ with $m' = m''$).

The tests are described by defining a statistic $L_n$ (typically a log like-
lihood ratio of some sort). Unless otherwise indicated, the test is carried
out as follows: for each $n = 1, 2, \ldots$, in turn, calculate $L_n$ and

terminate and accept $H_0$ if $L_n \leq b$,

terminate and reject $H_0$ if $L_n \geq a$,

proceed to stage $n+1$ otherwise,

where $b = \log[\beta/(1-\alpha)]$ and $a = \log[(1-\beta)/\alpha]$; $\alpha$ is the prescribed bound on
the type I error and $\beta$ is the prescribed bound on the type II error. Some of
the group tests are carried out by calculating a group-specific statistic $L_n^*$.
for each group (stage), and letting \( L_n = \sum_{i=1}^{n} L_i^* \).

The following five properties are frequently investigated for SPRT's of specified strength:

(i) certainty of termination;
(ii) monotonicity of the OC in some numerical parameter \( \gamma \) of interest;
(iii) approximability of the OC (by Wald's method);
(iv) approximability of the ASN (by Wald's method);
(v) minimality of ASN under \( H_0 \) and \( H_1 \) subject to strength requirements.

All tests considered here can be shown to terminate with certainty under broad assumptions—always broader than the hypotheses specifically tested. Relevant references are [31], [6], [18], [17] and [30]. No further comments about the certainty of termination will be made.

Comments on monotonicity have been made in Part I. Relevant references are [22] and [11].

Only when the test is a simple SPRT (SPRT of two simple hypotheses based on a sequence of independent and identically distributed observations) may the operating characteristic function (OC) be approximated (with the exception of Tests 2a and 2b); Wald's method is described in [31].

Whenever Wald's equation is valid, the ASN may be related to the OC. The OC is always known (approximately) under \( H_0 \) and \( H_1 \) and thereby the ASN may frequently be approximated under \( H_0 \) and \( H_1 \); otherwise, the approximability of the ASN depends on the approximability of the OC and is essentially confined to the case of simple SPRT's. When Wald's equation is not valid, some conjectured methods of approximation of the ASN under \( H_0 \) and \( H_1 \) have been commonly used; see Johnson [19] (and [16]) for discussion.

Wald's optimal property (v) is known to hold only when the test is in fact a simple SPRT.
NORMAL TESTS

One of the following two assumptions will be made throughout this section.

**ASSUMPTION A1:** The X-population is \( N(\lambda, \sigma^2) \), the Y-population is \( N(\lambda + \delta, \sigma^2) \), and the X- and Y-populations are independent \((-\infty < \lambda < \infty, -\infty < \delta < \infty, 0 < \sigma < \infty)\).

**ASSUMPTION A2:** \( Z = Y - X \) is \( N(\delta, 2\sigma^2) \) \((-\infty < \delta < \infty, 0 < \sigma < \infty)\).

Assumption A2 does not require independence of \( X \) and \( Y \), nor equal variances of \( X \) and \( Y \), nor constant joint distributions of \( X \) and \( Y \); it will only be invoked when sampling in pairs; the successive \( Z \)'s are then assumed independent and identically distributed.

The null hypothesis will always be that \( \delta = 0 \), and the alternative will be some specific form of \( \delta > 0 \).

1. **Normal mean, variance known:**

   In this section we shall always assume either A1 or A2, and in addition, assume that \( \sigma^2 \) is known (specified); \( \delta \) is the parameter of interest and \( \lambda \) the nuisance parameter, and \( \delta' \) is a specified positive number.

**TEST 1a:**

**Assumptions:** A1, S1, and \( \sigma^2 \) known.

**Hypotheses:** \( H_0: \delta = 0 \), \( H_1: \delta = \delta' \).

**Method:** Method (2)—invariance under common location shifts in all observations from all stages; test based on l.r. of difference between sample means of accumulated data.

**Test statistic:** \( L_n = \frac{\delta'}{\sigma^2} \cdot \frac{1}{n' + n} \left( n'y - n''x - \frac{1}{2} n'n'\delta' \right) \).

**Properties:** ASN and OC are only known when the sampling scheme is regular and \( m' = m'' \); then the test is a simple SPRT, method (1) being applicable.
**TEST 1b:**

Assumptions: A2, S2, and \( \sigma^2 \) known.

Hypotheses: \( H_0: \delta = 0, H_1: \delta = \delta' \).

Method: Methods (1) and (2); data reduction to differences \( Z_1, Z_2, \ldots \),
and then separation of \( \delta \) (see Part I); test based on l.r. of \( \overline{Z} \) (sample
mean of available Z's), and hence coincides with l.r. of Test 1a.

Test statistic: \( L_n = \frac{\delta'}{\sigma^2} \cdot \frac{1}{2} \left( \Sigma z - \frac{1}{2} n\delta' \right) \)

Properties: properties of simple SPRT, uniformly in \( \lambda \).

**TESTS 1c and 1d:** Group tests, analogous to Tests 1a and 1b, respectively,
based on a statistic \( L_n \) computed from the data from each stage and then
summed over stages. Permits variation in the nuisance parameter \( \lambda \) from
stage to stage, and uses group sampling \( S1^* \) (Test 1a) or \( S2^* \) (Test 2a).
Wald's methods for approximating the OC and ASN are applicable in the case
of regular sampling.

2. Normal mean, variance unknown:

We continue to assume A1 or A2, with \( \delta \) as the parameter of interest, but
now \( \lambda \) and \( \sigma^2 \) are nuisance parameters; \( \delta' \) is as before. The tests given here
are not really SPRT's; they are modifications of SPRT's developed in analogy
with Stein's two-stage test for such problems ([14] and [14]). An estimate of
\( \sigma^2 \) is based on the data in the first stage of sampling, and the tests may be
considered conditionally (on this variance estimate) as SPRT's with termination
boundaries depending on the variance estimate. Alternatively, unconditionally
they may be described by a test statistic \( L_n \) which is a l.r. except that \( \sigma^2 \)
is replaced by its estimate and the termination boundaries a and b are
replaced by

\[ a_\nu = \frac{1}{2} \nu (\alpha^{-2/\nu} - 1) \text{ and } b_\nu = -\frac{1}{2} \nu (\rho^{-2/\nu} - 1) \]

where \( \nu \) denotes the degrees of freedom of the variance estimate. See [14] for
details.
TEST 2a:

Assumptions: A1 and S1 with the restriction that the first stage has at least two observations from at least one of the populations.

Hypotheses: \( H_0: \delta = 0, \ H_1: \delta = 5' \).

Method: See above; \( v \) = number of observations in first stage less \( k \) where \( k = 2 \) if both populations are sampled in the first stage and \( k = 1 \) otherwise.

Test statistic: \( L_n \) same as in Test 1a with \( \sigma^2 \) replaced by \( s^2 = \text{pooled within-sample mean square from first stage data} \) \( (v \text{ degrees of freedom}) \)

\[
= \frac{1}{v} \left[ \sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 \right] \quad \text{(first stage data only); termination boundaries} \ a_v \quad \text{and} \quad b_v \quad \text{are used.}
\]

Properties: ASN and OC may be approximated when \( m' = m'' = m \), free of \( n \); see [14].

TEST 2b:

Assumptions: A2 and S2 with the modification that the first stage consists of \( v+1 \) pairs, rather than a single pair, of observations \( (v > 0) \).

Hypotheses: \( H_0: \delta = 0, \ H_1: \delta = 5' \).

Method: See above; based solely on differences.

Test statistic: \( L_n \) same as in Test 1b with \( \sigma^2 \) replaced by \( \frac{1}{2} s^2 \) and

\[
\hat{s}^2 = \frac{1}{v} \sum (z_i - \bar{z})^2 \quad \text{(first stage data only); termination boundaries} \ a_v \quad \text{and} \quad b_v \quad \text{are used.}
\]

Properties: ASN and OC may be approximated; see [14].

3. Normal mean in standard deviation units—t-tests:

We continue to assume A1 or A2 but re-parameterize, replacing \((\delta, \lambda, \sigma^2)\) by \((\gamma, \lambda, \sigma^2)\) where \( \gamma = \delta/\sigma \); \( \gamma \) is the parameter of interest, and \( \lambda \) and \( \sigma^2 \) are nuisance parameters; \( \gamma' \) is a specified positive number. Four tests are given, analogous to tests 1a, 1b, 1c and 1d, respectively. The first two are based on the invariance method (or Cox's method) and are analogs both of Rushton's [28] one-sided sequential t-test, with the degrees of freedom and non-centrality
parameter modified to adapt it to the two-sample problem, and Hajnal's [17] two-sample sequential t-test except that we consider one-sided rather than two-sided alternatives.

The exact test statistics involve confluent hypergeometric functions \( M(\cdot,\cdot,\cdot); \) suitable tables are given by Rushton and Lang [29]. Approximation formulas are also given, using Rushton's third approximation [25]; such approximations are probably satisfactory for large degrees of freedom, which are not available in Tests 3c and 3d.

The ASN's and OC's are not known, but the ASN is presumably close to that for the corresponding tests with variance known (for small \( \gamma' \)), and approximations are discussed by Johnson [19] for analogous sequential t- and F-tests. Wald's methods can be applied to approximate the OC and ASN in Tests 3c and 3d when sampling is regular.

**TEST 3a:**

**Assumptions:** \( A_l \) and \( S_l \).

**Hypotheses:** \( H_0: \gamma = 0, \quad H_1: \gamma = \gamma'. \)

**Method:** Method (2)—invariance under common location and scale changes; test based on l.r. of t-statistic based on difference between sample means of accumulated data and pooled within-sample variance estimate based on accumulated data.

**Test statistic:**

\[
L_n = -\frac{1}{2} k^2 \gamma^2 + \log[M(\frac{\nu+1}{2}, \frac{1}{2}, \frac{1}{2} \nu^2 \gamma^2, \nu^2 \gamma^2) + \sqrt{2} \nu^2 M(\frac{\nu+2}{2}, \frac{3}{2}, \frac{1}{2} \nu^2 \gamma^2) \\
\cdot \Gamma(\frac{\nu+2}{2})/\Gamma(\frac{\nu+1}{2})]
\]

where \( k = \sqrt{\frac{n'n''}{n' + n''}} \), \( \nu = n' + n'' - 2 \), and

\[
v = (n' \Sigma \gamma - n'' \Sigma \gamma)(n' + n'')^{-1/2} [(n' + n'') (\Sigma \gamma^2 + \Sigma \gamma^2) - (\Sigma \gamma + \Sigma \gamma)^2]^{-1/2}
\]

\((L_n = 0 \text{ if } n' \text{ or } n'' = 0)\);

approximate \( L_n = \frac{1}{2} k^2 \gamma^2 + \frac{1}{n' \nu^2} (\nu \gamma)^2 + (\nu + 1)^{1/2} (\nu \gamma') [1 - \frac{1}{4(\nu + 1)} + \frac{(\nu \gamma')^2}{2(\nu + 1)}].\)
Properties: See above.

TEST 3b:

Assumptions: A2 and S2.

Hypotheses: \( H_0: \gamma = 0, \quad H_1: \gamma = \gamma' \).

Method: Method (2)--data reduction to differences and invariance under scale changes of the differences. Test is Rushton's test applied to the differences.

Test statistic: As in Test 3a with \( k = \sqrt{(n/2)}, \quad v = n-1, \) and \( w = \Sigma (2 \Sigma x^2)^{-1/2} \) (approximations as in Test 3a).

Properties: See above.

TESTS 3c and 3d: Group tests, analogous to Tests 3a and 3b, respectively, based on a statistic \( L_n \) computed from the data from each stage and then summed over stages. Permits variation in the nuisance parameters from stage to stage. (See section 1.7 of [16] for comments on analogous two-sided tests.)

SIGN TESTS

In this section, the actual observations are replaced by dichotomous observations--either the X's and Y's are separately dichotomized, or the differences, the Z's, are dichotomized. Continuous distributions are not required. We introduce the following notation

\[ p_1 = \text{Prob}(X < c), \quad p_2 = \text{Prob}(Y < c), \quad p = \text{Prob}(Y > X) \quad (c \text{ fixed}). \]

Under Assumption A1, and with \( \Phi \) representing the standard normal distribution function, \( p_1 = \Phi[(c-\lambda)/\sigma] \) and \( p_2 = \Phi[(c-\lambda-\delta)/\sigma] \); while under Assumption A2, \( p = \Phi[(\delta/\sqrt{2\sigma})] \). Thus, tests about the parameters \( p_1 \) and \( p_2 \), or \( p \), may be related to tests about underlying normal models, but of course these tests do not require a normality assumption. We also let \( q_i = 1-p_i (i=1, 2) \) and

40
denote the odds ratio $\frac{p_1 q_2}{p_2 q_1}$ by $\rho$.

We introduce the assumptions:

**ASSUMPTION A3:** The X- and Y-populations are independent and the odds ratio $\rho$ is constant from stage to stage.

**ASSUMPTION A4:** When sampling in pairs, $p$ is constant from stage to stage.

As always, we assume independence among stages, but the distributions may vary from stage to stage subject to either A3 or A4.

The tests presented here are based on dichotomous classifications, but they could all be extended to permit more than two classifications; such an extension of Test 4 below is treated elsewhere [15]. Other double-dichotomy tests are given by Cox [8].

**TEST 4:** Here, the parameter of interest is $p$ and all other specification of the joint distributions of successive pairs is ascribed to nuisance parameters; $p'$ is a specified number between $1/2$ and 1.

**Assumptions:** A4 and S2.

**Hypotheses:** $H_0$: $p = 1/2$, $H_1$: $p = p'$.

**Method:** Method (2) -- data reduction to the signs of the differences. (May be justified by invariance under monotone transformations $q_n$ applied to each of the observations in stage $n$; see Section I.6 of [16].) The reduced data constitute a Bernoulli sequence with success (positive sign) probability $p$. The simple SPRT for Bernoulli data (Chapter 5 of [31]) is applicable, the L.R. at stage $n$ being that of a binomial random variable with parameters $n$ and $p$. In the context of normal data, this test was originally described by Armitage [2].

**Test statistic:** $L_n = s_n \log(p'/q') + n \log(2q')$ where $q' = 1-p'$ and $s_n =$ number of positive differences in the first $n$ pairs.

**Properties:** Approximations to the OC and ASN, as functions of $p$, are well-known (e.g., [31]).
**double-dichotomy tests:**

The parameter of interest is now \( \rho \), which equals unity if and only if \( p_1 = p_2 \) and exceeds unity if and only if \( p_1 > p_2 \). All other specification of the distributions of \( X \) and \( Y \) (assumed independent) is ascribed to nuisance parameters; \( \rho' \) is a specified constant exceeding unity. This test is different from all others so far in that it is not location-free; adding a constant to all observations would not leave this formulation of the two-sample problem invariant.

Test 5a is Wald's double-dichotomy test (Chapter 6 of [31]; see also Armitage [3]), and is a special case of Test 5b which follows.

**Test 5a:**

**Assumptions:** A3 and S2.

**Hypotheses:** \( H_0: \rho = 1 \), \( H_1: \rho = \rho' \).

**Method:** Methods (2) and (1), or methods (2) and (3). Data reduction to the trichotomy: "success" if \( X < c \) and \( Y \geq c \), "failure" if \( X \geq c \) and \( Y < c \), and "tie" if \( X-c \) and \( Y-c \) have the same signs; and then separation of the parameter \( \rho \) (see Part I). Alternatively, data reduction to the double dichotomy: "success" in the \( X \)-observation if \( X < c \) and "failure" otherwise, and likewise for the \( Y \)-observation; then condition on \( U = \) number of successes \((0, 1 \text{ or } 2)\) in the data for each stage. The first leads to a simple SPRT about \((\rho, \eta)\) where \( \eta = p_1 q_2 + p_2 q_1 \) and \( \eta \) is assigned an arbitrary but common value under both \( H_0 \) and \( H_1 \). The second leads to a CSPRT which is in effect a SPRT about the Bernoulli sequence of "successes" and "failures" as defined in the trichotomy with all "ties" ignored.

**Test statistic:** \( L_n = s_n \log \rho' - t_n \log \frac{1}{2}(1 + \rho') \)

where \( s_n = \) number of pairs among the first \( n \) stages with \( X \)-successes and \( Y \)-failures

and \( t_n = \) number of untied pairs among the first \( n \) stages (\( X \)-success and \( Y \)-failure or \( X \)-failure and \( Y \)-success).

**Properties:** Approximations to the OC and ASN are well-known [31]; since the test is a simple SPRT about the trichotomous data, it has minimal ASN in \( H_0 \) and in \( H_1 \) among all tests of this strength based on the reduced (trichotomous) data.
TEST 5b: This is the extended double-dichotomy test, described in Part I; the assumptions and notation will be similar to that in Test 5a. The sampling scheme is now $S_1^*$ (see accompanying notation in the Introduction of Part II).

Assumptions: $A_3$ and $S_1^*$.

Hypotheses: $H_0: \rho = 1$, $H_1: \rho = \rho'$.

Method: Methods (2) and (3)--data reduction to the stage-wise double-dichotomy, letting $s'_n$ and $s''_n$ denote the numbers of X-successes and Y-successes at stage $n$, and $m'_n - s'_n$ and $m''_n - s''_n$ the respective numbers of failures; then conditioning on $u_n = s'_n + s''_n$. The test statistic is the cumulative sum of stage-wise statistics, each being a conditional likelihood ratio of $s'_n$ given $u_n$ (parameterized by $\rho$). See Part I. All stages with $s'_n$ or $s''_n = 0$ are in effect ignored since their log c.l.r.'s are zero.

Test statistic: $L_n = \sum_{i=1}^{n} L_i^*$ and $L^*_n = s'_n \log \rho' - \log \sum_{j=0}^{U_n} \frac{u_n \binom{m}{r} \binom{m'-j}{r'} \rho'}{\binom{m''}{r''}}$.

Properties: OC is not known; ASN under $H_0$ and $H_1$ may be approximated (see Appendix I).

PERMUTATION TESTS

We now describe a test which is distribution-free with regard to the null hypothesis, but is directed against parametric alternatives. We here describe the alternatives as a certain family of normal hypotheses about the X- and Y-populations, but the alternatives may be enlarged to include certain other exponential family distributions; see Part I (p. 30).

We introduce the assumption:

ASSUMPTION A5: The X- and Y-populations are independent with continuous distribution functions, denoted $F$ and $G$, respectively.

Actually, in this section, $F$ and $G$ may vary from stage to stage.

We denote $\Delta = \delta / \sigma^2$, $\delta$ and $\sigma^2$ having been defined in the section on NORMAL TESTS; $\lambda$ and $\sigma^2$ are nuisance parameters under the alternative (and they may
vary from stage to stage), and $\Delta'$ is a specified number.

**TEST 6:**

**Assumptions:** A5 and Sl$^*$.  

**Hypotheses:** $H_0$: $G = F$ (unspecifed), $H_1$: $A$ and $\Delta = \Delta'$.  

**Method:** Method(3). A CSPRT is constructed by conditioning on the stage-wise order statistics, which may be shown to be sufficient for $H_1$ as well as for $H_0$. At a particular stage $n$, an unordered collection $u_n$ of the values of the $n'$ $X$-observations and $n''$ $Y$-observations is fixed ($u_n$ is 1 to 1 in the order statistic). Let $j$ index the $k_n = \binom{n'+n''}{n''}$ ways of selecting $n''$ numbers out of the collection $u_n$ to play the role of the $Y$-observations. The conditional probability that any particular $n''$ of them, say $y_1', \ldots, y_{n''}$, are the $Y$-observations is $1/k_n$ under $H_0$, and is found to be $e^{\Delta'}\sum_{j=1}^{k_n} e^{\Delta'\sum_{j=1}^{n''} y_{1}(j)}$ under $H_1$ where $y_{1}(j), \ldots, y_{n''}(j)$ are the (unordered) numbers making up the $j^{th}$ selection.

Hence, the c.l.r. for the stage $n$ data is this latter ratio with the sum in the denominator replaced by the average—that is, the sum is pre-multiplied by $1/k_n L_n^*$ is the log of this c.l.r.

**Test statistic:** $L_n = \sum_{i=1}^{n} L_i^*$ where $L_i^* = \Delta' s_{n1} - \log[\frac{1}{k_n} \sum_{j=1}^{n''} e^{\Delta's_{nj}}]$  

where $s_{n1}$ is the sum of the stage $n$ $Y$-observations,  

$s_{nj}$ is the sum of the numbers making up the $j^{th}$ selection of $n''$ numbers from among the $X$- and $Y$-observations at stage $n$ (the $k_n$ selections are ordered in an arbitrary way except that the actual $Y$-observations make up the first selection).  

$k_n = \binom{n'+n''}{n''}$.  

44
(An alternative form is: $L^*_n = -\log \left[ \frac{1}{k_n} \sum_{j=1}^{k_n} e^{\Delta'(s_{n_j} - s_{n_1})} \right]$.)

Properties: The OC is unknown, but may be shown to be monotone in $\Delta$ under $H_1$ (normal populations). The ASN is approximable under $H_0$ and $H_1$, at least for small $\Delta'$, where it approximates the ASN for Test 1 under $H_1$—that is, Test 6 is asymptotically efficient (as $\Delta' \to 0$).

See Appendix 2.

RANK TESTS

We now consider nonparametric rank tests, continuing to assume continuous distributions of $X$ and $Y$ (or of $Z = Y - X$). But now both $H_0$ and $H_1$ will be distribution-free, $H_1$ being a "power" (Lehmann) alternative. Instead of such power alternatives, other distribution-free families could be considered with $G = h(F)$ (rather than $G = F^{k'}$) for other specified monotone $h(\cdot)$. These are the first tests which are not completely valid if $X$ and $Y$ are indeed normally distributed; $H_0$ and $H_1$ are not both compatible with normality (see discussion in [32]).

The first of these tests is described (and derived) in [16], and also—by a different approach—in [26] and some of its properties derived in [30]. The second test (which may seem somewhat artificial in its formulation, especially with power alternatives) is new, but mentioned in [16], page 594. Group rank tests were introduced in [32].

$F$ and $G$ are the distribution functions of $X$ and $Y$; $H$ denotes the distribution function of $Z = Y - X$, $p = \text{Prob}(Z > 0)$, $H^+$ denotes the conditional distribution function of $Z$ given that $Z > 0$, and $H^-$ denotes the conditional distribution function of $-Z$ given that $Z < 0$; $k'$ is a specified number exceeding unity, and $p'$ is a specified number in the interval $(\frac{1}{2}, 1)$. We
shall use Assumption A5 or:

**ASSUMPTION A6:** Z has a continuous distribution function.

This assumption does not require the independence of X and Y; the joint distribution of (X, Y) need not remain constant from pair to pair, but the differences are assumed to be independent and identically distributed.

**TEST 7a:**

**Assumptions:** A5 and S1.

**Hypotheses:** $H_0: G = F$ (unspecified), $H_1: G = F^{k'}$.

**Method:** Method (2) -- invariance under a common monotone continuous transformation applied to every observation. The test is based on the l.r. of the ranks of the accumulated Y-observations in a combined ranking of all (accumulated) X- and Y-observations. The relevant distributions are given in [10], p. 192 (and in [22], p. 256, where $s_{j+1}$ is twice misprinted as $s_{j+1}$); an alternative expression appears in [30].

**Test statistic:**
\[
L_n = n'' \log k' + \sum_{j=1}^{n''} \log \left( \frac{\Gamma(s_{j+n''}+j'-j)}{\Gamma(s_{j+n''}+j'-j-k''+1)} \right)
- \log \left( \frac{\Gamma(n'+n''+1)}{\Gamma(n'+n''+1)} \right)
\]

where $s_{1n'', \ldots, s_{n''n}$ are the ranks (integers between 1 and n'+n''), in increasing order, of \{y_{1}, \ldots, y_{n''}\} (the accumulated Y-observations) among \{x_{1}, \ldots, x_{n}, y_{1}, \ldots, y_{n''}\} at stage n.

**Properties:** Sure termination is proved in [30], but the OC and ASN are unknown, nor is any monotonicity of the OC known.

**TEST 7b:**

**Assumptions:** A6 and S2.

**Hypotheses:** $H_0: p = \frac{1}{2}$ and $H_+ = H_-$ (Z is symmetrically distributed
about the origin), \( H_0: p = p' \) and \( H^+ = H^- = k' \).

**Method:** Method (2)—data reduction to the differences \( Z_1, Z_2, \ldots \), and invariance under a common monotone continuous transformation applied to the absolute value of every difference. The test is based on the l.r. of the number of positive Z-values and the ranks of the positive Z-values in a combined ranking of the absolute values of the accumulated Z-values. See [32], p. 204.

**Test statistic:**
\[
L_n = n^+ \log p' + n^- \log (1-p') + n^+ \log k'
+ \sum_{j=1}^{n^-} \log \frac{\Gamma(s_{jn} + j - 1)}{\Gamma(s_{jn} + j - 1 - k' + 1)} - \log[\Gamma(n^- + n^+ k' + 1)/n!]
\]

where \( n^+ = \) accumulated number of positive Z's at stage \( n \),
\( n^- = n - n^+ \), and
\( s_{jn}, \ldots, s_{n^+ n} \) are the ranks in increasing order of the positive Z's among the absolute values of all the accumulated Z's.

**Properties:** Unknown (but presumably termination can be proved as in [30]).

**TESTS 7c and 7d:**

Group tests, analogous to Tests 7a and 7b, respectively, based on a statistic \( L_n \) computed from the data from each stage separately and then summed over stages. (Test 7d uses sampling scheme S2\( ^* \).) Test 7c is described, and its properties investigated, in [32]; see also the comments in [16]. Test 7d could be investigated in an analogous way.

**TESTS 7e and 7f:**

Group tests, again analogous to Tests 7a and 7b, respectively, and to Tests 7c and 7d. But these tests are based on the stage-wise l.r. of the rank sum \( \sum s_{jn} \) (the sum of the ranks of the stage \( n \) Y's, or positive Z's, in a combined ranking of the stage \( n \) data only). Test 7e
is described, and its properties investigated, in [32]; Test 7f could likewise be investigated. These group rank sum tests are feasible since the exact distribution of the rank sum can be worked out under $H_0$ and under $H_1$—at least in small samples (small group sizes); non-group versions are not feasible since invariance and sufficiency theory is not applicable, and the required joint distributions quickly become unmanageable. The test statistics for carrying out these tests are not presented here; see [32].
Appendix I

ASN OF THE EXTENDED DOUBLE-DICHOTOMY TEST

In this appendix we shall sketch an investigation of the ASN of the extended double-dichotomy test (Test 5b, p. 43; see also Part I, pp. 24-27). A more thorough investigation seems warranted, however. We shall assume a regular sampling scheme, with \( m' \) X-observations and \( m'' \) Y-observations per stage, and assume the basic observations are independent and identically distributed.

It was noted in Part I (pp. 21-22) that, under the above assumptions, Wald's method of approximating the ASN is applicable to CSFRT's when \( H_0 \) or \( H_1 \) is true. For, ignoring overshoot, and with \( \log R_N = L_N^* \), \( \log r_n = L_n^* \),

\[
E_L N = \begin{cases} 
- c(\alpha, \beta) = - c_0 & \text{if } H_0 \text{ is true} \\
\quad c(\beta, \alpha) = c_1 & \text{if } H_1 \text{ is true}
\end{cases}
\]

where \( c(\alpha, \beta) = (1-\alpha) \log((1-\alpha)/\beta) - \alpha \log((1-\beta)/\alpha) \); also, Wald's equation is applicable and hence \( E_L N = E N \cdot E_L 1^* \) since \( L_N = \sum_{j=1}^{N} L_j^* \) and the \( L_j^* \)'s are independent and identically distributed. Here \( N \) represents the number of stages. If we let \( M = m' + m'' \), the number of observations per stage, then the expected number of observations is

\[
M E_i N = M c_i / E_i L_1^* \quad (i = 0, 1)
\]

where \( E_i \) denotes expectation when some specific distribution in \( H_i \) is true, and the subscript has been deleted from \( L_1^* \). It thus remains to evaluate \( E_1 L^*/M \) -- the expected increment (per observation) in the cumulative sum defining the test. We shall evaluate this when \( \rho' \), the hypothesized value of the odds ratio under \( H_1 \), is close to the null hypothesis value of unity.

Using the notation of pp. 40-43, we find the stage-wise log c.l.r. to be

\[
L^* = - \log \{ \rho'^{-s} \sum_{j=0}^{u} \left( \frac{M-u}{M} \right)^{u-j} \rho'^{j} \} \quad (2)
\]

(we have replaced \( s'_n \) by \( s \), \( u'_n \) by \( u \), and \( m' + m'' \) by \( M \)).

We now set \( \varepsilon = \rho' - 1 \) (> 0), expand \( (1+\varepsilon)^j \) in (2) as a binomial sum, interchange the order of summation, expand the \( \log(1+\varepsilon) \) term, and thus find

49
\[ L^* = -(a_1-s) \epsilon - \left( a_2 - \frac{1}{2} a_1^2 + \frac{1}{2} s \right) \epsilon^2 - \left( a_3 - a_1 a_2 + \frac{1}{3} a_1^3 - \frac{1}{3} s \right) \epsilon^3 \]
\[ - \left( a_4 - a_1 a_3 - \frac{1}{2} a_2^2 + a_1 a_2 - \frac{1}{4} a_1^4 + \frac{1}{4} s \right) \epsilon^4 + O(\epsilon^5) \] (3)

where \( a_j = \binom{u}{j} \binom{m'}{j} / \binom{M}{j} \).

Now \( S \) is (marginally) a binomial random variable with parameters \((m', p_1)\), and \( U \) is (marginally) the sum of two independent binomial random variables with parameters \((m', p_1)\) and \((m'', p_2)\), so that moments of \( S \) and \( U \) are readily available. Taking expectations in (3) term-by-term, we find (after some labor)

\[
\frac{1}{M} E L^* = \pi(1-\pi)(p_1-p_2) \epsilon + \frac{1}{2} \pi(1-\pi) \epsilon^2 \left( \frac{M}{M-1} \left[ \pi p_1 + (1-\pi) p_2 \right]^2 \right) \]
\[ - \frac{1}{M-1} \left[ \pi p_1^2 + (1-\pi) p_2^2 \right] - p_1 - \pi(p_1-p_2) \] + \( O(\epsilon^3) \) \hfill (4)

where \( \pi = m'/M \), the proportion of the observations in any stage which are from the \( X \)-population.

The modulus of the RHS of (4) is seen to be maximized (and hence \( E L^* \) minimized) when \( \pi = \frac{1}{2} \). Hereafter, we shall confine attention to this case for simplicity. We now examine \( E L^* \) separately for \( H_0 \) and for \( H_1 \).

**Case 1:** \( H_0 \) true. \( m' = m'' = M/2, p_1 = p_2 = p \) (say), and \( q = 1-p \).

Then (4) becomes, with some additional terms in (3) also evaluated,

\[
\frac{1}{M} E_0 L^* = - \frac{1}{8} pq \epsilon^2 \left[ 1 - \epsilon + \frac{1}{8} \epsilon^2 \left( 7 + \frac{M-2}{M-1} pq \right) \right] + O(\epsilon^5) \] \hfill (5)

This expression is increasing in \( M \) (for \( \epsilon \) small), but very slowly, and hence the approximate ASN under \( H_0 \) is seen to decrease (slowly) with increasing batch size \( M \); on the other hand, if \( M \) is too large, the overshoot cannot safely be ignored and the approximation to the ASN would not be valid. It would appear that there is an optimal batch size (for small \( \epsilon \), i.e., \( \rho' \) close to unity), presumably greater than \( M = 2 \), but that not much improvement over the case \( M = 2 \) is typically possible.

Let \( e_{M, M} \) denote the ratio of ASN approximations (equation (1)) when the batch size is 2 (numerator) and when the batch size is \( M \) (denominator), when \( H_0 \) is true and the common \( p_1 \) value is \( p \). (When the batch size is 2, the test is Wald's double-dichotomy test.) Then \( e_{M, M} \) is a measure of efficiency.
under $H_0$ of the extended test (with $m' = m''$), relative to Wald's test. We find

$$e_{M,p} = 1 + \frac{1}{8} pq \epsilon^2 \frac{M-2}{M-1} + O(\epsilon^3)$$

where it is to be recalled that $\epsilon = \rho' - 1$ and $\rho' (> 1)$ is the value of the odds ratio under $H_1$. Thus, the efficiency under $H_0$ is slightly improved with $M > 2$, uniformly in $p$; but it never exceeds $1 + pq \epsilon^2 / 8 + O(\epsilon^3)$, which is at most $1 + \epsilon^2 / 32 + O(\epsilon^3)$. This supports the conclusion drawn in the previous paragraph.

Case 2: $H_1$ true. $m' = m'' = M/2$, $p_1q_2/p_2q_1 = \rho'$. Then (4) becomes

$$\frac{1}{M} E_{L^*} = \frac{1}{4} (p_1 - p_2) \epsilon - \frac{1}{32} \epsilon^2 \left[ \frac{1}{M-1} (p_1 - p_2)^2 - (p_1 + p_2)^2 \right] + 2(p_1 + p_2) + \frac{1}{8} (p_1 - p_2) + O(\epsilon^3), \quad (6)$$

which is also increasing in $M$ (for $\epsilon$ small).

Letting $e_{M; p_1, p_2}$ be the relative efficiency under $H_1$ of the extended test (with $m' = m''$ and $p_1q_2/p_2q_1 = \rho'$) relative to Wald's test, we find from (1) and (6)

$$e_{M; p_1, p_2} = 1 + \frac{1}{8} (p_1 - p_2) \epsilon \frac{M-2}{M-1} + O(\epsilon^3). \quad (7)$$

Again, the efficiency is slightly increased by choosing $M > 2$. The least upper bound on the second order term in (7) may be shown to be $\epsilon^2 / 32$, the same value as obtained in case 1 above. Hence, only a small percentage reduction in the ASN, under both $H_0$ and $H_1$, can be achieved by increasing the batch size.

The ASN can be evaluated from (1), (5) and (6); a few values of the constant $c_1$, when $\alpha = \beta$, are given below:

$$\begin{array}{cccc}
\alpha & 0.1 & 0.05 & 0.01 & 0.001 \\
c_1 = c(\alpha, \alpha) & 1.76 & 2.6^* & 4.50 & 6.91 \\
\end{array}$$

51
APPENDIX 2

ASH OF THE PERMUTATION TEST

In this appendix we shall briefly investigate the ASN of the sequential permutation test (Test 6) when sampling in pairs and when the observations are actually normal. A more thorough investigation, under broader assumptions, is warranted, however.

Noting the remarks in Appendix 1 on the ASN's of CSPRT's, we only need evaluate the expected increment, $E L^*$, for the permutation test in order to approximate the ASN under $H_0$ and $H_1$. When sampling in pairs (S2), and with $Z = Y - X$, a typical increment $L^*$ is given by

$$L^* = \log[2/(1 + e^{-\Delta'Z})].$$

We shall approximate its expectation under the assumption that $Z$ is normal and $\Delta'$ is small. (Recall that $\Delta'$ is the hypothesized value for the mean of $Z$ divided by one-half the variance of $Z$ under $H_1$, and is positive.) We proceed as follows:

Let $g(w) = - \log \frac{1}{2}(1 + e^{-w})$ so that $L^* = g(\Delta'Z)$. Expanding $g$ in a Maclaurin series, we find

$$g(w) = -\frac{1}{2}w - \frac{1}{6}w^2 + \frac{1}{192}w^4 - \frac{1}{2880}w^6 + O(w^8)$$

as $w \to 0$. Let $W$ be $N(\mu, 2\sigma)$ and assume that $\mu = O(h)$ and $\sigma = O(h)$ for some $h \to 0$. Taking expectation of $g(W)$ term-by-term, we find (after some labor)

$$E g(W) = \frac{1}{2} \mu - \frac{1}{4} \sigma^2 - \frac{1}{8} \mu^2 + \frac{1}{16} \sigma^4 - \frac{1}{24} \mu^3 + O(h^4). \quad (6)$$

Now let $H_0'$ be the hypothesis that $Z$ is $N(0, 2\sigma^2)$ and $H_1'$ that $Z$ is $N(\delta', 2\sigma^2)$ where $\delta' = \Delta'\sigma^2$; these are compatible with the hypotheses $H_0$ and $H_1$, respectively, which are being tested by the permutation test. Letting $h = (\Delta'\sigma)^2$ and $W = \Delta'Z$, we have equivalently:

$$H_0': W \text{ is } N(0, 2h), \quad H_1': W \text{ is } N(h, 2h)$$

and $L^* = g(W)$. Hence, from (6), we have
\[ E_0 L^* = -\frac{1}{4} h^* + \frac{1}{16} h^2 - \frac{1}{24} h^3 + 0(h^4) \] 
\[ E_1 L^* = \frac{1}{4} h - \frac{1}{16} h^2 + \frac{1}{45} h^3 + 0(h^4) \] 
\[ (9) \]

where \( E_1 \) denotes expectation under \( H_1 \).

For the normal theory test of \( H_0 \) vs. \( H_1 \) for specified \( \sigma \) (Test 1a or 1b), we find the corresponding increment, say \( L^{**} \), is given by

\[ L^{**} = \frac{1}{2} \Delta'Z - \frac{1}{4} \Delta'^2 \sigma^2 = \frac{1}{2} U - \frac{1}{4} h \]

so that

\[ E_0 L^{**} = -\frac{1}{4} h, \quad E_1 L^{**} = \frac{1}{4} h. \] 
\[ (10) \]

Note that, to the first order of approximation, (9) and (10) agree. Consequently, to the first order of approximation, the ASN's of the two tests—the permutation test (test 6) and the normal SPRT (Test 1b)—agree when one of the normal test hypotheses is true. The approximation is valid for small \( h = (\Delta' \sigma)^2 = (\delta'/\sigma)^2 \).

The approximate ASN's (actually, expected number of stages, with two observations per stage) for the normal test are \( E_1 N = 4c_1/h \), where \( c_1 \) is defined in Appendix 1. For the permutation test, \( E_1 N = (4c_1/h) \cdot (1 + \frac{1}{4}h + 0(h^2)) \). The ratio of these ASN's provides a measure of efficiency: namely, \( 1 - \frac{1}{4} h + 0(h^2) \), of the permutation test relative to the normal test. The efficiency is less than unity, but is asymptotically \( (h \rightarrow 0) \) unity.

Subtracting the normal test ASN from the corresponding permutation test ASN, we obtain \( c_1 + O(h) \). Hence, the permutation test requires approximately \( c_1 \) additional pairs of observations, compared with the normal test, when sampling in pairs, when \( h \) is small, and when \( H_1 \) is true. Some values of \( c_1 \) are given at the end of Appendix 1 (p. 51).
REFERENCES


