SEQUENTIAL EVALUATION OF GRADED PREFERENCES FOR TWO TREATMENTS

By W. J. Hall
University of North Carolina

Institute of Statistics Mimeo Series No. 486

August 1966

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.

SEQUENTIAL EVALUATION OF GRADED PREFERENCES FOR TWO TREATMENTS

by W. J. Hall

University of North Carolina at Chapel Hill

Suppose a sequence of subjects each in turn indicates a graded preference for one of two treatments, A and B; for example, a strong or weak preference may be registered for either of the treatments, or no preference registered. A sequential test of the hypothesis that, in each preference grade, A is just as likely to be preferred as B, is presented. OC and ASN functions are derived. Allowance is made for certain subject-to-subject differences. The test is a conditional sequential probability ratio test and may be considered as a simple extension of Wald's sequential test for double dichotomies, applicable in this context if only one preference grade (and a no preference grade) were considered. Two-sided alternative versions and analogous non-sequential tests are also described. Possible applications include various kinds of consumer testing, psychological testing, and subjective aspects of clinical trials. The tests are also suitable for testing certain hypotheses about 2x2, or 2xk, tables.

1. THE PROBLEM

A sequence of subjects independently compare treatments A and B. Each subject in turn expresses either a strong preference (for A or for B), a weak preference (for A or for B), or no preference. (The extension to a larger number of preference categories will be obvious.) We denote the datum x from a single subject as

$x = \underline{A}$ (strong preference for A) , $x = \underline{B}$ (strong preference for B)

$x = \underline{a}$ (weak preference for A) , $x = \underline{b}$ (weak preference for B)

$x = \underline{0}$ (no preference) .

A subscript n will designate the n^{th} subject.

For example, hospitalized patients may be questioned on their preferences for two alternative ways of carrying out certain hospital routines, or their preferences for two alternative pain-relieving medications, or their preferences for capsule or pill versions of the same m dication. Or, in consumer testing of new designs for automobiles, preferences for alternative silhouette drawings may be solicited.

The preference ratings may reflect qualitative judgments. Alternatively, a (single) measurement might be made reflecting the effect of each treatment on each subject; letting y represent the difference between the A-measurement (a large y-value indicating the superiority of A, say), the y-values may be categorized by writing

$x(y) = \underline{A}$ if $y > \Delta$, $x(y) = \underline{B}$ if $y < -\Delta$,

$x(y) = \underline{a}$ if $\delta < y \leq \Delta$, $x(y) = \underline{b}$ if $-\Delta \leq y < -\delta$,

$x(y) = \underline{0}$ if $-\delta \leq y \leq \delta$

for specified $\Delta > \delta \geq 0$, determining graded preference ratings as before. Thus, the tests to be presented may be considered non-parametric tests for measurement data.

We introduce a grade indicator $t = t(x)$ which equals 2, 1, or 0, respectively, according as x is a strong, weak, or null preference.

The following probabilities are introduced, as parameters of the model:

p_{An}, p_{Bn}, q_{An}, q_{Bn} = probability that subject n registers a strong preference for A ($x_n = \underline{A}$), for B ($x_n = \underline{B}$), a weak preference for A ($x_n = \underline{a}$), for B ($x_n = \underline{b}$), respectively

p_n, q_n, r_n, = probability that subject n registers a strong ($t_n = 2$), weak ($t_n = 1$), null ($t_n = 0$) preference, respectively

σ_n, τ_n = conditional probability of a preference for A, given that subject n registers a strong (weak, respectively) preference.

Thus, $p_{An} + p_{Bn} = p_n$, $q_{An} + q_{Bn} = q_n$, $p_n + q_n + r_n = 1$, $\sigma_n = p_{An}/p_n$, and $\tau_n = q_{An}/q_n$, for each n. (To simplify the exposition, it will usually be assumed that both p_n and q_n are strictly positive, but such a restriction is certainly not necessary.) For each n, the four parameters (p_{An}, p_{Bn}, q_{An}, q_{Bn}), or equivalently (p_n, q_n, σ_n, τ_n), describe the model.

The hypotheses to be considered, for specified constants σ' and τ', each between 1/2 and 1, are:

H_0: $\sigma_n = \tau_n = 1/2$, all n (or $p_{An} = p_{Bn}$ and $q_{An} = q_{Bn}$)

H_A: $\sigma_n = \sigma'$ and $\tau_n = \tau'$, all n (or $p_{An} = \lambda' p_{Bn}$ and $q_{An} = \mu' q_{Bn}$ where $\lambda' = \sigma'/(1-\sigma')$ and $\mu' = \tau'/(1-\tau')$)

H_B: $\sigma_n = 1-\sigma'$ and $\tau_n = 1-\tau'$, all n (or $p_{Bn} = \lambda' p_{An}$ and $q_{Bn} = \mu' q_{An}$)

H: the union of H_A and H_B.

H_0 is the null hypothesis of no difference in treatments, H_A is a hypothesis of the superiority of treatment A, and H_B of the superiority of treatment B, and H is a (two-sided) negation of H_0. Typically, it would be reasonable to choose $\sigma' > \tau' > 1/2$. For example, if the preferences are based on measurements and the difference y between the A-measurement and the B-measurement

is normally distributed with positive mean, then the conditional probability σ that $y > \Delta$ given that $|y| > \Delta$ exceeds the conditional probability τ that $\delta < y \leq \Delta$ given that $\delta < |y| \leq \Delta$ (and both are greater than 1/2). It is to be emphasized that all hypotheses are composite; the parameters p_n and q_n (and $r_n = 1 - p_n - q_n$)are not specified by any of the hypotheses, and may vary from subject to subject.

It is a useful fact that the distribution of the grade indicator t_n depends on the nuisance parameters (p_n, q_n) while the conditional distribution of x_n given t_n depends only on the parameters of interest (σ_n, τ_n). (This is a type of conditional sufficiency of t_n introduced by Fraser [3], which implies a separability of the parameters [4].)

In Section 2, we consider the problem of sequentially testing H_0 vs. H_A, and present a sequential test of prescribed strength (α,β); in Section 3 we derive properties of this test. In Section 4 a non-sequential version of this one-sided test is presented, together with an analogous alternative sequential version. In Section 5 we consider the problem of testing H_0 against the two-sided alternative H (or the three-decision problem of choosing among H_0, H_A, H_B). A brief comment on accounting for effects due to the order of treatment is in Section 6.

All tests are based on an appropriate composite likelihood ratio, made up of conditional likelihood ratios for testing binomial hypotheses (about σ_n and about τ_n) within each preference grade. If $\sigma' = \tau'$ the strong and weak grades can be combined into one, or if $\tau' = 1/2$ the weak and no preference grades can be combined; in either of these cases, the tests reduce to known ones (Wald [11], Chapter 6).

Two extensions are easily made: one (already noted) is that of

including more than two preference grades in addition to the no preference grade; and the other is that of replacing the testing of H_0 vs. H_A with the testing of two hypotheses like H_A (or one like H_A and one like H_B) but with differing σ'- and τ'-values—say (σ_0, τ_0) vs. (σ_1, τ_1)—and replacing the testing of H_0 vs. H with the testing of two hypotheses like H (with differing σ'- and τ' -values). Actually, the tests as given are valid for somewhat more general hypotheses than those stated; explicitly, the one-sided test is valid for testing

H_0^* : σ_n constant and $\leq 1/2$ and τ_n constant and $\leq 1/2$

vs.

H_A^* : σ_n constant and $\geq \sigma'$ and τ_n constant and $\geq \tau'$;

it is presumed (but not proved) that the tests would remain valid even if σ_n and τ_n were not constant but nevertheless satisfied the inequalities of H_0^* or H_A^*.

The tests may also be considered as tests of certain hypotheses about the parameters of 2x2 (or 2 x k) tables. Specifically, ignoring the no preference category data, and with preference categories 1, 2,..., k (the columns of the table, A and B labelling the rows), one can test the hypothesis that $\sigma_{(1)} = \sigma_{(1)}', \ldots, \sigma_{(k)} = \sigma_{(k)}'$ against $\sigma_{(1)} = \sigma_{(1)}'', \ldots, \sigma_{(k)} = \sigma_{(k)}''$ where $\sigma_{(i)}$ is the conditional probability of preference for A, given that the preference is of type (strength) i.

2. THE ONE-SIDED SEQUENTIAL TEST

We now introduce a conditional sequential probability ratio test (CSPRT) of H_0 vs. H_A, conditional on the sequence of grade indicators $t_1, t_2, \ldots, t_n, \ldots$. (A general exposition of such CSPRT's is given in [4], as

an extension of the method introduced by Wald [12] in the double dichotomy problem.) At the n^{th} stage, we consider the conditional likelihood ratio of the two hypotheses (H_A in the numerator and H_0 in the denominator), given t_n. Each of these hypotheses becomes a simple hypothesis about the conditional distribution (since t_n may be shown to be sufficient for each of the composite hypotheses H_0 and H_A). Letting z_n represent the logarithm (natural) of the stage n ratio, the logarithm z_n of the joint conditional likelihood ratio of the data from the first n stages is the accumulated sum of the stage-wise log ratios: $Z_n = \Sigma_{i=1}^{n} z_i$. The CSPRT is a SPRT based on the conditional log ratio Z_n. Thus, at each successive stage n, the test is terminated in favor of H_A, continued on to the next stage, or terminated in favor of H_0 according as $Z_n \geq a$, $b < Z_n < a$, or $Z_n \leq b$, where $a = \log [(1-\beta)/\alpha]$ and $b = \log [\beta/(1-\alpha)]$, and α and β are the prescribed error probabilities.

According to the basic theory of SPRT's [11], the conditional error probabilities (still conditional on the sequence of indicators) will be approximately α and β, the approximation being due solely to the possibility of overshooting the stopping boundaries. Since these (approximate) conditional error probabilities do not depend on the t-sequence, they are also the unconditional error probabilities (still approximately). Thus, the test has the prescribed <u>strength</u> (α, β).

Such conditional tests are common in non-sequential theory (e.g., Fisher's exact test for 2x2 tables, conditional on fixed marginal totals; see Lehmann [8], Chapter 4, for this and other examples). A sequential example was introduced by Wald for comparing two Bernoulli sequences when sampling in pairs (see Wald [11], Chapter 6, and Armitage [2]; other

sequential examples are described in [4]). Wald's test may be considered as a test for ungraded preferences, the "tied pairs" (both successes or both failures) being the no preference category and the united pairs being the single preference category (which we have subdivided into strong and weak preferences). The test described below is a simple extension to the case of graded preferences, and it reduces to Wald's test if the grades are merged (it is sufficient to set $\sigma' = \tau'$ in H_A) or if the weak preferences are treated as null preferences (set $\tau' = 1/2$). In both Wald's test and ours, the conditional likelihood ratio for the no preference category (his tied pairs) is unity so that observations in this category are effectively ignored. In Wald's case, all that remains is a single Bernoulli sequence—the untied pairs, each favoring either A or B. In our case, two interspersed Bernoulli sequences remain—a strong preference sequence and a weak preference sequence, occurring in a random order. We, in effect, run a test for each Bernoulli sequence, but the stage-wise log likelihood ratios are accumulated in a single sum Z_n rather than keeping the two sums separate. We now develop a formula for Z_n.

The conditional likelihoods for the datum from the n^{th} subject are given the following table:

condition: t_n =	datum: x_n =	conditional likelihood
2	$\underline{A}$, $\underline{B}$	$p_{An}/p_n = \sigma_n$, $p_{Bn}/p_n = 1-\sigma_n$
1	$\underline{a}$, $\underline{b}$	$q_{An}/q_n = \tau_n$, $q_{Bn}/q_n = 1-\tau_n$
0	$\underline{0}$	1

All other combinations of conditions and data have a zero likelihood. The conditional likelihood ratios (H_A over H_0) for the n^{th} subject and their natural logarithms are thereby found and presented below:

datum: x_n =	conditional ratio	log ratio z_n =
$\underline{A}$	$\sigma' / \frac{1}{2}$	$\log 2\sigma' = c_A$
$\underline{B}$	$(1-\sigma') / \frac{1}{2}$	$\log 2(1-\sigma') = -c_B$
$\underline{a}$	$\tau' / \frac{1}{2}$	$\log 2\tau' = d_A$
$\underline{b}$	$(1-\tau') / \frac{1}{2}$	$\log 2(1-\tau') = -d_B$
$\underline{0}$	1 / 1	0

The joint log ratio, on which the test is based, is $Z_n = \Sigma_{i=1}^{n} z_i$, which may be expressed as

$$Z_n = s_{An} c_A - s_{Bn} c_B + w_{An} d_A - w_{Bn} d_B$$

where s_{An}, s_{Bn}, w_{An}, and w_{Bn} are the accumulated numbers of strong preferences for A, strong preferences for B , weak preferences for A, and weak preferences for B , respectively, at stage n. If weak preferences were merged with the strong or with the no preference category, Z_n would reduce to the appropriate Z_n for Wald's double dichotomy test.

The test may be carried out by plotting the cumulative sum Z_n on a chart and terminating as soon as Z_n intersects one of the horizontal rejection lines with ordinates a and b .

We now present an alternative derivation of this test which will be useful in studying its properties. We shall show that the test is not only a CSPRT, but a bona fide SPRT of two simple hypotheses H_0' and H_A' about the sequence of data, where H_0' and H_A' are included within the composite

hypotheses H_0 and H_A, respectively. (That this is possible is due to the fact that the parameters of interest (σ_n, τ_n) and the nuisance parameters (p_n, q_n) are separable, as discussed in Method (1) in [4].) To this end, let p' and q' be arbitrary positive numbers with $p' + q' \leq 1$, and make both of the hypotheses H_0 and H_A simple by affixing primes and adding the restriction that $p_n = p'$ and $q_n = q'$ (we could let p' and q' vary with n but have no need to). The model is then that each observation is classified into one of five categories (the range of x_n) with respective probabilities p_{An}, p_{Bn}, q_{An}, q_{Bn}, and r_n. Under H_A' these probabilities are $\sigma'p'$, $(1-\sigma')p'$, $\tau'q'$, $(1-\tau')q'$, and $1-p'-q'$, while under H_0 they are analogous with σ' and τ' replaced by $1/2$. The likelihood ratio (H_A' over H_0') is then seen to be identical with the conditional ratio given in the previous table—as it must since the marginal distribution of t_n (given by p', q' and $1-p'-q'$) is the same under both simple hypotheses. Thus, the CSPRT of composite hypotheses is a SPRT of simple hypotheses—for arbitrary but fixed values p' and q' of the nuisance parameters, common to the two hypotheses.

3. PROPERTIES OF THE ONE-SIDED SEQUENTIAL TEST

We shall now consider some elementary properties of the test under the assumption that none of the parameters vary from subject to subject; that is, the subscript n may be deleted from all parameters. Since the observations $x_1, x_2, \ldots, x_n, \ldots$ are then independent and identically distributed, and since the test may be considered as a SPRT of two simple hypotheses (see previous paragraph), Wald's methods [11] of studying the certainty of termination, the ASN function, and the OC function are applicable. Specifically, the test does terminate with certainty. (It is

easily seen that this occurs quite generally, so long as, for infinitely many values of n, $r_n < 1$.)

We next consider the operating characteristic (OC) function $L(\theta)$, the probability of accepting H_0 as a function of the parameter $\theta = (p_A, p_B, q_A, q_B)$ or(p, q, σ, τ), as convenient. Let $h(\theta)$ be implicitly defined as the unique non-zero solution of the equation $E(e^{zh}) = 1$. (That such a solution exists is well known [11]—except for parameter values for which $E(z) = 0$, and this case will be treated separately below.) The equation readily reduces to the equation of a hyperplane through the origin in the parameter space:

$$(e^{hc_A} - 1)\, p_A + (e^{-hc_B} - 1)\, p_B + (e^{hd_A} - 1)\, q_A + (e^{-hd_B} - 1)\, q_B = 0 .$$

The OC-function is then approximately given by the formula $L = (e^{ah} - 1)/(e^{ah} - e^{bh})$, for parameter values lying on this hyperplane. The OC-function may be determined then by specifying a value for h, determining the corresponding value for L, and then describing the hyperplane of parameter points where L has this value. (If q_A and q_B are taken to be zero, this reduces to the known fact that Wald's double dichotomy test has a constant OC along lines through the origin in the parameter space of (p_A, p_B)—that is, the OC is a function only of the ratio p_A/p_B.) For fixed h (and hence fixed L), and starting at any point on the corresponding hyperplane in the space of (p_A, p_B, q_A, q_B), an increase of either p_A, q_A or both must be accompanied by an increase in p_B, q_B or both.

The approximate OC will be examined in more detail below. We first note that the OC is approximately $a/(a-b)$ when $h = 0$ [11], and this occurs whenever $E(z) = 0$, namely on the hyperplane

$$c_A p_A - c_B p_B + d_A q_A - d_B q_B = 0 .$$

We also note that Wald's upper and lower bounds on the OC are also applicable (section A.2 of [1]) where his $\delta_\theta(h)$ and $\eta_\theta(h)$ are readily found to be

$$\delta_\theta = \max \left[e^{(c_A - d_A)h}, \frac{e^{c_A h} p_A + e^{d_A h} q_A}{p_A + q_A} \right],$$

$$\eta_\theta = \min \left[e^{(d_B - c_B)h}, \frac{e^{-c_B h} p_B + e^{-d_B h} q_B}{p_B + q_B} \right].$$

We now examine the OC as a function of the parameters of interest, namely the true conditional probabilities σ and τ (assumed free of n). Also denote $\rho = q/p$. Dividing the hyperplane equation through by p, we readily obtain the equivalent equation:

$$(e^{hc_A} - e^{-hc_B})\sigma + (e^{hd_A} - e^{-hd_B})\rho\tau - \rho(1 - e^{-hd_B}) - (1 - e^{-hc_B}) = 0.$$

Thus, h (and hence L) is completely determined by σ, τ and ρ. We now hold ρ fixed and consider h and L as functions of (σ, τ) for fixed ρ. We continue to use the approximate formula for L as a function of h, which incidentally is a monotone increasing function. The parameter space of (σ, τ) is the unit square.

Now the h-contours in the (σ, τ)-space—where L and h are constant—are given by the above equation and are seen to be straight lines with negative slopes, the slopes and intercepts being determined by ρ and h. Several properties of the h-contours are readily derived: (1) being distinct, the intersection of any pair of them occurs outside the unit

square; (2) the point (1/2,1/2), consistent with H_0, lies on the h=1 line, and the point (σ', τ'), consistent with H_A, lies on the h= -1 line; (3) h=+∞ at the origin and h= -∞ at (1,1); (4) the h=0 line is given by

$$(c_A + c_B)\ \sigma + (d_A + d_B)\ \rho\tau - (c_B + \rho d_B) = 0 ;$$

(5) the sides of the unit square along the axes correspond to cases in which the test reduces to Wald's test (effectively a binomial test) which has an OC-function and h-function decreasing in the parameter (σ or τ); hence, as one moves away from the origin along either axis, h (and L) is strictly decreasing; and since for suitable choice of ρ each point on the other two sides of the square is on the same h-contour as some point on the axes sides of the square, h must also decrease along these two sides as one moves toward (1,1). Hence, the h-contours within the square, for any fixed ρ , are ordered in decreasing order from (0,0) to (1,1). Hence, on any vertical or horizontal cut through the square, h (and L) are decreasing away from the axes.

The figure shows some h-contours—contours of approximate OC—for the case of testing H_0 against H_A: (σ,τ) = (.8, .6). Then

$$\sigma = -\frac{(1.2)^h - (.8)^h}{(1.6)^h - (.4)^h}\ \rho\tau + \frac{\rho - (.8)^h\rho + 1 - (.4)^h}{(1.6)^h - (.4)^h}$$

and $L(\sigma, \tau,\rho) \simeq (A^h - 1)/(A^h - B^h)$ when $h \neq 0$. (When h=0, the special case considered above applies.)

From the facts listed above we can conclude that, whatever the value of ρ —in fact, whatever the value of the nuisance parameters p and q — the OC is monotone (decreasing) in both σ and τ separately. Hence, any

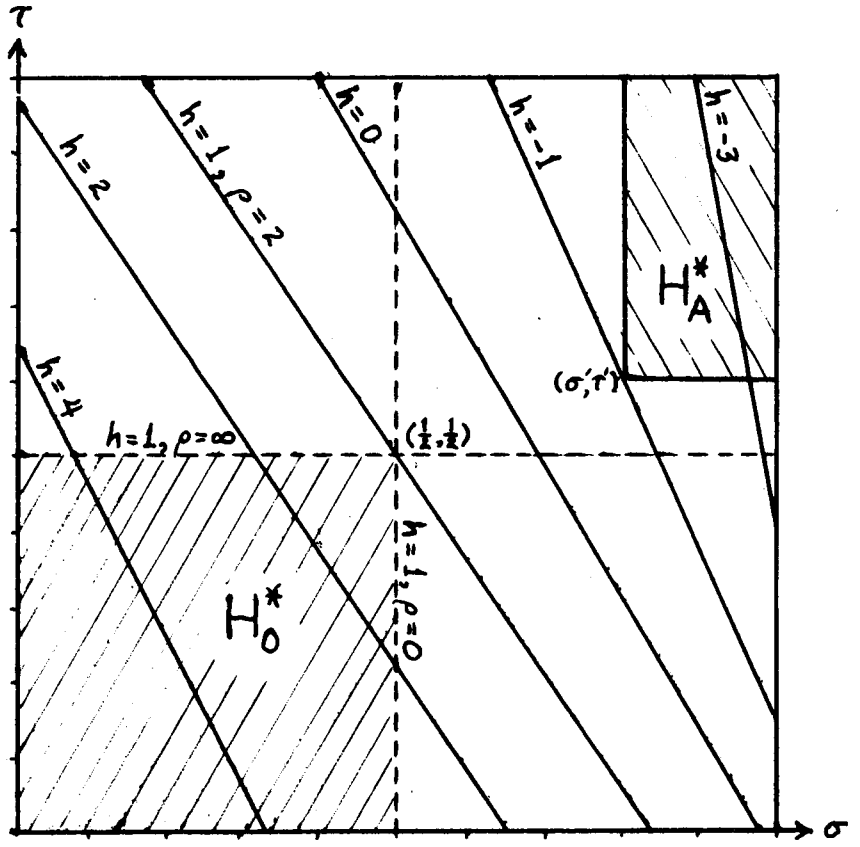

Figure: Contours of constant h (and approximate OC) when $(\sigma', \tau') = (.8,.6)$ and ρ $(= q/p) = 2$.

(For small α and β, the OC $\simeq 1 - \left(\frac{\alpha}{1-\beta}\right)^{h}$ for $h > 1$ and $\simeq \left(\frac{\beta}{1-\alpha}\right)^{-h}$ for $h < -1$.)

point to the lower left of the point (1/2,1/2) in the (σ,τ)-space has OC at least $1-\alpha$ (h at least unity) and any point to the upper right of the point (σ', τ') has OC at most β (h at most -1). We thus conclude that, when all parameters are assumed constant from subject to subject, the test has strength (α, β) for testing H_0^* vs. H_A^*.

It is conjectured that the OC is decreasing in each σ_n and τ_n when there is subject to subject variation in the parameters; if so, the test is valid for testing $\sigma_n \leq 1/2$ and $\tau_n \leq 1/2$ against $\sigma_n \geq \sigma'$ and $\tau_n \geq \tau'$, but this has not been proved. The conjecture seems apparent (when p and q are fixed) from the fact that the expected value of the stage n contribution to the cumulative sum Z_n is an increasing function of both σ_n and τ_n, so that an increase in either of these parameters would decrease the probability of crossing the upper boundary. The effect of a change in p_n at the expense of q_n or r_n is not so clear-cut, however, since an increase in p_n would increase both the chance of a positive contribution (of c_A) to the log likelihood ratio and the chance of a negative contribution (of $-c_B$). However, the direction of the effect on the OC is presumably determined by the direction of the effect on the expected contribution to the log likelihood, and this is easily evaluated.

We now turn to the average sample number (ASN) function, which, except when E(z) vanishes, is approximately related to the OC through Wald's equation: $E(Z_N) = E(N) \cdot E(z)$, together with the approximation $E(Z_N) = a(1 - L(\theta)) + bL(\theta)$, where N is the sample size required by the test and $E(z) = c_A p_A - c_B p_B + d_A q_A - d_B q_B$. When E(z) vanishes, E(N) may be approximated (see [11]) by $-ab/E(z^2)$ where $E(z^2) = c_A^2 p_A + c_B^2 p_B + d_A^2 q_A + d_B^2 q_B$.

Specifically, when H_0 is true, the ASN is approximately

$$\frac{2b + 2(a-b)\alpha}{p \log [4\sigma'(1-\sigma')] + q [\log 4\tau'(1-\tau')]}$$

and when H_A is true the ASN is approximately

$$\frac{a - (a-b)\beta}{p [\sigma' \log 2\sigma' + (1-\sigma') \log 2(1-\sigma')] + q [\tau' \log 2\tau' + (1-\tau') \log 2(1-\tau')]}.$$

Now if we had ignored the weak preferences (by setting $\tau' = 1/2$), the second term in each denominator above would vanish, thereby increasing the ASN; we thus may obtain some measure of the value of utilizing weak preferences. Specifically, the ratio of ASN's when H_0 is true, ignoring weak preferences (numerator) compared with using weak preferences (denominator), is $1 + \rho \log [4\tau'(1-\tau')]/\log[4\sigma'(1-\sigma')]$ which is approximately $1 + \rho\epsilon^2/\delta^2$ when $\epsilon = \tau' - \frac{1}{2}$ and $\delta = \sigma' - \frac{1}{2}$ and both δ and ϵ are small. When H_A is true, the corresponding ratio is also found to be approximately $1 + \rho\, \epsilon^2/\delta^2$.

Finally, we note that the test has Wald's optimal property, uniformly in the nuisance parameters; that is, among all tests of H_0 vs. H_A with the same strength, this test has the smallest ASN, whatever the values of the nuisance parameters p and q .

4. A ONE-SIDED NON-SEQUENTIAL TEST

We now presume that the number of observations is a fixed number n, and present the analogous most powerful conditional test of H_0 vs. H_A, conditional on $t_1, t_2, \ldots, t_n$; such tests have been called tests of Neyman structure [8]. The conditional likelihood ratio Z_n has already been

derived in Section 2. The critical region is of the form $Z_n \geq c$ where c may depend on $t_1,\ldots,t_n$ and is determined so that the test has (approximate) conditional significance level α —and hence unconditional significance level α. (We shall not describe a randomized version which would enable the dropping of the qualification "approximate".) Equivalently, the critical region is of the form $s_{An} \log \lambda' + w_{An} \log \mu' \geq k$ and k is determined as follows: Under H_0, s_{An} and w_{An} have, conditionally, independent binomial distributions with parameters $(1/2,s)$ and $(1/2,w)$ where $s = s_{An} + s_{Bn}$ = number of strong preferences and w = number of weak preferences; thus, k is determined (as a function of s and w) so that the following sum is approximately α:

$$\sum \binom{s}{i}\binom{w}{j}\left(\frac{1}{2}\right)^{s+w} \qquad \begin{array}{l}\text{(the sum being extended over } i \text{ and } j \\ \text{for which } i \log \lambda' + j \log \mu' \geq k).\end{array}$$

It should be noted that the test does depend on the specification of H_A—that is, on λ' and μ' or on σ' and τ'. The critical region is defined by a linear combination of the numbers of strong and weak preferences for A, weighted by $\log \lambda'$ and $\log \mu'$, respectively.

Alternatively, the critical region can be expressed as

$$(s_{An} - s_{Bn}) \log \lambda' + (w_{An} - w_{Bn}) \log \mu' \geq C(s,w) .$$

The left side has (under H_0) conditional mean zero and conditional variance $v^2 = s \log^2\lambda' + w \log^2\mu'$. Then $C(s,w)$ may be determined approximately from standard normal tables as equal to $z_\alpha v$ where z_α is exceeded with probability α according to the standard normal law.

The power of the test can be approximated from the approximate normal-

ity of $v^{-1}(s_{An} - s_{Bn}) \log \lambda' + v^{-1}(w_{An} - w_{Bn}) \log \mu'$; since v is also a random variable in this expression, the unconditional mean and variance formulas are cumbersome and omitted. Of course, the conditional power of the test is readily determinable since s_{An} and w_{An} are conditionally independent and binomially distributed (under any alternative of the form of H_A); this conditional power will clearly vary with s and w, however, in contrast to the sequential test, and therefore one cannot conclude that the power against H_A is a determinable constant. The test is, however, most powerful against H_A among conditional tests.

As already noted, the conditional power for any fixed (s,w) may be determined. One could readily prepare a list of (s,w) pairs which would permit a conditional test of level α and conditional power at least $1-\beta$. An alternative sequential test procedure is then to continue sampling until an (s,w) pair on the list is first obtained, and then perform the corresponding non-sequential conditional test. This procedure would have unconditional strength (α,β), just as the sequential test of the previous section. This test will not be pursued further here, but its ASN properties would be inferior to the previous sequential test, the comparison being conceptually analogous to the comparison of a non-sequential test with a sequential probability ratio test. Here, both tests are conditional, and the "non-sequential"one is in fact sequential since sampling is continued until a satisfactory "condition" is achieved. Tests of this type have been proposed by Lehmann ([8], pg. 140).

5. TESTING AGAINST A TWO-SIDED ALTERNATIVE

Sequential tests against two-sided alternatives of the form H — the alternative being that either A is superior to B or that B is superior to A—

may be developed in one of two ways, either by the intersection method of Armitage [1] whereby two one-sided tests are run simultaneously, or by the invariance method [6]. (A general exposition and comparison of these two methods in other contexts is planned elsewhere [5].) We describe the first somewhat briefly (the mothod was also used by Sobel and Wald [10] and extensively by Armitage [2] and Hogan [7]), and the second in a little more detail since it requires a combination of both the conditional and invariance methods of constructing sequential tests. A non-sequential analog of the latter test will also be given.

In the intersection method, one runs a test as given in Section 2 (denote it T_A) of H_0 vs. H_A with specified α, β, σ' and τ'. Simultaneously, using the same data, one runs a test (denote it T_B) of H_0 vs. H_B with the same values (for simplicity) of α, β, σ' and τ' . This test is identical with T_A except that the roles of A and B are reversed. It is based on the cumulative sum

$$Z_{Bn} = s_{Bn} c_A - s_{An} c_B + w_{Bn} d_A - w_{An} d_B .$$

The two cumulative sums may be plotted on the same chart. The combined test, say T, is terminated only after both tests T_A and T_B have terminated. If both tests accept H_0, then T accepts H_0; if either test rejects H_0, then T rejects H_0. (The tests considered in Armitage's book [2] are described slightly differently since in all of his examples it is possible to plot a single cumulative sum on a chart with two sets of boundary lines, rather than plotting two sums on a chart with a single set of boundary lines)

We shall now derive a sufficient condition on α, β, σ' and τ' to assure the impossiblity of both tests T_A and T_B leading to rejection of H_0. When

this condition is satisfied, the type I error probabilities of the two tests are additive so that the probability of rejecting H_0 when it is true using test T is approximately 2α. Also, if H_A is true, test T_B may be shown to lead to acceptance of H_0 with virtual certainty (when α and β are fairly small) so that test T will lead to acceptance of H_0 when H_A is true with approximate probability β' and likewise if H_B is true. Hence, we may conclude that test T has approximate strength $(2\alpha, \beta)$.

Designate as Condition 1: $\sigma' \geq \tau'$ and $(c_A + d_B - d_A)/c_A \geq -b/a$ (or $\sigma' \leq \tau'$ and $(d_A + c_B - c_A)/d_A \geq -b/a$). We consider the first case, assuming $\sigma' \geq \tau'$. The left side of the associated inequality depends on σ' and τ', and is easily seen to be greater than unity, while the right side (namely $-b/a$) depends on α and β. Condition 1 is certainly satisfied when $\alpha \leq \beta$ $(< 1/2)$, since then $a \geq -b$, and commonly α is chosen to be no greater than β.

We now proceed to show that, under Condition 1 (with $\sigma' \geq \tau'$), if the cumulative sum for test T_A (denote it Z_{An}) is $\geq a$ (rejection of H_0) then the cumulative sum Z_{Bn} for test T_B is $\leq b$ (acceptance of H_0), and similarly with A and B interchanged; hence, T_A and T_B cannot both reject H_0. (It is also assumed that both α and β are smaller than 1/2 so that $b < 0 < a$.)

Now $Z_{Bn} \leq s_n c_A + w_n d_A \leq (s_n + w_n) c_A$, where $s_n = s_{An} + s_{Bn}$ and

$w_n = w_{An} + w_{Bn}$, since $d_A < c_A$. Hence, if $a \leq Z_{Bn}$ then $a/c_A \leq s_n + w_n$. Also,

$Z_{An} + Z_{Bn} = s_n (c_A - c_B) + w_n (d_A - d_B) \leq (s_n + w_n) (d_A - d_B)$ since

$c_A - c_B < d_A - d_B < 0$. Hence, if $a \leq Z_{Bn}$, then $Z_{An} + a \leq -(s_n + w_n) (d_B - d_A)$

$\leq -a\,(d_B - d_A)/c_A$, that is, $Z_{An} \leq -a\,(c_A + d_B - d_A)/c_A \leq b$, by Condition 1. That $Z_{An} \geq a$ implies $Z_{Bn} \leq b$ may be shown analogously.

We now turn to the invariance method. The problem treated here—the model and the hypotheses H_0 and H—are completely symmetric in the roles of the two treatments A and B. Thus, if we transform the sequence of observations $x_1, x_2, \ldots$, by replacing all A-preferences with B-preferences and vice versa, the model will be the same except that the A and B subscripts on the parameters $p_{An}, p_{Bn}, q_{An}, q_{Bn}$ will be interchanged; H_0 will remain invariant but H_A and H_B will be interchanged so that H also remains invariant-that is, if H_0 were true about the original data, it is true about the transformed data, and if H_A were true about the original data, then H_B is true about the transformed data and conversely. Such a transformation, together with the identity transformation (since the transformation is its own inverse) form a group leaving the problem invariant in the sense described in [6].

We shall in fact apply the invariance method to the conditional models given the t-sequence. Then, the data consist of three sequences of trials, in the order prescribed by the t-sequence. When $t_n = 0$, $x_n = \underline{0}$ with certainty; when $t_n = 1$, $x_n = \underline{a}$ or $\underline{b}$ with (conditional) probabilities τ_n and $1-\tau_n$, and when $t_n = 2$, $x_n = \underline{A}$ or $\underline{B}$ with (conditional) probabilities σ_n and $1-\sigma_n$. The transformation interchanges $(\underline{a},\underline{A})$ and $(\underline{b},\underline{B})$ and leaves $\underline{0}$ invariant, and thereby yields the same conditional model with (σ_n, τ_n) and $(1-\sigma_n, 1-\tau_n)$ interchanged; p_n and q_n remain invariant. Under H_0, (σ_n, τ_n) and $(1-\sigma_n, 1-\tau_n)$ equals $(1/2,1/2)$ while under H one or the other equals (σ', τ').

Under H_0 or H, a sufficient statistic after n observations (still conditional on the t-sequence) is the accumulated numbers, s_{An} and w_{An}, of A-preferences in the strong and weak preference trials. Of course, s_n and w_n are

given constants (functions of the t-sequence). A maximal invariant function of the sufficient statistic is $v_n = (\{s_{An}, s_{Bn}\}, \{w_{An}, w_{Bn}\})$ where the curly brackets indicate lack of order—that is, the two numbers s_{An} and s_{Bn} are specified by v_n but the "labels are lost", it is not recorded which is the number of A preferences and which is the number of B preferences. The conditions for the Stein Theorem [6] are satisfied, and we can conclude that v_n is an <u>invariantly sufficient statistic</u>—sufficient for the (conditional) distribution of any invariant function of the first n observations in the sense that the conditional distribution of an invariant function of $x_1,\ldots x_n$, given $t_1,\ldots,t_n$ and v_n, is parameter-free. In particular, the conditional distribution of $(v_1,\ldots v_{n-1})$ given v_n and the t's is parameter-free; consequently the conditional likelihood ratio of $(v_1,\ldots,v_n)$ coincides with the conditional likelihood ratio of v_n.

A CSPRT can then be based on the sequence of v's. We only need evaluate the log, say Z_n, of the conditional likelihood ratio of v_n given the t's and determine at each stage whether it lies between b' and a', or whether it is $\geq$ a' or $\leq$ b' where $a' = \log[(1-\beta)/2\alpha]$ and $b' = \log[\beta/(1-2\alpha)]$. The test will then have prescribed strength $(2\alpha,\beta)$. (We have represented the strength this way to conform with that of the intersection test given earlier.)

Now the conditional likelihood of v_n under H is

$$\binom{s}{x}\binom{w}{y}[\sigma'^{x}(1-\sigma')^{s-x}\tau'^{y}(1-\tau')^{w-y} + \sigma'^{s-x}(1-\sigma')^{x}\tau'^{w-y}(1-\tau')^{y}]$$

(substituting (s,w) for (s_n,w_n) and (x,y) for (s_{An},w_{An})); the conditional likelihood under H_0 is the same with $\sigma' = \tau' = 1/2$. Setting $X = |\frac{s}{2} - x| = |s_{An}-s_{Bn}|$ and $Y = |\frac{w}{2} - y| = |w_{An}-w_{Bn}|$, Z_n is then found to be:

$$Z_n = \frac{s}{2} \log[4\sigma'(1-\sigma')] + \frac{w}{2} \log[4\tau'(1-\tau')] + \log\left[\frac{\lambda'^{X}\mu'^{Y} + \lambda'^{-X}\mu'^{-Y}}{2}\right]$$

$$= (s_{An} + s_{Bn})\frac{c_A - c_B}{2} + (w_{An} + w_{Bn})\frac{d_A - d_B}{2}$$

$$\log\cosh\ [|s_{An}-s_{Bn}|(c_A+c_B) + |w_{An}-w_{Bn}|(d_A + d_B)].$$

The test is performed by plotting Z_n and terminating in favor of H_0 whenever the horizontal line with ordinate b' is crossed and terminating in favor of H whenever the horizontal line with ordinate a' is crossed. Since Z_n is not a cumulative sum, the mechanics of carrying out the test are less convenient; however, Z_n is easily bounded between two easily computed functions Z_n^- and Z_n^+, defined below, and Z_n need not be computed until one of the boundaries is approached. Since $\frac{1}{2}\ e^{|x|} < \frac{e^x + e^{-x}}{2} \leq e^{|x|}$, $x - \log 2 < \log\cosh x \leq |x|$. Hence, Z_n^+ is defined as Z_n with the log cosh term replaced by its argument (which is non-negative) and $Z_n^- = Z_n^+ - \log 2$.

The certainty is readily established by approximating the distribution of Z_n. Other properties of the test are largely unknown, but some crude approximations are possible by replacing $Z_n^{\pm}$ by cumulative sums (removing the absolute value signs).

Finally, we describe a two-sided non-sequential test—the most powerful invariant conditional test of H_0 vs. H. The test has critical region $Z_n \geq c(t_1,\ldots,t_n)$ where c is so chosen that the conditional probability of the critical region is approximately α when H_0 is true (Z_n is the conditional likelihood ratio of v_n, given above). Alternatively, the critical region may be expressed as:

$$|s_{An} - s_{Bn}| \log\lambda' + |w_{An} - w_{Bn}| \log\mu' \geq C(s,w),$$

the complement of a diamond-shaped region in the space of the differences

$s_{An} - s_{Bn}$ and $w_{An} - w_{Bn}$. Since, under H_0, these differences have symmetric distributions, the probability of the critical region is bounded by four times the probability of the region defined by the same inequality but with the absolute value signs removed. The left side of the latter inequality has, conditionally, an approximate normal distribution with zero mean and conditional variance v^2 as noted in Section 4. Alternatively, the critical region lies between inscribed and circumscribed ellipses, with conditional probabilities that are found to be (asymptotically) $F(c^2/v^2)$ and $F(2c^2/v^2)$ where F is the distribution function of χ^2 with two degrees of freedom.

There is surely little difference between the intersection and invariance tests, though no formal comparison is presented here. The intersection test does not require complete symmetry in H_A and H_B, making up H, however, whereas the invariance test does.

6. COMMENT ON ORDER EFFECTS

No attention has so far been given to the possibility of order effects—that is, effects due to treatment A being applied before treatment B or vice versa. If some prior estimate of their likely magnitude is available, then the hypotheses may be modified slightly as noted below to account for such effects. But the simple models treated in this paper are not designed to explore elaborate hypotheses or to estimate various extraneous effects. More elaborate linear effects models have been considered by Scheffé [9].

If it is known that subjects have a slight preference for the first of two similar treatments, say, then one might test the hypothesis that $(\sigma_n, \tau_n) = (\frac{1}{2} \pm \delta, \frac{1}{2} \pm \epsilon)$ where the + sign is used whenever treatment A is applied first and the - sign whenever treatment B is applied first;

δ and ϵ are specified positive numbers. Likewise, the alternative hypothesis H_A would be modified in a suitable way. For a two-sided intersection test, positive constants would be added to the hypothesized values under H_O, H_A and H_B whenever treatment A was applied first, and negative constants added otherwise. It does not seem possible to make suitable modifications and yet preserve the symmetry required for the invariance test, so it cannot be recommended when order effects are known to be of consequence.

REFERENCES

1. Armitage, P. (1947). Some sequential tests of Student's hypothesis. J. Roy. Statist. Soc. Supp. 9 250-263.

2. Armitage, P. (1960). Sequential Medical Trials. Charles C. Thomas, Springfield, Ill.

3. Fraser, D. A. S. (1956). Sufficient statistics with nuisance parameters. Ann. Math. Statist. 27 838-842.

4. Hall, W. J. (1965). Methods of sequentially testing composite hypotheses with special reference to the two-sample problem. Inst. of Statist. Mimeo Series No. 441, Univ. of North Carolina, Chapel Hill. Abstract: Technometrics 8 206 (1966).

5. Hall, W. J., and Nemenyi, P. (1966). A comparison of the intersection and invariance methods of sequentially testing against two-sided alternatives. In preparation.

6. Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. Ann. Math Statist. 36 575-614.

7. Hogan, Michael D. (1963). Several test procedures (sequential and non-sequential) for normal means. Inst. of Statist. Mimeo Series No. 379, Univ. of North Carolina, Chapel Hill.

8. Lehmann, E. L. (1959). Testing Statistical Hypotheses. Wiley, New York.

9. Scheffé, Henry (1952). An analysis of variance for paired comparisons. Jour. Amer. Statist. Assoc. 47 381-400.

10. Sobel, Milton, and Wald, Abraham (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of normal distribution. Ann. Math. Statist. 20 502-522.

11. Wald, Abraham (1947). Sequential Analysis. Wiley, New York.