A SEQUENTIAL SIGNED-RANK TEST

BY

RUPERT G. MILLER, JR.

TECHNICAL REPORT NO. 14
JULY 18, 1969

SUPPORTED BY PUBLIC HEALTH SERVICE
GRANT USPHS-GM-14554-03

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A SEQUENTIAL SIGNED-RANK TEST

By

Rupert G. Miller, Jr.

Technical Report No. 14

July 18, 1969

Prepared under U. S. Public Health Service Grant USPHS GM-14554-03

Reproduction in whole or in part is permitted for any purpose of the United States Government

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A Sequential Signed-Rank Test

Rupert G. Miller, Jr.
Stanford University

Once upon a time there was a good little statistician. He lived with his colleagues in the deep dark medical school. One day a friendly client came to him for a design of a sequential clinical trial. The client wanted to use matched pairs of children to compare two therapies for cystic fibrosis. Immediately the little statistician thought of his good friend the sequential sign test. But the data would be quantitative, and his friend the sequential sign test did not know how to take advantage of this. Then the little statistician thought of his other friend, the sequential $t$ test. But the bad data would be very wicked and would have long tails. The little statistician grew afraid. He was puzzled about what to do so he thought and thought. Then, all of a sudden, he knew what he wanted: a sequential signed-rank test. Alas, there was no sequential signed-rank test, so the good little statistician wished and wished and wished ...

1. Introduction

Let $X_1, X_2, \ldots$ be a sequence of independent random variables, identically distributed according to the cdf $F$. Assume that $F$ has a density $f$ which is symmetric about $\theta$. The problem is to test the
null hypothesis $H_0: \theta = 0$ against the alternative $H_1: \theta \neq 0$ or possibly against some more general alternative.

In the fixed sample size case the three tests most commonly used in practice are (a) the sign test, (b) the Wilcoxon signed-rank test, and (c) the $t$ test. In sequential testing there exist analogs of (a) and (c), but not (b). Wald [5] derived a one-sided sequential binomial test; Armitage [1] and I. Bross, using Wald-type boundaries, fully developed the two-sided sequential binomial test for clinical trials. K. Arnold, G. Barnard, H. Goldberg, S. Rushton, and A. Wald have all worked on sequential $t$ tests. Rushton [4] gives a lucid discussion of the two-sided sequential $t$ test.

The sequential sign test has been tremendously useful in sequential experimentation, especially in medical clinical trials. Its only deficiency is that it fails to use all the information in quantitative (non-dichotomous) data. For badly behaved data it may be best to ignore this extra information, but for good data it is wasteful.

Some reservations come to mind with regard to the sequential $t$ test. In its exact form it is laborious to perform and requires special tables [2]. Its approximate form [4] is much more palatable, but it is still a little tedious to compute. The design is not truncated so sampling could continue for longer than desired. It is not easy to explain to nonstatisticians, in particular medical investigations. Finally, I have some doubts about its validity. For fixed samples the $t$ test is known to be reasonably robust against departures from normality. However, the sequential boundaries may be more
sensitive to nonnormality. If this is the case, it causes a dilemma because the sequential $t$ test may terminate sampling before enough observations are taken to investigate the question of normality.

The purpose of this paper is to fill the void beyond the sequential sign test (or the gap between it and the sequential $t$ test). Considerable work, both theoretical and applied, has been done on two-sample sequential rank tests, but very little on the one-sample problem. Parent [3] gave a one-sample test which uses signed-ranks, but the statistic at each stage $n$ differs from the usual Wilcoxon statistic. Weed and Bradley have done work on grouped one-sample tests which is unpublished at this date. However, grouping wastes the information between groups, and the test proposed in this paper avoids grouping.

The sequential signed-rank test proposed in this paper is presented in Section 2. Its distribution theory is described in Section 3, and the Monte Carlo rejection boundaries are tabled. Section 4 contains a discussion of the power and expected sample size for this test, and Section 5 gives an inner boundary for early stopping when $H_0$ is to be accepted.

2. A Sequential Signed-Rank Test

With complete freedom of choice confronting me, I had the opportunity to ponder on what properties I would like a sequential signed-rank test to have. Six thoughts gradually emerged and molded the eventual structure of the test. These are not put forth as fundamental truths or logical principles underlying statistical inference, and they
should not be viewed as such. They are simply the ideas which guided
the formation of this test.

Thoughts on the selection of a test:

1. At each stage $n$ the decision to continue sampling or
stop with a choice of $H_0$ or $H_1$ should depend on the
customary signed-rank statistic for the sample $X_1, \ldots, X_n$.

Parent [3] derived a test based on the statistic $\frac{\sum_{i=1}^{n} S_i}{n}$ where
$S_i$ is the signed sequential rank of $X_i$. ($|S_i|$ is the rank of $|X_i|$ relative to $|X_1|, \ldots, |X_n|$; the sign of $S_i$ is the sign of $X_i$.)
This statistic is not identical to the usual Wilcoxon signed-rank
statistic for $X_1, \ldots, X_n$, nor is it a monotonic function of the
usual statistic. From the sequence of Parent statistics for
$n = 1, 2, 3, \ldots$ it is possible to determine the corresponding
sequence of Wilcoxon statistics, and vice versa. However, since the
statistics are not monotonic functions of one another, a rejection
(acceptance) interval for one would not necessarily produce a rejection
(acceptance) interval for the other.

Since the Wilcoxon signed-rank test is a technique with well-known
desirable properties and is one of my favorite practical tools, I
prefer to base a sequential test on it. This is not to say that a
test based on the Parent statistic would be bad. The test devised by
Parent was specially designed for a process control problem.
2. The test should control \( \alpha \), the probability of a type I error.

This is the classical approach. In clinical trials the probability of a type I error is an important variable.

3. The test should be easy to apply and explain.

In medical applications statistical techniques which are difficult to use or to explain to nonstatisticians suffer from disuse. Unless it is absolutely mandatory to use a complicated technique, the simpler ones will prevail.

4. The structure of the test should not depend on special alternative hypotheses such as Lehmann alternatives.

There doesn't seem to me to be much point in structuring the test on a mathematically convenient alternative hypothesis so as to achieve prescribed power against an unreal alternative. However, abandonment of special alternatives means that very little can be obtained in the way of mathematical results. It also means abandoning the likelihood ratio since this is usually complicated for ranks except under special distributions.

5. The test should be of closed form.

In clinical trials it is unrealistic to envision unlimited sampling. Limitations of time, money, energy, etc. invariably impose an
upper bound on the number of patients that will be admitted to the 
study. If the statistician does not truncate the sampling in his 
design, other considerations will truncate it for him. Most medical 
investigators can specify an upper bound on the number of patients 
they can handle in the study. They seem far better able to give such 
a bound than to indicate what alternative they have in mind and what 
power they want at this alternative.

6. The test should have reasonable power against reasonable 
alternatives with reasonably early stopping.

Without reasonable power one might just as well not run the test. 
This guarantees that whatever alternative the medical investigator 
should have had in mind the power is reasonably good at it.

Decision-theoretic mathematical statisticians might append a 
seventh thought -- in some fashion maximize the power or minimize the 
expected sample size. This thought would be nice, but usually it is 
mathematically incompatible with several of the preceding six thoughts. 
Provided one is not employing a bad test, the question of whether one 
has the optimal test is relatively unimportant in applications. Other 
considerations usually rule out, or outweigh, optimality.

These thoughts led me to the following test. The test is by no 
means uniquely determined by the thoughts; other tests may well satisfy 
them.
A sequential signed-rank test:

Let $SR^+_n$ be the sum of the positive ranks for the sample $X_1, \ldots, X_n$. (Let $R_i$ be the rank of $|X_i|$ in $|X_1|, \ldots, |X_n|$; let $X^+_i = 1$ if $X_i > 0$, = 0 if $X_i < 0$. Then $SR^+_n = \sum_{i=1}^{n} X^+_i R_i$.)

1. Continue sampling as long as

   i) $SR^+_n \leq \frac{n(n+1)}{4} - 2\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}}$

   and

   ii) $n < N$.

2. Stop sampling as soon as i) or ii) is violated.

   a) If i) is violated, decide in favor of $H_1$.

   b) If ii) is violated and not i), decide in favor of $H_0$.

$N$ and $\alpha$ are selected by the investigator; these determine $\frac{z^\alpha}{N}$ (see Table 1).

$SR^+_n$ is easily computed sequentially from the following representation which is due to Tukey:

$$SR^+_n = \sum_{j=1}^{n} \sum_{i=1}^{j} (X^+_i + X^+_j)$$

$$= SR^+_n - \sum_{i=1}^{n} (X^+_i + X^n)$$

(1)

\[\text{† } a \in b + c \text{ means } b - c < a < b + c.\]
where

\[(x_i + x_j)^+ = \begin{cases} 1 & \text{if } x_i + x_j > 0, \\ 0 & \text{if } x_i + x_j < 0. \end{cases}\]

If the statistician simply keeps a list of the observations on a piece of paper, then by eye he can compute \((x_i + x_n)^+\) and sum to obtain \(\sum_{i=1}^{n}(x_i + x_n)^+\), which when added to \(SR_{n-1}^+\) yields \(SR_n^+\). Thus, successive calculation of the signed-rank sums \(SR_n^+\) is extraordinarily simple.

The test can be carried out by plotting the boundaries \(n(n+1)/4 + \frac{\alpha \sqrt{n(n+1)(2n+1)/24}}{N}\) as functions of \(n\) and sequentially recording the values of \(SR_n^+\) until they cross a boundary or reach \(n = N\). Graphically, it may be easier to plot the boundaries \(n(n+1)/4 + \frac{\alpha \sqrt{n(n+1)(2n+1)/24}}{N}\) which are symmetric about the \(n\)-axis, and sequentially record the values of \(SR_n^+ - n(n+1)/4\).

In addition to the outer rejection boundary there is an inner acceptance boundary which permits stopping in favor of \(H_0\) for \(n < N\). The inner stopping region consists of those points \((n, SR_n^+)\) for which it is not possible to reach the outer boundary no matter what the values of the next \(N - n\) observations. The inner boundary will be discussed in Section 5.

In closing this section it might be worthwhile to indicate how well, or poorly, the six founding thoughts are satisfied. Clearly, thoughts 1 (\(SR_n^+\)), 2 (\(\alpha\)), and 5 (\(n \leq N\)) are satisfied.

Thought 3 (easy to apply and explain) I also think is well taken care of. The \(SR_n^+\) are easily computed sequentially via the Tukey
representation. The fixed sample size Wilcoxon test is known to a sizeable number of medical investigators, and if it is not known, it is easily explained. Medical investigators also understand that many standard deviations away is bad for the null hypothesis and few standard deviations away is good. Since \( n(n+1)/4 \) is the mean of \( SR_n^+ \) under \( H_0 \) and \( \sqrt{n(n+1)(2n+1)/24} \) the standard deviation, the test simply rejects \( H_0 \) if \( SR_n^+ \) gets too many standard deviations away from its mean. The constant which equals "too many" is simply adjusted to allow for the sequential nature of the test. The idea of sequentially applying a fixed sample size test is not new with this paper.

Thought 4 (special alternatives) is clearly satisfied since no alternative hypothesis is specified. For this test \( N \), the upper limit on sampling, is specified in addition to \( \alpha \); these determine the test. In classical sequential analysis a particular alternative and the power at this alternative (in addition to \( \alpha \)) determine the test.

Thought 6 (reasonable power for reasonable alternatives) will be a matter of opinion. Results along this line are presented in Section 4. I believe the test performs rather well, but the reader is invited to decide for himself.

3. Sampling under \( H_0 \)

Let

\[
Y_n = \frac{SR_n^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}},
\]

9
\[ Z_N = \max_{1 \leq n \leq N} \left| Y_n \right| . \]

For the rejection boundary on $\text{SR}^+_n$ defined by $n(n+1)/4 + z\sqrt{n(n+1)(2n+1)/24}$, the test will decide in favor of $H_1$ if and only if $Z_N \geq z$. Thus, $z$ should be the upper $\alpha$-percentile point of the distribution of $Z_N$, i.e., $z = z^\alpha_N$ where $P(Z_N \geq z^\alpha_N) = \alpha$.

The exact joint distribution of $Y_1, \ldots, Y_N$ is a complicated discrete multivariate distribution, and consequently, the exact distribution of $Z_N$ appears intractable. This leads one either to simulation or approximation. I'll discuss simulation first.

Three separate Monte Carlo simulations were performed using the Rand Corporation random numbers. In the first study 1000 $Z_N$'s were generated for $N = 15, 20, 25, 30, 50$. After these results were examined, a second study was run with 2000 $Z_N$'s being generated at $N = 10, 20, 30, 40, 50, 60$. The sample .10, .05, and .01 upper percentile points for $N = 60$ were a bit smaller than what one would have predicted from the corresponding percentile points for the other $N$ so 4000 more $Z_N$'s were generated at $N = 60$. By starting the sequences at different places in the tape of Rand random numbers the random sequences used to generate the $Z_N$'s were different for each $N$ and each study.

At each $N$ the $Z_N$'s from each study were pooled to form a single sample for that $N$, and the sample upper $\alpha$ percentile points were obtained from the ordered sample. These Monte Carlo estimates of $z^\alpha_N$
are presented in Table 1 and Figure 1. Because several different Monte Carlo studies were combined, the accuracy of \( z^\alpha_N \) varies with \( N \). The size of the sample \( K \) for each \( N \) is listed on the last line of Table 1. Judging from Figure 1, I would say the monotonic agreement between the points is quite good. The rate of increase in \( z^\alpha_N \) as a function of \( N \) decreases with increasing \( N \) so values of \( N \) somewhat larger than those in the table should not have appreciably larger \( z^\alpha_N \).

<table>
<thead>
<tr>
<th>( \alpha ) ( N )</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>2.02</td>
<td>2.16</td>
<td>2.20</td>
<td>2.22</td>
<td>2.28</td>
<td>2.33</td>
<td>2.37</td>
<td>2.39</td>
</tr>
<tr>
<td>.05</td>
<td>2.20</td>
<td>2.39</td>
<td>2.40</td>
<td>2.46</td>
<td>2.55</td>
<td>2.55</td>
<td>2.62</td>
<td>2.62</td>
</tr>
<tr>
<td>.01</td>
<td>2.55</td>
<td>2.75</td>
<td>2.83</td>
<td>2.91</td>
<td>2.93</td>
<td>3.03</td>
<td>3.03</td>
<td>3.07</td>
</tr>
<tr>
<td>( K/10^3 )</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1

Values of \( z^\alpha_N \)

(estimated from a sample of \( K \) randomly generated \( Z_n \)'s).

It may be possible to approximate \( z^\alpha_N \) from the moments of \( SR^+ \) and its approximate normality for large \( n \). From well-known fixed sample size moments

\[
E(Y_n) = 0, \quad \text{Var}(Y_n) = 1.
\]

Under \( H_0 \)

\[
E((X_i + X_j)^+(X_k + X_j)^+) = \int_{-\infty}^{+\infty} (1 - \Phi(-x))^2 d\Phi(x) = \frac{1}{3}
\]

for \( i, j, k \) unequal, and
Figure 1
Values of $z_N^a$
\begin{equation}
\mathbb{E}((X_j + X_j^\dagger)(X_i + X_i^\dagger)) = \int_0^{+\infty} (1 - F(-x))dF(x) = \frac{3}{8}
\end{equation}

for \( i \neq j \). These two simple results together with the independence of \( X_i \) and \( X_j \) for \( i \neq j \) and the decomposition

\begin{equation}
\sum_{j=1}^{n} \sum_{i=1}^{\frac{m}{2}} (X_i + X_j^\dagger) + \sum_{j=m+1}^{n} \sum_{i=1}^{\frac{m}{2}} (X_i + X_j^\dagger)
\end{equation}

for \( m < n \), enable one to straightforwardly verify that under \( H_0 \)

\begin{equation}
\text{Cov}(SR^+_m, SR^+_n) = \frac{m(n+1)(2m+1)}{24}
\end{equation}

for \( m < n \). Thus, for \( m < n \)

\begin{equation}
\text{Cov}(Y_m, Y_n) = \sqrt{\frac{m(n+1)(2m+1)}{n(m+1)(2n+1)}},
\end{equation}

\begin{equation}
\approx \sqrt{\frac{m}{n}}.
\end{equation}

The statistic \( SR^+_n \) is asymptotically normally distributed and is approximately normal even for \( n \) as small as 10. For \( \alpha \leq .10 \) it is impossible for the boundary to reject \( H_0 \) until \( n \) is at least 5. This means that as a rough approximation the upper tail of \( Z_N \) is behaving as the maximum between 5 and \( N \) of a discrete time Gaussian process with covariance structure (10).

To push the approximation one step further we can imagine the Gaussian process being imbedded in a Wiener process. Define

\begin{equation}
Y'_n = \frac{SR^+_n - \frac{n(n+1)}{4}}{\sqrt{\frac{(n+1)(2n+1)}{24}}} = \sqrt{n} Y_n.
\end{equation}
The $Y'_n$ are approximately jointly normally distributed with

$$E(Y'_n) = 0, \ Var(Y'_n) = n,$$

(12) $$\text{Cov}(Y'_m, Y'_n) = m, \text{ for } m < n,$$

and for $\alpha \leq .10$

(13) $$P(Z_N \geq z) = P(\{Y'_n \geq z\sqrt{n} \text{ for some } n, 5 \leq n \leq N\}.$$

Let $W(t)$ be the Brownian motion process with no drift and variance $t; W(0) = 0$. Then, with the identifications $W(n/N) = Y'_n/\sqrt{N}$ and $t_0 = 5/N$, the Wiener approximation becomes

(14) $$P(Z_N \geq z) \approx P(\{|W(t)| \geq z\sqrt{t} \text{ for some } t, t_0 \leq t \leq 1\}.$$

How good this approximation is depends on how good the normality approximation is and how safe it is to ignore the fact that the Wiener process could remain within the boundary at times $t = n/N, n = 5, \ldots, N$, but exceed it at other times.

For small $\alpha$ (viz., $\alpha \leq .05$) it should also be reasonable to make the additional approximation

(15) $$P(Z_N \geq z) \approx 2P(W(t) \geq z\sqrt{t} \text{ for some } t, t_0 \leq t \leq 1).$$

To my knowledge an analytic expression for the probability on the right in (15) has not been published, but it seems far more susceptible to analytic treatment and asymptotic expansion than does the original discrete problem.
Since for $\alpha \leq .10$ the sequential signed-rank test cannot stop for $n < 5$, the test amounts to considering the fixed sample size signed-rank statistic for $N_1 \leq n \leq N_2$ (where $N_1 = 5$, $N_2 = N$). Vernon Johns has recently developed a method for approximating the probability of a type I error when a fixed sample size procedure is applied sequentially between two fixed integers $N_1$ and $N_2$. This method involves investigating probabilities similar to (15) and will be described in a forthcoming paper.

4. Power and Expected Sample Size

Using the percentile points in Table 1, Monte Carlo studies were run with $\alpha = .05$, .01 and $N = 20$, 50 to investigate the power and stopping time distribution for the sequential signed-rank test. The Rand random numbers were used to generate random variables $X_1$ with a double exponential distribution centered at $\Delta$, i.e.,

$$p_{X_1}(x) = \frac{1}{2} e^{-|x-\Delta|}, \quad -\infty < x < +\infty.$$  

(16)

The distribution (16) has mean $\Delta$, variance 2 (standard deviation 1.42).

This distribution was selected because it is easily generated in a computer from the uniform distribution by taking logarithms, it leads to integrals in a power approximation which can be easily evaluated, and it has thicker tails than the normal distribution.

Runs were performed at different values of the shift parameter: for $N = 20$, $\Delta = 0$, .5, 1, 1.5; for $N = 50$, $\Delta = 0$, .25, .5, .75, 1. The largest value $\Delta = 1.5$ roughly corresponds to a one standard
deviation shift away from the null hypothesis $\Delta = 0$. For each combination of $\Delta$, $N$, and $\alpha$, 500 sequences of random variables were substituted into the test. The results are displayed in Tables 2a, b, c, d.

| $N = 20$ | $\alpha = .05$ |
|---|---|---|---|---|
| $n$ | $\Delta$ | 0 | .5 | 1.0 | 1.5 |
| 1-2 | 0 | 0 | 0 | 0 |
| 3-4 | 0 | 0 | 0 | 0 |
| 5-6 | 0 | 0 | 0 | 0 |
| 7-8 | 3 | 28 | 96 | 209 |
| 9-10 | 6 | 31 | 79 | 119 |
| 11-12 | 5 | 26 | 60 | 70 |
| 13-14 | 3 | 24 | 59 | 42 |
| 15-16 | 3 | 20 | 47 | 25 |
| 17-18 | 3 | 22 | 35 | 17 |
| 19-20 | 5 | 19 | 34 | 8 |
| Never exceed boundary | 472 | 330 | 90 | 10 |

- Power
  - Approx. Power
    - .056
    - .34
    - .82
    - .98

- Expected $n$ | rejection
  - 13.3
  - 12.9
  - 12.2
  - 10.1

Table 2a

Monte Carlo stopping time frequencies, power, and expected sample sizes for $N = 20$, $\alpha = .05$. 

16
<table>
<thead>
<tr>
<th></th>
<th>( N = 20 )</th>
<th>( \alpha = .01 )</th>
<th>( \Delta )</th>
<th>( 0 )</th>
<th>( .5 )</th>
<th>( 1.0 )</th>
<th>( 1.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3-4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5-6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7-8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11-12</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
<td>98</td>
<td>236</td>
<td></td>
</tr>
<tr>
<td>13-14</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>13</td>
<td>69</td>
<td>95</td>
</tr>
<tr>
<td>15-16</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>14</td>
<td>45</td>
<td>54</td>
</tr>
<tr>
<td>17-18</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>13</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>19-20</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>14</td>
<td>46</td>
<td>27</td>
</tr>
<tr>
<td>Never exceed boundary</td>
<td>496</td>
<td>423</td>
<td>197</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|    |                  |                  |      |     |     |     |     |
|    | Power            | Approx. Power    |      |     |     |     |     |
|    | .008            | .002             |      |     |     |     |     |
|    | .15             | .09              |      |     |     |     |     |
|    | .61             | .50              |      |     |     |     |     |
|    | .92             | .91              |      |     |     |     |     |
| Expected n | rejection | 16.8          | 14.9  | 14.6 | 13.3 |      |     |
| Expected n |        | 20.0          | 19.2  | 16.7 | 13.8 |      |     |

Table 2b
Monte Carlo stopping time frequencies, power, and expected sample sizes for \( N = 20, \alpha = .01 \).

<table>
<thead>
<tr>
<th></th>
<th>( N = 50 )</th>
<th>( \alpha = .05 )</th>
<th>( \Delta )</th>
<th>( 0 )</th>
<th>( .25 )</th>
<th>( .5 )</th>
<th>( .75 )</th>
<th>( 1.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11-15</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>17</td>
<td>46</td>
<td>95</td>
<td>142</td>
</tr>
<tr>
<td>16-20</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>12</td>
<td>53</td>
<td>96</td>
<td>105</td>
</tr>
<tr>
<td>21-25</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>16</td>
<td>36</td>
<td>71</td>
<td>68</td>
</tr>
<tr>
<td>26-30</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>10</td>
<td>46</td>
<td>45</td>
<td>41</td>
</tr>
<tr>
<td>31-35</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>7</td>
<td>39</td>
<td>45</td>
<td>22</td>
</tr>
<tr>
<td>36-40</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>11</td>
<td>39</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>41-45</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>10</td>
<td>35</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>46-50</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>14</td>
<td>30</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Never exceed boundary</td>
<td>478</td>
<td>392</td>
<td>146</td>
<td>31</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|    |                  |                  |      |     |     |     |     |     |
|    | Power            | Approx. Power    |      |     |     |     |     |     |
|    | .044            | .004             |      |     |     |     |     |     |
|    | .22             | .12              |      |     |     |     |     |     |
|    | .70             | .56              |      |     |     |     |     |     |
|    | .94             | .93              |      |     |     |     |     |     |
|    | 1.0             | 1.0              |      |     |     |     |     |     |
| Expected n | rejection | 26.9          | 27.0  | 27.4 | 21.5 | 17.1 |      |     |
| Expected n |        | 49.0          | 45.0  | 34.1 | 23.2 | 17.1 |      |     |

Table 2c
Monte Carlo stopping time frequencies, power, and expected sample sizes for \( N = 50, \alpha = .05 \).
Table 2d

Monte Carlo stopping time frequencies, power, and expected sample sizes for \( N = 50, \alpha = .01 \).

For \( N = 20 \) the stopping times were grouped into intervals of 2 whereas for \( N = 50 \) the intervals contain 5 values of \( n \). For each column of \( \Delta \) the number of sequences out of 500 which exceeded the rejection boundary in each interval of \( n \) is displayed. The empirical power of the test is the number of sequences exceeding the rejection boundary divided by 500. Those sequences which exceeded the rejection boundary when \( \Delta > 0 \) always did so in the correct direction.

For \( N = 20, \alpha = .05 \), the power is good (.82) for a shift of \( \Delta = 1 \), which is roughly two-thirds of a standard deviation from \( H_0 \). For \( N = 50, \alpha = .05 \), the power is good (.94) for a one-half standard deviation departure (\( \Delta = .75 \)) from \( H_0 \). At the \( \alpha = .01 \) level the
power is still good at $\Delta = .75$ when $N = 50$, but for $N = 20$ the shift needs to be one standard deviation away to have good power. This to me seems to indicate a reasonably powerful test.

One can attempt to theoretically approximate the power from the following considerations. Both under the null and alternative hypotheses $SR_n^+$ is asymptotically normally distributed. For a general distribution $F$

$$E(SR_n^+) = \frac{n(n-1)}{2} E((X_i + X_j)^+ + nE((X_j + X_j)^+),$$

$$\text{Var}(SR_n^+) = n(n - 1)(n - 2) \text{Cov}((X_i + X_j)^+, (X_k + X_j)^+),$$

$$+ 2n(n - 1) \text{Cov}((X_j + X_j)^+, (X_i + X_j)^+),$$

$$+ \frac{n(n-1)}{2} \text{Var}((X_i + X_j)^+),$$

$$+ n \text{Var}((X_j + X_j)^+),$$

(17)

where $i, j, k$ are unequal. The variances and covariances in (17) involve the integrals

$$E((X_i + X_j)^+(X_k + X_j)^+) = \int_{-\infty}^{+\infty} (1 - F(-x))^2 dF(x),$$

$$E((X_i + X_j)^+(X_j + X_j)^+) = \int_{0}^{+\infty} (1 - F(-x)) dF(x),$$

(18)

$$E((X_i + X_j)^+) = \int_{-\infty}^{+\infty} (1 - F(-x)) dF(x),$$

$$E((X_j + X_j)^+) = \int_{0}^{+\infty} dF(x),$$

19
which are easily evaluated for some distributions (such as the double exponential) but not for others.

If one also contends that most sequences which reject $H_0$ for some $n \ (1 \leq n \leq N)$ have a final value $SR_N^+$ which also falls outside the rejection boundary, the power can be approximated by the probability in the tail of a normal distribution, namely,

$$
(19) \quad P(\text{rejection}) \approx 1 - \Phi \left( \frac{\frac{N(N+1)}{4} + \frac{\alpha}{N} \sqrt{\frac{N(N+1)(2N+1)}{24}} - E_A(SR_N^+)}{\sqrt{\text{Var}_A(SR_N^+)}}, \right),
$$

where $\Phi(\cdot)$ is the unit normal cdf and $0 < E_A(SR_N^+), \text{Var}_A(SR_N^+)$ are the mean and variance under the alternative hypothesis computed from (17) - (18). If the supposition that once the sequence falls outside the boundary it will tend to remain outside is inaccurate, then (18) will at least provide an approximate lower bound for the power.

The calculations in (17) - (19) were carried out for the double exponential distribution, and the approximate power is displayed in Tables 2a, b, c, d below the empirical power. Inspection of the tables reveals that the approximate power is considerably less than the empirical power except for the larger $\Delta$. Thus, the supposition "once outside, always outside" is faulty in general, and (19) should be regarded as an approximate lower bound.

Two different empirical expected sample sizes were computed. The first, a conditional expectation, is the empirical expected stopping time given that the sequence of random variables rejects $H_0$. This quantity indicates how many observations will be needed, on the average,
if the null hypothesis is going to be rejected. This to me is a somewhat more interesting number than the other, the unconditional expectation. The latter includes the former because

\[(20) \quad E(n) = E(n | H_0 \text{ rejected})P(H_0 \text{ rejected}) + N \cdot P(H_0 \text{ accepted})\]

where \( n \) is the stopping time.

Both empirical expectations are included in Tables 2a, b, c, d. These expectations were computed from the ungrouped frequencies. Examination of the tables reveals that although rejection tends to be earlier as \( \Delta \) increases the conditional expectations given rejection only vary moderately. For \( N = 20 \) \( E(n | H_0 \text{ rejected}) \) shifts as \( \Delta \) increases from 13 to 10 and 17 to 13 approximately for \( \alpha = .05 \) and .01, respectively. For \( N = 50 \) \( E(n | H_0 \text{ rejected}) \) shifts as \( \Delta \) increases from 27 to 17 and 33 to 23 approximately for \( \alpha = .05 \) and .01, respectively. There are much greater changes in \( E(n) \) due to the shift in the probabilities from acceptance to rejection. If the test is going to reject \( H_0 \), rejection seems to occur reasonably early in the trial which is nice.

5. **Inner Acceptance Boundary**

One would guess that if \( SR^+_n \) is hovering near its mean \( n(n+1)/4 \), there will come a time prior to \( N \) at which it would be impossible for \( SR^+_n \) to increase (or decrease) fast enough to reach the rejection boundary before \( N \) (since \( SR^+_{n+1} \) differs from \( SR^+_n \) by at most \( n + 1 \)). This is in fact true.
Determination of the inner acceptance boundary depends on the following proposition: if $\text{SR}^+_n$ exceeds the rejection boundary in the positive (negative) direction and $X_{n+1}$ is positive (negative) with the largest absolute value of $X_1, \ldots, X_{n+1}$, then $\text{SR}^+_{n+1}$ will exceed the rejection boundary in the positive (negative) direction. The proposition says that if $\text{SR}^+_n$ is outside the boundary, either positively or negatively, and the subsequent $X_i$ are the largest in absolute value and in the same direction, then the subsequent signed-ranks sums will continue to remain outside the boundary. If this proposition is true, then the inner acceptance region simply consists of those points $(n, \text{SR}^+_n)$ for which adding (subtracting) the sum of the remaining integers $\sum_{m=n+1}^{N} m$ will not increase (decrease) $\text{SR}^+_n$ above $N(N+1)/4 + z_N^\alpha N(N+1)(2N+1)/24$ (below $N(N+1)/4 - z_N^\alpha N(N+1)(2N+1)/24$).

If the proposition were false, then determination of the inner boundary would be far more complicated because it would be necessary to consider whether $\text{SR}^+_n$ could reach the boundary at points prior to $N$ as well as at $N$.

The proposition is true if

\begin{equation}
\frac{n(n+1)}{4} + z_N^\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}} + (n + 1) > \frac{(n+1)(n+2)}{4} + z_N^\alpha \sqrt{\frac{(n+1)(n+2)(2n+3)}{24}}
\end{equation}

for $1 \leq n \leq N - 1$. Algebraic manipulation reduces this to the inequality

\begin{equation}
\sqrt{\frac{n+1}{2}} + \frac{3}{2} + \frac{1}{2(n+1)} - \sqrt{\frac{(n+1)}{2}} - \frac{3}{2} + \frac{1}{2(n+1)} \leq \frac{\sqrt{3}}{z_N^\alpha}.
\end{equation}
Since $\sqrt{x+h} - \sqrt{x-h} \leq h(x - h)^{-\frac{1}{2}}$ for $0 < h \leq x$, (22) will hold if

$$\frac{3}{2} \left( (n + 1) - \frac{3}{2} + \frac{1}{2(n+1)} \right)^{-\frac{1}{2}} \leq \frac{\sqrt{3}}{2} \alpha_N.$$  \hspace{1cm} (23)

Inequality (23) is equivalent to

$$n + 1 \geq \frac{3(z_N^\alpha)^2}{4} + \frac{3}{2} - \frac{1}{2(n+1)}. \hspace{1cm} (24)$$

In Table 1 $z_N^\alpha$ varies from 2.02 for $N = 10$, $\alpha = .10$ to 3.07 for $N = 60$, $\alpha = .01$. For $z_N^\alpha = 2.02$ (24) gives $n \geq 4$, and for $z_N^\alpha = 3.07$ (24) gives $n \geq 8$. From Table 3 it is clear the test cannot terminate until $n \geq 5$ for $z_N^\alpha = 2.02$ and $n \geq 13$ for $z_N^\alpha = 3.07$. A check of the other $z_N^\alpha$ in Table 1 reveals that at least for the values of $N$ and $\alpha$ being considered in this paper, the proposition holds.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Max-Ave [SD]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.34</td>
</tr>
<tr>
<td>3</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>1.83</td>
</tr>
<tr>
<td>5</td>
<td>2.02</td>
</tr>
<tr>
<td>6</td>
<td>2.20</td>
</tr>
<tr>
<td>7</td>
<td>2.37</td>
</tr>
<tr>
<td>8</td>
<td>2.52</td>
</tr>
<tr>
<td>9</td>
<td>2.67</td>
</tr>
<tr>
<td>10</td>
<td>2.80</td>
</tr>
<tr>
<td>11</td>
<td>2.93</td>
</tr>
<tr>
<td>12</td>
<td>3.06</td>
</tr>
<tr>
<td>13</td>
<td>3.18</td>
</tr>
<tr>
<td>14</td>
<td>3.30</td>
</tr>
<tr>
<td>15</td>
<td>3.41</td>
</tr>
<tr>
<td>20</td>
<td>3.92</td>
</tr>
<tr>
<td>25</td>
<td>4.37</td>
</tr>
<tr>
<td>50</td>
<td>6.15</td>
</tr>
</tbody>
</table>

Table 3

Maximum possible value of $SR_n^+$ normalized by the mean and standard deviation of $SR_n^+$ under $H_0$. 
The inner acceptance boundary can be determined from the following simple identity:

\begin{equation}
\sum_{m=n+1}^{N} m = \frac{(N-n)(N+n+1)}{2}.
\end{equation}

If at stage \( n \) \( SR_n^+ \) has a value less than

\begin{equation}
\frac{N(N+1)}{4} + z N \sqrt{\frac{N(N+1)(2N+1)}{24}} - \frac{(N-n)(N+n+1)}{2},
\end{equation}

and greater than

\begin{equation}
\frac{N(N+1)}{4} - z N \sqrt{\frac{N(N+1)(2N+1)}{24}} + \frac{(N-n)(N+n+1)}{2},
\end{equation}

it will not be possible for \( SR_n^+ \) to reach the rejection boundary in the remaining \( N-n \) observations so the test can terminate with acceptance of \( H_0 \).

Had the Monte Carlo studies on power and expected sample size been run with the inner acceptance boundary included, the overall expected sample size would have been smaller, but none of the other numbers in Table 2 (viz., frequencies of stopping times, power, expected sample size given rejection) would have been changed.

If the values of \( SR_n^+ - n(n+1)/4 \) (instead of \( SR_n^+ \)) are recorded sequentially and compared with the boundaries \( + z N \sqrt{\frac{N(N+1)(2N+1)}{24}} \), then the inner acceptance boundary becomes

\begin{equation}
+ \left[ z N \sqrt{\frac{N(N+1)(2N+1)}{24}} - \frac{(N-n)(N+n+1)}{4} \right],
\end{equation}

provided the quantity in brackets is positive.
For $N = 20$, $\alpha = .05$, Table 1 gives $z_{\alpha}^N = 2.40$. As an illustration of the test, the rejection and acceptance boundaries on $SR_n^+ = n(n+1)/4$ are presented in Table 4 and Figure 2.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Positive Acceptance Boundary</th>
<th>Positive Rejection Boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>24</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>31</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>39</td>
</tr>
<tr>
<td>15</td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>16</td>
<td>27</td>
<td>47</td>
</tr>
<tr>
<td>17</td>
<td>35</td>
<td>51</td>
</tr>
<tr>
<td>18</td>
<td>44</td>
<td>56</td>
</tr>
<tr>
<td>19</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td>64</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 4

Acceptance and rejection boundaries for $N = 20$, $\alpha = .05$.  

25
\[ \text{SR}_n^+ - \frac{n(n+1)}{4} \]

Figure 2

Acceptance and rejection boundaries for \( N = 20, \alpha = .05 \).
6. **Acknowledgments**

I would like to thank Judy Grindle for programming the Monte Carlo experiments, Margaret Lof for typing the manuscript, and Betty Jo Prine for drawing the figures. I am indebted to Vernon Johns for valuable discussions on his approach to sequential confidence intervals and tests. These discussions led me to run the Monte Carlo studies in this paper.
References


