CONTRIBUTIONS TO THE THEORY OF RANK ORDER STATISTICS—THE ONE-SAMPLE CASE[1]

I. Richard Savage

University of Minnesota

0. Summary. The one-sample problem is considered using techniques developed earlier [2], [3]. Let $Z = (Z_1, \cdots, Z_N)$ be a random vector with $Z_i = 1(0)$ if the ith smallest in absolute value in a sample of N from the density $f(x)$ is positive (negative). Then

$$P(Z = z) = N! \int \underset{0 \leqq y_1 \leqq \cdots \leqq y_N \leqq \infty}{\cdots} \int \prod_{i=1}^{N} [f^{1-z_i}(-y_i) f^{z_i}(y_i)\, dy_i]$$

Conditions are found implying $P(Z = z) > P(Z = z')$ where z is derived from z' by replacing a 0 by a 1, or interchanging a 0 and 1 in z' by moving the 1 to the right. These conditions are met by the normal and other distributions.

The results are useful in finding good tests of such null hypotheses as $X_1, \cdots, X_N$ are independently and identically distributed symmetrically about zero against such alternatives as slippage to the right. The Wilcoxon one sample signed rank test is a typical nonparametric procedure used under these conditions [4].

1. Assumptions and notations. Throughout it is assumed that $X_1, \cdots, X_N$ are independently and identically distributed random variables with a continuous distribution function, $F(x, \theta)$ having a density function $f(x, \theta)$. θ will be a real valued parameter and under the *null hypothesis* $H_0 : \theta = 0$.

If $x_1, \cdots, x_N$ are the observations and $y_1, \cdots, y_N$ are the absolute values of the observations arranged from smallest to largest, then $z = (z_1, \cdots, z_N)$ is defined to be the observed *rank order* where $z_i = 1$ if y_i is the absolute value of a positive number and $z_i = 0$ if y_i is the absolute value of a negative number. Thus, $n = \sum_{i=1}^{N} z_i$ is the number of positive observations and $m = \sum_{i=1}^{N}(1 - z_i)$ is the number of negative observations. Corresponding to the observed $y = (y_1, \cdots, y_N)$ and $z = (z_1, \cdots, z_N)$ are the random variables $Y = (Y_1, \cdots, Y_N)$ and $Z = (Z_1, \cdots, Z_N)$. There are 2^N possible values of Z. For a specified value of n there are $\binom{N}{n}$ values of Z. For n fixed the conditional distribution of Z is that of the two sample problem [2] where the first population has the c.d.f.

$$F^{-}(x, \theta) = \begin{cases} \dfrac{F(0, \theta) - F(-x, \theta)}{F(0, \theta)}, & x \geqq 0, \\ 0, & x < 0, \end{cases}$$

Received June 23, 1958; revised January 22, 1959.

[1] Work supported in part by the Office of Naval Research.

and the second population has c.d.f.

$$F^{+}(x, \theta) = \begin{cases} \dfrac{F(x, \theta) - F(0, \theta)}{1 - F(0, \theta)}, & x \geqq 0, \\ 0, & x < 0. \end{cases}$$

Thus for fixed n the partial order problem is exactly that treated in [2] where $F(x) = F^{-}(x)$ and $G(x) = F^{+}(x)$. The previous results are not immediately applicable, however, since it is not clear how to impose conditions on $F(x, \theta)$ in order to get $F^{-}(x, \theta)$ and $F^{+}(x, \theta)$ to satisfy the conditions of [2]. In Section 2, the case of fixed n is considered. The notation $z'Lz$ denotes the following relationship: $z_k' = z_k$ for all $k = 1, \cdots, N$ except i and j $(i < j)$ and $z_i = z_j' = 0$, $z_j = z_i' = 1$. This notation is also used if there exists $z^1, \cdots, z^I$ such that $z'Lz^1 \cdots z^I Lz$, e.g., $(1010)L(0101)$ since $(1010)L(0110)$ and $(0110)L(0101)$.

In Section 3, the partial order of the probabilities of two rank orders having different values of n is considered. The notation $z'Sz$ denotes $z_k \geqq z_k'$ for $k = 1, \cdots, N$ and $>$ holds for at least one value of k.

The following formula is used repeatedly:

$$(1.1) \qquad P(Z = z) = N! \int \underset{0 \leqq y_1 \leqq \cdots \leqq y_N \leqq \infty}{\cdots} \int \prod_{i=1}^{N} [f^{z_i}(y_i, \theta) f^{1-z_i}(-y_i, \theta)\, dy_i]$$

The null hypothesis of concern is $F(-x, 0) + F(x, 0) = 1$, i.e., symmetry about 0. Under H_0, $P(Z = z) = 2^{-N}$ for each z. An alternative of particular interest is

$$F(x, \theta) = \int_{-\infty}^{x} (2\pi)^{-1/2} e^{-(t-\theta)^2/2}\, dt, \qquad \theta > 0.$$

All of the following results apply to this alternative hypothesis.

2. The case of fixed n.

THEOREM 2.1:

a) $f(x, \theta) = u(x) v(\theta) e^{a(x) b(\theta)}$

b) $v(\theta) \geqq 0$

c) $u(x) = u(-x) > 0$

d) If $x < y$ then $a(x) < a(y)$

e) $b(\theta) > 0$

then $z'Lz$ implies $\Delta = P(Z = z) - P(Z = z') > 0$.

PROOF: Using (1.1) obtain

$$\Delta = N! \int \underset{0 \leqq y_1 \leqq \cdots \leqq y_N \leqq \infty}{\cdots} \int A(y_i, y_j) \prod_{i=1}^{N} [f^{z_i}(y_i, \theta) f^{1-z_i}(-y_i, \theta)\, dy_i]$$

where

$$\begin{aligned} A(y_i, y_j) &= 1 - \frac{f(y_i, \theta)f(-y_j, \theta)}{f(-y_i, \theta)f(y_j, \theta)} \\ &= 1 - \exp\{b(\theta)[a(y_i) - a(-y_i) + a(-y_j) - a(y_j)]\} \end{aligned}$$

The theorem is proved by showing $A(y_i, y_j) \geqq 0$, which follows since the exponent is negative due to the monotonicity of $a(x)$.

Theorem 2.2: *If*

a) $f(x, \theta) = f(x - \theta) = f(\theta - x)$

b) *If* $x > y$ *and* $\theta > \delta$ *then,*

$$\begin{vmatrix} f(x, \theta) & f(x, \delta) \\ f(y, \theta) & f(y, \delta) \end{vmatrix} > 0$$

c) $\theta > 0$

then $z'Lz$ *implies* $\Delta = P(Z = z) - P(Z = z') > 0$.

Proof:

$$\Delta = N! \int \cdot \cdot \cdot_{0 \leqq y_1 \leqq \cdots \leqq y_N \leqq \infty} \int B(y_i, y_j) \left\{ \prod_{\substack{k=1 \\ i \neq k \neq j}}^{N} [f^{z_k}(y_k, \theta) f^{1-z_k}(-y_k, \theta)] \right\} \left[\prod_{k=1}^{N} dy_k \right]$$

where $B(y_i, y_j) = f(y_j, \theta)f(-y_i, \theta) - f(-y_j, \theta)f(y_i, \theta)$ and the proof is completed by showing $B(y_i, y_j) > 0$. In assumption b let $x = y_j$, $y = y_i$, and $\delta = -\theta$ so that

$$0 < \begin{vmatrix} f(y_j, \theta) & f(y_j, -\theta) \\ f(y_i, \theta) & f(y_i, -\theta) \end{vmatrix} = f(y_j - \theta)f(y_i + \theta) - f(y_j + \theta)f(y_i - \theta).$$

Now use $f(x - \theta) = f(\theta - x)$, assumption a, hence

$$\begin{aligned} 0 &< f(y_j - \theta)f(y_i + \theta) - f(y_j + \theta)f(y_i - \theta) \\ &= f(y_j, \theta)f(-y_i, \theta) - f(-y_j, \theta)f(y_i, \theta) \\ &= B(y_i, y_j). \end{aligned}$$

3. The case of variable n.

Theorem 3.1: *Under the assumptions of Theorem* 2.1, *if* $z'Sz$, *then* $\Delta = P(Z = z) - P(Z = z') > 0$.

Proof. It is sufficient to consider only the special case $z_k' = z_k$ for all $k = 1, \cdots, N$ except $k = i$ where $z_i = 1$ and $z_i' = 0$. Then,

$$\Delta = N! \int \cdot \cdot \cdot_{0 \leqq y_1 \cdots \leqq y_N \leqq \infty} \int C(y_i) \left\{ \prod_{k=1}^{N} [f^{z_k}(y_k, \theta) f^{1-z_k}(-y_k, \theta)\, dy_k] \right\}$$

and the proof is completed by showing $C(y_i) = 1 - f(-y_i, \theta) \times [f(y_i, \theta)]^{-1} > 0$. Using the special form of $f(x, \theta)$, $C(y_i) = 1 - \exp\{b(\theta)[a(-y_i) - a(y_i)]\}$ and again the exponent is negative because of the monotonicity of $a(y)$.

THEOREM 3.2: *If*

a) $f(x, \theta) = f_\theta(x - \theta) = f_\theta(\theta - x)$

b) *If* $x > y > 0$ *then* $f_\theta(y) > f_\theta(x)$

c) $\theta > 0$

then $z'Sz$ *implies* $\Delta = P(Z = z) - P(Z = z') > 0$.

PROOF:

$$\Delta = N! \int_{0 \leqq y_1 \leqq \cdots \leqq y_N \leqq \infty} \cdots \int D(y_i) \left\{ \prod_{\substack{k=1 \\ i \neq k}}^{N} [f^{z_k}(y_k, \theta) f^{1-z_k}(-y_k, \theta)] \right\} \left[\prod_{k=1}^{N} dy_k \right]$$

and it is sufficient to show that $D(y_i) = f(y_i, \theta) - f(-y_i, \theta) > 0$. First, using assumption a, $D(y_i) = f_\theta(y_i - \theta) - f_\theta(-y_i - \theta) = f_\theta(y_i - \theta) - f_\theta(y_i + \theta)$. Now if $y_i > \theta$ the result follows from b, since $y_i - \theta < y_i + \theta$. If $y_i < \theta$ the result follows from b when we write $D(y_i) = f_\theta(\theta - y_i) - f_\theta(y_i + \theta)$.

Remark 1. In Theorem 3.2 writing $f(x, \theta) = f_\theta(x - \theta)$ allows $f(x, \theta)$ not only to be translations of the H_0 but also other changes, such as changes in scale, can occur.

Remark 2. The assumptions of Theorem 2.2 imply those of Theorem 3.2 but not conversely. If in b of Theorem 2.2 we set $\delta = 0$, $2\theta = x + y$, and $0 < y < x$ we obtain b of Theorem 3.2. The Cauchy density is a counter example of the converse.

4. Some partial orderings. If the assumptions of Theorems 2.1 and/or of 2.2 and 3.2 hold, then the following diagrams are obtained:

$N = 1$

$$1 \rightarrow 0$$

(where $P(Z = z) > P(Z = z') \equiv z \rightarrow z'$)

$N = 2$

$$11 \rightarrow 01 \rightarrow 10 \rightarrow 00$$

$N = 3$

$$\begin{tikzcd}
111 \arrow[r] & 011 \arrow[r] & 101 \arrow[r] \arrow[dr] & 110 \arrow[dr] & & & \\
& & & 001 \arrow[r] & 010 \arrow[r] & 100 \arrow[r] & 000
\end{tikzcd}$$

$N = 4$

$$\begin{tikzcd}
1111 \arrow[r] & 0111 \arrow[r] & 1011 \arrow[r] \arrow[dd] & 1101 \arrow[r] \arrow[dd] & 1110 \arrow[d] & & & & \\
& & & & 0110 \arrow[dr] & & & & \\
& & 0011 \arrow[r] & 0101 \arrow[ur] \arrow[dr] & & 1010 \arrow[r] \arrow[dd] & 1100 \arrow[dd] & & \\
& & & & 1001 \arrow[ur] \arrow[d] & & & & \\
& & & & 0001 \arrow[r] & 0010 \arrow[r] & 0100 \arrow[r] & 1000 \arrow[r] & 0000
\end{tikzcd}$$

Now consider the uniform distribution $f(x, \theta) = 1$ for $\theta - \frac{1}{2} \leqq x \leqq \theta + \frac{1}{2}$ and 0 otherwise, $0 \leqq \theta \leqq \frac{1}{2}$. If n' is the length of the last run of 1's in z or n' = the number of the positive observations greater than the maximum of the absolute values of the negative observations, then

$$P(Z = z) = \sum_{i=0}^{n'} \binom{N}{i} (\tfrac{1}{2} - \theta)^{N-i}(2\theta)^i \tag{4.1}$$

To obtain (4.1), begin with

$$P(Z = z) = \sum_{i=0}^{n'} P(Z = z \mid i \text{ observations} > \tfrac{1}{2} - \theta) \times P(i \text{ observations} > \tfrac{1}{2} - \theta)$$

and use

$$P(Z = z \mid i \text{ observations } > \tfrac{1}{2} - \theta) = 2^{-(N-i)},$$

$$P(i \text{ observations } > \tfrac{1}{2} - \theta) = \binom{N}{i} (2\theta)^i(1 - 2\theta)^{N-i}.$$

Holding θ fixed, $P(Z = z)$ is an increasing function of n', and otherwise does not depend on z. Thus, the most powerful rank order tests depend solely on n'.

5. A statistical application. For the normal alternative hypotheses, mentioned at the end of Section 1, several test statistics have been proposed:

a. On intuitive grounds Wilcoxon proposed the statistic

$$T_W = \sum_{i=1}^{N} z_i i.$$

b. Fraser [1] showed the locally most powerful rank order test is of the form $T_F = \sum_{i=1}^{N} z_i E(X_{Ni})$ where X_{Ni} is the ith order statistic from the chi distribution with one degree of freedom.

Both of these statistics are of the form $T = \sum_{i=1}^{N} z_i a_i$ where the a_i form an increasing sequence. It is easily verified that if $z'Lz$ and/or $z'Sz$ then $T(z) > T(z')$. Thus statistics of this form take full advantage of the results of this paper, i.e., using these statistics the known more probable rank orders are put into the critical region first.

6. Normal slippage. The theorems of Sections 2 and 3 do not help in the ordering of $P_1 = P(Z = (0, 0, 1))$ and $P_2 = P(Z = (1, 1, 0))$, for normal alternatives. If $P_1 > P_2$ then the partial order for $N = 3$ given in Section 4 becomes the simple order:

$$111 \rightarrow 011 \rightarrow 101 \rightarrow 001 \rightarrow 110 \rightarrow 010 \rightarrow 100 \rightarrow 000$$

THEOREM 6.1.[2] *If $X_1, \cdots, X_N (N \geqq 3)$ are independently and normally distributed, each with mean $\theta(>0)$ and variance* 1, *then* $\Delta = P(Z = z) -$

[2] M. Sobel proved this result for $N = 3$ at the 1958 Summer Statistical Institute sponsored by the National Science Foundation.

$P(Z = z') > 0$ *where* z *and* z' *are identical except* $z_1 = z_2 = z_3' = 0$ *and* $z_1' = z_2' = z_3 = 1$.

Proof: Using (1.1)

$$\Delta = \frac{N!}{(2\pi)^{3/2}} \int \cdot \quad \cdot \quad \cdot \int \left\{ \prod_{i=4}^{N} [f^{z_i}(y_i, \theta) f^{1-z_i}(-y_i, \theta)] \right\} \left[\prod_{i=1}^{N} dy_i \right]$$
$$\times \{\exp [- \tfrac{1}{2} (y_1^2 + y_2^2 + y_3^2 + 3\theta^2)]\} \times [e^{\theta(y_3 - y_1 - y_2)} - e^{\theta(y_1 + y_2 - y_3)}]$$

Now make the transformation $y_1 = w_1$, $y_2 = w_1 + w_2$, $y_3 = w_1 + w_2 + w_3$, and $y_i = w_i$ for $i = 4, \cdots, N$. The Jacobian is 1 and the region of integration becomes $0 \leqq w_i \leqq w_4 \leqq w_5 \leqq \cdots \leqq w_N \leqq \infty$ for $i = 1, 2, 3$; and $\sum_{i=1}^{3} w_i \leqq w_4$. Then

$$\Delta = \frac{N!}{(2\pi)^{3/2}} \int \cdot \quad \cdot \quad \cdot \int \left\{ \prod_{i=4}^{N} [f^{z_i}(w_i, \theta) f^{1-z_i}(-w_i, \theta)] \right\} \left[\sum_{i=1}^{N} dw_i \right]$$
$$\times \{\exp [- \tfrac{1}{2} (w_1^2 + (w_1 + w_2)^2 + (w_1 + w_2 + w_3)^2 + 3\theta^2)]\}$$
$$\times [e^{(\theta w_3 - w_1)} - e^{\theta(w_1 - w_3)}]$$

The above integral is equivalent to the following integral, where the region of integration is $0 \leqq w_1 \leqq w_3 \leqq w_4 \cdots \leqq w_N \leqq \infty, 0 \leqq w_2 \leqq w_4$, and $\sum_{i=1}^{3} w_i \leqq w_4$.

$$\Delta = \frac{N!}{(2\pi)^{3/2}} \int \cdot \quad \cdot \quad \cdot \int \left[\prod_{i=4}^{N} [f^{z_i}(w_i, \theta) f^{1-z_i}(-w_i, \theta)] \right] \left[\prod_{i=1}^{N} dw_i \right]$$
$$\times \{\exp [- \tfrac{1}{2} (w_1^2 + (w_1 + w_2)^2 + (w_1 + w_2 + w_3)^2 + 3\theta^2)]$$
$$- \exp [- \tfrac{1}{2} (w_3^2 + (w_3 + w_2)^2 + (w_3 + w_2 + w_1)^2 + 3\theta^2)]\}$$
$$\times [e^{\theta(w_3 - w_1)} - e^{\theta(w_1 - w_3)}]$$

For the region of integration each of the factors in the above integrand is clearly >0 except for the $\{\quad\}$. To show $\{\quad\} > 0$, prove the equivalent inequality

$$w_3^2 + (w_3 + w_2)^2 > w_1^2 + (w_1 + w_2)^2 \equiv w_3(w_3 + w_2) > w_1(w_1 + w_2)$$

which is clearly the case since $w_3 > w_1 > 0$.

Theorem 6.1 implies a simple order for the five most probable rank orders against normal slippage.

REFERENCES

[1] D. A. S. Fraser, "Most powerful rank-type tests", *Ann. Math. Stat.*, Vol. 28 (1957), pp. 1040–1043.

[2] I. R. Savage, "Contributions to the theory of rank order statistics—the two sample cases", *Ann. Math. Stat.*, Vol. 27, (1956), pp. 590–615.

[3] I. R. Savage, "Contributions to the theory of rank order statistics—the 'trend case' ", *Ann. Math. Stat.*, Vol. 28, (1957), pp. 968–977.

[4] F. Wilcoxon, *Some Rapid Approximate Statistical Procedures*, American Cyanamid Co., Statistical Research Laboratories (July, 1949), p. 16.