ASYMPTOTICALLY MOST POWERFUL RANK TESTS FOR THE TWO-SAMPLE PROBLEM WITH CENSORED DATA[1]

By Joseph L. Gastwirth

The Johns Hopkins University

1. Introduction and summary. Although the theory of asymptotically most powerful rank tests is fairly well developed, only little attention (see [4]) has been paid to censored data. In this paper we derive the a.m.p.r.t. for censored data from the a.m.p.r.t. based on the complete sample.

For simplicity, we consider the two-sample location problem for data censored on the right. The methods used, however, can be applied to find the a.m.p.r.t. for multi-censored data and also the a.m.p.r.t. for the two-sample dispersion problem when the data are censored.

2. The statement of the problem and preliminary results. Let $X_1, \cdots, X_m$ be the ordered observations of a random sample from a population with absolutely continuous cdf $F(x)$ and density function $f(x)$. Let $Y_1, Y_2, \cdots, Y_n$ be the ordered observations of a random sample from a population with absolutely continuous cdf $G(x)$. Let $N = n + m$ and $\lambda_N = m/N$ and assume that for all N the inequalities $0 < \lambda_0 \leqq \lambda_N \leqq 1 - \lambda_0 \leqq 1$ hold for some fixed $\lambda_0 \leqq \frac{1}{2}$. Finally, let $Z_1, Z_2, \cdots, Z_N$ denote the combined sample, ordered from smallest to largest. The indicator random variables δ_i, $i = 1, \cdots, N$ are defined by

$$(2.1) \qquad \delta_i = 1, \quad \text{if } Z_i \text{ is an } x \text{ observation}$$
$$= 0, \quad \text{otherwise.}$$

The usual rank statistic (T_N) is of the form

$$(2.2) \qquad mT_N = \sum_{i=1}^{n} c_{Ni}\delta_i,$$

where c_{Ni} are given numbers. We shall assume that there is a function $J(u)$ defined on the interval [0, 1] and that $c_{Ni} = J(i/N)$. It is sometimes preferable, for practical purposes, to consider c_{Ni} to be given by $J(i/N + 1) \quad i = 1, \cdots, N$. The extremes are then given finite weights when $J(u)$ approaches infinity at 0 or 1. Clearly, this minor modification does not alter the asymptotic behavior of the statistics T_N[1]. We shall always consider the weighting function to be normalized so that

$$(2.3) \qquad \int_0^1 J(u)\, du = 0.$$

Received 14 December 1964; revised 18 March 1965.

[1] This research was supported in part by the Office of Naval Research, Contract No. Nonr 4010(09), awarded to the Department of Statistics, The Johns Hopkins University. This paper in whole or in part may be reproduced for any purpose of the United States Government.

Let $J(u)$ be defined by

(2.4) $$J(u) = -f'[F^{-1}(u)]/f[F^{-1}(u)].$$

For every finite N, we can define two functions in L^2 [0, 1],

(2.5) $$J_{1N}(u) = J(i/N + 1) \qquad (i - 1)/N < u < i/N,$$

and

(2.6) $$J_{2N}(u) = -E\{f'(v_i)/f(v_i)\} \quad (i - 1)/N < u < i/N,$$

when $v_1 < v_2 < \cdots < v_N$ are the order statistics of a sample of size N from $F(x)$. The function $J_{2N}(u)$ corresponds to the locally most powerful rank test (see [3] p. 277).

Hájek [2], generalizing a result of Chernoff and Savage (p. 983 of [2]), showed that if

(a) the density function $f(x) = F'(x)$ exists

(b) $[f(x)]^{\frac{1}{2}}$ possesses a square integrable derivative, i.e.,

$$\int_{-\infty}^{\infty} [f'(x)/f(x)]^2 f(x)\, dx = \int_0^1 J^2(u)\, du < \infty$$

(c) $$\lim_{N\to\infty} \int_0^1 [J_{1N}(u) - J(u)]^2\, du = 0,$$

then an a.m.p.r.t. for testing $H_0 : F(x) = G(x)$ against $H_1 : G(x) = F(x - \theta)$, where θ tends to zero at the rate $kN^{-\frac{1}{2}}$, is given by the test based on $J(u)$, i.e., on the sequence of functions $J_{1N}(u)$. Furthermore, he showed that $\lim_{N\to\infty} \int_0^1 [J_{2N}(u) - J(u)]^2\, du = 0$ and the functions $J_{2N}(u)$ also generate an a.m.p.r.t.

Sometimes we cannot observe the entire ordered sample $(Z_1, \cdots, Z_N)$ but only the first $N^*(<N)$. The question arises as to what is the a.m.p.r.t. if we are given $(Z_1, \cdots, Z_{N^*})$ and thus know how many x's and y's are larger than Z_{N^*}. We call such data observation censored because the censoring point is determined by the sample, in particular by the N^*th largest member of the combined array. We assume that N^*/N approaches $p(0 < p < 1)$ as N tends to ∞ so that the censoring is always done by the same percentile of the combined sample. It is convenient for our purposes to work with the function $J(u)$ [(2.4)] which generates the a.m.p.r.t. rather than handle the sequence of functions $J_{1N}(u)$.

3. The solution. In order to determine the weighting function $K(u)$ associated with the a.m.p.r.t. for censored data we shall use the results of Hájek [2]. In [2] Hájek proved that the limiting Pitman efficiency of two tests based on the weighting functions $J(u)$ and $L(u)$ is given by ρ^2, where

$$\rho = \left[\int_0^1 J(u)L(u)\, du\right] / \left\{\left[\int_0^1 J^2(u)\, du \cdot \int_0^1 L^2(u)\, du\right]^{\frac{1}{2}}\right\},$$

provided that one of the two tests is the a.m.p.r.t. for the problem considered. The value of ρ is simply the null correlation between the rank tests based respectively, on the weight functions $J(u)$ and $L(u)$. The result is analogous to the well known fact that if T is the best (minimum variance unbiased) estimate of a parameter θ and if S is another consistent estimate of θ, then $\mathrm{Var}(T)/\mathrm{Var}(S)$ is equal to the square of the correlation between S and T. This relationship has been discussed in detail by van Eeden [5].

A weighting function that corresponds to an a.m.p.r.t. for data censored at the sample 100pth percentile must be constant on the interval $(p, 1]$ because the last group of ranks clearly cannot be weighted since they are unknown. Therefore, the weighting function $K_p(u)$ for the data censored at the 100pth percentile of the combined sample must satisfy the following conditions:

(a) $K_p(u) \varepsilon L^2[0, 1]$

(b) $K_p(u)$ is constant on the interval $(p, 1]$.

(c) $K_p(u)$ maximizes ρ^2 where $J(u)$ is efficient for the uncensored problem.

We now state the main theorem.

THEOREM 3.1. *If $J(u)$ is the weighting function corresponding to the* a.m.p.r.t. *for testing $H_0 : F(x) \equiv G(x)$ against the alternative $H_1 : G(x) = F(x - \theta)$, then the weighting function $K_p(u)$ corresponding to the* a.m.p.r.t. *for the same problem where we have only the first* pth *fraction of the combined sample is given by*

$$K_p(u) = J(u), \qquad 0 \leqq u \leqq p, \tag{3.2}$$
$$= c, \qquad p < u \leqq 1,$$

where c satisfies

$$\int_p^1 [J(u) - c]\, du = 0, \qquad \text{i.e., } c(1 - p) = \int_p^1 J(u)\, du.$$

PROOF. Consider the Hilbert space $L^2[0, 1]$. Clearly $L^2[0, 1] = L^2[0, p] \oplus L^2(p, 1]$. The function $L(u)$, corresponding to rank tests for censored data are in the subspace $\bar{L}^2[0, 1] = L^2[0, p] \oplus K(p, 1]$ of $L^2[0, 1]$, where $K(p, 1]$ is the one dimensional subspace of $L^2(p, 1]$ spanned by the constant functions. The function $K_p(u)$ in $\bar{L}^2[0, 1]$ which maximizes ρ^2 is, therefore, the projection of $J(u)$ onto $\bar{L}^2[0, 1]$. The projection of $J(u)$ onto $L^2[0, p]$ is just the restriction of the function $J(u)$ to the interval $[0, p]$. The projection of $J(u)$ onto $K(p, 1]$ is that constant function which minimizes $\int_p^1 [J(u) - c]^2\, du$. It is easily seen that $c = (1 - p)^{-1} \cdot \int_p^1 J(u)\, du$. Since $\bar{L}^2[0, 1]$ is the direct sum of $L^2[0, p]$ and $K(p, 1]$, the projection of $J(u)$ onto $\bar{L}^2[0, 1]$ is the sum of the projections on the component spaces. Thus, $K_p(u)$ is given by Equation (3.2).

We now give the asymptotic distribution of our statistic (based on the function $K(u)$ given by (3.2)) assuming the null hypothesis is true.

THEOREM 3.2. *If there is a λ_0 such that $0 < \lambda_0 < \lambda_N = m/(m + n) < 1 - \lambda_0 < 1$ and if $F(x) \equiv G(x)$, then*

$$\lim_{N\to\infty} \Pr[((T_N - \mu_N)/\sigma_N) < t] = \int_{-\infty}^t e^{-u^2/2}/(2\pi)^{\frac{1}{2}}\, du, \tag{3.3}$$

$$\mu_N = \int_0^1 K(u)\, du \tag{3.4}$$

and

$$\lim_{N\to\infty} \lambda_N N \sigma_N^2/(1 - \lambda_N) = \int_0^1 K^2(u)\, du - [\int_0^1 K(u)\, du]^2. \tag{3.5}$$

PROOF. These results follow from a standard derivation [6] of the asymptotic distribution of T_N when $F = G$ where T_N is regarded as the average of a sample of m from the population of N numbers $K(1/N), \cdots, K(N/N)$.

4. Examples. In this section we shall consider the a.m.p.r.t.'s for right censored data which are derived from the standard rank tests. The results are presented in Table 1.

TABLE 1

The censored a.m.p.r.t.'s *corresponding to some commonly used rank tests*

Test	Standard Weight Function	Censoring Percentile	Censored Sample Weight Function
Wilcoxon	$J(u) = u - \frac{1}{2}$, $0 \leqq u \leqq 1$	100pth	$K(u) = u - \frac{1}{2}$, $0 \leqq u \leqq p$; $= p/2$, $p < u \leqq 1$
Normal scores	$J(u) = \Phi^{-1}(u)$, $0 \leqq u \leqq 1$	100pth	$K(u) = \Phi^{-1}(u)$, $0 \leqq u \leqq p$; $= (1 - p)^{-1}\exp\{-\Phi^{-1}(p)^2/2\}$, $p < u \leqq 1$
Normal scores	$J(u) = \Phi^{-1}(u)$, $0 \leqq u \leqq 1$	50th	$K(u) = \Phi^{-1}(u)$, $0 \leqq u \leqq \frac{1}{2}$; $= (2/\pi)^{\frac{1}{2}}$, $\frac{1}{2} < u \leqq 1$
Savage	$J(u) = -\ln(1 - u) - 1$	100pth	$K(u) = -\ln(1 - u) - 1$, $0 \leqq u \leqq p$; $= -\ln(1 - p)$, $p < u \leqq 1$

In practice, it is preferable to use the l.m.p.r.t. which is based on $J_{2N}(u)$. Theorem 3.2 can easily be modified to yield the asymptotic distribution of this statistic.

Finally, we also note that every weight function corresponds to an a.m.p.r.t. for some distribution (see [2], Section 3.8). In particular, the a.m.p.r.t.'s for censored data also correspond to a.m.p.r.t.'s for complete data from a different distribution. This distribution can be shown to be the original one up to the censored percentile and an exponential for the remainder. For instance, the censored Wilcoxon test, when $p = \frac{1}{2}$, corresponds to the density function $g(x)$ defined by

$$\begin{aligned} g(x) &= e^{-x}/(1 + e^{-x})^2, & x \leqq 0, \\ &= \tfrac{1}{4}e^{-x/2}, & x > 0. \end{aligned}$$

The censored normal scores test (censored at $p = \frac{1}{2}$) is the l.m.p.r.t. for the density function $h(x)$ defined by

$$\begin{aligned} h(x) &= (2\pi)^{-\frac{1}{2}} e^{-x^2/2}, & x \leqq 0, \\ &= (2\pi)^{-\frac{1}{2}} \exp\,[-x(\pi/2)^{\frac{1}{2}}], & x > 0. \end{aligned}$$

REFERENCES

[1] Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29** 972–994.

[2] Hájek, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Statist.* **33** 1124–1147.

[3] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

[4] RAO, U. V. R., SAVAGE, I. R. and SOBEL, M. (1960). Contributions to the theory of rank order statistic: the two sample censored case. *Ann. Math. Statist.* **31** 415–426.

[5] VAN EEDEN, CONSTANCE (1963). The relation between Pitman's asymptotic relative efficiency of two-tests and the correlation coefficient between their test statistics. *Ann. Math. Statist.* **34** 1442–1451.

[6] WALD, ABRAHAM and WOLFOWITZ, JACOB (1944). Statistical tests based on permutations of the observations. *Ann. Math. Statist.* **15** 358–372.