COMPLETENESS COMPARISONS AMONG SEQUENCES OF SAMPLES

by

N.L. Johnson*

Department of Statistics

University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1056

February, 1976

---

* This research was supported by the U.S. Army Research Office under Contract No. DAHCO4-74-C-0030.

Completeness Comparisons Among Sequences of Samples

By N.L. Johnson
University of North Carolina at Chapel Hill

1. Introduction

In previous papers [1] - [3] we have considered methods of testing whether a set (or sequence of sets) of observed values represents a complete random sample (or a sequence of complete random samples) or is the consequence of some form of censoring having been applied to a complete random sample (or sequence of such samples). The methods required a knowledge of the population distribution of the measured character(s); some assessment of the effects of imperfect knowledge were discussed in [4] and [5].

We now turn our attention to problems arising when there are two sets (or sequences of sets) of data and we wish to compare the incidence of censoring them. We will suppose the appropriate population distribution to be the same for each set, and also (as in the earlier papers) that the distribution is continuous. Certain special problems will be selected for detailed discussion from among the very wide range of possible problems. A particularly interesting feature of the enquiry is that distribution-free procedures can be used (see Section 3), thus freeing us from the dependence on knowledge of population distributions which characterized the earlier work.

A considerable variety of problems can arise

(i) We may have different types of censoring. We will restrict ourselves to symmetrical censoring of extreme values - omission of the

s greatest and s least sample values.

(ii) We may have different hypotheses in mind. For example we may know that one sample (or sequence of samples) is censored, and the other not, and wish to find which one.

(iii) We may, or may not, know the population distribution.

(iv) The sample sizes in each sequence may not stay the same.

2. "Parametric" Test Procedures

2.1 Fixed Numbers of Samples

We suppose that we have available two samples, one (A_{11}) of size r_1 and one (A_{21}) of size r_2. We denote the order statistics in A_{i1} $(i = 1,2)$ by

$$X_{i1} \leq X_{i2} \leq \ldots \leq X_{ir_i} .$$

We further suppose that it is known that one of the two samples (it is not specified which one) is a complete random sample, while the other is a random sample which has been censored by removal of some extreme observations. We will specialize the problem by supposing that the censoring is symmetrical, i.e. the s greatest and s least values in the original random samples are removed. The value of s may (or may not) be known.

Finally, we suppose that the population cumulative distribution function, $F(x)$, is the same for each of the two samples; and is absolutely continuous with $f(x) = dF/dx$.

Let H_i $(i = 1,2)$ be the hypothesis that the censored sample is A_{i1}. The likelihood function for H_i is

$$(1)\quad L_i = \frac{r_{3-i}!(r_i+2s)!}{(s!)^2}\left[F(X_{i,1})\{1 - F(X_{i,r_i})\}\right]^s \prod_{k=1}^{2}\prod_{j=1}^{r_h} f(X_{k,j})$$

and so the likelihood ratio (H_1/H_2) is

$$(2)\qquad L_1/L_2 = \frac{(r_1+1)^{[2s]}}{(r_2+1)^{[2s]}}\left[\frac{F(X_{1,1})\{1-F(X_{1,r_i})\}}{F(X_{2,1})\{1-F(X_{2,r_2})\}}\right]^2 .$$

Whatever the value of s, we see that discrimination between H_1 and H_2 will be based on the statistic

$$(3)\qquad T = \frac{F(X_{1,1})\{1-F(X_{1,r_1})\}}{F(X_{2,1})\{1-F(X_{2,r_2})\}}$$

$$(3')\qquad = \frac{Y_{11}(1-Y_{1r_1})}{Y_{21}(1-Y_{2r_2})}$$

(4) where $\qquad Y_{ij} = F(X_{ij})$.

(It is worth noting that if the X's have different distributions F_1 and F_2 in A_1, and A_2 respectively, we reach the same criterion (3') with $Y_{ij} = F_i(X_{ij})$ and the Y's have the same joint distribution.) The hypothesis H_1 will be accepted if $T > K$, and the hypothesis H_2 will be accepted if $T < K$. (If $T = K$, an arbitrary decision can be made. Since $Pr[T=K] = 0$, this will not affect the probability properties of the procedure. Choice of K can depend on s, r_1, r_2 and relative costs of different incorrect decisions.

If $r_1 = r_2$, the likelihood ratio is just T^s and it is natural to take $K = 1$, in the absence of information on relative costs of errors (whatever the value of s).

To evaluate probabilities of error, we need the distribution of T. A discussion of this distribution follows. We first evaluate the moments for purposes of record.

If both A_{11} and A_{21} have been censored with the s_i' least and s_i'' greatest observations removed from A_{i1} $(i = 1,2)$ then the joint density function of Y_{11}, Y_{1r_1}, Y_{21}, Y_{2r_2} is

$$(5)\quad f(y_{11},y_{1r_1},y_{21},y_{2r_2}) = \prod_{i=1}^{2} \left\{ \frac{(r_i+s_i'+s_i'')!}{s_i'!s_i''!} y_{il}^{s_i'} (y_{ir_i} - y_{il})^{r_i-2} (1-y_{ir_i})^{s_i''} \right\}$$

$$(0 \le y_{il} \le y_{ir_i} \le 1;\quad i = 1,2).$$

(Hypothesis H_i corresponds to $s_i' = s_i'' = s$; $s_{3-i}' = s_{3-i}'' = 0$.) and the h-th moment of T (h an integer) is

$$(6)\quad \mu_h'(T) = E[T^h] = \prod_{i=1}^{2} \left\{ \frac{(r_i+s_i'+s_i'')!}{s_i'!s_i''!} \right\} \cdot \frac{(s_i'+h)!(s_i''+h)!(s_2'-h)!(s_2''-h)!}{(r_1+s_1'+s_i''+2h)!(r_2+s_2'+s_2''-2h)!}$$

$$= \frac{(s_i'+1)^{[h]}(s_1''+1)^{[h]}}{s_2'^{(h)}s_2''^{(h)}} \cdot \frac{(r_2+s_2'+s_2'')^{(h)}}{(r_1+s_1'+s_i''+1)^{[h]}}.$$

Moments of order $(\min(s_2',s_2'') + 1)$ or higher are infinite. We do not propose to use the moments in approximating the distribution of T.

The distribution of $Z_i = Y_{il}(1 - Y_{ir_i})$ has been studied in [1] and [3]. It is quite a complicated distribution, and the distribution of $T = Z_1/Z_2$ is even more complicated. Here we describe two methods of obtaining an approximation.

As pointed out in [3] we can write

$$(7)\qquad Z_i = U_iW_i(U_i + V_i + W_i)^{-2}$$

where U_i, V_i and W_i are independent random variables distributed as χ^2 with $2(s_i' + 1)$, $2(r_i - 1)$ and $2(s_i'' + 1)$ degrees of freedom, respectively. Hence

$$T = \frac{U_1 W_1 (U_2+V_2+W_2)^2}{U_2 W_2 (U_1+V_1+W_1)^2} \tag{8}$$

all six variables on the right hand side being mutually independent.

If r_1 and r_2 are large compared with s_1', s_1'', s_2' and s_2'' then

$$\frac{(r_1-1)^2}{(r_2-1)^2} T = \frac{U_1 W_1}{U_2 W_2} \cdot \left\{ \frac{(r_2-1)^{-1}(U_2+W_2)+(r_2-1)^{-1}V_2}{(r_1-1)^{-1}(U_1+W_1)+(r_1-1)^{-1}V_1} \right\}^2. \tag{9}$$

Since $(r_i-1)^{-1}(U_i+W_i) \sim 0$ and $(r_i-1)^{-1}V_i \sim 2$ for r_i large, we would expect the large sample distribution of T to be approximated by that of

$$\left(\frac{r_2-1}{r_1-1}\right)^2 \frac{U_1 W_1}{U_2 W_2}$$

$$\text{i.e.} \quad \left(\frac{r_2-1}{r_1-1}\right)^2 \frac{(s_1'+1)(s_1''+1)}{(s_2'+1)(s_2''+1)} \cdot F_{2(s_1'+1),2(s_2'+1)} \, F_{2(s_1''+1),2(s_2''+1)} \cdot \tag{10}$$

The distribution of the product of two independent F-variables has been studied by Schumann and Bradley [6].

They give tables of upper percentage points of products of independent variables distributed as F_{ν_1,ν_2} and F_{ν_2,ν_1} respectively. These will only apply to (10) for the special case $s_1' = s_2''$, $s_2' = s_1''$, while we are at present interested in $s_1' = s_1'' = s$; $s_2' = s_2'' = 0$.

Alternatively we can consider the distribution of log T. The characteristic function of log T is (cf. (6))

$$E[T^{ih}] = \prod_{i=1}^{2} \left\{ \frac{(r_i+s_i'+s_i'')!}{s_i'!s_i''!} \right\} \cdot \frac{\Gamma(s_1'+1+ih)\Gamma(s_1''+1+ih)\Gamma(s_2'+1-ih)\Gamma(s_2''+1-ih)}{\Gamma(r_1+s_1'+s_1''+1+2ih)\Gamma(r_2+s_2'+s_2''+1-2ih)} \cdot \tag{11}$$

Taking derivatives of the cumulant generating function $(\log E[T^{ih}])$ we find

$$(12)\quad \kappa_t(\log T) = \psi^{(t-1)}(s_1' + 1) + \psi^{(t-1)}(s_1'' + 1) - 2^t\psi^{(t-1)}(r_1+s_1'+s_1''+1) + (-1)^t\{\psi^{(t-1)}(s_2'+1) + \psi^{(t-1)}(s_2''+1) - 2^t\psi^{(t-1)}(r_2+s_2'+s_2''+1)\}$$

where $\psi^{(s)}(x) = \dfrac{d^{s+1}\log\ (x)}{dx^{s+1}}$ is the (s+2)-gamma function when H_1 is true, and $r_1 = r_2 = r$, we have $s_1' = s_1'' = s$, and $s_2' = s_2'' = 0$, whence (12) becomes

$$(13)\quad \kappa_t(\log T|H) = 2\psi^{(t-1)}(x+1) - 2^t\psi^{(t-1)}(r+2s+1) + (-1)^t\{2\psi^{(t-1)}(1) - 2^t\psi^{(t-1)}(r+1)\}.$$

Also

$$(14)\quad \kappa_t(\log T|H_2) = 2\psi^{(t-1)}(1) - 2^t\psi^{(t-1)}(r+1) + (-1)^t\{2\psi^{(t-1)}(s+1) - 2^t\psi^{(t-1)}(r+2s+1)\}.$$

In particular $\kappa_1(\log T|H_1) = 2[\psi(s+1) - \psi(1) - \{\psi(r+2s+1) - \psi(r+1)\}]$

$$(15)\quad = -\kappa_1(\log T|H_2)$$

and $\kappa_2(\log T|H_1) = 2[\psi^{(1)}(s+1) + \psi^{(1)}(1) - 2\{\psi^{(1)}(r+1) + \psi^{(1)}(r+2s+1)\}]$

$$(16)\quad = \kappa_2(\log T|H_2).$$

Some values of $\{\kappa_1(\log T|H_1) - \psi_1(\log T|H_2)\}/\sqrt{\kappa_2(\log T|H_i)} = \rho \quad (i = 1,2)$ are shown in Table 1.

TABLE 1: Values of ρ

$r_1=r_2=r\backslash s$	1	2	$r\backslash s$	1	2
4	1.44	2.10	8	1.63	2.46
5	1.52	2.23	9	1.66	2.51
6	1.57	2.33	10	1.68	2.55
7	1.60	2.40	20	1.77	2.74
			∞	1.87	2.97

(The values for $r = \infty$ are the limiting values

$$\frac{4\{\psi^{(s+1)}-\psi(1)\}}{[2\{\psi^{(1)}(s+1)+\psi^{(1)}(1)\}]^{\frac{1}{2}}}\ .)$$

These values of ρ are of such a size as to indicate that even with a single pair of samples, fairly good discrimination is attainable, at least when $r_1 = r_2 = r$.

If we choose H_1 (H_2) when $\log T > (<) 0$ then approximating the distribution of $\log T$ by a normal distribution, the probability of correct decision would be $\Phi(\frac{1}{2}\rho)$.

If we have two *sequences* of samples

$$S_1 : A_{11}, A_{12}, \cdots A_{1m}$$

and

$$S_2 : A_{21}, A_{22}, \cdots A_{2m}$$

we would use as criterion

$$T_{(m)} = \prod_{j=1}^{m} T_j \tag{17}$$

where T_j is the value of T (as defined in (3)) for the j-th pair of samples A_{1j}, A_{2j} ($j = 1,2,\ldots m$). Denoting the order statistics

of sample A_{ij} by

$$X_{ij1} \leq X_{ij2} \leq \ldots \leq X_{ijr_i}$$

we have

$$T_j = \frac{Y_{ij1}(1-Y_{ijr_1})}{Y_{2j1}(1-Y_{2jr_2})} \tag{18}$$

where $Y_{ijh} = F(X_{ijh})$. The second method of approximation (described above for T) is the more simply extended to $T_{(m)}$. Since $T_1, T_2, \ldots T_m$ are mutually independent we have

$$\kappa_r(\log T_{(m)}) = \sum_{j=1}^{m} \kappa_r(\log T_j) = m\kappa_r(\log T) \tag{19}$$

(on the assumption that r_1, r_2, s_1', s_1'', s_2' and s_2'' remain constant throughout). It is not essential that the X's have the same distribution in each of the samples A_{ij}, provided the appropriate transformations to the Y's are used (see the remarks following (4)). (Nor is it essential that there be the same numbers of samples in the sequences S_1, S_2; the modifications in such a case are obvious and will not be discussed here.)

From (19) we can see that there will be increasing power (i.e. probability of reaching a correct decision) as the lengths of the sequences S_1, S_2 increase. In fact, results obtained in the next Section support the view, stated above, that the fixed sample procedure is likely to be quite effective, even with a single pair of samples.

2.2 A Sequential Procedure

In the circumstances described above if pairs of samples A_{1j}, A_{2j} from the sequences S_1, S_2 become available at about the same time,

while there is an appreciable interval between successive pairs - (A_{1j},A_{2j}) and $(A_{1,j+1},A_{2,j+1})$ - it is natural to consider using a sequential test procedure. A sequential probability ratio test discriminating between H_1 and H_2, using the pairs of samples as they become available, is based on the continuation region

$$(20) \qquad \frac{\alpha_1}{1-\alpha_2} < \left\{\frac{(r_1+1)^{[s]}}{(r_2+1)^{[s]}}\right\}^m \prod_{j=1}^{m} \left[\frac{Y_{1j1}(1-Y_{1jr_1})}{Y_{2j1}(1-Y_{2jr_2})}\right]^s < \frac{1-\alpha_1}{\alpha_2}$$

with acceptance of H_1 (H_2) if the right (left) hand inequality is violated. (α_1, α_2 are the nominal chances of error when the hypotheses H_1, H_2 respectively are valid.)

The remainder of this section will be devoted to the special case $r_1 = r_2 = r$, in which (15) becomes

$$(21) \qquad \left(\frac{\alpha_1}{1-\alpha_2}\right)^{1/s} < \prod_{j=1}^{m} \left\{\frac{Y_{1j1}(1-Y_{1jr})}{Y_{2jr}(1-Y_{2jr})}\right\} < \left(\frac{1-\alpha_1}{\alpha_2}\right)^{1/s} .$$

The limits depend on s, as well as on α_1 and α_2. The larger s, the closer together are the limits and the earlier a decision is reached. If an incorrect value of s - $\hat{s}$, say - is used in (17), then the approximate actual chances of error $\alpha_1(\hat{s})$ and $\alpha_2(\hat{s})$ can be obtained from the formulae

$$\left\{\frac{\alpha_1(\hat{s})}{1-\alpha_2(\hat{s})}\right\}^{1/s} \doteqdot \left(\frac{\alpha_1}{1-\alpha_2}\right)^{1/\hat{s}} ; \quad \left\{\frac{1-\alpha_1(\hat{s})}{\alpha_2(\hat{s})}\right\}^{1/} \doteqdot \left(\frac{1-\alpha_1}{\alpha_2}\right)^{1/\hat{s}}$$

whence

$$(22) \quad \alpha_i(\hat{s}) \doteqdot \left[\left(\frac{\alpha_i}{1-\alpha_{3-i}}\right)^{s/\hat{s}} - \left\{\frac{\alpha_1\alpha_2}{(1-\alpha_1)(1-\alpha_2)}\right\}^{s/\hat{s}}\right] \left[1 - \left\{\frac{\alpha_1\alpha_2}{(1-\alpha_1)(1-\alpha_2)}\right\}^{s/\hat{s}}\right]^{-1} .$$

From (15)

$$(23)\quad E_1 = E\left[s \log\left\{\frac{Y_{1j1}(1-Y_{1jr})}{Y_{2j1}(1-Y_{2jr})}\right\}\middle| H_1\right] = s[2\psi(s+1)-2\psi(r+2s+1)-2\psi(1)+2\psi(r+1)]$$

$$= 2s\left(\sum_{j=1}^{s} j^{-1} - \sum_{j=r+1}^{r+2s} j^{-1}\right)$$

and

$$(24)\quad E_2 = -E_1$$

is an obvious notation.

The standard approximate formula for the average sample number (ASN) of the procedure when H_1 is valid, if

$$(25)\quad E_1^{-1}\left\{\alpha_1 \log\frac{\alpha_1}{1-\alpha_2} + (1-\alpha_1)\log\frac{1-\alpha_1}{\alpha_2}\right\}.$$

Table 2 presents a few values of this quantity for the symmetrical case $\alpha_1 = \alpha_2 = \alpha$. The values for $\alpha = 0.05$ are not shown. They are so small that they clearly are only very rough approximations. (It is, of course, impossible for the ASN to be less than one, in reality.) They do indicate, however, that a decision can be expected very soon in the process.

TABLE 2: Approximate Average Sample Numbers

	$\alpha = 0.01$		$\alpha = 0.001$	
r\s	1	2	1	2
4	3.6	1.3	5.4	2.0
5	3.3	1.2	5.0	1.8
6	3.1	1.1	4.7	1.7
7	2.9	1.0	4.5	1.6
8	2.9	1.0	4.4	1.5
9	2.8	1.0	4.3	1.5
10	2.7	1.0	4.2	1.5
20	2.5	(0.9)	3.8	1.3
∞	2.3	(0.8)	3.4	1.1

(The "$r = \infty$" values are calculated from the formula $\frac{(1-2\alpha)\log\{(1-\alpha)/\alpha\}}{2s\sum_{j=1}^{s} j^{-1}}$.)

3. Distribution-Free Tests

3.1 Single Pair of Samples

We suppose we have a single pair of samples A_{11}, A_{21} as described at the beginning of Section 2.1. It is remarkable that it is possible to construct a test of H_1 (or H_2) in this case with a known significance level without knowing the population distribution function of the X's, provided only that it is continuous, and the same for both A_{11} and A_{21}. Heuristically one might "expect" this, on the grounds that one of the two samples (the one which is not censored) provides some information on the population distribution.

The only assumption we need (apart from continuity) is that the population distribution of X is the same for A_{11} and A_{21}.

Suppose H_1 is valid. Then the original sample sizes were $(r_1 + 2s)$ for A_{11} and r_2 for A_{21}. For the original set of $(r_1 + r_2 + 2s)$ sample values there would be $\binom{r_1+r_2+2s}{r_2}$ possible rankings of these values, according to origin, in order of ascending magnitude. Since the population distributions for A_{11} and A_{21} are identical, these orderings would be equally likely, each having probability $\binom{r_1+r_2+2s}{r_2}^{-1}$. (Under H_2, the number of equally likely orderings would be $\binom{r_1+r_2+2s}{r_1}$.)

To calculate the likelihood, based on ranks, corresponding to the specific sets of observed ranks we have to evaluate the numbers of rankings of the original (r_1+r_2+2s) observations which could have produced the observed ranks (of (r_1+r_2) observed values) after censoring. Under

H_1 , we can recover the original ranking by adding to A_{11} , s observations less than the least observed value in A_{11} , and s greater than the greatest observed value in A_{11} .

If, among the (r_1+r_2) observed values in A_{11} and A_{21} combined, the L_2 least and the G_2 greatest are from A_{21} then there are

$$\binom{L_2+s}{s}\binom{G_2+s}{s}$$

corresponding original rankings. Hence the likelihood function for H_1 is

$$L_1' = \binom{L_2+s}{s}\binom{G_2+s}{s}\Big/\binom{r_1+r_2+2s}{r_2}. \tag{26}$$

Similarly, the likelihood function for H_2 is

$$L_2' = \binom{L_1+s}{s}\binom{G_1+s}{s}\Big/\binom{r_1+r_2+2s}{r_1} \tag{27}$$

where L_1, G_1 are defined analogously to L_2, G_2 . The likelihood ratio (H_1/H_2) is

$$\frac{L_1'}{L_2'} = \frac{(r_1+1)^{[2s]}}{(r_2+1)^{[2s]}} \cdot \frac{(L_2+1)^{[s]}(G_2+1)^{[s]}}{(L_1+1)^{[s]}(G_1+1)^{[s]}} . \tag{28}$$

In the special case of equal observed sample sizes $(r_1 = r_2 = r)$ we have

$$\frac{L_1'}{L_2'} = \frac{(L_2+1)^{[s]}(G_2+1)^{[s]}}{(L_1+1)^{[s]}(G_1+1)^{[s]}} . \tag{29}$$

If we construct a rule: choose H_1 (H_2) if $L_1'/L_2' > (<)\ 1$, then the decision appears to depend on s, as well as L_1, L_2, G_1 and G_2 . This was not the case with the "parametric" test (see (3)).

However, we note that just one of L_1 and L_2 must be zero, and just one of G_1 and G_2 must be zero. If both L_1 and G_1, (or both L_2 and G_2) are zero then the test will accept H_1 (or H_2) whatever be the value of s. If L_1 and G_2 are zero then the test accepts H_1 (H_2) if $L_2 > (<) G_1$; and if L_2 and G_1 are zero then the test accepts H_1 (H_2) if $G_2 > (<) L_1$. In all these cases the decision is in fact not dependent on s. So we can present the test in the form: accept H_1 (H_2) if

$$L_2 > G_1 \quad \text{or} \quad G_2 > L_1 \quad (L_2 < G_1 \quad \text{or} \quad G_2 < L_1) \tag{30}$$

or even more simply: accept H_1 (H_2) if

$$L_2 + G_2 > (<) L_1 + G_1 \tag{30'}$$

whatever be the value of s. (Note that $L_2 > G_1$ implies $L_2 > 0$ and so also $L_1 = 0$.)

If $L_2 = G_1 > 0$ or $G_2 = L_1 > 0$, we are unable to reach a decision.

In order to obtain the distribution of L_1'/L_2' we have to take into account not only the probabilities (26) or (27) but also the number of possible orderings of the values between the $\max(L_1+1, L_2+1)$ and $\min(r_1+r_2-G_1, r_1+r_2-G_2)$ order statistics of the combined $(A_1 \cup A_2)$ sample. When $L_1 = r_1$ we must have $G_2 = r_2$, and when $G_1 = r_1$ we must have $L_2 = r_2$, and conversely. Apart from this case we have, in general, when $\max(L_1,G_1) < r_1$ and $\max(L_2,G_2)<r_2$ to multiply the probabilities in (26) (when H_1 is valid) or (27) (when H_2 is valid) by

$$\binom{r_1+r_2-2-L_1-L_2-G_1-G_2}{r_1-L_1-G_1-\phi} \tag{31}$$

where ϕ (= 0,1,2) is the number of zeroes in (L_1,G_1). We further note that $L_i + G_i \le r_i$ $(i = 1,2)$ and (31) is equal to 1 if any of L_i, G_i equals r_i or (r_i-1) for either $i = 1$ or 2.

Table 2 sets out the probabilities for the case H_1 valid. A similar table can easily be constructed for H_2 valid. (In calculating (28) we take $\binom{0}{0} = 1$.)

Table 2: Distribution of L_1'/L_2' under H_1

$L_1=$	$L_2=$	$G_1=$	$G_2=$	PROBABILITY$\times\binom{r_1+r_2+2s}{r_2}$	$(L_1'/L_2')\cdot(r_2+1)^{[2s]}/(r_1+1)^{[2s]}$
0	ℓ_2	0	g_2	$(s!)^{-1}(\ell_2+1)^{[s]}(g_2+1)^{[s]}\times(31)$	$(\ell_2+1)^{[s]}(g_2+1)^{[s]}/(s!)^2$
0	ℓ_2	g_1	0	$(s!)^{-1}(\ell_2+1)^{[s]}\times(31)$	$(\ell_2+1)^{[s]}/(g_1+1)^{[s]}$
ℓ_1	0	0	g_2	$(s!)^{-1}(g_2+1)^{[s]}\times(31)$	$(g_2+1)^{[s]}/(\ell_1+1)^{[s]}$
ℓ_1	0	g_1	0	$1\times(31)$	$(s!)^2/\{(\ell_1+1)^{[s]}(g_1+1)^{[s]}\}$

(For $\max(\ell_1,g_1) < r_1$; $\max(\ell_2,g_2) < r_2$. If $\max(\ell_1,g_1) = r_1$ then $\max(\ell_2,g_2) = r_2$ and the entry in the fifth column is $(r_2 + 1)^{[s]}/s!$.)

When $r_1 = r_2$ the last column gives the values of L_1'/L_2' . If acceptance of H_1 follows when $(L_2 > G_1)$ or $(G_2 > L_1)$ then

$$\begin{aligned}\Pr[\text{correct decision}|H_1] &= \Pr[(L_2>G_1) \cup (G_2>L_1)|H_1] \\ &= \Pr[L_2>G_1>0|H_1] + \Pr[G_2>L_1>0|H_1] + \Pr[L_1=G_1=0|H_1]\end{aligned} \tag{32}$$

(remembering that $L_2 > 0$, $G_2 > 0$ imply $L_1 = 0$, $G_1 = 0$ and conversely).

From (32) and Table 2 it is possible to evaluate the probability of a correct decision (and the probability of not reaching a decision) when

H_1 is valid. In the symmetrical case $(r_1=r_2=r)$ these probabilities have the same values when H_2 is valid.

The calculations for $r_1 = r_2 = 4$ are set out below in detail:

$$\binom{2r+2s}{r} = \binom{8+2s}{4} = 210(s=1);\ 495(s=2)$$

$$(\ell_2+1)^{[s]}(g_2+1)^{[s]}/(s!)^2$$

ℓ_1	ℓ_2	g_1	g_2	(31)	s=1	s=2			
4	0	0	4	-	5	15			(α)"
3	0	0	3	1	4	10			(α)"
2	0	0	3	1	4	10		(α)'	
1	0	0	3	1	4	10		(α)'	
0	1	0	3	1	8	30	(α)		
2	0	0	2	2	3	6			(α)"
1	0	0	2	3	3	6		(α)'	
0	1	0	2	3	6	18	(α)		
0	2	0	2	1	9	36	(α)		
1	0	0	1	6	2	3			(α)"
0	1	0	1	6	4	9	(α)		

		s=1	s=2
(α)	$p = \Pr[L_1=G_1=0\|H_1] = \frac{(2\times8)+(2\times18)+9+24}{210}$	$= \frac{85}{210}$	$\frac{258}{495}$
(α)'	$p' = \Pr[G_2>L_1>0\|H_1] = \frac{4+4+9}{210}$	$= \frac{17}{210}$	$\frac{38}{495}$
(α)"	$p'' = \Pr[G_2-L_1>0\|H_1] = \frac{5+4+6+12}{210}$	$= \frac{27}{210}$	$\frac{55}{495}$

	s=1	s=2
Probability of correct decision $(p+2p')$	$\frac{119}{210} = 0.567$	$\frac{334}{495} = 0.675$
Probability of no decision $(2p'')$	$\frac{54}{210} = 0.257$	$\frac{110}{495} = 0.222$
Probability of incorrect decision	$\frac{37}{210} = 0.176$	$\frac{51}{495} = 0.103$

Table 3 presents the results of similar calculations for $r = 4(1)9$.

TABLE 3: Properties of Distribution-Free Test Based on a Single Pair of Samples Each Containing r Observed Values

r\s	1			2		
	PROB. CORRECT	PROB. NO DEC.	PROB. INCORRECT	PROB. CORRECT	PROB. NO DEC.	PROB. INCORRECT
4	0.567	0.257	0.176	0.675	0.222	0.103
5	0.619	0.194	0.187	0.742	0.152	0.107
6	0.647	0.165	0.188	0.778	0.117	0.104
7	0.664	0.149	0.187	0.801	0.099	0.100
8	0.676	0.139	0.184	0.815	0.088	0.097
9	0.684	0.133	0.182	0.827	0.081	0.092

The figures in Table 3 are fairly encouraging, when it is realized that we are trying to reach a conclusion on the basis of a single pair of samples, without knowing anything of the population distribution, except that it is continuous.

3.2 Sequence of Pairs of Samples

If we have two sequences, S_1 and S_2, as described in Section 2 (pages 7-8), we can expect considerable increases in power. The likelihood ratio test criterion, based on two sequences of m samples each, would be

$$\prod_{j=1}^{m} (L'_{1j}/L'_{2j}) \tag{33}$$

in an obvious notation. If $r_1=r_2=r$ the criterion would be

$$\prod_{j=1}^{m} \left[\frac{(L_{2j}+1)^{[s]}(G_{2j}+1)^{[s]}}{(L_{1j}+1)^{[s]}(G_{1j}+1)^{[s]}} \right] \tag{34}$$

and it would be reasonable to assign to H_1 (H_2) if its value is $>$ ($<$) 1, reaching no conclusion if it equals 1.

Probabilities of correct and incorrect decision with this criterion, for the case $m = 2$, are shown in Table 4 for $r = 4(1)6$.

TABLE 4: Properties of Distribution Free Test Based on Two Pairs of Samples Each Containing r Observed Values

r\s	1			2		
	PROB. CORRECT	PROB. NO. DEC.	PROB. INCORRECT	PROB. CORRECT	PROB. NO. DEC.	PROB. INCORREC
4	0.791	0.117	0.093	0.897	0.076	0.027
5	0.832	0.084	0.085	0.936	0.044	0.020
6	0.851	0.066	0.083	0.954	0.029	0.017

These figures do, indeed, show considerable improvement over those in Table 3. They also confirm the conclusions in Section 2 on the power of discrimination when the population distribution(s) is (are) known, since the latter will be more powerful than the present procedure.

If it be supposed that the population distribution common to the two members of a pair of samples is the *same* for each pair, then a better distribution free procedure should be feasible, taking account of this additional knowledge.

To construct such a procedure, it is necessary to enumerate all arrangements of the original $m(r_1 + r_2 + 2s)$ observations, which could result in the observed sequences $S_1(\equiv A_{11}, \ldots, A_{1m})$ and $S_2(\equiv A_{21}, \ldots, A_{2m})$ after removal of the s greatest and s least observations from each of the members of S_j (for H_j valid; $j = 1,2$). Under H_j, the mr_{3-j} observations in S_{3-j} constitute a random sample of that size, but the mr_j observations in S_j have been censored in a special way.

Evaluation of the required probabilities appears to be rather complex, though not impossible. In view of the good power attainable with the

procedure already discussed in this section, (which can also be applied when variation in population distribution from pair to pair is suspected) we will not investigate further possibilities here.

A distribution free sequential probability ratio test has the continuation region

$$(35) \qquad \frac{\alpha_1}{1-\alpha_2} < \left\{\frac{(r_1+1)^{[2s]}}{(r_2+1)^{[2s]}}\right\}^m \prod_{j=1}^{m} \left\{\frac{(L_{2j}+1)^{[s]}(G_{2j}+1)^{[s]}}{(L_{1j}+1)^{[s]}(G_{1j}+1)^{[s]}}\right\} < \frac{1-\alpha_1}{\alpha_2}$$

with $H_1 (H_2)$ accepted if the left (right) inequality is the first to be violated.

4. Some Other Problems

In the situation described in Section 3.1, if we just want to test the hypothesis: "Is A_1 uncensored (assuming A_2 to be uncensored?",) the likelihood criterion is simply

$$(36) \qquad W = (L_2 + 1)^{[s]}(G_2 + 1)^{[s]}$$

with high values of W leading to rejection of the hypothesis.

(Note that $(L_2 + 1)$ = rank order of least member of A, and $(r_1 + r_2 - G_2)$ = rank order of greatest member of A in the combined set of $(r_1 + r_2)$ observed sample values.)

Under $H_{s,s}$ (using the notation of [3] for symmetrical censoring of the s extreme values from each end of the sample), the joint distribution of L_2 and G_2 is

$$(37)\qquad \Pr[(L_2=\ell_2)\cap(G_2=g_2)] = \frac{\binom{r_1+r_2-\ell_2-g_2-2}{r_2-2}\binom{\ell_2+s}{s}\binom{g_2+s}{s}}{\binom{r_1+r_2+2s}{r_2}}$$

$$(0 < \ell_2, g_2;\quad \ell_2 + g_2 \le r_2).$$

It is interesting to note that if r_2 is large, so that we have a lot of information on the population distribution provided by a large random sample, known to be uncensored, then

$$(38)\qquad Wr_2^{-2s} \doteqdot Y_{11}(1 - Y_{1r_1})$$

which is the criterion used to test for symmetrical censoring (see (2) of [3]) when the population distribution is known.

We now, for a moment, consider the problem of detecting which one, out of k sequences $S_1, S_2, \ldots, S_k$ of samples, has been symmetrically censored. We identify the hypothesis H_i with the statement "S_i is the censored sequence." For simplicity we will suppose that each available sample contains r observed values.

The method, of course, is to consider each set in turn as the censored one, the other $(k - 1)$ sets being uncensored. The appropriate likelihoods are those appropriate to the case of two samples (as in Sections 2 and 3) of sizes r and $(k - 1)r$ respectively, the smaller sample being the residue after removing the s greatest and s least values from a complete random sample of size $(r + 2s)$.

If the population distribution function(s) $F_j(\cdot)$ are known the sequence S_i indicated as the censored one by this method is that for which

$$(39) \qquad T_i = \prod_{j=1}^{m} [F_i(X_{ji})\{1 - F_i(X_{jr})\}]$$

is maximum among $T_1, T_2, \ldots, T_k$. The distribution-free approach leads to selection of that S_i for which

$$(40) \qquad V_i = \prod_{j=1}^{m} (L_{ij} + 1)(G_{ij} + 1)$$

is maximum among $V_1, V_2, \ldots, V_k$ where $L_{ij}(G_{ij})$ is the number of observations less (greater) than the least (greatest) observation in A_{ij} among $A_{1j}, A_{2j}, \ldots, A_{kj}$ for $j = 1,2,\ldots,m$.

Since we have a relatively large uncensored sample available (though we are not certain where it is) we should be able to construct good distribution-free procedures in this case. When H_i is valid and we are considering S_i versus the remaining sequences we are in effect in the situation, described earlier in this Section, wherein we have a relatively large uncensored random sample providing information on the population distribution. When any other sequence - S_i, say - is the one compared with the remainder, such bias as is introduced by the presence of S_i among the remainder tends to reduce apparent significance. It therefore seems that a procedure in which each sequence in turn is tested for symmetrical censoring against the remainder, in the way described at the beginning of this section, will provide useful information. The situation becomes more complicated if there can be more than one censored sequence, especially if the number of censored sequences is not known precisely.

Sequential procedures, choosing among the hypotheses $H_1, H_2, \ldots, H_k$ may be constructed in a straightforward way.

The discussion in this paper has been restricted to symmetrical censoring of extreme values. Analysis for other cases (e.g. censoring from above or below (i.e. on right or left)) follows parallel lines, and will be reported on in the next paper in this series. Topics concerned with censoring in more than two sequences of samples will be studied in the third paper in this series.

REFERENCES

[1] Johnson, N.L. (1966) "Sample censoring", *Proc. 12th Conf. Des. Exp. Army Res. Dev. Testing*, 403-424.

[2] Johnson, N.L. (1970) "A general purpose test of censoring of sample extreme values" (In *Essays in Probability and Statistics (S.N. Roy Memorial Volume)*) Chapel Hill: University of North Carolina Press, pp. 379-384.

[3] Johnson, N.L. (1971) "Comparison of some tests of sample censoring of extreme values", *Austral. J. Statist., 13*, 1-6.

[4] Johnson, N.L. (1973-4) "Robustness of certain tests of censoring of extreme sample values; I, II", *University of North Carolina Mimeo Series Nos. 866,940.*

[5] Johnson, N.L. (1974) "Study of possibly incomplete samples", *Proc. 5th Intern. Conf. Prob., Brasov, Romania*

[6] Schumann, D.E.W. and Bradley, R.A. (1959) "The comparison of the sensitivities of similar experiments: Model II of the analysis of variance", *Biometrics, 15*, 405-416.