EXTREME SAMPLE CENSORING PROBLEMS WITH MULTIVARIATE DATA - II

TECHNICAL REPORT

N.L. JOHNSON

JULY, 1979

U.S. ARMY RESEARCH OFFICE
RESEARCH TRIANGLE PARK, NORTH CAROLINA

DAAG29-77-C-0035

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27514

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Extreme Sample Censoring Problems with Multivariate Data - II		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL 6. PERFORMING ORG. REPORT NUMBER Mimeo Series No. 1237
7. AUTHOR(s) N.L. Johnson		8. CONTRACT OR GRANT NUMBER(s) DAAG29-77-C-0035
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of North Carolina Chapel Hill, North Carolina 27514		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Research Office Research Triangle Park, NC		12. REPORT DATE July 1979 13. NUMBER OF PAGES 21
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for Public Release: Distribution Unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Indirect censoring, Farlie-Gumbel-Morgenstern distribution, Likelihood ratio tests, Order statistics, moments.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Indirect censoring is defined as the effect on observed variables of censoring on unobserved variables. Methods of testing for indirect censoring are discussed, and exemplified, using a bivariate Farlie-Gumbel-Morgenstern distribution.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

EXTREME SAMPLE CENSORING PROBLEMS WITH MULTIVARIATE DATA - II

N.L. Johnson*

University of North Carolina at Chapel Hill

1. Introduction.

Johnson (1978a) has given a survey of various problems which can arise in testing for censoring of extreme values from univariate data. When data are multivariate, there is a much richer variety of possible problems; some possibilities are described in Johnson (1975). The present paper extends these possibilities and indicates lines of attack on certain of the problems. These are worked out in some detail for Farlie-Gumbel-Morgenstern bivariate distributions.

We suppose that observed values on m characters $X_1, X_2, \ldots, X_m$ are available for each of r individuals. We wish to investigate whether these represent a complete random sample, or are the remainder of such a sample (original size n>r) after some form of censoring of extreme values has been applied.

As in Johnson (1978a), we will restrict attention to random sampling from large populations in which the joint distribution of $X_1, \ldots, X_m$ is absolutely continuous, with joint probability density function (PDF)

$$f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = f_{\underset{\sim}{X}}(\underset{\sim}{x}) = f_{12\ldots m}(x_1,\ldots,x_m) \ .$$

We will denote the (unordered) observations on the i-th available individual by

$$\underset{\sim}{X}^* = (X^*_{1i},\ldots,X^*_{mi}) \quad (i=1,\ldots,r) \ .$$

*Research sponsored by the Army Research Office under Grant DAAG29-77-C-0035.

We also use the notation:

(i) $\Pr[\bigcap_{j=1}^{m}(X_{ji}^{*}\leq x_j)] = F_{12\ldots m}(x_1,\ldots,x_m)$ (in particular $\Pr[X_{ji}^{*}\leq x] = F_j(x)$)

(ii) for the conditional PDF of $X_{a_1}^{*},\ldots,X_{a_s}^{*}$ given $X_{b_1}^{*},\ldots,X_{b_t}^{*}$

$g_{a_1\ldots a_s;b_1\ldots b_t}(x_{a_1},\ldots,x_{a_s}|x_{b_1},\ldots,x_{b_t})$ (in particular

$g_{12}(x_1|x_2)$, $g_{21}(x_2|x_1)$)

and

(iii) for the order statistics corresponding to $X_{j1}^{*},\ldots,X_{jr}^{*}$

$$X_{j1} \leq X_{j2} \leq \ldots \leq X_{jr} .$$

We also will focus on the forms of censoring accorded special attention in Johnson (1978a):

(i) from above (exclusion of s_r greatest values) or below (exclusion of s_0 least values), and

(ii) symmetrical -(exclusion of equal number of greatest and least values ($s_0 = s_r$)).

We denote the hypothesis that the s_0 least and s_r greatest values of an original complete random sample of size $n(=r+s_0+s_r)$ have been excluded by H_{s_0,s_r}, so that

(i) corresponds to H_{0,s_r} or $H_{s_0,0}$ $(s_0,s_r > 0)$

(ii) corresponds to $H_{s,s}$ $(s > 0)$.

To indicate that the censoring is applied to the variable X_j we use the symbol $H_{s_0,s_r}^{(j)}$.

2. Problems.

Problems which will be discussed in the present report include:

(a) Censoring is known to be possible on some one of a certain subset of

m' variables. It is required to decide whether such censoring has occurred (Sections 3 and 4). In Section 3 we show that if m' = 1, the univariate techniques already developed can be used on the specified variable; they provide the likelihood ratio test of the hypothesis of no censoring. In Section 4 we discuss the related problem of indirect censoring: detecting whether there has been censoring on X_1 when this variable is not observed directly; only values of $X_2,\ldots,X_m$ are observed. (We take m = 2 for simplicity.) An Appendix sets out the application of the reuslts in a special case. This constitutes the major part of this report. In fact, the text of the report might even be regarded as an introduction to the Appendix.

(b) Censoring is known to have occured on some one of a certain subset of variables. It is required to decide which is the censored variable (Section 5).

(c) We also give an introduction to problems arising in connection with chain censoring. This is censoring in a succession of stages. At the first stage the $(s_0^{(1)}+s_r^{(1)})$ individuals with the $s_0^{(1)}$ least and $s_r^{(1)}$ greatest values of X_1 are removed; then those with the $s_0^{(2)}$ least and $s_r^{(2)}$ greatest values of X_2 among the remainder are removed, and so on. (Section 6 - again with m = 2, for reasons of simplicity.)

3. Censoring Possible Only on One of m' (Observed) Specified Variables.

In the special case of (a), Section 2, with m' = 1 we suppose (without loss of generality) that X_1 is the possibly censored variable.

Since

$$f_{\underset{\sim}{X}^*}(\underset{\sim}{x}) = f_{\underset{\sim}{X}_1^*}(x_1)\, g_{\underset{\sim}{X}_2^*,\ldots,\underset{\sim}{X}_m^*}(\underset{\sim}{x}_2,\ldots,\underset{\sim}{x}_m|\underset{\sim}{x}_1) \qquad (1)$$

and the last term is the same whether $H^{(1)}_{s_0,s_r}$ is valid or not, it follows that

the likelihood ratio

$$\frac{f_{\underset{\sim}{X}^*}(\underset{\sim}{x}|H^{(1)}_{s_0,s_r})}{f_{\underset{\sim}{X}^*}(\underset{\sim}{x}|H^{(1)}_{0,0})} \text{ is equal to } \frac{f_{X_1^*}(x_1|H^{(1)}_{s_0,s_r})}{f_{X_1^*}(x_1|H^{(1)}_{0,0})} . \tag{2}$$

This shows that the likelihood ratio tests appropriate for univariate data, described in Johnson (1978a) (see also Johnson (1966, 1971, 1972)) can be applied to X_1, ignoring the observed values of all the remaining variables $X_2,\ldots,X_m$. (For m = 2, this result is given in Johnson (1978a).)

As in Johnson (1966, 1971, 1972), application of this result requires a knowledge of the population distribution of X_1 (though not of $X_2,\ldots,X_m$). The methods (described in Johnson (1978a,b)) of utilizing partial knowledge of the distribution of X_1 are, of course, also relevant here. In fact, in this case we really do not have a multivariate, but only a univariate problem, so far as detection of censoring is concerned.

4. Censoring Possible Only on One (Unobserved) Variable: Indirect Censoring.

A truly multivariate problem arises if we suppose that values of the possibly censored variable (X_1) are not available. How should the (observed) values of $(X_{2i},\ldots,X_{mi})$ (i=1,...,r) be used to detect if there has been censoring on X_1? For simplicity, again, we consider the bivariate case (m = 2). We first derive a likelihood ratio test. As we shall see, there appear to be considerable technical difficulties in applying this test in many natural situations. Therefore, we also suggest some other procedures which may sometimes be applied more easily.

The available data consist of the r observed values of X_2, denoted by $X^*_{21},\ldots,X^*_{2r}$. Their joint PDF, if $H^{(1)}_{s_0,s_r}$ is valid, is

$$f_{\underset{\sim}{X}_2^*}(\underset{\sim}{x}_2|H^{(1)}_{s_0,s_r})$$

$$= \frac{(r+s_0+s_r)!}{r!s_0!s_r!}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\{F_1(\ell)\}^{s_0}\{1-F_1(u)\}^{s_r}\prod_{i=1}^{r}[f_1(x_{1i})g_{21}(x_{2i}|x_{1i})]dx_{11}\cdots dx_{1r} \tag{3}$$

where $\ell \equiv \min(x_{11},\ldots,x_{1r})$; $u \equiv \max(x_{11},\ldots,x_{1r})$.

Since

$$f_1(x_{1i})g_{21}(x_{2i}|x_{1i}) = f_{12}(x_{1i},x_{2i}) = f_2(x_{2i})g_{12}(x_{1i}|x_{2i})$$

we also have

$$f_{\underset{\sim}{X}_2^*}(\underset{\sim}{x}_2|H^{(1)}_{s_0,s_r})$$

$$= \frac{(r+s_0+s_r)!}{r!s_0!s_r!}\{\prod_{i=1}^{r}f_2(x_{2i})\}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\{F_1(\ell)\}^{s_0}\{1-F_1(u)\}^{s_r}\prod_{i=1}^{r}g_{12}(x_{1i}|x_{2i})dx_{11}\cdots dx_{1r}. \tag{3'}$$

In particular

$$f_{\underset{\sim}{X}_2^*}(\underset{\sim}{x}_2|H^{(1)}_{0,0}) = \prod_{i=1}^{r}f_2(x_{2i}) .$$

It follows that the likelihood ratio is

$$L = \frac{f_{\underset{\sim}{X}_2^*}(\underset{\sim}{X}_2^*|H^{(1)}_{s_0,s_r})}{f_{\underset{\sim}{X}_2^*}(\underset{\sim}{X}_2^*|H^{(1)}_{0.0})} = \frac{(r+s_0+s_r)!}{r!s_0!s_r!}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\{F_1(\ell)\}^{s_0}\{1-F_1(u)\}^{s_r}$$

$$\times\prod_{i=1}^{r}g_{12}(x_{1i}|x_{2i})dx_{11}\cdots dx_{1r}$$

$$= \frac{(r+s_0+s_r)!}{r!s_0!s_r!}E[\{F_1(X_{11})\}^{s_0}\{1-F_1(X_{1r})\}^{s_r}|\underset{\sim}{X}_2^*] \tag{4}$$

(remembering that $X_{11} = \min(X_{11}^*,\ldots,X_{1r}^*)$; $X_{1r} = \max(X_{11}^*,\ldots,X_{1r}^*)$).

Calculation of L from the observed values $\underset{\sim}{X}_2^*$ is usually quite difficult. When this is done, determination of the distribution of L (even when the null hypothesis, $H^{(1)}_{0,0}$ is valid) is likely to be even more difficult. In the

Appendix we use a Farlie-Gumbel-Morgenstern (see e.g. Johnson and Kotz (1975)) joint distribution for illustrative purposes. Calculation of L is not very difficult in this special case, but even here, the distribution of L is not easily derived. The conditional joint distribution of X_{11} and X_{1r} can be derived from

$$\Pr[\ell \le X_{11} \le X_{1r} \le u | \underset{\sim}{X}_2^*] = \prod_{i=1}^{r} [\int_{\ell}^{u} g_{12}(x_1 | x_{2i}^*) dx_1] \tag{5}$$

but this expression is usually quite complicated.

We note that the value of L (and so its distribution) is unchanged by any monotonic increasing transformations of X_1^* and X_2^*. This means that we can take, without loss of generality, each of the variables to have a standard uniform distribution ($f_i(x) = 1$ for $0 \le x \le 1$, $i = 1,2$). However, the joint PDF would then have to be that resulting from application of the appropriate transformations to the original joint PDF. The Appendix contains some analyses appropriate to a bivariate Farlie-Gumbel-Morgenstern distribution (e.g. Johnson and Kotz (1975)) which does have standard uniform marginal distributions.

A simpler criterion, suggested by the above analysis is

$$L_1 = \{F_1(\min_i E[X_1 | X_{2i}^*])\}^{s_0} \{1-F_1(\max_i E[X_1 | X_{2i}^*])\}^{s_r} . \tag{6}$$

If $E[X_1 | X_2]$ is a monotonic increasing function of X_2 then

$$L_1 = \{F_1(E[X_1 | \min(X_{21}^*, \ldots, X_{2r}^*)])\}^{s_0} \{1-F_1(E[X_1 | \max(X_{21}^*, \ldots, X_{2r}^*)])\}^{s_r} . \tag{7}$$

If $E[X_1 | X_2]$ is a monotonic decreasing function of X_2, then "min" and "max" in (7) are interchanged.

Another related criterion, generally more difficult to compute is

$$L_1' = E[\{F_1(X_{1(\ell)})\}^{s_0}\{1-F_1(X_{1(u)})\}^{s_r}] \tag{8}$$

where $X^*_{1(\ell)}$, $X^*_{1(u)}$ are independent with PDF's $g_{12}(x_1|X^*_{2(\ell)})$, $g_{12}(x_1|X^*_{2(u)})$ respectively and $i = (\ell),(u)$ resepctively minimize and maximize $E[X_1|X^*_{2i}]$ with respect to i.

If $E[X_1|X_2]$ is a monotonic increasing (decreasing) function of X_2, then

$$X^*_{2(\ell)} = \min(\max)\ (X^*_{21},\ldots,X^*_{2r})$$

$$X^*_{2(u)} = \max(\min)\ (X^*_{21},\ldots,X^*_{2r})\ .$$

5. Identification of Censored Variable.

Suppose we know that some one of the m' variables $X_1,X_2,\ldots,X_{m'}$ has been censored (and that there has been no other form of censoring). We wish to decide which one of these variables is the one in respect to which censoring has occurred.

Using the argument in Section 3, the likelihood approach is straightforward, if it can be assumed that (s_0,s_r) censoring has been used with s_0,s_r known. We take each of the m' variables in order and calculate the appropriate (univariate) likelihood ratio criterion for that variable. We choose that variable for which the likelihood ratio is greatest. This means that we choose X_h if

$$\{F_h(X_{h1})\}^{s_0}\{1-F_h(X_{hr})\}^{s_r} = \max_{j=1,\ldots,m'}[\{F_j(X_{j1})\}^{s_0}\{1-F_j(X_{jr})\}^{s_r}]$$

(with some arbitrary rule for deciding ties - which, anyway, have zero probability of occurrence).

We note that the same decision will be reached for all pairs of values s_0, s_r for which $s_0/s_r = \theta$ has the same value. In particular, the same decision will be reached for (i) censoring from above ($\theta=0$) with any s_r

(ii) " " below ($\theta=\infty$) " " s_0

(iii) symmetrical censoring($\theta=1$) " " $s_0 = s_r$.

(Of course, the decisions for (i), (ii) and (iii) are not, in general, identical, though they might be.)

By analogy with the results in Johnson (1971), when the values of s_0, s_r are unknown, we would choose X_h if

$$F_h(X_{h1}) + 1-F_h(X_{hr}) = \max_{j=1,\ldots,m'} [F_j(X_{j1}) + 1-F_j(X_{jr})] .$$

Hasofer and Davis (1979) have considered a similar type of problem where truncation, rather than censoring is applied to one of a number of variables, and it is desired to identify which of the variables is being truncated.

6. Chain Censoring.

A further type of censoring of an essentially multivariate nature occurs when two or more variables, in sequence, are used for censoring. For example, from a sample of size n, with m variables $X_1, X_2, \ldots, X_m$ measured on each individual (i) the $s_0^{(1)}$ individuals with the least, and $s_r^{(1)}$ with the greatest values of X_1 are removed and then (ii) from the remainder the $s_0^{(2)}$ with the least, and $s_r^{(2)}$ with the greatest values of X_2 are removed, leaving a set of

$$r = n-s_0^{(1)} - s_0^{(2)} - s_r^{(1)} - s_r^{(2)}$$

values.

Problems arising from this type of situation include:

1) Given that there has been chain censoring involving two specified variables X_i and X_j, which variable has been used for the first censoring operation?

2) Given that X_i is the first and X_j the second (if any), is there evidence that second stage censoring has in fact been applied?

3) Which variables have been used in censoring? (Given that just two (or three or more) are used and possibly given an order of precedence such that for any two specified variables it is known which would be used before the other, if both were used.)

APPENDIX: Detection of Indirect Censoring in Farlie-Gumbel-Morgenstern (FGM) Distributions.

A1. Relevant Properties of FGM Distributions

Here we will illustrate calculations and use of the criteria introduced in Section 4, using FGM population distributions with joint distribution

$$\Pr[(X_1^* \le x_1) \cap (X_2^* \le x)] = x_1 x_2\{1+\theta(1-x_1)(1-x_2)\} \quad (0 \le x_j \le 1;\ j=1,2;\ |\theta|<1). \tag{A1}$$

Each X_j^* has a marginal standard uniform distribution ($F_j(x_j) = x_j$ $(0 \le x_j \le 1)$). This distribution has been chosen for analytical convenience. It is not claimed that the results will apply for other joint distributions, even after transformation to make the marginals to be standard uniform. However, there are some speculative analogies which might be drawn.

From (A1) it follows that the joint PDF is

$$f(x_1,x_2) = 1 + \theta(1-2x_1)(1-2x_2) \quad (0 \le x_j \le 1;\ j=1,2) \tag{A2}$$

and the conditional PDF's are

$$\left.\begin{aligned} g_{12}(x_1|x_2) &= 1 + \theta(1-2x_2)\cdot(1-2x_1) \quad (0 \le x_1 \le 1) \\ g_{21}(x_2|x_1) &= 1 + \theta(1-2x_1)\cdot(1-2x_2) \quad (0 \le x_2 \le 1) \end{aligned}\right\}. \tag{A3}$$

Hence

$$\Pr[\ell \le X_1^* \le u | x_2] = (u-\ell)[1+\theta(1-2x_2)(1-u-\ell)] \quad (0 \le \ell \le u \le 1)$$

and so

$$\Pr[\ell \le X_{11} \le X_{1r} \le u | \underset{\sim}{x}_2] = (u-\ell)^r \prod_{j=1}^{r} [1+\theta(1-2x_{2j})(1-u-\ell)] .$$

A2. Derivation of L.

The conditional joint PDF of X_{11} and X_{1r} is therefore

$$\frac{-\partial^2 \Pr[\ell \le X_{11} \le X_{1r} < u \mid \underset{\sim}{X}(2)]}{\partial \ell \partial u} = r(r-1)(u-\ell)^{r-2} \prod_{j=1}^{r} \{1+\theta z_{2j}(1-u-\ell)\}$$

$$-2\alpha^2 (u-\ell)^r \sum_{j<j'} \sum z_{2j} z_{2j'} \prod_{h \neq j,j'} \{1+\theta z_{2h}(1-u-\ell)\}$$

(A5)

where $z_{2j} = 1-2x_{2j}$.

From (4),

$$L = \frac{(r+s_0+s_r)!}{r!s_0!s_r!} E[X_{11}^{s_0}(1-X_{1r})^{s_r} \mid \underset{\sim}{X}^*(2)]$$

$$= \frac{(r+s_0+s_r)!}{r!s_0!s_r!} \iint_{0<\ell<u<1} \ell^{s_0}(1-u)^{s_r} \left[\frac{-\partial^2 \Pr[\ell \le X_{11} \le X_{1r} \le u \mid \underset{\sim}{X}^*(2)]}{\partial \ell \partial u}\right] d\ell du$$

$$= \frac{(r+s_0+s_r)!}{r!s_0!s_r!} \{r(r-1) \sum_{h=0}^{r} J(r-2,s_0,s_r;h)\theta^h Y_h - 2\alpha^2 \sum_{h=0}^{r} J(r,s_0,s_r;h)\binom{h+2}{2}\theta^h Y_{h+2}\}$$

(A5)

where $Y_0 = 1$; $Y_h = \sum_{j_1<\ldots<j_h} \cdots \sum \prod_{i=1}^{h} Z^*_{2j_i}$; $Z^*_{2j} = 1-2X^*_{2j}$ $(h,j=1,\ldots,r)$ and (with ε a positive integer)

$$J(\beta,\gamma,\delta;\varepsilon) = \int_0^1 \int_0^u (u-\ell)^{\beta} \ell^{\gamma} (1-u)^{\delta} (1-u-\ell)^{\varepsilon} d\ell du$$

$$= \sum_{i=0}^{\varepsilon} (-1)^i \binom{\varepsilon}{i} \int_0^1 (1-u)^{\delta+\varepsilon-i} \int_0^u \ell^{\gamma+i} (u-\ell)^{\beta} d\ell dg$$

$$= \sum_{i=0}^{\varepsilon} (-1)^i \binom{\varepsilon}{i} B(\gamma+i+1,\beta+1) B(\beta+\gamma+i+2,\delta+\varepsilon-i+1) .$$

If, also, β, γ and δ are positive integers

$$J(\beta,\gamma,\delta;\varepsilon) = \sum_{i=0}^{\varepsilon} (-1)^i \binom{\varepsilon}{i} \frac{\beta!(\gamma+i)!(\delta+\varepsilon-i)!}{(\beta+\gamma+\delta+\varepsilon+2)!} = \frac{\beta!\gamma!\delta!}{(\beta+\gamma+\delta+\varepsilon+2)!} G(\gamma,\delta;\varepsilon) \qquad (A6)$$

where

$$G(\gamma,\delta;\varepsilon) = \sum_{i=0}^{\varepsilon} (-1)^{\varepsilon} \binom{\varepsilon}{i} (\gamma+1)^{[i]} (\delta+1)^{[\varepsilon-i]} \qquad (A7)$$

and $a^{[b]} = a(a+1)\ldots(a+b-1)$ is the b-th ascending factorial of a.

Formula (A5) can be written

$$L = \sum_{h=0}^{r} K(r,s_0,s_r;h)\theta^h Y_h \tag{A8}$$

with

$$K(r,s_o,s_r;h) = \{G(s_0,s_r;h) - h(h-1)G(s_0,s_r;h-2)\}/(r+s_0+s_r+1)^{[h]}. \tag{A9}$$

(If $\varepsilon < 0$, $G(s_0,s_r;\varepsilon)$ can be defined arbitrarily.)

We note that

$$K(r,s_0,s_r;0) = G(s_0,s_r;0) = 1$$

so the first term on the right hand side of (A8) is 1.

We also note that

$$G(s,0;h) = h!\psi_h(s+1) = (-1)^h G(0,s;h) \tag{A10}$$

where

$$\psi_h(y) = 1 - \frac{y}{1!} + \frac{y^{[2]}}{2!} - \ldots + (-1)^h \frac{y^{[h]}}{h!}. \tag{A11}$$

So, for censoring from below ($s_r = 0$)

$$(r+s_0+1)^{[h]}K(r,s_0,0;h) = h!\{\psi_h(s_0+1) - \psi_{h-2}(s_0+1)\} = (-1)^h s_0^{[h]} \tag{A12}$$

and for censoring from above ($s_0 = 0$)

$$(r+s_r+1)^{[h]}K(r,0,s_r;h) = s_r^{[h]}. \tag{A13}$$

For symmetrical censoring ($s_0=s_r=s$) we have

$$G(s,s;h) = \begin{cases} 0 & \text{if } h \text{ is odd} \\ (k+1)^{[k]}(s+1)^{[k]} & \text{if } h = 2k. \end{cases} \tag{A14}$$

Hence, for symmetrical censoring

$$(r+2s+1)^{[h]}K(r,s,s;h) = \begin{cases} 0 & \text{if } h \text{ is odd} \\ (k+1)^{[k]}s^{[k]} & \text{if } h = 2k. \end{cases} \tag{A15}$$

Summarizing, we have the following expressions for the likelihood ratios:-

For detecting censoring from below: $$L = 1 + \sum_{h=1}^{r} (-1)^r \frac{s_0^{[h]}}{(r+s_0+1)^{[h]}} \theta^h Y_h \qquad (A16)$$

" " " " above: $$L = 1 + \sum_{h=1}^{r} \frac{s_r^{[h]}}{(r+s_r+1)^{[h]}} \theta^h Y_h \qquad (A17)$$

" " symmetrical censoring: $$L = 1 + \sum_{k \le r/2} \frac{(k+1)^{[k]} s^{[k]}}{(r+2s+1)^{[2k]}} \theta^{2k} Y_{2k} \qquad (A18)$$

In each case large values of the statistic are to be regarded as significant of censoring of the relevant type. Some numerical values for calculating the coefficients of $\theta^h Y_h$ in (A16)-(A18) are shown in Table 1.

A3. Moments of L.

From the general theory of testing hypotheses, we have

$$E[L|H_{0,0}^{(1)}] = 1 \qquad (5)'$$

Under $H_{0,0}^{(1)}$, the Z_2^*'s are mutually independent and each is distributed uniformly over the interval (-1,1) so, for all θ

$$E[(Z_2^*)^q|H_{0,0}^{(1)}] = \begin{cases} 0 & \text{if } q \text{ is odd} \\ (q+1)^{-1} & \text{if } q \text{ is even} . \end{cases} \qquad (A19)$$

It follows that for any h, h' ($\neq$ h)

$$\left.\begin{aligned} &E[Y_h|H_{0,0}^{(1)}] = 0 = E[Y_h Y_{h'}|H_{0,0}^{(1)}] \\ \text{and} \\ &\text{var}(Y_h|H_{0,0}^{(1)}) = E[Y_h^2|H_{0,0}^{(1)}] = \binom{r}{h}\left(\tfrac{1}{3}\right)^h \end{aligned}\right\} \qquad (A20)$$

Hence when using the statistics (A16), (A17) testing for censoring from below or above ($s_0 = s$, $s_r = 0$ or $s_0 = 0$, $s_r = s$)

$$\mathrm{var}(L|H_{0,0}^{(1)}) = \sum_{h=1}^{r} \left\{\frac{s^{[h]}}{(r+s+1)^{[h]}}\right\}^2 \binom{r}{h} \left(\tfrac{1}{3}\theta^2\right)^h \tag{A21}$$

while when testing for symmetrical censoring ($s_0=s_r=s$)

$$\mathrm{var}(L|H_{0,0}^{(1)}) = \sum_{k\le r/2} \left\{\frac{(k+1)^{[k]}s^{[k]}}{(r+2s+1)^{[2k]}}\right\}^2 \binom{r}{2k} \left(\tfrac{1}{3}\theta^2\right)^{2k} . \tag{A22}$$

Approximate significance limits for L may be obtained by supposing the distribution under $H_{0,0}^{(1)}$ is approximately normal.

A4. Alternative Tests.

Since

$$E[Z_1^*|Z_2^*] = E[1-2X_1^*|Z_2^*] = \int_0^1 (1-2x_1)\{1 + \theta Z_2^*(1-2x_1)\}dx_1$$

$$= \tfrac{1}{3}\theta Z_2^* \tag{A23}$$

it follows that

$$E[X_1^*|X_2^*] = \tfrac{1}{2}[1 - \tfrac{1}{3}\theta(1-2X_2^*)]$$

$$= \tfrac{1}{2} - \tfrac{1}{6}\theta(1-2X_2^*) \tag{A24}$$

Hence, the simplified test statistic, L_1, defined in (6) is, for our FGM distribution and with $\theta > 0$

$$L_1 = \left\{\tfrac{1}{2} + \tfrac{1}{6}\theta(2X_{21}-1)\right\}^{s_0}\left\{\tfrac{1}{2} - \tfrac{1}{6}(2X_{2r}-1)\right\}^{s_r} . \tag{A25}$$

We note that if s_0 (s_r) = 0 (and $\theta > 0$), the critical region becomes simply $X_{2r}<(X_{21}>)K$, with an appropriate value for the constant K. This would, of course, be the appropriate likelihood ratio test of the null hypothesis, with the alternative that X_2 itself has been subjected to censoring from above (below).

In the case of symmetric censoring ($s_0=s_r=s$), the critical region (for all $s > 0$) is of form

$$\{\tfrac{1}{2} + \tfrac{1}{6}\theta(2X_{21}-1)\}\{\tfrac{1}{2} - \tfrac{1}{6}\theta(2X_{2r}-1)\} > K \qquad \text{(A26)}$$

or equivalently

$$\tfrac{1}{9}\theta^2 X_{21}(1-X_{2r}) + \tfrac{1}{3}\theta(\tfrac{1}{2} - \tfrac{1}{6}\theta)(X_{21} + \overline{1-X_{2r}}) > K'. \qquad \text{(A26)'}$$

This can be compared with the critical regions

$X_{21}(1-X_{2r})$ (for symmetrical censoring)

$X_{21} + (1-X_{2r})$ (for general censoring - see Johnson (1971))

for likelihood ratio tests of $H_{0,0}^{(2)}$.

The values in Table 1 suggest that useful tests might be constructed by taking as test statistics the first terms only in the summations in (A16)-(A18). This would lead to critical regions (which do not depend on θ).

For censoring from below: $Y_1 < C$ (A27)

" " " above: $Y_1 > C$ (A28)

" symmetrical censoring: $Y_2 > C$. (A29)

Since $Y_1 = \sum_{j=1}^{r} Z_{2j}^* = \sum_{j=1}^{r}(1-2X_{2j}^*)$, (A27) and (A28) are equivalent to

$$\sum_{j=1}^{r} X_{2j}^* > C' \qquad \text{(A27)'}$$

$$\sum_{j=1}^{r} X_{2j}^* < C' \qquad \text{(A28)'}$$

respectively. (The signs of the inequalities would be reversed if $\theta < 0$.)

On the null hypothesis $H_{0,0}^{(1)}$ (no censoring) the X_{2j}^*'s are mutually

independent standard uniform variables. Therefore, even for r as small as 5, the distribution of their sum is closely approximated by a normal distribution with expected value $\frac{1}{2}r$ and variance $\frac{1}{12}r$ (e.g. Johnson and Kotz (1972, p. 64)). So we obtain an approximate significance level α by taking

$$C' = \tfrac{1}{2}r + \lambda_\alpha\sqrt{\tfrac{r}{12}} \quad \text{in (A27)'}$$

$$C' = \tfrac{1}{2}r - \lambda_\alpha\sqrt{\tfrac{r}{12}} \quad \text{in (A28)'}$$

where $\Phi(\lambda_\alpha) = 1 - \alpha$.

From (A20)

$$E[Y_2|H^{(1)}_{0,0}] = 0;\ \mathrm{var}(Y_2|H^{(1)}_{0,0}) = \tfrac{1}{18}r(r-1). \tag{A30}$$

Assuming Y_2 has an approximately normal distribution under $H^{(1)}_{0,0}$, we obtain an approximate significance level α for the test for symmetrical censoring (A29) by taking

$$C = \lambda_\alpha\sqrt{\frac{r(r-1)}{18}}\ .$$

The moments of the Y_h's under $H^{(1)}_{s_0',s_r'}$ may be evaluated by the following steps:-

(i) find the conditional expected value, given $\underset{\sim}{X}^*_1$, and

(ii) find the expected value of (i) when the joint distribution of the $\underset{\sim}{X}^*_1$'s is that of the $(s_0'+1)$-th, $(s_0'+2)$-th,..., $(s_0'+r)$-th order statistics among $(r+s_0'+s_r')$ variables, each with PDF $f_1(x)$.

For (i) we use the result:

$$E[(Z^*_2)^q|X^*_1] = \int_0^1 (1-2x)^q\{1+\theta Z^*_1(1-2x)\}dx$$

$$= \begin{cases} \theta(q+2)^{-1}Z^*_1 & \text{if } q \text{ is odd} \\ (q+1)^{-1} & \text{if } q \text{ is even} \end{cases} \tag{A31}$$

(cf (A23))

where $Z^*_1 = 1 - 2X^*_1$.

For (ii) we use the result (for ordered variables)

$$E[\prod_{i=1}^{p} X_{1h_i}^{q_i} | H^{(1)}_{s_0',s_r'}] = \frac{1}{(r+s_0'+s_r'+1)^{[\sum_1^p q_i]}} \prod_{i=1}^{p} (s_0'+h_i+\sum_{j=1}^{i} q_j - q_i)^{[q_i]} . \quad (A32)$$

In fact, in view of (A31), and since, conditionally on $\underset{\sim}{X}_1^*$, the Z_2^*'s are mutually independent, we need only the special case $q_i = 1$ for all i,

$$E[\prod_{i=1}^{p} X_{1h_i} | H^{(1)}_{s_0',s_r'}] = \{\prod_{i=1}^{p} (s_0'+h_i+i-1)\}/(r+s_0'+s_r'+1)^{[p]} . \quad (A33)$$

We find

$$E[Y_1 | H^{(1)}_{s_0',s_r'}] = \frac{1}{3}\theta \sum_{j=1}^{r} E[1-2X_{1j}^* | H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{3}\theta \sum_{h=1}^{r} E[1-2X_{1h} | H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{3}\theta[r - 2\sum_{h=1}^{r} \frac{s_0'+h}{r+s_0'+s_r'+1}]$$

$$= \frac{r(s_r'-s_0')}{3(r+s_0'+s_r'+1)}\theta . \quad (A34)$$

In particular

$$E[Y_1 | H^{(1)}_{s,0}] = -\frac{rs}{3(r+s+1)}\theta = -E[Y_1 | H^{(1)}_{0,s}] . \quad (A35)$$

Next,

$$Y_2 = \sum_{j<j'} \sum Z_{2j}^* Z_{2j}^*, \quad \text{and} \quad E[Z_{2j}^* Z_{2j}^*, | X_{1j}^*, X_{1j}^*,] = \frac{1}{9}\theta^2 Z_{ij}^* Z_{ij}^* .$$

So

$$E[Y_2 | H^{(1)}_{s_0',s_r'}] = \frac{1}{9}\theta^2 \sum_{j<j'} \sum E[1-2(X_{1j}^*+X_{1j}^*,) + 4X_{1j}^* X_{1j}^*, | H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{9}\theta^2[\binom{r}{2} - 2\sum_{h<h'} \sum E[X_{1h}+X_{1h'} | H^{(1)}_{s_0',s_r'}] + 4\sum_{h<h'} \sum E[X_{1h}X_{1h'} | H^{(1)}_{s_0,s_r}]]$$

$$= \frac{1}{9}\theta^2\Big[\binom{r}{2} - \frac{2}{r+s_0'+s_r'+1}\sum_{h<h'}\sum(2s_0'+h+h') + \frac{4}{(r+s_0'+s_r'+1)^{[2]}}$$

$$\times \sum_{h<h'}\sum(s_0'+h)(s_0'+h'+1)\Big] .$$

Performing the summations (see, e.g. David and Johnson (1954, pp. 239-40)) we obtain

$$E[Y_2|H^{(1)}_{s_0',s_r'}] = \frac{r(r-1)\{(s_0'-s_r')^2+s_0'+s_r'\}}{18(r+s_0'+s_r'+1)^{[2]}}\theta^2. \tag{A36}$$

We can use (A30) together with

$$E[Z^{*2}_{2j}|X^*_j] = \tfrac{1}{3} \quad (\text{for all } X^*_{1j}) \tag{A37}$$

to calculate the variance of Y_1 $(= \sum_{j=1}^{r} Z^*_{2j})$ under $H^{(1)}_{s_0',s_r'}$. We have

$$E[Y_1^2|H^{(1)}_{s_0',s_r'}] = \sum_{j=1}^{r} E[Z^{*2}_{2j}|H^{(1)}_{s_0',s_r'}] + 2E[Y_2|H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{3}r + \frac{r(r-1)\{(s_0'-s_r')^2+s_0'+s_r'\}}{9(r+s_0'+s_r'+1)^{[2]}}\theta^2$$

and so, using (A28)

$$\mathrm{var}(Y_1|H^{(1)}_{s_0',s_r'}) = \frac{1}{3}r + \Big[\frac{r(r-1)\{(s_0'-s_r')^2+s_0'+s_r'\}}{9(r+s_0'+s_r'+1)^{[2]}} - \frac{r^2(s_0'-s_r')^2}{9(r+s_0'+s_r'+1)^2}\Big]\theta^2$$

$$= \frac{1}{3}r + \frac{r\{(r-1)(r+s_0'+s_r'+1)(s_0'+s_r') - (2r+s_0'+s_r'+1)(s_0'-s_r')^2\}}{9(r+s_0'+s_r'+1)^2(r+s_0'+s_r'+2)}\theta^2 . \tag{A38}$$

For censoring from below ($s_0'=s$, $s_r'=0$) or above ($s_0'=0$, $s_r=s$) this gives

$$\mathrm{var}(Y_1|H^{(1)}_{s,0}) = \mathrm{var}(Y_1|H^{(1)}_{0,s}) = \frac{1}{3}r + \frac{rs\{(r-1)(r+s+1)-(2r+s+1)s\}}{9(r+s+1)^2(r+s+2)}\theta^2$$

$$= \frac{1}{3}r + \frac{rs\{r(r-s) - (s+1)^2\}}{9(r+s+1)^2(r+s+2)}\theta^2 . \tag{A39}$$

Note that as $s \to \infty$ (with r fixed)

$$E[Y_1|H^{(1)}_{s,0}] \to -\frac{1}{3} r\theta; \quad \text{var}(Y_1|H^{(1)}_{s,0}) \to \frac{1}{3} r(1-\frac{1}{3}\theta^2) .$$

Evaluation of variance of Y_2, expected values of Y_h ($h > 2$) and higher moments and product-moments of all Y's, under $H^{(1)}_{s_0',s_r'}$ requires evaluation of quantities

$$(T_p(r,s_0',s_r')=)T_p = \sum_{j_1<\ldots<j_p}^{r-p+1 \quad r} \cdots \sum E[\prod_{i=1}^{p} Z^*_{1j_i}|H^{(1)}_{s_0',s_r'}]$$

$$= \sum_{h_1<\ldots<h_p}^{r-p+1 \quad r} \cdots \sum E[\prod_{i=1}^{p}(1-2X_{1h_i})|H^{(1)}_{s_0',s_r'}]$$

$$= \sum_{h_1<\ldots<h_p}^{r-p+1 \quad r} \cdots \sum \sum_{u=0}^{p} (-2)^u \sum_{i_1<\ldots<i_u}^{p-u+1} \cdots \sum E[\prod_{\alpha=1}^{u} X_{1h_{i_\alpha}}|H^{(1)}_{s_0',s_r'}]$$

$$= \sum_{u=0}^{p} (-2)^u \{(r+s_0'+s_r'+1)^{[u]}\}^{-1} \sum_{h_1<\ldots<h_p}^{r-p+1} \cdots \sum$$

$$\times \sum_{i_1<\ldots<i_u}^{p-u+1} \cdots \sum \prod_{\alpha=1}^{p\ u} (s_0'+h_{i_\alpha}+\alpha-1) . \qquad \text{(A40)}$$

(The term corresponding to u = 0 is 1. The final multiple sum of products can be expressed as a polynomial in s_0' and r with coefficients calculated from tables like those in David and Johnson (1954, pp. 239-240).) For example

$$E[Y_1Y_2|H^{(1)}_{s_0',s_r'}]$$

$$= E[(\sum_{j=1}^{r} Z^*_{2j})(\sum_{j}^{r-1}\sum_{j'}^{r} Z^*_{2j}Z^*_{2j'})|H^{(1)}_{s_0',s_r'}]$$

$$= E[\sum_{j<j'}^{r-1\ r} (Z^*_{2j}+Z^*_{2j'})Z^*_{2j}Z^*_{2j'} + 3\sum_{j<j'<j''}^{r-2\ r-1\ r} Z^*_{2j}Z^*_{2j'}Z^*_{2j''}|H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{9}\theta \sum_{j<j'}^{r-1\ r} E[Z^*_{1j}+Z^*_{1j'}|H^{(1)}_{s_0',s_r'}] + \frac{1}{9}\theta^3 \sum_{j<j'<j''}^{r-2\ r-1\ r} E[Z^*_{1j}Z^*_{1j'}Z^*_{1j''}|H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{9}\theta(r-1)T_1 + \frac{1}{9}\theta^3 T_3 \qquad \text{(A41)}$$

and

$$E[Y_2^2|H^{(1)}_{s_0',s_r'}]$$

$$= \sum_{j<j'}^{r-1}\sum^{r} E[Z^{*2}_{2j}Z^{*2}_{2j'}|H^{(1)}_{s_0',s_r'}] + 2\sum_{j<j'<j''}^{r-2}\sum^{r-1}\sum^{r} E[(Z^{*}_{2j}+Z^{*}_{2j'})Z^{*}_{2j'}Z^{*}_{2j''}|H^{(1)}_{s_0',s_r'}]$$

$$+ 6\sum_{j<j'<j''<j'''}^{r-3}\sum^{r-2}\sum^{r-1}\sum^{r} E[Z^{*}_{2j}Z^{*}_{2j'}Z^{*}_{2j''}|H^{(1)}_{s_0',s_r'}]$$

$$= \frac{1}{9}\sum_{j<j'}^{r-1}\sum^{r} 1 + \frac{2\theta^2}{27}\sum_{j<j'<j''}^{r-2}\sum^{r-1}\sum^{r} E[Z^{*}_{1j}Z^{*}_{1j''}+Z^{*}_{1j}Z^{*}_{1j''}+Z^{*}_{1j}Z^{*}_{1j'}|H^{(1)}_{s_0',s_r'}]$$

$$+ \frac{6\theta^4}{81}\sum_{j<j'<j''<j'''}^{r-3}\sum^{r-2}\sum^{r-1}\sum^{r} E[Z^{*}_{1j}Z^{*}_{1j'}Z^{*}_{1j''}Z^{*}_{1j'''}|H^{(1)}_{s_0',s_r'}]$$

$$= \frac{r(r-1)}{18} + \frac{2\theta^2}{27}(r-2)T_2 + \frac{2\theta^4}{27}T_4 \,. \qquad \text{(A42)}$$

TABLE I. Values of $(r+s_0+s_r+1)^{[h]}K(r,s_0s_r;h)$

s_0	s_r	h = 1	2	3	4	5	6
0	1	1	2	6	24	120	720
0	2	2	6	24	120	720	5040
1	1	0	2	0	24	0	720
0	3	3	12	60	360	2520	20160
1	2	1	4	12	72	360	2880
0	4	4	20	120	840	6720	60480
1	3	2	8	36	216	1440	11520
2	2	0	4	0	72	0	2880

Note: (i) To obtain the coefficient of $\theta^h Y_h$ in the formula for L (see (16)) these numbers must be divided by $(r+s_0+s_r+1)^{[h]}$. Thus for $r = 5$, $s_0 = 1$, $s_5 = 2$ we have

$$L = 1 + \frac{1}{9}\theta Y_1 + \frac{4}{90}\theta^2 Y_2 + \frac{12}{990}\theta^3 Y_3 + \frac{72}{1180}\theta^4 Y_4 + \frac{360}{154440}\theta^5 Y_5$$

$$= 1 + 0.111\theta Y_1 + 0.0444\theta^2 Y_2 + 0.0121\theta^3 Y_3 + 0.00606\theta^4 Y_4 + 0.00233\theta^5 Y_5 .$$

(ii) Values of s_0 and s_r can be interchanged by multiplying entries by $(-1)^h$.

(iii) A convenient formula for computation is (with $s_r > s_0$)

$$G(s_0,s_r;h) = h! \sum_{i\le h/2} \frac{(s_0+1)^{[i]}}{i!} \cdot \frac{(s_r-s_0)^{[h-2i]}}{(h-2i)!} .$$

In particular, $G(s,s+1;h) = h! \sum_{i\le h/2} \frac{(s+1)^{[i]}}{i!}$, whence

$$(r+2s+2)^{[h]}K(r,s,s+1;h) = h!\frac{(s+1)^{[k]}}{k!} = (k+1)^{[h-k]}(s+1)^{[k]}$$

with $$k = \begin{cases} \frac{1}{2}(h-1) & \text{if } h \text{ is odd} \\ \frac{1}{2}h & \text{if } h \text{ is even} . \end{cases}$$

REFERENCES

David, F.N. and Johnson, N.L. (1954). Statistical treatment of censored data. Part I: Fundamental formulae, Biometrika, 41, 228=240.

Hasofer, A.M. and Davis, R.B. (1970). On the identification of the selection variable in Pearson's selection model, Austral. J. Statist., 21, 71-77.

Johnson, N.L. (1966) Sample censoring, Proc. 12th Conf. Des. Exp., Army Res. Des. Testing, 403-424.

_____ (1971). Comparison of some tests of sample censoring of extreme values, Austral. J. Statist., 13, 1-6.

_____ (1972). Inferences on sample size: Sequences of samples, Trab. Estad., 23, 85-110.

_____ (1975). Extreme sample censoring problems with multivariate data - I, Institute of Statistics, University of North Carolina, Mimeo Series No. 1010.

_____ (1978a). Tests for censoring of extreme values (especially when population distributions are incompletely defined, Proc. Army Res. Office Workshop on Robustness.

_____ (1978b). Completeness comparisons among sequences of samples, Contributions to Survey Sampling and Applied Statistics (H.O. Hartley 65th Birthday Volume) (pp. 259-275). New York: Academic Press.

Johnson, N.L. and Kotz, S. (1972). Distributions in Statistics. Continuous Univariate Distributions - 2, New York: Wiley.

_____ (1975). On some generalized Farlie-Gumbel-Morgenstern distributions, Comm. Statist., 4, 415-427.