A GENERAL PURPOSE TEST FOR INDIRECT CENSORING

by

Norman L. Johnson

University of North Carolina

ABSTRACT

A general purpose test for censoring of extreme values of an unobserved variable, in unknown proportions, is suggested. The test uses observed values of a variable correlated with the censored variable, and the joint distribution of the two variables has to be known for application of the test.

Key Words & Phrases: Multivariate censoring; Farlie-Gumbel-Morgenstern distribution; order statistics; U-statistics

Norman L. Johnson's work was supported by the National Science Foundation under Grant MCS-8-21704.

1. INTRODUCTION

In [1], I discussed testing whether data consisting of r sample values $\underset{\sim}{X}_1=(X_{11},\ldots,X_{1r})$ on a variable X_1 represent a complete random sample, or the remainder of a random sample of size $n = r+s_o + s_r$ from which the individuals with the s_o least and s_r greatest values of *another* variable X_2 have been removed, so that the X_1 values have been 'indirectly censored'. It was assumed that the joint distribution of X_1 and X_2 is absolutely continuous, and that their joint distribution is known.

Using a likelihood ratio approach, the criterion appropriate for testing the hypothesis $H_{o,o}$ (that there has been no censoring) against the alternative H_{s_o,s_r} (s_o least and s_r greatest X_2 values censored) was found to be

$$E[\{F_2(\min \underset{\sim}{X}_2)\}^{s_o}\{1-F_2(\max \underset{\sim}{X}_2)\}^{s_r}|\underset{\sim}{X}_1]$$

when $F_2(\cdot)$ is the cumulative distribution function of each X_2, while the corresponding test for direct censoring of an observed variable X has critical region of form $\{F(\min X)\}^{s_o}\{1-F(\max X)\}^{s_r} > K$.

I now propose a procedure for use when s_o and s_r are not known.

2. A GENERAL PURPOSE TEST

If s_o and s_r are not known, analogy with [2], wherein the criterion $F(\min \underset{\sim}{X})+ 1-F(\max \underset{\sim}{X})$ is indicated as a 'general purpose' test for direct censoring an observed variable X, suggests consideration of a criterion

$$T = E[F_2(\min \underset{\sim}{X}_2) + 1 - F_2(\max \underset{\sim}{X}_2)|\underset{\sim}{X}_1] \qquad (1)$$

as a basis for a general purpose test of indirect censoring on X_2, using observed values $\underset{\sim}{X}_1$ of X_1.

The conditional cumulative distribution functions, given $\underset{\sim}{X}_1$, of $L = \min \underset{\sim}{X}_2$ and $G = \max \underset{\sim}{X}_2$ are, respectively

$$1 - \prod_{j=1}^{r} \{1-F_{2|1}(\ell|X_{1j})\} \text{ and } \prod_{j=1}^{r} F_{2|1}(g|X_{1j})$$

and the corresponding density functions are:

$$\sum_{i=1}^{r} f_{2|1}(\ell|X_{1i}) \prod_{j\neq i}^{r} \{1-F_{2|1}(\ell|X_{1j})\} \text{ and } \sum_{i=1}^{r} f_{2|1}(g|X_{1i}) \prod_{j\neq i}^{r} F_{2|1}(g|X_{1j})$$

where $F_{2|1}(\cdot|X_1)$, $f_{2|1}(\cdot|X_1)$ denote the conditional cumulative distribution function and density function, respectively of X_2 given X_1. So in (1)

$$T = 1 + \sum_{i=1}^{r} \int_{-\infty}^{\infty} F_2(y) f_{2|1}(y|X_{1i}) \left[\prod_{j\neq i} \{1-F_{2|1}(y|X_{1j})\} - \prod_{j\neq i} F_{2|1}(y|X_{1j})\right] dy \quad (2)$$

If $F_{2|1}(y|X_{1j}) = H_2(F_2(y), X_{1j})$ so that

$$f_{2|1}(y|X_{1j}) = f_2(y) h_2(F_2(y), X_{1j}) \quad (3)$$

with $$h_2(u,v) = \frac{\partial H_2(u,v)}{\partial u}$$

then

$$T = 1 + \sum_{i=1}^{r} \int_0^1 z h(z,X_{1i}) \left[\prod_{j\neq i} \{1-H_2(z,X_{1j})\} - \prod_{j\neq i} H_2(z,X_{1j})\right] dz. \quad (4)$$

3. FARLIE-GUMBEL-MORGENSTERN POPULATION

In the special case where (X_{1j}, X_{2j}) have a joint Farlie-Gumbel-Morgenstern distribution with cumulative distribution function

$$F_1(x_1)F_2(x_2)[1 + \theta\{1-F_1(x_1)\}\{1-F_2(x_2)\}] \qquad (|\theta| \leq 1) \quad (5)$$

we have

$$H_2(y,x_1) = y[1 + \theta\{1-2F_1(x_1)\}(1-y)]$$

$$1- H_2(y,x_1) = (1-y)[1-\theta\{1-2F_1(x_1)\}y]$$

$$h_2(y,x_1) = 1 + \theta\{1-2F_1(x_1)\}(1-2y)$$

and

$$T = 1 + \sum_{i=1}^{r} \int_0^1 z\{1 + \theta Z_i(1-2z)\}[(1-z)^{r-1} \prod_{j\neq i} (1-\theta Z_j z) - z^{r-1} \prod_{j\neq i} \{1+\theta Z_j(1-z)\}]dz$$

where $$Z_h = 1-2F_1(X_{1h}) \qquad (6)$$

After some manipulation, we obtain

$$T = \frac{1}{r+1} + 2 \sum_{j=1}^{[r/2]} \frac{(2j)!r!}{(r+2j+1)!} \theta^{2j} Y_{2j} \qquad (7)$$

where $[\cdot]$ denotes integer part of $Y_{2j} = \sum_{i_1<\cdots<i_{2j}} \cdots \sum \prod_{h=1}^{2j} Z_{1i_h}$ (8)

There are $\binom{r}{2j}$ terms in the summation for Y_{2j}. It is interesting to note that, introducing the arithmetic mean

$$\bar{Y}_{2j} = Y_{2j}/\binom{r}{2j}$$

(7) can be written

$$(r+1)T = 1 + 2 \sum_{j=1}^{[r/2]} \frac{r^{(2j)}}{(r+2)^{[2j]}} \theta^{2j} \bar{Y}_{2j} \qquad (7)'$$

where $a^{(b)} = a(a-1) \ldots (a-b+1)$ and $a^{[b]} = a(a+1)\ldots+(a+b-1)$ are the ascending and descending factorials, respectively.

It is also interesting to note that the criterion obtained in testing for direct symmetrical ($s_o=s_r=s$) censoring is

$$1 + \sum_{j=1}^{[r/2]} \frac{(j+1)^{[j]}s^{[j]}}{(r+2s+1)^{[2j]}} \theta^{2j} Y_{2j} \qquad \text{(eq. (25) of [1])}$$

which has evident points of similarity with (7).

Formula (7) (or (7)') requires knowledge of θ for its evaluation (though it is not necessary to know its *sign*). Noting that $\bar{Y}_2 \geq \bar{Y}_4 \geq \ldots.$, and that the coefficients $\theta^{2j} r^{(2j)}/(r+2)^{[2j]}$ decrease as j increases, it seems

reasonable to consider using a test with critical region $Y_2 > K$ when θ is not known, as was suggested in [2] for the case $s_o = s_r$.

In [2] it was suggested that a normal distribution might be used to approximate the null hypothesis ($H_{o,o}$) distribution of Y_2, but this does not seem to be justifiable, even for large r. (For Y_1, on the other hand, it *is* a good approximation

The approximation (see Appendix)

$1 + 6Y_2 r^{-1}$ distributed as χ^2 with one degree of freedom seems to be appropriate. The corresponding critical region would be

$$Y_2 > \frac{1}{6} r(\lambda^2_{\frac{1}{2}\alpha} - 1)$$

where $\lambda_{\frac{1}{2}\alpha}$ is the upper $\frac{1}{2}\alpha$ point of the unit normal distribution ($\Phi(\lambda_{\frac{1}{2}\alpha}) = 1-\frac{1}{2}\alpha$).

REFERENCES

[1] Johnson, N.L. (1980). "Extreme sample censoring problems with multivariate data: Indirect censoring and the Farlie-Gumbel-Morgenstern distribution", *J. Multiv. Anal.* 10, 351-362.

[2] Johnson, N.L. (1970). "A general purpose test of censoring of extreme sample values", *S. N. Roy Memorial Volume*, University of North Carolina Press, pp. 377-384.

APPENDIX

We first show that if $Y_2 = \sum_{i<j}^{r-1} \sum^{r} Z_i Z_j$ and the Z's are independent and identically distributed, with zero expected value and all moments ($\mu'_q = \mu_q$) finite, then

(i) the limit of the moment ratio $\{\mu_q(Y_2)\}/\{\mu_2(Y_2)\}^{\frac{1}{2}q}$ as $r \to \infty$ is

finite and does not depend on the common distribution, and

(II) if the Z's are normal then the limiting distribution of

$$Y_2/\{\tfrac{1}{2}r(r-1)\mu_2'^2\}^{\frac{1}{2}} \text{ is that of } \frac{1}{\sqrt{2}}\{(\chi^2 \text{unit 1 d.f.})-1\}.$$

Result (I) suggests that this result applies generally.

Proof of (I) $E[Y_2) = 0$;

$$\mu_q(Y_2) = E[(\sum_{i<j}^{r-1}\sum^{r} Z_iZ_j)^q] \text{ and so, if } r > q,$$

$$\mu_q(Y_2) = \binom{r}{q} g_q \mu_2'^q + \text{terms in } r^{q-1}, r^{q-2}, \ldots r, 1$$

where g_q is the coefficient of $Z_1^2Z_2^2\ldots Z_q^2$ in $(\sum_{i<j}^{r-1}\sum^{r} Z_iZ_j)^q$ - the same for any $r > q$. Hence

$$\lim_{r\to\infty} \frac{\mu_q(Y_2)}{\{\mu_2(Y_2)\}^{\frac{1}{2}q}} = \frac{2^q}{q!}\frac{g_q}{g_2^{\frac{1}{2}q}} \tag{A1}$$

Proof of (II) If the Z's are normal $N(0,\mu_2)$, then

$$Y_2 = \tfrac{1}{2}\{(\sum_{i=1}^{r} Z_i)^2 - \sum_{i=1}^{r} Z_i^2] = \tfrac{1}{2}\{(r-1)(\bar{Z}\sqrt{r})^2 - \sum_{i=1}^{r}(Z_i - \bar{Z})^2\} \text{ when } \bar{Z} = r^{-1}\sum_{i=1}^{r} Z_i$$

$$= \tfrac{1}{2}\{(r-1)\chi_1^2 - \chi_{r1}^2\}\mu_2 = \tfrac{1}{2}\{(r-1)\chi_1^2 - \chi_{r-1}^2\}$$

where χ_1^2, χ_{r-1}^2 are mutually independent and $\mu_2 = E[Z^2] = 1$

$$\frac{Y_2}{\sqrt{\{\frac{1}{2}r(r-1)\}}} = \frac{1}{\sqrt{2}}\left[\chi_1^2\sqrt{\frac{r-1}{r}} - \frac{\chi_{r-1}^2}{\sqrt{\{r(r-1)\}}}\right] \tag{A2}$$

and the limiting distribution as $\tau \to \infty$, is that of $\frac{1}{\sqrt{2}}(\chi_1^2-1)$. Dr. W. Hoeffding has pointed out to me that Y_2 is proportional to a *degenerate* U-statistic, and that the general result can be established by a similar analysis.

In our case, when $Z = 1-2F_1(X_1)$ the common distribution of the Z's (under $H_{o,o}$) is uniform (-1,1) and

$$\mu_q = \begin{cases} 0 & \text{if } q \text{ is odd} \\ (q+1)^{-1} & \text{if } q \text{ is even} \end{cases} \tag{A3}$$

This suggests using the approximation

$$\frac{Y_2}{\sqrt{\{\frac{1}{18}\, r(r-1)\}}} \text{ approximately distributed as } \frac{1}{\sqrt{2}}(\chi_1^2 - 1)$$

or equivalently

$$1 + \frac{6Y_2}{\sqrt{\{r(r-1)\}}} \text{ approximately distributed as } \chi_1^2 \,. \tag{A4}$$

The $\sqrt{\{r(r-1)\}}$ may be conveniently replaced by r.

The first four moments of Y_2 are, in general

$$E[Y_2] = 0 \tag{A5.1}$$

$$\text{var}(Y_2) = \mu_2(Y_2) = \binom{r}{2}\mu_2^2 = \tfrac{1}{2} r^{(2)} \mu_2^2 \tag{A5.2}$$

$$\mu_3(Y_2) = \binom{r}{3}\binom{3}{1,1,1}\mu_2^3 + \binom{r}{2}\mu_3^2 = r^{(3)}\mu_2^3 + \tfrac{1}{2} r^{(2)}\mu_3^2 \tag{A5.3}$$

$$\begin{aligned} \mu_4(Y_2) &= \binom{r}{4}\{\binom{3}{2}\binom{4}{1,1,1,1} + 3\}\mu_2^4 + \binom{r}{3}3\binom{4}{2,1,1}\mu_3^2\mu_2 \\ &\quad + \binom{r}{3}3\binom{4}{2,2}\mu_4\mu_2^2 + \binom{r}{2}\mu_4^2 \\ &= \tfrac{15}{4} r^{(4)}\mu_2^4 + 6r^{(3)}\mu_3^2\mu_2 + 3r^{(3)}\mu_4\mu_2^2 + \tfrac{1}{2} r^{(2)}\mu_4^2 \end{aligned} \tag{A5.4}$$

Inserting the special values (A3), we obtain

$$E[Y_2 \,|H_{o,o}] = 0 \tag{A6.1}$$

$$\text{var}(Y_2|H_{o,o}) = \frac{1}{18} r^{(2)} \tag{A6.2}$$

$$\mu_3(Y_2|H_{o,o}) = \frac{1}{27}\, r^{(3)} \tag{A6.3}$$

$$\begin{aligned} \mu_4(Y_2|H_{o,o}) &= \frac{1}{50} r^{(2)} + \frac{1}{15} r^{(3)} + \frac{5}{108}\, r^{(4)} \\ &= \frac{1}{2700} r^{(2)}(125r^2 - 445r + 444) \end{aligned} \tag{A6.4}$$