This work was partially supported by the United States Air Force Office of Scientific Research under Grant No. AFOSR-68-1415.

COMPARISON OF SOME TESTS OF SAMPLE CENSORING OF EXTREME VALUES

by

N. L. JOHNSON
Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 665

JANUARY 1970

COMPARISON OF SOME TESTS OF SAMPLE CENSORING OF EXTREME VALUES

by

N. L. JOHNSON*
University of North Carolina at Chapel Hill
and University of New South Wales

1. There are currently available a number of methods designed to reduce the possible effects of "wild" ("maverick") observations on the analysis of sample values. Among these may be mentioned "trimming" and "Winsorisation". These methods involve the possible or sometimes automatic exclusion of extreme values among those observed. Apart from these methods, for which appropriate statistical analyses, taking proper account of the omission of sample values, are available, samples may be incomplete owing to inadequate recording, or, unfortunately, biassed selection of values which accord best with some preconceived ideas or desires.

While, under properly regulated conditions, information on any censoring of sample values should accompany the records of the values themselves, this is not always the case. Indeed, with the last situation described with the preceding paragraph, such information is not to be expected; but also, even in more respectable cases, information may be omitted by negligence.

The problems to be considered in this paper are those arising when it is suspected that there has been some form of censoring, i.e., omission of

* This work was partially supported by the United States Air Force Office of Scientific Research under Grant No. AFOSR-68-1415.

certain order statistics of the original sample. Complete, and reasonably tidy solutions are obtained only on the assumption that the population distribution of an observed character is known. However, study of this situation does give some clue as to what can be done when knowledge of the population distribution is incomplete.

Problems of a similar kind have been discussed in an earlier paper (Johnson (1962)). They were of a rather simple nature in that there was usually a direct choice between two possible sample sizes.

The tests discussed in this paper have been developed for use in situations where observed values $X_1', X_2', \ldots, X_r'$ are available which may represent measurements on a complete random sample, or may be the remaining part of such a sample after censoring. Denoting by $X_1, X_2, \ldots, X_r$ (with $X_1 \leq X_2 \leq \ldots \leq X_r$) the order statistics corresponding to $X_1', X_2', \ldots, X_r'$ the hypothesis $H_{s_0,s_1,\ldots,s_r}$ states that the original random sample was of size $(r + \sum_{j=0}^{r} s_j)$, with s_j values censored between X_j and X_{j+1} $(j = 0,1,\ldots,r;\ X_0 = -\infty, X_{r+1} = +\infty)$. The hypothesis $H_{0,0,\ldots,0}$ states that there has been no censoring.

We will be concerned here only with censoring of extreme values, so we assume $s_1 = s_2 = \ldots = s_{r-1} = 0$, and for brevity denote $H_{s_0,0,\ldots,s_r}$ by H_{s_0,s_r}. Construction of tests of $H_{0,0}$ with respect to alternatives of type H_{s_0,s_r} has been discussed by Johnson (1969) and will be briefly recapitulated here.

2.

It is supposed that each X' is continuous has the same density function, $f(x)$. For the application of the tests it is necessary that this function be known so that the corresponding probability integrals, $Y' = \int_{-\infty}^{X'} f(x)dx$ can be calculated.

The most powerful test of $H_{0,0}$ with respect to H_{s_0,s_r} has a critical region of form

(1) $$Y_1^{\theta}(1 - Y_r) > K_{\alpha}(\theta)$$

where $\theta = s_0/s_r$ $Y_1, Y_2, \ldots, Y_r$ are the ordered probability integrals corresponding to $X_1, X_2, \ldots, X_r$ respectively and α is the significance level of the test. (If $s_r = 0$, the region is of form $Y_1 >$ constant.) This test is uniformly most powerful with respect to all H_{s_0, s_r} for which $s_0/s_r = \theta$. In particular, when $s_0 = s_r(\theta=1)$, we have a critical region of form

(2) $$T_1 = Y_1(1 - Y_r) > A_{\alpha},$$

which is uniformly most powerful with respect to symmetrical censoring. When $s_0 = 0(\theta=0)$, we have a critical region of form

$$1 - Y_r > \text{constant}$$

(3) $$\text{i.e.,} \quad T_2 = Y_r < B_{\alpha}$$

which is uniformly most powerful with respect to censoring from above. (The region '$Y_1 >$ constant' is uniformly most powerful with respect to censoring from below.)

In order to construct a 'general purpose' test which should be good (though necessarily not most powerful) with respect to all H_{s_0,s_r} whatever the ratio $\theta = s_0/s_r$, an attempt was made by Johnson (1969) to use the union of critical regions of form (1),

$$Y_1^{\theta}(1-Y_r) > R_{\alpha'}(\theta)$$

where

$$\Pr[Y_1^{\theta}(1-Y_r) > R_{\alpha'}(\theta)|H_{0,0}] = \alpha'$$

with α' fixed. (The actual level of significance, α, will of course, not be equal to α'.) By using some approximations, a test with critical region of form

$$T_3 = Y_1 + (1-Y_r) > C_{\alpha} \tag{4}$$

was suggested by this approach.

The remainder of this paper is devoted to comparisons among the three tests just described.

3. We first state some results on the joint distribution of the ordered probability integrals $Y_1, Y_2, \ldots, Y_r$ which we will use later.

The joint density function of Y_1 and Y_r, when H_{s_0,s_r} is valid, is

$$p_{Y_1,Y_r}(y_1,y_r|H_{s_0,s_r}) = \frac{(r+s_0+s_r)!}{(r-2)!s_0!s_r!} y_1^{s_0}(1-y_r)^{s_r}(y_r-y_1)^{r-2} \tag{5}$$

$$(0 < y_1 < y_r).$$

Also, the quantities

$$\{Y_{j+1} - Y_j\} \quad (j = 0,1,\ldots,r; \quad Y_0 = 0, \quad Y_1 = 1)$$

are jointly distributed as

$$V_j \left(\sum_{i=0}^{r} V_i \right)^{-1}$$

where the V's are mutually independent, V_0 is distributed as $\chi^2_2(s_0+1)$, V_r as $\chi^2_2(s_r+1)$ and $V_1, V_2, \ldots, V_{r-1}$ each as χ^2_2. (χ^2_r means 'χ^2 with r degrees of freedom'.)

4. The power of test T_1, with respect to H_{s_0,s_r} is

$$(6)\quad \Pr[Y_1(1-Y_r) > A_\alpha | H_{s_0,s_r}] = \frac{(r+s_0+s_r)!}{(r-2)!s_0!s_r} \iint_R y_1^{s_0}(1-y_r)^{s_r}(y_r-y_1)^{r-2}dy_r dy_1$$

where the region of integration (R) is $y_1(1-y_r) > A_\alpha$ and A_α satisfies the equation

$$(7)\qquad r(r-1)\int_R\int (y_r-y_1)^{r-2}dy_r dy_1 = \alpha.$$

By writing $(1-y_r)$ as $(1 - y_1 - [y_r-y_1])$ in (6) and expanding the various binomial expressions, the formula can be expressed as a linear function of quantities:

$$(8)\qquad I_{a,b} = \int_{y_-}^{y_+} y^a(1-A_\alpha y^{-1}-y)^b dy$$

where $y_\pm = \frac{1}{2}[1\pm\sqrt{1-4A_\alpha}]$.

Calculation of $I_{a,b}$ for integer values of a and b is direct but rather tedious. It is aided by the recurrence relations

$$(9.1)\qquad I_{a,b+1} = I_{a,b} - I_{a+1,b} - A_\alpha I_{a-1,b}$$

$$(9.2)\qquad I_{a,b+1} = (a+1)^{-1}(b+1)[I_{a+1,b}-A_\alpha I_{a-1,b}] \qquad (a \neq -1)$$

and

$$(9.3) \qquad I_{a,b} = A_\alpha^{a+1} I_{-(a+2),b} .$$

Equation (9.2) is obtained by integration by parts (noting that $1-A_\alpha y_\pm^{-1}-y_\pm = 0$). Combining (9.1) and (9.2), we have

$$(9.4) \qquad I_{a+1,b} = (a+b+2)^{-1} [(a+1)I_{a,b} + (b-a)A_\alpha I_{a-1,b}].$$

As initial values we have

$$(9.5) \qquad I_{a,0} = [2^a(a+1)]^{-1} \sqrt{1-4A_\alpha} \sum_{j=0}^{[\frac{1}{2}a+1]} \binom{a+1}{2j-1} (1-4A_\alpha)^j \qquad (a \neq -1)$$

$$(9.6) \qquad I_{-1,0} = \log(1 + \sqrt{1-4A_\alpha}) - \log(1 - \sqrt{1-4A_\alpha}) .$$

Equation (7) can be written in the form

$$(7)' \qquad r\, I_{0,r-1} = \alpha.$$

Some values of A_α satisfying this equation are shown in Table 1. This table also shows values given by empirical approximate formulae.

TABLE 1

Values of A_α

	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
r	A_α	$2.65(r+\frac{3}{2})^{-2}$	A_α	$4.1(r+2)^{-2}$	A_α	$9.15(r+\frac{7}{2})^{-2}$
2	0.182	-	0.207	-	0.235	-
3	0.122	-	0.150	-	0.195	-
4	0.0841	0.0876	0.109	0.114	0.156	0.163
5	0.0611	0.0627	0.0822	0.0837	0.125	0.127
6	0.0463	0.0471	0.0633	0.0641	0.101	0.101
7	0.0362	0.0367	0.0503	0.0506	0.0830	0.0830
8	0.0289	0.0294	0.0408	0.0410	0.0692	0.0692
9	0.0238	0.0240	0.0338	0.0339	0.0585	0.0586
10	0.0199	0.0200	0.0285	0.0285	0.0500	0.0502
Large	$2.583r^{-2}$		$3.997r^{-2}$		$8.315r^{-2}$	

For large values of r an approximate formula for A_α can be obtained by the following argument. From Section 2, we have

$$\Pr[V_0 V_r (V_0+V_r+V')^{-2} > A_\alpha] = \alpha \tag{10}$$

where V_0, V_r and $V'(= \sum_{j=1}^{r-1} V_j)$ are independent variables distributed as χ^2_2, χ^2_2 and $\chi^2_{2(r-1)}$ respectively. Putting $A_\alpha = A'_\alpha/r^2$ we have

$$\Pr[V_0 V_r \{r^{-1}(V_0+V_r+V')\}^{-2} > A'_\alpha] = \alpha. \tag{10}'$$

Since $r^{-1}(V_0+V_r+V')$ tends to 2 as $r \to \infty$, with probability 1, it follows that the limiting distribution of $V_0V_r\{r^{-1}(V_0+V_r+V')\}^{-2}$ is that of $\frac{1}{4} \times$ (product of two independent χ^2_2 variables). Hence for large r, A'_α tends to the solution of the equation

$$\int_0^\infty \exp(-u-A'_\alpha u^{-1})du = \alpha$$

i.e.,

$$2\sqrt{A'_\alpha}\, K_1(2\sqrt{A'_\alpha}) = \alpha \tag{11}$$

where $K_1(.)$ is a Bessel function (see, for example, British Association tables (1950)). Some values of A'_α from (11) are shown in Table 1.

A similar argument leads to the conclusion that the limiting power, as $r \to \infty$ (s_0 and s_r remaining constant) is (for $s_r \geq s_0$)

$$\frac{2A'^{\frac{1}{2}(s_r+1)}_\alpha}{s_r!} \sum_{j=0}^{s_0} \frac{A'^{\frac{1}{2}j}_\alpha}{j!} K_{s_r-j+1}(2\sqrt{A'_\alpha}) \tag{12}$$

($K_\nu(.)$ denotes a Bessel function of order ν.)
Note that if $s_0 = 0$, the limiting power is $2A'^{\frac{1}{2}(s_r+1)}_\alpha (s_r!)^{-1} K_{s_r+1}(2\sqrt{A'_\alpha})$.

The power of test T_1 with respect to H_{0,s_r} is

$$\beta_{s_r} = \frac{(r+s_r)!}{(r-1)!s_r!} I_{s_r,r-1}.$$

Since $I_{s_r+1,r-1} / I_{s_r,r-1} < y_+ < 1$, it follows that for s_r sufficiently large

$$\beta_{s_r+1} / \beta_{s_r} = \left(\frac{r+s_r+1}{s_r+1}\right) (I_{s_r+1,r-1} / I_{s_r,r-1}) < 1$$

and so the power tends to zero as $s_r \to \infty$, r being kept constant.

5. The critical region for test T_2 is

$$Y_r < \alpha^{1/r} \tag{13}$$

and the power with respect to H_{s_0,s_r}

$$[B(r+s_0,s_r+1)]^{-1} \int_0^{\alpha^{1/r}} y^{r+s_0-1} (1-y)^{s_r} dy = I_{\alpha^{1/r}}(r+s_0,s_r+1). \tag{14}$$

As $r \to \infty$ (s_0 and s_r remaining constant) the power tends to $\Pr[\chi^2_{2(s_r+1)} > -2 \log \alpha]$. Note that this does not depend on s_0.

As $s_0 \to \infty$ (r and s_r remaining constant) the power tends to zero, and as $s_r \to \infty$ (r and s_0 remaining constant) the power tends to 1. If s_0 and s_r both tend to infinity, with $s_0/s_r = \theta$ and r both kept constant, a rather curious situation arises. The power tends to zero or one, depending on the values of α (the level of significance), and of r.

To see how this happens, we consider the case $\theta = 1$. For $s_0 = s_r = s$ large, the beta distribution with parameters $(r+s)$, $(s+1)$ is approximately normal, with expected value tending to $\frac{1}{2}$ and variance of order s^{-1}. The power therefore tends to zero or one according as $\alpha^{1/r}$ is less, or greater, than $\frac{1}{2}$. For a given value of r, this means that the power tends to zero or one according as α is less, or greater, than 2^{-r}. For a given α, the power tends to zero or one according as r is less or greater than $(-\log \alpha)/(\log 2)$. For $\alpha = 0.05$, the power tends to zero as $s \to \infty$ if $r \leq 4$. Table 2 shows some values of the power for $\alpha = 0.05$ and $r = 4$,

with various values of s. The power increases with s to a flat maximum, and then decreases.

TABLE 2

Power of T_2 test with respect to $H_{s,s}$,
($\alpha = 0.05$; $r = 4$)

s	Power	s	Power
7	0.174	19	0.207
9	0.185	21	0.208
11	0.193	23	0.209
13	0.199	25	0.209
15	0.202	27	0.209
17	0.205	29	0.208

(For lower values of s, see Table 3.)

6. The theory associated with the general purpose test T_3 is simple, because $Y_1 + (1-Y_r)$ is distributed as $(V_0+V_r)(V_0+V'+V_r)^{-1}$ (see equation (10) for distributions of V's). That is, the distribution is beta with parameters $s_0 + s_r + 2$, $r - 1$. The critical region is

$$Y_1 + (1-Y_r) > C_\alpha$$

where

$$I_{C_\alpha}(2, r-1) = 1 - \alpha$$

and the power with respect to H_{s_0,s_r} is $1 - I_{C_\alpha}(s_0+s_r+2, r-1)$.

The power depends only on (s_0+s_r), and not on s_0 and s_r separately. As $r \to \infty$, with (s_0+s_r) remaining constant, the power tends to

$$\Pr[\chi^2_2(s_0+s_r+2) > (\text{upper } 100\alpha\% \text{ point of } \chi^2_4)] .$$

As $(s_0+s_r) \to \infty$, r remaining constant, the power tends to one.

7. The results of calculations based on formulae developed in the last three sections are shown in Table 3. These figures appear to indicate the practical usefulness of T_3 as a 'general purpose' test. This test insures against the possibility of having a very low power without sacrificing too much relative to the most powerful tests available. Of course, the latter should be used when the type of censoring suspected (i.e., the value of θ) is very definitely known.

TABLE 3

Comparison of Powers ($\alpha = 0.05$)

Test Criterion	r	s_0+s_r =	2		6		10	
		(s_0,s_r) =	(0,2)	(1,1)	(0,6)	(3,3)	(0,10)	(5,5)
$T_1 = Y_1(1-Y_r)$	4		0.124	0.206	0.120	0.594	0.080	0.841
	∞		0.239	0.339	0.533	0.925	0.644	0.998
$T_2 = Y_r$	4		0.294	0.086	0.780	0.131	0.955	0.157
	30		0.381	0.177	0.949	0.549	0.999	0.821
	∞		0.424	0.200	0.967	0.648	0.999	0.996
$T_3 = Y_1+(1-Y_r)$	4		0.167		0.470		0.716	
	30		0.281		0.845		0.989	
	∞		0.303		0.892		0.996	

In interpreting the figures shown in Table 3, it should be noted that considerably higher powers will be obtained when series of two or more samples, each possibly subject to the same system of censoring, are available.

It may be felt that the condition stated at the beginning of Section 2, namely that the true probability density function $f(x)$ must be known, is unlikely to be satisfied in practice. While this is so, in the strict sense that it is very rarely the case that a theoretically formulated model gives an exact representation of reality, it will sometimes be the case that there is sufficiently massive evidence to establish $f(x)$, from observed relative frequencies, with adequate accuracy. It may be noted that it is not essential that $f(x)$ have a simple, or indeed any explicit, mathematical form -- a graphical representation can suffice. Slight variations in form of $f(x)$ can be tolerated without serious effect. (Since the test criteria depend only on Y_1 and Y_r, inaccuracy in $\int_{-\infty}^{X} f(x)dx$ for values of X in the central part of the distribution have little effect.)

It would, however, be interesting, but beyond the scope of the present investigation, to inquire into the robustness of these tests with respect to variation in $f(x)$.

REFERENCES

Johnson, N.L. (1962). "Estimation of sampel size", *Technometrics*, 4, 59-67.

Johnson, N.L. (1969) "A general purpose test of censoring of extreme sample values", *S.N. Roy Memorial Volume*, Indian Statistical Institute.

British Association Mathematical Tables (1950). Volume 6, *Bessel Functions* (Part I).