[1]This research was supported in part by the U.S. Army Research Office, Durham, under Contract No. DAHC04-71-C-0042.

ROBUSTNESS OF CERTAIN TESTS OF CENSORING OF EXTREME SAMPLE VALUES[1]

N.L. Johnson

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 866

March, 1973

ROBUSTNESS OF CERTAIN TESTS OF CENSORING
OF EXTREME SAMPLE VALUES [1]

N.L. Johnson

University of North Carolina at Chapel Hill

ABSTRACT

The effects of inaccuracy in parameters of a hypothesized normal distribution on certain tests of symmetrical sample censoring are investigated.

[1] This research was supported in part by the U.S. Army Research Office, Durham, under Contract No. DAHC04-71-C-0042.

Robustness of Certain Tests of Censoring of Extreme Sample Values

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

1. Introduction

Tests of censoring of sample values, and in particular of extreme sample values, have been described in [1]-[3]. All these tests use the probability integral transformation

$$Y = \int_{-\infty}^{X} f(x)dx$$

of sample values of a continuous random variable, X, with density function f(x). If this density function is not accurately known, the test procedure might be based on an estimated f(x). It is the purpose of this paper to give an assessment of the effects of errors in f(x) on the properties of some of the tests described in [1]-[3]. In particular we will consider tests with critical regions

(i) $Y_1 \geq C_\alpha$

(ii) $Y_1(1-Y_r) \geq C_\alpha$

(iii) $Y_1 + (1-Y_r) \geq C_\alpha$

where $Y_1 \leq Y_2 \leq \ldots \leq Y_r$ are the ordered probability integral transforms corresponding to an available sample set of r observed values of X, and the C_α's are chosen to give a significance level α. The C_α's in (i), (ii) and (iii) do not, of course, have the same values.

If it be supposed that the r observed values are the surviving members of a complete random sample of size $n = r+s_0+s_r$, after the s_0 least and s_r greatest values have been censored, the tests (i) and (ii) are uniformly most powerful tests of the hypothesis that $n = n_0(\geq r)$ with respect to alternatives $n > n_0$ when censoring is restricted to be from below (i.e. $s_r = 0$) in case (i), or symmetrical (i.e. $s_0 = s_r$) in case (ii). Test (iii) was suggested, on heuristic grounds, in [1] as being suitable for use when the ratio s_0/s_r is unknown. Subsequent investigation [2] has tended to support the validity of this suggestion.

2. General Theory

If the correct density function $f(x)$ is used in calculating

$$Y'_j = \int_{-\infty}^{X_j} f(x)dx \qquad (j = 1,\ldots,r) \tag{1}$$

then the joint density function of the corresponding order statistics $Y_1 \leq Y_2 \leq \ldots \leq Y_r$ is

$$\frac{(r+s_0+s_r)!}{s_0!\ s_r!} y_1^{s_0}(1-y_r)^{s_r} \qquad (0\leq y_1 \leq y_2 \leq \ldots \leq y_r \leq 1) \tag{2}$$

If a density function $g(x)$ is used, in place of the correct $f(x)$, then the order statistics $Y_1[g] \leq Y_2[g] \leq \ldots \leq Y_r[g]$ corresponding to

$$Y'_j[g] = \int_{-\infty}^{X_j} g(x)dx \tag{3}$$

will in general have a distribution different from (2).

If Y'_j and $Y'_j[g]$ are each strictly increasing functions of X_j, then $Y_j[g]$ is also a strictly increasing function of Y_j. In fact

$$Y_j[g] = h(Y_j)$$

where $h(y)$ is defined by

$$h(y) = \int_{-\infty}^{t(y)} g(t)dt \quad \text{with} \quad y = \int_{-\infty}^{t(y)} f(t)dt$$

It follows that any set of inequalities involving the $Y_j[g]$'s can be expressed in terms of inequalities involving the Y_j's, which have known joint distributions.

In particular

(5.1) $$\Pr[Y_1[g] \geq C_\alpha | n] = \Pr[h(Y_1) \geq C_\alpha | n]$$

(5.2) $$\Pr[Y_1[g]\{1-Y_r[g]\} \geq C_\alpha | n] = \Pr[h(Y_1)\{1-h(Y_r)\} \geq C_\alpha | n]$$

and

(5.3) $$\Pr[Y_1[g] + \{1-Y_r[g]\} \geq C_\alpha | n] = \Pr[h(Y_1) + \{1-h(Y_r)\} \geq C_\alpha | n] \quad .$$

Calculation of these probabilities may sometimes be technically awkward, but is straightforward in principle.

3. Censoring from Below

This is the simplest case, since (5.1) can be written

(5.1)' $$\Pr[Y_1[g] \geq C_\alpha | n] = \Pr[Y_1 \geq h^{-1}(C_\alpha) | n]$$

and Y_1 has a beta distribution with parameters (s_0+1) and $(r+s_r)$. (Of course, if censoring *is* from below, then $s_r = 0$.)

Hence

(6) $$\Pr[Y_1[g] \geq C_\alpha | n] = 1 - I_{h^{-1}(C_\alpha)}(s_0+1, r+s_r)$$

where $I(\cdot)$ is the incomplete beta function ratio [2].

In any particular case (i.e. for any particular $f(\cdot)$ and $g(\cdot)$) we need study only the appropriate function $h^{-1}(y)$.

Since $Y_1[g] \geq C_\alpha$

is equivalent to $Y_1 \geq h^{-1}(C_\alpha)$

it is clear that if $g(x)$ is used in place of $f(x)$, we are, in effect, using the *correct* density function $(f(x))$ with an *actual* significance level α^k defined by

(7) $$C_{\alpha^*} = h^{-1}(C_\alpha)$$

instead of the required value α.

As an example, suppose that $f(x)$ is the normal density function

$$(\sqrt{2\pi}\ \sigma)^{-1} \exp\{-\tfrac{1}{2}(\tfrac{x-\xi}{\sigma})^2\}$$

and that $g(x)$ is also normal, but with an incorrect pair of values ξ^*, σ^* in place of ξ, σ. Then, in (4)

$$t(y) = \xi + \sigma\ \Phi^{-1}(y)$$

and

(8) $$\begin{aligned} h(y) &= \Phi(\sigma^{*-1}\{\xi + \sigma\ \Phi^{-1}(y) - \xi^*\}) \\ &= \Phi(\tfrac{\sigma}{\sigma^*}\ \Phi^{-1}(y) - \tfrac{\xi^*-\xi}{\sigma^*}) \end{aligned}$$

where $\Phi(u) = (\sqrt{2\pi})^{-1} \int_{-\infty}^{u} e^{-\frac{1}{2}x^2} dx$.

Conversely,

(9) $$h^{-1}(y) = \Phi(\tfrac{\sigma^*}{\sigma}\ \Phi^{-1}(y) + \tfrac{\xi^*-\xi}{\sigma})$$

We will obtain some numerical results for the case $n_0 = r$ -i.e. test of completeness of the available data. In this case

$$C_\alpha = 1 - \alpha^{1/r} \quad . \tag{10}$$

Taking $\alpha = 0.05$ we have the following values of $C_{0.05}$:

$r =$	5	10	15	20
$C_{0.05} =$	0.4507	0.2589	0.1810	0.1391

From (7) and (9)

$$C_{\alpha*} = \Phi\left(\frac{\sigma*}{\sigma} \; \Phi^{-1}(C_\alpha) + \frac{\xi*-\xi}{\sigma}\right) \tag{11}$$

Given $C_{\alpha*}$, we can calculate $\alpha* = (1-C_{\alpha*})^r$.

Values of $\Phi^{-1}(C_{0.05})$ are:

$r =$	5	10	15	20
$\Phi^{-1}(C_{0.05}) =$	-0.1239	-0.6469	-0.9116	-1.0844

Table 1 gives values of $\alpha*$ for various combinations of values of $\sigma*/\sigma$ and $(\xi*-\xi)/\sigma$. These should be compared with the nominal significance level of 5%, on which the calculations are based.

Table 1 : Values of α^* (actual significance level)(Nominal level: 5%)
(Test for one-sided censoring)

	σ^*/σ	$\frac{\xi^*-\xi}{\sigma}$ = -0.4	-0.2	0	0.2	0.4
r=5	2	0.224	0.138	0.075	0.038	0.017
	1	0.168	0.097	0.050	0.023	0.009
	½	0.143	0.080	0.040	0.018	0.007
r=10	2	0.630	0.496	0.357	0.229	0.130
	1	0.202	0.109	0.050	0.019	0.006
	½	0.069	0.028	0.009	0.002	0.001
r=15	2	0.803	0.720	0.594	0.446	0.299
	1	0.224	0.117	0.050	0.017	0.004
	½	0.038	0.012	0.003	-	-
r=20	2	0.904	0.828	0.738	0.603	0.456
	1	0.239	0.136	0.050	0.015	0.003
	½	0.022	0.006	0.001	-	-

(- denotes "less than 0.0005")

The importance of having a reasonably accurate estimate of σ is clear. If σ is not accurately estimated, deviation of actual from nominal significance level increases rapidly with σ^*.

4. Symmetrical and General Censoring

In the other two cases ((ii) and (iii)) it is usually necessary to employ numerical quadrature. Formula (5.2) can be evaluated as

$$\Pr[h(Y_1)\{1-h(Y_r)\} \ge C_\alpha | n] = E[\Pr[h(Y_1) \ge C_\alpha\{1-h(Y_r)\}^{-1} | n, Y_r]]$$

(5.2)'

$$= E[\Pr[Y_1 \ge h^{-1}(C_\alpha\{1-h(Y_r)\}^{-1}) | n, Y_r]$$

$$= E[I_{1-\gamma_2(Y_r)}(r-1, s_0+1)]$$

where

$$\gamma_2(Y_r) = \begin{cases} h^{-1}(C_\alpha\{1-h(Y_r)\}^{-1})/Y_r & \text{for } h(Y_r)\{1-h(Y_r)\} \ge C_\alpha, \\ 1 & \text{otherwise} \end{cases}$$

and expectation is taken with respect to Y_r. (Y_r has a standard beta distribution with parameters (s_0+r), (s_r+1); the conditional distribution of Y_1, given Y_r, is beta with parameters (s_0+1), $(r-1)$, **and range $(0,Y_r)$.**

Similarly, from (5.3),

(5.3)' $$\Pr[h(Y_1) + 1 - h(Y_r) \ge C_\alpha | n] = E[I_{1-\gamma_3(Y_r)}(r-1, s_0+1)]$$

where

$$\gamma_3(Y_r) = \begin{cases} h^{-1}(C_\alpha - 1 + h(Y_r))/Y_r & \text{for } h(Y_r) \ge 1 - C_\alpha, \\ 0 & \text{for } h(Y_r) < 1 - C_\alpha. \end{cases}$$

When there is no censoring $(s_0 = s_r = 0)$ then

$$I_{1-\gamma_j(Y_r)}(r-1, s_0+1) = \{1-\gamma_j(Y_r)\}^{r-1}$$

and (5.2)', (5.3)' become

(12) $$E[\{1-\gamma_j(Y_r)\}^{r-1}] = r\int_0^1 \{y[1-\gamma_j(Y_r)]\}^{r-1}\,dy$$

$(j = 2,3$ respectively)

Note that $\gamma_2(y) = 1$ for $y < h^{-1}(\frac{1}{2}[1-\sqrt{1-4C_\alpha}\,])$

and $\gamma > h^{-1}(\frac{1}{2}[1+\sqrt{1-4C_\alpha}\,])$;

and $\gamma_3(y) = 0$ for $y > h^{-1}(1-C_\alpha)$.

Hence the limits of the integral for $E[\{1-\gamma_2(Y_r)\}^{r-1}]$ can be changed from 0 to 1 to $h^{-1}(\frac{1}{2}[1-\sqrt{1-4C_\alpha}\,])$, and also

$$(13)\quad E[\{1-\gamma_3(Y_r)\}^{r-1}] = \{h^{-1}(1-C_\alpha)\}^r + r\int_{h^{-1}(1-C_\alpha)}^{1} \{y[1-\gamma_3(y)]\}^{r-1}\,dy$$

$$= \{h^{-1}(1-C_\alpha)\}^r + r\int_{h^{-1}(1-C_\alpha)}^{1} \{y-h^{-1}(C_\alpha-1+h(y))\}^{r-1}dy$$

The integrals in (12) and (13) will usually have to be evaluated by quadrature. As r increases it may be difficult to retain adequate accuracy.

We now turn to the special case considered in Section 3, with

$$f(x) = (\sqrt{2\pi}\,\sigma)^{-1}\exp\{-\tfrac{1}{2}(\tfrac{x-\xi}{\sigma})^2\}$$

$$g(x) = (\sqrt{2\pi}\,\sigma^*)^{-1}\exp\{-\tfrac{1}{2}(\tfrac{x-\xi^*}{\sigma^*})^2\}$$

$$h(y) = \Phi(\tfrac{\sigma}{\sigma^*}\,\Phi^{-1}(y) - \tfrac{\xi^*-\xi}{\sigma^*})$$

$$h^{-1}(y) = \Phi(\tfrac{\sigma^*}{\sigma}\,\Phi^{-1}(y) + \tfrac{\xi^*-\xi}{\sigma}) \quad .$$

Using (8) and (9) we find

$$(14.1)\qquad Y_r\gamma_2(Y_r) = \Phi\left(\frac{\sigma^*}{\sigma}\,\Phi^{-1}\left[\frac{C_\alpha}{1-\Phi(\frac{\sigma}{\sigma^*}\,\Phi^{-1}(Y_r) - \frac{\xi^*-\xi}{\sigma^*})}\right] + \frac{\xi^*-\xi}{\sigma}\right)$$

for Y_r in the limits

$$\Phi(\tfrac{\sigma^*}{\sigma}\,\Phi^{-1}(\tfrac{1}{2}[1-\sqrt{1-4C_\alpha}\,]) + \tfrac{\xi^*-\xi}{\sigma})$$

and

$$(14.2)\qquad Y_r\gamma_3(Y_r) = \Phi(\tfrac{\sigma^*}{\sigma}\,\Phi^{-1}[C_\alpha-1+\Phi(\tfrac{\sigma}{\sigma^*}\,\Phi^{-1}(Y_r) - \tfrac{\xi^*-\xi}{\sigma^*})] + \tfrac{\xi^*-\xi}{\sigma})$$

for

$$Y_r \geq \Phi(\tfrac{\sigma^*}{\sigma}\,\Phi^{-1}(1-C_\alpha) + \tfrac{\xi^*-\xi}{\sigma})$$

Introducing the function (defined for $0 \le z \le 1$)

$$\delta(z;\ A,B) = \Phi(A\ \Phi^{-1}(z) + B)$$

we have

$$h(y) = \delta\left(y;\ \frac{\sigma}{\sigma^*},\ -\frac{\xi^*-\xi}{\sigma^*}\right)$$

and (14.1), (14.2) can be re-written

$$(14.1)' \qquad Y_2\gamma_2(Y_r) = \delta\left(\frac{C_\alpha}{1-\delta(Y_r;\ \frac{\sigma}{\sigma^*},\ -\frac{\xi^*-\xi}{\sigma^*})};\ \frac{\sigma^*}{\sigma},\ \frac{\xi^*-\xi}{\sigma}\right)$$

for Y_r in the limits

$$\delta\left(\tfrac{1}{2}[1 \pm \sqrt{1-4C_\alpha}]\ \frac{\sigma^*}{\sigma};\ \frac{\xi^*-\xi}{\sigma}\right);$$

$$(14.2)' \qquad Y_r\gamma_3(Y_r) = \delta\left(C_\alpha - 1 + \delta(Y_r;\ \frac{\sigma}{\sigma^*},\ -\frac{\xi^*-\xi}{\sigma^*});\ \frac{\sigma^*}{\sigma},\ \frac{\xi^*-\xi}{\sigma}\right)$$

$$\text{for } Y_r \ge \delta\left(1-c_\alpha;\ \frac{\sigma^*}{\sigma},\ \frac{\xi^*-\xi}{\sigma}\right).$$

Note that the inverse of $\delta(z;\ A,\ B)$ is $\delta(a;\ A^{-1},\ -BA^{-1})$.

The actual significance levels of the tests are then, for the test for symmetrical censoring

$$(15.1) \qquad r\int_{Y^-}^{Y^+} \left\{y - \delta\left(\frac{C_\alpha}{1-\delta(y;\ \frac{\sigma}{\sigma^*},\ -\frac{\xi^*-\xi}{\sigma^*})};\ \frac{\sigma^*}{\sigma},\ \frac{\xi^*-\xi}{\sigma}\right)\right\}^{r-1} dy$$

where $Y^{\pm} = \delta(\tfrac{1}{2}[1 \pm \sqrt{1-4C_\alpha}]);\frac{\sigma^*}{\sigma};\ \frac{\xi^*-\xi}{\sigma})$,

and, for the "general purpose" test

$$(15.2) \qquad \left[\delta\left(1-C_\alpha;\frac{\sigma^*}{\sigma},\frac{\xi^*-\xi}{\sigma}\right)\right]^r + r\int_Y^1 \left\{y - \delta\left(C_\alpha - 1 + \delta(y;\frac{\sigma}{\sigma^*},-\frac{\xi^*-\xi}{\sigma^*});\ \frac{\sigma^*}{\sigma},\ \frac{\xi^*-\xi}{\sigma}\right)\right\}^{r-1} dy$$

where

$$Y = \delta\left(1-C_\alpha;\ \frac{\sigma^*}{\sigma},\ \frac{\xi^*-\xi}{\sigma}\right).$$

(Note that these formulae will apply, with appropriate choice of function to replace $\Phi(\circ)$, to any distribution depending only on a location parameter and a scale parameter σ, i.e. with a density function of form $\frac{1}{\sigma} w(\frac{x-\xi}{\sigma})$.)

Tables 2 and 3 are analogous to Table 1, and are based on (15.1) and (15.2) respectively. The values of $C_{0.05}$ used are as follows:

Value of $C_{0.05}$ in

r =	Table 2	Table 3
5	0.08183	0.65741
10	0.02842	0.39416
15	0.01413	0.27940
20	0.00841	0.21611

Table 2: Values of α^* (actual significance level) (Nominal level: 5%)

(Test for symmetrical censoring)

	σ^*/σ \ $\frac{\vert\xi^*-\xi\vert}{\sigma}$ = 0	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
r=5	2	0.466	0.461	0.445	0.418	0.381	0.336	0.285	0.230
	1	0.050	0.048	0.041	0.043	0.023	0.018	0.014	0.008
	$\frac{1}{2}$	0.002	0.002	0.002	0.001	0.001	-	-	-
r=10	2	0.815	0.811	0.801	0.783	0.756	0.719	0.672	0.612
	1	0.050	0.047	0.041	0.031	0.021	0.012	0.006	0.003
	$\frac{1}{2}$	-	-	-	-	-	-	-	-
r=15	2	0.937	0.931	0.926					
	1	0.050	0.047	0.040					
	$\frac{1}{2}$	-	-	-					
r=20	2	0.977	0.971	0.969					
	1	0.050	0.047	0.040					
	$\frac{1}{2}$	-	-	-					

Table 3: Values of α^* (actual significance level) (Nominal level 5%)

(General purpose test for censoring)

	σ^*/σ \ $\frac{\lvert\xi^*-\xi\rvert}{\sigma}$ = 0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
r=5	2: 0.305	0.309	0.323	0.346	0.378	0.420	0.472	0.531
	1: 0.050	0.056	0.075	0.111	0.165	0.241	0.336	0.445
	$\frac{1}{2}$: 0.030	0.039	0.069	0.123	0.202	0.304	0.421	0.542
r=10	2: 0.716	0.722	0.740	0.767	0.801	0.838	0.874	0.907
	1: 0.050	0.060	0.094	0.155	0.247	0.364	0.495	0.623
	$\frac{1}{2}$: 0.006	0.011	0.030	0.072	0.146	0.253	0.385	0.525
r=15	2: 0.893	0.897	0.906					
	1: 0.050	0.063	0.106					
	$\frac{1}{2}$: 0.001	0.004	0.015					
r=20	2: 0.957	0.958	0.962					
	1: 0.050	0.066	0.116					
	$\frac{1}{2}$: 0.001	0.002	0.008					

In Table 2, it is noteworthy that when $\sigma = \sigma^*$ and $|\xi^*-\xi|/\sigma$ is fixed, the actual significance level varies only slowly with r, the apparent sample size. In Table 3, we notice that for $r = 5$ and $|\xi^*-\xi| > 0.5$, there appears to be a minimum value of α^* as σ^*/σ varies. There is a similar effect for larger values of r, but it does not show within the confines of this table.

It is clear that for reasonable control of significance level, σ^*/σ must be quite close to 1, though some variation (say up to 0.25) in $|\xi^*-\xi|/\sigma$ can be tolerated.

REFERENCES

[1] 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.

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

[3] Johnson, N.L. (1972) Inferences on sample size: Sequences of samples, *Communication in Statistics*, **1**, 17-26.