EXTREME SAMPLE CENSORING PROBLEMS WITH MULTIVARIATE DATA - I

by

N. L. Johnson[1]

Department of Statistics

University of North Carolina at Chapel Hill

June 1975

Institute of Statistics Mimeo Series No. 1010

[1] This research was supported by the AROD under Grant No. DAHCO4-74-0030.

Extreme Sample Censoring Problems with Multivariate Data - I

by

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

1. Introduction. This paper contains a discussion of ways in which problems of the kind described in [1] - [3] may arise in connection with multivariate data. Censoring of univariate data can be "generalized" in many different ways, and there is a corresponding variety of problems connected with the detection of such censoring. In this paper, we will describe some classes of problems and indicate methods for their solution. Later reports will contain details for specific problems, and relevant numerical tables.

2. Bivariate Data. Suppose we have r pairs of observed values (X_{1j}, X_{2j}) $(j = 1, 2, \ldots, r)$ on two variables X_1 and X_2, and we wish to investigate whether these have been obtained by some form of censoring from an original camplete random sample of size n.

We distinguish the following cases:

(i) It is known that, if there is censoring, it is with respect to values of a specified variable - X_1, say.

(ii) It is known that, if there is censoring, it is with respect to values of one (only) of X_1 and X_2, but it is not known which one.

(iii) It is known that, if there is censoring it is with respect to some linear function - $aX_1 + \sqrt{1 - a^2}\, X_2$, say $(-1 < a \leq 1)$ - of X_1 and X_2 - with a unknown. (In this case, in addition to the problem of testing for censoring, there is that of estimating a, assuming that censoring has occurred.)

(iv) A related problem arises when we have r values $Z_1, Z_2, \ldots, Z_r$ of a variable Z which may represent a complete random sample, or one which has been censored with respect to values of an unrecorded, related variable X.

We will discuss considerations relating to (i) - (iv) in the next four sections. Some multivariate problems will be described in the final section. Our discussions will, in all cases, be restricted to variables with absolutely continuous distributions. In accordance with the approach in [1] - [3] we shall suppose, in this paper, that the population distributions are known. Later reports (referred to in Section 1) will endeavor to examine the effects of incomplete knowledge of population distributions and how this may be, to some extent, allowed for. Extension of methods described in [3] will enable an appreciation of the effects of inaccuracies, in the population distributions used, to be assessed.

3. <u>Censoring on a Specified Variable</u>. We will denote the joint density function of $X_1, X_2, \ldots, X_r$ by

$$p_{X_1,\ldots,X_r}(x_1, \ldots, x_r) = f_{1\ldots r}(x_1, x_2, \ldots, x_r) ,$$

and the conditional density function of $X_{a_1}, \ldots, X_{a_s}$ given $X_{b_1}, \ldots, X_{b_t}$ ($a_i \neq b_j$ for any i,j) by

$$g(x_{a_1}, \ldots, x_{a_s} | x_{b_1}, \ldots, x_{b_t}) .$$

We note that $f_{1,2}(x_1,x_2)=f_1(x_1)g(x_2|x_1)$ etc.

The hypothesis denoting censoring of the s_0 smallest and s_r largest valued - (s_0, s_r) censoring - will be denoted by H_{s_0,s_r}. If it is necessary to specify the variable(s) in respect to which censoring operates we will use symbols of the form $H^{(X)}_{s_0,s_r}$.

We now consider how to test the hypothesis $H^{(X_1)}_{s_0,s_r}$. We have for the joint distribution of $\underset{\sim}{X}_j = (X_{1j}, X_{2j})$ with $X_{11} \le X_{12} \le \ldots \le X_{1r}$ (i.e. ordered in regard to X_1)

$$(1) \quad f(\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_r | H^{(X_1)}_{s_0,s_r}) = \frac{(r+s_0+s_r)!}{s_0!(r-2)!s_r!} \{F_1(x_{11})\}^{s_0}\{1 - F_1(x_{1r})\}^{s_r} \times \prod_{j=1}^{r} f_1(x_{1j})g(x_{2j}|x_{1j}) .$$

where $F_1(t) = \int_{-\infty}^{t} f_1(t)dt$. Testing $H^{(X_1)}_{s_0,s_r}$ against $H_{0,0}$ we have the likelihood ratio

$$(2) \quad \frac{f(\underset{\sim}{x}_1,\underset{\sim}{x}_2,\ldots,\underset{\sim}{x}_r|H^{(X_1)}_{s_0,s_r})}{f(\underset{\sim}{x}_1,\underset{\sim}{x}_2,\ldots,\underset{\sim}{x}_r|H_{0,0})} = \frac{(r+s_0+s_r)!}{s_0!(r-2)!s_r!} \{F_1(x_{11})\}^{s_0}\{1 - F_1(x_{1r})\}^{s_r} .$$

In case (i) of Section 3 therefore we use only the data on the possibly censored variable, so the methods described in [1] - [3] are applicable.

4. <u>Censoring on One of Two Variables</u>. If we suppose (s_0, s_r) fixed the alternative hypothesis to that of no censoring $(H_{0,0})$ is

$$(3) \qquad H^{(X_1)}_{s_0,s_r} \cup H^{(X_2)}_{s_0,s_r} \equiv H^{(X_1 \cup X_2)}_{s_0,s_r} \quad \text{say} .$$

The likelihood approach leads to a critical region of form

$$(4) \quad \max[\{F_1(X_{11})\}^{s_0}\{1 - F_1(X_{1r})\}^{s_r} \; ; \; \{F_2(X_{21})\}^{s_0}\{1 - F_2(X_{2r})\}^{s_r}] \geq C_\alpha$$

where
$$\left.\begin{aligned} X_{h1} &= \min(X'_{h1}, \ldots, X'_{hr}) \\ X_{hr} &= \max(X'_{hr}, \ldots, X'_{hr}) \end{aligned}\right\} (h = 1,2) .$$

We here encounter a difficulty which was not present in the univariate case. Although the distribution of $\{F_h(X_{h1})\}^{s_0}\{1 - F_h(X_{hr})\}^{s_r} = T_h(\underset{\sim}{X})$ $(h = 1,2)$ does not depend on the distributions of X_1 or X_2, the joint distribution of $T_1(\underset{\sim}{X})$ and $T_2(\underset{\sim}{X})$ does, in general, depend on the joint distribution of X_1 and X_2.

In the special case when X_1 and X_2 are independent this is not so, and $T_1(\underset{\sim}{X})$ and $T_2(\underset{\sim}{X})$ have distributions which do not depend on the distributions of X_1 and X_2, and are also independent. Under $H_{0,0}$ their distributions are identical, and we can use the univariate results referred to above, but with significance level equal to $1 - (1 - \alpha)^{\frac{1}{2}}$, where α is the overall significance level required.

When, as we are supposing, the joint density function of X_1 and X_2 is known, it may be possible to find a transformation to two new variates W_1, W_2 which are independent. Although it will no longer be the case that the expected censoring is in regard to either W_1 or W_2 alone, statistics based on $T_1(\underset{\sim}{W})$, $T_2(\underset{\sim}{W})$ may be useful.

In order to discuss the joint distribution of $T_1(\underset{\sim}{X})$ and $T_2(\underset{\sim}{X})$ in the general case, we need to have an expression for the joint distribution of X_{11}, X_{1r}, X_{21}, and X_{2r}. This is usually quite complicated. It is, of course, easy to obtain the joint distribution of X_{11} and X_{1r}, and of X_{21} and X_{2r}. The essential difficulty is that if $X_{11} = X'_{1j}$ then it is not

necessarily the case that $X_{21} = X'_{2j}$.

Some theoretical results useful in the analysis will be found in Section 6, but the matter will not be pursued further here.

5. "Linear Censoring". If possible censoring is in accordance with the values of a variable

$$Y^{(a)} = aX_1 + \sqrt{1 - a^2}\, X_2$$

then if $\underline{a}$ is known, we can use univariate techniques applied to

$$Y_j^{(a)\prime} = aX'_{1j} + \sqrt{1 - a^2}\, X'_{2j} .$$

If a is unknown, it is natural to use Roy's [4] union-intersection method, and to reject the hypothesis $H_{0,0}$ (no censoring) if it is rejected (at a specified significance level, α') for *any* value of a in the range $-1 < a \leq 1$.

Each application of the test requires evaluation of the distribution of $Y^{(a)}$. Given the joint distribution of X_1 and X_2 this is quite feasible. In the case when X_1 and X_2 have a bivariate normal distribution, for instance, $Y^{(a)}$ has a normal distribution with mean $aE[X_1] + \sqrt{1 - a^2}\, E[X_2]$ and variance $a^2\mathrm{var}(X_1) + 2a\sqrt{1 - a^2}\, \mathrm{cov}(X_1, X_2) + (1 - a^2)\mathrm{var}(X_2)$.

Although the calculation of the criterion appears feasible, appropriate significance limits may have to be derived from Monte Carlo experiments.

A similar procedure can be used in estimating $\underline{a}$. If s_0, s_r are known we find the value of $\underline{a}$ for which

$$\{F(Y_1^{(a)})\}^{s_0}\{1 - F(Y_r^{(a)})\}^{s_r}$$

(in an obvious notation) is maximum.

6. Data Available Only for a Variable Related to a Possibly Censored Variable.

The available data $Z_1, Z_2, \ldots, Z_r$ correspond to unrecorded values $X_1', X_2', \ldots, X_r'$ of a variable X which may have been censored. Note that $X_1', X_2', \ldots, X_r'$ are themselves unordered; though they may be the $(s_0 + 1)$-through $(s_0 + r)$-th order statistics (in some order) of a complete random sample of size $(r + s_0 + s_r)$. We use the X-values in this form, because in general the Z-values will not be in the same order as the X's.

The joint density function of the *unordered* values $X_1', \ldots, X_r'$, if there is (s_0, s_r) censoring is

$$\frac{(r+s_0+s_r)!}{s_0!s_r!} \cdot \frac{1}{r!} \left\{F_X\big(\min(x_1, \ldots, x_r)\big)\right\}^{s_0}\left\{1 - F_X\big(\max(x_1, \ldots, x_r)\big)\right\}^{s_r} \prod_{j=1}^{r} f_X(x_j)$$

where $f_X(\cdot)$, $F_X(\cdot)$ are the population density and cumulative distribution functions, respectively, of X'.

We suppose that the density function of Z, given X is $g(z|x)$.

The joint density of $X_1', \ldots, X_r', Z_1', \ldots, Z_r'$ is obtained by multiplying by $\prod_{j=1}^{r} g(z_j|x_j)$, and so the joint density of $Z_1', Z_2', \ldots, Z_r'$ is

$$f_{\underset{\sim}{Z}'}(\underset{\sim}{z}|H_{s_0,s_r}) = \frac{(r+s_0+s_r)!}{s_0!r!s_r!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{F_X\big(\min(x_1, \ldots, x_r)\big)\right\}^{s_0} \cdot$$

$$\cdot \left\{1 - F_X\big(\max(x_1,\ldots,x_r)\big)\right\}^{s_r}\left\{\prod_{j=1}^{r} f_X(x_j)g(z_j|x_j)\right\}dx_1 \ \ldots \ dx_r$$

$$= \frac{(r+s_0+s_r)!}{s_0!r!s_r!} \binom{r}{2} \int_{-\infty}^{\infty}\int_{0}^{x_2} \{F_X(x_1)\}^{s_0}\{1 - F_X(x_2)\}^{s_r}$$

$$\left\{\prod_{j=1}^{2} f(x_j)g(z_j|x_j)\right\} \prod_{j=3}^{r} u(z_j|x_1,x_2)dx_1dx_2$$

where $u(z|x_1,x_2) = \int_{x_1}^{x_2} f_X(x)g(z|x)dx \quad (x_1 < x_2)$. The likelihood ratio is proportional to

$$f_{\underset{\sim}{Z}'}(\underset{\sim}{z}|H_{s_0,s_r})/[\prod_{j=1}^{r} \int_{-\infty}^{\infty} f_X(x_j)g(z_j|x_j)dx_j] = f_{\underset{\sim}{Z}'}(\underset{\sim}{z}|H_{s_0,s_r})/[\prod_{j=1}^{r} f_{Z'}(z_j)] .$$

If Z and X are closely related and the regression of Z on X is monotonic we can see intuitively, that it will be reasonable to use univariate tests based on the order statistics of $Z_1', Z_2', \ldots, Z_r'$. A particular example of such a situation is when there is linear regression of Z on X and high correlation. In cases of this kind it is likely that the rank ordering of the Z's will be very similar to that of the X's , and so censoring of the X's leads to a very similar censoring of the Z's .

The null hypothesis $(H_{0,0})$ distribution of the appropriate statistic based on the Z's is the same as that for the corresponding statistic based on the X's . (For example even though the Z corresponding to the maximum X need not be the maximum Z , the statistics

$$Y_{1(X)} = \int_{-\infty}^{\min(X_1,\ldots,X_r)} f_X(x)dx \quad \text{and} \quad Y_{r(X)} = \int_{-\infty}^{\max(X_1,\ldots,X_r)} f_X(x)dx \text{ ; and}$$

$$Y_{1(Z)} = \int_{-\infty}^{\min(Z_1,\ldots,Z_r)} f_Z(z)dz \quad \text{and} \quad Y_{r(Z)} = \int_{-\infty}^{\max(Z_1,\ldots,Z_r)} f_Z(z)dz \quad \text{have}$$

the same joint distributions.)

Relevant results of [1] - [3] are thus directly applicable and valid tests, based on the Z's , can easily be constructed (even when there is not high correlation between X and Z). There will, however, usually be some loss of power, because the ordering of the Z's will not in general be identical with that of the X's .

7. <u>Multivariate Problems</u>. When there are $m(> 2)$ variables the possible variations on the problems described become much more numerous, but their intrinsic nature remains the same. We distinguish

(i) Possible censoring with regard to one of a specified sub-set of variables (as special cases, the sub-set may be all, or just one, of the variables).

(ii) Possible repeated censoring, first with respect to one, and then (of the remaining observations) with respect to another of the variates.

(iii) Possible censoring with respect to a function (or functions) of the variables. The function may be completely specified, or only its form may be given, with some parameters which need to be estimated. In particular, linear function(s) call for special attention.

(iv) We may have r sets of values of m variates $Z_1, \ldots, Z_m$ which may represent a complete random sample, or a sample which has been censored in respect of an unrecorded variable (or variables).

In many cases, useful practical procedures will be based on uni- or bivariate analyses applied to selected variables, or pairs of variables. For this reason, the problems described in Sections 2 - 6 will be attacked first. Analysis of problems of the kind described in the present section will then be primarily based on appropriate combinations of those developed in [1] - [3] for univariate data with the later developed bivariate techniques. Although "portmanteau" procedures, dealing simultaneously with many variates will be studied, it appears unlikely (by analogy with situations in multivariate analysis of variance, for example) that they will be of great use on their own even if it is technically possible to apply them.

REFERENCES

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

[2] Johnson, N. L. (1973-4) Robustness of certain tests of censoring of extreme values I & II. University of North Carolina, Institute of Statistics Mimeo Series Nos. 866, 940.

[3] Johnson, N. L. (1974) Study of possibly incomplete samples. *Proc. Fifth Brasov Conference* (To appear).

[4] Roy, S. N. (1957) *Some Aspects of Multivariate Analysis*. Wiley, New York.