The Annals of Statistics
1977, Vol. 5, No. 3, 547-553

ON THE COMPLETENESS OF MINIMAL SUFFICIENT STATISTICS WITH CENSORED OBSERVATIONS[1]

BY G. K. BHATTACHARYYA, RICHARD A. JOHNSON AND K. G. MEHROTRA

University of Wisconsin and Syracuse University

The property of completeness, for the minimal sufficient statistics, is investigated in the context of life testing when the set of observations is censored at a fixed time or at a fixed order statistic. Nonparametric families are shown to retain completeness for the observed order statistics and some implications regarding unbiased estimators and similar tests are presented. Most of the common parametric models fail to possess completeness of minimal sufficient statistics under censored sampling in the one-sample, two-sample and regression situations.

1. Introduction. The concept of completeness and its relation to unbiased estimators and similar tests first received careful attention in a basic paper by Lehmann and Scheffé [6]. The completeness property of order statistics ([2], [4], [5]) and other minimal sufficient statistics has been widely used for estimation and tests of hypotheses with complete samples. Our primary concern in this note is with life testing situations where the set of observations is either time censored (Type I) or fixed order statistics censored (Type II). Except for [9], [10], which treat unbiased estimation and completeness for time censored exponential and geometric distributions, no results are available concerning the manner in which the completeness property carries over to the censored case. The purpose of this note is twofold, first to establish the existence or nonexistence of complete sufficient statistics and then to draw important implications regarding point estimators and similar test. This exposition divides naturally into two parts with nonparametric families being studied in the following section and common parametric families in the last. Attention is given to one- and two-sample problems as well as regression models.

It was shown in [8] that the property of completeness is preserved when distributions are truncated so that all observations must fall in a prescribed interval. In marked contrast to the truncated situation, we show that censoring with a parametric family will usually produce minimal sufficient statistics which are not complete. Completeness is preserved, however, in the nonparametric families considered below. Unfortunately, this then precludes the existence of unbiased estimators of many important population characteristics which are otherwise estimable.

Received August 1975; revised September 1976.

[1] This research was sponsored by the Air Force Office of Scientific Research under Grant No. AFOSR 72-2363C.

AMS 1970 *subject classifications.* Primary 62B05, 62N05; Secondary 62G30.

Key words and phrases. Completeness, reliability, sufficiency, censoring, order statistics.

2. Completeness of nonparametric families. Our specific conclusions with regard to completeness of censored observations and unbiased estimation from censored data may be drawn from the following two general results. Theorem A is well known and the proof follows directly from the definition of completeness.

THEOREM A. *Let $\mathscr{P}$ be a family of probability distributions for a real or vector random variable X and let $Y = h(X)$ be a measurable function with $\mathscr{P}^Y$ denoting the induced family of distributions of Y. If X is complete* w.r.t. *the family $\mathscr{P}$, then Y is complete* w.r.t. *$\mathscr{P}^Y$.*

Let X have the probability model $(\mathscr{X}, \mathscr{B}, P)$, $P \in \mathscr{P}$. A parameter $\phi(P)$ is said to be (unbiasedly) estimable from X if there exists a measurable function $T(X)$ such that $E_P[T(X)] = \phi(P)\ \forall\ P \in \mathscr{P}$.

THEOREM B. *Assume that X is complete* w.r.t. *$\mathscr{P}$, ϕ is estimable from X and let $Y = h(X)$ be a measurable function. Let $T(X)$ be the unbiased estimator of ϕ. Then ϕ is estimable from Y if and only if T is a function of Y with probability* 1.

PROOF. The "if" part is trivial. For the "only if" part let $T^*(Y)$ be an unbiased estimator of ϕ based on Y. We then have

$$E_P[T^* \circ h(X)] = \phi(P) \quad \forall\ P \in \mathscr{P}.$$

From the completeness of $\mathscr{P}$, $T^* \circ h(x) = T(x)\ \forall\ x \in \mathscr{X} - N$ where $P(N) = 0$ $\forall\ P \in \mathscr{P}$. Hence, outside a $\mathscr{P}$-null subset of $\mathscr{X}$, T depends on x only through $y = h(x)$.

We now consider ordered observations from a univariate cdf $F \in \mathscr{F}$ where $\mathscr{F}$ is the family of all cdf's absolutely continuous with respect to Lebesgue measure. The order statistics from a random sample of size n from F are denoted by $W_1 \leqq W_2 \leqq \cdots \leqq W_n$. The above theorems may be used to draw some implications based on the fact that the full vector $\mathbf{W} = (W_1, \cdots, W_n)$ of the order statistics is complete w.r.t. $\mathscr{F}$.

CASE (a). *Order statistics censoring.* Consider the procedure which censors all observations after W_r, that is, the observable data consist of $(W_1, \cdots, W_r)$, $1 \leqq r \leqq n$. The induced family of distributions $\mathscr{F}^{(r)}$ of $(W_1, \cdots, W_r)$ is given by the pdf's

$$\frac{n!}{(n-r)!} \prod_{i=1}^{r} f(w_i) \cdot [1 - F(w_r)]^{n-r}, \qquad w_1 < \cdots < w_r;\ F \in \mathscr{F}.$$

By considering the function $h(W_1, \cdots, W_n) = (W_1, \cdots, W_r)$, Theorem A yields:

COROLLARY 2.1. *The vector of uncensored observations $(W_1, \cdots, W_r)$ is complete* w.r.t. *$\mathscr{F}^{(r)}$.*

The same reasoning extends to multiple order statistic censoring which includes left censoring as a special case. For instance, if the censoring is in blocks so that $W_{a+1}, \cdots, W_b$ and $W_{c+1}, \cdots, W_d$ are observed, these order statistics are complete with respect to the corresponding induced family of distributions.

Of interest in life testing, is the question of the existence of an unbiased estimator of the reliability $R_a = P_F[X > a]$, $F \in \mathscr{F}$ based on censored samples. Based on the full set of order staistics $\mathbf{W}^{(n)} = (W_1, \cdots, W_n)$, the unique unbiased estimator of R_a is $\tilde{R}_a = n^{-1}\#\{W_i > a, i = 1, \cdots, n\}$. In order that R_a be estimable from $\mathbf{W}^{(r)} = (W_1, \cdots, W_r)$ with $r < n$, Theorem B requires that $\tilde{R}_a$ must be a function of $\mathbf{W}^{(r)}$. This is not the case since $\tilde{R}_a$ is a symmetric function of $(W_1, \cdots, W_n)$ but $\mathbf{W}^{(r)}$ is not. Consequently, we have:

COROLLARY 2.2. *With fixed a and the observations censored at the r*th *order statistic* $(1 \leqq r < n)$, *there does not exist an unbiased estimator of* $R_a = P_F[X > a]$. *More generally,* R_a *is not estimable whenever any of the n order statistics is censored.*

The following alternative proof explicitly relates the null sets. Based on the full sample $\tilde{R}_a = \#\{W_i > a\}/n$ is an unbiased estimator of R_a. Suppose, for a contradiction, that there exists an unbiased estimator $T(W_1, \cdots, W_r)$. Let $T^*(w_1, \cdots, w_n) \equiv T(w_1, \cdots, w_r)$ and set $N = \{(w_1, \cdots, w_n): T^* \neq \tilde{R}_a\}$, $S_n = \{(w_1, \cdots, w_n): w_n \leqq a\}$ and $S_n^* = \{(w_1, \cdots, w_n): w_r \leqq a, w_{r+1} > a\}$. Then $T^*(\mathbf{w}) = 0$ for $\mathbf{w} \in S_n - N$ and $T^*(\mathbf{w}) = \tilde{R}_a(\mathbf{w}) \neq 0$ on $S_n^* - N$. However, $T^*(\mathbf{w})$ only depends on $(w_1, \cdots, w_r)$ and thus is zero on $S_n^* - N$ which is a contradiction since the n-dimensional Lebesgue measure of S_n^* is infinite while that of N is zero by the completeness with the full sample.

We consider next the two-sample problem where m units receive treatment 1 and n units treatment 2. Suppose that the experiment is terminated at the rth failure so that the observable data are the ordered failure times $X_1 \leqq \cdots \leqq X_{m_r}$; $Y_1 \leqq \cdots \leqq Y_{n_r}$, $m_r + n_r = r$, where the X's are based on a random sample from F and Y's on a sample from G. The reasoning leading to Corollary 2.1 yields:

COROLLARY 2.3. *With r fixed,* $r < (m + n)$, *the set of uncensored order statistics* $X_1 < \cdots < X_{m_r}$; $Y_1 < \cdots < Y_{n_r}$ *is complete with respect to the family of distributions induced by* (F, G), $F \in \mathscr{F}$, $G \in \mathscr{F}$.

Again, the completeness property implies that many of the usual functionals do not have unbiased estimators when any of the observations are censored. For instance, $\Delta = P[X > Y]$ is a measure of the difference in treatment effects. In the context of a stress-strength model, Δ represents the reliability of a component, having a random strength X, when it is subjected to a random stress Y. Nonestimability of Δ from the abovementioned censored samples follows by an application of Theorem B and noting that the unique unbiased estimator $(mn)^{-1}\#\{X_i > Y_j, i = 1, \cdots, m, j = 1, \cdots, n\}$, based on full samples is not a function of the uncensored observations.

In the context of testing hypotheses, in the two sample problem we let $W_1 < \cdots < W_r < \cdots < W_{m+n}$ denote the combined sample order statistics and set $Z_i = 1(0)$ if W_i is an $X(Y)$. We then have the one-to-one correspondence $(\mathbf{X}, \mathbf{Y}) \leftrightarrow (\mathbf{Z}, \mathbf{W})$. Under the null hypothesis $H_0: F = G$, the set of uncensored order statistics $(W_1, \cdots, W_r)$ is sufficient and complete according to Corollary

2.3. Consequently, all similar tests are conditional on $\mathbf{W}^{(r)} = (W_1, \cdots, W_r)$. The conditional distribution of $\mathbf{Z}$ given $\mathbf{W}^{(r)}$ is

$$P_{F,G}[\mathbf{Z} = \mathbf{z} \mid \mathbf{w}^{(r)}] = \frac{\dfrac{m!\, n!}{(m-m_r)!\,(n-n_r)!} \prod_1^r [f(w_i)]^{z_i}[g(w_i)]^{1-z_i}[1-F(w_r)]^{m-m_r}[1-G(w_r)]^{n-n_r}}{\sum_z (\text{numerator})}$$

and it is free of F under H_0. A most powerful similar test against the single alternative $H_1\colon F \neq G$ is then a permutation test which rejects H_0, for

$$\prod_1^r \left[\frac{g(w_i)}{f(w_i)}\right]^{1-z_i} \left[\frac{1-G(w_r)}{1-F(w_r)}\right]^{n-n_r} > c(\mathbf{w}),$$

which depends on F and G. For the exponential scale alternatives $f(x) = \theta_1 \exp(-\theta_1 x)$, $g(y) = \theta_2 \exp(-\theta_2 y)$, this reduces to

$$\left(\frac{\theta_2}{\theta_1}\right)^{n_r} \exp[-(\theta_2 - \theta_1)[\textstyle\sum_1^{n_r} y_i + (n - n_r)w_r]] > c(\mathbf{w})$$

which depends both on θ_1 and θ_2. This is in contrast to the full sample situation where a UMP similar test exists for $H_0\colon F = G$, $F \in \mathscr{F}$, $G \in \mathscr{F}$ against one-sided scale alternatives in the parametric family of exponential distributions.

CASE (b). *Time censoring.* We now consider the censoring scheme where n items with i.i.d. life distribution $F \in \mathscr{F}$ are simultaneously tested until a prescribed time t and the ordered failure times during this interval are recorded. Here $R = \#\{W_i \leqq t, i = 1, \cdots, n\}$ is a random variable. The observable data consist of $(R; W_1, \cdots, W_R)$ whose induced family of distributions $\mathscr{F}_t$ has pdf's

$$\begin{aligned} f(r; w_1, \cdots, w_r) &= [1 - F(t)]^n, \quad r = 0 \\ &= \binom{n}{r} r! \prod_{i=1}^r f(w_i)[1 - F(t)]^{n-r} I_{[w_r < t]}, \quad r \geqq 1. \end{aligned}$$

Using Theorem A and taking the function h as $h(w_1, \cdots, w_n) = (r; w_1, \cdots, w_r)$ where $r = \#\{w_i \leqq t, i = 1, \cdots, n\}$, we have

COROLLARY 2.4. $(R; W_1, \cdots, W_R)$ *is complete* w.r.t. *the induced family of distributions* $\mathscr{F}_t$.

Our next result shows that the reliability may be unbiasedly estimated provided that the censoring time is greater than the mission time a.

These results follow by an application of Theorem B and using an argument similar to that for Corollary 2.2. When $a \leqq t$, the unbiased estimator $\tilde{R}_a$ based on the full sample is indeed a function of $(R; W_1, \cdots, W_R)$

$$\tilde{R}_a = \frac{(n - R)}{n} + \frac{1}{n} \#\{W_i > a, i = 1, \cdots, R\}.$$

However, it is not so for $a > t$.

COROLLARY 2.5. *For a sample censored at time t,*

(i) *if $a \leqq t$, $\#\{W_i > a\}/n$ is a* UMVU *estimator of $R_a = P[X > a]$*
(ii) *if $a > t$, there is no unbiased estimator of R_a.*

An alternative proof, by contradiction, helps establish the completeness of the mixed exponential in the next section. Suppose that $g_R(W_1, \cdots, W_R)$ is an unbiased estimator of $1 - R_a$. Then it is also unbiased for the subclass $\{F_c(x) = cF_0(x), x \leqq t$ and $= cF_0(t) + [1 - cF_0(t)][F_0(x) - F_0(t)][1 - F_0(t)]^{-1}, x > t; 0 < c < 1\}$ where $F_0 \in \mathscr{F}$ and $0 < F_0(t) < 1$. That is,

$$\sum_{r=0}^{n} d_r \binom{n}{r} [cF_0(t)]^r [1 - cF_0(t)]^{n-r} \equiv_c cF_0(t) + [1 - cF_0(t)] \frac{[F_0(a) - F_0(t)]}{1 - F_0(t)}$$

where $d_r = r! \int g_r \prod_1^r f_0(w_i) I_{[w_r \leqq t]} F_0^{-r}(t)\, dw_1 \cdots dw_r$. Writing $p = cF_0(t)$, this becomes $a_n p^n + a_{n-1} p^{n-1} + \cdots + a_1 p + a_0 \equiv_p 0$, $0 < p < F_0(t)$ where $a_0 = d_0 - [F_0(a) - F_0(t)]/[1 - F_0(t)]$. Thus $a_0 = 0$ but d_0 is independent of F_0.

3. Completeness in some parametric families.

Exponential distribution. The negative exponential is perhaps the most extensively employed model for life testing, especially when dealing with censored observations. With order statistics censoring, it is well known that $\sum_{i=1}^{r} W_i + (n - r)W_r$ is a complete sufficient statistic. With time censoring, Torgersen [9] characterized unbiased estimators of 0 and also the functions admitting UMVU estimators. One conclusion is that $\{R, \sum_{i=1}^{R} W_i + (n - R)t\}$ is minimal sufficient but not complete. As a sidelight, it is also interesting to note that if the model is enriched to be of the form $c\theta \exp[-\theta x]$, $0 < x < t$ with $0 < c \leqq 1$, $0 < \theta < \infty$ and arbitrary on $x > t$, then $(R, \sum_{i=1}^{R} W_i)$ is sufficient and complete.

In the following, we consider only order statistic censoring.

Weibull distribution. In the Weibull model with pdf

$$\frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\theta}\right)^{\beta}\right], \qquad x > 0,$$

the minimal sufficient statistic $(W_1, \cdots, W_r)$ is not complete when $r \geqq 3$. To see this consider any bounded function of $\ln(W_1/W_2)/\ln(W_1/W_3)$ and note that the distribution of this variable is free of θ and β.

Type I *extreme value distribution.* One example of a location-scale family where the sufficient statistic does not collapse is the extreme value distribution with cdf

$$1 - \exp\left[-\exp\left(\frac{x - \theta_1}{\theta_2}\right)\right].$$

This distribution is related by a log transformation to the three parameter Weibull. The sufficient statistic $(W_1, \cdots, W_r)$ is not complete for $r \geqq 3$ which follows from consideration of functions of $(W_1 - W_2)/(W_1 - W_3)$. The same method works for many other location and scale families.

Normal distribution. With a censored sample from $N(\mu, \sigma)$, $\{\sum_1^{r-1} W_i, \sum_1^{r-1} W_i^2, W_r\}$ is a minimal sufficient statistic that is not complete for $r \geqq 3$. Consideration of $[W_r - \sum_1^{r-1} W_i/(r-1)]/[\sum_1^r (W_i - W_r)^2]^{\frac{1}{2}}$, whose distribution is free of μ and σ, establishes the lack of completeness. A similar construction can be used to show lack of completeness in a straight line regression model where each replicated point has its own order statistic censoring. The minimal sufficient statistics are not complete if there are observations at three or more points. Moreover, this construction, subtracting the last observation from the average, also extends to one-way ANOVA and replicated factorials. Lack of completeness becomes the norm when censoring occurs.

Exponential regression model. A special case of Cox's proportional hazard model [3] involves a regression of the scale parameter of an exponential life distribution on a concomitant variable x according to the model

$$f(y \mid x) = \lambda e^{\beta x} e^{-\lambda e^{\beta x} y} .$$

For a value x_i of the concomitant variable, suppose n_i items are put on test and the smallest r_i failure times $Y_{i1} < \cdots < Y_{ir_i}$ are observed, $i = 1, \cdots, k$. The likelihood function is then proportional to

$$\lambda^{\sum_1^k r_i} \exp[\beta \textstyle\sum_1^k r_i x_i] \exp[\lambda \textstyle\sum_1^k T_i \exp(\beta x_i)]$$

where $T_i = \sum_{j=1}^{r_i} y_{ij} + (n_i - r_i) y_{ir_i}$. The minimal sufficient statistic $(T_1, \cdots, T_k)$ is not complete for $k > 2$. In order to see this, consider any two x points which are different from $\bar{x}$, say x_1, x_2, and note that the distribution of $(x_1 - \bar{x})^{-1}(U_1 - \bar{U}) - (x_2 - \bar{x})^{-1}(U_2 - \bar{U})$ is free of λ and β where $U_i = \ln T_i$. The parametric models demonstrate the need for further work, along the lines initiated in Bahadur [1] and Linnik [7], regarding inferences when the minimal sufficient statistics are not complete.

REFERENCES

[1] BAHADUR, R. (1957). On unbiased estimators of uniformly minimum variance. *Sankhyā* **18** 211–224.

[2] BELL, C. B., BLACKWELL, DAVID and BREIMAN, L. (1960). On the completeness of order statistics. *Ann. Math. Statist.* **31** 794–797.

[3] COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* 187–220.

[4] FRASER, D. A. S. (1954). Completeness of order statistics. *Canad. J. Math.* **6** 42–45.

[5] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[6] LEHMANN, E. L. and SCHEFFÉ, H. (1950). Completeness similar regions and unbiased estimation. *Sankhyā* **10** 305–340.

[7] LINNIK, JU. (1968). Statistical problems with nuisance parameters. Translations Monographs **20**, Amer. Math. Soc., Providence.

[8] SMITH, W. L. (1957). A note on truncation and sufficient statistics. *Ann. Math. Statist.* **28** 247–252.

[9] TORGERSEN, ERIK N. (1973). Uniformly minimum variance unbiased (UMVU) estimators based on samples from right truncated and right accumulated exponential distributions. Tech. Report. No. 3, Univ. of Oslo.

[10] TORGERSEN, ERIK N. (1973). Uniformly minimum variance (UMVU) estimators based on samples from a right truncated and right accumulated geometric distribution. Tech. Report No. 4, Univ. of Oslo.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 WEST DAYTON STREET
MADISON, WISCONSIN 53706