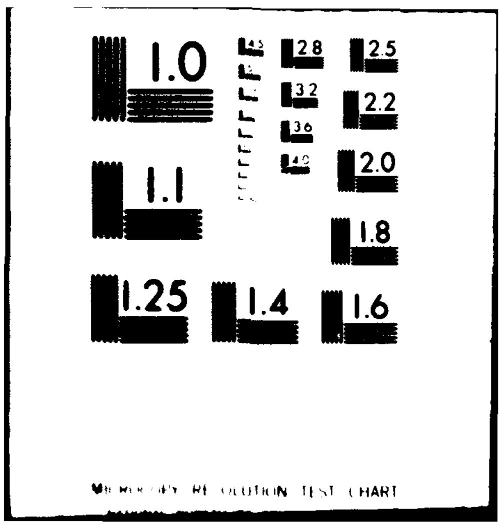
AD-A089 925 SOUTH CAROLINA UNIV COLUMBIA DEPT OF MATHEMATICS AND--ETC F/G 12/1
ESTIMATION OF THE RATIO OF SCALE PARAMETERS IN THE TWO SAMPLE P--ETC(U)
JUN 80 W J PADGETT, L J WEI F49620-79-C-0140

UNCLASSIFIED TR-60 AFOSR-TR-80-0960 NL

1 of 1
AD A
089925

END
DATE
FILMED
11-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6

Department of Mathematics
and Statistics
The University of South Carolina
Columbia, South Carolina 29208

(4)

AFOSR-TR- 80-0960

LEVEL II

AD A089925

F 49620-79-C-0140

80 9 25 083

ESTIMATION OF THE RATIO OF SCALE PARAMETERS IN THE TWO SAMPLE PROBLEM WITH ARBITRARY RIGHT CENSORSHIP*

by

W. J. Padgett and L. J. Wei

University of South Carolina
Statistics Technical Report No. 60
62G05-4

June, 1980

Department of Mathematics and Statistics
University of South Carolina
Columbia, South Carolina 29208

*Research supported by the United States Air Force Office of Scientific Research under Contract No. F49620-79-C-0140.

A

ABSTRACT

A two-sample version of the Cramér-von Mises statistic for right censored observations is used to obtain an estimator of the ratio of scale parameters of two distributions. It is shown that this estimator is consistent. For small samples, simulations are performed which show the superiority of the estimator over the maximum likelihood estimator for exponential distributions.

KEY WORDS: Cramér-von Mises distance; Kaplan-Meier estimators; Right censorship; Scale parameter; Hodges and Lehmann estimator; Consistency.

1. INTRODUCTION

Point estimation for the squared ratio of the scale parameters associated with the class of ranklike statistics has been discussed in many elementary statistics books (for example, Hollander and Wolfe 1973). In this article, we consider a special situation in which the two underlying distributions with positive support only differ by their scale parameters. The quantity in which we are interested is the ratio of these two parameters. A simple and obvious estimator of this quantity based on two independent samples would be the median of all possible ratios of the observations in one sample to those in the other sample. This estimator is essentially a Hodges and Lehmann (1960) estimator (Randles and Wolfe 1979) through a logarithmic transformation.

Unfortunately, in many situations the observations may be censored or truncated. For example, the parameter of interest may be the length of survival. It is common that at the end of the trial there may be incomplete survival information on certain individuals. An (nonparametric) estimator of the ratio of the scale parameters is proposed based on arbitrarily right-censored observations in this article. The proposed estimator, after a logarithmic transformation, is a minimum Cramér-von Mises distance estimator (Fine 1968, Parr 1980) and reduces to the Hodges and Lehmann type of estimator mentioned in the previous paragraph for the uncensored case. The parametric estimation procedure under arbitrary censorship is generally rather complicated. For example, only approximate solutions to the likelihood equations may be obtained for two censored samples from Weibull distributions with the same unknown shape parameter and different scale parameters.

Specifically, in this paper suppose that two positive random variables

S^0 and T^0 differ in distribution only by their scale parameters. That is, there exists a positive constant θ such that θS^0 and T^0 have the same distribution. Let $S_1^0, \ldots, S_m^0$ be independently distributed as S^0, and let $T_1^0, \ldots, T_n^0$ be independent and distributed as T^0 and also independent of $S_1^0, \ldots, S_m^0$. Furthermore, it is assumed that S_i^0 and T_j^0 may be censored from the right by random variables (or constants) U_i' and V_j', respectively, $i=1,\ldots,m$, $j=1,\ldots,n$. We wish to estimate θ based on the observations (s_i,δ_i), $i=1,\ldots,m$, and (t_j,ϵ_j), $j=1,\ldots,n$, where

$$s_i = \min\{s_i^0, u_i'\} \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } s_i = s_i^0 \\ 0 & \text{otherwise,} \end{cases} \qquad i=1,\ldots,m,$$

and

$$t_j = \min\{t_j^0, v_j'\} \quad \text{and} \quad \epsilon_j = \begin{cases} 1 & \text{if } t_j = t_j^0 \\ 0 & \text{otherwise,} \end{cases} \qquad j=1,\ldots,n.$$

In the development of the estimator of θ, we consider the transformation

$$x_i^0 = \ell n\ s_i^0,\ x_i = \ell n\ s_i,\ u_i = \ell n\ u_i', \quad i=1,\ldots,m$$

and

$$y_j^0 = \ell n\ t_j^0,\ y_j = \ell n\ t_j,\ v_j = \ell n\ v_j',\ j=1,\ldots,n.$$

Then the observations are (x_i, δ_i), $i=1,\ldots,m$ and (y_j, ϵ_j), $j=1,\ldots,n$, where

$$\delta_i = \begin{cases} 1 & \text{if } x_i = x_i^0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \epsilon_j = \begin{cases} 1 & \text{if } y_j = y_j^0 \\ 0 & \text{otherwise}. \end{cases}$$

Thus, the problem becomes one of estimating the shift parameter $\Delta = \ell n\ \theta$, based on (x_i, δ_i), $i=1,\ldots,m$, and (y_j, ϵ_j), $j=1,\ldots,n$, where $X_i^0 + \Delta$ has the same distribution as Y_j^0 for each i and j. We denote the distribution function of X_i^0 by F and that of Y_j^0 by G so that $F(y - \Delta) = G(y)$ for all y.

The derivation of the proposed estimator of θ (or Δ) is given in Section 2. The consistency of the estimator is discussed in Section 3. For small samples, the comparison of our estimator with the maximum likelihood estimator of the ratio of the scale parameters of two exponential distributions is made in Section 4 based on the results of computer simulations. The simulations indicate that our estimator is superior to the maximum likelihood estimator even for this "nice" parametric case.

2. THE MINIMUM DISTANCE ESTIMATOR OF $\Delta = \ln \theta$

Without loss of generality we assume that $x_1 \le x_2 \le \ldots \le x_m$ and $y_1 \le y_2 \le \ldots \le y_n$. Define the Kaplan-Meier estimators of $F^c = 1-F$ and $G^c = 1-G$, respectively, by (Efron 1967)

$$\hat{F}_m^c(x) = \begin{cases} \prod_{i=1}^{k-1} \left(\frac{m-i}{m-i+1} \right)^{\delta_i}, & x \in (x_{k-1}, x_k],\ k=2,\ldots,m \\ 1, & x \le x_1 \\ 0, & x > x_m \end{cases}$$

and

$$\hat{G}_n^c(y) = \begin{cases} \prod_{j=1}^{k-1} \left(\frac{n-j}{n-j+1} \right)^{\epsilon_j}, & y \in (y_{k-1}, y_k],\ k=2,\ldots,n \\ 1, & y \le y_1 \\ 0, & y > y_n. \end{cases}$$

Also, define the jumps of $\hat{F}_m^c$ at the x_i's by a_i (if the x_i is censored, the jump is zero), that is,

$$a_i = \begin{cases} \hat{F}_m^c(x_i) - \hat{F}_m^c(x_{i+1}), & i=1,\ldots,m-1 \\ \hat{F}_m^c(x_m), & i=m. \end{cases} \tag{2.1}$$

Similarly, let b_j denote the jump of $\hat{G}_m^c$ at y_j for each $j=1,\ldots,n$.

We define $\hat{\Delta}_{mn}$ to be a value of Δ which minimizes the Cramér-von Mises distance $\int_{-\infty}^{\infty} [\hat{F}_m(t-\Delta) - \hat{G}_n(t)]^2 dt$ where $\hat{F}_m = 1 - \hat{F}_m^c$ and $\hat{G}_n = 1 - \hat{G}_n^c$. Letting A be some real number such that $A > \max\{x_1 + \Delta, \ldots, x_m + \Delta, y_1, \ldots, y_n\}$ and letting I_B be the indicator function of the set B, we can write

$$\int_{-\infty}^{\infty} [\hat{F}_m(t-\Delta) - \hat{G}_n(t)]^2 dt = \int_{-\infty}^{A} [\sum_{i=1}^{m} a_i \, I_{(-\infty,\, t-\Delta]}(x_i)$$

$$- \sum_{j=1}^{n} b_j \, I_{(-\infty, t]}(y_j)]^2 dt. \qquad (2.2)$$

Squaring and collecting terms, (2.2) equals

$$A \, [\sum_{i=1}^{m} \sum_{k=1}^{m} a_i a_k - 2 \sum_{i=1}^{m} \sum_{k=1}^{n} a_i b_j + \sum_{j=1}^{n} \sum_{\ell=1}^{n} b_j b_\ell]$$

$$- \sum_{i=1}^{m} \sum_{k=1}^{m} a_i a_k \max\{x_i + \Delta, x_k + \Delta\}$$

$$- \sum_{j=1}^{n} \sum_{\ell=1}^{n} b_j b_\ell \max\{y_j, y_\ell\} + 2 \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \max\{x_i + \Delta, y_j\}. \qquad (2.3)$$

The first term in (2.3) is zero, and the third term does not involve Δ. Hence, the problem is to minimize over Δ the expression

$$2 \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \max\{x_i + \Delta, y_j\} - \sum_{i=1}^{m} \sum_{k=1}^{m} a_i a_k \max\{x_i + \Delta, x_k + \Delta\}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j (x_i + y_j) + \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j |y_j - x_i - \Delta|.$$

Therefore, we wish to find

$$\min_{\Delta} \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j |y_j - x_i - \Delta|. \qquad (2.4)$$

Since $\sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j = 1$, it is easy to see that a value $\hat{\Delta}_{mn}$ of Δ which solves (2.4) is a <u>median</u> of the (discrete) probability distribution function H_{mn} whose probability density function is defined by

$$h_{mn}(v) = \begin{cases} a_i b_j, \ v = y_j - x_i, & i=1,\dots,m \\ & j=1,\dots,n \\ 0, \quad \text{otherwise.} & \end{cases} \tag{2.5}$$

Therefore, the corresponding estimator of θ is $\hat{\theta}_{mn} = \exp(\hat{\Delta}_{mn})$.

Note that if the two samples $X_1,\dots,X_m$ and $Y_1,\dots,Y_n$ are not censored, the median of the distribution (2.5) which solves (2.4) is exactly the median of all the differences $y_j - x_i$ since $a_i = \frac{1}{m}$ and $b_j = \frac{1}{n}$ for all i and j (Fine 1968).

It is also interesting to note that $1-H_{mn}$ can be represented as the convolution of the two Kaplan-Meier estimators of F^c and G^c, that is,

$$H_{mn}(v) = \int_{-\infty}^{\infty} \hat{F}_m^c(t-v)d\, \hat{G}_n(t) = -\int_{-\infty}^{\infty} \hat{F}_m^c(t-v)d\, \hat{G}_n^c(t).$$

This representation of H_{mn} will be used in the next section to show that $\hat{\Delta}_{mn}$ is a consistant estimator of Δ.

3. CONSISTENCY OF THE ESTIMATOR

For mathematical convenience we assume that the censoring variables $\{U_i\}$ and $\{V_j\}$ are independent random variables with common distribution functions J and K, respectively. Also, the U_i and V_j are assumed to be independent of the X_i^0 and Y_j^0 and independent of each other.

For any real number v, it follows from Theorem 8.2 of Efron (1967) that

$$H_{mn}(v) = -\int_{-\infty}^{\infty} \hat{F}_m^c(x-v)d\, \hat{G}_n^c(x) \xrightarrow{P} P[X_i^0 + v \geq Y_j^0] = H(v) \tag{3.1}$$

as $m, n \to \infty$ so that $\frac{m}{m+n} \to \lambda$ and $\frac{n}{m+n} \to 1-\lambda$, $0<\lambda<1$, where H is the distribution function of $Y_j^0 - X_i^0$, which is symmetric about Δ.

We now prove the following theorem.

THEOREM 3.1 Let $m,n \to \infty$ so that $\frac{m}{m+n} \to \lambda$ and $\frac{n}{m+n} \to 1-\lambda$, $0<\lambda<1$. Then $\hat{\Delta}_{mn} \xrightarrow{P} \Delta$.

PROOF. Let $\epsilon > 0$ and $\epsilon' = \epsilon/2$. Since Δ is the unique median of H, we have $H(\Delta - \epsilon') < \frac{1}{2}$. Since $\hat{\Delta}_{mn}$ is a median of H_{mn}, $H_{mn}(\hat{\Delta}_{mn} + \epsilon') \geq \frac{1}{2}$. By (3.1), $H_{mn}(\Delta-\epsilon') \xrightarrow{P} H(\Delta-\epsilon')$ as $m, n \to \infty$. But the event that $\hat{\Delta}_{mn} + \epsilon' < \Delta - \epsilon'$ implies the event that $\frac{1}{2} \leq H_{mn}(\Delta - \epsilon')$. Thus, $P(\hat{\Delta}_{mn} + \epsilon' < \Delta - \epsilon') \leq P(\frac{1}{2} \leq H_{mn}(\Delta - \epsilon')) \to 0$ as $m, n \to \infty$ which implies that $P(\hat{\Delta}_{mn} \geq \Delta - \epsilon) \to 1$.

On the other hand, $H(\Delta + \epsilon) > \frac{1}{2}$ and $H_{mn}(\Delta + \epsilon) \xrightarrow{P} H(\Delta + \epsilon)$ as $m, n \to \infty$ under the conditions of the theorem. Also, $H_{mn}(\hat{\Delta}_{mn}) \leq \frac{1}{2}$, and since $\hat{\Delta}_{mn} > \Delta + \epsilon$ implies that $\frac{1}{2} \geq H_{mn}(\Delta + \epsilon)$, we have $P(\hat{\Delta}_{mn} > \Delta + \epsilon) \leq P(\frac{1}{2} \geq H_{mn}(\Delta + \epsilon)) \to 0$ as $m, n \to \infty$.

Thus, $$P(\Delta - \epsilon \leq \hat{\Delta}_{mn} \leq \Delta + \epsilon) \to 1 \text{ as } m, n \to \infty. \qquad ///$$

Therefore, the estimator $\hat{\theta}_{mn}$ of the ratio of scale parameters for the original problem is a consistent estimator of θ.

4. SMALL SAMPLE COMPARISONS

In addition to the nonparametric nature, an obvious advantage of the estimator $\hat{\theta}_{mn}$ (or $\hat{\Delta}_{mn}$) is that it is easily computed in closed form, whereas for arbitrarily right censored data in many parametric models the scale parameter cannot be estimated in closed form by maximum likelihood or other procedures.

One parametric case which yields a closed form maximum likelihood estimate of the ratio of the two scale parameters for censored observations is the exponential distributions. In this section we will compare our estimator $\hat{\theta}_{mn}$ with the mle for small size censored samples from two exponential distributions with different scale parameters through computer simulations.

In the notation of Section 1, let S_i^0, $i=1,\ldots,m$ and T_j^0, $j=1,\ldots,n$ be samples from the exponential densities $f^0(s) = \lambda^{-1} \exp(-s/\lambda)\, I_{[0,\infty)}(s)$ and $g^0(t) = (\lambda\theta)^{-1} \exp(-t/\lambda\theta)\, I_{[0,\infty)}(t)$, respectively. The maximum likelihood estimator of θ is found to be $\tilde{\theta}_{mn} = \sum_{j=1}^{n} t_j \sum_{i=1}^{m} \delta_i \,/\, (\sum_{i=1}^{m} s_i \sum_{j=1}^{n} \epsilon_j)$, which is the ratio of the two estimates of average lifetime from the respective censored samples.

In order to compare the mle $\tilde{\theta}_{mn}$ with our estimator $\hat{\theta}_{mn}$ of θ, we have performed Monte Carlo simulations to indicate the bias and mean squared error of each estimator. Taking $\lambda = 1$ in f^0 and g^0 for all the computations, the censoring variables U_i', $i=1,\ldots,m$ and V_j', $j=1,\ldots,n$ described in Section 1 were taken to be independent uniformly distributed random variables on the intervals $[0,s_\xi]$ and $[0,t_\eta]$, respectively, where s_ξ was the 100ξ th percentile of f^0 and t_η was the 100η th percentile of g^0 $(0 < \xi, \eta < 1)$. For $\xi = 0.9$, for example, slightly more than 10% of the S_i's will be censored.

For each fixed set of values for m,n,θ,ξ, and η, 1000 samples $(s_1,\delta_1), \ldots, (s_m,\delta_m)$ and $(t_1,\epsilon_1), \ldots, (t_n,\epsilon_n)$ were generated and $\hat{\theta}_{mn}$ and $\tilde{\theta}_{mn}$ were calculated for each. The average and estimated mean squared errors (mse) were calculated for the 1000 repetitions. The resulting estimated biases and mean squared errors for several cases are reported in Table 1.

TABLE 1. Small Sample Comparison for Exponential Distributions

m	n	θ	ξ	η	$\hat{\theta}_{mn}$ bias	$\hat{\theta}_{mn}$ mse	$\tilde{\theta}_{mn}$ bias	$\tilde{\theta}_{mn}$ mse
			0.90	0.50	-0.073	0.182	0.313	0.851
5	10	0.5	0.75	0.75	0.189	0.345	0.319	1.058
			0.50	0.50	0.174	0.226	0.252	0.693
			0.90	0.90	0.080	0.133	0.123	0.245
10	10	0.5	0.90	0.50	-0.163	0.069	0.227	0.476
			0.75	0.75	0.060	0.111	0.181	0.487
			0.50	0.50	0.024	0.039	0.252	0.693
			0.90	0.90	0.160	0.531	0.245	0.978
10	10	1.0	0.90	0.50	-0.326	0.276	0.454	1.903
			0.75	0.75	0.122	0.445	0.363	1.947
			0.50	0.50	0.048	0.157	0.505	2.772
			0.90	0.90	0.320	2.123	0.490	3.914
10	10	2.0	0.90	0.50	-0.653	1.104	0.909	7.613
			0.75	0.75	0.242	1.780	0.726	7.787
			0.50	0.50	0.097	0.628	1.009	11.087
			0.90	0.90	0.052	0.067	0.081	0.102
15	15	0.5	0.90	0.50	-0.162	0.051	0.185	0.292
			0.75	0.75	0.038	0.047	0.094	0.129
			0.50	0.50	0.016	0.017	0.171	0.327
15	15	2.0	0.90	0.90	0.209	1.069	0.325	1.635
			0.90	0.50	-0.647	0.810	0.741	4.667
15	10	2.0	0.90	0.90	0.119	1.035	0.324	2.223
			0.90	0.50	-0.750	0.872	0.672	4.366

These simulations indicate some clear patterns in the behavior of $\hat{\theta}_{mn}$. As m and n increase for fixed θ, ξ, and η, the bias and mean squared error tend to decrease for both $\hat{\theta}_{mn}$ and $\tilde{\theta}_{mn}$ (since these estimators are consistent). As the censoring in the two samples becomes more severe for fixed m,n, and θ, the bias and mse tend to <u>decrease</u> for $\hat{\theta}_{mn}$ and <u>increase</u> for the mle. Also, for fixed m,n,ξ, and η, the bias and mse of both estimators tend to increase as θ increases, more so for the mle than for $\hat{\theta}_{mn}$. Therefore, even for the exponential distributions when both estimates are easily calculated from small samples our estimator is always superior to the maximum likelihood estimator in performance.

5. SUMMARY AND CONCLUSIONS

An estimator $\hat{\theta}_{mn}$ of the ratio of scale parameters has been developed for the two sample problem when the observations are arbitrarily censored from the right. The estimator was obtained by logarithmically transforming the problem to one of estimating a shift parameter by minimizing a Cramér-von Mises measure of the distance between two Kaplan-Meier estimates of the underlying distribution functions. This approach was taken since the minimum distance estimation does not seem to yield a closed form solution for the original scale problem.

The estimator $\hat{\theta}_{mn}$ is shown to be consistent, and Monte Carlo simulation results for two exponential distributions with different scale parameters indicate that the estimator is superior in performance to the maximum likelihood estimator for that case.

REFERENCES

Efron, Bradley (1967), "The Two Sample Problem with Censored Data," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 4, 831-853.

Fine, Terrence (1968), "On the Hodges and Lehmann Shift Estimator in the Two Sample Problem," Annals of Mathematical Statistics, 39, 1814-1818.

Hodges, J. L., Jr., and Lehmann, E. L. (1960), "Comparison of the Normal Scores and Wilcoxon Tests," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 307-317.

Hollander, Myles, and Wolfe, Douglas A. (1973), Nonparametric Statistical Methods, John Wiley & Sons, New York.

Parr, William C. (1980). "Minimum Distance Estimation: A Bibliography," Technical Report, Institute of Statistics, Texas A&M Univ.

Randles, R. H., and Wolfe, D. A. (1979), Introduction to the Theory of Nonparametric Statistics, John Wiley & Sons, New York.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE — READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 80-0960
2. GOVT ACCESSION NO.: AD-A089925
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): ESTIMATION OF THE RATIO OF SCALE PARAMETERS IN THE TWO SAMPLE PROBLEM WITH ARBITRARY RIGHT CENSORSHIP
5. TYPE OF REPORT & PERIOD COVERED: Interim rept.
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): W. J. Padgett and L. J. Wei
8. CONTRACT OR GRANT NUMBER(s): F49620-79-C-0140
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of South Carolina, Dept. of Mathematics, Computer Science & Statistics, Columbia, South Carolina 29208
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, D.C. 20332
12. REPORT DATE: Jun[redacted]980
13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): TR-60
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Cramér-von Mises distance; Kaplan-Meier estimators; Right censorship; Scale parameter; Hodges and Lehmann estimator; Consistency.
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

A two-sample version of the Cramér-von Mises statistic for right censored observations is used to obtain an estimator of the ratio of scale parameters of two distributions. It is shown that this estimator is consistent. For small samples, simulations are performed which show the superiority of the estimator over the maximum likelihood estimator for exponential distributions.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

410442 JOB

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)