Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 43–50

A Comparison of CAN and UMVU Estimators in Inliers-Prone Distributions

K. Muralidharan
Department of Statistics, Faculty of Science
The Maharaja Sayajirao University of Baroda, Vadodara, 390 002 India

Received: 18 June 2023; Revised: 30 June 2023; Accepted: 05 July 2023

Abstract

Among various classical estimations procedures, a relatively better estimation is provided by Consistent Asymptotic Estimators (CAN) method. The method of CAN provides estimators for parametric functions of regular and non-regular or degenerate families of distributions. In this article, we present CAN estimators for parametric functions of inlier-prone (a case of degenerate) distribution models. The estimates are also compared numerically.

Key words: Consistency; Degenerate family of distributions; Inlier-prone models; Minimum variance unbiased estimators.

1. Introduction

In statistical estimation theory, one starts with the data $(x_1, x_2, \dots, x_n)$ of a random variable X, which are assumed to be independent and identically distributed with a common probability distribution $f(x, \theta)$ characterized by an unknown population parameter $\theta \in \Omega$, where θ can be real-valued scalar or vector. The objective is to propose a best inference for θ or $\psi(\theta)$ which satisfies good statistical properties. If the probability model is uniquely defined, one can suggest suitable estimators for the parameter or parametric functions explicitly. Let $T = T(x_1, x_2, \dots, x_n)$ be an estimator of θ based on the observed sample values $(x_1, x_2, \dots, x_n)$. By using the techniques of transformation or form the basic principles of distribution theory, one could, at least theoretically, obtain the sampling distribution of the estimator T and thus begin the inference of the population parameter θ.

There are many criteria and procedures available for deciding the best estimator for θ or $\psi(\theta)$ in Statistics literature. The best estimator in a statistical sense is decided based on a comparison of the variance or mean square error (MSE) of the estimator of one method over the other. For this we assume that T, a real-valued statistic, is to be used as an estimator of real parameter θ based on a random sample of size n from $\{f(x, \theta), \theta \in \Omega\}$, $\Omega \subset R^1$. One of the criteria based on a large sample size is the consistency of an estimator.

Definition 1: An estimator T_n is said to be consistent for θ if $T_n \to \theta$ for each $\theta \in \Omega$ in probability and the convergence in probability is taken under the distribution indexed by θ.

Corresponding Author: K. Muralidharan
Email: lmv_murali@yahoo.com
This paper is based on the B.K. Kale Memorial Endowment Lecture "A comparison of CAN and UMVU Estimators in inliers-prone distributions" delivered by the author on 15 February 2023 during the conference.

A very important property of a consistent estimator is the invariance under continuous transformation, a property not enjoyed by an unbiased estimator. Thus, if $\psi(\theta)$ is a continuous function and if T is consistent for θ, then the invariance property says that $\psi(T)$ is consistent for $\psi(\theta)$. Because of the invariance property of consistent estimators, for all practical purposes one need to consider consistent estimators of θ only for further study of the estimators. The invariance property can be extended to the case of vector valued T and θ as follows:

Definition 2: Let T be jointly consistent for θ and let ψ be k-dimensional continuous functions from Ω to R^k, then $\psi(T)$ is jointly consistent for $\psi(\theta)$ (Kale and Muralidharan, 2015).

To choose between consistent estimators one can compare the MSE's of the estimators, where MSE is defined as $MSE(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$, where $\hat{\theta}=T(x)$ is the unbiased estimate of θ. For instance, if T_1 and T_2 are both consistent for θ then we would prefer T_1 to T_2 if $MSE(T_1) \leq MSE(T_2)$, $\forall \theta \epsilon \Omega$. This comparison generally results into the comparison of the sample sizes of the two estimators. Thus, if T_1 is preferred over T_2 then by Tchebychev inequality it follows that $P[|T_1 - \theta| < \epsilon]$ converges to unity faster than $P[|T_2 - \theta| < \varepsilon] \to 1$ as $n \to \infty, \forall \theta \epsilon \Omega$ and $\varepsilon > 0$. For large n, it is easy to show that $a_n(T - \theta) \to N(0, \sigma_T^2(\theta)$ or $T \sim AN\left(0, \frac{\sigma_T^2(\theta)}{a_n^2}\right)$, where a_n is the blow-up factor (Kale and Muralidharnan, 2015). Such an estimator is called Consistent Asymptotic Normal or CAN estimator. As discussed above, if $\psi(\theta)$ is a continues differentiable function then according to invariance property of consistent estimators the CAN estimator for $\psi(\theta)$ is defined as follows:

Definition 3: Let T be CAN for θ so that $T \sim AN\left(\theta, \frac{\sigma_T^2(\theta)}{a_n^2}\right)$ and let ψ be differentiable functions such that $\frac{d\psi}{d\theta}$ is continuous and nonvanishing then $\psi(T)$ is CAN for $\psi(\theta)$ and $\psi(T) \sim AN\left(\psi(\theta), \sigma_T^2(\theta)\left(\frac{d\psi}{d\theta}\right)^2 / a_n^2\right)$ (Kale and Muralidharan, 2015).

We now propose CAN estimators for parametric functions by considering a family of distributions which are degenerated at some random point. This degeneracy may be due to the occurrence of *instantaneous* or *early* failures together known as inliers are usually seen in life testing experiments. In the instantaneous failure cases, the random variable will have discrete probability mass at the origin (that is lifetime will be zero) and some positive lifetimes, and in the early failure case the failure times may be small in relation to other lifetimes. For modeling positive lifetimes, we have used exponential distribution, as it has been widely used as a model in areas ranging from studies on the lifetimes of manufactured items to research involving survival or remission times in chronic diseases. The exponential distribution has the pdf

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \; x \geq 0 \qquad (1)$$

The maximum likelihood estimator of θ is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The desirable properties of $\hat{\theta}$ are numerous. In particular $\hat{\theta}$ is exactly distributed as $\left(\frac{\theta}{2n}\right)\chi^2_{(2n)}$ and it is a sufficient, efficient, and minimum variance estimator of θ.

The article is organized as follows: The model presentations along with some distributional results are given in Section 2. Along with the CAN estimation, we also propose uniformly minimum variance unbiased estimate (UMVUE) for various parametric functions in Section 3. The numerical illustration is presented in the last section.

2. Inliers-prone model

If the underlying distribution is exponential as given in (1.1), then the inliers-prone model with instantaneous failures is shown as

$$g(x; p, \theta) = \begin{cases} 1-p, & x = 0 \\ \frac{p}{\theta} e^{-\frac{x}{\theta}}, & x > 0 \end{cases} \tag{2}$$

Let $X_1, X_2, \dots, X_n$ be a random sample from (2) then the pdf of X_i is

$$g(x_i; p, \theta) = \begin{cases} (1-p)^{I(x_i)} \left(\frac{p}{\theta} e^{-\frac{x}{\theta}}\right)^{1-I(x_i)} & x_i \geq 0, 0 < p \leq 1, \theta > 0, i = 1, 2, \dots, n \\ 0, & o.w. \end{cases}$$

where,

$$I(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{o.w.} \end{cases} \tag{3}$$

Aitchison (1955) had proposed various unbiased functions for parametric function in (3). Kale and Muralidharan (2000) were the first authors to introduce the term inliers in connection with the estimation of (p, θ) of early failure model with modified failure time distribution (FTD) being (1) with mean θ. A similar problem was attempted by Lai et al. (2007), wherein they have defined nearly instantaneous through the sample configurations, considering Weibull as the underlying FTD. For a detailed review of inliers prone models and their inferences, refer to Muralidharan (2010).

If $p = P(x > 0)$ and further, if we denote $\sum_{i=1}^{n} I(x_i) = n - r$, where r is number of positive observations, then the joint pdf is given by

$$g(\underline{x}; p, \theta) = \begin{cases} (1-p)^{n-r} \left(\frac{p}{\theta}\right)^r e^{-\frac{1}{\theta}\sum_{i=1}^{n}\left(1-I(x_i)\right)x_i}, & x_i \geq 0, r = 0, 1, \dots, n \\ 0, & o.w. \end{cases} \tag{4}$$

The following results are now obvious:

Result 1: The joint density function given in (4) is a two-parameter exponential family of distribution.

Result 2: $\left(\sum I(x_i), \sum\left(1 - I(x_i)\right)x_i\right)$ are jointly sufficient for p and θ.

Result 3: The MLE of p and θ are respectively given by $\hat{p}_{MLE} = \frac{r}{n}$ and $\hat{\theta}_{MLE} = \frac{1}{r}\sum_{x_i>0} x_i$.

Result 4: $\left(\hat{p}_{MLE}, \hat{\theta}_{MLE}\right)' \sim AN^{(2)}\left[(p,\theta)',\ diag\left(\frac{p(1-p)}{n}, \frac{\theta^2}{np}\right)\right]$.

Result 5: The parameters p and θ are orthogonal.

Result 6: The true reliability or survival function for the model at time t is given by

$$S(t) = pe^{-\frac{t}{\theta}}, t > 0, \theta > 0$$

Result 7: $g_{Z|R}(z;\theta|r) = \begin{cases} \frac{e^{-\frac{z}{\theta}} z^{r-1}}{\Gamma r\ \theta^r}, & z > 0, r > 0 \\ 1, & z = 0, r = 0 \end{cases}$,

where $z = \sum_{i=1}^{n}[1 - I(x_i)]x_i \left(= \sum_{x_i>0} x_i\right)$.

3. UMVUE and CAN estimators

It is observed that, obtaining conditional distribution given the sufficient statistics is bit difficult in the above model. Therefore, we use exponential family approach to study the distributional properties.

The equation (4) is written as

$$g(x;p,\theta) = \frac{\left[e^{-\frac{1}{\theta}}\right]^{\left(1-I(x)\right)d(x)} \left[\frac{\theta(1-p)}{p}\right]^{I(x)}}{\left(\frac{\theta}{p}\right)}$$

$$= [a(x)]^{\left(1-I(x)\right)}[h(\theta)]^{\left(1-I(x)\right)d(x)} \left[\frac{g(\theta)(1-p)}{p}\right]^{I(x)} \left(\frac{g(\theta)}{p}\right)^{-1} \quad (5)$$

where $a(x) = 1,\ h(\theta) = e^{-\frac{1}{\theta}},\ d(x) = x,\ g(\theta) = \theta$. The density in (5) is so obtained is defined with respect to measure $\mu(x)$ which is the sum of Lebesgue measure over $(0, \infty)$ and a singular measure at $\{0\}$, is a well-known form of two parameter exponential family with natural parameters $(\eta_1, \eta_2) = \left(log\left(\frac{\theta(1-p)}{p}\right), log\left(e^{-\frac{1}{\theta}}\right)\right)$ generated by the underlying indexing parameters (p, θ). Here $\left(I(x), \left(1 - I(x)\right)x\right)$ is jointly minimal sufficient for (p, θ) as $I(x)$ and $(1 - I(x))x$ do not satisfy any linear restriction. Hence the natural parameter space is convex set in E_2 containing a two-dimensional rectangle making (5) a full rank family. The statistic $\left(I(x), \left(1 - \mathrm{I(x)}\right)x\right)$ is thus complete (Lehmann and Casella, 1998, p 42). Kale and Muralidharan (2000) considered the above mixture and obtained optimal estimating equation for θ ignoring p in the case of exponential failure time distribution.

Further, if we denote $z = \sum_{i=1}^{n}[1 - I(x_i)]x_i \left(= \sum_{x_{i>0}} x_i\right)$, then the joint density function can be expressed as

$$g(\underline{x}; p, \theta) = \binom{n}{r}(1-p)^{n-r}\left(\frac{p}{\theta}\right)^{r} e^{-\frac{z}{\theta}} \tag{6}$$

Hence $(n - R, Z)$ are jointly complete sufficient for (p, θ). Also, the variable $(Z|R = r, r > 0)$ is distributed as a Gamma random variable with parameter (r, θ). Since, $n - R$ is binomial which is same as that of R with parameter (n, p), the joint distribution of $(n - R, Z)$ is

$$\begin{aligned} g(z, n-r; p, \theta) &= P(n - R = n - r)\, g(z; \theta|n - r) \\ &= P(R = r)\, g(z; \theta|r) \\ &= \binom{n}{r}(1-p)^{n-r}p^{r}\frac{1}{\Gamma r\, \theta^{r}} z^{r-1}e^{-\frac{z}{\theta}} \end{aligned}$$

$$= \begin{cases} (1-p)^{n}, & z = 0;\ r = 0 \\ \binom{n}{r}\dfrac{z^{r}}{\Gamma r}e^{-\frac{z}{\theta}}\left(\dfrac{\theta(1-p)}{p}\right)^{n-r}\left(\dfrac{\theta}{p}\right)^{-n}, & z > 0;\ r > 0 \end{cases}$$

$$= \begin{cases} (1-p)^{n}, & z = 0;\ r = 0 \\ B(z, r, n)[h(\theta)]^{z}\left[\dfrac{g(\theta)(1-p)}{p}\right]^{n-r}\left(\dfrac{g(\theta)}{p}\right)^{-n}, & z > 0;\ r > 0 \end{cases} \tag{7}$$

where

$$B(z, r, n) = \begin{cases} 1, & z = 0; r = 0 \\ \binom{n}{r} B(z|r), & z > 0; r > 0 \end{cases} \tag{8}$$

is such that $(1-p)^{n} + \sum_{r=1}^{n}\int_{z>0}\binom{n}{r}B(z|r)\left[e^{-\frac{1}{\theta}}\right]^{z}\left(\frac{\theta(1-p)}{p}\right)^{n-r}\left(\frac{\theta}{p}\right)^{-n} dz = 1$ and $B(z|r) = \frac{z^{r-1}}{\Gamma r}$. Following Roy and Mitra (1957) and Jani and Singh (1995), it is possible to obtain the uniformly minimum variance unbiased estimates (UMVUE) for some parametric functions. Note that, the UMVUE's of parametric function $\phi(p, \theta)$exits if and only if $\phi(p, \theta)$ can be expressed in the form

$$\phi(p, \theta) = \alpha(0,0,n)(1-p)^{n} + \sum_{r=1}^{n}\int_{z>0}\frac{\alpha(z,r,n)e^{-\frac{z}{\theta}}\left(\frac{\theta(1-p)}{p}\right)^{n-r}}{\left[\frac{\theta}{p}\right]^{n}} dz.$$

Below we consider some estimates for the parametric functions:

Result 8: For $m \leq n$, the UMVUE of $(1-p)^{m}$ is $G_m(Z, R, n)$ as given by

$$G_m(z, r, n) = \begin{cases} \dfrac{\binom{n-m}{r}}{\binom{n}{r}}, r = 0,1, \ldots, n-m \\ 0, \quad o.w. \end{cases}$$

Result 9: *For* $m = 1$, Result 8 reduces to the UMVUE of $(1-p)$ as

$$G_1(z,r,n)=\begin{cases}\dfrac{n-r}{n}, & r>0; z>0\\ 1, & r=0, z=0\end{cases}$$

Result 10: $\psi(T_1)=(1-\frac{r}{n})^m$ is CAN estimator of $\psi(p)=(1-p)^m$ with asymptotic variance

$$\frac{m^2}{n}p(1-p)^{2m-1}.$$

Result 11: For $m\le\frac{n}{2}$, the UMVUE of the variance of $G_m(Z,R,n)$ is computed as

$$\widehat{var}[G_m(z,r,n)]=\begin{cases}G_m^2(z,r,n)-G_{2m}(z,r,n), & r=1,2,\dots,(n-2m)\\ G_m^2(z,r,n), & r=(n-2m+1),\dots,(n-m)\\ 0, & otherwise\end{cases}$$

$$=\begin{cases}\left[\dfrac{\binom{n-m}{r}}{\binom{n}{r}}\right]^2-\dfrac{\binom{n-2m}{r}}{\binom{n}{r}}, & r=1,2,\dots,(n-2m)\\ \left[\dfrac{\binom{n-m}{r}}{\binom{n}{r}}\right]^2, & r=(n-2m+1),\dots,(n-m)\\ 0, & o.w.\end{cases}$$

Result 12: For $m=1$, the UMVUE of the variance of UMVUE of $(1-p)$ is given by

$$\widehat{var}[G_1(z,r,n)]=\begin{cases}\dfrac{r(n-r)}{n^2(n-1)}, & r=1,2,\dots,(n-1)\\ 0, & o.w.\end{cases}$$

Result 13: For $k>0$ the UMVUE of parametric function $(1-p)^n+\left(\frac{p}{\theta}\right)^k[1-(1-p)^{n-k}]$ is given by

$$H_k(z,r,n)=\begin{cases}\dfrac{(r)_k(r-1)_k}{(n)_k z^k}, & r=1,2,\dots,n;\ z>0\\ 1, & r=0;\ z=0\end{cases}$$

where $(a)_k=a(a-1)\dots(a-k+1)$, and $z=\sum_{x_{i>0}}x_i$.

For various values of $k\ge1$, one can obtain the UMVUE of the parametric function. Unfortunately, it is impossible to find a unbiased estimate for the parameter θ alone. Aitchison (1955) through the usual classical approach obtain the UMVUE of the parametric function $(1-p)^2\theta^2$ as

$$\varphi(z,r,n)=\begin{cases}\frac{(2n-r-1)z^2}{n(n-1)(r+1)}, & r>0;\ z>0\\ 0, & r=0;\ z=0\end{cases}$$

Result 15: $\psi(T_2) = (\sum_{x_i>0} x_i)^m$ is CAN estimator of $\psi(\theta) = \theta^m$ with asymptotic variance $\frac{m^2\theta^{2m}}{np}$.

Result 16: For fixed x, the UMVUE of pdf $g(x; p, \theta)$ is shown as

$$\phi_x(z,r,n) = \begin{cases} \frac{r(r-1)}{nz}\left(1+\frac{x}{z}\right)^{r-2}, & 0 < x < z;\ r = 1,2,\dots,n \\ \frac{n-r}{n}, & x = 0;\ r = 0,1,\dots,n-1 \\ 0, & o.w. \end{cases}$$

Result 17: For $r = n$, that is when all the observations are coming from the density, then the UMVUE of the density $f(x;\theta)$ is simplified as

$$\phi_x(z,r,n) = \begin{cases} \frac{n-1}{z}\left(1+\frac{x}{z}\right)^{n-2}, & 0 < x < z;\ n > 1 \\ 0, & o.w. \end{cases}$$

Result 18: For fixed x, the UMVUE of variance of pdf $g(x; p, \theta)$ is obtained as

$\widehat{var}[\phi_x(z,r,n)]$

$$= \begin{cases} \left[\frac{r(r-1)}{nz}\left(1-\frac{x}{z}\right)^{r-2}\right]^2 \\ -\frac{r(r-1)^2(r-2)}{n(n-1)z(z-x)}\left(1-\frac{x}{z}\right)^{r-2}\left(1-\frac{x}{z-x}\right)^{r-3}, & 0 < x < z;\ r = 2\dots,n \\ \left[\frac{r(r-1)}{nz}\left(1-\frac{x}{z}\right)^{r-2}\right]^2, & 0 < x < z;\ r = 2,\dots,n \\ \frac{r(n-r)}{n^2(n-1)}, & x = 0;\ r = 0,1,\dots,n-1 \\ 0, & o.w. \end{cases}$$

For $r = n$, all the results will reduces to that of the estimates of an exponential distribution, without inliers.

Result 19: For fixed z and r, the UMVUE of the survival function $S(t) = P(X > t)$, $t \geq 0$ is obtained as

$$\hat{S}(t) = \begin{cases} \frac{r}{n}\left(1-\frac{t}{z}\right)^{r-1}, & t < z \\ 0, & o.w. \end{cases}$$

Result 20: For fixed z and r, the *UMVUE* of the variance of $\hat{S}(t)$ is obtained as

$$\widehat{var}\left[\hat{S}(t)\right] = \begin{cases} \left[\frac{r}{n}\left(1-\frac{t}{z}\right)^{r-1}\right]^2 - \frac{r(r-1)}{n(n-1)}\left(1-\frac{2\,t}{z}\right)^{r-1}, & z > 2t \\ \left[\frac{r}{n}\left(1-\frac{t}{z}\right)^{r-1}\right]^2, & t < z < 2t \\ 0, & o.w. \end{cases}$$

For $r = n$, both the above results reduce to the case of an exponential distribution.

Result 21: $\psi(T_3) = (r/n)e^{-t/\sum_{x_i>0} x_i}$ is CAN for the survival function $S(t) = P(X > t) = pe^{-t/\theta}$ with asymptotic variance $\frac{pe^{-2t/\theta}}{n\theta^2}$.

Definition 3 can be extended to multiparameter case so that CAN estimator for linear combination of parameters can be made possible. Let $T = (T_1, T_{2,\ldots,} T_m)'$ be a vector valued estimator which is consistent for a vector parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$ then T_i is CAN for θ_i with asymptotic variance $\frac{\lambda_{ii}(\theta)}{n}$ and any linear combination $T' = \sum_{i=1}^{m} l_i T_i$ is CAN for $\sum_{i=1}^{m} l_i \theta_i$ with asymptotic variance $\frac{1}{n} l' \Lambda(\theta) l$, where $\Lambda(\theta)$ is the variance-covariance matrix of vector of parameters θ (Kale and Muralidharan, 2015).

Result 22: Let $\psi(p, \theta) = l_1 p + l_2 \theta$, then the estimator $T' = l_1 \left(\frac{r}{n}\right) + l_2 \sum_{x_i>0} x_i$ is CAN for $\psi(p, \theta)$ with asymptotic variance $\frac{1}{n}(l_1^2 p(1-p)/n + l_2^2 \theta^2/(np))$.

We now investigate the MVU estimation of θ *or* $\psi(\theta)$ based on Cramer-Rao Lower Bound (CRLB) to the variance of an unbiased estimator. Let $\{f(x, \theta), \theta \in \Omega\}$, $\Omega c R^1$ be a class of distributions $I_X(\theta)$ is the Fisher Information, then under some regularity conditions (refer to Kale and Muralidharan, 2015) the CRLB for $V(T) \geq \left(\frac{d\psi(\theta)}{d\theta}\right)^2 / I_X(\theta)$. For instance, if $\psi(\theta) = \theta^2$ then the CRLB for $V(T)$ is $\frac{4\theta^4}{np}$. Similarly, the CRLB for $V(T)$ for estimating $\psi(p) = (1-p)^m$ is obtained as $\frac{mp(1-p)^{2m-1}}{n}$.

References

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**, 901–908.

Jani, P. N. and Singh, A. K. (1995). Minimum variance unbiased estimation in multi-parameter exponential family of distributions. *Metron*, **53**, 93–106.

Kale, B. K. and Muralidharan, K. (2015). *Parametric Inference: An Introduction*. Narosa Publishing House, New Delhi.

Lai, C. D., Khoo, B. C., Muralidharan, K., and Xie, M. (2007). Weibull model allowing nearly instantaneous failures. *Journal of Applied Mathematics and Decision Sciences*, Article ID 90842: 11 pages.

Muralidharan, K. (2000). The UMVUE and Bayes estimate of reliability of mixed failure time distribution. *Communications in Statistics - Simulations and Computations*, **29**, 603 – 619.

Muralidharan, K. (2010). Inliers prone models: A review. *Prob Stat Forum*, **3**, 38–51.

Roy, J. and Mitra, S. K. (1957). Unbiased minimum variance estimation in a class of discrete distributions. *Sankhya*, **18**, 371–378.