UNIVERSIDADE DE SÃO PAULO

A m.v.u. estimator vs a m.l. estimator: A way of comparing their precisions by using the Pareto Distribution.

ELIZABETE CORRÊA LEME

Nº 004

NOTAS

Instituto de Ciências Matemáticas de São Carlos

Instituto de Ciências Matemáticas de São Carlos

ISSN - 0103-2577

A m.v.u. estimator vs a m.l. estimator: A way of comparing their precisions by using the Pareto Distribution.

ELIZABETE CORRÊA LEME

Nº 004

NOTAS DO ICMSC
Série Estatística

São Carlos
Jul./1993

A m.v.u. estimator vs a m.l. estimator: A way of comparing their precisions by using the Pareto Distribution.

Elizabete Corrêa Leme
Departamento de Ciências de Computação e Estatística
Instituto de Ciências Matemáticas de São Carlos
Universidade de São Paulo
São Carlos - S.P.

Summary:

Estimating the fraction of population in an income bracket by using the Pareto Distribution, Shanmugam (1987) points out that a maximum likelihood (m.l) estimate is clearly inefficient as compared to a minimum variance unbiased (m.v.u) estimate when the income inequality parameter is known. In this note we define a family of m.v.u. estimators which includes the estimator of Shanmugam and by comparing the precision of this family and m.l. estimators we find a way to evaluate how much more efficient the estimator proposed by Shanmugam is when compared to the m.l. estimator.

Keywords: Pareto Distribution, Survival Function, Maximum Likelihood - Minimum Variance Unbiased Estimates, Reduced Random Sample, Mean Square Error.

1. Introduction

Denoting by X the person's income in a population, Vilfredo Pareto (1897) showed that, if

$$R(x) = P(X \geq x) \tag{1.1}$$

then, at least for large values of x,

$$lnR(x) \cong \alpha(ln\beta - lnx) \tag{1.2}$$

where α and β are parameters, α being representative of distribution shape and known as "constant of Pareto" and β meaning the minimum income.

A random variable X is called a classic Pareto type if its probability density function (p.d.f.) is

$$f(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \qquad x \geq \beta > 0, \quad \alpha > 0 \tag{1.3}$$

It is a known result that, if $X_1, X_2, \ldots, X_n$ is a random sample from the density (1.3), then $S_2 = min(X_1, X_2, \ldots, X_n)$ is a complete sufficient statistic for parameter β, exhibiting the p.d.f.

$$g(s_2; n) = \frac{n\alpha\beta^{n\alpha}}{s_2^{n\alpha+1}}, \qquad s_2 \geq \beta > 0, \quad \alpha > 0 \tag{1.4}$$

Furthermore, since S_2 is also the m.l. estimator of β then, the m.l. estimator of the fraction $R(x)$ when the parameter α is known is, due to the invariance property of the m.l. principle, given by

$$\hat{R}(x) = \left(\frac{S_2}{x}\right)^{\alpha} \tag{1.5}$$

The expectation, variance and mean square error of $\hat{R}(x)$ obtained are

$$E\hat{R}(x) = \frac{n}{n-1}\left(\frac{\beta}{x}\right)^{\alpha} \tag{1.6}$$

$$Var\hat{R}(x) = \frac{n}{(n-2)(n-1)}\left(\frac{\beta}{x}\right)^{2\alpha} \tag{1.7}$$

$$m.s.e.\hat{R}(x) = \frac{2}{(n-1)(n-2)}\left(\frac{\beta}{x}\right)^{2\alpha} \tag{1.8}$$

The m.l. principle provides as the m.l. estimators of those quantities, respectively,

$$\widehat{E\hat{R}(x)} = \frac{n}{n-1}\left(\frac{S_2}{x}\right)^{\alpha}$$

$$\widehat{Var\hat{R}(x)} = \frac{n}{(n-2)(n-1)^2}\left(\frac{S_2}{x}\right)^{2\alpha}$$

$$\widehat{m.s.e.\hat{R}(x)} = \frac{2}{(n-1)(n-2)}\left(\frac{S_2}{x}\right)^{2\alpha}$$

Among the (m.v.u.) estimates which have been proposed for $R(x)$ in the especific case when α is the parameter known, the one established by Shanmugam from

$Prob(X_1 > x/S_2)$, is obtained by adopting an approach of transformation of random variables which brings out the estimator

$$\hat{R}(x) = \left(1 - \frac{1}{n}\right)\left(\frac{S_2}{x}\right)^{\alpha} \quad , x > s_2 \tag{1.9}$$

Although we agree with Shanmugam's opinion that the m.l. estimator has been proved not so much efficient as compared to the estimator (1.9), we do believe that further attention should be given to the result he achieved once his argumentation is based on an analytic expression which is incorrect for the m.s.e. of the estimator as well as it doesn't clarify how much more efficient the proposed estimator is when compared to the m.l. estimator. By reducing the sample and using the correct analytic expression, we make it possible to establish a way of quantifying the Shanmugam's result and reach the conclusion that his estimator is more efficient than the m.l. estimator even in cases when the size of the sample has been considerably reduced.

2. A class of estimators for $R(x)$ when α is known.

Let $X_{i_1}, X_{i_2}, \ldots, X_{i_{n-j}}$ the random sample resulting of one obtained from the density (1.3) after j of its n components had been disregarded.

The statistic $S_{2,j} = min(X_{i_1}, X_{i_2}, \ldots, X_{i_{n-j}})$ is, like S_2, a complete sufficient statistic for the parameter β having as p.d.f. the density $g(s_{2,j}; n-j)$ given in accordance with (1.4).

Making use of the Rao Blackwell theorem, the (m.v.u) estimator for $R(x)$, built on $S_{2,j}$, is given by $Prob(X_{i_1} > x/S_{2,j})$. Then, employing the Shanmugam approach, we obtain as such estimator the expression

$$\tilde{R}_j(x) = \left(\frac{n-j-1}{n-j}\right)\left(\frac{S_{2,j}}{x}\right)^{\alpha} \quad , \quad x > s_{2,j} \; , \quad 0 \leq j \leq n-2 \tag{2.1}$$

The variances of these estimators are obtained as follows:

$$Var\tilde{R}_j(x) = \left(\frac{n-j-1}{n-j}\right)^2 Var\left(\frac{S_{2,j}}{x}\right)^{\alpha}$$

As we have

$$
\begin{aligned}
Var\left(\frac{S_{2,j}}{x}\right)^{\alpha} &= \int_{\beta}^{\infty}\left[\left(\frac{s_{2,j}}{x}\right)^{\alpha} - \left(\frac{n-j}{n-j-1}\right)\left(\frac{\beta}{x}\right)^{\alpha}\right]^2 g\left(s_{2,j}; n-j\right) ds_{2,j} \\
&= \frac{n-j}{(n-j-2)(n-j-1)^2}\left(\frac{\beta}{x}\right)^{2\alpha} \qquad (2.2)
\end{aligned}
$$

then

$$
Var\tilde{R}_j(x) = \left(\frac{n-j-1}{n-j}\right)^2 \frac{(n-j)}{(n-j-2)(n-j-1)^2}\left(\frac{\beta}{x}\right)^{2\alpha}
$$

That is,

$$
Var\tilde{R}_j(x) = \frac{1}{(n-j)(n-j-2)}\left(\frac{\beta}{x}\right)^{2\alpha}, \qquad 0 \le j \le n-3 \qquad (2.3)
$$

Now, denoting by $\mathcal{U}(S_{2,j}; n-j)$ the (m.v.u.) estimator of Var $\tilde{R}_j(x)$ and making use of the expectation inversion technique, the expression for $\mathcal{U}(s_{2,j}; n-j)$ is obtained as follows:

$$
\int_{\beta}^{\infty} u(s_{2,j}; n-j) g(s_{2,j}; n-j) ds_{2,j} = \frac{1}{(n-j)(n-j-2)}\left(\frac{\beta}{x}\right)^{2\alpha}
$$

That is,

$$
\int_{\beta}^{\infty} u(s_{2,j}; n-j) \frac{(n-j)^2 \alpha \beta^{(n-j)\alpha}(n-j-2)}{s_{2,j}^{(n-j)\alpha+1}\left(\frac{\beta}{x}\right)^{2\alpha}} ds_{2,j} = 1
$$

Or then

$$
\int_{\beta}^{\infty} u(s_{2,j}; n-j)(n-j)^2\left(\frac{x}{s_{2,j}}\right)^{2\alpha} g(s_{2,j}; n-j-2) ds_{2,j} - 1 \equiv 0
$$

Hence

$$
\int_{\beta}^{\infty}\left[u(s_{2,j}; n-j)(n-j)^2\left(\frac{x}{s_{2,j}}\right)^{2\alpha} - 1\right] g(s_{2,j}; n-j-2) ds_{2,j} \equiv 0
$$

As we have

$$
u(s_{2,j}; n-j)(n-j)^2\left(\frac{x}{s_{2,j}}\right)^{2\alpha} - 1 \equiv 0
$$

then

$$\mathcal{U}(S_{2,j}; n-j) = \frac{1}{(n-j)^2}\left(\frac{S_{2,j}}{x}\right)^{2\alpha}, \qquad x > s_{2,j}\ , \quad 0 \leq j \leq n-3 \tag{2.4}$$

Note that the expression (2.3) and (2.4) which have been obtained improve those (3.5) and (3.7) exhibited in Shanmugam (1987) when $j = 0$.

3. Comparing the precision of m.v.u. and m.l. estimators.

In order to establish a comparison of the efficiencies of the (m.v.u.) and the (m.l.) estimators, we examine the precision of each one which is given by the inverse of the mean square error which is equal to the variance plus the square of the bias.

Since the m.s.e. of $\tilde{R}_j(x)$ is its variance, exhibited in (2.3), and the m.s.e. of $\hat{R}(x)$ is the one given in (1.8), we observe that, in order to have $\tilde{R}_j(x)$ more efficient than $\hat{R}(x)$, we must have

$$\frac{1}{(n-j)(n-j-2)} \leq \frac{2}{(n-1)(n-2)}\ , \qquad 0 \leq j \leq n-3 \tag{3.1}$$

That is,

$$\varphi_1(n;j) \equiv 2j^2 - 4(n-1)j + (n^2 - n - 2) \geq 0\ , \qquad 0 \leq j \leq n-3$$

A straightforward calculation shows that the roots of that polinomial expression are

$$j_{1,1}(n) = (n-1) - \delta_1 \qquad \text{and} \qquad j_{1,2}(n) = (n-1) + \delta_1$$

where

$$\delta_1 = \frac{[2\,(n^2 - 3n + 4)]^{1/2}}{2} \tag{3.2}$$

So, in order to get the validity of the relationship (3.1), the value of j must be such that $j \leq j_{1,1}(n)$ or $j \geq j_{1,2}(n)$. As $j_{1,2}(n) > n-3$ then, by restriction in (3.1), it must be ignored. And since $j_{1,1}(n) \leq n-3$ for $n \geq 4$, then we can establish that

$$m.s.e.\tilde{R}_j(x) \leq m.s.e.\hat{R}(x)$$

for values of j in the interval $[0, [j_{1,1}(n)]]$ with $n \geq 4$, where $[j_{1,1}(n)]$ is the greatest integer minor or equal to $j_{1,1}(n)$.

4. Analysis of the m.s.e. estimates

Since the m.s.e. $\tilde{R}_j(x)$ and m.s.e. $\hat{R}(x)$ estimates are based, respectively, on the statistics S_{2j} and S_2, it also becomes necessary an analysis of those quantities.

We have, all the time, $s_{2,j} \geq s_2$. Thus:

(i) if $s_{2,j} = s_2$ then, the requirement

$$u(s_{2,j}; n-j) \ \leq m.s.\widehat{e.\hat{R}}(x) \tag{4.1}$$

that is,

$$\frac{1}{(n-j)^2}\left(\frac{s_{2,j}}{x}\right)^{2\alpha} \leq \frac{2}{(n-1)(n-2)}\left(\frac{s_2}{x}\right)^{2\alpha}$$

or then

$$\varphi_2(n;j) \equiv 2j^2 - 4nj + (n^2 + 3n - 2) \geq 0 \ ,$$

would bring to values of j satisfying $j \leq j_{2,1}(n)$ or $j \geq j_{2,2}(n)$ where

$$j_{2,1}(n) = n - \delta_2 \qquad \text{and} \qquad j_{2,2}(n) = n + \delta_2$$

with

$$\delta_2 = \frac{\left[2\,(n^2 - 3n + 2)\right]^{1/2}}{2} \tag{4.2}$$

Now, as we have $j_{2,2}(n) > n-3$ then it must be ignored and, in order to assure that $j_{2,1}(n) \leq n-3$, we must have $n \geq 6$. Furthermore, $j_{2,1}(n) > j_{1,1}(n), \forall n \geq 0$.

Thence, if $s_{2,j} = s_2$, in order that the validity of both the (3.1) and (4.1) relations be established, the values of j must be in the interval $[0, [j(n)]]$, with $n \geq 6$, where $[j(n)] \equiv [j_{1,1}(n)]$.

(ii) if $s_{2,j} > s_2$ then, denoting $\left(\frac{s_{2,j}}{s_2}\right)^{2\alpha} = q$, the requirement (4.1) becomes

$$\frac{2(n-j)^2}{(n-1)(n-2)} \geq q$$

or then

$$\phi(n;j;q) \equiv j^2 - 2nj - \left[\left(\frac{q}{2} - 1\right) n^2 - \frac{3}{2}nq + q\right] \geq 0$$

The roots of that expression are given by

$$j_1(n;q) = n - \delta \qquad \text{and} \qquad j_2(n;q) = n + \delta$$

where

$$\delta = \frac{\left[2\left(n^2 - 3n + 2\right) q\right]^{1/2}}{2} \tag{4.3}$$

Here we also have $j_2(n;q) > n - 3$ and, in order to make sure that $j_1(n;q) \leq n - 3$, we must require $n \geq \dfrac{3 + (1 + 72/q)^{1/2}}{2}$ which, in extreme case when $q = 1$ results in $n \geq 6$.

Consequently, when $s_{2,j} > s_2$, the validity of (4.1) is verifyed for values of j in the interval $[0, [j(n;q)]]$ with $n \geq 6$, where $[j(n,q)] \equiv [j_1(n,q)]$.

Now, through a comparative analysis with the expressions $j(n)$ and $j(n;q)$, we obtain

$$\begin{cases} j(n) \leq j(n;q) \ \text{ for } \ 1 \leq q \leq \eta_1(n), \ \text{ with } \ n \geq 6 \\ \\ \text{and} \\ \\ j(n;q) \leq j(n) \ \text{ for } \ \eta_1(n), \leq q \leq \eta_2(n) \ , \text{ with } \ n \geq 6 \end{cases}$$

where

$$\begin{cases} \eta_1(n) = 1 + \frac{4+2[2(n-1)(n-2)+4]^{1/2}}{(n-1)(n-2)} \\ \\ \eta_2(n) = \dfrac{2n^2}{(n-1)(n-2)} \end{cases} \tag{4.4}$$

Therefore, if $s_{2,j} > s_2$, the requirements (3.1) and (4.1) must be valid for values of j in the interval $[0, [j(n)]]$ when $1 \leq q \leq \eta_1(n)$, and for values of j in the interval

$[0, [j(n,q)]]$ when $\eta_1(n) \leq q \leq \eta_2(n)$.

5. Analysis of the results.

Considering the expression (4.3) we find that the quotient $[j(n,q)]/n$ is close to $1 - \sqrt{\frac{q}{2}}$ when n has large values. When $q = 1$ (or $s_{2_j} = s_2$), that is, when you discard $[j(n)]$ data of the sample and the minimum s_2 is maintained, such quotient remains around 0.29 for large values of $n (n > 100)$. For the other values, straight forward calculi can be made. You will find an expressive reduction in the size of the sample. Some examples will be found in Table (A) in the next page.

On the other hand, the expressions given in (4.4) assure that $\eta_i(n) \to i (i = 1,2)$ as n gets bigger. This fact indicates that, for $q > 1$ and $\eta_1(n) \leq q \leq \eta_2(n)$, the values of the minimum $s_{2,j}$ remain under control. Particularly for $\alpha = 1.5$ and for large values of n, you will find $s_{2,j}/s_2 < 1.28$. For small values of n, Table (B) that follows, shows possibilities of reduction as well as it displays the domains of q in each case.

TABLE

	A		B				
n	$[j(n)]$	$[j(n)]/n$	$\eta_1(n)$	$\eta_2(n)$	domain of q	$[j(n,q)]$	$[j(n,q)]/n$
6	1	0,1666	1.863	3.600	1.863-2.500	1	0,1666
					2.501-3.600	0	
10	2	0,2000	1.393	2.777	1.393-1.777	2	0.2000
					1.778-2.250	1	0.1000
					2.251-2.777	0	
20	5	0.2500	1.165	2.339	1.165-1.315	5	0.2500
					1.316-1.497	4	0.2000
					1.498-1.690	3	0.1500
					1.691-1.894	2	0.1000
					1.895-2.111	1	0.0500
					2.112-2.339	0	
30	8	0.2666	1.104	2.216	1.104-1.192	8	0.2666
					1.193-1.302	7	0.2333
					1.303-1.418	6	0.2000
					1.419-1.539	5	0.1666
					1.540-1.665	4	0.1333
					1.666-1.795	3	0.1000
					1.796-1.931	2	0.0666
					1.932-2.071	1	0.0333
					2.072-2.216	0	
50	14	0.2800	1.060	2.125	1.060-1.102	14	0.2800
					1.165-1.227	12	0.2400
					1.294-1.360	10	0.2000
					1.430-1.500	8	0.1600
					1.647-1.721	5	0.1000
					1.800-1.878	3	0.0600
					2.042-2.125	0	
100	29	0.2900	1.029	2.061	1.029-1.039	29	0.2900
					1.129-1.159	25	0.2500
					1.287-1.319	20	0.2000
					1.455-1.489	15	0.1500
					1.633-1.669	10	0.1000
					1.822-1.860	5	0.0500
					2.021-2.061	0	

References:

Pareto, V. (1987) - Cours d'Economie Politique. Lausanne and Paris: Rouge and Cie.

Shanmugam, R. (1987) - Estimating the fraction of population in an income bracket using Pareto distribution. Revista Brasileira de Probabilidade e Estatística, 1, 139-156.

NOTAS DO ICMSC

SÉRIE ESTATÍSTICA

Nº 001/93 - ACHCAR, J.A.- Some aspects of reparametrization for statistical models.

Nº 002/93 - RODRIGUES, J.; BABA, M.Y.- Bayesian estimation of a simple regression model with measurement errors

Nº 003/93 - ACHCAR, J.A.- Some practical aspects of approximate Bayesian inference

Nº 004/93 - LEME, E.C.- A m.v.u. estimator vs a m.l. estimator: A way of comparing their precisions by using the Pareto Distribution.

Observação: A partir do nº 134/93, a publicação "Notas do ICMSC" subdividiu-se em três séries: COMPUTAÇÃO, ESTATÍSTICA E MATEMÁTICA.