Original Article

Confidence interval estimation of log of scale parameter of pareto distribution

Sandeep Singh Charak[1*], Rahul Gupta[2]

[1,2]Department of Statistics, University of Jammu, Jammu-180006, INDIA.
Email: sndpcharak@gmail.com, rahulgupta68@yahoo.com

Abstract

The paper considers the problem of constructing confidence interval for the scale parameter of Pareto distribution. We prove the failure of the fixed sample size procedures to handle the estimation problem. Purely sequential procedure is developed to tackle the situation, thus leading us to variable sample selection procedure. The second-order approximations are obtained for the suggested variable sample selection procedure. Section 1 is introductory in nature and in section 2, is described the set-up of the estimation problem and prove the failure of the fixed sample size procedures to deal with it. In section 3, is proposed a equential procedure to construct fixed-width confidence interval for the log of scale parameter of Pareto Distribution and second order properties of the procedure are developed.

Keywords: Sequential procedure, Fixed-width confidence Interval, Second-order approximation, Stopping rule etc.

***Address for Correspondence:**
Dr. Sandeep Singh Charak, Department of Statistics, University of Jammu, Jammu-180006, INDIA.
Email: sndpcharak@gmail.com

Received Date: 18/12/2015 Revised Date: 10/01/2016 Accepted Date: 19/02/2016

Access this article online	
Quick Response Code:	Website: www.statperson.com
	DOI: 25 February 2016

INTRODUCTION

The Pareto distribution, named after the Italian civil engineer, economist, and sociologist V. Pareto, is a power law probability distribution that is used in description of social, scientific, geophysical, actuarial, and many other types of observable phenomena related to econometrics. Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. He also used it to describe distribution of income. The probability density function (PDF) graph of this distribution shows that the "probability" or fraction of the population that owns a small amount of wealth per person is rather high, and then decreases steadily as wealth increases. This distribution is not limited to describing wealth or income, but to many situations in which an equilibrium is found in the distribution of the "small" to the "large". Mukhopadhyay and Ekwo (1987) proposed some sequential estimation problems for the scale parameter of a Pareto Distribution and have time and again provided solutions for some related given precision problems. Castillo and Daoudi (2009) and De Zea Bermudez and Kotz, S. (2010), focused on new methods for Parametric estimation of the generalized Pareto distribution. Nadarajah and Ali (2008) applied Pareto random variables for hydrological modeling. In this paper we consider the problems of constructing confidence interval for the log of scale parameter of Pareto distribution. We prove the failure of the fixed sample size procedures to handle the estimation problems. Purely sequential procedure is developed to tackle the situation and second-order approximations are obtained. In section 2, we describe the set-up of the estimation problems and prove the failure of the fixed sample size procedures to deal with them. In section 3, we develop sequential procedure to construct fixed-width confidence interval for the log of scale parameter of Pareto Distribution.

THE ESTIMATION PROBLEM AND THE FAILURE OF THE FIXED SAMPLE SIZE PROCEDURE

How to site this article: Sandeep Singh Charak, Rahul Gupta. Confidence interval estimation of log of scale parameter of pareto distribution. *International Journal of Statistika and Mathemtika* February to April 2016; 18(1): 60-63. http://www.statperson.com (accessed 28 February 2016).

Let us consider a sequence $\{X_i\}, i = 1,2, \ldots$ of independent random variables from a first kind of Pareto Distribution $f(X;\mu,\sigma) = \sigma^{-1}\mu^{-1/\sigma}X^{-1/\sigma-1}$; $X \geq \mu > 0, \sigma > 0$, where μ and σ are respectively, the unknown scale and shape parameters. Given an observed random sample $X_1, \ldots\ldots, X_n$ of size n(≥2), for $X_{n(1)} = \min(X_1, \ldots\ldots, X_n)$, Let $u_{n(1)} = \log X_{n(1)}$ and $\widehat{\sigma}_n = (n-1)\sum_{i=1}^{n} \log\{X_i/X_{n(1)}\}$ be the estimators of log μ and σ, respectively.

Let$(\log\mu, \sigma)\, \epsilon\, \Re \times \Re^+$, where $\Re$ and $\Re^+$denote, respectively, the one –dimensional euclidean space and the positive-half of the real line satisfying the following assumptions:

(i) : $n\sigma^{-1}[(u_{n(1)} - log\mu)] \sim \chi^2_{(2)}$, where $\chi^2_{(r)}$ denotes a chi-square random variable with 2 degrees of freedom.

(ii) For all $n \geq 2$, $u_{n(1)}$ and $\hat{\sigma}_n$ are stochastically independent.

(iii) : $2(n-1)\hat{\sigma}_n/\sigma = \sum_{j=1}^{n-1} Z_j^{(2)}$, $where\ Z_j^{(2)} \sim \chi^2_{(2)}$.

Our problem is to construct a fixed-size confidence interval for $\log\mu$. For specified $d\,(>0)$ and $\alpha\, \epsilon\, (0,1)$, suppose one wishes to construct a confidence interval R_n for $\log\mu$ such that its maximum width is 2d and $P(\log\mu\ \epsilon\ R_n) \geq \alpha$. We define

$$R_n = [\,Z : \{(\,u_{n(1)} - Z\,)\} \leq d^2] \tag{2.1}$$

We note that R_n is ellipsoidal confidence interval.

Since $P(\log\mu\ \epsilon\ R_n) = P\left[\left(u_{n(1)} - \log\mu\right)^2 \leq d^2\right] \equiv P\left[\left|u_{n(1)} - \log\mu\right| \leq d\right]$,

results based on the confidence interval (2.1) are equivalent to those based on the confidence interval of width 2d. Denoting by $G^{(2)}(.)$, the cumulative distribution function (c.d.f) of a $\chi^2_{(2)}$ random variable and utilizing (i), we obtain from (2.1),

$$P(\log\mu\, \epsilon\, R_n) = G^{(2)}(n\,\sigma^{-1}d^2). \tag{2.2}$$

Let 'a' be the constant, determined by the relation

$$G^{(2)}(a^2) = \alpha. \tag{2.3}$$

Using monotonicity property of cumulative distribution function (cdf), it can be seen from (2.2) and (2.3) that, for known σ, in order to achieve

$P(\log\mu\ \epsilon\ R_n) \geq \alpha$,

the fixed sample size required is the smallest positive integer $n \geq n^*$,

where

$$n^* = (a/d)^2\sigma. \tag{2.4}$$

However, as we have already assumed, that σ is unknown, there does not exist any fixed sample size procedure which achieves the goals of 'specified width and coverage probability' for all values of σ. Thus we propose a sequential procedure to tackle the problem as demonstrated in the next section.

SEQUENTIAL PROCEDURE TO CONSTRUCT FIXED-SIZE CONFIDENCE INTERVAL FOR *log μ*

Let us start with the sample of size $m \geq 2$. It is worth mentioning here that we take $m \geq 2$ in order to ensure the assumptions $(i) - (iii)$. Then, the stopping time $N \equiv N(d)$ associated with the sequential procedure is defined by

$$N = inf\left[n \geq m :\ n \geq \left(\frac{a}{d}\right)^2 \hat{\sigma}_n\right]. \tag{3.1}$$

After stopping, we construct the interval

$$R_n = [\,Z : \{(\,u_{n(1)} - Z\,)\} \leq d^2] \tag{3.2}$$

for $\log\mu$. Utilizing $(i)\ and\ (ii)$, the coverage probability associated with this sequential procedures defined at (3.1), comes out to be

$$P(\log\mu\ \epsilon\ R_N) = \sum_{n=m}^{\infty} P\left[n\sigma^{-1}\left\{\left(u_{n(1)} - log\mu\right)^2\right\}^{1/2} \leq a^2\left(\frac{n}{n^*}\right); N = n\right]$$

$$= \sum_{n=m}^{\infty} P\left[n\sigma^{-1}\{(u_{n(1)} - log\mu)\} \leq a^2\left(\frac{n}{n^*}\right); N = n\right]$$

$$= \sum_{n=m}^{\infty} G^{(2)}\left(\frac{a^2 n}{n^*}\right) P(N = n)$$

$$= E\left[G^{(2)}\left(\frac{a^2 n}{n^*}\right)\right]. \tag{3.3}$$

In what follows, we obtain second-order approximations for the expected sample size and coverage probability associated with the sequential procedures. Before proving the main results, we establish some lemmas. We denote by

$V_n = 2(n-1)\hat{\sigma}_n/\sigma$.

Lemma 3.1: N terminates with probability one. (3.4)

$$\lim_{d\to 0} N = \infty\ a.s. \tag{3.5}$$

$$\lim_{d\to 0} \frac{N}{n^*} = 1\ a.s. \tag{3.6}$$

Proof: Denoting by $Z_n = \{V_n - 2(n-1)\}/\sqrt{4(n-1)}$, it follows from the definition of N that

$$P(N > n) \leq P\left[V_n - 2(n-1)\left(\frac{n}{n^*}\right)\right]$$

$$= P\left[Z_n \geq \left\{\frac{2(n-1)}{2}\right\}^{\frac{1}{2}}\left\{\left(\frac{n}{n^*}\right) - 1\right\}\right]. \qquad (3.7)$$

It follows from (iii) and the central limit theorem that $Z_n \overset{L}{\rightarrow} Z$ as $n \rightarrow \infty$, where $Z \sim N(0,1)$ and from Zacks (1971,p.561),

$1 - \Phi(x) \approx x^{-1}\,\phi(x)\; as\; x \rightarrow \infty$,

where $\Phi(x)$ and $\phi(x)$ denote, respectively, the cumulative distribution function and probability density function of a N(0,1) random variable.

Thus we obtain from (3.7),

$P(N > n) = O\left(n^{-3/2}\right)\; as\; n \rightarrow \infty$,and (3.4) follows.

Result (3.5) is a direct consequence of the definition of N.

We notice the basic inequality

$$\left(\frac{a}{d}\right)^2 \hat{\sigma}_N \leq N \leq \left(\frac{a}{d}\right)^2 \hat{\sigma}_N + (m-1), \qquad (3.8)$$

or

$$\left(\frac{\hat{\sigma}_N}{\sigma}\right) \leq \frac{N}{n^*} \leq \left(\frac{\hat{\sigma}_N}{\sigma}\right) + \frac{(m-1)}{n^*}.$$

From (iii) and strong law of large numbers [See Bhat (1981,p187)], we conclude that $\hat{\sigma}_N \overset{a.s.}{\longrightarrow} \sigma\; as\; n \rightarrow \infty$. Result (3.6) now follows from (3.9) on taking the limit as $d \rightarrow 0$ and using (3.5). In the following lemma, we provide a simple and direct method of obtaining asymptotic distribution of the stopping time.

Lemma 3.2 : As $d \rightarrow 0$, $(n^*)^{-1/2}(N - n^*) \overset{L}{\rightarrow} N(0,1)$.

Proof : It follows from (iii) that

$E(\hat{\sigma}_n) = \sigma$

and

$E\left(\hat{\sigma}_n{}^2\right) = \sigma^2[1 + 2(2n)^{-1} + o(n^{-1})]$,

$or\; E\left(\hat{\sigma}_n{}^2\right) = \sigma^2[1 + 1/n + o(n^{-1})]$,

so that,

$var(\hat{\sigma}_n) = 2\sigma^2(2n)^{-1} + o(n^{-1})$

$or\; var(\hat{\sigma}_n) = \sigma^2/n + o(n^{-1})$.

Thus, from central limit theorem,

$\sqrt{n^*}\,\frac{(\hat{\sigma}_{n^*} - \sigma)}{\sigma} \overset{L}{\rightarrow} N(0,1)$, as $n^* \rightarrow \infty$,

which on using lemma 1 and Theorem 1 of Anscombe (1952), leads us to that, as $d \rightarrow 0$,

$$\frac{\sqrt{N}(\hat{\sigma}_N - \sigma)}{\sigma} \overset{L}{\rightarrow} N(0,1).$$

We obtain from (3.8) that

$$\frac{\sqrt{n^*}(\hat{\sigma}_N - \sigma)}{\sigma} \leq \frac{(N - n^*)}{(n^*)^{1/2}} \leq \frac{\sqrt{n^*}(\hat{\sigma}_N - \sigma)}{\sigma} + \frac{(m-1)}{(n^*)^{1/2}},$$

which on applying (3.10) gives the desired result.

Lemma 3.3_: $(N - n^*)^2/n^*$ is uniformly integrable for all $m > 2$.

Proof: Denoting with F (.), the cumulative distribution function of $Z_j^{(2)}$,

we have for some B(>0),

$F(X) = P\left(Z_j^{(2)} \leq X\right)$

$= B\int_0^x e^{-Y/2}\; Y^{2/2-1}dy$

$\leq BX$

Thus, in Woodroofe's (1977) notations, $a = 1$. The lemma is now a direct consequence of Theorem 2.3 of Woodroofe (1977).

Lemma 3.4_: For $\eta\, \epsilon (0,1)$, as $d \rightarrow 0$,

$P(N = m) = P(m + 1 \leq N \leq \eta\, n^*)$

$= O\left(d^{2(m-1)}\right)$.

Proof: The proof is similar to that of Lemma 3 in Chaturvedi, Pandey, and Gupta (1991). The main results of this section are now stated and proved in the following theorem, which provides second-order approximations for the expected sample size and coverage probability associated with the sequential procedures.

Theorem 3.1 : For the sequential procedures defined at (3.1), and all $m > max\{1,2\}$, as $d \rightarrow 0$ (3.9)

$$E(N) = n^* + \upsilon - 2 + o(1), \qquad (3.11)$$

and

$$P(\log \mu\, \epsilon\, R_N) = \alpha + \left(\frac{a^2}{n^*}\right)\left[\upsilon - 1 + \frac{1}{4}\{2 - (a^2 + 6)\}\right] g^{(2)}(a^2) + o(d^2) \qquad (3.12)$$

where $g^{(2)}(.)$ denotes the probability density function (pdf) of a $\chi^2_{(2)}$ random variable and υ is specified.

Proof: Utilizing (iii), the stopping rule (3.1) can be rewritten as

$$N = inf\left[n \geq m : \sum_{j=1}^{n-1}\left\{{}^1\!/_2\, Z_j^{(2)}\right\} \leq (n-1)\left(\frac{n}{n^*}\right)\right].$$

Let us define a new stopping variable N^* by

$$N^* = inf\left[n \geq m - 1 : \textstyle\sum_{j=1}^{n}\left\{{}^1\!/_2\, Z_j^{(2)}\right\} \leq n^2(1 - n^{-1})(n^*)^{-1}\right]. \qquad (3.13)$$

Along the lines of proof of Lemma 1 in Swanepoel and van Wyk (1982), it can be shown that the stopping variables $N\; and\; N^*$ follows the same probability distribution. From (3.13) (3.10) and equation (1.1) of Woodroofe (1977), $\lambda = n^*, L(n) = 1 + sn^{-1}, L_0 = s, \alpha = 2, \beta = 1, \mu = 1$ and $\tau^2 = 2q^{-1}$. Result (3.11) is now a direct consequence of Theorem 2.4 of Woodroofe (1977) that, for all $m > max(1,2)$, as $d \rightarrow 0$,

$E(N) = \lambda + \beta\mu^{-1}\upsilon - \beta L_0 - (1/2)\alpha\beta^2\tau^2\mu^{-2} + o(1)$.

Expanding $G^{(2)}(.)$ around 'a^2' by Taylor's series expansion, we obtain from (3.3), for

$|a^2 - W| \leq a^2|(N/n^*) - 1|$,

$P(\log \mu\; \epsilon\; R_N) = G^{(2)}(a^2) + a^2 G^{(2)'}(a^2)E[(N/n^*) - 1]$

$$+(a^4/2)E\left[\{(N/n^*)-1\}^2 G^{(2)''}(W)\right], \qquad (3.14)$$

where $G^{(2)'}(X)$ and $G^{(2)''}(X)$ denote, respectively, the first and second derivatives of $G^{(2)}(X)$. It can be seen that $G^{(2)'}(X) = g^{(2)}(X)$ and

$$G^{(2)''}(X) = [-1/2]g^{(2)}(X). \qquad (3.15)$$

On the event '$N > \eta\, n^*$', $|W-1| \le |(N/n^*)-1|$ gives, $\eta \le W \le 2-\eta$, that is, both positive and negative powers of W are bounded. Thus, from (3.15), $G^{(2)''}(W)$ is bounded on the event '$N > \eta\, n^*$'. Moreover, on the event '$N < \eta\, n^*$', $(m/n^*) \le W \le (2-m/n^*)$, so that, for d sufficiently small, denoting by K-any positive generic constant independent of d, we have

$$E\left[G^{(2)''}(W)I(N \le \eta\, n^*)\right] \le K\, n^* P(N \le \eta\, n^*),$$

which on using Lemma 3.4 gives, as $d \to 0$,

$$E\left[G^{(2)''}(W)I(N \le \eta\, n^*)\right] = o\left(d^{2(m-1)-2}\right)$$

$= o(1)$, for all $m > 2$

Thus we conclude that $G^{(2)''}(W)$ is bounded for all $m > 2$. It follows from (3.6) and the definition of W that $W \xrightarrow{a.s.} a^2$ as $d \to 0$.

Using these results, lemmas 3.2, 3.3 and (3.11), we obtain from (3.14) and (3.15), for all $m > \max(1,2)$, as $d \to 0$,

$$P(\log\mu\ \epsilon\ R_N) = \alpha + \left(\frac{\alpha^2}{n^*}\right)[v-2+o(1)]g^{(2)}(\,a^2)$$

$+\left(\frac{\alpha^4}{2n^*}\right)[-1/2]g^{(2)}(\,a^2)$.Result (3.12) now follows after some algebraic manipulations, thus completing the proof of the theorem.

REFERENCES

1. Anscombe, F.J (1952): Large sample theory of sequential estimation. Proc. Cambridge Philos.Soc., 48, 600-607.
2. Bhat, B.R. (1981): Modern probability theory.Wiley Eastern Limited, New Delhi.
3. Castillo, J. and Daoudi, J. (2009), Estimation of the generalized Pareto distribution. Statistics and Probability Letters, 79, 684 – 688.
4. Chaturvedi, A., Pandey, S.K. and Gupta, M. (1991): On a class of asymptotically risk-efficient sequential procedures.Scandinavian Actura. Jour., 1 : 87-96
5. De Zea Bermudez, P. and Kotz, S. (2010), "Parameter estimation of the generalized Pareto distribution—Part I," Journal of Statistical Planning and Inference, Volume 140, Issue 6 (2010), Pages 1353-1373.
6. Mukhopadhyay, N. and Ekwo, M.E. (1987): Sequential Estimation problems for the Scale Parameter of a Pareto Distribution. Scand. Actuarial Journal, Vol.1987, 1-2, 83-103.
7. Nadarajah S, Ali MM (2008) Pareto random variables for hydrological modeling. Water Resour Manag 22: 1381–1393.
8. Swanepoel, J.W.H. and Van Wyk, J.W.J. (1982): Fixed width confidence intervals for the location parameter of an exponential distribution. Commun. Statist. Theor. Meth. A11 (11), 1279-1289.
9. Woodroofe, M. (1977): Second-order approximations for sequential point and interval estimation. Ann Statist., 5, 984-995.

Source of Support: None Declared
Conflict of Interest: None Declared