American Journal of Mathematics and Statistics 2013, 3(6): 332-336
DOI: 10.5923/j.ajms.20130306.05

Sequential Estimation of the Shape Parameter of the Pareto Distribution

Mohamed Tahir

Department of Mathematics, College of Sciences, University of Sharjah

Abstract The problem addressed is that of sequentially estimating θ, the shape parameter of the type I Pareto distribution, subject to the loss function $L_c = (\hat{\theta}_n - \theta)^2 + cn$, where c is a known cost per observation and $\hat{\theta}_n$ is the maximum likelihood estimator of θ. We propose a stopping time t and provide a second-order asymptotic expansion, as $c \to 0$, for the regret incurred by the sequential procedure $(t, \hat{\theta}_t)$ under the loss L_c. We also show that the point estimator $\hat{\theta}_t$ is asymptotically unbiased for θ.

Keywords Anscombe's Theorem, Doob's Maximal Inequality, Excess over the Stopping Boundary, Hölder's Inequality, Regret, Shape Parameter, Stopping Time, Uniform Integrability

1. Introduction

The Type I Pareto distribution is a continuous distribution with probability density function (p.d.f.)

$$f_\theta(x) = \begin{cases} \dfrac{\theta}{x^{\theta+1}} & \text{if } x \geq 1 \\ 0 & \text{if not,} \end{cases} \qquad (1)$$

where θ is a positive number.

The Pareto distribution was first formulated in the late 1800s by the Italian economist Vilfredo Pareto, who 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 the distribution of income. The Pareto distribution is not limited to describing wealth or income. It can be used as the distribution for insurance loss, oil reserve in a oil field, standardized price return on an individual stock, area burnt in a forest fire, etc…

Let $X_1, , \ldots, X_n$ denote independent observations to be taken sequentially up to a predetermined stage n from the Pareto distribution with p.d.f. (1). It is desired to estimate the parameter θ, subject to the loss function

$$L_c(\hat{\theta}_n, \theta) = (\hat{\theta}_n - \theta)^2 + cn, \qquad (2)$$

where c is a known cost per observation and $\hat{\theta}_n$ is the maximum likelihood estimator of θ. Since the log-likelihood function is $l_n(\theta) = n \ln \theta - (\theta + 1) \sum_{i=1}^{n} \ln x_i$ for observed values $x_1 > 1, \ldots, x_n > 1$, of $X_1, , \ldots, X_n$, it follows that the maximum likelihood estimator of θ is

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} \ln X_i} = \frac{1}{\overline{Y}_n},$$

where $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ with $Y_i = \ln X_i$, $i = 1, \ldots, n$, and where the random variables $Y_1, \ldots, Y_n$ are independent with common distribution the Exponential distribution with mean $\mu_Y = 1/\theta$. Expanding $h(y) = 1/y$ about $y = 1/\theta$ and substituting $y = 1/\overline{Y}_n$ in the obtained expansion yields $\hat{\theta}_n - \theta \approx -\theta^2 (\overline{Y}_n - \mu_Y)$ for large n. It follows that the risk incurred by estimating θ by $\hat{\theta}_n$ under the loss (2) is

$$R_c(n, \theta) = E[L_c(\hat{\theta}_n, \theta)]$$

$$= E[(\hat{\theta}_n - \theta)^2] + cn \approx \frac{\theta^2}{n} + cn$$

for large n. The approximate risk is minimized with respect to n by choosing n as the greatest integer less than or equal to

$$n^* = \theta / \sqrt{c}. \qquad (3)$$

The minimum risk is

* Corresponding author:
tahir_stat@yahoo.com (Mohamed Tahir)
Published online at http://journal.sapub.org/ajms

$$R_c^*(\theta) = R_c(n_c, \theta) \approx 2\theta\sqrt{c} = 2cn^*$$

for sufficiently small c. Since n^* depends on the unknown value of θ, there is no fixed-sample-size procedure that attains the minimum risk $R_c^*(\theta)$. Therefore, we propose to use the sequential procedure $(t, \hat{\theta}_t)$ which stops the sampling process after observing $Y_1, \ldots, Y_t$ and estimates θ by $\hat{\theta}_t = 1/\overline{Y}_t$, where

$$t = \inf\left\{ n \geq m : n > c^{-1/2}\overline{Y}_n^{-1} \right\} \quad (4)$$

with m being a positive integer. The performance of the procedure $(t, \hat{\theta}_t)$ is measured by its regret, which is defined as

$$r_c(\hat{\hat{\theta}}_t) = E[L_c(\theta_t, \theta] - R_c^*(\theta) = E\left[\left(\frac{1}{\overline{Y}_t} - \theta\right)^2 + ct\right] - 2cn^* \quad (5)$$

for $c > 0$.

In this paper we propose a stopping time, t, based on (3), and provide a second-order asymptotic expansion, as $c \to 0$, for the regret incurred by the sequential procedure $(t, \hat{\theta}_t)$ under the loss (2). We also show that the estimator $\hat{\theta}_t$ is asymptotically unbiased for the shape parameter θ.

The problem of sequentially estimating the mean of a population was studied by many authors. Starr and Woodroofe[7] considered the case in which $X_1, X_2, \ldots$ are i.i.d. Normal random variables and showed that the regret of his procedure is O(1). Then, Woodroofe[10] showed that the regret is 0.5 + o(1) if $m \geq 4$. Martinsek[8] extended Woodroofe's[10] result to the nonparametric case. For the distribution-free case, Ghosh and Mukhopadhyay[3] and Chow and Yu[5] established asymptotic risk efficiency under certain conditions. Tahir[9] proposed a class of bias-reduction estimators of the mean of the one-parameter exponential family and provided an asymptotic second-order lower bound for the regret.

2. Asymptotic Expansion for the Regret of the Sequential Procedure $(t, \hat{\theta}_t)$

Theorem 1: Let t be as in (4) with $m \geq 1$, then $E[t] = n^* + 1 + o(1)$ as $c \to 0$.

Proof: The stopping time t can be rewritten as

$$t = \inf\left\{ n \geq m : Z_n > n^* \right\}, \text{ where } Z_n = n\theta\overline{Y}_n. \quad (6)$$

Furthermore, let $U_i = \theta Y_i - 1$ for i = 1, 2, … and let $\overline{U}_n = n^{-1}(U_1 + \cdots + U_n),\ n \geq 1.$ Then, $Z_n = ng(\overline{U}_n)$ with $g(u) = u + 1$. Since g is a convex function and

$$E\left[\left(g(U_1)^+\right)^3\right] = E\left[\left(\theta Y_1\right)^+\right)^3\right] = \theta^3 E\left[Y_1^3\right] < \infty,$$

where $U^+ = \max\{U, 0\}$, it follows from Proposition 5 and Theorem 1 of[2] that $E[t] = n^* + \rho + o(1)$ as $c \to 0$, where ρ denotes the asymptotic mean of the excess over the stopping boundary, $B_c = tZ_t - n^*$. Moreover,

$$\rho = \frac{E\left[(\tau + S_\tau)^2\right]}{2E[\tau + S_\tau]} = \frac{E\left[(1 + U_1)^2\right]}{2E[1 + U_1]} = \frac{E[(\theta Y_1)^2]}{2E[\theta Y_1]} = 1,$$

where

$$\tau = \inf\left\{ n \geq 1 : n + S_n > 0 \right\} = 1$$

$$\text{and } S_n = U_1 + \cdots + U_n.$$

The theorem follows.

Expanding $h(y) = 1/y$ about $y = 1/\theta$ and substituting $y = 1/\overline{Y}_t$ in the obtained expansion yields

$$\hat{\theta}_t = \frac{1}{\overline{Y}_t} = \theta - \theta^2\left(\overline{Y}_t - \frac{1}{\theta}\right) + \frac{1}{W_t^3}\left(\overline{Y}_t - \frac{1}{\theta}\right)^2, \quad (7)$$

where W_t is a random variable lying between $\overline{Y}_t$ and $1/\theta$.

Lemma 1: Let t be as in (4) and let $p > 1$. Then,

(i) $\frac{t}{n^*} \to 1$ w.p.1 as $c \to 0$.

(ii) $\left(\frac{t}{n^*}\right)^{-p}, c > 0,$ is uniformly integrable;

(iii) $\left(\frac{t}{n^*}\right)^{p}, 0 < c \leq c_0,$ is uniformly integrable if $m > p$.

Proof. Assertion (i) holds by Proposition 2 of Aras and Woodroofe[2]. For the second assertion,

$$\left(\frac{t}{n^*}\right)^{-p} \leq \left(\theta\overline{Y}_t\right)^p \Rightarrow \sup_{c>0}\left(\frac{t}{n^*}\right)^{-p} \leq \sup_{c>0}\left(\theta\overline{Y}_t\right)^p \leq \sup_{n\geq 1}\left(\theta\overline{Y}_n\right)^p$$

since $t > c^{-1/2}\overline{Y}_t$, by the definition of t. Thus, Assertion (ii) holds since

$$E\left[\sup_{n\geq 1}\left(\theta\overline{Y}_n\right)^p\right] \leq \left(\frac{p}{p-1}\right)^p E[Y_1] < \infty,$$

By Doob's maximal inequality (see[4]). To establish (iii), observe that $(t-1)\theta\overline{Y}_{t-1} \leq n^*$ on $\{t > m\}$, by the definition of t; so that

$$\frac{t}{n^*} \leq \left[\left(\theta\overline{Y}_{t-1}\right)^{-1} + \frac{1}{n^*}\right] I_{\{t>m\}} + \frac{m}{n^*} I_{\{t=m\}}$$

$$\leq \frac{(\overline{Y}_{t-1})^{-1}}{\theta} I_{\{t>m\}} + (m+1),$$

where $I_{\{.\}}$ denotes the indicator function. This implies that

$$\left(\frac{t}{n^*}\right)^p \le \frac{2^{p-1}}{\theta^p}\left(\overline{Y}_{t-1}\right)^{-p} I_{\{t>m\}} + 2^{p-1}(m+1)^p,$$

by the c_r-inequality (see Loève[6]). Since

$$\sup_{0<c\le c_0}\left(\overline{Y}_{t-1}\right)^{-p} I_{\{t>m\}} \le \sup_{n\ge m}\left(\overline{Y}_m\right)^{-p} \text{ and}$$

$$E\left[\sup_{n\ge m}\left(\overline{Y}_n\right)^{-p}\right] \le AE\left[\left(\overline{Y}_m\right)^{-p}\right] < \infty,$$

for some A >0, by Doob's maximal inequality, Assertion (ii) follows.

Lemma 2: Let $p > 1$. If $m > p$, then,

(i) $\left|(n^*)^{-1/2}\sum_{i=1}^{t}\left(Y_i - \frac{1}{\theta}\right)\right|^p$, $c > 0$, is uniformly integrable and

(ii) $W_t^{-p}, c > 0,$ is uniformly integrable.

Proof. The first assertion follows by using Assertion (i) of Lemma 1 and Theorem 2 of[3]. The second assertion follows since $E[W_t^{-p}] \le \theta^q + E[(\overline{Y}_t)^{-p}]$ and

$$\sup_{c>0} E[(\overline{Y}_t)^{-p}] \le E\left[\sup_{n\ge m}(\overline{Y})^{-p}\right] \le BE[(\overline{Y}_m)^{-p}] < \infty$$

for some constant $B > 0$, by Doob's maximal inequality.

Theorem 2: Let $r_c(\hat{\theta}_t)$ be as in (5). Then, $r_c(\hat{\theta}_t) = 10c + o(c)$ as $c \to 0$.

Proof. It follows from (5) that

$$\begin{aligned} r_c(\hat{\theta}_t) = E\left[\left(\frac{1}{\overline{Y}_t} - \theta\right)^2\right] + \mathrm{cE[t]} - 2cn^* &= E\left[\theta^4\left(\overline{Y}_t - \frac{1}{\theta}\right)^2\right] + cE[t] - 2cn^* \\ &\quad - 2\theta^2 E\left[\frac{1}{W_t^3}\left(\overline{Y}_t - \frac{1}{\theta}\right)^3\right] + E\left[\frac{1}{W_t^6}\left(\overline{Y}_t - \frac{1}{\theta}\right)^4\right] \\ &= c\left[(n^*)^2 E\left[\overline{U}_t^2\right] + E[t] - 2n^*\right] \\ &\quad - 2\theta^2 E\left[\frac{1}{W_t^3}\left(\overline{Y}_t - \frac{1}{\theta}\right)^3\right] + E\left[\frac{1}{W_t^6}\left(\overline{Y}_t - \frac{1}{\theta}\right)^4\right] \end{aligned} \tag{8}$$

for $c > 0$. Next, by Corollary 1 of Aras and Woodroofe[2],

$$(n^*)^2 E\left[\overline{U}_t^2\right] + E[t] - 2n^* = 3 + 2E[U_1^3] + o(1) = 7 + o(1) \tag{9}$$

as $c \to 0$. Now, let

$$Q_t = \frac{S_t^* - \mu_t t}{\sigma_Y\sqrt{t}} = \frac{S_t^* - \frac{t}{\theta}}{\frac{\sqrt{t}}{\theta}} \text{ with } S_t^* = Y_1 + \cdots + Y_t. \tag{10}$$

and let Z denote a random variable having the Standard Normal distribution. Then,

$$\begin{aligned} \frac{1}{c}E\left[\frac{1}{W_t^3}\left(\overline{Y}_t - \frac{1}{\theta}\right)^3\right] &= \frac{1}{c^{1/4}\theta^3}E\left[\frac{1}{W_t^3}\left(\frac{t}{n^*}\right)^{-3/2} Q_t^3\right] \to E[Z^3] = 0 \\ \frac{1}{c}E\left[\frac{1}{W_t^6}\left(\overline{Y}_t - \frac{1}{\theta}\right)^4\right] &= \frac{1}{(n^*)^2\theta^4}E\left[\frac{1}{W_t^6}\left(\frac{t}{n^*}\right)^{-2} Q_t^4\right] \to E[Z^4] = 3 \end{aligned} \tag{11}$$

as $c \to 0$, by Assertion (i) of Lemma 1, the fact that $Q_t \to Z$ in distribution as $c \to 0$ by Anscombe's theorem (see[1]) and the fact that $W_t \to 1/\theta$ w.p.1 as $c \to 0$. The uniform integrability of $Q_t^p, c > o,$ follows from Assertions (ii) of Lemmas 1 and 2 since

$$Q_t = \frac{1}{\sigma_Y}\left(\frac{t}{n^*}\right)^{-1/2}\left[(n^*)^{-1/2}\sum_{i=1}^{t}\left(Y_i - \frac{1}{\theta}\right)\right].$$

Now take the limit as $c \to 0$ in (8) and use (9) and (11) to complete the proof of the theorem.

Lemma 3: Let t be defined by (4). If $m > 4$, then $E[\hat{\theta}_t] = \theta + o(\sqrt{c})$ as c $\to 0$.

Proof: It follows from (7) that

$$\frac{\hat{\theta}_t - \theta}{\sqrt{c}} = -\frac{\theta^2}{\sqrt{c}}\left(\overline{Y}_t - \mu_Y\right) + \frac{1}{W_t^3\sqrt{c}}\left(\overline{Y}_t - \mu_Y\right)^2, \tag{12}$$

where W_t is a random variable lying between $\overline{Y}_t$ and $\mu_Y = 1/\theta$. Next,

$$E\left[\frac{\theta}{\sqrt{c}}\left(\overline{Y}_t - \mu_Y\right)\right] = E\left[\left(\frac{n^*}{t} - 1\right)t\left(\overline{Y}_t - \mu_Y\right)\right] \tag{13}$$

by (3) and Wald's lemma. Moreover, it follows from (6) that

$$\left(\frac{n^*}{t} - 1\right)t\left(\overline{Y}_t - \mu_Y\right) = \theta t\left(\overline{Y}_t - \mu_Y\right)^2 - U_c\left(\overline{Y}_t - \mu_Y\right) = \frac{1}{\theta}Q_t^2 + I_c,$$

say, where $U_c = Z_t - n^*$ denotes the excess over the stopping boundary and Q_t is as in (10). Now, $E\left[\theta^{-1}Q_t^2\right] \to \theta^{-1}$ as $c \to 0$, by Anscombe's theorem and the fact that $Q_t^2, c > o,$ is uniformly integrable. Also, by Hölder's inequality,

$$|E[I_c]| = \frac{1}{\sqrt{n^*}}E\left[U_c\left(\frac{t}{n^*}\right)^{-1}\frac{\sum_{i=1}^{t}(Y_i - \mu_Y)}{\sqrt{n^*}}\right] \le \frac{1}{\sqrt{n^*}}\left(E[U_c^2]\right)^{1/2}\left(E\left[\left|(n^*)^{-1/2}\sum_{i=1}^{t}\left(Y_i - \frac{1}{\theta}\right)\right|^4\right]\right)^{1/4}\left(E\left[\left(\frac{t}{n^*}\right)^{-4}\right]\right)^{1/4}$$

$$\to 0$$

as $c \to 0$, by Proposition 7 of[2] and Lemmas 1 and 2.Using these results in (13) yields

$$E\left[\frac{\theta}{\sqrt{c}}\left(\overline{Y}_t - \mu_Y\right)\right] = \frac{1}{\theta} + o(1) \quad \text{as } c \to 0. \tag{14}$$

For the second term in (12), observe that

$$E\left[\frac{1}{W_t^3\sqrt{c}}\left(\overline{Y}_t - \frac{1}{\theta}\right)^2\right] = \frac{1}{\theta^3}E\left[\frac{1}{W_t^3}\left(\frac{t}{n*}\right)^{-1}Q_t^2\right] \to 1 \text{ as c } \to 0, \tag{15}$$

by the fact that $W_t \to 1/\theta$ w.p.1 as $c \to 0$, Assertions (i) and (ii) of Lemma 1, Anscombe's theorem and the uniform integrability of $Q_t^2, c > 0,$. The lemma follows by using (14) and (15) in (12).

3. Conclusions

1- We have proposed a sequential procedure for estimating the shape parameter of the type I Pareto distribution and provided a second-order asymptotic expansion for the regret. The procedure can be used for the data representing the losses for insurance policyholders or the insured values of homes. It specifies how many policyholders or homes should be selected and provides an estimate of θ based on this number.

2- Since the asymptotic regret is positive for small values of the cost of sampling, it would be interesting to find a sequential procedure whose asymptotic regret is negative.

REFERENCES

[1] Anscombe, F., 1952, Large sample theory of sequential estimation, Proeedings Cambridge Philos. Soc., 48, 600-607.

[2] Aras, G. and Woodroofe, M., 1993, Asymptotic expansions for the moments of a randomly stopped average, Annals of Statistics, 21, 503-519.

[3] Chow, Y. S., Hsiung, C. A., and Lai, T. L., 1979, Extended renewal theory and Moment convergence in Anscombe's theorem, the Annals of Probability, 7, 304-318.

[4] Doob, J.L., 1953, Stochastic Processes, Wiley, New York

[5] Ghosh, M. and Mukhopadhyay, N.,1979, Sequential point estimation of the mean when the distribution is unspecified, Commun. Statist.-Theory & Methods, A8, 637-652.

[6] Loève, M., 1977, Probability Theory, 4th Ed., Springer-Verlag, New York.

[7] Starr, N. and Woodroofe, M., 1969, Remarks on sequential point estimation, Proc. Nat. Acad. Sci., U.S.A., 63, 285-288.

[8] Martinsek, A.T., 1983, Second order approximation to the risk of a sequential Procedure, Annals of Statistics, 11, 827-836.

[9] Tahir, M., 1989, An asymptotic lower bound for the local minimax regret in sequential point estimation, Annals of Statistics, 17, 1335-1346

[10] Woodroofe, M., 1977, Second order approximations for sequential point and interval estimation, Annals of Statistics, 5, 984-995.