Application of Generalized Confidence Interval in the Study of Web Performance

Dais George[1], Pit Pichappan[2], Sebastian George[3]
[1]Catholicate College
Pathanamthitta
Kerala, India
[2]Faculty of Computer and Information Sciences
Al Imam University
Riyadh
[3]St.Thomas College
Palai, Kerala
India
{daissaji@rediffmail.com, pichappan@dirf.org, sthottom@gmail.com}

ABSTRACT : *Much of the recent research has been aimed at improving web performance and scalability. For attaining the goal of improving web performance the basic need is the understanding of WWW work loads. In this paper we present a method useful for the system engineer to improve the service performance of a Web server through session-based Web workload, the best indicator of the users perception of the Web quality. Bytes transferred per session is one of the characteristics of intra-session which collectively describe session-based Web workload. This characteristic exhibits heavy-tailed behavior and its distribution match well with the Pareto Type I distribution [Goseva-Popstojanova et al. (2006)]. So for the performance study, we estimate the probability,* $R = P(X > Y)$*, when X and Y are two independent but not identically distributed random variables following Pareto Type I distribution, using the maximum likelihood estimator and Hill estimator. Extensive simulation studies are carried out to study the performance of these estimators. A generalized two-sided confidence interval for R of the Pareto type I distribution is constructed. The derived confidence interval suits both small samples and large samples. The average width and the coverage probability of this confidence interval is compared with the usual asymptotic confidence intervals through simulations. Using real data, we illustrate how R and generalized confidence interval of R can be used for improving the service performance of a Web server.*

Keywords: Generalized confidence interval, Generalized pivotal quantity, Heavy tail distributions, Hill Estimator, Maximum Likelihood Estimator, Pareto distribution

Received: 17 November 2010, Revised 22 July 2011, Accepted 25 July 2011

1. Introduction

Recently, we see a robust development in information technology and its impact on networks such as the World Wide Web (WWW) so that it acts as an information super highway. Millions of people all around the world access miscellaneous information scattered over the Web. WWW traffic has increased from 74 megabytes per month in 1992 to several million tera-bytes by the beginning of 21st century. Its tremendous growth brought huge challenges to system engineers, Web site designers, maintainers and content producers. A clear understanding of the WWW workloads and its characterization is fundamental to the goal of improving Web performance.

The alarming growth of Web traffic has sparked much research activity on improving the World Wide Web. Though there are several studies reported in the literature [Braun and Claffy (1994), Bestavros et al. (1995), Arlitt and Williamson (1997), Fengbin et al. (2007)] most studies focus on characterizing Web clients, rather than Web servers. In our earlier work [Dais and Sebastian (2007)], we studied the workload characteristics of Internet Web servers, using a college Web server data.

In this paper, we are considering the performance study, especially service performance of a Web server with due importance to user sessions. A session is defined as a sequence of requests from the same user during a single visit to the Web site. We can see considerable amount of research work focussing on characterizing Web user sessions for different purposes such as capacity planning, finding user navigational patterns etc. in the literature. Arlitt (2000) presented a detailed characterization of user sessions of the 1998 World Cup Web site and showed how these characteristics can be utilized in improving Web server performance. Goseva-Popstojanova et al. (2006) introduced several inter-session and intra-session characteristics which collectively describe session-based workload. In this work we are concentrating on one intra-session characteristic, bytes transferred per session. This characteristic exhibits heavy-tailed behavior and its distribution match well with the Pareto Type I distribution [Goseva-Popstojanova et al. (2006)].

The Pareto distribution, is a power law probability distribution that coincides with social, scientific, geophysical, actuarial and many other types of observable phenomena. The univariate Pareto distribution is a simple model for non- negative data with a power law probability tail. It is a useful model in the analysis of income data, reliability studies, risk modeling and business failure data [Lomax (1954)]. Arnold and Press (1983) gave an extensive historical survey of its uses in the context of income distribution. Jan Beirlant et al. (1996), Embrechts et al. (1997), Reed (2003) and Vandewalle et al. (2007) discuss the applications of the Pareto distribution in various fields. The sizes of human settlements, file size distribution of internet traffic which uses the TCP Protocol, clusters of Bose-Einstein condensate near absolute zero, the values of oil reserver in oil fields, the length of distribution in jobs assigned in super computers, the standardized price returns in individual stocks, sizes of sand particles, sizes of materiorites, number of species per genes, areas burn in forest fires and severity of large casuality losses for certain lines of business such as general liability, commercial auto and workers compensation are examples of random variables following the Pareto Type I distribution.

Power laws have been discovered for Web file sizes, Web site connectivities and the router connection degrees. A Web file size distribution is important for Web servers scheduling. More importantly, these discoveries motivate us to identify the mechanisms behind the observed power laws and thereby enable us to design mechanisms to improve the current operational structure of the Web server.

In most cases $R = P (X > Y)$ where X and Y are two independent but not identically distributed random variables following the same distribution is used as a measure of reliability or for conducting stress-strength analysis. It may be noted that R is of greater importance because it provides a general measure of the difference between two populations and has applications in many areas. For instance, R can be a measure of the effect of the treatment, if Y is the response of a control group and X refers to a treatment group. The function $P (X > Y) - P (X < Y)$is practically important in many situations including clinical trials and genetics. Bamber (1975) demonstrates that $A(X,Y) = P (X < Y) -1/2P (X = Y)$ is a useful measure of the size of the difference between two populations. The receiver operating characteristic (ROC) curves plot the true positive rate versus the false positive rate for a classification rule based on a continuously increasing sequence of cut off values [Venkataraman and Begg (1996)]. The area under the ROC curve (AUC) which is equivalent to the probability that a random observation X coming from the diseased population is larger than that from the non-diseased population, is commonly used as a measure of forecast quality. If this probability is extreme, it is inferred that the sample contains information useful for discrimination see, Briggs and Zaretzki (2008).

Here we are using $R = P (X > Y)$where X and Y are two independent but not identically distributed random variables following Pareto type I distribution, for studying the service performance of a Web server. This study is important because if one could fit a set of Web file size data as Pareto type I tailed distribution, which depends only on the tail index parameter, he can give it to the simulation engineer to estimate queueing delays. The simulation engineer would generate service times according to the given power law tail and we know that the simulation results will depend heavily on how long the simulation is run. So as long simulation length runs, the greater the probability of generating an unrealistically large service time.

In this paper, we consider, Pareto type I distribution with parameters α_1 and α_2 and estimate the desired probability $R = P(X > Y)$.This paper is organized as follows. Section 2 discusses about Pareto type I distribution. Estimation of tail index using

maximum likelihood estimation method and Hill estimation method is included in Section 3. Section 4 focus on the evaluation of R, when X and Y are two independent but not identically distributed random variables belonging to Pareto Type I distribution and the various steps for estimating R. In Section 5, we estimate R using the above mentioned tail index estimators and performance of these estimators of R has been successfully studied. In Section 6, the generalized confidence interval for R is obtained and a comparative study is carried out with the asymptotic confidence intervals. Using real data, we illustrate how R and generalized confidence interval of R can be used for improving the service performance of a Web server in Section 7.

2. Pareto Type-I Distribution

Heavy tails are characteristics of many phenomena where the probability of a single huge value impacts heavily. Record-breaking insurance losses, financial log returns, file sizes stored on a server, transmission rates of files are examples of heavy tailed phenomena.

The distribution of a random variable X is said to have heavy tail if :

$$P[X > x] = x^{-\alpha} L(x), \text{ where L is slowly varying.}$$

That is, for $x > 0$

$$\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1$$

This means that regardless of the distribution for small values of the random variable if the asymptotic shape of the distribution is hyperbolic, it is heavy tailed. The simplest heavy tailed distribution is the Pareto distribution which is hyperbolic over its entire range and has the probability density function,

$$f(x) = \alpha c^{\alpha} x^{-\alpha-1}, \alpha, c > 0, x \geq c \qquad (2.1)$$

and cumulative distribution function is given by:

$$F(x) = P[X \leq x] = 1 - (c/x)^{\alpha}, x \geq c$$

where α is the shape parameter and tail index and *C* represents the smallest value, that the random variable can take. As α decreases, an arbitrarily large portion of the probability mass may be present in the tail of the distribution.

3. Estimation of Tail Index α

To obtain the estimate of R, we have to estimate α_1 and α_2. For that we are using three estimators, viz. maximum likelihood estimator, Hill estimator and harmonic moment estimator.

3.1 The Maximum Likelihood Estimator

For a sample $X_1, X_2, ..., X_n$ from Pareto (α) as described by (2.1), the MLE of is readily derived, (see, Arnold (1983)) as

$$\hat{\alpha}_{MLE} = \frac{1}{n^{-1}\sum_{i=1}^{n} \ln X_i - \log c} \qquad (3.1)$$

For typical parametric models, the MLE proves highly efficient. For large sample size n, it attains the minimum possible variance among a large class of competing estimators and is approximately normally distributed. But the non-robust nature of MLE, in the presence of departures of the actual data from the assumed parametric model degrades the performance of the MLE severely. One thus seek to replace the MLE by a competitor which trades of some degree of efficiency in return for a favorable degree of robustness. That is why we desire an estimator which maintains satisfactorily high performance over a specified range of departures from the ideal model, which also being not too less efficient than the MLE Thus we prefer Hill estimator.

3.2 The Hill Estimator α

Suppose $X_1, X_2, \ldots X_n$ are iid from a distribution F and $X_{(1)} > X_{(2)} > X_{(n)}$ be their order statistics. If *F* has an exact Pareto distribution,

$$1 - F(x) = x^{-\alpha}, x > 1,$$

then

$$H_n = 1 / n \sum_{i=1}^{n} \ln X_{(i)} \tag{3.2}$$

is the MLE of α^{-1}.

If instead of assuming a Pareto distribution, we only assume

$$1 - F(x) = x^{-\alpha} L(x), x \to \infty,$$

where L is slowly varying. That is, for x > 0,

$$\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1$$

then we may pick k < n, so that the Hill estimator [Hill(1975)] is

$$H_{k,n} = (1 / k) \sum_{i=1}^{k} \ln \left(\frac{X_{(i)}}{X_{(k+1)}} \right) \tag{3.3}$$

where *k* is the number of upper order statistics used in the estimation. The rough idea behind using only *k* upper order statistics is that we should only sample from that part of the distribution which looks most Pareto like.

3.2.1 Properties of Hill Estimator

The Hill estimator of the tail index α is a pseudo-maximum likelihood estimator based on the exponential approximation of the normalized log-spacings,

$$Y_j = \ln \left(\frac{X_{(i)}}{X_{(k+1)}} \right) \text{ for } j = 1, \ldots, k.$$

The following are some of the important properties of the Hill estimator.

• If $n \to \infty$, $k \to \infty$, $(k / n) \to 0$, and the Hill estimator is a consistent estimate of $(1/\alpha)$ [Hill (1975)].

• The Hill estimator can be surprisingly sensitive to changes in location. A shift in location does not theoretically affect the tail index, but may throw the hill estimate way off.

• The Hill estimator is asymptotically normal [Resnick (1997)].

$$\text{That is } \sqrt{(k)}[H(k, n)] - \alpha^{-1}] \Rightarrow N(0, \alpha^{-1})$$

In practice, the hill estimator is used as follows. We graph $\{ (k, H_{k,n}^{-1}), 1 \leq k \leq n \}$ and the graph looks stable so that we pick out a value of $\alpha = H_{k,n}^{-1}$.

The Hill estimator has optimality properties only when the underlying distribution is close to Pareto. If the distribution is far from Pareto there may be outrageous bias even for sample sizes such as 100000. The hill plot is not always so revealing. One of the difficulty that we are facing using the hill estimator is that the graph may exhibit considerable volatility and/or the true answer may be hidden in the graph.

There are certain improved versions of Hill estimators which overcomes the limitations of the Hill estimator to a great extent.

They are smooHill estimator (Resnick and Starica (1995)), weighted Hill estimator (Gomes et al. (2008)) and smooweighted Hill estimator (Dais and Sebastian (2009)).

4. Estimation of R

Let X and Y be random variables distributed as Pareto type I distribution with parameters α_1 and α_2 respectively. The probability density functions of X and Y are given by

$$f_1(x) = \alpha_1 c^{\alpha_1} x^{-(\alpha_1+1)}, \ \alpha_1, c > 0, x \geq c \qquad (4.1)$$

and

$$f_2(x) = \alpha_2 c^{\alpha_2} x^{-(\alpha_2+1)}, \ \alpha_2, c > 0, x \geq c \qquad (4.2)$$

Using equations (4.1) and (4.2), the function R will be

$$R = P(X > Y) = \frac{\alpha_2}{\alpha_1 + \alpha_2} \qquad (4.3)$$

Clearly R depends only on α_1 and α_2.

As $\alpha_1 \to \alpha_2$, using L'Hospital rule, $R = P(X > Y) \to 1/2$.

Note that

$$R = \frac{1}{1+a}, a = \frac{\alpha_1}{\alpha_2}$$

and

$$\frac{dR}{da} = \frac{-1}{(1+a)^2} \leq 0$$

Therefore, R is a decreasing function in $a = \alpha_1 / \alpha_2$. That is for fixed $\alpha_1(\alpha_2)$, R is increasing(decreasing) function in $\alpha_2(\alpha_1)$.

We have conducted a simulation study to estimate R of the Pareto type I distribution using simulated observations from the distributions with $\alpha_1 = 0.8$ and $\alpha_1 = 1.6$ and $\alpha_2 = 1.5$.

The following steps will be considered for obtaining the numerical results.

Step I

Generate 1000 random samples X_i , i =1,2,...,n from the Pareto type I distribution with sample sizes n =15, 20, 30, 50 and 100 with $\alpha_1 = 0.8$ and $\alpha_1 = 1.6$.

Step II

Using the maximum likelihood estimation technique and Hill estimation technique, obtain 1000 estimates of $\hat{\alpha}_1$.

Step III

Similarly generate 1000 random samples from the Pareto distribution with parameter $\alpha_2$and obtain $\hat{\alpha}_2$.

Step IV

Using equation (4.3), we obtain the estimate of R using the maximum likelihood estimates of α_1 and α_2 respectively.

5. A Performance Study

An extensive numerical investigation is carried out to study the performance of these estimators using the simulated random samples from the Pareto Type I distribution in terms of the following measures.

• Average bias of the simulated N estimates of R:

$$\frac{1}{N} \sum_{i=1}^{N} (\hat{R}_i - R)$$

• Average bias of the simulated N estimates of R:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{R}_i - R)^2$$

	ML Estimate of R			Hill Estimate of R		
(n,m)	$\hat{R}_{MLE}$	Bias	MSE	$\hat{R}_{Hill}$	Bias	MSE
(15,15)	0.6477	-0.0049	0.0067	0.6483	-0.0051	0.0071
(15,20)	0.6463	-0.0071	0.0061	0.6472	-0.0061	0.0063
(15,30)	0.6452	-0.0065	0.0054	0.6457	-0.0058	0.0056
(15,50)	0.6450	-0.0091	0.0047	0.6451	-0.0078	0.0049
(15,100)	0.6427	-0.0093	0.0043	0.6423	-0.0084	0.0044
(20,15)	0.6502	-0.0024	0.0060	0.6517	0.0024	0.0056
(20,20)	0.6480	-0.0041	0.0051	0.6497	-0.0035	0.0052
(20,30)	0.6475	-0.0045	0.0044	0.6482	-0.0043	0.0043
(20,50)	0.6457	-0.0051	0.0037	0.6461	-0.0054	0.0036
(20,100)	0.6451	-0.0067	0.0033	0.6458	-0.0066	0.0031
(30,15)	0.6519	0.0009	0.005	0.6527	-0.0008	0.0052
(30,20)	0.6507	-.00026	0.0042	0.6511	-0.0003	0.0044
(30,30)	0.6492	-0.0029	0.0034	0.6496	-0.0026	0.0035
(30,50)	0.6482	-0.0027	0.0027	0.6486	-0.0023	0.0031
(30,100)	0.6475	-0.0039	0.0023	0.6478	-0.0032	0.0024
(50,15)	0.6552	0.0021	0.0043	0.6558	-0.0033	0.0046
(50,20)	0.6535	0.0004	0.0036	0.6539	-0.0004	0.0041
(50,30)	0.6515	-0.0009	0.0027	0.6518	-0.0008	0.0032
(50,50)	0.6507	-0.0014	0.0020	0.6509	-0.0022	0.0029
(50,100)	0.6500	-0.0020	0.0016	0.6502	-0.0024	0.0016
(100,15)	0.6551	0.0031	0.0038	0.6554	0.0028	0.0034
(100,20)	0.6536	0.0023	0.0031	0.6541	0.0027	0.0028
(100,30)	0.6526	0.0017	0.0021	0.6529	0.0018	0.0023
(100,50)	0.6525	-0.0003	0.0015	0.6526	-0.0003	0.0018
(100,100)	0.6515	-0.0014	0.001	0.6526	-0.0018	0.0011

Table 1.1 Comparison of Estimators of R
with $\alpha_1 = 0.8$ and $\alpha_2 = 1.5$, R = 0.6522, k = 10

The results are given in Tables 1.1-1.2. The absolute bias of the maximum likelihood estimate and Hill estimate is very small if the sample size of both the populations are small and if the sample size of the population with parameter α_2 is small. The MSE is symmetric with respect to α_1 and α_2 . Though we have considered different choice for the parameter α_1, there is no appreciable difference in the average bias and average MSE accordingly.

	ML Estimate of R			Hill Estimate of R		
(n,m)	$\hat{R}_{MLE}$	Bias	MSE	$\hat{R}_{Hill}$	Bias	MSE
(15,15)	0.4852	0.00001	0.008	0.4847	0.00002	0.0083
(15,20)	0.4833	-0.0025	0.0071	0.4835	-0.0028	0.0076
(15,30)	0.4786	-0.0052	0.006	0.4794	-0.0055	0.0065
(15,50)	0.4785	-0.005	0.0054	0.4789	-0.0051	0.0052
(15,100)	0.4754	-0.0062	0.0047	0.4768	-0.0064	0.0049
(20,15)	0.4853	0.0026	0.0071	0.4867	-0.0030	0.0076
(20,20)	0.4852	0.0021	0.006	0.4858	-0.0022	0.0064
(20,30)	0.4845	-0.0014	0.0053	0.4842	-0.0016	0.0057
(20,50)	0.4822	-0.0029	0.0044	0.4826	0.0028	0.0048
(20,100)	0.4794	-0.0047	0.0037	0.4799	0.0048	0.0037
(30,15)	0.4874	0.0049	0.0064	0.4887	-0.0042	0.0062
(30,20)	0.4864	0.0041	0.0051	0.4872	-0.0038	0.0050
(30,30)	0.4837	-0.0005	0.0041	0.4843	-0.0005	0.0040
(30,50)	0.4831	-0.0006	0.0033	0.4836	0.0006	0.0031
(30,100)	0.4812	-0.0024	0.0027	0.4819	0.0021	0.0024
(50,15)	0.4901	0.0065	0.0054	0.4896	-0.0059	0.0053
(50,20)	0.4881	0.0033	0.0042	0.4885	-0.0034	0.0042
(50,30)	0.4862	0.0022	0.0033	0.4869	-0.0023	0.0035
(50,50)	0.4842	0.0008	0.0024	0.4840	0.0007	0.0026
(50,100)	0.4833	-0.0012	0.0018	0.4834	0.0014	0.0013
(100,15)	0.4907	0.0073	0.0048	0.4903	-0.0072	0.0046
(100,20)	0.4894	0.0052	0.0038	0.4898	-0.0053	0.0036
(100,30)	0.4864	0.0038	0.0027	0.4861	-0.0037	0.0025
(100,50)	0.4846	0.0015	0.0018	0.4839	0.0014	0.002
(100,100)	0.4833	0.0004	0.0012	0.4834	0.0004	0.0013

Table 1.2 Comparison of Estimators of R with $\alpha_1 = 1.6$, $\alpha_2 = 1.5$, R = 0.4839, k =10

6. Generalized Confidence Interval of R

Generalized confidence interval for R is computed using the percentiles of a generalized pivotal quantity (GPQ), say G(**X, x**, α_1 ,α_2) function of **X, x**,α_1 and α_2 satisfying the following conditions.

• For a given x, the distribution of G(**X, x**,α_1 ,α_2) is free all unknown parameters and

• The "observed value" of G(**X, x,** α_1 ,α_2), namely its value at **X = x** is, R = $\frac{\alpha_2}{\alpha_1+\alpha_2}$, the parameter of interest.

When the conditions (1) and (2) hold, appropriate quantiles of G(**X, x**, R) form a 1-α confidence interval for R.

If G_p is the P^{th} quantile of G (**X, x,** R), then $(G_{\alpha/2}, G_{1-\alpha/2})$ is 1-α confidence interval for R. Such confidence intervals are referred to as generalized confidence intervals for R.

Numerous applications of generalized confidence intervals have appeared in the literature. Several such applications are given in Weerahandi (1993,1995a).

Here we are constructing the generalized confidence interval for R. Let $X_1, X_2,...X_n$ be a random sample from Pareto(α) distribution. The maximum likelihood estimate of α is

$$\hat{\alpha} = \frac{n}{\Sigma \ln X_i}$$

$$= \frac{n}{T} \text{ Where } T = \Sigma \ln X_i$$

The GPQ of α is obtained as

$$G_\alpha = \frac{T}{U} \text{ Where } T = \Sigma \ln X_i \text{ and } U = \frac{T^*}{\alpha} \text{Gamma } (1, n) \qquad (6.1)$$

This G_α satisfies the two conditions of GPQ mentioned earlier. Replacing the parameters in (4.3) by their GPQ's we get a GPQ for R as

$$G_R = \frac{G\alpha_2}{G\alpha_1 + G\alpha_2} \qquad (6.2)$$

For a given $\hat{R}_0$, the distribution of G_R does not depend on any unknown parameters and so Monte Carlo simulation can be used to estimate the percentiles of G_R. We compute a 95% confidence interval and hence the average length of the confidence interval of R, for various values of α_1 and α_2=1.5 and for various sample sizes can be computed. The coverage probabilities of each interval is calculated. This is shown in Table 2.1.

Through simulation study, using ML estimation method and Hill estimation method we can obtain the average length of the asymptotic 95% confidence intervals of R for α_1=0.8,1.6 and α_2=1.5 using the expression

$$\frac{1}{N}\sum_{i=1}^{N} 2\,(1.96)\,\text{Var}\,(\hat{R}_i),$$

for various sample sizes. Also we can find out the coverage probabilities of each interval.

Suppose $X_1, X_2,... ,X_m$ is a random sample of size m from the Pareto type I distribution with unknown parameter α_1 and is an $Y_1,Y_2,...,Y_n$ independent random sample of size n from the Pareto type I distribution with unknown parameter α_2.

The log-likelihood function based on the two independent random samples is given by

$$1(\alpha_1,\alpha_2) = \sum_{i=1}^{m} \ln[g(x_1, \alpha_1)] + \sum_{j=1}^{n} \ln[g(y_j, \alpha_2)]$$

$$\alpha\ m \ln(\alpha_1) + n \ln(\alpha_2) + (\alpha_1,\alpha_2) \ln c$$

$$-(\alpha_1 + 1)\sum_{i=1}^{m} \ln x_i - (\alpha_2 + 1)\sum_{j=1}^{n} \ln y_j$$

so that the information matrix,

$$I = \begin{bmatrix} \frac{m}{\alpha_1^2} & 0 \\ 0 & \frac{n}{\alpha_2^2} \end{bmatrix}$$

Theorem: *As* $m \to \infty, n \to \infty$

$$[\sqrt{m}(\hat{\alpha} - \alpha_1), \sqrt{n}(\hat{\alpha} - \alpha_2)] \xrightarrow{d} N\left(0, \text{diag}\left\{\frac{1}{a_{11}}, \frac{1}{a_{22}}\right\}\right) \text{ where}$$

$$l_{11} = -E\left(\frac{\partial^2 l}{\partial \alpha_1^2 l}\right) \text{and } l_{11} = -E\left(\frac{\partial^2 l}{\partial \alpha_1^2}\right)$$

so that

$$a_{11} = \lim_{m, n \to \infty} \frac{1}{m} l_{11} = \frac{1}{\alpha_1^2},$$

$$a_{22} = \lim_{m, n \to \infty} \frac{1}{n} l_{22} = \frac{1}{\alpha_2^2},$$

The point estimator of R is obtained using the maximum likelihood estimates (MLE) of α_1 and α_2 as $\hat{R}_{MLE} = \hat{R}(\hat{\alpha}_1, \hat{\alpha}_2)$

Then using delta method,

$$b_1(\alpha_1 \text{ and } \alpha_2) = \frac{-\alpha_2}{(\alpha_{1+} \alpha_2)^2}$$

$$b_2(\alpha_1 \text{ and } \alpha_2) = \frac{-\alpha_1}{\alpha_2} b_1(\alpha_1, \alpha_2) \tag{6.3}$$

and

$$\text{Var}(\hat{R}_{MLE}) \alpha_1^2 b_1^2(\alpha_1, \alpha_2)\left(\frac{1}{m}, \frac{1}{n}\right) \tag{6.4}$$

Thus we have the following result

As $m \to \infty, n \to \infty$

$$\frac{\hat{R}_{MLE} - R}{\alpha_1 b_1(\alpha_1, \alpha_2)\sqrt{\frac{1}{m}, \frac{1}{n}}} \xrightarrow{d} N(0,1)$$

Then the asymptotic 95% confidence interval for R is given by

$$\hat{R}_{MLE} \mp 1.96 \hat{\alpha}_1 a_1 b_1(\alpha_1, \alpha_2)\sqrt{\frac{1}{m} + \frac{1}{n}} \tag{6.5}$$

and the average length of the asymptotic 95% confidence interval of R is

$$\frac{1}{N}\sum_{i=1}^{n} 2(1.96)\hat{\alpha}_1 a_1 b_1(\hat{\alpha}_1, \hat{\alpha}_2)\sqrt{\frac{1}{m} + \frac{1}{n}}$$

Similarly we can find out the point estimators of R using Hill estimates α_1 of α_2 and .

If we are using Hill estimates of α_1 and α_2,

$$\text{Var}(\hat{R}_{\text{Hill}}) = 2\alpha_1^2 b_1^2(\alpha_1, \alpha_2)$$

and as $m \to \infty, n \to \infty, k \to \infty, (k/n) \to 0$,

$$\frac{\hat{R}_{\text{Hill}} - R}{\sqrt{2}\alpha_1^2 b_1^2(\alpha_1, \alpha_2)} \xrightarrow{d} N(0,1).$$

The asymptotic 95% confidence interval for R is given by

$$\hat{R}_{Hill} \mp 1.96\hat{\alpha}_1 a_1 b_1(\hat{\alpha}_1,\hat{\alpha}_2)\sqrt{2} \qquad (6.6)$$

and the average length of the asymptotic 95% confidence interval of R is

$$\frac{1}{N}\sum_{i=1}^{n} 2\sqrt{2}(1.96)\hat{\alpha}_1 a_1 b_1(\hat{\alpha}_1,\hat{\alpha}_2)$$

Now we compare the average confidence length of the generalized confidence interval with that of the asymptotic confidence intervals. The results are given in Table 2.

	Generalized Confidence Interval for R		**Asymptotic Confidence Interval for R Using ML Estimator**		**Asymptotic Confidence Interval for R Using Hill Estimator**	
(n,m)	Average Confidence Length	Coverage Probability	Average Confidence Length	Coverage Probability	Average Confidence Length	Coverage Probability
(15,15)	0.1129	0.9530	0.3182	0.9276	0.3182	0.9276
(15,20)	0.1121	0.9504	0.3012	0.9264	0.3012	0.9264
(15,30)	0.1108	0.9472	0.2816	0.9313	0.2816	0.9313
(15,50)	0.1109	0.9475	0.2622	0.9339	0.2622	0.9339
(15,100)	0.1095	0.9482	0.2574	0.9341	0.2574	0.9341
(20,15)	0.1129	0.9526	0.3005	0.9216	0.3005	0.9216
(20,20)	0.1099	0.9490	0.2811	0.9309	0.2811	0.9309
(20,30)	0.1100	0.9452	0.2663	0.9354	0.2663	0.9354
(20,50)	0.1110	0.9502	0.2423	0.9398	0.2423	0.9398
(20,100)	0.1100	0.9484	0.2204	0.9381	0.2204	0.9381
(30,15)	0.1110	0.9491	0.2864	0.9291	0.2864	0.9291
(30,20)	0.1102	0.9464	0.2638	0.9324	0.2638	0.9324
(30,30)	0.1094	0.9499	0.2334	0.9389	0.2334	0.9389
(30,50)	0.1101	0.9458	0.2102	0.9442	0.2102	0.9442
(30,100)	0.1111	0.9520	0.2062	0.9452	0.2062	0.9452
(50,15)	0.1113	0.9478	0.2610	0.9328	0.2610	0.9328
(50,20)	0.1082	0.9431	0.2394	0.9321	0.2394	0.9321
(50,30)	0.1092	0.9488	0.2208	0.9408	0.2208	0.9408
(50,50)	0.1097	0.9531	0.1712	0.9453	0.1712	0.9453
(50,100)	0.1115	0.9525	0.1601	0.9425	0.1601	0.9425
(100,15)	0.1092	0.9500	0.2462	0.9310	0.2462	0.9310
(100,20)	0.1104	0.9464	0.2184	0.9310	0.2184	0.9310
(100,30)	0.1093	0.9500	0.1886	0.9398	0.1886	0.9398
(100,50)	0.1105	0.9511	0.1572	0.9411	0.1572	0.9411
(100,100)	0.1091	0.9502	0.1268	0.9472	0.1268	0.9472

Table 2.1. Comparison of Confidence Intervals of R with $\alpha_1 = 0.8$, $\alpha_2 = 1.5$, R = 0.6522

	Generalized Confidence Interval for R		Asymptotic Confidence Interval for R Using ML Estimator		Asymptotic Confidence Interval for R Using Hill Estimator	
(n,m)	Average Confidence Length	Coverage Probability	Average Confidence Length	Coverage Probability	Average Confidence Length	Coverage Probability
(15,15)	0.1126	0.9453	0.3460	0.9283	0.3496	0.9253
(15,20)	0.1139	0.9535	0.3247	0.9293	0.3289	0.9275
(15,30)	0.1132	0.9466	0.3020	0.9357	0.3052	0.9366
(15,50)	0.1122	0.9463	0.2818	0.9349	0.2844	0.9343
(15,100)	0.1119	0.9443	0.2657	0.9352	0.2671	0.9344
(20,15)	0.1145	0.9489	0.3250	0.9287	0.3282	0.9289
(20,20)	0.1125	0.9481	0.3022	0.9346	0.3052	0.9341
(20,30)	0.1123	0.9453	0.2766	0.9301	0.2791	0.9295
(20,50)	0.1139	0.9522	0.2544	0.9352	0.2570	0.9350
(20,100)	0.1132	0.9472	0.2361	0.9350	0.2380	0.9332
(30,15)	0.1118	0.9452	0.3019	0.9254	0.3048	0.9252
(30,20)	0.1133	0.9460	0.2770	0.9373	0.2793	0.9364
(30,30)	0.1137	0.9487	0.2486	0.9407	0.2502	0.9401
(30,50)	0.1107	0.9405	0.2231	0.9398	0.2248	0.9394
(30,100)	0.1143	0.9519	0.2015	0.9425	0.2032	0.9420
(50,15)	0.1127	0.9449	0.2822	0.9321	0.2862	0.9332
(50,20)	0.1140	0.9542	0.2547	0.9385	0.2583	0.9382
(50,30)	0.1126	0.9487	0.2232	0.9412	0.2258	0.9408
(50,50)	0.1125	0.9495	0.1939	0.9464	0.1960	0.9453
(50,100)	0.1124	0.9492	0.1683	0.9434	0.1694	0.9429
(100,15)	0.1127	0.9501	0.2661	0.9362	0.2687	0.9365
(100,20)	0.1109	0.9426	0.2363	0.9368	0.2371	0.9369
(100,30)	0.1118	0.9451	0.2017	0.9421	0.2026	0.9411
(100,50)	0.1128	0.9503	0.1684	0.9480	0.1690	0.9475
(100,100)	0.1122	0.9496	0.1378	0.9447	0.1375	0.9446

Table 2.2 Comparison of Confidence Intervals of R with $\alpha_1 = 1.6$, $\alpha_2 = 1.5$, R = 0.4839

From tables 2.1-2.2, it is clear that for the generalized confidence interval, the average length of the confidence interval is smaller than the average length of asymptotic confidence intervals for small samples as well as for large samples and the difference is large for small samples. Moreover the coverage probability is more close to the nominal value in the case of generalized confidence interval than in the case of asymptotic confidence intervals. So generalized confidence interval is preferable and it performs better in small and large samples.

7. Application

This study is very useful in improving the service performance of Web servers and thereby reducing the queueing delay. For the study we used the access log obtained from a college Web server. The data were collected from May 24, 2006 for 7 days and from May 24, 2007 for 7 days. Table 3.1 summarizes the raw data from the access log.

Item	College Web Server	
	I Data Set	**II Data Set**
Access Log Duration	1 week	1 week
Access Log Start Date	May 24/2006	May 24/2007
Access Log Size(MB)	4182	4182
Total Requests	54715	126,324
Average Requests/Day	7816	18046
Average Requests/Hour	326	752
Average Requests/ Peak Hour	674	1521
Total Bytes Transferred(MB)	317,295	600,577
Average Bytes/Day(MB)	45315	85796.71
Average Bytes/Hour(MB)	1888.661	3574.863
Total Bytes Transferred in Peak Hour(MB)	120,236	305,798
Average Bytes/Peak Hour(MB)	4294.143	10899.93

Table 3.1 Summary of Access Log Characteristics (Raw Data)

For data analysis first we have to study the response codes in the Web server access log. The HTTP server logs record all processed requests and the corresponding responses. The response status codes are returned to the client making the request and also recorded in the server's log file. Based on HTTP error standards, responses wit response codes between 400 and 599 are classified as failures. The HTTP response code standards are listed below:

100 to 199 Informational status codes, rarely used; 200 to 299 Successful; only 200 frequently used; 300 to 399 Warning - but the request may still be success; 400 to 499 Client Error, the request was invalid in some way and 500 to 599 Server Error, the server could not fulfill the (valid) request.

Here only 93.8% server responses are successful for the I data set and 89.1% are successful for the II data set.

Since the successful responses are responsible for all the documents transferred by the server, only these responses are used for the remaining analysis. Table 3.2 provides the statistical summary of the reduced data.

Item	College Web Server	
	I Data Set	II Data Set
Access Log Duration	1 week	1 week
Access Log Start Date	May 24/2006	May 24/2007
Access Log Size(MB)	4182	4182
Total Requests	51323	112,485
Average Requests/Day	7332	16070
Average Requests/Hour	306	670
Average Requests/ Peak Hour	604	1402
Total Bytes Transferred(MB)	317,295	600,577
Average Bytes/Day(MB)	45315	85796.71
Average Bytes/Hour(MB)	1888.661	3574.863
Total Bytes Transferred in Peak Hour(MB)	120,236	305,798
Average Bytes/Peak Hour(MB)	4294.143	10899.93

Table 3.2 Summary of Access Log Characteristics (Reduced Data Sets)

Considering first data set, for the the file size transferred (X), most files appear to be in the range 100-30000 bytes where as in the second data set, for file size transferred (Y), most files appear to be in the range 100-62000 bytes and since the tail of the distribution can be modeled by the Pareto type I distribution [Goseva-Popstojanova et al. (2006)], we consider file sizes larger than 100 bytes only. We estimated the tail index using Maximum Likelihood estimator and Hill estimator using formulae (3.1) and (3.2) and the results are given in Table 4.

Estimates of α	$\hat{\alpha}$	
	$\hat{\alpha}_1$	$\hat{\alpha}_2$
$\hat{\alpha}_{Hill}$	0.9512	0.8376
$\hat{\alpha}_{MLE}$	0.97	0.85

Here $\hat{R}$ = 0.4683. If R < .5 means that the random observation (file size transferred per unit time) from the first set is smaller than the second set. That is the tail index of the second data set $\alpha_2 < \alpha_1$, the tail index of the first data set. This means that the number of very large files transferred remains significant for the second data set. So if the traffic capacity of the Web server remains the same degradation in service performance will result. When the modification of the traffic capacity of the server is to be done, can be decided through the generalized confidence interval. Then as described in section 6, through Monte Carlo simulation we can find the 95% generalized confidence interval of R. Here the 95% confidence interval of R is obtained as (0.4488, 0.5512). If the value of R lies in this interval then no need of thinking about modification in traffic capacity. Here the value of R lies in the interval. So no need of modification now. But if the value of R lies outside the interval (R the lower limit), then the traffic capacity of the server has to be increased. Otherwise the server will take unrealistically large service time thereby degradation in service performance will result. So in order to avoid degradation in service performance, we have to calculate R at regular intervals (better at short intervals) and check whether it lies in the generalized confidence interval.

8. Conclusion

In this paper we have estimated the probability R = P(X > Y) with respect to two independent Pareto type I distributions with shape parameters α_1 and α_2. This study can be used in two dimensions. Firstly for comparing two distributions with common base distribution, this study can be used. Such problems arises mainly in medical studies, for studying the effectiveness of drugs. Extreme values of the estimated R will indicate the difference between the populations. Here we used this concept of R for improving the service performance of a Web server and it will be very helpful for the system designers and simulation engineers to meet their challenge by reducing queueing delay and improving service performance of Web servers. Secondly this study can be used as a measure of reliability or for conducting stress-strength analysis. As a failure time distribution, this measure can be used as a measure of reliability for Pareto type I distribution. Again in practical situations sometimes we have to deal with small samples and sometimes with large samples. In this study we considered both sample sizes and the simulation study suggested the effectiveness of the Maximum Likelihood estimate and Hill estimate of R for both samples. We also introduced the generalized confidence interval for R and proved that it is better than the asymptotic confidence interval of R for both small and large samples.

References

[1] Arlitt, M., Williamson, C. (1997). Internet Web servers: workload characterization and performance implications, *IEEE/ACM Trans. Netw*., 5 (5) 631-645.

[2] Arlitt, M. (2000). Characterizing Web user sessions, SIGMETRICS Perform. Eval. Rev., Vol.28,2, 50-63.

[3] Arnold, B.C., Press, S.J. (1983). Bayesian inference for Pareto population, *Journal of Econometrics*, 21, 287-306.

[4] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *J. Math. Psychol*., 12, 387-415.

[5] Bestavros, A., Caeter, R., Crovella, M. (1995). Application level of document catching in the Internet, *In*: Second Int'l Workshop on Services in Distributed and Networked Environments,166-173.

[6] Braun, H. and Claffy, K. (1994). Web server characterization, an assessment of the impact catching docements from NCSA's Web server, Chicago, Illinois.

[7] Briggs, W.M., Zaretzki, R. (2008). The skill plot: a graphical technique for evaluating continuous diagnostic tests, Biometrics, 63, 250-261.

[8] Dais George, Sebastian George (2008). A study on the characteristics of Web server data, Int'l Conference on Data Engineering and Management, Thiruchirappally, 170-173.
[9] George, Dais., Sebastian George (2009). Analyzing tail heaviness of Web server data, *International Journal of Web Applications,* 1 (4) 201-214.
[10] Embrechts, P., Kluppelberg, C., Mikosh, T. (1997). Modeling extremal events for insurance and finance, Springer-Verlag, Berlin Heidelberg.
[11] Fengbin, Li, Goseva-Popstojanova, K., Arun Ross (2007). Discovering Web workload characteristics through cluster analysis, IEEE Int'l Symposium on Network Computing and Applications, Cambridge, USA, 61-68.
[12] Goseva-Popstojanova, K., Li, F., Wang, X., Sangle, A. (2006). A contribution towards solving the Web workload Puzzle, Int'l Conf. Dependable systems and Networks, 505-514.
[13] Hill, B.M. (1975). A simple general approach to interface about the tail of a distribution, *Ann. Statistica*, 3, 1163-1174.
[14] Ivette Gomes, M., Laurens de Hann and Ligia Rodrigues (2008). Tail index estimation for heavy tailed model: accomodation of bias in weighted log excesses, *Journal of Royal Statistical Society:Series B* (Statistical Methodology), V. 70, 31-52.
[15] Jan Beirlant, Petra Vynckier, Jozel, L. (1996). Tail index estimation, Pareto quantile plots and regression diagnostics, J*ournal of the American Statistical Association*, V. 91.
[16] Lomax, K.S. (1954). Business failures: another example of the analysis of failure data, *J. Am. Statist. Assoc.*, 49, 847-852.
[17] Reed, W. J. (2003). The Pareto law of incomes - an explanation and an extension, *Physica*, A 319 , 469-485.
[18] Resnick, S.I. and Starica (1995). Consistency of Hill estimator for Department data, *Journal of Appl.Probab.* 32, 139-167.
[19] Vandewalle, B., Beirlant, J., Christmann, A., Hubert, M. (2007). A robust estimator for the tail index of Pareto-type distributions, Elsevier Science, January 8.
[20] Venkataraman, E.S., Begg, C.B. (1996). A distribution free procedure for comparing receiver operating characteristic curves from a paired experiment, Biometrica, 83 (4) 835-848.
[21] Weerahandi, S. (1993). Generalized confidence intervals, *Journal of the American Statistical Association*, 88, 899-905 .
[22] Weerahandi, S. (1995). Exact Statistical Methods for Data Analysis, Springer-Verlag, New York.