International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 5, No. 2 (Nov-2018)

http://dx.doi.org/10.12785/ijcts/050205

Generalized Inference for the Overlapping Coefficientof two Pareto Distributions

Sibil Jose[1] and Seemon Thomas[2]

[1]*Department of Statistics, St. George's College Aruvithura, Kottayam, Kerala, India.*
[2]*Department of Statistics, St. Thomas College Pala, Kottayam, Kerala, India.*

Received March 2, 2018, Revised August 2, 2018, Accepted September 4, 2018. Published November 1, 2018

Abstract: This paper introduces a new method, called GPQ method, for the computation of overlapping coefficient of two Pareto distributions. Expected lengths and coverage probabilities of the confidence intervals are also calculated using the generalized pivotal quantity. The comparison of the method is done with the best available method, that is, bootstrap percentile method. The general performance of the proposed method is better than the existing methods. An illustrative example is also presented.

Keywords: Pareto distribution, Overlapping coefficient, Generalized pivotal quantity, Percentile bootstrap, Coverage probability.

1. INTRODUCTION

Overlapping coefficient (OVL) is a statistical measure used to measure the degree of overlap between two statistical populations. It is the common area under two probability density functions. Reference [1] measured the overlap of income distributions of White and Negro families in the United States using the formula:

$$\text{OVL}=\int \min\{f_1(x), f_2(x)\}dx \qquad (1)$$

where $f_1(.)$ and $f_2(.)$ are the respective probability densities of the populations. Obviously, the value of OVL ranges from 0 to 1, where a value 0 indicates that there is no overlap or similarity and a value 1 indicates that the two populations are identical or coincident. If the characteristic under consideration is discrete in nature, the integral in (1) can be replaced with summation. In the literature one can find other measures of OVL, see for example, [2], [3], [4] and [5]. Reference [6] used the concept of OVL in testing the equality of two Pareto distributions.

A researcher often wants to study the similarity of distribution of income in two populations. Pareto distribution is a heavy tailed distribution that is a good choice for modeling income above a threshold value. It is a good choice in insurance applications for modeling extreme loss. Reference [7] summarises the distribution of top incomes in the UK using Pareto models.

When the probability distribution under consideration contains two or more parameters conventional inference procedures may not be applicable as one cannot find a statistic that is free of these parameters. Usually, OVL is a function of two or more parameters and hence the statistic for OVL consists of nuisance parameters. So conventional methods based on sufficient statistics are not available and hence it is necessary to consider alternative methods to deal with the inference of OVL. In this study we apply the method of Generalized Pivotal Quantity (GPQ) proposed by [8] and [9] to obtain confidence interval for OVL of two Pareto distributions. Reference [10] constructed generalized confidence intervals for the OVL of two normal distributions with equal variance. Reference [11] constructed generalized lower confidence limit for the reliability function of two parameter exponential distribution.

E-mail address: sibiljose60@yahoo.com, seemonpala@rediffmail.com

2. OVL OF ONE-PARAMETER PARETO DISTRIBUTIONS

Consider two Pareto distributions with the following probability density function:

$$f_i(x) = \begin{cases} \alpha_i x^{-(\alpha_i+1)}; x > 1, \alpha_i > 0 \\ 0; elsewhere \end{cases}$$

for i=1,2. Parameter α_i is the shape parameter ('tail index') describing the heaviness of the right tail of the distribution, with smaller values corresponding to greater tail heaviness. We shall denote the OVL defined in (1) by ρ and one can obtain the expression for ρ in this case as given below:

$$\rho = 1 - r^{\frac{1}{(1-r)}} \left| 1 - \frac{1}{r} \right| \qquad (2)$$

where $r=\alpha_1/\alpha_2$. It is obvious that ρ= 1 if r=1.

A. *GPQ Method*

Let X_{ij}, j=1,2,…,n_i, i=1,2 be two independent random samples of sizes n_1 and n_2 taken from two independent Pareto populations with parameters α_1 and α_2 respectively.

Define

$$U_i = 2\alpha_i \sum_{j=1}^{n_i} \ln(X_{ij})$$

for i=1,2. One can see that U_i's are independent chi-square random variables with $2n_i$ degrees of freedom, i=1,2. According to substitution method by [12], the corresponding GPQ for the parameter α_i is the following:

$$T_{\alpha_i} = \frac{U_i}{2\sum_{j=1}^{n_i} \ln(x_{ij})}; \text{ i=1,2.} \qquad (3)$$

Note that (3) is independent of nuisance parameters and its observed value is the parameter itself. Thus the GPQ of the OVL, ρ can be obtained by substituting (3) in (2).

B. *Bootstrap Percentile Method*

In percentile bootstrap method we shall first generate b bootstrap samples, say X_1^*, …,X_b^*. In the next step estimate OVL values $\hat{\rho}_1^*, \ldots, \hat{\rho}_b^*$. Then identify $100(\alpha/2)^{th}$ and $100(1-(\alpha/2))^{th}$ percentiles of the $\hat{\rho}^*$ as the percentile points of ρ and those points are taken to be the respective lower and upper limits of a 100(1 - α)% confidence interval for ρ.

C. *Simulation Study*

TABLE 1 gives the estimated coverage probabilities of the confidence intervals using the GPQ and the percentile bootstrap methods for the OVL of two Pareto distributions. The 95% nominal level confidence interval is constructed for different sample sizes and parameter values. Numerical results are obtained using 10,000 simulated samples and are computed using R codes. For each simulated sample, 10,000 values of the GPQ are generated in order to compute the confidence limits and for the bootstrap method 10,000 parametric bootstrap samples are generated. It can be observed that GPQ based confidence intervals provide much better coverage for most of the sample sizes. If the value of ρ is large, then the expected length of the GPQ intervals is found to be smaller than that of the percentile bootstrap intervals for almost all sample sizes.

TABLE 1. COVERAGE PROBABILITIES AND EXPECTED LENGTHS IN ONE PARAMETER CASE

Parameters	(n_1,n_2)	GPQ Method		Bootstrap Percentile Method	
		Coverage	*Length*	*Coverage*	*Length*
	(2,4)	0.9626	0.6590	0.9609	0.7183
α_1=1	(5,5)	0.9515	0.5195	0.9517	0.5196
	(10,10)	0.9476	0.3628	0.9471	0.3628
α_2=10	(10,20)	0.9529	0.3080	0.9475	0.3227
	(20,20)	0.9481	0.2515	0.9478	0.2516
ρ = 0.3030	(20,30)	0.9460	0.2280	0.9445	0.2318
	(50,50)	0.9488	0.1568	0.9495	0.1569
	(50,100)	0.9523	0.1351	0.9507	0.1364
	(100,100)	0.9505	0.1103	0.9506	0.1103
	(100,200)	0.9543	0.0953	0.9499	0.0958
	(2,4)	0.9765	0.7617	0.9937	0.7398
α_1=2	(5,5)	0.9751	0.5953	0.9749	0.5953
	(10,10)	0.9514	0.4590	0.9464	0.4171
α_2=5	(10,20)	0.9625	0.4583	0.9742	0.5018
	(20,20)	0.9483	0.3921	0.9479	0.3921
ρ = 0.6743	(20,30)	0.9464	0.3635	0.9449	0.3637
	(50,50)	0.9486	0.2590	0.9500	0.2590
	(50,100)	0.9527	0.2244	0.9510	0.2253
	(100,100)	0.9506	0.1834	0.9507	0.1834
	(100,200)	0.9542	0.1589	0.9538	0.1592
	(2,4)	0.9430	0.7399	0.9490	0.7311
α_1=4	(5,5)	0.9609	0.5831	0.9607	0.5835
	(10,10)	0.9686	0.4358	0.9692	0.4380
α_2=5	(10,20)	0.9721	0.3916	0.9827	0.3814
	(20,20)	0.9755	0.3288	0.9748	0.3239
ρ = 0.9180	(20,30)	0.9742	0.3019	0.9798	0.2924
	(50,50)	0.9753	0.2208	0.9763	0.2209
	(50,100)	0.9762	0.1990	0.9743	0.2095
	(100,100)	0.9743	0.1695	0.9706	0.1563
	(100,200)	0.9765	0.1533	0.9804	0.1461

3. OVL OF TWO PARAMETER PARETO DISTRIBUTIONS

Let us consider two independent Pareto distributions with the following probability density function:

$$f_i(x) = \begin{cases} \dfrac{\alpha_i \beta_i^{\alpha_i}}{x^{\alpha_i+1}}; x > \beta_i, \alpha_i > 0 \\ 0; elsewhere \end{cases}$$

for i=1,2. The point of intersection of the two probability density functions is the following:

$$x_0 = \left[\frac{\alpha_1}{\alpha_2} \left(\frac{\beta_1^{\alpha_1}}{\beta_2^{\alpha_2}} \right) \right]^{1/(\alpha_1 - \alpha_2)} \quad \text{if} \quad x_0 \in (\max(\beta_1, \beta_2), \infty).$$

Then the OVL of the two probability density functions can be expressed as follows:

On simplification the final expression for ρ reduces to:

$$\rho = \begin{cases} \left(\frac{\beta_2}{\beta_1}\right)^{\alpha_2} - \left(\frac{\beta_2}{x_0}\right)^{\alpha_2} + \left(\frac{\beta_1}{x_0}\right)^{\alpha_1} ; \beta_1 > \beta_2 \\ \left(\frac{\beta_1}{\beta_2}\right)^{\alpha_1} + \left(\frac{\beta_2}{x_0}\right)^{\alpha_2} - \left(\frac{\beta_1}{x_0}\right)^{\alpha_1} ; \beta_1 < \beta_2 \end{cases} \quad (4)$$

Let X_{ij}, j=1,2,…,n_i, i=1,2 be two independent random samples of sizes n_1 and n_2 taken from two independent Pareto populations. Let $X_{i(1)}$ be the smallest observation in the i^{th} sample. Then the estimators of the parameters are the following (see [13]):

$$\hat{\beta}_i = X_{i(1)}$$

and

$$\hat{\alpha}_i = n_i \left[\sum_{j=1}^{n_i} \ln\left(\frac{X_{ij}}{X_{i(1)}} \right) \right]^{-1}, \text{ i=1,2.}$$

Define,

$$U_i = \frac{2n_i\alpha_i}{\hat{\alpha}_i} = 2\alpha_i \left[\sum_{j=1}^{n_i} \ln\left(\frac{X_{ij}}{X_{i(1)}} \right) \right]$$

and

$$V_i = 2n_i\alpha_i \ln\left(\frac{X_{i(1)}}{\beta_i} \right), \text{ i=1,2.}$$

Note that U_i s are chi-square random variables with $2(n_i - 1)$ degrees of freedom and V_i s are also chi-square random variables with 2 degrees of freedom. Then the GPQs of the parameters can be expressed as follows:

$$T_{\alpha_i} = \frac{U_i}{2\left[\sum_{j=1}^{n_i} \ln\left(\frac{x_{ij}}{x_{i(1)}} \right) \right]}, \quad \text{i=1,2} \quad (5)$$

$$T_{\beta_i} = x_{i(1)} \exp\left\{ \frac{-V_i}{2n_i T_{\alpha_i}} \right\}, \text{ i= 1,2.} \quad (6)$$

Note that the probability distributions of (5) and (6) are free of nuisance parameters and their observed values are the respective parameters. Now the GPQ of ρ, say T_ρ, is obtained by substituting (5) and (6) in (4) according as $T_{\beta 1} > T_{\beta 2}$ or $T_{\beta 1} < T_{\beta 2}$ as the case may be. Then the $100(\alpha/2)^{th}$ and $100(1-(\alpha/2))^{th}$ percentile values of T_ρ will be the respective lower and upper limits of the $100(1 - \alpha)$% generalized confidence interval for ρ.

A. Simulation Study

TABLE 2 gives the results of a simulation study conducted to assess the performance of GPQ and percentile bootstrap methods in computing the coverage probabilities and expected length of the confidence intervals for ρ. The study is conducted for two different values of ρ. For each simulated sample, 10,000 values of the GPQ are generated in order to compute the confidence limits and for the bootstrap method, 10,000 parametric bootstrap samples are generated. It can be observed that GPQ method provide confidence intervals having much better coverage for ρ and shortest expected length.

TABLE 2. COVERAGE PROBABILITIES AND EXPECTED LENGTHS IN TWO PARAMETER CASE

Parameters	(n_1,n_2)	GPQ Method				Bootstrap perc. Method			
		Tail Coverage				*Tail Coverage*			
		Equal	Left	Right	Length	Equal	Left	Right	Length
α_1=3, β_1=2	(20,20)	0.9674	0	0.0323	0.1782	0.9373	0.0046	0.0581	0.3245
α_2=2, β_2=1	(20,30)	0.9658	0	0.0342	0.1735	0.9246	0.0030	0.0724	0.2825
ρ =0.2477	(50,50)	0.9591	0.0002	0.0407	0.1387	0.9386	0.0105	0.0509	0.2245
	(50,100)	0.9658	0.0004	0.0338	0.1136	0.9406	0.0063	0.0531	0.1717
	(100,100)	0.9521	0.0076	0.0403	0.1123	0.9480	0.0067	0.0453	0.1447
α_1=1, β1=1	(20,20)	0.9452	0.0002	0.0546	0.2227	0.9581	0.0270	0.0149	0.3324
α_2 =2, β2=2	(20,30)	0.9579	0.0012	0.0409	0.1825	0.9471	0.0411	0.0118	0.3323
ρ =0.4375	(50,50)	0.9467	0.0072	0.0461	0.1118	0.9454	0.0246	0.030	0.2052
	(50,100)	0.9555	0.0099	0.0346	0.0967	0.9457	0.0333	0.0210	0.1912
	(100,100)	0.9484	0.0135	0.0381	0.0807	0.9433	0.0268	0.0299	0.1471

4. EXAMPLE

To illustrate the method let us consider the data on the district wise per capita income of two states in India, namely, Kerala and Uttarakhand, for the financial year 2013-14. The values of 14 districts in Kerala are 114495, 84900, 104243, 104424, 114708 , 151210, 96647, 79552, 71727, 82243, 72909, 93906, 65216, 98246 and the values of 13 districts in Uttarakhand are 59791, 90173, 85156, 122804, 91708, 69401, 79981, 86699, 105960, 68730 , 72922, 122172, 115543. The Kolmogrov-Smirnov one sample test is used to check whether the data sets follow Pareto distribution. The maximum likelihood estimates of the parameters α_1 and β_1 of Pareto distribution for the first data set are 2.8129 and 65216 repectively.. The value of K-S test statistic is 0.2128 against the table value of 0.314 corresponding to 5% level of significance. The maximum likelihood estimates of the parameters α_2 and β_2 for the second data set are 2.5964 and 59791 respectively. The value of K-S test statistic obtained in this case is 0.2224 and the table value is 0.325 corresponding to 5% level of significance. Thus both the data sets follow Pareto distribution. The estimated value of the OVL of the two distributions is 0.7966 and the 95% confidence interval based on GPQ is (0.5555, 0.9035). Of course, an economist can make several inferences about these two populations based on the value of OVL.

ACKNOWLEDGMENT

The authors acknowledge the referees for their comments and suggestions that substantially improved this present manuscript.

REFERENCES

[1] M.S. Weitzman, “Measures of overlap of income distributions of White and Negro families in the United States”. Technical paper No.22, Department of Commerce, Bureau of Census, Washington, U.S.A, 1970.

[2] K. Matusita, “Decision rules based on the distance for problem of fit, two samples, and estimation,” Ann. Math. Statist., 26, 1955, pp. 631-640.

[3] K. Matusita, A distance and related statistics in multivariate analysis I. (P.R. Krishnaiah Ed.), Academic Press New York., 1966, pp. 187-200.

[4] M. Morisita, “Measuring interspecific association and similarity between communities,” Memoirs of the faculty of Kyushu University. Series E. Biology 3, 1959, pp. 64-80.

[5] R. Mac Arthur and R. Levins, “The limiting similarity, convergence and divergence of co existing species,” Amer. Nat. 101, Sep- Oct 1967, pp. 377-385.

[6] H. A. Bayoud, “Testing the equality of two Pareto distributions,” Proceedings of the World Congress on Engineering, II, London, UK, July 2017.

[7] S.P. Jenkins, “Pareto distributions, top incomes and recent trends in UK income inequality”, Economica, special issue, March 2016.

[8] S. Weerahandi, “Generalized confidence intervals,” J. Am. Stat. Assoc., 88, 1993, pp. 899-905,

[9] S.Weerahandi, Exact statistical methods for data analysis. Springer series in Statistics, New York, 1994.

[10] B. Reiser and D. Faraggi, "Confidence intervals for the overlapping coefficient: the normal equal variance case," The Statistician 48, September 1999, pp.413-418

[11] A. Roy, and T. Mathew, "A generalized confidence limit for the reliability function of a two- parameter exponential distribution," J. Statist. Plan. Inf. 128, 2005, pp.509-517.

[12] S. Weerahandi, Generalized inference in repeated measures. Wiley series in probability and statistics, New Jersey, 2004.

[13] N.L. Johnson, S. Kotz and N. Balakrishnan, Continuous univariate distributions, Vol.1, 2nd edition, Wiley Series in Probability and Statistics, 1994.