Discussiones Mathematicae
Probability and Statistics 38 (2018) 5–13
doi:10.7151/dmps.1199

INTERVAL ESTIMATION OF THE OVERLAP COEFFICIENT OF TWO NORMAL DISTRIBUTIONS

Sibil Jose

Department of Statistics
St. George's College Aruvithura
Kottayam, Kerala, India

e-mail: sibiljose60@yahoo.com

and

Seemon Thomas

Department of Statistics
St. Thomas College Pala
Kottayam, Kerala, India

e-mail: seemonpala@rediffmail.com

Abstract

Matusita's measure of overlap is considered for two normal distributions, without assuming equal variances, and a confidence interval is proposed using the generalized pivotal quantity approach. Simulation results show that the proposed method provides better coverage than bootstrap methods.

Keywords: coverage probability, bootstrap t, generalized pivotal quantity, Matusita's measure of overlap, percentile bootstrap.

2010 Mathematics Subject Classification: 62F99.

1. Introduction

The study of distance measures is important in addressing practical problems such as: hypothesis testing, classification, outlier detection and density estimation, to name a few. Distance or similarity criteria describe how close two distributions are. Similarity can sometimes be assessed based on random samples using graphical summaries such as histograms and boxplots, plotted side-by-side, or calculating summary measures such as the correlation coefficient. However

these measures do not always adequately represent the similarity, and this motivates the use of formal measures such as the overlap coefficient (OVL). In this note we shall consider the OVL to study the similarity of two normal populations, whose means could be different, and whose variances could also be different. The study of the OVL of two normal distributions typically assume equal variances; it is this assumption that is relaxed in our work. Furthermore, the OVL measure that we shall use is the one proposed by [5].

By definition, OVL is the area of intersection of the graphs of two probability density functions. Let $f_1(x)$ and $f_2(x)$ be probability density functions of two populations. Matusita's measure of OVL is defined as

$$\rho = \int \sqrt{f_1(x) f_2(x)} dx. \tag{1}$$

For discrete distributions, one can replace the integral in (1) with summation. Further, this can be generalized to multivariate distributions as well. As OVL gives the common area under two probability density functions, it's scale is from 0 to 1. It has value 0 if the two distributions are entirely different, and has value 1 if the two distributions are identical. Note that Matusita's measure is invariant under any one-to-one differentiable transformation.

Inference concerning the OVL has been addressed by various authors. [2] considered hypothesis testing and interval estimation of the overlap of two normal distributions with equal variances. [6] estimated Matusita's measure of similarity between two multivariate normal distributions and they calculated asymptotic variance and bias of the estimator. [7] addressed the problem of making inferences about the overlap coefficients based on normal densities with equal means using jackknife, bootstrap, Taylor series approximation and transformation methods; the authors conclude that the bootstrap method is to be preferred.

In the present work, we shall explore the method based on the generalized pivotal quantity (GPQ) to address inference problems involving the OVL of two normal distributions with unequal means and unequal variances. It appears that for the parameter ρ defined in equation (1), there is no conventional pivotal quantity that will facilitate the computation of confidence limits. Thus we shall appeal to the concept of a *generalized pivotal quantity* (GPQ) due to [9, 10] and [11] for computing confidence limits. Numerous applications have shown that the GPQ idea can provide accurate confidence limits in situations where conventional methods appear to be non-existent. For example, the GPQ concept is used in [4] for computing accurate confidence limits for a log-normal mean, and in [3] for assessing occupational exposure using confidence limits for various parametric functions in the one way random effects model. [8] used the GPQ idea to construct exact confidence limits for the reliability function of a two parameter exponential distribution.

In the next section, we shall explain the construction of confidence intervals for ρ using the GPQ idea, and also mention the percentile bootstrap and bootstrap t methods for computing confidence limits. The solutions obtained by the three methods are compared in Section 3 using a simulation study to estimate the coverage probability and expected length of the confidence intervals. The simulation study indicates that overall, the GPQ method exhibits better performance compared to the bootstrap approaches. An illustrative example is given in Section 4.

2. CONFIDENCE INTERVALS FOR ρ

Matusita's measure of OVL of two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ simplifies to

$$\rho = \frac{\sqrt{2\sigma_1\sigma_2}}{(\sigma_1^2+\sigma_2^2)^{1/2}} exp\left\{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2)}\right\}. \tag{2}$$

If $\mu_1 = \mu_2$, we have

$$\rho = \frac{\sqrt{2\sigma_1\sigma_2}}{(\sigma_1^2+\sigma_2^2)^{1/2}}. \tag{3}$$

We shall now introduce the different confidence intervals for ρ, starting with the generalized confidence interval, i.e., the confidence interval constructed using the GPQ.

2.1. Generalized confidence interval

Let X_{ij}, $j = 1, \ldots, n_i$ be a random sample of size n_i from $N(\mu_i, \sigma_i^2), i = 1, 2$. Let

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \quad \text{and} \quad S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2, \ i = 1, 2.$$

Then

$$Z_i = \frac{(\overline{X}_i - \mu_i)\sqrt{n_i}}{\sigma_i} \sim N(0,1) \ \text{ and } U_i = \frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1},$$

where χ_r^2 denotes a chi-square distribution with r degrees of freedom. Let $\overline{x}_i$ and s_i^2 be the observed values of $\overline{X}_i$ and S_i^2, respectively. The GPQs of σ_i^2 and μ_i, say $T_{\sigma_i^2}$ and T_{μ_i}, respectively, are given by

$$T_{\sigma_i^2} = \frac{(n_i-1)s_i^2}{U_i}, \ \text{and} \ \ T_{\mu_i} = \overline{x}_i - Z_i\sqrt{\frac{T_{\sigma_i^2}}{n_i}} \tag{4}$$

for i=1,2; see [9]. In particular, we note that given the observed data, the probability distributions of $T_{\sigma_i^2}$ and T_{μ_i} are free of unknown parameters and that their observed values are σ_i^2 and μ_i, respectively, for $i = 1, 2$. A GPQ for ρ can be obtained by substituting $T_{\sigma_i^2}$ and T_{μ_i} in the place of σ_i^2 and μ_i, respectively. We shall denote the GPQ for ρ by T_ρ.

With T_ρ so defined, let $T_\rho(\alpha)$ denote the $100\alpha^{th}$ percentile of T_ρ. Then the $100(1-\alpha)\%$ generalized lower confidence limit for ρ is $T_\rho(\alpha)$. Also, a two-sided $100(1-\alpha)\%$ generalized confidence interval for ρ is $(T_\rho(\alpha/2), T_\rho(1-\alpha/2))$.

2.2. Bootstrap t confidence interval

Let $\hat{\rho}$ denote the estimate of ρ obtained by replacing μ_i with $\bar{X}_i$ and σ_i^2 with S_i^2 $(i = 1, 2)$ in the expression for ρ given in equation (2). The asymptotic variance of $\hat{\rho}$ is derived in [6], and is given by

$$V(\hat{\rho}) = \frac{\rho^2}{8} \sum_{i=1}^{2} \left\{ \frac{2\beta_i}{n_i} + \frac{1}{n_i - 1}(1 - 4\alpha_i + 4\alpha_i^2 + 2\beta_i + \beta_i^2 - 4\alpha_i\beta_i) \right\},$$

where

$$\alpha_i = \frac{\sigma_i^2}{\sigma_1^2 + \sigma_2^2} \quad \text{and} \quad \beta_i = \frac{(\mu_1 - \mu_2)^2 \sigma_i^2}{(\sigma_1^2 + \sigma_2^2)^2},$$

for $i = 1, 2$. In order to develop the bootstrap t confidence interval for ρ, let $(\bar{X}_{ib}^*, S_{ib}^{2*})$, $b = 1, 2, \ldots, B$, be a parametric bootstrap sample of size B generated according to

$$\bar{X}_{ib}^* \sim N\left(\bar{X}_i, \frac{S_i^2}{n_i}\right), \ \frac{(n_i - 1)S_{ib}^{2*}}{S_i^2} \sim \chi_{n_i - 1}^2,$$

$i = 1, 2$. Now let $\hat{\rho}^*(b)$ denote the estimate of ρ obtained from the bth parametric bootstrap sample, and let $se^*(b)$ denote its asymptotic standard error; the latter is obtained as the square root of the asymptotic variance given above, estimated using the bth parametric bootstrap sample. Now define the bootstrap t statistic

$$t^*(b) = \frac{\hat{\rho}^*(b) - \hat{\rho}}{se^*(b)},$$

$b = 1, 2, \ldots, B$, and let $\hat{t}_\gamma$ denote the upper γ percentile of the $t^*(b)-$values. The $100(1-\alpha)\%$ bootstrap t confidence interval for ρ is given by

$$\big(\hat{\rho} - \hat{t}_{\alpha/2} \times se(\hat{\rho}), \quad \hat{\rho} + \hat{t}_{\alpha/2} \times se(\hat{\rho})\big),$$

where $se(\hat{\rho})$ is the square root of the estimated asymptotic variance.

2.3. Percentile bootstrap

The percentile bootstrap consists of using the $100(\alpha/2)^{th}$ and $100\,(1-\alpha/2)^{th}$ percentiles of the $\hat{\rho}^*(b)$ values as the respective lower and upper limits of a $100(1-\alpha)\%$ confidence interval for ρ.

3. Simulation study

We shall now report estimated coverage probabilities and expected lengths of the 95% two sided confidence intervals for ρ, computed by the methods described in the previous paragraph. Two scenarios were considered for the simulation; we shall refer to them as simulation scenarios (I) and (II), given below.

Simulation scenario (I):
$$\mu_1 = 0,\ \mu_2 = 0,\ \sigma_1^2 = 1,\ \sigma_2^2 = 900,\ \rho = 0.2581$$
$$\mu_1 = 0,\ \mu_2 = 0,\ \sigma_1^2 = 9,\ \sigma_2^2 = 36,\ \rho = 0.8944$$
$$\mu_1 = 2,\ \mu_2 = 2,\ \sigma_1^2 = 1,\ \sigma_2^2 = 36,\ \rho = 0.5695$$

Simulation scenario (II):
$$\mu_1 = 0,\ \mu_2 = 1,\ \sigma_1^2 = 1,\ \sigma_2^2 = 2,\ \rho = 0.8933$$
$$\mu_1 = 0,\ \mu_2 = 3,\ \sigma_1^2 = 1,\ \sigma_2^2 = 2,\ \rho = 0.4587$$
$$\mu_1 = 2,\ \mu_2 = 3,\ \sigma_1^2 = 1,\ \sigma_2^2 = 2,\ \rho = 0.1209$$

Tables 1–2 give the results for a few choices of the sample size (n_1, n_2). The results are tabulated against the values of ρ only. The computations are done using `R`, based on 5,000 simulated samples. For implementing the bootstrap, 5000 parametric bootstrap samples were used. For computing the GPQ based confidence interval, 5,000 values of the GPQ were generated for each simulated sample. Tables 1–2 give the estimated coverage probability as well as the left tail and right tail coverages of the different confidence intervals. In the tables, 'Length' denotes the expected length of the confidence intervals, estimated by simulation.

Table 1. Coverage probabilities and expected lengths of 95% confidence intervals under simulation scenario (I)

		GPQ				Bootstrap t				Percentile bootstrap			
ρ	(n_1, n_2)	Coverage	Left tail	Right tail	Length	Coverage	Left tail	Right tail	Length	Coverage	Left tail	Right tail	Length
0.2581	(20,20)	0.9501	0.0127	0.0372	0.1193	0.9421	0.0520	0.0059	0.1195	0.9484	0.0122	0.0394	0.1189
	(20,30)	0.9493	0.0176	0.0331	0.1086	0.9372	0.0601	0.0027	0.1101	0.9375	0.0128	0.0497	0.1060
	(50,50)	0.9489	0.0185	0.0326	0.0729	0.9372	0.0512	0.0116	0.0735	0.9495	0.0175	0.0330	0.0722
	(50,100)	0.9492	0.0218	0.0290	0.0631	0.9549	0.0136	0.0315	0.0637	0.9431	0.0118	0.0451	0.0624
	(100,100)	0.9487	0.0204	0.0309	0.0509	0.9601	0.0281	0.0118	0.0514	0.9476	0.0195	0.0329	0.0507
0.5695	(20,20)	0.9510	0.0122	0.0368	0.2443	0.9561	0.0221	0.0218	0.2663	0.9475	0.0113	0.0412	0.2400
	(20,30)	0.9474	0.0164	0.0362	0.2225	0.9661	0.0162	0.0177	0.2433	0.9378	0.0073	0.0549	0.2195
	(50,50)	0.9494	0.0178	0.0328	0.1513	0.9459	0.0071	0.0470	0.1622	0.9524	0.0154	0.0322	0.1501
	(50,100)	0.9474	0.0212	0.0314	0.1309	0.9468	0.0024	0.0508	0.1408	0.9426	0.0105	0.0469	0.1319
	(100,100)	0.9512	0.0166	0.0322	0.1062	0.9568	0.0053	0.0379	0.1084	0.9476	0.0194	0.0330	0.1059
0.8944	(20,20)	0.9479	0.0042	0.0479	0.2344	0.9120	0.0863	0.0017	0.2574	0.9473	0.0020	0.0507	0.1955
	(20,30)	0.9532	0.0068	0.0400	0.2098	0.9274	0.0621	0.0105	0.2350	0.9361	0.0019	0.0620	0.1721
	(50,50)	0.9470	0.0115	0.0415	0.1481	0.9612	0.0283	0.0105	0.1616	0.9517	0.0102	0.0381	0.1470
	(50,100)	0.9480	0.0158	0.0362	0.1272	0.9587	0.0214	0.0199	0.1356	0.9425	0.0063	0.0512	0.1313
	(100,100)	0.9498	0.0168	0.0334	0.1049	0.9624	0.0262	0.0114	0.1103	0.9466	0.0157	0.0377	0.1045

Table 2. Coverage probabilities and expected lengths of 95% confidence intervals under simulation scenario (II)

ρ	(n_1, n_2)	GPQ				Bootstrap t				Percentile bootstrap			
		Coverage	Left tail	Right tail	Length	Coverage	Left tail	Right tail	Length	Coverage	Left tail	Right tail	Length
0.1209	(20,20)	0.9475	0.0317	0.0208	0.2915	0.9120	0.0880	0	0.3776	0.9344	0.0055	0.0601	0.2446
	(20,30)	0.9527	0.0299	0.0174	0.2572	0.9090	0.0910	0	0.3585	0.9449	0.0052	0.0499	0.2259
	(50,50)	0.9504	0.0293	0.0230	0.1792	0.9230	0.0770	0	0.2329	0.9457	0.0110	0.0433	0.1657
	(50,100)	0.9492	0.0291	0.0217	0.1435	0.9433	0.0567	0	0.1458	0.9342	0.0120	0.0538	0.1598
	(100,100)	0.9510	0.0271	0.0219	0.1246	0.9467	0.0533	0	0.1198	0.9419	0.0131	0.0450	0.1193
0.4587	(20,20)	0.9491	0.0256	0.0253	0.4242	0.9591	0.0390	0	0.6122	0.9321	0.0068	0.0611	0.4261
	(20,30)	0.9521	0.0261	0.0218	0.3868	0.9546	0.0454	0	0.5619	0.9486	0.0037	0.0477	0.4111
	(50,50)	0.9500	0.0263	0.0237	0.2821	0.9736	0.0264	0	0.3758	0.9438	0.0119	0.0443	0.2833
	(50,100)	0.9474	0.0269	0.0257	0.2365	0.9723	0.0277	0	0.2893	0.9671	0.0066	0.0263	0.2650
	(100,100)	0.9512	0.0275	0.0213	0.2029	0.9717	0.0283	0	0.2533	0.9492	0.0124	0.0384	0.2033
0.8933	(20,20)	0.9466	0.0039	0.0495	0.2571	0.9312	0.0688	0	0.3450	0.9322	0.0022	0.0656	0.2770
	(20,30)	0.9505	0.0066	0.0429	0.2301	0.9377	0.0623	0	0.2740	0.9492	0.0013	0.0495	0.2326
	(50,50)	0.9501	0.0123	0.0376	0.1625	0.9722	0.0228	0	0.2305	0.9410	0.0101	0.0489	0.1882
	(50,100)	0.9495	0.0162	0.0343	0.1394	0.9758	0.0242	0	0.1710	0.9674	0.0041	0.0285	0.1648
	(100,100)	0.9502	0.0183	0.0315	0.1151	0.9810	0.0190	0	0.1464	0.9482	0.0130	0.0388	0.1241

The numerical results in Table 1 and Table 2 show that in terms of coverage probability, the GPQ method is to be preferred; the resulting confidence interval has coverages very close to the nominal level of 0.95 in all the cases considered for the simulation. Furthermore, only the GPQ based confidence interval gives left and right tail coverages close to 0.025 (in most cases). The percentile bootstrap does provide reasonable coverages in most cases. However, as can be seen from Table 2, the confidence interval based on the bootstrap t has poor coverages in several cases. In terms of expected length, the comparison between the GPQ method and the percentile bootstrap is mixed; but they are mostly comparable. Our overall recommendation is to use the confidence interval based on the GPQ approach.

4. An Example

We shall now illustrate our methodology using an example taken from [1] on the nitrogen content of Iowa soils with and without the bacterium *Azetobactor*. The data consists of 13 soil samples with the bacterium, and 10 samples without the bacterium. It can be verified that both sets of observations follow normal distributions. The maximum likelihood estimates of the parameters are $\hat{\mu}_1$=44.92, $\hat{\sigma}_1^2$=147.24, $\hat{\mu}_2$=20.8 and $\hat{\sigma}_2^2$=24.18. The estimate of Matusita's measure of similarity is $\hat{\rho} = 0.3571$. We computed 95% confidence intervals for ρ using (i) the GPQ method (using 10,000 GPQ values), (ii) bootstrap t and (iii) the percentile bootstrap based on 10,000 parametric bootstrap samples. The confidence intervals constructed using these methods are (0.1579, 0.6511), (0.1909, 0.7959) and (0.1100, 0.5265), respectively. We note that the interval based on the percentile bootstrap is shifted to the left of the interval based on the GPQ method, and the latter is shifted to the left of the bootstrap t interval. This behavior among the three intervals reflects the differences noticed among the coverage probabilities reported in Tables 1-2, especially the left and right tail coverages. Given the significant differences among the intervals, it is important to use an interval that performs satisfactorily in terms of the coverage probabilities and expected lengths. Our recommendation is to use the interval based on the GPQ methodology.

5. Discussion

The OVL coefficient is a widely used measure to assess the similarity of two distributions. For two normal populations with unequal means and unequal variances, we have considered the interval estimation of the OVL coefficient, and have assessed the performance of different confidence intervals using simulations.

The simulation results and an illustrative example have brought out the differences among the confidence intervals based on the GPQ method, bootstrap t and percentile bootstrap. Even though the bootstrap is a well established methodology, the GPQ based confidence interval appears to be the one with the most satisfactory performance in terms of coverage probability.

REFERENCES

[1] G.M. Cox and W.P. Martin, *Use of discriminant function for differentiating soils withdifferent Azetobactor populations*, Journal paper No. *J 451* of the Iowa aricultural experiment station, Ames USA (1937).

[2] H.F. Inman and E.L.Bradley, *Hypothesis tests and confidence interval estimates for the overlap of two normal distribution with equal variances*, Environmetrics **5** (1994) 167–189.

[3] K. Krishnamoorthy and T. Mathew, *Assessing Occupational exposure via the one way random effects model with balanced data*, Journal of Agricultural Biological and Environmental Statistics **4** (2002) 440–457.

[4] K. Krishnamoorthy and T. Mathew, *Inferences on the means of log normal distributions using generalized p-values and generalized confidence intervals*, Journal of Statistical Planning and Inference **115** (2003) 103–121.

[5] K. Matusita, *Decision rules based on the distance for problem of fit, two samples, and estimation*, The Annals of Mathematical Statistics **26** (1955) 631–640.

[6] M. Minami and K. Shimizu, *Estimation of similarity measure for multivariate normal distributions*, Environmetal and Ecological Statistics **6** (1999) 229–248.

[7] M.S. Mulekar and S.N. Mishra, *Confidence interval estimation of the overlap: equal means case*, Computational Statistics and Data Analysis **34** (2000) 121–137.

[8] A. Roy and T. Mathew, *A generalized confidence limit for the reliability function of a two parameter exponential distribution*, Jornal of Statistical Planning and Inference **128** (2) (2005) 509–17.

[9] S. Weerahandi, *Generalized Confidence Intervals*, Journal of the American Statistical Association **88** (1993) 899–905.

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

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

Received
Revised
Accepted