Comparing Mean Vectors Via Generalized Inference in Multivariate Log-Normal Distributions

Kamel Abdollahnezhad*[1], Ali Akbar Jafari[2], Nasrin Tatari[1]
1. Golestan Universiy
2. Yazd University

Received: 26 Nov. 2017 Accepted: 28 May 2018

Extended Abstract

Paper pages (187-196)

Introduction

The skewed distributions are particularly common when mean values are small, variances are large and values cannot be negative (for example lengths of latent periods of infectious diseases), and often closely fit the log-normal distribution. The log-normal distribution has been widely used in medical, biological and economic studies, where data are positive and have a right-skewed distribution. In this paper, we develop a procedure that is readily applicable for both hypothesis testing and confidence region construction for comparing the mean vectors of several multivariate log-normal distributions. The classical theory of statistical analysis is based on the assumption that the data are normally distributed. However, when the data are positive and skewed, the techniques of normality are inadequate to support inferential tests. For these data, the log-normal distribution is potentially a more suitable candidate to describe the positive and skewed random variables. Furthermore, once the data contains multiple and dependent variables, a multivariate approach instead of a univariate approach is practically necessary.

The mean of a log-normal random variable involves a linear combination of the mean and variance of the normal distribution of the log-data, and thus the inference procedures are more complicated concerning the log-normal mean. There are many applications for confidence intervals and tests regarding the log-normal means. Therefore, inference procedures concerning a single log-normal mean, or that of comparing the means of two or more independent log-normal distributions, attract more attention in the literature; see Zhou and Gao (1997), Taylor et al. (2002), Wu et al. (2002, 2003), Zhou et al. (1997), Chen and Zhou (2006), Gupta and Li (2006), Krishnamoorthy and Mathew (2003), Li (2009), Lin and Wang (2013) and Jafari and Abdollahnezhad (2015, 2017).

The applications of the multivariate log-normal distribution are similar to those of the univariate and bivariate log-normal distributions including, for example, the study of the size distribution of aerosol particles, airborne fibers, biomedical applications, etc. So far, there are no existing methods available in the literature for comparing mean vectors of multivariate log-normal distributions when several multivariate log-normal populations are not homogeneous. The multivariate analysis of variance (MANOVA) can only be applied when several log-data sets are homogeneous. Thus, the purpose of this study is to fill the gap by developing a procedure for comparing the mean vectors of several multivariate log-normal distributions based on the generalized variable approach (GVA).

Material and methods

The GVA, including the generalized p-values and generalized confidence interval, was introduced by Tsui and Weerahandi (1989). The method has turned out to be extremely fruitful for obtaining tests and confidence intervals involving non-standard parameters. Numerous articles have revealed that this approach can satisfactorily deal with the problems which are heteroscedasticity, high dimensions or small sample sizes; see the books by Weerahandi for a detailed discussion along with numerous examples.

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be a random sample from a p-variate normal population with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ and covariance matrix $\Sigma = (\sigma_{ij})$, $i = 1, \dots, p$, $j = 1, \dots, p$. Let $\overline{\boldsymbol{X}}$ and S denote the sample mean and sample covariance matrix, respectively. That is,

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{h=1}^{n} \boldsymbol{X}_h, \qquad S = (S_{lj}) = \sum_{h=1}^{n} (\boldsymbol{X}_h - \overline{\boldsymbol{X}})(\boldsymbol{X}_h - \overline{\boldsymbol{X}})'.$$

It is well-known that $\overline{\boldsymbol{X}}$ and S are independent with

$$\overline{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right), \qquad S = W_p(n-1, \frac{1}{n-1}\Sigma).$$

where $W_p(r, \Sigma)$ denotes the p-dimentsional Wishart distribution with r degrees of freedom and scale parameter matrix Σ**.**

We give a generalized pivotal variable for Σ. By Cholesky decomposition we can find a lower triangular matrix Λ such that $\Sigma = \Lambda\Lambda'$. Let A be Cholesky decomposition of S such that $S = AA'$. Also, consider $B = (B_{lj}) = \Lambda^{-1}A$, then

$$BB' = \Lambda^{-1}A(\Lambda^{-1}A)' = \Lambda^{-1}S\Lambda'^{-1} \sim W_p(n-1, I_p).$$

Therefore, B is a lower triangular matrix and B_{ij}'s are independent (Muirhead, 1982) and

$$B_{ll}^2 \sim \chi^2_{(n-l)}, \qquad B_{lj} \sim N(0,1), \quad l > j.$$

Let s and a be the observed matrices of S and A, respectively. Then

$$V = (V_{lj}) = aB^{-1}B'^{-1}a' = aA^{-1}\Sigma A'^{-1}a',$$

is a generalized pivotal variable for Σ,

Now, we present a generalized pivotal variable for $\boldsymbol{\mu}$. Consider $\overline{\boldsymbol{x}}$ is observed value for $\overline{\boldsymbol{X}}$. Since $\boldsymbol{Z} = \sqrt{n}\Lambda^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, I_p)$, a generalized pivotal variable for $\boldsymbol{\mu}$ **is**

$$\boldsymbol{T}_{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} - \frac{aB^{-1}}{\sqrt{n}}\boldsymbol{Z} = \overline{\boldsymbol{x}} - aA^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}).$$

Consider $\boldsymbol{T}_{\boldsymbol{\sigma}} = (\sqrt{V_{11}}, \dots, \sqrt{V_{PP}})'$. Therefore, $\boldsymbol{T}_{\boldsymbol{\theta}} = \boldsymbol{T}_{\boldsymbol{\mu}} + \frac{1}{2}\boldsymbol{T}_{\sigma}$ is a generalized pivotal variable for $\left(\mu_1 + \frac{1}{2}\sigma_1, \dots, \mu_p + \frac{1}{2}\sigma_p\right)$.

Results and discussion

Let $\boldsymbol{Y}_{i1}$‘ … ‘$\boldsymbol{Y}_{in_i}$ be a random sample from multivariate log-normal distribution with parameters $\boldsymbol{\mu}_i$ and Σ_i $(i = 1, \dots, k)$ where

$$\boldsymbol{\mu_i} = (\mu_{i1}, \dots, \mu_{ip})', \qquad \Sigma_i = [\sigma_{i,st}], \quad i = 1, \dots, k.$$

Our problem is testing equality of means of this k populations which equivalent to $H_0 : H\boldsymbol{\theta} = \boldsymbol{0}$, vs. $H_1 : H\boldsymbol{\theta} \neq \boldsymbol{0}$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_{\boldsymbol{k}})'$, $\boldsymbol{\theta_i} = \boldsymbol{\mu_i} + \frac{1}{2}\boldsymbol{\sigma_i}$, $\boldsymbol{\sigma}_i = (\sigma_{i,11}, \dots, \sigma_{i,pp})'$, $H = Q \otimes I_p$, $Q = [-\boldsymbol{1}_{k-1} : I_{k-1}]$, $\boldsymbol{1}_{k-1}$ is a column vector of 1's

A generalized pivotal variable for the mean of ith population is $\boldsymbol{T}_i = \boldsymbol{T}_{\boldsymbol{\mu}_i} + \frac{1}{2}\boldsymbol{T}_{\boldsymbol{\sigma_i}}$**.** Therefore, $\boldsymbol{T}^* = H(\boldsymbol{T}'_1, \dots, \boldsymbol{T}'_k)'$, is a generalized pivotal variable for $H\boldsymbol{\theta}$**.**

Let $\widetilde{\boldsymbol{T}}^*$ denote the standard expression of $\boldsymbol{T}^*$ with $\widetilde{\boldsymbol{T}}^* = \Sigma_{\boldsymbol{T}^*}^{-\frac{1}{2}}(\boldsymbol{T}^* - \boldsymbol{\mu}_{\boldsymbol{T}^*})$, where $\boldsymbol{\mu}_{\boldsymbol{T}^*}$ and $\Sigma_{\boldsymbol{T}^*}$ are the conditional expectation and conditional variance-covariance matrix of $\boldsymbol{T}^*$. The observed value of $\boldsymbol{T}^*$ is $\tilde{\boldsymbol{t}}^* = \Sigma_{\boldsymbol{T}^*}^{-\frac{1}{2}}(H\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{T}^*})$. Therefore, the generalized p-value for the test can be given by

$$p = P\big((\boldsymbol{T}^* - \boldsymbol{\mu}_{\boldsymbol{T}^*})'\Sigma_{\boldsymbol{T}^*}^{-1}(\boldsymbol{T}^* - \boldsymbol{\mu}_{\boldsymbol{T}^*}) > \boldsymbol{\mu}'_{\boldsymbol{T}^*}\Sigma_{\boldsymbol{T}^*}^{-1}\boldsymbol{\mu}_{T^*}\big).$$

and H_0 will be rejected when p is less than the level α.

We performed a simulation study to compare the proposed generalized p-value and multivariate analysis of variance (MANOVA) to test the equality of means of several multivariate log-normal populations. The simulations show that the actual sizes of the both approaches are very smaller than the nominal level and there is not different between them. But the power of the proposed approach is larger than the power of MANOVA in small sample sizes.

Conclusion

We consider the problem of means in several multivariate log-normal distributions and propose a useful method called as generalized variable method. Simulation studies show that suggested method has an appropriate size and power regardless sample size. To evaluation this method, we compare this method with traditional MANOVA such that the actual sizes of the two methods are close but the power of test and coverage probability of proposed methods are better than MANOVA in most cases specially when the sample sizes are small. Therefore, we can use this method when the variance-covariance matrices are not equal and there is not a suitable method.

Keywords: Generalized Variable, Multivariate log-normal distribution, Size and power of test, Coverage probability, Cholesky decomposition.

[*] Corresponding Author: kamel_abdollahnezhad@yahoo.com