Comparing the equality of K mean vectors on several multivariate log-normal distributions

Shu-Hui Lin*

National Taichung University of Science and Technology

129 Sanmin Road Sec. 3, Taichung 404, Taiwan.

Abstract

In this study, we extend our research experience on studying mean vector of one and two multivariate log-normal populations to further consider the mean vectors of several independent multivariate log-normal populations. The log-normal distribution is one of good candidates to describe positive and skewed data. If the data contain many characteristic values, the multivariate log-normal distribution is a good choice to fit such data. In this stdudy, we will derive the testing procedure to test the equality of K-mean vectors based on the generalized variable method (GVM). The proposed method will be compared with the classical F-test and the classical χ^2 - test which are available in the literature under under different groups, dimensions and parameters configuration.

Keywords: generalized *p*-value; generalized variable method; K mean vectors; multivariate log-normal distribution; pairwise mean vector difference

1. Introduction

Traditionally, the statistical analysis of the classical theory is mostly based on the assumption that the data are normally distributed. However, if the data are departed from bell shape and symmetric, then the techniques of normality are inadequate to support inferential tests. In the real world, the data is usually positive and skew; the log-normal distribution is one of the potential models to describe such data. For the log-transformed data, the mean of a log-normal distribution involves a linear combination of the mean and variance of the normal distribution, and thus the inference procedures for the log-normal mean are more complicated than mean of the normal random variable. Nevertheless, inference procedures concerning the log-normal mean are in great demand, thus in the literature, inferences on a single log-normal mean, two or several independent log-normal means attract much attention. For example, for one population: Zhou and Gao (1997) constructed a confidence interval for the log-normal mean; Taylor, Kupper and Muller (2002) provided improved approximate confidence interval for the mean of a log-normal random variable; Wu, Wong and Jiang (2003) applied likelihood-based method to construct confidence interval for a log-normal mean. For two log-normal populations: Zhou, Gao and Hui (1997) and Krishnamoorthy and Mathew (2003) provided methods to compare two independent log-normal means; Wu et al. (2002) used likelihood analysis on the ratio of two independent log-normal means; Chen and Zhou (2006) constructed interval estimation for the ratio and difference of two log-normal means; Gupta and Li (2006) provided inference on the common mean of two log-normal distributions. For serval populations: Lin (2013) applied higher order likelihood method for making inference on the common mean of several log-normal distributions; Lin and Wang (2013) used modified method on comparing several log-normal means.

If the data contain more than one characteristic, then single variate inference is inadequate to fit the data. Several recent researches focused on comparing the means for a bivariate log-normal distribution. For example, Bebu and Mathew (2008) compared the means and variances of a bivariate log-normal distribution; Zhou, Gao and Tierney (2001) provided inference on testing the equality of means of a bivariate log-normal distribution; Hawkins (2002) diagnosed for conformity of paired quantitative measurements based on bivaraite log-normal distribution, etc. However, if the data contain many characteristic values or the interest is not merely to compare these two elements of bivariate log-normal distributions, the properties of multivariate log-normal distributions deserve further research. Lin (2014) applied generalized variable method to compare the mean vectors of two independent multivariate log-normal distributions. Hence, it is of practical and theoretical importance to extend the achievement to further develop a procedure for comparing the mean vectors of several independent multivariate log-normal distributions. The applications of the multivariate log-normal distribution are similar to those of the univariate and bivaraite log-normal distributions, and can be applicable to the exploration of the size distribution of aerosol particles, airborne fibers, biomedical applications, etc.

Most importantly, the purpose of this study is to develop a procedure for comparing the mean vectors of several multivariate log-normal distributions based on the concepts of the generalized variable method (GVM) which were derived by Tsui and Weerahandi (1989) and Weerahandi (1993), respectively. GVM method has been applied to solve many statistical problems involving nuisance parameters and many of the results are satisfactory. The reader is referred to the books by Weerahandi (1995, 2004) for a detailed discussion along with numerous examples.

In this paper, we develop procedures that are readily applicable for testing the equality of several independent multivariate log-normal mean vectors. We will first derive the generalized test variable (GTV) to test the equality of K mean vectors, and then the proposed method will be compared under different groups, dimensions and parameters configuration with the other methods available in the literature.

The rest of the article is organized as follows. The theory of generalized *p*-values and generalized confidence interval are briefly introduced in Section 2. The property of the multivariate log-normal distribution will also briefly reviewed in Section 2. The proposed procedure will be presented in Section 3. The classical *F*-test and the classical χ^2-test are also briefly introduced in Section 3. The numerical study will be presented in Section 4 to compare our proposed method with the other two methods in different combinations of sample sizes and parameter configurations. Finally, some conclusions are presented in Section 5.

2. Preliminary

2.1.1 Generalized *p*-values

The setup of the generalized *p*-value and the generalized confidence interval will be briefly introduced as follows. Let $\mathbf{X}$ be a random quantity having a density function $f(\mathbf{X}|\zeta)$, where $\zeta=(\theta,\boldsymbol{\pi})$ is a vector of unknown parameters and θ is the parameter of interest, and $\boldsymbol{\pi}$ is a vector of nuisance parameters. Suppose we are interested in testing

$$H_0:\theta\le\theta_0 \text{ vs. } H_1:\theta>\theta_0, \qquad (2.1)$$

where θ_0 is a pre-specified value.

Let $\mathbf{x}$ denote the observed value of $\mathbf{X}$ and the ***generalized test variable*** **(GTV)**, $T(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})$, which depends on the observed value $\mathbf{x}$ and the parameters ζ, and satisfies the following requirements:

(A)
- (i) For fixed $\mathbf{x}$ and $\boldsymbol{\zeta}=(\theta_0,\boldsymbol{\pi})$, the distribution of $\mathrm{T}(\mathbf{X};\mathbf{x},\theta_0,\boldsymbol{\pi})$ is free of the nuisance parameters $\boldsymbol{\pi}$.
- (ii) The observed value $\mathrm{T}(\mathbf{x};\mathbf{x},\theta_0,\boldsymbol{\pi})$ of $\mathrm{T}(\mathbf{X};\mathbf{x},\theta_0,\boldsymbol{\pi})$ does not depend on unknown parameters $\boldsymbol{\pi}$.
- (iii) For fixed $\mathbf{x}$ and $\boldsymbol{\pi}$, $\Pr[\mathrm{T}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})\ge \mathrm{t}]$ is either increasing or decreasing in θ for any given t.

Under the above conditions, if $T(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})$ is stochastically increasing in θ, then the generalized *p*-value for testing the hypothesis in (2.1) can be defined as

$$p=\sup_{\theta\le\theta_0}\Pr[T(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})\ge t]=\Pr[T(\mathbf{X};\mathbf{x},\theta_0,\boldsymbol{\pi})\ge t], \qquad (2.2)$$

where $t=T(\mathbf{x};\mathbf{x},\theta_0,\boldsymbol{\pi})$.

2.1.2 Generalized confidence intervals

Under the same set up, suppose $\mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})$ satisfies the following conditions:

$$\textbf{(B)}\ \begin{cases}\text{(i) The distribution of } \mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi}) \text{ does not depend on any unknown parameters.}\\ \text{(ii) The observed value } \mathrm{Q}(\mathbf{x};\mathbf{x},\theta,\boldsymbol{\pi}) \text{of } \mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi}) \text{ is free of nuisance parameters } \boldsymbol{\pi}.\end{cases}$$

Then we say $\mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})$ is a ***generalized pivotal quantity*** **(GPQ)**. Furthermore, if c_1 and c_2 are such that

$$\Pr[c_1 \le \mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi}) \le c_2] = 1-\alpha, \tag{2.3}$$

then $\{\theta : c_1 \le \mathrm{Q}(\mathbf{x};\mathbf{x},\theta,\boldsymbol{\pi}) \le c_2\}$ is a $100(1-\alpha)\%$ generalized confidence interval for θ. Specially, if the value of $\mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})$ at $\mathbf{X}=\mathbf{x}$ is θ, the parameter of interest, and $h_{\{Q(\mathbf{X});\ 1\text{-}\alpha\}}$ represents the $100(1-\alpha)^{\text{th}}$ percentile of $\mathrm{Q}(\mathbf{X};\mathbf{x},\theta,\boldsymbol{\pi})$, then $\left\{h_{\{Q(\mathbf{x});\alpha/2\}},\ h_{\{Q(\mathbf{x});1-\alpha/2\}}\right\}$ is a $100(1-\alpha)^{\text{th}}$ confidence interval of θ.

2.2 The property of the multivariate log-normal distribution

Suppose $(\underset{\sim}{\mathbf{Y}}_1,\ldots,\underset{\sim}{\mathbf{Y}}_n)$ is d-variate multivariate log-normal population. Let $\mathbf{X}=\ln\mathbf{Y}$ and so that $(\underset{\sim}{\mathbf{X}}_1,\ldots,\underset{\sim}{\mathbf{X}}_n)$ follows d-variate multivariate normal distributions with mean vector $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix $\boldsymbol{\Sigma}$, where

$$\underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix}\mu_1\\ \vdots\\ \mu_d\end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_{11} & \cdots & \sigma_{1d}\\ \vdots & \ddots & \vdots\\ \sigma_{d1} & \cdots & \sigma_{dd}\end{pmatrix}. \tag{2.4}$$

The mean vector and the covariance matrix of $\mathbf{Y}_j$ are

$$E(\underset{\sim}{\mathbf{Y}}_j) = \underset{\sim}{\boldsymbol{\upsilon}} = \begin{pmatrix}\exp(\mu_1+\sigma_{11}/2)\\ \vdots\\ \exp(\mu_d+\sigma_{dd}/2)\end{pmatrix} \text{ and } Var(\underset{\sim}{\mathbf{Y}}_j) = \boldsymbol{\Psi} = [e_{st}], \tag{2.5}$$

where $e_{st} = \exp[\mu_s+\mu_t+(\sigma_{ss}+\sigma_{tt})/2][\exp(\sigma_{st})-1]$, $s,t=1,\ldots,d.$ For simplicity, if $d=2$, then the mean vector and the covariance matrix are $\underset{\sim}{\boldsymbol{\upsilon}} = \begin{pmatrix}\exp(\mu_1+\sigma_{11}/2)\\ \exp(\mu_2+\sigma_{22}/2)\end{pmatrix}$ and

$$\boldsymbol{\Psi} = \begin{bmatrix}\exp[2\mu_{11}+\sigma_{11}]\cdot[\exp(\sigma_{11})-1] & \exp[\mu_2+\mu_1+(\sigma_{22}+\sigma_{11})/2]\cdot[\exp(\sigma_{21})-1]\\ \exp[\mu_1+\mu_2+(\sigma_{11}+\sigma_{22})/2]\cdot[\exp(\sigma_{12})-1] & \exp[2\mu_2+\sigma_{22}]\cdot[\exp(\sigma_{22})-1]\end{bmatrix}.$$

Since we are interested in making inference for the mean vector $\underset{\sim}{\boldsymbol{\upsilon}}$, which can be obtained equivalently through the parameter $\underset{\sim}{\boldsymbol{\eta}}$, where

$$\underset{\sim}{\boldsymbol{\eta}} = \begin{pmatrix} \mu_1 + \sigma_{11}/2 \\ \vdots \\ \mu_d + \sigma_{dd}/2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \sigma_{11} \\ \vdots \\ \sigma_{dd} \end{pmatrix}. \tag{2.6}$$

Based on the log-transformed data $X_j = \ln Y_j$, two sufficient statistics, the sample mean vector and sample covariance matrix, are denoted by $\underset{\sim}{\bar{\mathbf{X}}}$ and $\mathbf{S}$, where

$$\underset{\sim}{\bar{\mathbf{X}}} = \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{\mathbf{X}}_j \quad \text{and} \quad \mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\underset{\sim}{\mathbf{X}}_j - \underset{\sim}{\bar{\mathbf{X}}})(\underset{\sim}{\mathbf{X}}_j - \underset{\sim}{\bar{\mathbf{X}}})'. \tag{2.7}$$

It is easily to verify that

$$\underset{\sim}{\bar{\mathbf{X}}} \sim N_d(\underset{\sim}{\boldsymbol{\mu}},\ \boldsymbol{\Sigma}/n) \quad \text{and} \quad \mathbf{A} \equiv (n-1)\mathbf{S} \sim W_d(n-1, \boldsymbol{\Sigma}), \tag{2.8}$$

where $N_d(\underset{\sim}{\boldsymbol{\mu}},\ \boldsymbol{\Sigma}/n)$ is the *d*-variate multivariate normal distribution with mean vector $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix $\boldsymbol{\Sigma}/n$ and $W_d(n-1, \boldsymbol{\Sigma})$ denotes the *d*-dimensional Wishart distribution with degrees of freedom $n-1$ and scale matrix $\boldsymbol{\Sigma}$. Besides, since $\underset{\sim}{\bar{\mathbf{X}}}$ and $\mathbf{S}$ are two sufficient statistics and both are affine invariant, and thus, we will can make inference of $\underset{\sim}{\boldsymbol{\eta}}$ based on those two statistics.

Furthermore, if the covariance matrix $\boldsymbol{\Sigma}$ is known, then from (2.8) we can derive

$$\underset{\sim}{\mathbf{Z}} = \left(\boldsymbol{\Sigma}/n\right)^{-1/2}\left(\underset{\sim}{\bar{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}\right) \sim N_d(\underset{\sim}{\mathbf{0}},\ \mathbf{I}_d) \tag{2.9}$$

and

$$\mathbf{B} = \left(\mathbf{a}^{-1/2}\boldsymbol{\Sigma}\mathbf{a}^{-1/2}\right)^{-1/2}\left(\mathbf{a}^{-1/2}\mathbf{A}\mathbf{a}^{-1/2}\right)\left(\mathbf{a}^{-1/2}\boldsymbol{\Sigma}\mathbf{a}^{-1/2}\right)^{-1/2} \sim W_d(n-1,\ \mathbf{I}_d), \tag{2.10}$$

where $\mathbf{a}$ are the observed value of $\mathbf{A}$. Let

$$\boldsymbol{\Omega} = \mathbf{a}^{1/2}\mathbf{B}^{-1}\mathbf{a}^{1/2}, \tag{2.11}$$

then it can be easily seen that the observed value of $\boldsymbol{\Omega}$ is $\boldsymbol{\Sigma}$ and the distribution of $\boldsymbol{\Omega}$ is free of any unknown parameter, thus we can used it to draw the information about the nuisance parameter $\boldsymbol{\Sigma}$. Next, define

$$\begin{aligned} \underset{\sim}{\mathbf{T}} &= \left[\underset{\sim}{\bar{\mathbf{x}}} - \left(\boldsymbol{\Omega}/n\right)^{1/2}\ \left(\boldsymbol{\Sigma}/n\right)^{-1/2}\left(\underset{\sim}{\bar{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}\right)\right] + \frac{1}{2}\left(diag\left(\boldsymbol{\Omega}\right)\right)' \\ &= \underset{\sim}{\bar{\mathbf{x}}} - \left(\boldsymbol{\Omega}/n\right)^{1/2}\cdot\underset{\sim}{\mathbf{Z}} + \frac{1}{2}\left(diag\left(\boldsymbol{\Omega}\right)\right)' \end{aligned} \tag{2.12}$$

where $diag\left(\boldsymbol{\Omega}\right)$ means the diagonal element of $\boldsymbol{\Omega}$, $\underset{\sim}{\bar{\mathbf{x}}}$ represents the observed value of $\bar{\underset{\sim}{\mathbf{X}}}$, and $\underset{\sim}{\mathbf{Z}}$ and $\mathbf{\Omega}$ are independent distributed.

It can be seen that $\underset{\sim}{\mathbf{T}}$ is a function of the independent random variables $\underset{\sim}{\mathbf{Z}}$, $\mathbf{\Omega}$, and the observed quantities $\bar{\underset{\sim}{\mathbf{x}}}$ and $\mathbf{a}$. The observed value of $\underset{\sim}{\mathbf{T}}$ is just the parameter of interest $\underset{\sim}{\boldsymbol{\eta}}$ and its distribution is free of unknown parameters. The property of $\underset{\sim}{\mathbf{T}}$ is conformed with the requirements of GTV and GPQ, therefore we can use it used to perform the hypothesis testing and construct a confidence region for $\underset{\sim}{\boldsymbol{\eta}}$.

3. Comparing mean vectors of several multivariate log-normal populations

3.1 The proposed method

Suppose $(\underset{\sim}{\mathbf{Y}}_{11},...,\underset{\sim}{\mathbf{Y}}_{1n_1}),\ldots,(\underset{\sim}{\mathbf{Y}}_{K1},...,\underset{\sim}{\mathbf{Y}}_{Kn_K})$ are K independent *d*-variate log-normal populations. Let $\underset{\sim}{\mathbf{X}}_{ij}=\ln\underset{\sim}{\mathbf{Y}}_{ij}$, so that $(\underset{\sim}{\mathbf{X}}_{11},...,\underset{\sim}{\mathbf{X}}_{1n_1})$,…, $(\underset{\sim}{\mathbf{X}}_{K1},...,\underset{\sim}{\mathbf{X}}_{Kn_K})$ follow *d*-variate normal distributions with mean vector $\underset{\sim}{\boldsymbol{\mu}}_i$ and covariance matrices $\mathbf{\Sigma}_i$,

where

$$\underset{\sim}{\boldsymbol{\mu}}_i=\begin{pmatrix}\mu_{i1}\\ \vdots\\ \mu_{id}\end{pmatrix}\text{ and }\mathbf{\Sigma}_i=\begin{pmatrix}\sigma_{i,11} & \cdots & \sigma_{i,1d}\\ \vdots & \ddots & \vdots\\ \sigma_{i,d1} & \cdots & \sigma_{i,dd}\end{pmatrix},\ i=1,...,K. \tag{3.1}$$

The mean vector and the covariance matrix of Y_{ij} are

$$E(\underset{\sim}{\mathbf{Y}}_{ij})=\underset{\sim}{\boldsymbol{\upsilon}}_i=\begin{pmatrix}\exp(\mu_{i1}+\sigma_{i,11}/2)\\ \vdots\\ \exp(\mu_{id}+\sigma_{i,dd}/2)\end{pmatrix}\text{ and }Var(Y_{ij})=\mathbf{\Psi}_i=\left[e_{i,st}\right], \tag{3.2}$$

where $e_{i,st}=\exp\left[\mu_{is}+\mu_{it}+0.5(\sigma_{i,ss}+\sigma_{i,tt})\right]\left[\exp(\sigma_{i,st})-1\right], i=1,...,K; s,\ t=1,...,d.$

In order to compare K means, $\underset{\sim}{\boldsymbol{\upsilon}}_1,...,\underset{\sim}{\boldsymbol{\upsilon}}_K$, it can be obtained equivalently through

comparing $\underset{\sim}{\boldsymbol{\eta}}_1,...,\underset{\sim}{\boldsymbol{\eta}}_K$, where $$\underset{\sim}{\boldsymbol{\eta}}_i=\begin{pmatrix}\mu_{i1}+0.5\sigma_{i,11}\\ \vdots\\ \mu_{id}+0.5\sigma_{i,dd}\end{pmatrix},\ i=1,...,K. \tag{3.3}$$

Since it is known that $\underset{\sim}{\bar{\mathbf{X}}}_i$ and $\mathbf{S}_i$ are mutually independent with

$$\underset{\sim}{\bar{\mathbf{X}}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\underset{\sim}{\mathbf{X}}_{ij} \sim N_d(\underset{\sim}{\boldsymbol{\mu}}_i,\, \boldsymbol{\Sigma}_i / n_i) \tag{3.4}$$

and

$$\mathbf{A_i} \equiv (n_i - 1)\mathbf{S}_i = \sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{X}}_{ij} - \underset{\sim}{\bar{\mathbf{X}}}_i)(\underset{\sim}{\mathbf{X}}_{ij} - \underset{\sim}{\bar{\mathbf{X}}}_i)' \sim W_d(n_i - 1, \boldsymbol{\Sigma}_i),\ i = 1,...,K. \tag{3.5}$$

3.2 The generalized *p*-value

Comparing the equality of mean vectors of K multivariate independent log-normal distributions

$$H_0 : \underset{\sim}{\boldsymbol{\eta}}_1 = ... = \underset{\sim}{\boldsymbol{\eta}}_K \text{ vs. } H_1 : \underset{\sim}{\boldsymbol{\eta}}_i\text{'s not all the same} \tag{3.6}$$

can be equivalently set as testing the hypothesis

$$H_0 : \mathbf{G}\boldsymbol{\eta}_{(C)} = \underset{\sim}{\mathbf{0}} \text{ vs. } H_1 : \mathbf{G}\boldsymbol{\eta}_{(C)} \neq \underset{\sim}{\mathbf{0}}, \tag{3.7}$$

where $\mathbf{G} = \begin{pmatrix} \mathbf{I_d} & -\mathbf{I_d} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{I_d} & \mathbf{0} & -\mathbf{I_d} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & & & & \\ \mathbf{I_d} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & -\mathbf{I_d} \end{pmatrix}$ and $\underset{\sim}{\boldsymbol{\eta}}'_{(C)} = (\underset{\sim}{\boldsymbol{\eta}}_1{}', ..., \underset{\sim}{\boldsymbol{\eta}}'_K).$ (3.8)

It is noted that $\mathbf{G}$ is $(k-1)d \times kd$ matrix, $\mathbf{I_d}$ stand for the $d \times d$ identity matrix, $\boldsymbol{\eta}_{(C)}$ means the "combined" block of $\underset{\sim}{\boldsymbol{\eta}}_1, ..., \underset{\sim}{\boldsymbol{\eta}}_K$ and then

$$\mathbf{G}\underset{\sim}{\boldsymbol{\eta}}_{(c)} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\eta}}_1 - \underset{\sim}{\boldsymbol{\eta}}_2 \\ \underset{\sim}{\boldsymbol{\eta}}_1 - \underset{\sim}{\boldsymbol{\eta}}_3 \\ \vdots \\ \underset{\sim}{\boldsymbol{\eta}}_1 - \underset{\sim}{\boldsymbol{\eta}}_K \end{pmatrix}. \tag{3.9}$$

According to (2.12) that the observed value of $\underset{\sim}{\mathbf{T}}$ is the parameter of interest $\underset{\sim}{\boldsymbol{\eta}}$ and its distribution is free of unknown parameters. Thus the generalized test variable can be defined (GTV) as

$$\begin{aligned} \underset{\sim}{\mathbf{T}}^* &\equiv \mathbf{G}\underset{\sim}{\mathbf{T}}_{(c)} = \mathbf{G}(\underset{\sim}{\mathbf{T}}_1{}', ..., \underset{\sim}{\mathbf{T}}'_K)' \\ &= \mathbf{G}\underset{\sim}{\bar{\mathbf{x}}}_{(c)} - \left(\mathbf{G} \cdot (\boldsymbol{\Omega}/n)_{(c)}\mathbf{G}'\right)^{1/2} \cdot \underset{\sim}{\mathbf{Z}}_{(K-1)d} + \frac{1}{2}\mathbf{G}\left(\left(diag(\boldsymbol{\Omega})\right)_{(c)}\right)', \end{aligned} \tag{3.10}$$

where $\underset{\sim}{\mathbf{T}}_i = \underset{\sim}{\bar{\mathbf{x}}}_i - (\boldsymbol{\Omega}_i / n_i)^{1/2} \cdot \underset{\sim}{\mathbf{Z}}_i + \frac{1}{2}\left(diag(\boldsymbol{\Omega}_i)\right)$ and $\boldsymbol{\Omega}_i = \mathbf{a}_i^{1/2}\mathbf{B}_i^{-1}\mathbf{a}_i^{1/2}$. It is noted that $\mathbf{B}_i$ follows $W_d(n_i - 1,\, \mathbf{I}_d)$ distribution, $\underset{\sim}{\mathbf{Z}}_i$ is $N_d(\underset{\sim}{\mathbf{0}},\, \mathbf{I}_d)$ distribution, $\mathbf{a}_i$ and $\underset{\sim}{\bar{\mathbf{x}}}_i$

are the observed values of $\mathbf{A}_i$ and $\underset{\sim}{\bar{\mathbf{X}}}_i$, accordingly the distribution of $\underset{\sim}{\mathbf{T}}^*$ is free of any nuisance parameter and the observed value of $\underset{\sim}{\mathbf{T}}^*$ does not depend on any unknown parameter and thus the property of $\underset{\sim}{\mathbf{T}}^*$ satisfies the requirement of GTV and can be used to test (3.7).

Suppose $\underset{\sim}{\boldsymbol{\mu}}_{\underset{\sim}{\mathbf{T}}^*}$ and $\mathbf{S}_{\underset{\sim}{\mathbf{T}}^*}$ are the mean and covariance matrix of $\underset{\sim}{\mathbf{T}}^*$, and $\underset{\sim}{\mathbf{T}}_s^*$ represents for the standardized expression of $\underset{\sim}{\mathbf{T}}^*$ with

$$\underset{\sim}{\mathbf{T}}_s^* = \mathbf{S}_{\underset{\sim}{\mathbf{T}}^*}^{-1/2}(\underset{\sim}{\mathbf{T}}^* - \underset{\sim}{\boldsymbol{\mu}}_{\underset{\sim}{\mathbf{T}}^*}), \tag{3.11}$$

then the generalized p-value for testing (3.7) can be obtained through computing

$$p_v = \Pr\left\{ \left\| \underset{\sim}{\mathbf{T}}_s^* \right\| > \left\| \underset{\sim}{\mathbf{0}}_s \right\| \,|\, \mathbf{a}, \underset{\sim}{\bar{\mathbf{x}}} \right\}, \tag{3.12}$$

whether $\underset{\sim}{\mathbf{0}}_s = \mathbf{S}_{\underset{\sim}{\mathbf{T}}^*}^{-1/2}(\underset{\sim}{\mathbf{0}} - \underset{\sim}{\boldsymbol{\mu}}_{\underset{\sim}{\mathbf{T}}^*})$. $\left\|\underset{\sim}{\mathbf{x}}\right\|$ means the Euclidean norm of $\underset{\sim}{\mathbf{x}}$ with $\left\|\underset{\sim}{\mathbf{x}}\right\| = \sqrt{\underset{\sim}{\mathbf{x}}'\underset{\sim}{\mathbf{x}}}$ and (3.7) will be rejected whenever $p_v \le \alpha$, where α is the significant level.

3.3 The classical F-test

In the classical procedure, for mathematical tractability practitioners usually assume that the covariance matrices among populations are homogenous. That is, we will assume $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_\mathrm{K} \equiv \boldsymbol{\Sigma}$ and then the hypothesis testing (3.7) is equivalent to

$$H_0 : \mathbf{G}\underset{\sim}{\boldsymbol{\mu}}_{(\mathrm{C})} = \underset{\sim}{\mathbf{0}} \text{ vs. } H_1 : \mathbf{G}\underset{\sim}{\boldsymbol{\mu}}_{(\mathrm{C})} \neq \underset{\sim}{\mathbf{0}}, \tag{3.13}$$

for the fact that under $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_\mathrm{K} \equiv \boldsymbol{\Sigma}$, $\underset{\sim}{\boldsymbol{\eta}}_i - \underset{\sim}{\boldsymbol{\eta}}_j$ is the same as $\underset{\sim}{\boldsymbol{\mu}}_i - \underset{\sim}{\boldsymbol{\mu}}_j$ and $\underset{\sim}{\boldsymbol{\mu}}'_{(\mathrm{C})} = (\underset{\sim}{\boldsymbol{\mu}}_1{}', \ldots, \underset{\sim}{\boldsymbol{\mu}}'_K)$. Furthermore, the point estimator of $\mathbf{G}\underset{\sim}{\boldsymbol{\mu}}_{(\mathrm{C})}$ is $\mathbf{G}\underset{\sim}{\bar{\mathbf{x}}}$, where

$$\mathbf{G}\underset{\sim}{\bar{\mathbf{x}}} \sim N_d(\mathbf{G}\underset{\sim}{\boldsymbol{\mu}}_{(\mathrm{C})}, \mathbf{G}\boldsymbol{\Phi}\mathbf{G}'), \tag{3.14}$$

$\underset{\sim}{\bar{\mathbf{x}}}' = (\underset{\sim}{\bar{\mathbf{x}}}'_1, \ldots, \underset{\sim}{\bar{\mathbf{x}}}'_K)$, $\boldsymbol{\Phi} = \begin{pmatrix} n_1^{-1}\boldsymbol{\Sigma} & \\ & n_K{}^{-1}\boldsymbol{\Sigma} \end{pmatrix}$ which stands for the block diagonal of $(n_1^{-1}\boldsymbol{\Sigma}_1, \ldots, n_K{}^{-1}\boldsymbol{\Sigma}_K) = (n_1^{-1}, \ldots, n_K{}^{-1})\boldsymbol{\Sigma}$. If the covariance matrices $\boldsymbol{\Sigma}$'s are known, according to (3.14), we have

$$(\mathbf{G}\boldsymbol{\Phi}\mathbf{G}')^{-1/2}\mathbf{G}(\underset{\sim}{\bar{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}}_{(\mathrm{C})}) \equiv \underset{\sim}{\mathbf{Z}}_{(K-1)d} \sim N(\underset{\sim}{\mathbf{0}}, \mathbf{I}_{(K-1)d}). \tag{3.15}$$

Let $\mathbf{S}_{\Phi}=\begin{pmatrix} n_1^{-1}\mathbf{S} & \\ & n_K^{-1}\mathbf{S} \end{pmatrix}$ which stands for the block diagonal of $(n_1^{-1}\mathbf{S},...,n_K^{-1}\mathbf{S})$ and $\mathbf{S}$ is the pool covariance matrix with $\mathbf{S}=\frac{1}{N-K}\sum_{i=1}^{K}(n_i-1)\mathbf{S}_i$, where $\mathbf{S}_i$'s are defined in (3.5) and $N=\sum_{i=1}^{K}n_i$. Hotelling's T-squared statistic (Anderson, 2003) is then defined as

$$t^2=\left(\mathbf{G}(\underset{\sim}{\bar{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}_{(C)})\right)'(\mathbf{G}\mathbf{S}_{\Phi}\mathbf{G}')^{-1}\left(\mathbf{G}(\underset{\sim}{\bar{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}_{(C)})\right), \tag{3.16}$$

and then

$$\frac{t^2}{N-K}\cdot\frac{N-d(K-1)-1}{d(K-1)}\sim F(d(K-1),\ N-d(K-1)-1). \tag{3.17}$$

The p-value for testing (3.13) is

$$p\text{-value}=\Pr\left\{F_{d(K-1),\ N-d(K-1)-1}>\left(\mathbf{G}\underset{\sim}{\bar{\mathbf{x}}}\right)'(\mathbf{G}\mathbf{S}_{\Phi}\mathbf{G}')^{-1}\left(\mathbf{G}\underset{\sim}{\bar{\mathbf{x}}}\right)\cdot\frac{N-d(K-1)-1}{d(N-K)(K-1)}\right\}, \tag{3.18}$$

and null hypothesis wii be rejected if $p\text{-value}\le\alpha$.

3.4 The classical χ^2 - test

On the other hand, the classical χ^2 - test is also a widely applid method in statistical analysis. The classical chi-square method is valid when the covariance matrices are known. If the covaricne matrices are unknown, the researchers usually use the plug-in method to get the approximated solution. In order to test $H_0:\mathbf{G}\boldsymbol{\eta}_{(C)}=\underset{\sim}{\mathbf{0}}$ vs. $H_1:\mathbf{G}\boldsymbol{\eta}_{(C)}\neq\underset{\sim}{\mathbf{0}}$ based on the χ^2-test, we define a statistics H_{chi}^2 with

$$\mathrm{H}_{chi}^2=\left(\mathbf{G}(\underset{\sim}{\bar{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}_{(C)})\right)'(\mathbf{G}\mathbf{S}_{chi}\mathbf{G}')^{-1}\left(\mathbf{G}(\underset{\sim}{\bar{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}_{(C)})\right), \tag{3.19}$$

where $\mathbf{S}_{chi}=\begin{pmatrix} n_1^{-1}\mathbf{S}_1 & \\ & n_K^{-1}\mathbf{S}_K \end{pmatrix}$ is the block diagonal of $(n_1^{-1}\mathbf{S}_1,...,n_K^{-1}\mathbf{S}_K)$. If the sample sized is large, the distributin of H_{chi}^2 is distributed approximately to chi-square distribution with $d(K-1)$ degrees of freedom. The p-value for testing $H_0:\mathbf{G}\boldsymbol{\eta}_{(C)}=\underset{\sim}{\mathbf{0}}$ is

$$p\text{-value} = \Pr\left\{\chi^2_{d(K-1)} > \left(\mathbf{G}\underset{\sim}{\bar{\mathbf{x}}}\right)' (\mathbf{G}\mathbf{S}_{chi}\mathbf{G}')^{-1}\left(\mathbf{G}\underset{\sim}{\bar{\mathbf{x}}}\right)\right\}. \qquad (3.20)$$

4. Numerical Studies

In this section, we present simulation studies by using a variety of parameter configurations and different settings of sample sizes for $K = 2$ and $K = 3$ to perform the hypothesis testing and compare the simulated sizes of the proposed procedure with the classical *F*-test and the classical χ^2-test. According to *Eigen Decomposition Theorem,* for any positive definite matrix $\mathbf{\Sigma}$, there exists an orthogonal matrix $\mathbf{O}$ such that $\mathbf{O}'\mathbf{\Sigma}\mathbf{O}$ is diagonal (Rao, 2001). Thus we choose $\mathbf{\Sigma}_1,...,\mathbf{\Sigma}_K$ to be diagonal in simulation studies. The results were shown in Table 1 to Table 3.

From the above tables, we find the similar pattern in $K = 2$ and $K = 3$ that the type I error rates obtained by the classical F-test is affected by the sample sizes and the degrees of homogeneity. When those populations are not homogenous and the sample sizes are not all equal, the type I error rates based on the classical F-test deteriorate as the degree of heteroscedasticity increases. When $K = 2$ and the nominal level is set to be .05, the type I errors rates by the classical *F*-test are as high as .145 when smaller sample sizes are associated with larger variances, and as low as .010 when larger sample sizes are associated with larger variances. Furthermore, when comparing K mean vectors of several populations, the type I error rates obtained by the classical F-test are unstable, they are as high as .294 and as low sas .000 when the nominal level is .05. Similarly, the performance of the classical χ^2-test is also not well. The type I error rates obtained by the classical χ^2-test are all higher than the nominal level .05 in all combinations which means that the classical χ^2-test tends to reject the true hypothesis in all cases. On the contrary, our proposed test has stable and satisfied type I error rates overall. The type I error rates obtained by our proposed method are around .05 in all cases regardless of the sample sizes, population numbers and heteroscedasticity among groups.

Table 1: Simulated sizes for $H_0 : \mathbf{G}\underset{\sim}{\boldsymbol{\eta}}=\underset{\sim}{\mathbf{0}}$ vs. $H_1 : \mathbf{G}\underset{\sim}{\boldsymbol{\eta}} \neq \underset{\sim}{\mathbf{0}}$ at $d = 2$ and $\mathbf{\Sigma}_1 = \mathbf{I}_2$, $\mathbf{\Sigma}_2 = a \cdot \mathbf{I}_2$.

$(n_1,..,n_K)$		(10, 10)	
a	1	5	10
GP	0.050	0.047	0.048
Classical F test	0.050	0.063	0.068
Classical χ^2-test	0.088	0.103	0.110
$(n_1,..,n_K)$		(10,15)	
a	1	5	10
GP	0.049	0.044	0.044

Classical F test	0.051	0.029	0.025
Classical χ^2-test	0.082	0.085	0.082

$(n_1,..,n_K)$		(15,10)	
a	1	5	10
GP	0.048	0.043	0.042
Classical F test	0.051	0.119	0.140
Classical χ^2-test	0.084	0.112	0.113

$(n_1,..,n_K)$		(50, 30)	
a	1	5	10
GP	0.048	0.051	0.048
Classical F test	0.049	0.125	0.145
Classical χ^2-test	0.058	0.068	0.068

$(n_1,..,n_K)$		(30, 50)	
a	1	5	10
GP	0.051	0.047	0.047
Classical F test	0.051	0.018	0.010
Classical χ^2-test	0.051	0.999	1.000

Results are based on 5,000 repetitions and $\alpha = 0.05$

Table 2: Simulated sizes for $H_0:\mathbf{G}\underset{\sim}{\boldsymbol{\eta}}=\underset{\sim}{\mathbf{0}}$ vs. $H_1:\mathbf{G}\underset{\sim}{\boldsymbol{\eta}}\neq\underset{\sim}{\mathbf{0}}$ at $d=3$ and $\boldsymbol{\Sigma}_1=\mathbf{I}_3,\boldsymbol{\Sigma}_2=\text{diag}(1,1,a)$.

$(n_1,..,n_K)$		(10, 10)	
a	1	5	10
GP	0.052	0.049	0.053
Classical F test	0.050	0.053	0.056
Classical χ^2-test	0.116	0.117	0.122

$(n_1,..,n_K)$		(10,15)	
a	1	5	10
GP	0.050	0.053	0.048
Classical F test	0.052	0.044	0.038
Classical χ^2-test	0.109	0.104	0.103

$(n_1,..,n_K)$		(15,10)	
a	1	5	10
GP	0.048	0.050	0.045
Classical F test	0.052	0.074	0.082
Classical χ^2-test	0.104	0.120	0.118

$(n_1,..,n_K)$		(50, 30)	
a	1	5	10
GP	0.052	0.051	0.048
Classical F test	0.050	0.077	0.090
Classical χ^2-test	0.068	0.068	0.077

$(n_1,..,n_K)$		(30, 50)	
a	1	5	10
GP	0.051	0.053	0.049
Classical F test	0.052	0.034	0.037
Classical χ^2-test	0.052	0.989	1.000

Results are based on 5,000 repetitions and $\alpha = 0.05$

Table 3: Simulated sizes for $H_0:\mathbf{G}\underset{\sim}{\boldsymbol{\eta}}=\underset{\sim}{\mathbf{0}}$ vs. $H_1:\mathbf{G}\underset{\sim}{\boldsymbol{\eta}}\neq\underset{\sim}{\mathbf{0}}$ at $\boldsymbol{\Sigma}_1=a_1\mathbf{I}_d,\boldsymbol{\Sigma}_2=\text{diag}(1,...,1,a_2),\boldsymbol{\Sigma}_3=a_3\mathbf{I}_d$.

$(n_1,..,n_K)$	(10,10,10)			
(a_1,a_2,a_3)	(1, 0.2, 0.2)	(1,1,1)	(1,4,4)	(1,25,25)
GP	0.050	0.049	0.046	0.048
Classical F test	0.111	0.055	0.033	0.035
Classical χ^2-test	0.113	0.093	0.090	0.110
$(n_1,..,n_K)$	(10,15,20)			
(a_1,a_2,a_3)	(1, 0.2, 0.2)	(1,1,1)	(1,4,4)	(1,25,25)
GP	0.045	0.047	0.054	0.051
Classical F test	0.213	0.066	0.009	0.003
Classical χ^2-test	0.153	0.119	0.094	0.096
$(n_1,..,n_K)$	(20,15,10)			
(a_1,a_2,a_3)	(1, 0.2, 0.2)	(1,1,1)	(1,4,4)	(1,25,25)
GP	0.051	0.050	0.044	0.045
Classical F test	0.051	0.051	0.089	0.148
Classical χ^2-test	0.102	0.095	0.112	0.131
$(n_1,..,n_K)$	(30, 50, 100)			
(a_1,a_2,a_3)	(1, 0.2, 0.2)	(1,1,1)	(1,4,4)	(1,25,25)
GP	0.053	0.048	0.050	0.051
Classical F test	0.294	0.071	0.003	0.000
Classical χ^2-test	0.105	0.091	0.066	0.067

Results are based on 5,000 repetitions and $\alpha=0.05$

These simulations support the expected result that the tests based on the generalized variable approach assure the level of the accuracy in all cases. Thus, for overall comparisons, we conclude that our proposed method is stable and suitable for practical use.

5. Concluding remarks

In this article, we have used the generalized variable approach to compare the difference of several independent multivariate log-normal mean vectors. The numerical results have shown that the proposed methods assure the level of the accuracy and the tests are more robust and efficient than other method that is currently available in the literature. Log-normal data are very common in applications, and thus our procedures should be of interest whenever multivariate log-normality holds. The proposed procedures are applicable regardless of the sample sizes and heteroscedasticity. These features should make the generalized variable approach an attractive option for application to practical problems involving the multivariate log-normal distribution.

References

Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis, 3rd edition. Wiley, New York.

Bebu, I. and Mathew, T. (2008) Comparing the means and variances of a bivariate log-normal distribution, *Statistics in Medicine* 27 (2008) 2684-2696.

Chen, Y. H. and Zhou, X. H. (2006). Interval estimates for the ratio and difference of two log-normal means. *Statistics in Medicine*, 25, 4099–4113.

Gupta, R. C. and Li, X. (2006).Statistical inference for the common mean of two log-normal distributions and some applications in reliability, *Computational Statistics & Data Analysis* 50, 3141-3164.

Hawkins, D. M. (2002).Diagnostics for conformity of paired quantitative measurements.*Statistics in Medicine*21, 1913–1935.

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

Lin, S. H. (2013). The higher order likelihood method for the common mean of several log-normal distributions, *Metrika* 76(3), 381-392.

Lin, S. H. and Wang, R. S. (2013). Modified method for several log-normal distributions based on the generalized variables, *Journal of Applied Statistics* 40, 194-208.

Lin, S. H. (2014). Comparing the mean vectors of two independent multivariate log-normal distributions, *Journal of Applied Statistics* 41(2), 259-274.

Rao, C. R. (2001). Linear statistical inference and its applications, 2ndedition. Wiley: New York.

Taylor, D. J., Kupper, L. L. and Muller, K. E. (2002).Improved approximate confidence intervals for the mean of a log-normal random variable.*Statistics in Medicine* 21, 1443–1459.

Tsui, K. W. and Weerahandi, S. (1989). Generalized *p*-value in significance testing of hypotheses in the presence of nuisance parameters.*Journal of the American Statistical Association* 84, 602-607.

Weerahandi S. (1993). Generalized confidence intervals.*Journal of the American Statistical Association* 88, 899–905.

Weerahandi S. (1995). Exact statistical methods for data analysis. Springer: New York.

Weerahandi S. (2004). Generalized inference in repeated measure: Exact methods in MANOVA and mixed models. Wiley: New Jersey.

Wu, J., Jiang, G., Wong A, C. M. and Sun, X. (2002). Likelihood analysis for the ratio of means of two independent log-normal distributions, *Biometrics* 58, 463-469.

Wu, J., Wong A. C. M. and Jiang, G. (2003). Likelihood-based confidence intervals for a log-normal mean, *Statistics in Medicine* 22, 1849-1860.

Zhou, X. H. and Gao, S. (1997). Confidence interval for the log-normal mean, *Statistics in Medicine* 16, 783-790.

Zhou, X. H., Gao, S. and Hui, S. L. (1997). Methods for comparing the means of two independent log-normal samples, *Biometrics* 53, 1129-1135.

Zhou, X. H., Gao, S. and Tierney, W. M. (2001). Methods for testing equality of means of health care costs in a paired design study.*Statistics in Medicine* 20, 1703–1720.