Far East Journal of Mathematical Sciences (FJMS)

http://www.pphmj.com
http://dx.doi.org/10.17654/MS102020361
Volume 102, Number 2, 2017, Pages 361-372 **ISSN: 0972-0871**

COMPARISON OF TWO LOG-LOGISTIC POPULATION MEDIANS IN THE PRESENCE OF UNDETECTED DATA

Aceng Komarudin Mutaqin and Siti Sunendiari

Department of Statistics
Faculty of Mathematics and Natural Sciences
Universitas Islam Bandung
Indonesia

Jl. Ranggamalela No. 1 Bandung 40116
West Java, Indonesia

Abstract

A test procedure is derived for comparing two log-logistic population medians in the presence of undetected data. It is assumed that two independent samples of sizes n_1 and n_2 are available from two log-logistic populations $LLD(\theta_1, \gamma_1)$ and $LLD(\theta_2, \gamma_2)$. The expectation maximization (EM) algorithm is used to estimate the parameters in four cases: $H_1(\theta_1 = \theta_2, \gamma_1 = \gamma_2)$, $H_2(\theta_1 \neq \theta_2, \gamma_1 = \gamma_2)$, $H_3(\theta_1 = \theta_2, \gamma_1 \neq \gamma_2)$ and $H_4(\theta_1 \neq \theta_2, \gamma_1 \neq \gamma_2)$. A guidance is described for testing the equality of the two medians $(H_0 : \theta_1 = \theta_2$ versus $H_r : \theta_1 \neq \theta_2)$. Two procedures are recommended for this test, depending on whether the coefficients of variation are equal

Received: March 1, 2017; Accepted: April 19, 2017

2010 Mathematics Subject Classification: 62F03, 62F10, 62N03.

Keywords and phrases: chi-square test, EM algorithm, log-logistic distribution, undetected data.

Communicated by Suk-Bok Kang; Editor: Far East Journal of Theoretical Statistics: Published by Pushpa Publishing House, Allahabad, India

$(H_0 : \gamma_1 = \gamma_2)$ or not $(H_0 : \gamma_1 \neq \gamma_2)$. Asymptotic chi-square tests are used in the recommended guidance. A case study example of the method is provided using data of vehicle exhaust emissions test.

1. Introduction

In general, there are two approaches to compare two populations containing undetected data, which are the parametric approach and the nonparametric approach. In the parametric approach and the case of two normal populations, Stoline [5] proposed using the test of equality of two population medians to compare two populations containing undetected data. Meanwhile, Zhong et al. [7] used the information of likelihood function to compare two populations containing undetected data. For the same case, a standard test such as the *T* test is often used by researchers (Zhong et al. [7]). In the nonparametric approach, Zhong et al. [7] used the permutation test.

The undetected data are frequently found in environmental data. In addition to the lognormal distribution, other distribution that can be used to model environmental data is a log-logistic distribution (Warsono [6]). Mutaqin and Kudus [3] discussed the permutation test to compare the two log-logistic population medians for the undetected data. Research results of Mutaqin [2] showed that for the data containing a single detection limit, the permutation test is appropriate for the case of a large sample size, low percentages of the undetected observations and big difference of the coefficients of variation. Meanwhile, for the data containing multiple detection limits, the permutation test is appropriate for the case of a large sample size, unequal percentages of the undetected observations and big difference of the coefficients of variation.

In this paper, a test procedure to compare two log-logistic population medians in the presence of undetected data is proposed. The expectation maximization (EM) algorithm is used to estimate the parameters of two log-logistic distributions. Asymptotic chi-square tests are used in the test procedure. A case study example of the method is provided using data of vehicle exhaust emissions test.

2. Log-logistic Distribution

The probability density function of the log-logistic distribution with a scale parameter $\tau > 0$ and a location parameter $-\infty < \beta < \infty$ is (Warsono [6])

$$g(x;\, \tau,\, \beta) = \frac{\tau}{x}\left[\frac{e^{\beta}x^{\tau}}{(1+e^{\beta}x^{\tau})^2}\right]; \quad x > 0.$$

Mutaqin et al. [4] discussed EM algorithm to estimate the parameters τ and β. Let $\gamma = \tau$, and $\theta = e^{-\beta/\tau}$. According to the reparameterization, we obtain the probability density function of the log-logistic distribution

$$f(x;\, \theta,\, \gamma) = \frac{\gamma(x/\theta)^{\gamma}}{x[1+(x/\theta)^{\gamma}]^2}; \quad x > 0,$$

where $\gamma > 0$ is a shape parameter, and $\theta > 0$ is a scale parameter (Klugman et al. [1]). It can be shown that the median of the log-logistic distribution is $M = \theta$, and the coefficient of variation is

$$CV = \frac{(E[X^2] - (E[X])^2)^{1/2}}{E[X]}$$

$$= \frac{\left[\Gamma\left(1+\frac{2}{\gamma}\right)\Gamma\left(1-\frac{2}{\gamma}\right) - \left[\Gamma\left(1+\frac{1}{\gamma}\right)\Gamma\left(1-\frac{1}{\gamma}\right)\right]^2\right]^{1/2}}{\Gamma\left(1+\frac{1}{\gamma}\right)\Gamma\left(1-\frac{1}{\gamma}\right)}.$$

It is noted that the median depends only on θ, and the coefficient of variation depends only on γ.

3. Permutation Test for the Equality of Two Log-logistic Population Medians

Let θ_1, γ_1 and θ_2, γ_2 be the parameters of two log-logistic populations,

respectively. Let the medians of the two log-logistic populations be M_1 and M_2. The two medians are statistically equal whenever hypothesis $H_0 : \theta_1 = \theta_2$ is accepted. For the homogeneous $(\gamma_1 = \gamma_2)$ case, the means and the variances of two log-logistic populations may differ, but the coefficients of variation are equal. Whenever hypothesis $H_0 : \theta_1 = \theta_2$ is accepted in the homogeneous case, it can be inferred that the two log-logistic populations are identical. For the heterogeneous $(\gamma_1 \neq \gamma_2)$ case, whenever hypothesis $H_0 : \theta_1 = \theta_2$ is accepted, it can be inferred that the two log-logistic medians are identical.

Mutaqin and Kudus [3] discussed hypothesis test $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$ using permutation test. Let $X_{11}, ..., X_{1n_1}$ and $X_{21}, ..., X_{2n_2}$ denote independent samples of sizes n_1 and n_2 from the two log-logistic populations $LLD(\theta_1, \gamma_1)$ and $LLD(\theta_2, \gamma_2)$, respectively. For each realization, x_{ij}, it is assumed that there exists a detection limit L_{ij}, for $i = 1, 2$ and $j = 1, 2, ..., n_i$. If x_{ij} is quantified above the detection limit, then the observation value x_{ij} is reported. Otherwise, if x_{ij} is quantified below the detection limit, then a left-censored realization $x_{ij} = L_{ij}$ is reported. Let us assume that the first r_i realizations for sample i are uncensored and that the last $n_i - r_i$ realizations for sample i are each left-censored for $i = 1, 2$. The detailed steps of the permutation test for hypothesis $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$ can be found in Mutaqin and Kudus [3].

4. The Proposed Test Procedure

According to the research results of Stoline [5], we define four hypotheses:

$$H_1 : \theta_1 = \theta_2 = \theta, \gamma_1 = \gamma_2 = \gamma;$$

$$H_2 : \theta_1 \neq \theta_2, \gamma_1 = \gamma_2 = \gamma;$$

$$H_3 : \theta_1 = \theta_2 = \theta,\ \gamma_1 \neq \gamma_2;$$

$$H_4 : \theta_1 \neq \theta_2,\ \gamma_1 \neq \gamma_2.$$

Four tests are proposed using the hypotheses defined above. These are:

$$\text{Test 1: } H_1 \text{ versus } H_4; \tag{1}$$

$$\text{Test 2: } H_2 \text{ versus } H_4; \tag{2}$$

$$\text{Test 3: } H_1 \text{ versus } H_2; \tag{3}$$

$$\text{Test 4: } H_3 \text{ versus } H_4. \tag{4}$$

The following are the three steps for the recommended test strategy:

Step 1. Test the overall homogeneity of the two log-logistic populations using test 1 with p-value p_1.

Step 2. Preliminary test of homogeneity of the coefficients of variation (at level α^*) using test 2 with p-value p_2.

Step 3. Test of the equality of the two medians: use test 3 if $p_2 > \alpha^*$ (a homogeneous case); use test 4 if $p_2 < \alpha^*$ (a heterogeneous case).

Asymptotic chi-square tests are used for tests 1-4 above. The expectation maximization (EM) algorithm is used to estimate the parameters in four cases: H_1 $(\theta_1 = \theta_2, \gamma_1 = \gamma_2)$, H_2 $(\theta_1 \neq \theta_2, \gamma_1 = \gamma_2)$, H_3 $(\theta_1 = \theta_2, \gamma_1 \neq \gamma_2)$ and H_4 $(\theta_1 \neq \theta_2, \gamma_1 \neq \gamma_2)$.

The expectation of the log-likelihood function (E-step in EM algorithm) for case $H_1 : \theta_1 = \theta_2 = \theta,\ \gamma_1 = \gamma_2 = \gamma$ is

$$l_1 = \sum_{i=1}^{2} \left[\sum_{j=1}^{r_i} \left\{ \ln\gamma - \gamma \ln\theta + (\gamma - 1)\ln x_{ij} - 2\ln\left[1 + \left(\frac{x_{ij}}{\theta} \right)^{\gamma} \right] \right\} + \sum_{j=r_i+1}^{n_i} \{\gamma E_{ij}^1 - \gamma \ln\theta - E_{ij}^2\} \right], \tag{5}$$

where

$$E_{ij}^{1} = E[\ln(X_{ij}) | X_{ij} < L_{ij}]$$

$$= \ln\theta + \frac{\ln(F(L_{ij};\, \theta,\, \gamma))}{\gamma} + \frac{\left[\dfrac{1}{F(L_{ij};\, \theta,\, \gamma)} - 1\right]\ln(1 - F(L_{ij};\, \theta,\, \gamma))}{\gamma}$$

and

$$E_{ij}^{2} = E[\ln[1 + (X_{ij}/\theta)^{\gamma}] | X_{ij} < L_{ij}]$$

$$= \ln\left(1 + \frac{F(L_{ij};\, \theta,\, \gamma)}{1 - F(L_{ij};\, \theta,\, \gamma)}\right) + \frac{\ln(1 - F(L_{ij};\, \theta,\, \gamma))}{F(L_{ij};\, \theta,\, \gamma)} + 1,$$

where $F(\cdot;\, \theta,\, \gamma)$ is the cumulative distribution function of the log-logistic distribution. The next step (M-step) is to maximize l_1 obtaining parameter estimates in the case of H_1. The parameters that maximize l_1 are solutions of the following equations:

$$-(n_1 + n_2)\frac{\gamma}{\theta} + \sum_{i=1}^{2}\left[\sum_{j=1}^{r_i} \frac{2\left(\dfrac{\gamma}{\theta}\right)\left(\dfrac{x_{ij}}{\theta}\right)^{\gamma}}{1 + \left(\dfrac{x_{ij}}{\theta}\right)^{\gamma}}\right] = 0$$

and

$$\frac{(r_1 + r_2)}{\gamma} - (n_1 + n_2)\ln\theta$$

$$+ \sum_{i=1}^{2}\left[\sum_{j=1}^{r_i}\left\{\ln x_{ij} - \frac{2\ln\left(\dfrac{x_{ij}}{\theta}\right)\left(\dfrac{x_{ij}}{\theta}\right)^{\gamma}}{1 + \left(\dfrac{x_{ij}}{\theta}\right)^{\gamma}}\right\} + \sum_{j=r_i+1}^{n_i} E_{ij}^{1}\right] = 0.$$

The expectation of the log-likelihood function (E-step in EM algorithm) for case $H_2 : \theta_1 \neq \theta_2, \gamma_1 = \gamma_2$ is

$$l_2 = \sum_{i=1}^{2}\left[\sum_{j=1}^{r_i}\left\{\ln\gamma - \gamma\ln\theta_i + (\gamma-1)\ln x_{ij} - 2\ln\left[1+\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma}\right]\right\} + \sum_{j=r_i+1}^{n_i}\{\gamma E_{ij}^3 - \gamma\ln\theta_i - E_{ij}^4\}\right], \quad (6)$$

where

$$E_{ij}^3 = E[\ln(X_{ij}) | X_{ij} < L_{ij}]$$

$$= \ln\theta_i + \frac{\ln(F(L_{ij};\theta_i,\gamma))}{\gamma} + \frac{\left[\frac{1}{F(L_{ij};\theta_i,\gamma)} - 1\right]\ln(1-F(L_{ij};\theta_i,\gamma))}{\gamma}$$

and

$$E_{ij}^4 = E[\ln[1+(X_{ij}/\theta_i)^{\gamma}] | X_{ij} < L_{ij}]$$

$$= \ln\left(1+\frac{F(L_{ij};\theta_i,\gamma)}{1-F(L_{ij};\theta_i,\gamma)}\right) + \frac{\ln(1-F(L_{ij};\theta_i,\gamma))}{F(L_{ij};\theta_i,\gamma)} + 1.$$

The next step (M-step) is to maximize l_2 obtaining parameter estimates in the case of H_2. The parameters that maximize l_2 are solutions of the following equations:

$$-n_i\frac{\gamma}{\theta_i} + \sum_{j=1}^{r_i}\frac{2\left(\frac{\gamma}{\theta_i}\right)\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma}}{1+\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma}} = 0; \quad i = 1, 2$$

and

$$\frac{(r_1+r_2)}{\gamma} - (n_1\ln\theta_1 + n_2\ln\theta_2)$$

$$+\sum_{i=1}^{2}\left[\sum_{j=1}^{r_i}\left\{\ln x_{ij}-\frac{2\ln\left(\frac{x_{ij}}{\theta_i}\right)\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma}}{1+\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma}}\right\}+\sum_{j=r_i+1}^{n_i}E_{ij}^{3}\right]=0.$$

The expectation of the log-likelihood function (E-step in EM algorithm) for case $H_3 : \theta_1 = \theta_2,\ \gamma_1 \neq \gamma_2$ is

$$l_3=\sum_{i=1}^{2}\left[\sum_{j=1}^{r_i}\left\{\ln\gamma_i-\gamma_i\ln\theta+(\gamma_i-1)\ln x_{ij}-2\ln\left[1+\left(\frac{x_{ij}}{\theta}\right)^{\gamma_i}\right]\right\}\right.$$

$$\left.+\sum_{j=r_i+1}^{n_i}\{\gamma_i E_{ij}^{5}-\gamma_i\ln\theta-E_{ij}^{6}\}\right], \tag{7}$$

where

$$E_{ij}^{5}=[\ln(X_{ij})\,|\,X_{ij}<L_{ij}]$$

$$=\ln\theta+\frac{\ln(F(L_{ij};\,\theta,\,\gamma_i))}{\gamma_i}+\frac{\left[\frac{1}{F(L_{ij};\,\theta,\,\gamma_i)}-1\right]\ln(1-F(L_{ij};\,\theta,\,\gamma_i))}{\gamma_i}$$

and

$$E_{ij}^{6}=[\ln[1+(X_{ij}/\theta)^{\gamma_i}]\,|\,X_{ij}<L_{ij}]$$

$$=\ln\left(1+\frac{F(L_{ij};\,\theta,\,\gamma_i)}{1-F(L_{ij};\,\theta,\,\gamma_i)}\right)+\frac{\ln(1-F(L_{ij};\,\theta,\,\gamma_i))}{F(L_{ij};\,\theta,\,\gamma_i)}+1.$$

The next step (M-step) is to maximize l_3 obtaining parameter estimates in the case of H_3. The parameters that maximize l_3 are solutions of the following equations:

$$-\frac{(n_1\gamma_1+n_2\gamma_2)}{\theta}+\sum_{i=1}^{2}\left[\sum_{j=1}^{r_i}\frac{2\left(\frac{\gamma_i}{\theta}\right)\left(\frac{x_{ij}}{\theta}\right)^{\gamma_i}}{1+\left(\frac{x_{ij}}{\theta}\right)^{\gamma_i}}\right]=0$$

and

$$\frac{r_i}{\gamma_i} - n_i \ln \theta + \sum_{j=1}^{r_i} \left\{ \ln x_{ij} - \frac{2 \ln\left(\frac{x_{ij}}{\theta}\right)\left(\frac{x_{ij}}{\theta}\right)^{\gamma_i}}{1 + \left(\frac{x_{ij}}{\theta}\right)^{\gamma_i}} \right\} + \sum_{j=r_i+1}^{n_i} E_{ij}^5 = 0;\ i = 1,\ 2.$$

The expectation of the log-likelihood function (E-step in EM algorithm) for case $H_4 : \theta_1 \neq \theta_2,\ \gamma_1 \neq \gamma_2$ is

$$l_4 = \sum_{i=1}^{2} \left[\sum_{j=1}^{r_i} \left\{ \ln \gamma_i - \gamma_i \ln \theta_i + (\gamma_1 - 1) \ln x_{ij} - 2 \ln\left[1 + \left(\frac{x_{ij}}{\theta_i}\right)^{\gamma_i} \right] \right\} \right.$$

$$\left. + \sum_{j=r_i+1}^{n_i} \{ \gamma_i E_{ij}^7 - \gamma_i \ln \theta_i - E_{ij}^8 \} \right], \tag{8}$$

where

$$E_{ij}^7 = E[\ln(X_{ij}) | X_{ij} < L_{ij}]$$

$$= \ln \theta_i + \frac{\ln(F(L_{ij};\ \theta_i,\ \gamma_i))}{\gamma_i} + \frac{\left[\frac{1}{F(L_{ij};\ \theta_i,\ \gamma_i)} - 1 \right] \ln(1 - F(L_{ij};\ \theta_i,\ \gamma_i))}{\gamma_i}$$

and

$$E_{ij}^8 = E[\ln[1 + (X_{ij}/\theta_i)^{\gamma_i}] | X_{ij} < L_{ij}]$$

$$= \ln\left(1 + \frac{F(L_{ij};\ \theta_i,\ \gamma_i)}{1 - F(L_{ij};\ \theta_i,\ \gamma_i)} \right) + \frac{\ln(1 - F(L_{ij};\ \theta_i,\ \gamma_i))}{F(L_{ij};\ \theta_i,\ \gamma_i)} + 1.$$

The next step (M-step) is to maximize l_4 obtaining parameter estimates in the case of H_4. The parameters that maximize l_4 are solutions of the following equations:

$$-n_i \frac{\gamma_i}{\theta_i} + \sum_{j=1}^{r_i} \frac{2\left(\frac{\gamma_i}{\theta_i}\right)\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma_i}}{1+\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma_i}} = 0; \quad i = 1, 2$$

and

$$\frac{r_i}{\gamma_i} - n_i \ln \theta_i + \sum_{j=1}^{r_i} \left\{ \ln x_{ij} - \frac{2 \ln\left(\frac{x_{ij}}{\theta_i}\right)\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma_i}}{1+\left(\frac{x_{ij}}{\theta_i}\right)^{\gamma_i}} \right\} + \sum_{j=r_i+1}^{n_i} E_{ij}^7 = 0; \quad i = 1, 2.$$

Asymptotic α-level chi-square tests can be described as follows (for tests 1-4, respectively):

$$\chi_1^2 = -2(l_1 - l_4) \sim \chi_{2,1-\alpha}^2,$$

$$\chi_2^2 = -2(l_2 - l_4) \sim \chi_{1,1-\alpha}^2,$$

$$\chi_3^2 = -2(l_1 - l_2) \sim \chi_{1,1-\alpha}^2,$$

$$\chi_4^2 = -2(l_3 - l_4) \sim \chi_{1,1-\alpha}^2,$$

where $\chi_{r,1-\alpha}^2$ is the upper α-point for a chi-square random variable with r degree of freedom. The values of l_1, l_2, l_3, and l_4 are defined in equations (5), (6), (7) and (8), respectively.

5. Numerical Example

In this section, we apply the proposed test procedure on the data of vehicle exhaust emissions test. The data are concentration of carbon monoxide (CO) of automobiles from two manufacturers A and B. The data values (in percentage) are presented in Table 1.

Our results show that the *p*-values of test 1, test 2 and test 3 are 0.5925, 0.6162 and 0.9615, respectively. Based on the results, it can be inferred that the two distributions of CO concentration of automobiles from two manufacturers A and B are identical.

Table 1. Concentration of CO (in percentage)

Manufacturer A	Manufacturer B
0.00*	0.00*
0.01	0.00*
0.01	0.01
0.01	0.01
0.02	0.01
0.02	0.01
0.03	0.01
0.05	0.02
0.09	0.02
2.21	0.02
2.29	0.17
	0.18
	0.47
	1.50
	2.53

*undetected (detection limit 0.01)

6. Conclusion

In this paper, we proposed a procedure to test for the equality of two log-logistic population medians in the presence of undetected data. The procedure can also be used to test for the equality of two log-logistic population coefficients of variation and to test for the equality of two log-logistic population distributions. Asymptotic chi-square tests are used to test the hypotheses in the procedure. The EM algorithm is used to estimate the parameters in the hypotheses in the procedure.

References

[1] S. A. Klugman, H. H. Panjer and G. E. Willmot, Loss Models: From Data to Decisions, Fourth Edition, Wiley, New York, 2012.

[2] A. K. Mutaqin, The performance of hypothesis testing for two log-logistic populations in the presence of undetected data, Statistika: Journal of Theoretical Statistics and its Applications 15(1) (2015), 31-37.

[3] A. K. Mutaqin and A. Kudus, Comparison of two log-logistic populations in the presence of undetected data, Proceedings of SNaPP: Science, Technology, and Health, 2014, pp. 89-94.

[4] A. K. Mutaqin, A. Kudus and F. T. Safitri, Parameter estimation for log-logistic population in the presence of undetected data, Proceedings of National Conference on Industrial Engineering, Universitas Malikussaleh, 2013, pp. 149-156.

[5] M. R. Stoline, Comparison of two medians using a two-sample lognormal model in environmental contexts, Environmetrics 4(3) (1993), 323-339.

[6] Warsono, Analysis of environmental pollutant data using generalized log-logistic distribution, Dissertation, University of Alabama, Birmingham, 1996.

[7] W. Zhong, R. Shukla, P. Succop, L. Levin, J. Welge and S. Sivaganesan, Statistical approaches to analyze censored data with multiple detection limits, Dissertation, Philosophy University of Cincinnati, 2005.