UW Biostatistics Working Paper Series

3-30-2005

Empirical Likelihood Intervals for the Mean Difference of Two Skewed Populations with Additional Zero Values

W. Zhou
National University of Singapore

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Suggested Citation
Zhou, W. and Zhou, Xiao-Hua, "Empirical Likelihood Intervals for the Mean Difference of Two Skewed Populations with Additional Zero Values" (March 2005). *UW Biostatistics Working Paper Series.* Working Paper 246.
http://biostats.bepress.com/uwbiostat/paper246

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Empirical likelihood intervals for the mean difference of two skewed populations with additional zero values

W. Zhou [1], X.H. Zhou [2,3]

Abstract

We considered the problem of constructing nonparametric confidence intervals for the difference in the means of two independent skewed populations which contain zero values. To account for zero values, we used a two-part model to separately estimate the probability of having any non-zero value and the expected value of positive observations. Under such a two-part model we developed the empirical likelihood (EL) based interval for the difference in the two population means. We then derived asymptotic properties of the proposed method. In a simulation study, we showed that the EL-based interval outperforms the existing normal approximation method and the bootstrap method. Finally, we illustrated the application of the proposed method in a study that assessed the relationship between the excess charges among older patients and the burden of their medical illness.

Key words and Phrases: Empirical Likelihood; Health Economics; Non-parametric Estimation; Skewed Distributions; Zero Costs

1 Introduction

In health economics, the parameter of interest is often the expected value of one population or a sub-population. For example, in a prospective payment model, such as capitation, which has a long history in the financing of private and public sector health care, capitated payments are set at the expected cost of a patient (Maciejewski et al, 2004). In our real example of this paper, we are interested in the difference between the expected diagnostic

[1]Department of Statistics and Applied Probability, National University of Singapore, Singapore
[2]Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195
[3]HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98108

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

charge of patients with depression and patients without depression when all patients have the same comorbid condition, as defined by the Ambulatory Care Group (ACG) system. Such an analysis is complicated by two characteristics of the diagnostic charge data in this study: (1) a certain proportion of patients had zero diagnostic testing charges because these patients didn't have any diagnostic tests done during the study; (2) non-zero diagnostic testing charge observations were highly skewed to the right, and their distributions may be unknown.

When there are no zero values and when non-zero values can be assumed to follow log-normal distributions, several authors have proposed appropriate tests and confidence intervals for comparing the means of two log-normal distributions (Zhou et al, 1997; Zhou et al, 2001; Krishnamoorthy and Mathew, 2003; Wu et al, 2002).

When there are zero values in populations, the most appropriate model for such the data is a two-part model (Duan, 1983; Diehr et al, 1999). Under a two-part model, if the distributions of non-zero values can be assumed to be log-normal, Zhou and Tu (1999) proposed a likelihood ratio test, and Zhou and Tu (2000) provided several confidence intervals for the ratio in the means of two populations.

When non-zero costs cannot be approximated by log-normal distributions, there are no published methods available for constructing good confidence intervals for the difference of means of two skewed populations with additional zero values. We needed to develop an approach that could address both the problem of zero values and the problem of unknown skewed distributions. In this paper we use two-part models to address the problem of zero versus non-zero values. Under the assumed two-part models, we develop the empirical likelihood method to address the second issue that non-zero values are skewed.

Empirical likelihood (EL) methods (Owen, 2001) are particularly suitable for skewed populations. First, empirical likelihood (EL) methods do not assume a symmetry shape, and instead its shapes are determined by data. Second, EL methods allow for confidence interval construction without an information/variance estimator. Third, the EL methods allow us to employ likelihood methods without having to pick a parametric family for the data.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

This paper is organized as follows. In Section 2, we formulate a model for the data and define the parameters of interest. In Section 3 we develop an empirical likelihood-based method for the construction of confidence intervals of the parameters of interest. In Section 4, we conduct simulation studies to assess coverage accuracy, interval length, and bias of the proposed intervals in finite sample sizes. In Section 6, we analyze the motivating example, introduced in the beginning of this section, with the proposed method.

2 Data and model setup

We assume that the two populations of interest contain both zero and positive observations with unknown but skewed distributions. Let $W_1, W_2, \cdots, W_n$ and $V_1, V_2, \cdots, V_m$ be two random samples from these two populations with corresponding means μ and ν, respectively. To deal with zero costs, we use a two-part model for each population. Assume $\delta = P(W_1 = 0) > 0, \eta = P(V_1 = 0) > 0$. We can show that

$$\mu = E(W_i) = (1-\delta)E(W_i \mid W_i > 0), \nu = (1-\eta)E(V_i \mid V_i > 0).$$

We are interested in finding a confidence interval for

$$\theta = \nu - \mu.$$

3 Main Theorem

In this section we develop an empirical likelihood based interval for θ without assuming a parametric distribution for the nonzero observations. Let $\mu^* = E(W_1|W_1 > 0)$ and $\nu^* = E(V_1|V_1 > 0)$. Let n_0 and n_1 be the number of zero and non-zero observations in the first sample $\{W_1, W_2, \cdots, W_n\}$, respectively, and let m_0 and m_1 be the number of zero and non-zero values in the second sample $\{V_1, V_2, \cdots, V_m\}$, respectively. For convenience, we denote the non-zero values as $x_1, \cdots, x_{n_1}$ for the first sample, $y_1, \cdots, y_{m_1}$ for the second sample.

A combined binomial likelihood for δ, η and the empirical likelihood for θ is defined as

$$L(\delta, \eta, \theta) = \delta^{n_0}(1-\delta)^{n_1} \prod_{i=1}^{n_1} p_i \eta^{m_0}(1-\eta)^{m_1} \prod_{j=1}^{m_1} q_j,$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1,\ \sum_{i=1}^{n_1} p_i(x_i - \frac{\mu}{1-\delta}) = 0,\ p_i \geq 0,$$
$$\sum_{j=1}^{m_1} q_j = 1,\ \sum_{j=1}^{m_1} q_j(y_j - \frac{\nu}{1-\eta}) = 0,\ q_j \geq 0.$$

Lagrange multiplier method gives the log-likelihood

$$\begin{aligned} l(\delta,\eta,\mu,\theta) &= n_0 \log\delta + n_1 \log(1-\delta) - \sum_{i=1}^{n_1} \log\Big(1 + \lambda(x_i - \frac{\mu}{1-\delta})\Big) \\ &+ m_0 \log\eta + m_1 \log(1-\eta) - \sum_{j=1}^{m_1} \log\Big(1 + \xi(y_j - \frac{\theta+\mu}{1-\eta})\Big), \end{aligned}$$

where λ, ξ, μ are determined by

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{x_i - \frac{\mu}{1-\delta}}{1 + \lambda(x_i - \frac{\mu}{1-\delta})} = 0, \tag{3.1}$$

$$\frac{1}{m_1}\sum_{j=1}^{m_1} \frac{y_j - \frac{\theta+\mu}{1-\eta}}{1 + \xi(y_j - \frac{\theta+\mu}{1-\eta})} = 0, \tag{3.2}$$

$$\frac{\lambda n_1}{1-\delta} + \frac{\xi m_1}{1-\eta} = 0. \tag{3.3}$$

The likelihood ratio statistic is given by

$$R(\theta) = 2\Big(\sup_{\delta,\eta,\mu,\theta} l(\delta,\eta,\mu,\theta) - \sup_{\delta,\eta,\mu} l(\delta,\eta,\mu,\theta)\Big).$$

THEOREM **3.1** *Let $W_1, W_2, \cdots, W_n$ and $V_1, V_2, \cdots, V_m$ be two random samples from two different populations which consist of both zero and positive observations. Suppose $0 < \delta = P(W_1 = 0) < 1, 0 < \eta = P(V_1 = 0) < 1$. Assume that $EW_1^2 < \infty, EV_1^2 < \infty$ and $0 < n/m \to \rho < 1$ as $n, m \to \infty$,. Then we have*

$$R(\theta) \to \chi_1^2.$$

For a proof, see Appendix.

4 Simulation Results

We conducted a simulation study to assess the coverage accuracy and the average length of empirical likelihood confidence intervals in comparison with existing intervals. We generated

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

non-zero values from three different types of skewed distributions, including exponential, log-normal, and chi-squared distributions; we generated zero values from binomial distributions with different proportions, with $\delta = \eta = 0.2, 0.3, 0.5$, respectively. The numerical results were based on 10,000 pseudo-random samples of various sizes.

In Tables 1 and 2, non-zero values in the first sample (X) were generated from the exponential distribution, $e^{-x}, x > 0$, and non-zero values in the second sample (Y) were generated from the exponential distribution, $e^{-(x-1)}, x \geq 1$. In Tables 3 and 4, non-zero values in the first sample (X) were generated from the log-normal distribution with parameters 0 and 1, and non-zero values in the second sample (Y) were generated from the log-normal distribution with parameters 1 and 2. In Tables 5 and 6, non-zero values in the first sample (X) were generated from a chi-square distribution with one degree of freedom, and non-zero values in the second sample (Y) were generated from a chi-square distribution with three degrees of freedom.

For the purpose of comparison, we also report the confidence intervals based on the asymptotic normality of the nonparametric maximum likelihood (ML) estimator $\hat{\theta}_0$. It is easy to show that the nonparametric ML estimator for θ has the following form:

$$\hat{\theta}_0 = m^{-1} \sum_{j=1}^{m_1} y_j - n^{-1} \sum_{i=1}^{n_1} x_i.$$

Its variance is

$$\sigma^2 = (\mu^*)^2 \delta(1-\delta)/n + (1-\delta)\sigma_X^2/n + (\nu^*)^2 \eta(1-\eta)/m + (1-\eta)\sigma_Y^2/m,$$

where $\sigma_X^2 = E(W_1^2 | W_1 > 0) - (\mu^*)^2, \sigma_Y^2 = E(V_1^2 | V_1 > 0) - (\nu^*)^2$. Replacing μ^*, ν^* by $\bar{x} = n_1^{-1} \sum_{i=1}^{n_1} x_i, \bar{y} = m_1^{-1} \sum_{j=1}^{m_1} y_j$ and σ_X^2, σ_Y^2 by $\sum_{i=1}^{n_1} (x_i - \bar{x})^2/n_1, \sum_{j=1}^{m_1} (y_j - \bar{y})^2/m_1$, we get the sample variance $\hat{\sigma}^2$. Therefore the $(1-\alpha)100\%$ confidence interval based on normal approximation is

$$(\hat{\theta}_0 - z_\alpha \hat{\sigma}, \hat{\theta}_0 + z_\alpha \hat{\sigma}),$$

where z_α is the upper α quantile of $N(0,1)$. For the log-normal case, we also included confidence intervals based the parametric maximum likelihood and bootstrap methods that can be easily derived from the methods in Zhou and Tu (2000).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

TABLES 1-6 GO HERE

From the results in Tables 1-6, we see that in the exponential and chi-squares cases, the empirical likelihood (E.L.) method is very competitive with the non-parametric normal approximation (N.A.) method. In the log-normal case, the E.L. method greatly outperforms the N.A. method in terms of both coverage probability and the length of confidence intervals. For the log-normal case, the performance of the E.L. method is also much better than the parametric maximum likelihood (M.L.) and bootstrap (B.T.) methods. The reason why the EL-based interval outperforms the parametric ML interval may be that the parametric ML interval uses a symmetric form for θ while the EL-based interval does not and allows the data to determine its shape.

5 An application to an real example

Callahan, et al. (1997) studied the relationship between depression and the expected cost of diagnostic testing for a patient. Here, the focus of statistical analysis was on the mean of diagnostic testing cost because the mean can be used to recover the total cost, which reflects the entire diagnostic expenditure in a given patient population. We re-analyzed the real data set in Callahan's study. To illustrate the proposed methods, we analyzed a subset of patients who had a chronic medical condition, as defined by Ambulatory Care Group (ACG) system. We were interested in comparing expected costs between depressed patients and non-depressed patients in this subset. The data set consists of 13 depressed patients and 112 non-depressed patients. The sample means for the depression and non-depression groups are \$588.7 and \$487.9, respectively, with respective standard deviations of 1116.3 and 1097.7. In this data set, there are some zero costs. In the depression group, 4 patients has zero costs, and in the non-depression group, 18 patients had zero costs. In addition, non-zero costs are highly skewed; the sample skewness is 6.49 for the depression group and 2.47 for the non-depression group. Applying the existing normal approximation and our EL-based methods, we obtain that the 95% confidence interval for the difference of the expected means for the depression and non-depression groups; the resulting confidence intervals are

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

$(-314.166, 780.848)$ using the normal approximation method, and $(-374.172, 851.084)$ using the empirical likelihood method. The empirical confidence interval is wider than the interval based on the normal approximation. The result is consistent with our simulation results which have shown that the normal approximation interval has a coverage probability that is lower than the nominal level while the empirical interval has a coverage probability that is close to the nominal level.

6 Discussion

In this paper we have developed an empirical likelihood (EL) based interval for the difference between two skewed populations with additional zero values. The main advantage of the EL method is that it employs likelihood methods without having to pick a parametric family for the data. Our simulation studies showed that the EL-based interval outperforms the normal approximation-based and the bootstrap methods, and the improvement can be huge when non-zero values are highly skewed.

Acknowledgements

This work is supported in part by AHRQ R01HS013105-01.

References

Callahan, C. M., Kesterson, J. G., Tierney, W. M. (1997). Association of symptoms of depression with diagnostic test charges. *Annals of Internal Medicine*, 126, 426-432.

Maciejewski, M. L., Zhou, X. H., Fortney, J. C., and Burgess, J. F. (2004). Alternative Methods for Modeling Heteroskedastic Non-normally Distributed Costs. *Health Economics*, **submitted**.

Owen, A. (2001). Empirical Likelihood. *Chapman & Hall/CRC.* New York.

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.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Wu, J., Jiang, G.Y., Wong, A.C.M., and Sun, X. (2002).Likelihood analysis for the ratio of means of two independent log normal distributions. *Biometrics,* 58:144-50.

Zhou, X. H. and Tu, W. (1999). Comparison of several independent population means when their sample contain log-normal and possibly zero observations. *Biometrics,* 55, 645-651.

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., Li, C., Gao, S., and Tierney, W. (2001). Methods for testing equality of means of health care costs in a paired design study. *Stat in Med,* 20, 1703-1720.

Zhou, X. H., and Tu, W. (2000). Confidence Intervals for the ratio of the log-normal means. *Computational Statistics and Data Analysis,* 35, 201-210.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

Appendix. A proof for Theorem 3.1

We first derive $\sup_{\delta,\eta,\mu,\theta} l(\delta,\eta,\mu,\theta)$.

Differentiating l with respect to δ,η,μ and θ, we obtain that

$$\frac{\partial l}{\partial \delta} = \frac{n_0}{\delta} - \frac{n_1}{1-\delta} + \frac{\lambda\mu n_1}{(1-\delta)^2}, \tag{6.4}$$

$$\frac{\partial l}{\partial \eta} = \frac{m_0}{\eta} - \frac{m_1}{1-\eta} + \frac{\xi\nu m_1}{(1-\eta)^2}, \tag{6.5}$$

$$\frac{\partial l}{\partial \mu} = \frac{\lambda n_1}{1-\delta} + \frac{\xi m_1}{1-\eta}, \tag{6.6}$$

$$\frac{\partial l}{\partial \theta} = \frac{\xi m_1}{1-\eta}. \tag{6.7}$$

Setting the expressions in (6.6) and (6.7) to zero, we obtain that $\xi = \lambda = 0$. Setting the expressions in (6.4) and (6.5) to 0, we obtain the following non-parametric ML estimators:

$$\hat{\delta}_0 = n_0/n, \hat{\eta}_0 = m_0/m, \hat{\mu}_0 = n^{-1}\sum_{i=1}^{n_1} x_i, \hat{\theta}_0 = m^{-1}\sum_{j=1}^{m_1} y_j - n^{-1}\sum_{i=1}^{n_1} x_i. \tag{6.8}$$

Hence

$$\begin{aligned}\sup_{\delta,\eta,\mu,\theta} l(\delta,\eta,\mu,\theta) &= n_0 \log\frac{n_0}{n} + n_1 \log\frac{n_1}{n} \\ &+ m_0 \log\frac{m_0}{m} + m_1 \log\frac{m_1}{m}.\end{aligned}$$

We now turn to $\sup_{\delta,\eta,\mu} l(\delta,\eta,\mu,\theta)$ for a fixed value of θ. In order to get the supremum, we need to solve equations (3.1), (3.2), (6.4), (6.5) and (6.6), whose solutions are denoted by $\hat{\lambda},\hat{\delta},\hat{\xi},\hat{\eta}$ and $\hat{\mu}$. Using standard arguments in the empirical likelihood literature, we may show that $\hat{\lambda} = O_p(n^{-\frac{1}{2}}), \hat{\delta} - \delta = O_p(n^{-\frac{1}{2}}), \hat{\xi} = O_p(n^{-\frac{1}{2}}), \hat{\eta} - \eta = O_p(n^{-\frac{1}{2}}), \hat{\mu} - \mu = O_p(n^{-\frac{1}{2}})$. Also note that $\frac{n_0}{n} - \delta = O_p(n^{-\frac{1}{2}}), \frac{m_0}{m} - \eta = O_p(n^{-\frac{1}{2}})$. Applying Taylor's expansions to (6.4), (3.1), (6.5), (3.2), (6.6) at $\lambda = 0, \delta = \frac{n_0}{n}, \xi = 0, \eta = \frac{m_0}{m}, \mu$, respectively, we obtain that

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \frac{\mu}{1-\frac{n_0}{n}}) \\ 0 \\ \frac{1}{m_1}\sum_{j=1}^{m_1}(y_j - \frac{\nu}{1-\frac{m_0}{m}}) \\ 0 \end{pmatrix} + S \begin{pmatrix} \hat{\lambda} \\ \hat{\delta} - \frac{n_0}{n} \\ \hat{\xi} \\ \hat{\eta} - \frac{m_0}{m} \\ \hat{\mu} - \mu \end{pmatrix} + o_p\left(n^{-\frac{1}{2}}\right) + o_p\left(m^{-\frac{1}{2}}\right),$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where

$$S =: \begin{pmatrix} \frac{\mu}{(1-\frac{n_0}{n})^2} & -\frac{n_0}{n_1(\frac{n_0}{n})^2} - \frac{1}{(1-\frac{n_0}{n})^2} & 0 & 0 & 0 \\ -s_x^2 & -\frac{\mu}{(1-\frac{n_0}{n})^2} & 0 & 0 & -\frac{1}{1-\frac{n_0}{n}} \\ 0 & 0 & \frac{\nu}{(1-\frac{m_0}{m})^2} & -\frac{m_0}{m_1(\frac{m_0}{m})^2} - \frac{1}{(1-\frac{m_0}{m})^2} & 0 \\ 0 & 0 & -s_y^2 & -\frac{\nu}{(1-\frac{m_0}{m})^2} & -\frac{1}{1-\frac{m_0}{m}} \\ \frac{1}{1-\frac{n_0}{n}} & 0 & \frac{m_1}{n_1(1-\frac{m_0}{m})} & 0 & 0 \end{pmatrix}.$$

Here

$$s_x^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \frac{\mu}{1-\frac{n_0}{n}})^2,\ s_y^2 = \frac{1}{m_1}\sum_{j=1}^{m_1}(y_j - \frac{\nu}{1-\frac{m_0}{m}})^2.$$

Noting that

$$\frac{n_0}{n} - \delta = O_p\left(n^{-\frac{1}{2}}\right), \frac{m_0}{m} - \eta = O_p\left(m^{-\frac{1}{2}}\right), s_x^2 \to \sigma_x^2, s_y^2 \to \sigma_y^2,$$

we have

$$\begin{pmatrix} 0 \\ \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \frac{\mu}{1-\frac{n_0}{n}}) \\ 0 \\ \frac{1}{m_1}\sum_{j=1}^{m_1}(y_j - \frac{\nu}{1-\frac{m_0}{m}}) \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{\mu}{(1-\delta)^2} & \frac{1}{\delta(1-\delta)^2} & 0 & 0 & 0 \\ \sigma_x^2 & \frac{\mu}{(1-\delta)^2} & 0 & 0 & \frac{1}{1-\delta} \\ 0 & 0 & -\frac{\nu}{(1-\eta)^2} & \frac{1}{\eta(1-\eta)^2} & 0 \\ 0 & 0 & \sigma_y^2 & \frac{\nu}{(1-\eta)^2} & \frac{1}{1-\eta} \\ \rho & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{\lambda} \\ \hat{\delta} - \frac{n_0}{n} \\ \hat{\xi} \\ \hat{\eta} - \frac{m_0}{m} \\ \hat{\mu} - \mu \end{pmatrix} + o_p\left(n^{-\frac{1}{2}}\right) + o_p\left(m^{-\frac{1}{2}}\right).$$

Therefore,

$$\hat{\delta} - \frac{n_0}{n} = \mu\delta\hat{\lambda} + o_p\left(n^{-\frac{1}{2}}\right) + o_p\left(m^{-\frac{1}{2}}\right), \tag{6.9}$$

$$\hat{\eta} - \frac{m_0}{m} = \nu\eta\hat{\xi} + o_p\left(n^{-\frac{1}{2}}\right) + o_p\left(m^{-\frac{1}{2}}\right), \tag{6.10}$$

$$\rho\hat{\lambda} + \hat{\xi} = o_p\left(n^{-\frac{1}{2}}\right) + o_p\left(m^{-\frac{1}{2}}\right), \tag{6.11}$$

$$\hat{\lambda} = \frac{(1-\delta)(\bar{x} - \frac{\mu}{1-\frac{n_0}{n}}) - (1-\eta)(\bar{y} - \frac{\nu}{1-\frac{m_0}{m}})}{(1-\delta)(\sigma_x^2 + \frac{\mu^2}{(1-\delta)^2}\delta) + \rho(1-\eta)(\sigma_y^2 + \frac{\nu^2}{(1-\eta)^2}\eta)} + o_p\left(n^{-\frac{1}{2}}\right) + o_p\left(m^{-\frac{1}{2}}\right). \tag{6.12}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

Again, using Taylor expansions we obtain that

$$
\begin{aligned}
& n_0 \log \frac{n_0/n}{\hat{\delta}} + n_1 \log \frac{n_1/n}{1-\hat{\delta}} + \sum_{i=1}^{n_1} \log \left(1 + \hat{\lambda}\Big(x_i - \frac{\hat{\mu}}{1-\hat{\delta}}\Big)\right) \\
= \; & -n_0 \log \left(1 + \frac{1}{n_0/n}(\hat{\delta} - \frac{n_0}{n})\right) - n_1 \log \left(1 - \frac{1}{n_1/n}(\hat{\delta} - \frac{n_0}{n})\right) \\
& + \sum_{i=1}^{n_1} \log \left(1 + \hat{\lambda}\Big(x_i - \frac{\mu}{1-\frac{n_0}{n}} - \frac{\hat{\mu}-\mu}{1-\frac{n_0}{n}} - \frac{\mu}{(1-\frac{n_0}{n})^2}(\hat{\delta} - \frac{n_0}{n})\Big) + o_p(n^{-1})\right) \\
= \; & -n_0 \left(\frac{1}{n_0/n}(\hat{\delta} - \frac{n_0}{n}) - \frac{1}{2(n_0/n)^2}(\hat{\delta} - \frac{n_0}{n})^2 + o_p(n^{-1})\right) \\
& -n_1 \left(-\frac{1}{n_1/n}(\hat{\delta} - \frac{n_0}{n}) - \frac{1}{2(n_1/n)^2}(\hat{\delta} - \frac{n_1}{n})^2 + o_p(n^{-1})\right) \\
& +n_1 \hat{\lambda} \left(\bar{x} - \frac{\mu}{1-\frac{n_0}{n}} - \frac{\hat{\mu}-\mu}{1-\frac{n_0}{n}} - \frac{\mu}{(1-\frac{n_0}{n})^2}(\hat{\delta} - \frac{n_0}{n})\right) \\
& -\frac{1}{2}\hat{\lambda}^2 \sum_{i=1}^{n_1} \left(x_i - \frac{\mu}{1-\frac{n_0}{n}} - \frac{\hat{\mu}-\mu}{1-\frac{n_0}{n}} - \frac{\mu}{(1-\frac{n_0}{n})^2}(\hat{\delta} - \frac{n_0}{n})\right)^2 + o_p(1) \\
= \; & \frac{n^2}{2}\left(\frac{1}{n_0} + \frac{1}{n_1}\right)(\hat{\delta} - \frac{n_0}{n})^2 - \frac{1}{2} n_1 \sigma_x^2 \hat{\lambda}^2 \\
& +n_1 \hat{\lambda} \left(\bar{x} - \frac{\mu}{1-\frac{n_0}{n}} - \frac{\hat{\mu}-\mu}{1-\frac{n_0}{n}} - \frac{\mu}{(1-\frac{n_0}{n})^2}(\hat{\delta} - \frac{n_0}{n})\right) + o_p(1) \\
= \; & \frac{n}{2}\frac{\delta \mu^2}{1-\delta}\hat{\lambda}^2 + \frac{1}{2} n_1 \sigma_x^2 \hat{\lambda}^2 + o_p(1), \\
& m_0 \log \frac{m_0/m}{\hat{\eta}} + m_1 \log \frac{m_1/m}{1-\hat{\eta}} + \sum_{j=1}^{m_1} \log \left(1 + \hat{\xi}\Big(y_j - \frac{\hat{\nu}}{1-\hat{\eta}}\Big)\right) \\
= \; & \frac{m}{2}\frac{\eta \nu^2}{1-\eta}\hat{\xi}^2 + \frac{1}{2} m_1 \sigma_y^2 \hat{\xi}^2 + o_p(1).
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\frac{1}{2}R(\theta) &= \hat{\lambda}^2 \frac{n}{2}\left((1-\delta)(\sigma_x^2 + \frac{\mu^2}{(1-\delta)^2}\delta) + \rho(1-\eta)(\sigma_y^2 + \frac{\nu^2}{(1-\eta)^2}\eta)\right) + o_p(1) \\
&= \frac{n}{2}\frac{\left((1-\delta)(\bar{x} - \frac{\mu}{1-\frac{n_0}{n}}) - (1-\eta)(\bar{y} - \frac{\nu}{1-\frac{m_0}{m}})\right)^2}{(1-\delta)(\sigma_x^2 + \frac{\mu^2}{(1-\delta)^2}\delta) + \rho(1-\eta)(\sigma_y^2 + \frac{\nu^2}{(1-\eta)^2}\eta)} + o_p(1) \\
&= \frac{\left(\sqrt{n}(1-\delta)\frac{n}{n_1}(\frac{n_1}{n}\bar{x} - \mu) - \sqrt{\rho}\sqrt{m}(1-\eta)\frac{m}{m_1}(\frac{m_1}{m}\bar{y} - \nu)\right)^2}{(1-\delta)(\sigma_x^2 + \frac{\mu^2}{(1-\delta)^2}\delta) + \rho(1-\eta)(\sigma_y^2 + \frac{\nu^2}{(1-\eta)^2}\eta)} + o_p(1).
\end{aligned}
$$

Noting that

$$
Var\left(\frac{n_1}{n}\bar{x} - \mu\right) = E\left(\frac{n_1^2}{n^2} Var(\bar{x}\|n_1)\right) + Var\left(\frac{n_1}{n} E(\bar{x}\|n_1)\right)
$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$
\begin{aligned}
&= E\left(\frac{n_1}{n^2}\sigma_x^2\right) + \frac{\mu^2}{n^2(1-\delta)^2}Var(n_1) \\
&= \frac{(1-\delta)\sigma_x^2}{n} + \frac{\delta\mu^2}{n(1-\delta)}, \\
Var\left(\frac{m_1}{m}\bar{y} - \nu\right) &= \frac{(1-\eta)\sigma_y^2}{m} + \frac{\eta\nu^2}{m(1-\eta)},
\end{aligned}
$$

we have

$$
\begin{aligned}
&\sqrt{n}(1-\delta)\frac{n}{n_1}(\frac{n_1}{n}\bar{x} - \mu) - \sqrt{\rho}\sqrt{m}(1-\eta)\frac{m}{m_1}(\frac{m_1}{m}\bar{y} - \nu) \\
\rightarrow_d\ & N\left(0, (1-\delta)(\sigma_x^2 + \frac{\mu^2}{(1-\delta)^2}\delta) + \rho(1-\eta)(\sigma_y^2 + \frac{\nu^2}{(1-\eta)^2}\eta)\right),
\end{aligned}
$$

which implies

$$R(\theta) \rightarrow_d \chi_1^2$$

as $n, m \rightarrow \infty$. ∎

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

Table 1: Coverage accuracy with nominal level 0.90 when skewed data are from exponential distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation.)

	Method	$\delta = \eta = 0.2$ (length)	$\delta = \eta = 0.3$ (length)	$\delta = \eta = 0.5$(length)
$n = m = 50$	E.L.	0.891 (0.714)	0.888 (0.707)	0.889 (0.692)
	N.A.	0.897 (0.716)	0.894 (0.724)	0.899 (0.695)
$n = m = 70$	E.L.	0.890 (0.608)	0.890 (0.606)	0.891 (0.589)
	N.A.	0.895 (0.606)	0.897 (0.613)	0.896 (0.588)
$n = m = 100$	E.L.	0.894 (0.511)	0.896 (0.512)	0.895 (0.494)
	N.A.	0.896 (0.507)	0.899 (0.513)	0.899 (0.492)

Table 2: Coverage accuracy with nominal level 0.95 when skewed data are from exponential distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation.)

	Method	$\delta = \eta = 0.2$(length)	$\delta = \eta = 0.3$(length)	$\delta = \eta = 0.5$(length)
$n = m = 50$	E.L.	0.943 (0.854)	0.942 (0.845)	0.945 (0.831)
	N.A.	0.948 (0.853)	0.948 (0.863)	0.952 (0.829)
$n = m = 70$	E.L.	0.943 (0.728)	0.946 (0.725)	0.947 (0.706)
	N.A.	0.947 (0.722)	0.950 (0.730)	0.951 (0.701)
$n = m = 100$	E.L.	0.944 (0.611)	0.945 (0.612)	0.944 (0.592)
	N.A.	0.948 (0.605)	0.948 (0.612)	0.948 (0.587)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 3: Coverage accuracy with nominal level 0.90 when skewed data are from log-normal distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation; M.L.: maximum likelihood; B.T.: bootstrap)

	Method	$\delta = \eta = 0.2$(length)	$\delta = \eta = 0.3$(length)	$\delta = \eta = 0.5$(length)
$n = m = 50$	E.L.	0.822 (3.860)	0.811 (3.282)	0.794 (2.792)
	N.A.	0.789 (6.226)	0.783 (5.769)	0.782 (4.790)
	M.L.	0.732 (4.607)	0.745 (4.411)	0.745 (3.918)
	B.T.	0.815 (5.094)	0.830 (4.986)	0.744 (3.626)
$n = m = 70$	E.L.	0.842 (3.777)	0.832 (3.076)	0.812 (2.695)
	N.A.	0.806 (5.405)	0.800 (5.066)	0.797 (4.259)
	M.L.	0.731 (3.838)	0.735 (3.669)	0.751 (3.253)
	B.T.	0.795 (4.059)	0.805 (3.930)	0.742 (3.030)
$n = m = 100$	E.L.	0.858 (2.967)	0.855 (2.768)	0.830 (2.443)
	N.A.	0.824 (4.672)	0.821 (4.397)	0.811 (3.689)
	M.L.	0.718 (3.182)	0.729 (3.035)	0.746 (2.683)
	B.T.	0.788 (3.359)	0.797 (3.209)	0.753 (2.532)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

Table 4: Coverage accuracy with nominal level 0.90 when skewed data are from log-normal distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation; M.L.: maximum likelihood; B.T.: bootstrap)

	Method	$\delta = \eta = 0.2$(length)	$\delta = \eta = 0.3$(length)	$\delta = \eta = 0.5$(length)
$n = m = 50$	E.L.	0.888 (4.017)	0.875 (3.427)	0.855 (2.848)
	N.A.	0.834 (7.419)	0.831 (6.874)	0.829 (5.708)
	M.L.	0.788 (5.490)	0.797 (5.256)	0.795 (4.669)
	B.T.	0.888 (6.216)	0.896 (6.042)	0.811 (4.304)
$n = m = 70$	E.L.	0.908 (3.949)	0.898 (3.203)	0.876 (2.799)
	N.A.	0.852 (6.440)	0.848 (6.037)	0.845 (5.075)
	M.L.	0.792 (4.573)	0.795 (4.371)	0.801 (3.876)
	B.T.	0.861 (4.832)	0.878 (4.768)	0.800 (3.586)
$n = m = 100$	E.L.	0.920 (3.129)	0.913 (2.918)	0.895 (2.580)
	N.A.	0.869 (5.567)	0.868 (5.239)	0.859 (4.397)
	M.L.	0.783 (3.792)	0.792 (3.616)	0.802 (3.197)
	B.T.	0.860 (4.046)	0.865 (3.839)	0.820 (3.015)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 5: Coverage accuracy with nominal level 0.90 when skewed data are from χ^2 distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation.)

	Method	$\delta = \eta = 0.2$(length)	$\delta = \eta = 0.3$(length)	$\delta = \eta = 0.5$(length)
$n = m = 50$	E.L.	0.857 (1.314)	0.856 (1.278)	0.869 (1.185)
	N.A.	0.865 (1.315)	0.861 (1.288)	0.877(1.176)
$n = m = 70$	E.L.	0.859(1.120)	0.854(1.093)	0.879 (1.010)
	N.A.	0.862(1.113)	0.858(1.092)	0.880 (1.000)
$n = m = 100$	E.L.	0.852(0.940)	0.860 (0.921)	0.879 (0.848)
	N.A.	0.863(0.933)	0.864 (0.917)	0.881(0.839)

Table 6: Coverage accuracy with nominal level 0.95 when skewed data are from χ^2 distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation.)

	Method	$\delta = \eta = 0.2$(length)	$\delta = \eta = 0.3$(length)	$\delta = \eta = 0.5$(length)
$n = m = 50$	E.L.	0.917 (1.560)	0.919 (1.519)	0.926 (1.415)
	N.A.	0.923 (1.567)	0.920(1.535)	0.928(1.401)
$n = m = 70$	E.L.	0.921(1.337)	0.917(1.304)	0.931(1.210)
	N.A.	0.922(1.326)	0.920 (1.301)	0.937 (1.191)
$n = m = 100$	E.L.	0.923(1.125)	0.926(1.102)	0.934(1.017)
	N.A.	0.924 (1.112)	0.926(1.092)	0.935 (0.999)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper246

Table 7: Coverage accuracy with $\delta = 18/112$, $\eta = 4/13$ with various skewed distributions. (E.L.: smoothed empirical likelihood; N.A.: normal approximation.)

Nominal level	Method	Exponential(length)	Lognormal(length)	Chisquare(length)
0.9	E.L.	0.864 (0.980)	0.759 (7.717)	0.829(2.151)
	N.A.	0.868 (1.123)	0.673 (8.888)	0.837(2.187)
0.95	E.L.	0.923 (1.154)	0.828 (8.485)	0.891(2.537)
	N.A.	0.920 (1.338)	0.713 (10.591)	0.889(2.606)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press