Appl. Math. J. Chinese Univ.
2020, 35(2): 203-219

A new two-part test based on density ratio model for zero-inflated continuous distributions

LU Ya-hui[1] LIU Ai-yi[2] JIANG Meng-jie[1] JIANG Tao[1,3,*]

Abstract. In this paper, we consider testing the hypothesis concerning the means of two independent semicontinuous distributions whose observations are zero-inflated, characterized by a sizable number of zeros and positive observations from a continuous distribution. The continuous parts of the two semicontinuous distributions are assumed to follow a density ratio model. A new two-part test is developed for this kind of data. The proposed test takes the sum of one test for equality of proportions of zero values and one conditional test for the continuous distribution. The test is proved to follow a χ^2 distribution with two degrees of freedom. Simulation studies show that the proposed test controls the type I error rates at the desired level, and is competitive to, and most of the time more powerful than two popular tests. A real data example from a dietary intervention study is used to illustrate the usefulness of the proposed test.

§1 Introduction

In many research fields, data from a semicontinuous distribution are commonly encountered, i.e., with a clump of observations at zero and positive continuous data. For example, in a dietary intake study, some food components are eaten daily by almost every study participant, while others are consumed episodically, so that the food intake record data are characterized by many zeros for the latter food components[8]; in meteorology study, the clump of zero observations may correspond to the the number of zero rainfall measurements recorded over several years[13]; in household expenditure study, some households spend nothing on a certain commodity during the period of investigation.

A random variable that follows a semicontinuous distribution can be defined as a response variable $y = (x, d)$, where $d = 1$ if y is observed (or positive), and 0 if it is zero (or missing or

Received: 2019-11-10. Revised: 2020-01-09.
MR Subject Classification: 62F03.
Keywords: two-part test, zero-inflated continuous distributions, density ratio model.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-020-3957-x.
Supported by the National Natural Science Foundation of China (No.11971433), the First Class Discipline of Zhejiang-A (Zhejiang Gongshang University-Statistics), the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.
*Corresponding author.

below the limits of detection) and x is the response if $d = 1$ and is undefined (or 0) if $d = 0$. We define the probability distribution function in the ith group as

$$f_i(x, d) = \left[{p_i}^{1-d} \{(1 - p_i) h_i(x)\}^d \right].$$

This is the (conditional) distribution of x (the continuous response) multiplied by the (binomial) probability of d (the indicator of a 0, nonresponse outcome) in the ith population. For a two-sample test, the null hypothesis is $H_0 : (p_1 = p_2) \cap (\mu_1 = \mu_2)$ where μ_i represents the location parameter of $h_i(x)$. Thus, we can test equality of the proportion of zeros and mean equality of the distributions of non-zeros.

In many literatures, two-part tests have been widely used to compare two samples from the non-standard mixture distribution. For example, Lachenbruch[10,11] comprehensively studied two-part tests for two populations. The two-part model tests are defined as $V^2 = B^2 + T^2$ where B is the usual binomial test and T can be the t-test (TBT) or the Wilcoxon test (TBW). The two-part tests use a χ^2 test with two degrees of freedom based on the sum of one test for equality of proportions of zero values and one conditional test for the continuous responses. For large samples, it was shown that B and T are independent under the assumptions of independent errors of the binomial and continuous parts of the distribution. It is also easy to see that the statistic V^2 has a χ^2 distribution with two degrees of freedom[10,11]. The two-part tests and their extensions have been successfully implemented in various applications[2,17,18]. Further ideas and comparisons of some existing one- and two-part procedures may be found in[5, 7, 20]. The existing TBT and TBW tests are either inefficient when no parametric assumptions are made for the positive components or are not robust when the parametric models are assumed. It is therefore desirable to borrow efficiency across similar populations to improve power of the test. At the same time, we also hope that a test is robust to deviations from the model assumptions. The semiparametric density ratio model (DRM) of Anderson[1], which gained popularity after Qin and Zhang[15], is a natural tool to use here. Therefore, under DRM, we propose a new semiparametric hypothesis testing method on the difference of two population means for the continuous responses with a clump of zero.

Some semiparametric statistical methods were recently developed under DRM, and these methods are usually more robust than parametric methods and more effective than nonparametric methods. The DRM is flexible and includes many parametric distribution families, such as the log-normal and gamma distributions, as special cases. In the literature, the DRM has been recognized as a powerful semiparametric tool in many statistical problems. For example, Qin and Zhang[15] and Zhang[19] showed the close relationship between the DRM and logistic regression models, and they further developed procedures to assess the goodness of fit of the logistic regression models based on case-control data. Qin[16] and Zou et al.[21] applied the DRM to a semiparametric mixture model. Fokianos et al.[6] and Cai et al.[4] considered hypothesis testing problems under the DRM without excess zero observations. Therefore, for a two-sample with excess zeros test, the null hypothesis is $H_0 : (p_1 = p_2) \cap (\mu_1 = \mu_2)$, we propose a new two-part model test (TBSE) by defined B as the usual binomial test for equality of proportions and T as a semiparametric hypothesis test under a semiparametric density ratio model for the continuous responses.

The rest of the article is arranged as follows. In Section 2, we describe the mathematical formulation of the testing problem, and describe test statistics including binomial test and Wilcoxon test. In Section 3, we introduce the proposed semiparametric hypothesis testing method based on DRM, and propose a two-part semiparametric Wald test statistic under the null hypothesis. We further provide the asymptotic properties of the proposed ststistic. In Section 4, we conduct simulation studies to evaluate the type I error and power of the proposed TBSE against the TBT and TBW methods. Real data from a dietary intervention study (the CHEF trial) are used to illustrate the methods in Section 5, and some concluding remarks are given in Sections 6. For the convenience of presentation, all proofs are given in the Appendix.

§2 Mathematical Formulation and Methodology

Consider a two-sample problem with m samples drawn from experimental Group 1 and n samples drawn from experimental Group 2. Let $x_1, x_2, \ldots, x_m$ be independent, identically-distributed bivariate observations of $X = (Z, D)$ from Group 1. $x = (z, d)$ denotes an observed value of X, where z is a non-negative real number and d is an indicator variable with the value $d = 1$ if $z > 0$ and $d = 0$ if $z = 0$. Then, the probability distribution of X is $h(x, d) = \left[{p_1}^{1-d}\{(1-p_1)\, g(z, \mu_1)\}^d\right]$, where $g(z, \mu_1)$ is a parametric density with mean μ_1. Similarly, let $y_1, y_2, \ldots, y_n$ be independent, identically-distributed bivariate observations of $Y = (W, E)$ from Group 2 where W is a non-negative random variable and E is a random indicator variable with values $E = 1$ if $W > 0$ and 0 otherwise. Then the density of Y at $W = w, E = e$ is $h(y, e) = \left[{p_2}^{1-e}\{(1-p_2)\, f(w, \mu_2)\}^e\right]$, where $f(w, \mu_2)$ is a parametric density with mean μ_2.

2.1 Binomial test

The binomial test statistic is

$$B = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}_c\,(1-\hat{p}_c)\left(\frac{1}{m_0} + \frac{1}{n_0}\right)}},$$

where $\hat{p}_1 = \frac{m_0}{m}$, $\hat{p}_2 = \frac{n_0}{n}$, $\hat{p}_c = \frac{m_0+n_0}{m+n}$, and m_0 and n_0 are the numbers of zero value observed in Groups 1 and 2, respectively.

2.2 Wilcoxon test

The Wilcoxon test statistic is

$$U = m_c n_c + \frac{m_c\,(m_c+1)}{2} - R$$

where m_c and n_c are the numbers of observations in the continuous component of Groups 1 and 2 respectively, and R is the sum of the ranks in Group 1. We normalize with continuity correction this statistic to $T = \frac{|U-\mu|-0.5}{\sigma}$ where $\mu = \frac{m_c n_c}{2}$ and $\sigma = \sqrt{\frac{m_c n_c (m_c+n_c+1)}{12}}$.

Under the setup of two-part model tests $V^2 = B^2 + T^2$, we use a binomial test for the equality of proportions of zero values (B), and a standard t test (TBT) or Wilcoxon test (TBW) for

the positive continuous responses (T). Because B and T are independent and asymptotically normal, the sum of the squared statistics V^2 asymptotically follows a χ^2_2 distribution[10,11].

§3 A Semiparametric Hypothesis Testing Method

We focus, in the presence of zero-inflated continuous data, on testing the hypothesis concerning the difference of two population means. For the two-part model tests, T is generally the standard t test or the non-parametric Wilcoxon test. In recent years, there have been some reports on the establishment of semiparametric statistical analysis methods under the DRM, which are usually more robust than parametric methods and more effective than non-parametric methods. Therefore we propose here a two-part model test based on binomial test and semiparametric hypothesis test. Specifically, the binomial test is used to test the probability value of the occurrence of zero value, and a hypothesis test is proposed to test the mean difference of positive continous value parts under the semiparametric DRM. Namely, B is the binomial test and T is the semiparametric test. This method is mainly based on the semiparametric estimation of the mean difference of positive value parts of the two populations.

3.1 The density ratio model

Let us first introduce some notation. Let m_0 and m_c denote the (random) numbers of zero values and positive observations for Group 1. Let n_0 and n_c denote the (random) numbers of zero and positive observations for Group 2. Define $l_0 = m_0 + n_0$ and $l_c = m_c + n_c$ as the total numbers of zero and positive values, respectively, and let $l = l_0 + l_c$ denote the total sample size. Without loss of generality, we use the first m_c observations $x_c = \{x_{c1}, x_{c2}, \ldots, x_{cm_c}\}$ to denote the positive observations in the Group 1, and use the first n_c observations $y_c = \{y_{c1}, y_{c2}, \ldots, y_{cn_c}\}$ to denote the positive observations in the Group 2.

Let μ_1 and μ_2 respectively be the mean of nonzero value of the Group 1 and Group 2. In this section, the main goal is to develop a test statistic for the positive values mean hypothesis $H_0{}' : \mu_1 = \mu_2$.

The most commonly used method is two sample t test. It assumes that the two data groups are from the two normal distributions, and assumes that the overall variance of each group is equal or unequal. However, in real applications we often encounter data that do not follow normal distributions, and thus the traditional two sample t test may not be efficient. Denote $\hat{\mu}_1$ and $\hat{\mu}_2$ as the sample mean of x_c and y_c, and denote $S^2_{x_c}$ and $S^2_{y_c}$ as the sample variance of x_c and y_c. Then a nonparametric test statistic is

$$\hat{\Lambda} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{S^2_{x_c}}{m_c} + \frac{S^2_{y_c}}{n_c}}},$$

where the asymptotic normal distribution of $\hat{\Lambda}$ can be used to test the hypothesis.

Next we develop a semiparametric testing method based on the DRM for the null hypothesis $H_0{}' : \mu_1 = \mu_2$.

Let G and F respectively be the distribution functions of x_c and y_c, and denote $g(x)$ and $f(x)$ as the corresponding density functions. We propose to model the distributions of the

nonzero values by the DRM to exploit information from all available samples. The DRM postulates that

$$f(x) = \exp\left\{\alpha + \boldsymbol{\beta}^T \boldsymbol{\gamma}(x)\right\} g(x) \tag{1}$$

for a non-trivial, pre-specified, basis function $\boldsymbol{\gamma}(x)$ of dimension p, and unknown parameters α and $\boldsymbol{\beta}$. Without specifying the baseline density $g(x)$, we propose a test based on the DRM defined in (1) that does not depend on the form of $g(x)$ and hence is robust to the assumptions on $g(x)$. As for how to choose function $\boldsymbol{\gamma}(x)$, Fokianos[6] and Kay and Little[9] provided a good reference.

3.2 The main method

Let $\{T_1, \ldots, T_{l_c}\}$ denote the combined sample data $\{x_{c1}, x_{c2}, \ldots, x_{cm_c}; y_{c1}, y_{c2}, \ldots, y_{cn_c}\}$, and denote $l_c = m_c + n_c$ the total nonzero sample sizes. Under the DRM (1), the likelihood function is written as

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}, G) &= \prod_{i=1}^{m_c} dG(x_{ci}) \prod_{j=1}^{n_c} \exp\left\{\alpha + \boldsymbol{\beta}^T \boldsymbol{\gamma}(y_{cj})\right\} dG(y_{cj}) \\ &= \prod_{i=1}^{l_c} \pi_i \prod_{j=1}^{n_c} \exp\left\{\alpha + \boldsymbol{\beta}^T \boldsymbol{\gamma}(y_{cj})\right\}, \end{aligned} \tag{2}$$

where $\pi_i = dG(T_i)$ is a jump of probability, and the sum is 1.

We have the following natural constrains:

$$\sum_{i=1}^{l_c} \pi_i = 1, \pi_i \geq 0, \sum_{i=1}^{l_c} \pi_i \left\{\exp\left(\alpha + \boldsymbol{\beta}^T \boldsymbol{\gamma}(T_i)\right) - 1\right\} = 0.$$

Following similar profiling procedures, using the Lagrangian multipliers as in Qin and Zhang[15], the maximum likelihood function (2) of $(\alpha, \boldsymbol{\beta})$ is gained at the point

$$\tilde{\pi}_i = \frac{1}{m_c} \frac{1}{1 + \rho \exp\left(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{\gamma}(T_i)\right)},$$

where $\rho = n_c/m_c$, $\left(\tilde{\alpha}, \tilde{\boldsymbol{\beta}}\right)$ is the maximum semiparametric likelihood estimator of $(\alpha, \boldsymbol{\beta})$, with $\left(\tilde{\alpha}, \tilde{\boldsymbol{\beta}}\right)$ solving

$$\frac{\partial l(\alpha, \boldsymbol{\beta})}{\partial \alpha} = n_c - \sum_{i=1}^{l_c} \frac{\rho \exp\left(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{\gamma}(T_i)\right)}{1 + \rho \exp\left(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{\gamma}(T_i)\right)} = 0,$$

$$\frac{\partial l(\alpha, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{n_c} \boldsymbol{\gamma}(y_{cj}) - \sum_{i=1}^{l_c} \frac{\rho \exp\left(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{\gamma}(T_i)\right)}{1 + \rho \exp\left(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{\gamma}(T_i)\right)} \boldsymbol{\gamma}(T_i) = 0.$$

Here $l(\alpha, \boldsymbol{\beta})$ is the empirical log-likelihood function of (2) associated with the positive obser-

vation

$$l(\alpha,\boldsymbol{\beta}) = \sum_{j=1}^{n_c}\left[\alpha+\boldsymbol{\beta}^T\boldsymbol{\gamma}(y_{cj})\right] - l_c\log m_c - \sum_{i=1}^{l_c}\log\left[1+\rho\exp\left(\alpha+\boldsymbol{\beta}^T\boldsymbol{\gamma}(T_i)\right)\right]. \tag{3}$$

Under the DRM (1), we get the maximum semiparameter likelihood estimator of the $G(t)$ and $F(t)$

$$\begin{aligned}\tilde{G}(t) &= \sum_{i=1}^{l_c}\pi_i I(T_i\le t)\\ &= \frac{1}{m_c}\sum_{i=1}^{l_c}\frac{1}{1+\rho\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(T_i)\right)}I(T_i\le t),\\ \tilde{F}(t) &= \sum_{i=1}^{l_c}\pi_i\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(T_i)\right)I(T_i\le t)\\ &= \frac{1}{m_c}\sum_{i=1}^{l_c}\frac{\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(T_i)\right)}{1+\rho\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(T_i)\right)}I(T_i\le t).\end{aligned}$$

Then we can obtain the maximum likelihood estimator of the mean difference of the continuous part of two populations $\mu_1-\mu_2=\int tdG-\int tdF$ by substituting the estimator of $G(t)$ and $F(t)$

$$\tilde{\mu}_1-\tilde{\mu}_2 = \frac{1}{m_c}\sum_{i=1}^{l_c}\frac{1-\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(T_i)\right)}{1+\rho\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(T_i)\right)}T_i. \tag{4}$$

In what follows, we study the asymptotic distribution of the semiparametric estimators (4). Suppose that the true value of $(\alpha,\boldsymbol{\beta})$ is $(\alpha_0,\boldsymbol{\beta}_0)$, and the asymptotic result relies on the condition that $\rho=n_c/m_c$ remains the same when $l_c=m_c+n_c\to\infty$. In addition, the following notations are introduced for the convenience of the presentation

$$\omega(t)=\exp\left(\alpha_0+\boldsymbol{\beta}_0^T\boldsymbol{\gamma}(t)\right), A_0=\int_{-\infty}^{\infty}\frac{\omega(y)}{1+\rho\omega(y)}dG(y),$$

$$\boldsymbol{A}_1=\int_{-\infty}^{\infty}\frac{\omega(y)}{1+\rho\omega(y)}\boldsymbol{\gamma}(y)\,dG(y),$$

$$\boldsymbol{A}_2=\int_{-\infty}^{\infty}\frac{\omega(y)}{1+\rho\omega(y)}\boldsymbol{\gamma}(y)\{\boldsymbol{\gamma}(y)\}^T dG(y),$$

$$\boldsymbol{A}=\begin{pmatrix}A_0 & \boldsymbol{A}_1^T\\ \boldsymbol{A}_1 & \boldsymbol{A}_2\end{pmatrix}, \boldsymbol{S}=\frac{\rho}{1+\rho}\boldsymbol{A}, B_0=\int_{-\infty}^{\infty}\frac{\omega(y)\,y}{1+\rho\omega(y)}dG(y),$$

$$\boldsymbol{B}_1=\int_{-\infty}^{\infty}\frac{\omega(y)\,y}{1+\rho\omega(y)}\boldsymbol{\gamma}(y)\,dG(y), B_2=\int_{-\infty}^{\infty}\frac{\omega(y)\,y^2}{1+\rho\omega(y)}dG(y).$$

The following theorem establishes the asymptotic normality of $\tilde{\mu}_1-\tilde{\mu}_2$. A proof of the theorem is given in the appendix A.

Theorem 3.1. *If DRM (1) holds and* $\boldsymbol{A}^{-1}$ *exists, then we have*

$$\sqrt{l_c}\left(\tilde{\mu}_1-\tilde{\mu}_2-(\mu_1-\mu_2)\right)\to N\left(0,\sigma_{semi}^2\right),$$

where

$$\begin{aligned}\sigma_{semi}^2=&(1+\rho)\left\{\int_{-\infty}^{\infty}u^2dG(u)-\left[\int_{-\infty}^{\infty}udG(u)\right]^2\right\}\\&+\frac{1+\rho}{\rho}\left\{\int_{-\infty}^{\infty}u^2dF(u)-\left[\int_{-\infty}^{\infty}udF(u)\right]^2\right\}\\&-\frac{(1+\rho)^3}{\rho}\left\{B_2-\left(B_0,\boldsymbol{B}_1^T\right)\boldsymbol{A}^{-1}\left(\begin{array}{c}B_0\\\boldsymbol{B}_1\end{array}\right)\right\}.\end{aligned}$$

The consistent estimates of σ_{semi}^2 and $\tilde{\sigma}_{semi}^2$ can be constructed by substituting $\left(\tilde{\alpha},\tilde{\boldsymbol{\beta}}\right)$ in the maximum semiparameter likelihood estimator. Then $\tilde{\sigma}_{semi}^2$ can be written as

$$\begin{aligned}\tilde{\sigma}_{semi}^2=&(1+\rho)\left\{\sum_{i=1}^{l_c}{T_i}^2\tilde{\pi}_i-\left[\sum_{i=1}^{l_c}T_i\tilde{\pi}_i\right]^2\right\}+\frac{1+\rho}{\rho}\left\{\sum_{i=1}^{l_c}{T_i}^2\tilde{\omega}(T_i)\tilde{\pi}_i-\left[\sum_{i=1}^{l_c}T_i\tilde{\omega}(T_i)\tilde{\pi}_i\right]^2\right\}\\&-\frac{(1+\rho)^3}{\rho}\left\{\tilde{B}_2-\left(\tilde{B}_0,\tilde{\boldsymbol{B}}_1^T\right)\tilde{\boldsymbol{A}}^{-1}\left(\begin{array}{c}\tilde{B}_0\\\tilde{\boldsymbol{B}}_1\end{array}\right)\right\},\end{aligned}$$

where

$$\tilde{\omega}(t)=\exp\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{\gamma}(t)\right),\tilde{A}_0=\sum_{i=1}^{l_c}\frac{\tilde{\omega}(T_i)}{1+\rho\tilde{\omega}(T_i)}\tilde{\pi}_i,$$

$$\tilde{\boldsymbol{A}}_1=\sum_{i=1}^{l_c}\frac{\tilde{\omega}(T_i)}{1+\rho\tilde{\omega}(T_i)}\boldsymbol{\gamma}(T_i)\tilde{\pi}_i,\tilde{\boldsymbol{A}}_2=\sum_{i=1}^{l_c}\frac{\tilde{\omega}(T_i)}{1+\rho\tilde{\omega}(T_i)}\boldsymbol{\gamma}(T_i)\{\boldsymbol{\gamma}(T_i)\}^T\tilde{\pi}_i,$$

$$\tilde{\boldsymbol{A}}=\left(\begin{array}{cc}\tilde{A}_0&\tilde{\boldsymbol{A}}_1^T\\\tilde{\boldsymbol{A}}_1&\tilde{\boldsymbol{A}}_2\end{array}\right),\tilde{B}_0=\sum_{i=1}^{l_c}\frac{\tilde{\omega}(T_i)T_i}{1+\rho\tilde{\omega}(T_i)}\tilde{\pi}_i,$$

$$\tilde{\boldsymbol{B}}_1=\sum_{i=1}^{l_c}\frac{\tilde{\omega}(T_i)T_i}{1+\rho\tilde{\omega}(T_i)}\boldsymbol{\gamma}(T_i)\tilde{\pi}_i,\tilde{B}_2=\sum_{i=1}^{l_c}\frac{\tilde{\omega}(T_i){T_i}^2}{1+\rho\tilde{\omega}(T_i)}\tilde{\pi}_i.$$

Under the condition DRM (1), the semiparametric Wald test for testing null hypothesis $H_0{}':\mu_1=\mu_2$ is defined as

$$\tilde{\Lambda}=\frac{\sqrt{l_c}\left(\tilde{\mu}_1-\tilde{\mu}_2\right)}{\tilde{\sigma}_{semi}^2}\sim N(0,1).\tag{5}$$

Clearly, we have $\tilde{\Lambda}^2\to\chi_1^2$ as $n\to\infty$, where χ_1^2 denotes a chi-squared random variable with 1 degree of freedom.

Suppose we have two groups of samples from zero-inflated continuous data and condition (1) is satisfied for the positive value observations. Under the null hypothesis $H_0:(p_1=p_2)\cap(\mu_1=\mu_2)$, we present a new two-part semiparametric test statistic $V^2=B^2+\tilde{\Lambda}^2$, referred to as TBSE, where B is the usual binomial test for equality of zero value proportions and $\tilde{\Lambda}$ is the proposed semiparametric test under a semiparametric DRM for the positive continuous responses. Like TBT and TBW tests, we can obviously see that $V^2=B^2+\tilde{\Lambda}^2\to\chi_2^2$, where χ_2^2 denotes a chi-squared random variable with 2 degree of freedom.

Table 1: Parameter settings for simulation studies. In the first column, each $LN_1 - LN_8$ and each $GAM_1 - GAM_8$ and each $EXP_1 - EXP_7$ denote mixture models whose nonnegative part follows $LN(a_i, b_i)$ and $GAM(a_i, b_i)$ and $EXP(a_i)$, respectively, for $i = 1, 2$. The last two columns are the means and the variances corresponding to the nonnegative part of each model.

model	(p_1, p_2)	(a_1, a_2)	(b_1, b_2)	(μ_1, μ_2)	(σ_1^2, σ_2^2)
LN_1	(0.2,0.2)	(0.0,0.0)	(1.0,1.0)	(1.65,1.65)	(4.67,4.67)
LN_2	(0.5,0.5)	(0.0,0.0)	(1.0,1.0)	(1.65,1.65)	(4.67,4.67)
LN_3	(0.2,0.2)	(0.0,0.25)	(1.0,0.5)	(1.65,1.65)	(4.67,1.76)
LN_4	(0.5,0.5)	(0.0,0.25)	(1.0,0.5)	(1.65,1.65)	(4.67,1.76)
LN_5	(0.4,0.3)	(0.0,0.0)	(1.0,1.0)	(1.65,1.65)	(4.67,4.67)
LN_6	(0.4,0.3)	(0.0,0.25)	(1.0,0.5)	(1.65,1.65)	(4.67,1.76)
LN_7	(0.4,0.4)	(0.0,0.5)	(1.0,2.25)	(1.65,5.08)	(4.67,218.9)
LN_8	(0.4,0.3)	(0.0,0.5)	(1.0,2.25)	(1.65,5.08)	(4.67,218.9)
GAM_1	(0.2,0.2)	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)
GAM_2	(0.5,0.5)	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)
GAM_3	(0.2,0.2)	(1.0,2.0)	(1.0,0.5)	(1.0,1.0)	(1.0,0.5)
GAM_4	(0.5,0.5)	(1.0,2.0)	(1.0,0.5)	(1.0,1.0)	(1.0,0.5)
GAM_5	(0.4,0.3)	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)
GAM_6	(0.4,0.3)	(1.0,2.0)	(1.0,0.5)	(1.0,1.0)	(1.0,0.5)
GAM_7	(0.4,0.4)	(1.0,1.0)	(1.0,0.5)	(1.0,0.5)	(1.0,0.25)
GAM_8	(0.4,0.3)	(1.0,1.0)	(1.0,0.5)	(1.0,0.5)	(1.0,0.25)
EXP_1	(0.2,0.2)	(1.0,1.0)	–	(1.0,1.0)	(1.0,1.0)
EXP_2	(0.5,0.5)	(1.0,1.0)	–	(1.0,1.0)	(1.0,1.0)
EXP_3	(0.2,0.2)	(0.5,0.5)	–	(2.0,2.0)	(4.0,4.0)
EXP_4	(0.5,0.5)	(0.5,0.5)	–	(2.0,2.0)	(4.0,4.0)
EXP_5	(0.4,0.3)	(1.0,1.0)	–	(1.0,1.0)	(1.0,1.0)
EXP_6	(0.4,0.4)	(1.0,0.5)	–	(1.0,2.0)	(1.0,4.0)
EXP_7	(0.4,0.3)	(1.0,0.5)	–	(1.0,2.0)	(1.0,4.0)

§4 Simulation Studies

In this section, we assess the finite-sample performance of the proposed TBSE test through simulation. To verify the proposed test method, we further compare the proposed method with two existing methods:

• the two-part t test (TBT) $V^2 = B^2 + T^2$ where B is the usual binomial test and T is the t-test;

• the two-part Wilcoxon test (TBW) $V^2 = B^2 + T^2$ where B is the usual binomial test and T is the Wilcoxon test.

We generate two-sample random observations, given that p_i'$s \neq 0$ or 1, from (1) with $f(x)$ and $g(x)$ being log-normal, gamma, or all exponential distribution. In the following, we use $LN(a_i, b_i)$ to denote a log-normal distribution with mean a_i and variance b_i (i.e., the mean and variance of the associated normal random variable) and $GAM(a_i, b_i)$ to denote a gamma distribution with shape parameter a_i and scale parameter b_i, and $EXP(a_i)$ to denote an exponential distribution with rate a_i. The parameters set-up under the null ($LN_1 - LN_4$, $GAM_1 - GAM_4$ and $EXP_1 - EXP_4$) and alternative ($LN_5 - LN_8$, $GAM_5 - GAM_8$ and $EXP_5 - EXP_7$) models are given in Table 1. We consider the case with equal sample sizes by setting n_i to be 25, 50 or 100 for $i = 1, 2$; and also consider the case with unequal sample sizes that $(n_1, n_2) = (50, 100)$.

Table 2: Type I error rates (%) for testing H_0 at significance level 0.05 when data are generated from a log-normal mixture model with parameter settings giving in Table 1. The TBSE test is under $\boldsymbol{\gamma}(x) = \{log(x), log^2(x)\}^T$.

model	(m, n)	TBT	TBW	TBSE
LN_1	(25,25)	2.72	2.77	2.77
	(50,50)	3.11	3.17	3.02
	(50,100)	4.20	3.78	3.74
	(100,100)	3.64	4.04	3.46
LN_2	(25,25)	3.37	2.52	3.62
	(50,50)	3.54	3.53	3.34
	(50,100)	4.17	3.51	3.79
	(100,100)	3.74	3.93	3.65
LN_3	(25,25)	4.26	8.04	4.32
	(50,50)	4.17	15.57	4.21
	(50,100)	6.23	22.27	5.65
	(100,100)	4.66	31.30	4.81
LN_4	(25,25)	5.53	5.91	5.06
	(50,50)	4.95	10.75	4.97
	(50,100)	6.94	14.46	6.16
	(100,100)	5.51	20.68	5.48

Table 3: Type I error rates (%) for testing H_0 at significance level 0.05 when data are generated from a gamma mixture model with parameter settings giving in Table 1. The TBSE test is under $\boldsymbol{\gamma}(x) = \{x, log(x)\}^T$.

model	(m, n)	TBT	TBW	TBSE
GAM_1	(25,25)	3.11	2.69	3.49
	(50,50)	3.50	3.20	3.65
	(50,100)	4.14	3.75	4.25
	(100,100)	3.65	3.59	3.75
GAM_2	(25,25)	4.08	2.96	4.75
	(50,50)	3.80	3.66	4.16
	(50,100)	4.28	3.52	4.43
	(100,100)	4.42	4.22	4.56
GAM_3	(25,25)	3.84	5.30	4.18
	(50,50)	3.99	8.98	4.11
	(50,100)	4.95	12.61	5.00
	(100,100)	4.10	14.92	4.23
GAM_4	(25,25)	4.56	4.54	5.22
	(50,50)	4.53	6.77	4.92
	(50,100)	5.55	9.79	5.69
	(100,100)	4.64	11.62	4.88

For all tests, the type I error rates and power at the 5% significance level are calculated based on 10000 repetitions. In the simulation, we only present the results of TBSE test under the correctly specified basis function $\boldsymbol{\gamma}(x)$. That is, TBSE test under $\boldsymbol{\gamma}(x) = \{log(x), log^2(x)\}^T$ for the LN models; and under $\boldsymbol{\gamma}(x) = \{x, log(x)\}^T$ for the GAM models, and under $\boldsymbol{\gamma}(x) = x$ for the EXP models. For more choices of $\boldsymbol{\gamma}(x)$, interested readers may refer to Fokianos[6] and Kay and Little[9].

Table 4: Type I error rates (%) for testing H_0 at significance level 0.05 when data are generated from an exponential mixture model with parameter settings giving in Table 1. The TBSE test is under $\boldsymbol{\gamma}(x) = \{x\}$.

model	(m, n)	TBT	TBW	TBSE
EXP_1	(25,25)	3.25	3.02	3.67
	(50,50)	3.62	3.46	3.77
	(50,100)	4.05	3.52	4.11
	(100,100)	3.76	3.80	3.82
EXP_2	(25,25)	4.14	3.29	4.65
	(50,50)	3.95	3.49	4.31
	(50,100)	4.42	3.82	4.51
	(100,100)	4.19	4.05	4.42
EXP_3	(25,25)	3.35	2.91	3.71
	(50,50)	3.61	3.33	3.75
	(50,100)	4.18	3.76	4.16
	(100,100)	3.87	3.84	3.92
EXP_4	(25,25)	4.28	3.11	4.91
	(50,50)	4.20	3.64	4.52
	(50,100)	4.55	4.10	4.58
	(100,100)	4.19	3.78	4.40

4.1 Type I error

The simulated type I error rates for the three selected representative tests are summarized in Tables 2-4. A two-part test is a two-degree of freedom test based on a test statistic for the equality of the proportions of zero counts and a conditional χ^2 test statistic for the positive part. From the results of Tables 2-4, we can see that in general both TBSE and TBT are able to control the type I error rates at the designed level of 5%. In the presence of heterogeneity (unequal variances, corresponding to LN_3, LN_4, GAM_3 and GAM_4), however, the two tests may have slightly inflated type I error rates.

The same cannot be said about the TBW test. When the two samples have equal variances, the TBW test performs well in controlling the type I error rates. But when the two samples have unequal variances, the TBW test may possess much higher type error rates, mostly unacceptable. Hence, when heterogeneity exits between the two samples, it is inappropriate and often misleading to compare the powers of TBW with that of TBT or TBSE due to the largely inflated type I error rates. Comparison of powers of tests is appropriate only if the tests have the same control of type I error rates.

Table 5: Scenarios categorized according to the alternative model settings in Table 1.

Scenarios I	Scenarios II	Scenarios III
LN_5, LN_6	LN_7	LN_8
GAM_5, GAM_6	GAM_7	GAM_8
EXP_5	EXP_6	EXP_7

Table 6: Simulated testing powers (%) of rejecting H_0 at significance level 0.05 when data are generated from a log-normal mixture model with parameter setting in Table 1.

model	(m, n)	TBT	TBW	TBSE
LN_5	(25,25)	6.11	5.76	6.20
	(50,50)	10.69	10.68	10.39
	(50,100)	15.47	14.43	15.17
	(100,100)	20.73	20.53	20.41
LN_6	(25,25)	8.17	10.67	8.43
	(50,50)	12.13	22.14	12.26
	(50,100)	17.61	30.89	17.17
	(100,100)	21.26	43.63	21.29
LN_7	(25,25)	11.54	10.86	12.79
	(50,50)	28.67	21.94	30.02
	(50,100)	60.01	26.46	60.35
	(100,100)	62.76	41.97	64.80
LN_8	(25,25)	18.82	13.96	19.68
	(50,50)	44.48	31.01	45.78
	(50,100)	73.73	41.40	74.52
	(100,100)	81.68	61.82	83.00

Table 7: Simulated testing powers (%) of rejecting H_0 at significance level 0.05 when data are generated from a gamma mixture model with parameter setting in Table 1.

model	(m, n)	TBT	TBW	TBSE
GAM_5	(25,25)	6.31	5.73	6.83
	(50,50)	10.41	9.74	10.66
	(50,100)	14.92	14.25	15.14
	(100,100)	20.67	20.55	20.87
GAM_6	(25,25)	7.00	7.83	7.51
	(50,50)	11.38	16.22	11.66
	(50,100)	15.65	22.34	15.86
	(100,100)	21.21	32.43	21.36
GAM_7	(25,25)	28.45	21.81	31.29
	(50,50)	59.17	47.44	61.07
	(50,100)	65.14	61.49	67.59
	(100,100)	91.27	81.53	91.69
GAM_8	(25,25)	32.82	27.55	35.87
	(50,50)	68.73	58.83	70.09
	(50,100)	76.85	73.36	78.56
	(100,100)	96.10	90.69	96.21

4.2 Testing power

Powers of the three tests were simulated under the selected alternatives LN_5-LN_8, GAM_5-GAM_8, and EXP_5-EXP_7. For clarity and to aid the discussion, we categorize the alternative model specifications into a 2×3 table given in Table 5. For the columns, the specifications can be divided into scenarios in which: the means of the positive components are held constant (Scenario I), the zero value proportions are held constant (Scenario II), or the zero values proportions and the means of the positive components are all different (Scenario III). Especially, Scenario I could be divided into two settings: the variances of the positive components are equal (LN_5, GAM_5 and EXP_5), the variances of the positive components are unequal (LN_6

Table 8: Simulated testing powers (%) of rejecting H_0 at significance level 0.05 when data are generated from an exponential mixture model with parameter setting in Table 1.

model	(m,n)	TBT	TBW	TBSE
EXP_5	(25,25)	6.30	5.66	6.84
	(50,50)	10.65	10.24	10.94
	(50,100)	14.82	13.91	14.89
	(100,100)	20.63	20.77	20.73
EXP_6	(25,25)	27.81	21.00	30.72
	(50,50)	58.45	47.13	60.29
	(50,100)	81.19	61.71	81.39
	(100,100)	91.38	81.49	91.79
EXP_7	(25,25)	37.04	27.88	39.33
	(50,50)	71.04	59.27	72.20
	(50,100)	88.70	74.37	89.03
	(100,100)	95.35	90.43	96.44

and GAM_6).

The simulated testing powers for the three tests at the 5% significance level are summarized in Tables 6-8. Together with the simulation results, we make the following comments and discussion.

1. (Homogeneity) When the two samples have equal variances (LN_5, GAM_5, EXP_5-EXP_7), the three tests have comparable type I error rates, and hence their powers can be compared. For the lognormal (LN_5) and gamma (GAM_5) models, the powers of the three tests are very close to each other and hence the three tests are almost equally efficient. For the exponential model (EXP_5-EXP_7), the proposed test TBSE performs slightly better than the TBT; both tests have substantially higher power than the TBW test.

2. (Heterogeneity) When the two samples have unequal variances (LN_6-LN_8, GAM_6-GAM_8), the TBW test, as mentioned earlier, has much inflated type I error rates and does not warrant a comparison with the other two tests (although for convenience the powers of TBW are also provided in the tables). Furthermore, the proposed test TBSE appears to outperform the TBT test in all cases.

In general, we have observed that the two-part semiparametric TBSE test is robust on the assumption of the positive components distribution, and the rejection rates are very close to the nominal level when sample sizes n is relatively large. In addition, the TBSE test has about the same of higher power than the other tests based on the correctly specified basis functions under the premise that the type I error rates are controlled. Hence in practice we recommend the proposed TBSE to compare two independent zero-inflated continuous samples.

§5 Application: The CHEF trial data

In the section, we further illustrate the proposed two-part semiparametric test (TBSE) with the CHEF study. The study is an 18-month randomized trial to evaluate the efficacy of a family-based behavioral intervention that integrated motivational interviewing, active learning, and applied problem-solving to increase intake of whole plant foods (fruits, vegetables, whole

grains, legumes, nuts, and seeds) among youth with type 1 diabetes[14]. At the CHEF study, a total of 136 children were enrolled into the study with 66 randomized to the intervention group and the remaining to the control group. Families in the intervention condition received sessions on healthy eating, with a focus on increasing intake of whole plant foods. Sessions subsequently applied intervention content to each meal time and other eating contexts. Participants in the control condition received no additional dietary advice beyond that provided as part of the standard type 1 diabetes care. Dietary data were collected at 6 time points, including a baseline, during the 18-month study duration based on 3-day diet records. Details of the study design, randomization procedures and treatment conditions can be found in [14] .

Table 9: The percents of zero values, mean and variance of non-zeros, and the p-values of Shapiro-Wilk normality test for each variable of each cluster.

Variable	$pr(X_{ij}=0)$	$E(X_{ij}>0)$	$Var(X_{ij}>0)$	p-values
Control				
TF	0.19	0.47	0.14	4.00×10^{-4}
WF	0.25	0.38	0.15	1.25×10^{-5}
DOL	0.28	0.18	0.05	1.44×10^{-8}
WG	0.04	1.13	0.68	7.00×10^{-4}
Intervention				
TF	0.16	0.39	0.08	1.00×10^{-3}
WF	0.34	0.27	0.03	0.02
DOL	0.26	0.15	0.04	5.17×10^{-9}
WG	0.07	0.85	0.49	1.82×10^{-5}

Dietary records were collected at baseline prior to the start of intervention from the study participants. Among the twelve food variables, eight are daily consumed foods whose intakes are continuous. The remaining four, total fruit (TF), whole fruit (WF), Dark Green/Orange Vegetables & Legumes (DOL), and whole grain (WG), are characterized by excess zero values due to episodical consumption of the foods. At 18-month follow-up, our aim is to investigate whether the efficacy of a family-based behavioral intervention to the four semicontinuous variables between control group and intervention group. In addition to the continuous positive measurements, there are substantial proportions of zero values. Some summary statistics are showed in Table 9.

Table 10: AIC for the positive value of each variable for five commonly used basis functions $\boldsymbol{\gamma}(x)$.

$\boldsymbol{\gamma}(x)$	x	$log(x)$	$\{x, log(x)\}^T$	$\{log(x), log^2(x)\}^T$	$\{x, log(x), log^2(x)\}^T$
TF	**138.96**	139.00	140.84	140.87	139.42
WF	118.30	120.78	**118.17**	121.27	118.28
DOL	123.89	123.29	122.87	**122.80**	124.79
WG	159.17	159.48	159.11	**158.98**	160.94

Since the distribution of control group and intervention group data are unknown, the Shapiro-Wilk normality test is performed on the positive values of each group. According to the p-values in table 9, the positive values of control group and intervention group do not follow normal distribution, so t test becomes inefficient and is not recommended to be used for the positive measurements. As discussed in Section 4, we use the proposed TBSE test to

Table 11: TBSE test statistics and corresponding p-values of each variable under basis function $\boldsymbol{\gamma}(x)$.

Variable	$\boldsymbol{\gamma}(x)$	Test statistics	p-value
TF	x	4.85	0.0887
WF	$\{x, log(x)\}^T$	8.91	0.0116
DOL	$\{log(x), log^2(x)\}^T$	2.83	0.2426
WG	$\{log(x), log^2(x)\}^T$	1.33	0.5153

discover if the mean measurements differ between the control and intervention group. Therefore we need to select a basis function $\boldsymbol{\gamma}(x)$ in a DRM that provides a reasonable fit to the four semicontinuous variables, respectively. We apply the AIC to select a basis function in the DRM for the positive data in this example[6]. The results are given in the Table 10. It can be seen that the DRM with $\boldsymbol{\gamma}(x) = x$ for TF variable, with $\boldsymbol{\gamma}(x) = \{x, log(x)\}^T$ for WF variable, with $\boldsymbol{\gamma}(x) = \{log(x), log^2(x)\}^T$ for DOL and WG variable provide the best fit for the data; these functions respectively have the smallest AIC among the five commonly used basis functions, and hence they are recommended in this example.

The results of Shapiro-Wilk normality test performed on the four variables indicate that the TBT test cannot be used to the example, and the TBW test exceeded the 0.05 level when the homoscedastic variances assumption are difficult to justify for multiple groups of the positive components as discussed in Section 4. Subsequently we applied the proposed TBSE test for the efficacy of the family-based behavioral intervention to the four semicontinuous variables of the intervention group in comparison to the control group. The observed test statistics and their corresponding p-values are reported at 5% significance level in Table 11.

From the results in Table 11, we see that there is a significant difference for the WF variable between control group and intervention group. But there does not seem to be a significant difference for the remaining variables between control group and intervention group. Therefore, the family-based behavioral intervention demonstrated efficacy for the WF variable, but the intervention may be noneffective to the TF, DOL and WG variables. A natural follow-up concern is to detect an improvement family-based behavioral intervention to the CHEF study.

§6 Concluding remarks

In this paper, we discussed the problem of making statistical inferences on the means of two group samples with excess zero observations. Under the semiparametrics framework developed by Fokianos[6], we proposed a TBSE statistic and derived its limiting distribution based on the two-part tests. Simulation studies showed that the proposed TBSE test has desired type I error control and is powerful to detect departures from the null hypothesis. Also the TBSE test is made computationally fast by using logistic regression routines available in standard statistical softwares. In addition, we use the TBSE test to a real dietary data for testing the effectiveness of a family-based behavioral intervention on increasing intake of diabetes-friendly foods, and the results illustrate the advantages of the proposed method.

As an important area of application, it is interesting to further consider the model selection

problem for DRM, and to employ TBSE test to deal with zero-inflated count data. We first adapt the idea of the hurdle model for zero-inflated count data, which models the zero and positive counts separately[3,12]. That is, $g(x)$ is the probability mass function for the positive counts in DRM (1). Note that the commonly used zero-truncated Poisson and zero-truncated negative binomial distributions both satisfy the DRM condition (1). As discussed in Bedrick et al.[3], testing homogeneity under the mixture structures of the zero-inflated Poisson and the Poisson-hurdle model is equivalent. A similar conclusion also applies to the negative binomial distribution. Hence, testing the homogeneity in the zero-inflated count data under our setup is equivalent to testing the null hypothesis $H_0 : (p_1 = p_2) \cap (\mu_1 = \mu_2)$. In the case, the TBSE test developed in Section 3 may be directly applied.

The DRM is a useful semiparametric tool for the comparison of independent samples. However, its application relies heavily on the basis function $\boldsymbol{\gamma}(x)$. In practice proper transformations such as logarithm are often applied to the original scale to better approximate the model. Kay and Little[9] discussed the forms of $\boldsymbol{\gamma}(x)$ applicable to common probability distributions. For example, the basis function $\boldsymbol{\gamma}(x)$ can be set to $\{\log(x), \log^2(x)\}^T$ for log-normal distributions, $\{x, \log(x)\}^T$ for gamma distributions, and x for exponential distributions. However, as Fokianos[6] pointed out, misspecification of the basis function could lead to biased estimators and loss of efficiency. In practice, we suggest consider a few competing choices of basis function and use Akaike's information criterion (AIC) to select the best one; see Fokianos[6].

Appendix A. *Proof of Theorem 3.1.*

Based on the Taylor expansion to $\tilde{\mu}_1 - \tilde{\mu}_2$, and combining with the consistency of $\tilde{\alpha}$ and $\tilde{\boldsymbol{\beta}}$ in Qin and Zhang[15], we have

$$\begin{aligned}\tilde{\mu}_1 - \tilde{\mu}_2 =& \frac{1}{m_c}\sum_{i=1}^{l_c}\frac{1-\omega(T_i)}{1+\rho\omega(T_i)}T_i - \frac{1+\rho}{m_c}\left(\sum_{i=1}^{l_c}\frac{\omega(T_i)T_i}{(1+\rho\omega(T_i))^2}, \sum_{i=1}^{l_c}\frac{\omega(T_i)\boldsymbol{\gamma}^T(T_i)T_i}{(1+\rho\omega(T_i))^2}\right)\begin{pmatrix}\tilde{\alpha}-\alpha_0\\ \tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\end{pmatrix}\\ &+ O_p\left({l_c}^{-1}\right).\end{aligned}$$

By direct operations, we can get

$$E\left\{\frac{1}{m_c}\sum_{i=1}^{l_c}\frac{\omega(T_i)T_i}{(1+\rho\omega(T_i))^2}\right\} = B_0,$$

$$E\left\{\frac{1}{m_c}\sum_{i=1}^{l_c}\frac{\omega(T_i)\boldsymbol{\gamma}^T(T_i)T_i}{(1+\rho\omega(T_i))^2}\right\} = \boldsymbol{B}_1^T.$$

Therefore, from the law of large numbers and the asymptotic expression $\begin{pmatrix}\tilde{\alpha}-\alpha_0\\ \tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\end{pmatrix} = \frac{1}{l_c}\boldsymbol{S}^{-1}\begin{pmatrix}\frac{\partial l(\alpha_0,\beta_0)}{\partial\alpha}\\ \frac{\partial l(\alpha_0,\beta_0)}{\partial\boldsymbol{\beta}}\end{pmatrix} + o_p({l_c}^{-1/2})$ discussed in Qin and Zhang[15], we have

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \frac{1}{m_c}\sum_{i=1}^{l_c}\frac{1-\omega(T_i)}{1+\rho\omega(T_i)}T_i - \frac{1}{m_c}\left(B_0, \boldsymbol{B}_1^T\right)\boldsymbol{S}^{-1}\begin{pmatrix}\frac{\partial l(\alpha_0,\beta_0)}{\partial\alpha}\\ \frac{\partial l(\alpha_0,\beta_0)}{\partial\boldsymbol{\beta}}\end{pmatrix} + o_p({l_c}^{-1/2}).$$

We obtained $E\left(\tilde{\mu}_1 - \tilde{\mu}_2\right) = \mu_1 - \mu_2$ by simple calculation, which shows that $\tilde{\mu}_1 - \tilde{\mu}_2$ is asymptotically unbiased. To derive the variance of $\sqrt{l_c}\left(\tilde{\mu}_1 - \tilde{\mu}_2\right)$, we denote $\Delta = \frac{1}{m_c}\sum_{i=1}^{l_c}\frac{1-\omega(T_i)}{1+\rho\omega(T_i)}T_i$

and split the variance into three parts,

$$\begin{aligned} Var\left(\sqrt{l_c}\left(\tilde{\mu}_1-\tilde{\mu}_2\right)\right) =& Var\left(\sqrt{l_c}\Delta\right)-2\left(1+\rho\right)\left(B_0,\boldsymbol{B}_1^T\right)\boldsymbol{S}^{-1}\left(\begin{array}{c} Cov\left(\Delta,\frac{\partial l(\alpha_0,\beta_0)}{\partial\alpha}\right) \\ Cov\left(\Delta,\frac{\partial l(\alpha_0,\beta_0)}{\partial\boldsymbol{\beta}}\right)\end{array}\right) \\ &+\left\{\frac{1+\rho}{m_c}\left(B_0,\boldsymbol{B}_1^T\right)\boldsymbol{S}^{-1}Var\left(\begin{array}{c}\frac{\partial l(\alpha_0,\beta_0)}{\partial\alpha} \\ \frac{\partial l(\alpha_0,\beta_0)}{\partial\boldsymbol{\beta}}\end{array}\right)\boldsymbol{S}^{-1}\left(\begin{array}{c}B_0 \\ \boldsymbol{B}_1\end{array}\right)\right\} \\ =& V_1+V_2+V_3. \end{aligned}$$

And through a little bit of derivation, we get

$$\begin{aligned} V_1 =& \left(1+\rho\right)\left[\int_{-\infty}^{\infty}u^2dG\left(u\right)+\frac{1}{\rho}\int_{-\infty}^{\infty}u^2dF\left(u\right)-\frac{\left(1+\rho\right)^2}{\rho}B_2\right] \\ &-\left(1+\rho\right)\left[\int_{-\infty}^{\infty}udG\left(u\right)-\left(1+\rho\right)B_0\right]^2-\frac{1+\rho}{\rho}\left[\int_{-\infty}^{\infty}udF\left(u\right)-\left(1+\rho\right)B_0\right]^2, \\ V_2 =& -2(1+\rho)^2B_0\int_{-\infty}^{\infty}udG\left(u\right)+\frac{2(1+\rho)^4}{\rho}B_0^2-\frac{2(1+\rho)^2}{\rho}B_0\int_{-\infty}^{\infty}udF\left(u\right), \\ V_3 =& \frac{\left(1+\rho\right)^3}{\rho}\left(B_0,\boldsymbol{B}_1^T\right)\boldsymbol{A}^{-1}\left(\begin{array}{c}B_0 \\ \boldsymbol{B}_1\end{array}\right)-\frac{\left(1+\rho\right)^4}{\rho}B_0^2. \end{aligned}$$

So the variance is

$$Var\left(\sqrt{l_c}\left(\tilde{\mu}_1-\tilde{\mu}_2\right)\right)=V_1+V_2+V_3=\sigma_{semi}^2.$$

It thus follows from the central limit theorem that

$$\sqrt{l_c}\left(\tilde{\mu}_1-\tilde{\mu}_2-\left(\mu_1-\mu_2\right)\right)\rightarrow N\left(0,\sigma_{semi}^2\right).$$

The proof is completed.□

Acknowledgement. The authors thank Dr. Tonja Nansel for helpful discussions on the CHEF study.

References

[1] J A Anderson. *Multivariate logistic compounds*, Biometrika, 1979, 66: 17-26.

[2] C Bascoul-Mollevi, S Gourgou-Bourgade, A Kramar. *Two-part statistics with paired data*, Statistics in Medicine, 2005, 24: 1435-1448.

[3] E J Bedrick, A Hossain. *Conditional tests for homogeneity of zero-inflated Poisson and Poisson-hurdle distributions*, Computational Statistics and Data Analysis, 2013, 61: 99-106.

[4] S Cai, J Chen, J V Zidek. *Hypothesis test in the presence of multiple samples under density ratio models*, Statistica Sinica, 2017, 27: 761-783.

[5] K L Delucchi, A Bostrom. *Methods for analysis of skewed data distributions in psychiatric clinical studies: working with many zero values*, American Journal of Psychiatry, 2004, 161: 1159-1168.

[6] K Fokianos. *Density ratio model selection*, Journal of Statistical Computation and Simulation, 2007, 77: 805-819.

[7] A P Hallstrom. *A modified Wilcoxon test for non-negative distributions with a clump of zeros*, Statistics in Medicine, 2010, 29: 391-400.

[8] W Kassahun-Yimer, P S Albert, L M Lipsky, A Liu. *A joint model for multivariate hierarchical semicontinuous data with replications*, Statistical Methods in Medical Research, 2019, 28: 858-870.

[9] R Kay, S Little. *Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data*, Biometrika, 1987, 74(3): 495-501.

[10] P A Lachenbruch. *Comparisons of two-part models with competitors*, Statistics in Medicine, 2001, 20: 1215-1234.

[11] P A Lachenbruch. *Analysis of data with excess zeros*, Statistical Methods in Medical Research, 2002, 11: 297-302 (2002).

[12] Y Min, A Agresti. *Modeling nonnegative data with clumping at zero: A survey*, Journal of The Iranian Chemical Society, 2002, 1: 7-33.

[13] K Muralidharan, B K Kale. *Modified gamma distribution with singularity at zero*, Communications in Statistics: Simulation and Computation, 2002, 31: 143-158.

[14] TR Nansel, LMB Laffel, DL Haynie, SN Mehta, LM Lipsky, LK Volkening, DA Butler, LA Higgins, A Liu. *Improving dietary quality in youth with type 1 diabetes: randomized clinical trial of a family-based behavioral intervention*, International Journal of Behavioral Nutrition and Physical Activity, 2015, 12(1): 58.

[15] J Qin, B Zhang. *A goodness-of-fit test for logistic regression models based on case-control data*, Biometrika, 1997, 84: 609-618.

[16] J Qin. *Empirical likelihood ratio based confidence intervals for mixture proportions*, The Annals of Statistics, 1999, 27: 1368-1384.

[17] S Taylor, K Pollard. *Hypothesis tests for point-mass mixture data with application to omics data with many zero values*, Statistical Applications in Genetics and Molecular Biology, 2009, 8: 1-43.

[18] B D Wagner, C E Robertson, J K Harris. *Application of two-part statistics for comparison of sequence variant counts*, PLoS One, 2011, 6: 20-26.

[19] B Zhang. *Assessing goodness-of-fit of generalized logit models based on case-control data*, Journal of Multivariate Analysis, 2002, 82: 17-38.

[20] L Zhang, J Wu, W D Johnson. *Empirical study of six tests for equality of populations with zero-inflated continuous distributions*, Communications in Statistics: Simulation and Computation, 2010, 39: 1196-1211.

[21] F Zou, J P Fine, B S Yandell. *On empirical likelihood for a semiparametric mixture model*, Biometrika, 2002, 89: 61-75.

[1]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China.
Email: Luyahui92@163.com, jiangmengjie24@163.com

[2]Biostatistics and Bioinformatics Branch, *Eunice Kennedy Shriver* National Institute of China Health and Human Development, Bethesda, MD 20817, USA.
Email: liu@mail.nih.gov

[3]Hangzhou Commercial College, Zhejiang Gongshang University, Hangzhou 311508, China.
Email: jtao@263.net