

关于两总体均值差的一种半参数假设检验方法

万树文*, 方芳

南京财经大学应用数学学院, 南京 210046
E-mail: wanshuwen@yahoo.com.cn, jiaodawcome@163.com

收稿日期: 2012-02-08; 接受日期: 2012-04-17; * 通信作者
国家自然科学基金 (批准号: 11001119) 资助项目

摘要 近年来, 已有一些在半参数密度函数比模型下建立半参数统计分析方法的报道, 这些方法往往比参数方法稳健, 比非参数方法有效. 在本文里, 我们提出一种半参数的假设检验方法用于对两总体均值差进行假设检验. 该方法主要建立在对两总体均值差进行半参数估计的基础上. 我们报告了一些理论和统计模拟的结果, 得出该方法在数据符合正态性假设时, 比常用的参数和非参数方法略好; 而在数据不符合正态性假设时, 它的优势就非常明显. 我们还将提出的方法用到了两组真实数据的分析上.

关键词 密度函数比模型 经验似然 假设检验 半参数统计学 两样本 t 检验

MSC (2010) 主题分类 62G10, 62G20

1 引言

我们探讨关于两总体均值差进行假设检验的经典统计问题. 以 $X_1, \dots, X_{n_0}$ 表示从第一个总体中获得的独立同分布的样本数据, 以 $Y_1, \dots, Y_{n_1}$ 表示从与第一个总体独立的第二个总体中获得的独立同分布的样本数据. 另外, 设 $\mu_1 = E(X_1)$, $\mu_2 = E(Y_1)$. 现考虑假设检验 $H_0: \mu_1 - \mu_2 = \theta_0$, θ_0 是一个常数. 最常用的方法是两样本 t 检验. 它假设数据是来自两个正态分布的总体, 这两个总体的方差可以假设相等或不相等, 在不相等时可以用 Welch 方法近似算出 t 分布的自由度. 但在实际研究中, 我们经常遇到数据不服从正态分布的情形, 这时传统的两样本 t 检验就不能使用了. 以 $\hat{\mu}_1, \hat{\mu}_2$ 分别代表两个样本的样本均值, 以 S_X^2, S_Y^2 分别代表样本方差, 则一个非参数的检验统计量

$$\hat{\Lambda} = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \theta_0}{\sqrt{\frac{S_X^2}{n_0} + \frac{S_Y^2}{n_1}}}$$

服从渐近的正态分布可以用来进行假设检验.

在本文里, 我们致力于在一个半参数密度函数比模型下建立对 $H_0: \mu_1 - \mu_2 = \theta_0$ 进行假设检验的方法. 在文献里, 已有多在半参数密度函数比模型下建立方法进行各种统计分析的报道, 这些半参数方法通常比传统的参数方法稳健, 比非参数方法有效. 例如, Qin 和 Zhang^[1], Wan 和 Zhang^[2] 建立了在密度函数比模型下进行 ROC 曲线估计的半参数方法. Wan 和 Zhang^[3] 建立了在密度函数比模型下进行 ROC 曲线比较的半参数方法. Folkianos^[4], Cheng 和 Chu^[5], 以及 Qin 和 Zhang^[6] 建立了半参数的密度函数估计方法. Qin^[7] 研究了在密度函数比模型下如何对混合比例进行半参数的统计推断. Zhang^[8] 研究了半参数的分位数估计方法.

英文引用格式: Wan S W, Fang F. A semiparametric hypothesis testing method on the difference of two population means (in Chinese). Sci Sin Math, 2012, 42(7): 671–679, doi: 10.1360/012012-104

以 G 和 F 分别代表 X_1 和 Y_1 的分布函数, 且以 $g(x)$ 和 $f(x)$ 代表相应的密度函数, 则半参数密度函数比模型假设 $g(x)$ 和 $f(x)$ 之间有一个指数形式的连接 $\exp\{\alpha + \beta^T r(x)\}$:

$$X_1, \dots, X_{n_0} \sim g(x), \quad Y_1, \dots, Y_{n_1} \sim f(x) = \exp\{\alpha + \beta^T r(x)\}g(x), \quad (1)$$

这里, α 是一个标量参数, β 是一个 $p \times 1$ 的向量参数, $r(x)$ 是一个 $p \times 1$ 的关于 x 的光滑的向量函数. Kay 和 Little^[9] 讨论过适用于常用概率分布的密度函数比模型的一些形式. 在这个模型里, 我们没有要求密度函数 $g(x)$ 和 $f(x)$ 是某个具体的分布, 只是要求它们之间有一个指数形式的连接. 这不是传统的参数模型, 但由于在模型中有参数的存在, 故也不是一个非参数的模型. 模型 (1) 是介于参数和非参数之间的一种半参数模型. 在实践中, 我们常常选择 $r(x) = x$ 或者 $r(x) = (x, x^2)^T$. 当 $r(x) = x$ 时, 模型 (1) 包含许多常用分布包括两个有相同方差不同均值的正态分布, 以及两个有不同均值的指数分布; 当 $r(x) = (x, x^2)^T$ 时, 模型 (1) 包含两个有不同均值和不同方差的正态分布. 有时, $r(x) = \log(x)$ 也是一种较好的选择, 此时模型 (1) 包含两个对数正态分布 $\log N(\theta_1, \sigma^2)$ 和 $\log N(\theta_2, \sigma^2)$, $\theta_1 \neq \theta_2$. 至于在实际应用中如何选择 $r(x)$, Kay 和 Little^[9] 提供了较好的参考. 对于一个给定的 $r(x)$, 我们可以采用 Qin 和 Zhang^[10] 提出的方法构造 Kolmogorov-Smirnov 型的统计量进行拟合优度检验.

本文的结构如下: 第 2 节, 我们提出一种半参数的假设检验方法并报告相关的一些理论结果; 第 3 节, 我们汇报一个统计模拟实验的结果; 第 4 节, 我们将提出的方法应用到两组真实数据的分析中去; 最后, 在第 5 节里我们提供了有关定理的详细证明.

2 主要方法

以 $\{T_1, \dots, T_n\}$ 表示合并的样本 $\{X_1, \dots, X_{n_0}; Y_1, \dots, Y_{n_1}\}$, 且记 $n = n_0 + n_1$. 根据模型 (1) 里观察到的样本数据, 我们可以写出相应的半参数经验似然函数为:

$$\begin{aligned} \mathcal{L}(\alpha, \beta, G) &= \prod_{i=1}^{n_0} dG(X_i) \prod_{j=1}^{n_1} \exp(\alpha + \beta^T r(Y_j)) dG(Y_j) \\ &= \prod_{i=1}^n p_i \prod_{j=1}^{n_1} \exp(\alpha + \beta^T r(Y_j)), \end{aligned}$$

这里, $p_i = dG(T_i)$, $i = 1, \dots, n$, 是概率的跃变, 且总和为 1. 类似于文献 [10], 我们采用 Lagrange 乘子法发现在如下约束条件

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i \{\exp(\alpha + \beta^T r(T_i)) - 1\} = 0$$

下, 似然函数 $\mathcal{L}$ 的最大值在

$$\tilde{p}_i = \frac{1}{n_0} \frac{1}{1 + \rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))}$$

时获得, 这里 $\rho = n_1/n_0$, $(\tilde{\alpha}, \tilde{\beta})$ 是 (α, β) 的最大半参数似然估计量, 通过解如下的计分方程组而获得:

$$\begin{aligned} \frac{\partial l(\alpha, \beta)}{\partial \alpha} &= n_1 - \sum_{i=1}^n \frac{\rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))}{1 + \rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))} = 0, \\ \frac{\partial l(\alpha, \beta)}{\partial \beta} &= \sum_{j=1}^{n_1} r(Y_j) - \sum_{i=1}^n \frac{\rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))}{1 + \rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))} r(T_i) = 0, \end{aligned}$$

这里, $l(\alpha, \beta)$ 是关于 (α, β) 的剖面对数似然函数,

$$l(\alpha, \beta) = \sum_{j=1}^{n_2} [\alpha + \beta^T r(Y_j)] - \sum_{i=1}^n \log[1 + \rho \exp(\alpha + \beta^T r(T_i))] - n \log n_0.$$

进而, 模型 (1) 下的 $G(t)$ 和 $F(t)$ 的最大半参数似然估计量为:

$$\begin{aligned}\tilde{G}(t) &= \sum_{i=1}^n p_i I(T_i \leq t) = \frac{1}{n_0} \sum_{i=1}^n \frac{1}{1 + \rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))} I(T_i \leq t), \\ \tilde{F}(t) &= \sum_{i=1}^n p_i \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i)) I(T_i \leq t) = \frac{1}{n_0} \sum_{i=1}^n \frac{\exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))}{1 + \rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))} I(T_i \leq t).\end{aligned}$$

于是, 我们可以通过代入 $G(t)$ 和 $F(t)$ 的估计量获得两总体均值差 $\mu_1 - \mu_2 = \int t dG - \int t dF$ 的最大似然估计量为:

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \frac{1}{n_0} \sum_{i=1}^n \frac{1 - \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))}{1 + \rho \exp(\tilde{\alpha} + \tilde{\beta}^T r(T_i))} T_i.$$

注意: 在模型 (1) 的 $r(x) = x$ 这种特殊情形里, 可以通过一些算术证明半参数估计量 $\tilde{\mu}_1 - \tilde{\mu}_2$ 跟常用的非参数估计量 $\hat{\mu}_1 - \hat{\mu}_2$ 的值相同. 当 $r(x)$ 不包含 x 时, 这两者估计量的值通常是不同的.

接下来, 我们研究半参数估计量 $\tilde{\mu}_1 - \tilde{\mu}_2$ 的渐近分布. 为此, 以 (α_0, β_0) 代表参数 (α, β) 在模型 (1) 下的真实值. 同时, 我们假设 $\rho = n_1/n_0$ 在 $n = n_0 + n_1 \rightarrow \infty$ 时保持不变. 另外, 为了表达的简洁引入如下的记号:

$$\begin{aligned}\omega(t) &= \exp(\alpha_0 + \beta_0^T r(t)), \quad A_0 = \int_{-\infty}^{\infty} \frac{\omega(y)}{1 + \rho \omega(y)} dG(y), \\ A_1 &= \int_{-\infty}^{\infty} \frac{\omega(y)}{1 + \rho \omega(y)} r(y) dG(y), \quad A_2 = \int_{-\infty}^{\infty} \frac{\omega(y)}{1 + \rho \omega(y)} r(y) \{r(y)\}^T dG(y), \\ A &= \begin{pmatrix} A_0 & A_1^T \\ A_1 & A_2 \end{pmatrix}, \quad S = \frac{\rho}{1 + \rho} A, \quad B_0 = \int_{-\infty}^{\infty} \frac{\omega(y)y}{1 + \rho \omega(y)} dG(y), \\ B_1 &= \int_{-\infty}^{\infty} \frac{\omega(y)y}{1 + \rho \omega(y)} r(y) dG(y), \quad B_2 = \int_{-\infty}^{\infty} \frac{\omega(y)y^2}{1 + \rho \omega(y)} dG(y).\end{aligned}$$

下面的定理建立了 $\tilde{\mu}_1 - \tilde{\mu}_2$ 的渐近正态性, 其详细证明在第 5 节里提供.

定理 2.1 如果模型 (1) 成立, 且 A^{-1} 存在, 我们有

$$\sqrt{n}(\tilde{\mu}_1 - \tilde{\mu}_2 - (\mu_1 - \mu_2)) \rightarrow N(0, \sigma_{\text{semi}}^2),$$

这里,

$$\begin{aligned}\sigma_{\text{semi}}^2 &= (1 + \rho) \left\{ \int_{-\infty}^{\infty} u^2 dG(u) - \left[\int_{-\infty}^{\infty} u dG(u) \right]^2 \right\} + \frac{1 + \rho}{\rho} \left\{ \int_{-\infty}^{\infty} u^2 dF(u) - \left[\int_{-\infty}^{\infty} u dF(u) \right]^2 \right\} \\ &\quad - \frac{(1 + \rho)^3}{\rho} \left\{ B_2 - (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \right\}.\end{aligned}$$

我们下面的定理表明半参数估计量 $\tilde{\mu}_1 - \tilde{\mu}_2$ 比常用的非参数估计量 $\hat{\mu}_1 - \hat{\mu}_2$ 渐近有效. 定理 2.2 的详细证明将在第 5 节里提供.

定理 2.2 在定理 2.1 的条件下, 我们有 $\sqrt{n}(\tilde{\mu}_1 - \tilde{\mu}_2 - (\mu_1 - \mu_2)) \rightarrow N(0, \sigma_{\text{semi}}^2)$ 和 $\sqrt{n}(\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)) \rightarrow N(0, \sigma_{\text{non}}^2)$, 这里 σ_{semi}^2 和 σ_{non}^2 满足

$$\sigma_{\text{non}}^2 - \sigma_{\text{semi}}^2 \geq 0.$$

进而, $\hat{\mu}_1 - \hat{\mu}_2$ 对于 $\tilde{\mu}_1 - \tilde{\mu}_2$ 的渐近相对效率小于 1.

σ_{semi}^2 的相合估计量 $\hat{\sigma}_{\text{semi}}^2$ 可以通过代入有关最大半参数似然估计量 $(\tilde{\alpha}, \tilde{\beta}, \tilde{G}, \tilde{F})$ 的方法来构建. 具体来说,

$$\begin{aligned} \hat{\sigma}_{\text{semi}}^2 = & (1 + \rho) \left\{ \sum_{i=1}^n T_i^2 \tilde{p}_i - \left[\sum_{i=1}^n T_i \tilde{p}_i \right]^2 \right\} + \frac{1 + \rho}{\rho} \left\{ \sum_{i=1}^n T_i^2 \tilde{\omega}(T_i) \tilde{p}_i - \left[\sum_{i=1}^n T_i \tilde{\omega}(T_i) \tilde{p}_i \right]^2 \right\} \\ & - \frac{(1 + \rho)^3}{\rho} \left\{ \tilde{B}_2 - (\tilde{B}_0, \tilde{B}_1^T) \tilde{A}^{-1} \begin{pmatrix} \tilde{B}_0 \\ \tilde{B}_1 \end{pmatrix} \right\}, \end{aligned}$$

这里,

$$\begin{aligned} \tilde{\omega}(t) &= \exp(\tilde{\alpha} + \tilde{\beta}^T r(t)), \quad \tilde{A}_0 = \sum_{i=1}^n \frac{\tilde{\omega}(T_i)}{1 + \rho \tilde{\omega}(T_i)} \tilde{p}_i, \\ \tilde{A}_1 &= \sum_{i=1}^n \frac{\tilde{\omega}(T_i)}{1 + \rho \tilde{\omega}(T_i)} r(T_i) \tilde{p}_i, \quad \tilde{A}_2 = \sum_{i=1}^n \frac{\tilde{\omega}(T_i)}{1 + \rho \tilde{\omega}(T_i)} r(T_i) \{r(T_i)\}^T \tilde{p}_i, \\ \tilde{A} &= \begin{pmatrix} \tilde{A}_0 & \tilde{A}_1^T \\ \tilde{A}_1 & \tilde{A}_2 \end{pmatrix}, \quad \tilde{B}_0 = \sum_{i=1}^n \frac{\tilde{\omega}(T_i) T_i}{1 + \rho \tilde{\omega}(T_i)} \tilde{p}_i, \\ \tilde{B}_1 &= \sum_{i=1}^n \frac{\tilde{\omega}(T_i) T_i}{1 + \rho \tilde{\omega}(T_i)} r(T_i) \tilde{p}_i, \quad \tilde{B}_2 = \sum_{i=1}^n \frac{\tilde{\omega}(T_i) T_i^2}{1 + \rho \tilde{\omega}(T_i)} \tilde{p}_i. \end{aligned}$$

现在, 我们来考虑在模型 (1) 下进行假设检验 $H_0: \mu_1 - \mu_2 = \theta_0$, $H_1: \mu_1 - \mu_2 \neq \theta_0$. 根据定理 2.1, 一个半参数的 Wald 型统计量构建为

$$\tilde{\Lambda} = \frac{\sqrt{n}(\tilde{\mu}_1 - \tilde{\mu}_2 - \theta_0)}{\tilde{\sigma}_{\text{semi}}}.$$

很显然, 当 $|\tilde{\Lambda}|$ 的值很大时, 我们倾向于拒绝 $H_0: \mu_1 - \mu_2 = \theta_0$. 根据定理 2.1 有, $\tilde{\Lambda} \rightarrow N(0, 1)$, 故在显著性水平为 λ 时, 如果 $|\tilde{\Lambda}| > z_{1-\lambda/2}$, 这里 $z_{1-\lambda/2}$ 满足 $P(Z \leq z_{1-\lambda/2}) = 1 - \lambda/2$, $Z \sim N(0, 1)$, 我们拒绝零假设 $H_0: \mu_1 - \mu_2 = \theta_0$ 而接受备择假设 $H_1: \mu_1 - \mu_2 \neq \theta_0$. $\tilde{\Lambda}$ 当然也可以用来检验 $H_0: \mu_1 - \mu_2 = \theta_0$, $H_1: \mu_1 - \mu_2 > \theta_0$. 如果 $\tilde{\Lambda} > z_{1-\lambda}$, 我们就拒绝零假设 H_0 而接受备择假设 H_1 .

3 统计模拟

在这节里, 我们用统计模拟实验比较了在有限样本情形下我们提出的半参数检验统计量 $\tilde{\Lambda}$ 与相应的参数和非参数检验统计量.

在统计模拟实验的第一种情形下, 我们假定 $g(x)$ 和 $f(x)$ 分别是正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的密度函数. 则此时模型 (1) 成立, 且有 $r(x) = (x, x^2)^T$, $\beta = (\beta_1, \beta_2)^T$ 和

$$\alpha = \log \frac{\sigma_1}{\sigma_2} + \frac{1}{2} \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right), \quad \beta_1 = \frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}, \quad \beta_2 = \frac{1}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right).$$

我们设置的样本容量为 $(n_0, n_1) = (40, 60)$, $(n_0, n_1) = (50, 50)$ 和 $(n_0, n_1) = (60, 40)$. 我们这样设置的原因是在第 4 节里提到的两个实例里, 样本容量之和在 100 左右. 另外, 设置 $\mu_1 = 0$, $\mu_2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, $\sigma_1 = 1$, $\sigma_2 = 2$. 对于每种组合 $(\mu_1, \mu_2, \sigma_1, \sigma_2)$, 我们从 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 中产生 1000 组相互独立的样本. 我们考虑在显著性水平为 0.05 时检验假设 $H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 \neq 0$, 相应的统计模拟结果总结在表 1 中.

表 1 中的统计模拟结果显示, 三种方法实现的显著性水平都与名义显著性水平 0.05 很接近. 另外, 三种假设检验方法的效能也很相近, 但是在大多数模拟情形下, 半参数方法有着最大的效能. 同时, 我们发现在样本容量的三种设置方式下, 有相似的统计模拟结果. 因此, 在数据满足正态性假设的前提下, 三种假设检验方法的效果很接近, 而半参数方法比其它两种方法在检验效能上略好.

我们在统计模拟中也考虑了分布是偏态的情形. 我们假设 $g(x)$ 和 $f(x)$ 分别是对数正态分布 $\log N(\theta_1, \sigma^2)$ 和 $\log N(\theta_2, \sigma^2)$ 的密度函数. 因此, 通常的正态性假设不再成立. 但是在此种情形下, 模型 (1) 仍然成立, 且有 $r(x) = \log x$ 和

$$\alpha = \frac{\theta_1^2 - \theta_2^2}{\sigma^2}, \quad \beta = \frac{\theta_2 - \theta_1}{\sigma^2}.$$

我们设置的样本容量为 $(n_0, n_1) = (40, 60)$, $(n_0, n_1) = (50, 50)$ 和 $(n_0, n_1) = (60, 40)$. 另外, 设置 $\theta_1 = 0$, $\theta_2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, $\sigma = 1$. 对于每种组合 $(\theta_1, \theta_2, \sigma)$, 我们从 $\log N(\theta_1, \sigma^2)$ 和 $\log N(\theta_2, \sigma^2)$ 中产生 1000 组相互独立的样本. 我们考虑在显著性水平为 0.05 时检验假设 $H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 \neq 0$, 相应的统计模拟结果总结在了表 2 中.

表 2 中的统计模拟结果显示: 当数据是对数正态分布时, 两样本 t 检验无论是从实现的显著性水平还是检验效能上都比其它两种方法明显的差. 具体地说, 两样本 t 检验实现的显著性水平不如其它两种方法离名义水平 0.05 近; 另外从检验效能上看, 两样本 t 检验的效能是三种方法中最低的. 就半参数方法和非参数方法比较而言, 两者实现的显著性水平都离 0.05 较近, 但是半参数方法的效能要比非参数方法的效能大, 这与定理 2.2 中的理论结果是一致的. 因此, 我们得出当数据是非正态时, 例如这里提到的对数正态分布, 我们提出的半参数方法较优.

表 1 数据是正态分布时, 三种假设检验方法实现的显著性水平和检验效能

μ_2	0	0.2	0.4	0.6	0.8	1.0
$(n_0, n_1) = (40, 60)$						
两样本 t 检验	0.048	0.096	0.283	0.520	0.748	0.898
$\tilde{\Lambda}$	0.054	0.105	0.291	0.515	0.753	0.903
$\hat{\Lambda}$	0.053	0.099	0.282	0.518	0.750	0.902
$(n_0, n_1) = (50, 50)$						
两样本 t 检验	0.049	0.100	0.219	0.475	0.704	0.875
$\tilde{\Lambda}$	0.054	0.106	0.228	0.486	0.718	0.880
$\hat{\Lambda}$	0.053	0.102	0.222	0.481	0.710	0.875
$(n_0, n_1) = (60, 40)$						
两样本 t 检验	0.051	0.101	0.243	0.426	0.621	0.842
$\tilde{\Lambda}$	0.054	0.108	0.244	0.444	0.648	0.855
$\hat{\Lambda}$	0.054	0.102	0.244	0.435	0.640	0.849

表 2 当数据是对数正态分布时, 三种假设检验方法实现的显著性水平和检验效能

θ_2	0	0.2	0.4	0.6	0.8	1.0
$(n_0, n_1) = (40, 60)$						
两样本 t 检验	0.065	0.151	0.404	0.676	0.899	0.955
$\tilde{\Lambda}$	0.052	0.188	0.498	0.834	0.959	0.995
$\hat{\Lambda}$	0.061	0.157	0.415	0.685	0.902	0.958
$(n_0, n_1) = (50, 50)$						
两样本 t 检验	0.041	0.120	0.344	0.673	0.883	0.961
$\tilde{\Lambda}$	0.045	0.148	0.446	0.817	0.966	0.988
$\hat{\Lambda}$	0.049	0.125	0.362	0.684	0.888	0.965
$(n_0, n_1) = (60, 40)$						
两样本 t 检验	0.039	0.100	0.278	0.592	0.833	0.949
$\tilde{\Lambda}$	0.045	0.130	0.418	0.781	0.941	0.989
$\hat{\Lambda}$	0.046	0.115	0.298	0.615	0.846	0.955

综合两种情形下的统计模拟实验, 我们发现提出的半参数方法在数据符合正态性假设时仍然比其它两者方法略优, 而在数据不符合正态性假设时, 它比其它两种方法明显的好.

4 实例分析

第一组真实数据来源于文献 [11], 它是采用糖类抗原 CA19-9 对 51 个胰腺炎病人和 90 个胰腺癌病人血清的测试结果. 这里我们运用提出的方法研究两类病人的测试结果是否有差异. 因为原始数据横跨几个数量级, 而且数据的 Shapiro-Wilk 正态性检验的 p 值接近于 0, 故对原始数据采用对数变换. 但是, 变换后的数据在显著性水平为 0.05 时仍然不能通过 Shapiro-Wilk 正态性检验, 故传统的两样本 t 检验是不能采用的. Qin 和 Zhang^[10] 对该组数据进行了模型 (1) 在 $r(x) = (x, x^2)^T$ 时的拟合优度检验, 汇报了较好的拟合. 参数 α 和 β 的最大半参数似然估计值为 $\tilde{\alpha} = 0.56002$ 和 $\tilde{\beta} = (-1.91417, 0.45057)^T$. 我们考虑显著性水平为 0.05 时的假设检验 $H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 \neq 0$. 我们提出的半参数检验统计量的值为 $\tilde{\Lambda} = -10.77409$, 故我们拒绝零假设, 从而得出胰腺炎病人和胰腺癌病人的 CA19-9 测试结果是具有显著差异的. 此时非参数统计量 $\hat{\Lambda} = -10.70513$, 我们会得出与半参数分析相同的结论.

第二组真实的数据来源于文献 [12] 关于糖尿病的研究. 我们想检验 76 个正常人与 33 个糖尿病人的空腹血浆葡萄糖水平 (PLG) 是否有显著差异. Shapiro-Wilk 正态性检验显示, 正常人群的 PLG 数据不符合正态分布. 运用文献 [10] 里的拟合优度检验发现模型 (1) 在 $r(x) = x$ 时有较好的拟合. 参数 α 和 β 的最大半参数似然估计值为 $\tilde{\alpha} = -6.27848$ 和 $\tilde{\beta} = 0.03084$. 我们考虑显著性水平为 0.05 时的假设检验 $H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 \neq 0$. 我们提出的半参数检验统计量的值为 $\tilde{\Lambda} = -12.43398$, 故我们拒绝零假设, 从而得出正常人和糖尿病人的空腹血浆葡萄糖水平是有显著差异的. 此时, 从非参数检验统计量 $\hat{\Lambda} = -12.24514$ 会得出相同的结论.

5 定理证明

定理 2.1 的证明 对 $\tilde{\mu}_1 - \tilde{\mu}_2$ 进行 Taylor 展开, 并结合文献 [10] 中 $\tilde{\alpha}$ 和 $\tilde{\beta}$ 的 $\sqrt{n}$ 一致性, 我们

有

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \frac{1}{n_0} \sum_{i=1}^n \frac{1 - \omega(T_i)}{1 + \rho\omega(T_i)} T_i - \frac{1 + \rho}{n_0} \left(\sum_{i=1}^n \frac{\omega(T_i)T_i}{(1 + \rho\omega(T_i))^2}, \sum_{i=1}^n \frac{\omega(T_i)r^T(T_i)T_i}{(1 + \rho\omega(T_i))^2} \right) \begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} + O_p(n^{-1}).$$

通过直接的运算可得

$$E \left\{ \frac{1}{n_0} \sum_{i=1}^n \frac{\omega(T_i)T_i}{(1 + \rho\omega(T_i))^2} \right\} = B_0, \quad E \left\{ \frac{1}{n_0} \sum_{i=1}^n \frac{\omega(T_i)r^T(T_i)T_i}{(1 + \rho\omega(T_i))^2} \right\} = B_1^T.$$

于是, 由大数定律和文献 [10] 中的渐近表达式

$$\begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} = \frac{1}{n} S^{-1} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} + o_p(n^{-1/2}),$$

我们有

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \frac{1}{n_0} \sum_{i=1}^n \frac{1 - \omega(T_i)}{1 + \rho\omega(T_i)} T_i - \frac{1}{n_0} (B_0, B_1^T) S^{-1} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} + o_p(n^{-1/2}).$$

简单推算就得到 $E(\tilde{\mu}_1 - \tilde{\mu}_2) = \mu_1 - \mu_2$, 这显示 $\tilde{\mu}_1 - \tilde{\mu}_2$ 是渐近无偏的. 记 $\Delta = \frac{1}{n_0} \sum_{i=1}^n \frac{1 - \omega(T_i)}{1 + \rho\omega(T_i)} T_i$. 为了推导 $\sqrt{n}(\tilde{\mu}_1 - \tilde{\mu}_2)$ 的方差, 我们将其剖分成三部分,

$$\begin{aligned} \text{Var}(\sqrt{n}(\tilde{\mu}_1 - \tilde{\mu}_2)) &= \text{Var}(\sqrt{n}\Delta) - 2(1 + \rho)(B_0, B_1^T) S^{-1} \begin{pmatrix} \text{Cov}(\Delta, \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha}) \\ \text{Cov}(\Delta, \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta}) \end{pmatrix} \\ &\quad + \left\{ \frac{1 + \rho}{n_0} (B_0, B_1^T) S^{-1} \text{Var} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} S^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \right\} \\ &= V_1 + V_2 + V_3. \end{aligned}$$

通过一些繁琐的推导知,

$$\begin{aligned} V_1 &= (1 + \rho) \left[\int_{-\infty}^{\infty} u^2 dG(u) + \frac{1}{\rho} \int_{-\infty}^{\infty} u^2 dF(u) - \frac{(1 + \rho)^2}{\rho} B_2 \right] \\ &\quad - (1 + \rho) \left[\int_{-\infty}^{\infty} u dG(u) - (1 + \rho) B_0 \right]^2 - \frac{1 + \rho}{\rho} \left[\int_{-\infty}^{\infty} u dF(u) - (1 + \rho) B_0 \right]^2, \\ V_2 &= -2(1 + \rho)^2 B_0 \int_{-\infty}^{\infty} u dG(u) + \frac{2(1 + \rho)^4}{\rho} B_0^2 - \frac{2(1 + \rho)^2}{\rho} B_0 \int_{-\infty}^{\infty} u dF(u), \\ V_3 &= \frac{(1 + \rho)^3}{\rho} (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} - \frac{(1 + \rho)^4}{\rho} B_0^2. \end{aligned}$$

于是,

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

由中心极限定理知,

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

□

定理 2.2 的证明 很容易验证

$$\sigma_{\text{non}}^2 = (1 + \rho) \left\{ \int_{-\infty}^{\infty} u^2 dG(u) - \left[\int_{-\infty}^{\infty} u dG(u) \right]^2 \right\} + \frac{1 + \rho}{\rho} \left\{ \int_{-\infty}^{\infty} u^2 dF(u) - \left[\int_{-\infty}^{\infty} u dF(u) \right]^2 \right\},$$

于是, 我们有

$$\sigma_{\text{non}}^2 - \sigma_{\text{semi}}^2 = \frac{(1+\rho)^3}{\rho} \left\{ B_2 - (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \right\}.$$

设

$$\Gamma = \frac{1}{n_0} \sum_{i=1}^n \frac{T_i}{1 + \rho\omega(T_i)} - \frac{1}{n_0} \sum_{i=1}^{n_0} X_i - \frac{\rho}{n} (B_0, B_1^T) S^{-1} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix},$$

并记

$$\Delta = \frac{1}{n_0} \sum_{i=1}^n \frac{T_i}{1 + \rho\omega(T_i)} - \frac{1}{n_0} \sum_{i=1}^{n_0} X_i.$$

为了推导 Γ 的方差, 我们将其剖分成三部分,

$$\begin{aligned} \text{Var}(\Gamma) &= \text{Var}(\Delta) - \frac{2\rho}{n} (B_0, B_1^T) S^{-1} \begin{pmatrix} \text{Cov}(\Delta, \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha}) \\ \text{Cov}(\Delta, \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta}) \end{pmatrix} \\ &\quad + \left\{ \frac{\rho^2}{n^2} (B_0, B_1^T) S^{-1} \text{Var} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} S^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \right\} \\ &= V_1 + V_2 + V_3. \end{aligned}$$

通过一些繁琐的推算可得

$$\begin{aligned} V_1 &= \frac{\rho}{n_0} B_2 - \frac{\rho(1+\rho)}{n_0} B_0^2, \\ V_2 &= -\frac{2\rho(1+\rho)}{n} (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} + \frac{2\rho(1+\rho)^2}{n} B_0^2, \\ V_3 &= \frac{\rho(1+\rho)}{n} (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} - \frac{\rho(1+\rho)^2}{n} B_0^2. \end{aligned}$$

于是,

$$\text{Var}(\Gamma) = \frac{\rho}{n_0} \left\{ B_2 - (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \right\}.$$

因为 $\text{Var}(\Gamma) \geq 0$, 所以有 $\{B_2 - (B_0, B_1^T) A^{-1} \begin{pmatrix} B_0 \\ B_1 \end{pmatrix}\} \geq 0$, 也就是 $\sigma_{\text{non}}^2 - \sigma_{\text{semi}}^2 \geq 0$. $\square$

致谢 作者感谢两位审稿人和编辑提出的宝贵意见.

参考文献

- 1 Qin J, Zhang B. Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika*, 2003, 90: 585–596
- 2 Wan S, Zhang B. Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Statist Med*, 2007, 26: 2565–2586
- 3 Wan S, Zhang B. Comparing correlated ROC curves for continuous diagnostic tests under density ratio models. *Comput Statist Data Anal*, 2008, 53: 233–245
- 4 Fokianos K. Merging information for semiparametric density estimation. *J Roy Statist Soc Ser B*, 2004, 66: 941–958
- 5 Cheng K, Chu C. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 2004, 10: 583–604
- 6 Qin J, Zhang B. Density estimation under a two-sample semiparametric model. *J Nonparametr Statist*, 2005, 17: 665–683
- 7 Qin J. Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann Statist*, 1999, 27: 1368–1384

- 8 Zhang B. Quantile estimation under a two-sample semi-parametric model. *Bernoulli*, 2000, 6: 491–511
- 9 Kay R, Little S. Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 1987, 74: 495–501
- 10 Qin J, Zhang B. A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 1997, 84: 609–618
- 11 Wieand S, Gail M H, James B R, et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 1989, 76: 585–592
- 12 Reaven G M, Miller R G. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 1979, 16: 17–24

A semiparametric hypothesis testing method on the difference of two population means

WAN ShuWen & FANG Fang

Abstract In the literature, some semiparametric statistical methods were recently developed under density ratio models, and these methods are usually more robust than parametric methods and more effective than nonparametric methods. In this paper, we propose a semiparametric hypothesis testing method on the difference of two population means under a semiparametric density ratio model. This method is essentially built upon semiparametric estimation of the difference of two population means. Some theoretical results as well as simulation results are provided, which suggest that our proposed method is slightly superior to some commonly used parametric and nonparametric methods when data are normal, and is significantly better than them when data are not normal. We also apply our proposed method to analysis to two real examples.

Keywords density ratio model, empirical likelihood, hypothesis test, semiparametric statistics, two-sample t test

MSC(2010) 62G10, 62G20

doi: 10.1360/012012-104