DOI: 10.16016/j.1000-5404.1999.03.019

两小样本比较的正态先验 Bayes推断

罗明奎* 樊爱军 王开发 (第三军医大学基础医学部数学教研室) 重庆,400038

提 要 目的:两大样本比较问题,传统上对同方差正态情形可采用双样本 t 检验法,对非正态情形可采用非参数秩检验法;但对两小样本比较问题,若采用以上方法其可靠性较差。本研究旨在建立一种两小样本比较的统计处理方法。方法:假定(1)样本来自同方差正态总体(2)参数具有正态先验(3)在平方损失下,采用了经验 Bayes方法。结果:建立了参数的 Bayes估计及 Bayes置信区间,并给出了两小样本比较的 Bayes检验法则。结论:利用实例进一步说明了对两小样本问题经验 Bayes方法较双样本 t 检验法具有更高可靠性。

关键词 Bayes估计; Bayes置性区间; Bayes检验法则

中图法分类号 R311

在许多医学实验中由于种种限制(如实验经费过于昂贵,病例罕见等)只能搜集到少量的样本,称作小样本,此时我们常常希望能通过两小样本去比较两个总体间是否有显著差异,称作两小样本比较问题。对于两小样本比较问题,我们不能简单地套用两大样本比较方法(如双样本 t 检验法,非参数秩检验法等)。小样本问题在实际中经常遇到,但在统计上却较难处理。近年来小样本问题成为统计学讨论的热门话题之一。随着经验 Bayes方法的产生,小样本问题的讨论更趋活跃。本研究就两小样本比较问题介绍一种正态先验 Bayes统计推断方法。

1 Bayes思想追溯

Bayes统计起源于 1763年,在本世纪二、三十年代对 Bayes观点的讨论空前激烈,自 1950年 A. Wald统计判决理论产生后,Bayes方法成为统计判决理论的重要工具。到本世纪六、七十年代产生了经验 Bayes方法。经验 Bayes方法认为对一个实验问题采取什么样的行动不仅依赖于现有的样本信息(不在乎样本容量的大小),而且还应当考虑样本中参数的某些先验信息以及采取行动所带来的损失大小。经验 Bayes方法用于估计问题产生了经验 Bayes估计,经验 Bayes方法用于检验问题则产生了经验 Bayes检验。

由于经验 Bayes方法着眼于现有的样本而不象传统的统计方法那样将样本看着总体中无穷多样本的一部分去考虑,因此,经验 Bayes方法更适用于小样本问题。同时经验 Bayes方法需要利用参数的先验信息和考虑行动所带来的损失,这点也较能贴合小样本问题的实际。例如,要估计好转率,试验前根据经验对好转率取值的可能性有一个初步的认识,即为参数的先验信息;另外,一个估计值如果离真值较远则损失较大,一个估计值如果离真值较远则损失较大,根据 Bayes思想可以看出,经验 Bayes方法有如下要求:

1)样本的分布形式 $f(x,\theta)$已知,其中 θ 为未知参数;

2)确定出未知参数 θ 合理的先验分布 $H(\theta)$;

3)确定出行动 d 合理的损失函数 $L(d,\theta)$,其中 d 为样本的函数;

4)选择使后验平均损失最小的行动作为最终采取的行动。

2 问题及方法

2.1 两小样本比较问题

假定 $X_1, X_2, \cdots X_m$ 来自 X服从正态分布 $N(\theta_1,\sigma^2)$, $Y_1, Y_2, \cdots Y_n$ 来自 Y服从正态分布 $N(\theta_2,\sigma^2)$,共中 m, n均较小,问题是:

1)估计未知参数 θ_1,θ_2;

2)检验假设:

$$H_0:\theta_1=\theta_2 \leftrightarrow H_1:\theta_1\neq\theta_2 \qquad (1)$$

2.2 先验信息的确定

利用共轭先验方法,假定 θ_1,θ_2具有如下先验分布:

$$\theta_1\sim N(\mu_1,\tau_1^2),\ \theta_2\sim N(\mu_2,\tau_2^2)$$

相应的密度函数分别记为:

$$f(\theta_1;\mu_1,\tau_1^2),\ \theta_2\sim N(\mu_2,\tau_2^2)$$

$\mu_1,\tau_1^2,\mu_2,\tau_2^2$可以用如下估计值近似替代:

假定以往有关 X的 k组小样本均值为: $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_k$;以往有关 Y的 r 组小样本值为: $\bar{Y}_1, \bar{Y}_2, \cdots, \bar{Y}_r$,

* 罗明奎,男,32岁,讲师,硕士

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

则

$$\hat{\mu_1}=\frac{1}{k}\sum_{i=1}^{k}\bar{X}_i,\qquad \hat{\tau_1^2}=\frac{1}{k-1}\sum_{i=1}^{k}(\bar{X}_i-\hat{\mu_1})^2$$
$$\hat{\mu_2}=\frac{1}{r}\sum_{i=1}^{r}\bar{Y}_i,\qquad \hat{\tau_2^2}=\frac{1}{r-1}\sum_{i=1}^{r}(\bar{Y}_i-\hat{\mu_2})^2 \tag{2}$$

2.3 损失函数的确定

对 θ_1,θ_2的估计 (行动)$d_1(X)$和 $d_2(Y)$,采用平方损失

$$L(d_1(X),\theta_1)=(d_1(X)-\theta_1)^2$$
$$L(d_2(Y),\theta_2)=(d_2(Y)-\theta_2)^2 \tag{3}$$

按照如上损失函数,将 $d_1(X)$作为 θ_1的估计,若 $d_1(X)$与 θ_1相差越大则损失越大,将 $d_2(Y)$作为 θ_2的估计,若 $d_2(Y)$与 θ_2相差越大则损失越大。

2.4 θ_1,θ_2的 Bayes估计

根据样本 X, Y的分布及未知参数 θ_1,θ_2的先验分布,可计算出在 $X_1,X_2,\cdots X_m$给定下 θ_1的后验分布及在 $Y_1,Y_2,\cdots Y_n$给定下 θ_2的后验分布。

因为

$X_i|\theta_1\sim N(\theta_1,\sigma^2)$,相应的密度函数记为 $f(x_i;\theta_1,\sigma^2)$, i= 1, 2,⋯ , m

$Y_j|\theta_2\sim N(\theta_2,\sigma^2)$,相应的密度函数记为 $f(y_j;\theta_2,\sigma^2)$, j= 1, 2,⋯ , n

从而

在给定 $(X_1,X_2,\cdots X_m)=(x_1,x_2,\cdots,x_m)$的条件下 θ_1的密度函数为

$$\frac{[\prod_{i=1}^{m}f(x_i;\theta_1,\sigma^2)]f(\theta_1;\mu_1,\tau_1^2)}{\int_{-\infty}^{+\infty}[\prod_{i=1}^{m}f(x_i;\theta_1,\sigma^2)]f(\theta_1;\mu_1,\tau_1^2)\,d\theta_1} \tag{4}$$

同理

在给定 $(Y_1,Y_2,\cdots Y_n)=(y_1,y_2,\cdots,y_n)$的条件下 θ_2的密度函数为

$$\frac{[\prod_{j=1}^{n}f(y_j;\theta_2,\sigma^2)]f(\theta_2;\mu_2,\tau_2^2)}{\int_{-\infty}^{+\infty}[\prod_{j=1}^{n}f(y_j;\theta_2,\sigma^2)]f(\theta_2;\mu_2,\tau_2^2)\,d\theta_2} \tag{5}$$

整理 (4)式、(5)式知

$$\theta_1|(X_1,X_2,\cdots,X_m)\sim N\left(\frac{m\bar{X}\tau_1^2+\sigma^2\mu_1}{m\tau_1^2+\sigma^2},\frac{\tau_1^2\sigma^2}{m\tau_1^2+\sigma^2}\right)$$
$$\theta_2|(Y_1,Y_2,\cdots,Y_m)\sim N\left(\frac{n\bar{Y}\tau_2^2+\sigma^2\mu_2}{n\tau_2^2+\sigma^2},\frac{\tau_2^2\sigma^2}{n\tau_2^2+\sigma^2}\right) \tag{6}$$

其中 $\bar{X}=\frac{1}{m}\sum_{i=1}^{m}X_i$, $\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$

则 $d_1(X_1,X_2,\cdots,X_m)$作为 θ_1的估计带来的后验风险 $R(d_1)$, $d_2(Y_1,Y_2,\cdots,Y_n)$作为 θ_2的估计带来的后验风险 $R(d_2)$有如下形式:

$$R(d_1)=E_{\theta_1}(d_1-\theta_1)^2\qquad R(d_2)=E_{\theta_2}(d_2-\theta_2)^2 \tag{7}$$

$E_{\theta_i}(\cdot)$表示在 θ_i的后验分布下求数学期望, i= 1, 2

可以证明当 $d_1=\frac{m\bar{X}\tau_1^2+\sigma^2\mu_1}{m\tau_1^2+\sigma^2}$时 $R(d_1)$达到最小,当 $d_2=\frac{n\bar{Y}\tau_2^2+\sigma^2\mu_2}{n\tau_2^2+\sigma^2}$时 $R(d_2)$达到最小

即 θ_1,θ_2的 Bayes估计为:

$$\hat{\theta_1}=\frac{m\bar{X}\tau_1^2+\sigma^2\mu_1}{m\tau_1^2+\sigma^2},\qquad \hat{\theta_2}=\frac{n\bar{Y}\tau_2^2+\sigma^2\mu_2}{n\tau_2^2+\sigma^2} \tag{8}$$

这里 $\mu_1,\tau_1^2,\mu_2,\tau_2^2$可用 (2)式给出的估计值替代,$\sigma^2$常常未知,可用如下 X, Y的合样本方差 $\hat{\sigma^2}$替代

$$\hat{\sigma^2}=\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}$$
$$s_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,\ s_2^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \tag{9}$$

2.5 $\theta_1-\theta_2$的后验置信区间及 Bayes检验法则

根据 (6)式由正态分布的再生性知 $\theta_1-\theta_2$的后验分布为:

$$(\theta_1-\theta_2)|_{(X_1,X_2,\cdots,X_m,Y_1,Y_2,\cdots,Y_n)}\sim N(\hat{\theta_1}-\hat{\theta_2},\hat{\sigma_1^2}+\hat{\sigma_2^2})$$

其中

$$\hat{\sigma_1^2}=\frac{\tau_1^2\sigma^2}{m\tau_1^2+\sigma^2},\qquad \hat{\sigma_2^2}=\frac{\tau_2^2\sigma^2}{n\tau_2^2+\sigma^2} \tag{10}$$

则 $\theta_1-\theta_2$的 (1-α)100% 后验置信区间为:

$$I=[(\hat{\theta_1}-\hat{\theta_2})-u_{\alpha/2}\sqrt{\hat{\sigma_1^2}+\hat{\sigma_2^2}},(\hat{\theta_1}-\hat{\theta_2})+u_{\alpha/2}\sqrt{\hat{\sigma_1^2}+\hat{\sigma_2^2}}] \tag{11}$$

这里 $u_{\alpha/2}$为标准正态分布的 α/2上侧分位数。

根据 (11)式,在给定显著水平 α下可构造如下 Bayes检验法则:

(1)若置信区间 I包含 0,则应接受 (1)式中的零假设 H_0,认为样本 X与样本 Y间无显著差别:

(2)若置信区间 I不包含 0,则应拒绝 (1)式中的零假设 H_0,认为样本 X与样本 Y间有显著差别。

3 实例分析

为了进一步说明如上小样本 Bayes统计方法的应用,以下利用实际例子进行分析。

[例]第三军医大学组胚教研室为研究癌症患者与非癌症患者间 SDH酶活性有无差异,共收集了 100例癌症患者和 50例非癌症患者的 SDH酶活性数据,见表 1

表 1 SDH酶活性数据表

Tab 1 Activity of SDH

Number	Cancerous patients(X)	Number	Non-cancerous patients(Y)
1	96	1	145
2	93	2	174
3	78	3	159
4	80	4	146
…	…	…	…
97	124	47	134
98	133	48	138
99	100	49	135
100	114	50	158

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

以上问题属于两大样本比较问题,可用双样本 t-检验法处理,结果显示: 两样本方差具有齐性,两样本均值间有显著差异。说明非癌症患者比癌症患者 SDH酶活性显著偏高。

为了说明如上统计方法,特构造如下小样本问题。我们将上述数据中样本 X分为 20组、样本 Y分为 16组,将其中各选一组作小样本比较的现有资料,其余作为小样本经验资料,具体数据如下:

现有资料:

X: 159, 148, 96, 148　　Y: 145, 174, 159　　(*)

经验资料均值:

$\bar{X}_i$: 88. 6, 90, 5, 94. 16, 96. 4, 97. 5, 134. 3, 113. 4, 113. 8, 114. 6, 91. 4, 86. 4, 105. 8, 93. 2, 86. 72, 79. 6, 68, 77. 8, 67, 54

$\bar{Y}_i$: 156, 172, 143. 6, 167. 6, 145. 6, 155. 5, 144. 3, 148. 6, 166. 7, 161. 6, 144. 6, 156. 5, 170, 134. 6, 143. 6

按照上述经验 Bayes方法可以算出:

$\bar{X}$= 137. 5　　$\bar{Y}$= 159. 33　　$\hat{\sigma^2}$= 565. 08

$\hat{\mu_1}$= 92. 27　　$\hat{\tau_1^2}$= 364　　$\hat{\mu_2}$= 154. 06　　$\hat{\tau_2^2}$= 133. 67

$\hat{\theta_1}$= 125. 034　　$\hat{\sigma_1^2}$= 101. 771　　$\hat{\theta_2}$= 156. 24

$\hat{\theta_2^2}$= 78. 185

则 $\theta_1-\theta_2$的 (1- 0. 05) 100% 后验置信区间为:

I= [4. 914, 57. 49]

可以看出后验置信区间 I不包含 0,说明 X与 Y间有显著差异,再根据 $\hat{\theta_1}$,$\hat{\theta_2}$的大小可得出结论: 非癌症患者 SDH酶活性显著偏高。这与大样本时双样本 t检验结论吻合。

但如果对小样本资料 (*)直接采用双样本 t检验法,其结论是: X与 Y间无显著差异。这与大样本时双样本 t检验结论不吻合。

对两小样本比较问题,采用本研究介绍的经验 Bayes方法比传统的双样本 t检验法更吻合实际。因此,对两小样本比较问题,当先验信息选择得当时,宜采用上述经验 Bayes方法。

参考文献

1 陈希孺,倪国熙.数理统计教程.上海: 上海科学技术出版社, 1988. 192~ 239

2 苑锡光.医用统计分析.成都: 四川科学技术出版社, 1985. 89~ 91

3 郭祖超.医用数理统计方法.北京: 人民出版社, 1988. 631~ 647

4 Rorbert E, McCullcch, Ruey S T. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. JASA, 1993, 88(3): 968

5 Thomas A S. Approximately Bayesian inference. JASA, 1994, 89(1): 425

(收稿: 1998-11-04;修回: 1999-02-04)

(编辑　张大春)

A normal prior Bayesian inference for the comparison of two small samples

Luo Mingkui, Fan Aijun, Wang Kaifa (Departmentof Mathematics, Third Military Medical University, Chongqing, 400038)

Abstract Objective: The-two-sample comparison, the two-sample-t-test and the nonparameter rank test are reliable when the sizes of samples are large, but they are not reliable when the sizes are small. The purpose of this paper is to establish a suitable statistical procedure for the small sample. Methods: The empirical Bayesian method is applied and the follows are supposed: ① two samples have normal distribution with the same variance② the parameters have normal prior distribution③ there is square loss. Results: A Bayes estimation and a Bayes posterior confidence interal are established and a Bayes test's rule is constructed for the comparison of two small samples. Conculusion: The emperical Bayesian method is more reliable than the two-sample-t-test method in comparison of two small samples.

Key words Bayes estimation; Bayes posterior confidence interal; Bayes inference

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net