全国中文核心期刊（2004版）

2015 6

大学数学

College Mathematics

第三十一卷 VOL. 31

Daxue Shuxue

ISSN 1672-1454

9 771672 145009 12>

教学研究

一个瑕积分的三种解法及推广 …… 李凤彦，戚晓秋(104)
一类数列极限收敛问题的探讨 …… 杨志林(108)
第一型线面积分的研究型教学方案探析 …… 苏永美(110)
功能性统计软件在统计学专业实践环节中的运用探索 …… 束　慧，熊萍萍(116)
混合样本方差的优良性 …… 崔艳丽，吕　文(123)

[期刊基本参数] CN 34-1221/O1 * 1984 * b * A4 * 128 * zh+en * P * ¥15.00 * 1200 * 25 * 2015-12

本期责任编辑：涂振坤，孙　琳，周　玲

第31卷第6期 大 学 数 学 Vol. 31, No. 6
2015年12月 COLLEGE MATHEMATICS Dec. 2015

混合样本方差的优良性

崔艳丽， 吕 文
(烟台大学 数学与信息科学学院，山东 烟台 264005)

[摘 要] 常见的教科书中通常采用混合样本方差作为方差相等的两个正态总体方差的估计，而对其优良性却鲜有解释. 在本文中，我们证明了混合样本方差是总体方差的UMVUE，即一致最小方差无偏估计.

[关键词] 混合样本方差；无偏；一致最小方差无偏估计
[中图分类号] O212.1 [文献标识码] C [文章编号] 1672-1454(2015)06-0123-04

1 引 言

实际中，我们经常需要处理两个正态总体均值的比较问题，例如检验两种类型的纸箱的平均抗断强度是否相等、品种改良后小麦的平均亩产量是否显著提高等. 通常来讲，总体的方差均未知，如果通过检验可以认为两个方差相等，一般采用混合样本方差作为总体方差的估计，然后作 t 检验. 直观上混合样本方差综合了两个样本提供的信息应该比采用单个样本方差要好，但是对其优越性，教材中鲜有严格的论证.

本文将对混合样本方差在实际应用中的优越性作出详细说明. 首先，利用统计中的基本定理证明了混合样本方差是总体方差的无偏估计. 其次，注意到样本的密度函数是指数型分布族(其定义参见文献[1])，而此分布族具有许多良好的性质，其中一个重要的性质就是积分的计算与求偏导的运算可以交换次序. 受此性质的启发，证明了混合样本方差是总体方差的一致最小方差无偏估计，即在无偏估计类中，如果以方差最小作为衡量标准，则混合样本方差是最好的.

2 主要结果

假设 $X=(X_1,X_2,\cdots,X_m)$ 与 $Y=(Y_1,Y_2,\cdots,Y_n)$ 分别是来自两个相互独立的正态总体 $N(\mu_1,\sigma^2)$，$N(\mu_2,\sigma^2)$ 的简单随机样本，样本方差分别用 S_1^2，S_2^2 表示，其中

$$S_1^2=\frac{\sum_{i=1}^{m}(X_i-\overline{X})^2}{m-1},\quad S_2^2=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}.$$

用 S_ω^2 表示混合样本方差，其中 $S_\omega^2=\dfrac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$. 下面首先证明 S_ω^2 是 σ^2 的无偏估计.

定理 2.1 S_ω^2 是 σ^2 的无偏估计.

证 由基本定理 $\dfrac{(m-1)S_1^2}{\sigma^2}\sim\chi^2(m-1)$，$\dfrac{(n-1)S_2^2}{\sigma^2}\sim\chi^2(n-1)$，有

$$E\frac{(m-1)S_1^2}{\sigma^2}=m-1,\quad E\frac{(n-1)S_2^2}{\sigma^2}=n-1.$$

[收稿日期] 2015-04-10
[基金项目] 山东省高校科研计划项目(J13LI06)；国家自然科学基金(11401513)

从而

$$ES_\omega^2 = E\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2} = \frac{E(m-1)S_1^2+E(n-1)S_2^2}{m+n-2}$$
$$= \frac{(m-1)\sigma^2+(n-1)\sigma^2}{m+n-2} = \sigma^2,$$

即 S_ω^2 是 σ^2 的无偏估计.

事实上，S_ω^2 是无偏估计中最好的一个，即一致最小方差无偏估计. 为了证明这个结论，首先引入下列判断一个无偏估计是一致最小方差无偏估计的等价条件.

引理 2.2[1] 设 $T(X)$ 是 θ 的无偏估计，$\mathrm{Var}(T(X))<\infty$，则 $T(X)$ 为 θ 的 UMVUE 的充要条件是对 0 的任一无偏估计 $\varphi(X)$，若 $\mathrm{Var}(\varphi(X))<\infty$，则 $\mathrm{cov}(\varphi(X),T(X))=0$.

证 必要性. 设 $\varphi(X)$ 是 0 的任一无偏估计，对任意常数 k，考虑 $T(X)+k\varphi(X)$，由 $T(X)$ 的一致性知

$$\mathrm{Var}(T(X)+k\varphi(X)) = V(\varphi(X))k^2+2\mathrm{cov}(\varphi(X),T(X))k+\mathrm{Var}(T(X)) \geqslant \mathrm{Var}(T(X)),$$

即

$$\mathrm{Var}(\varphi(X))k^2+2\mathrm{cov}(\varphi(X),T(X))k \geqslant 0$$

恒成立. 因此抛物线的顶点 $k=-\dfrac{\mathrm{cov}(\varphi(X),T(X))}{\mathrm{Var}(\varphi(X))}=0$，得 $\mathrm{cov}(\varphi(X),T(X))=0$.

充分性. 设 $\varphi_1(X)$ 是 θ 的任一无偏估计，只需证 $\mathrm{Var}(\varphi_1(X)) \geqslant \mathrm{Var}(T(X))$. 容易知道，$\varphi_1(X)-T(X)$ 是 0 的无偏估计. 从而

$$\mathrm{Var}(\varphi_1(X)) = \mathrm{Var}(\varphi_1(X)-T(X)+T(X))$$
$$= \mathrm{Var}(\varphi_1(X)-T(X))+\mathrm{Var}(T(X))+2\mathrm{cov}(\varphi_1(X)-T(X),T(X))$$
$$= Var(\varphi_1(X)-T(X))+Var(T(X)) \geqslant Var(T(X)).$$

下面是本文的主要定理：

定理 2.3 S_ω^2 是 σ^2 的一致最小方差无偏估计.

证 由引理 2.2 知，只需证明对 0 的任一无偏估计 $\varphi(X,Y)$，均有 $\mathrm{cov}(\varphi(X,Y),S_\omega^2)=0$ 即可，而这显然等价于证明 $E\varphi(X,Y)S_\omega^2=0$. 根据 S_ω^2 的定义只需证明

$$E\Big(\sum_{i=1}^m (X_i-\mu_1)^2+\sum_{i=1}^n (Y_i-\mu_2)^2\Big)\varphi(X,Y)=0, \quad E\bar{X}\varphi(X,Y)=0,$$
$$E\bar{Y}\varphi(X,Y)=0, \quad E\bar{X}^2\varphi(X,Y)=0, \quad E\bar{Y}^2\varphi(X,Y)=0$$

即可.

首先，由 $\varphi(X,Y)$ 是 0 的无偏估计得到，对 $\forall \mu_1,\mu_2 \in \mathbb{R}$，$\sigma^2>0$ 均有

$$E\varphi(X,Y) = \int\cdots\int \varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n = 0, \tag{1}$$

其中 $\forall x=(x_1,x_2,\cdots,x_m)\in\mathbb{R}^m$，$y=(y_1,y_2,\cdots,y_n)\in\mathbb{R}^n$.

$$f(x,y) = (2\pi\sigma^2)^{-\frac{m+n}{2}}\exp\left\{-\frac{\sum_{i=1}^m (x_i-\mu_1)^2+\sum_{i=1}^n (y_i-\mu_2)^2}{2\sigma^2}\right\}.$$

对(1)式两端分别关于 μ_1,μ_2,σ^2 求偏导数，利用指数型分布族的性质，即积分计算与求偏导数运算可以交换次序，整理得到

$$\int\cdots\int\sum_{i=1}^m (x_i-\mu_1)\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n = 0, \tag{2}$$

$$\int\cdots\int\sum_{i=1}^n (y_i-\mu_2)\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n = 0, \tag{3}$$

$$\int\cdots\int\Big(\sum_{i=1}^m (x_i-\mu_1)^2+\sum_{i=1}^n (y_i-\mu_2)^2\Big)\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n = 0. \tag{4}$$

结合(1)－(3)，得到

$$\int\cdots\int\Big(\sum_{i=1}^{m}x_i\Big)\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_n\mathrm{d}y_1\cdots\mathrm{d}y_n=0,\tag{5}$$

$$\int\cdots\int\Big(\sum_{i=1}^{n}y_i\Big)\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n=0,\tag{6}$$

由(4)－(6)得

$$\mathrm{E}\bar{X}\varphi(X,Y)=\mathrm{E}\bar{Y}\varphi(X,Y)=0,\quad \mathrm{E}\Big(\sum_{i=1}^{m}(X_i-\mu_1)^2+\sum_{i=1}^{n}(Y_i-\mu_2)^2\Big)\varphi(X,Y)=0.$$

对(5)，(6)分别关于 μ_1，μ_2 求偏导数，整理得到

$$\int\cdots\int\Big(\sum_{i=1}^{m}x_i\Big)^2\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n=0,\tag{7}$$

$$\int\cdots\int\Big(\sum_{i=1}^{n}y_i\Big)^2\varphi(x,y)f(x,y)\mathrm{d}x_1\cdots\mathrm{d}x_m\mathrm{d}y_1\cdots\mathrm{d}y_n=0,\tag{8}$$

由(7)－(8)得，$\mathrm{E}\bar{X}^2\varphi(X,Y)=\mathrm{E}\bar{Y}^2\varphi(X,Y)=0$．从而

$$\begin{aligned}(m+n-2)\mathrm{E}\varphi(X,Y)S_\omega^2=&\ \mathrm{E}\Big(\sum_{i=1}^{m}(X_i-\mu_1)^2+\sum_{i=1}^{n}(Y_i-\mu_2)^2\Big)\varphi(X,Y)\\&-m\mathrm{E}(\bar{X}-\mu_1)^2\varphi(X,Y)-n\mathrm{E}(\bar{Y}-\mu_2)^2\varphi(X,Y)\\=&-[m\mathrm{E}\bar{X}^2\varphi(X,Y)-2m\mu_1\mathrm{E}\bar{X}\varphi(X,Y)+m\mu_1^2\mathrm{E}\varphi(X,Y)\\&+n\mathrm{E}\bar{Y}^2\varphi(X,Y)-2n\mu_2\mathrm{E}\bar{Y}\varphi(X,Y)+n\mu_2^2\mathrm{E}\varphi(X,Y)]\\=&\ 0,\end{aligned}$$

由引理 2.2 知，S_ω^2 是 σ^2 的一致最小方差无偏估计．

3　模拟结果

对 σ^2 的四种估计 S_1^2，S_2^2，S_3^2，S_ω^2 的样本方差进行了随机模拟，其中

$$S_3^2=\frac{m}{m+n}S_1^2+\frac{n}{m+n}S_2^2.$$

这四个估计中 S_1^2，S_2^2 只利用了单个样本信息，S_3^2，S_ω^2 结合了两个样本提供的信息，不过加权方式不同，S_3^2 直接按样本容量进行加权．

分别从正态总体 N(0.5,2)，N(1,2) 中产生了两个容量为 m，n 的样本，当 m，n 取不同值时，重复进行 10000 次重样，然后算得 S_1^2，S_2^2，S_3^2，S_ω^2 的样本方差．结果如表 1 所示：

表 1　四种估计的偏差与样本方差

样本方差 \ 样本容量	(5,6)	(8,10)	(10,12)	(20,30)
$\hat{\mathrm{D}}(S_1^2)$	1.9678	1.1910	0.9036	0.4202
$\hat{\mathrm{D}}(S_2^2)$	1.5998	0.8735	0.7050	0.2736
$\hat{\mathrm{D}}(S_3^2)$	0.8882	0.5055	0.3901	0.1668
$\hat{\mathrm{D}}(S_\omega^2)$	0.8881	0.5050	0.3899	0.1667

易见，表 1 中的模拟结果与我们的理论结果是相吻合的．具体来讲，结合了两样本信息的 S_3^2，S_ω^2 明显优于只采用单样本信息得到的 S_1^2，S_2^2，尤其是样本容量较小时，S_1^2，S_2^2 的表现较差，例如 $m=5$ 时，$\hat{\mathrm{D}}(S_1^2)=1.9678$，$\hat{\mathrm{D}}(S_1^2)=1.5998$．对不同的加权方式得到的 S_3^2 和 S_ω^2，S_ω^2 优于 S_3^2，例如 $m=8$，$n=10$ 时，$\hat{\mathrm{D}}(S_3^2)=0.5055$，$\hat{\mathrm{D}}(S_\omega^2)=0.5050$．

[参　考　文　献]

[1]　茆诗松，王静龙，濮晓龙．高等数理统计[M]．北京：高等教育出版社，2003.

[2]　Jay L. Devore. Probability and Statistics[M]. Beijing: Higher Education Press, 2008.

The Superiorities of the Combined Sample Variance

CUI Yan-li, *LV Wen*

(School of Mathematics and Information Science, Yantai University, Yantai Shandong 264005, China)

Abstract: For the two normal populations that have the same variances, we always use the combined sample variance to estimate the variance. However, there are little explanations about its superiorities in our teaching books. In this paper we shall prove that it is the UMVUE.

Key words: combined sample variance; unbiased; UMVUE