DOI：10.13718/j.cnki.xsxb.2019.05.007

# 总体均值的分别比率估计方法改进及应用①

乔松珊[1]，　张建军[2]

1. 中原工学院 信息商务学院，郑州 450007；
2. 河南农业大学 信息与管理科学学院，郑州 450002

**摘要**：辅助信息可以在抽样设计和估计量设计两个阶段同时使用，在分层抽样下采用排序集样本代替随机样本，以辅助变量的多个指标线性组合为辅助信息，改进了总体均值的分别比率估计方法，计算了估计量的近似偏差和均方误差，比较了两种不同抽样方法下比率估计的精度. 结果表明，改进的分别比率估计均方误差较小. 最后借助随机模拟和算例分析进一步验证了结论的可靠性.

**关　键　词**：分层排序集抽样；偏斜系数；变异系数；分别比率估计；有效性

**中图分类号**：O212.1　　**文献标志码**：A　　**文章编号**：1000－5471(2019)05－0034－07

辅助信息可以提高参数估计的精度，具体来说，利用辅助信息可以在抽样设计阶段改进抽样方法，获得更高代表性的样本. 在估计量设计阶段，应用辅助变量可以改进估计方法. 为了提高估计精度，可以在两个阶段充分使用辅助信息，这些信息可以是相同的，也可以考虑把多种信息结合起来. 分层比率估计便是抽样设计和估计方法的一种结合. 设 $U$ 为由 $L$ 层组成的研究总体，各层单元数记为 $N_1$，$N_2$，…，$N_L$，并且 $N=\sum_{h=1}^{L}N_h$，$Y$ 为研究变量，$X$ 为辅助变量，$\bar{y}_h$ 与 $\bar{x}_h$ 为两变量在第 $h$ 层的样本均值，$\bar{y}_{RS}$ 为变量 $Y$ 总体均值的分别比率估计，则当辅助变量在第 $h$ 层的总体均值 $\mu_{Xh}$ 已知时，$\bar{y}_{RS}=\sum_{h=1}^{L}W_h\frac{\bar{y}_h}{\bar{x}_h}\mu_{Xh}$，其中 $W_h=\frac{N_h}{N}$ 为第 $h$ 层的层权，$h=1,2,\cdots,L$[1].

除均值外，如果还有辅助变量的其它信息能够利用，这些信息也可以用来提高分别比率估计的精度[2－4]. 在进行总体均值的比率估计时，以上的研究结果采用的都是分层随机抽样，而各层中样本的获取还可以利用排序集抽样方法进行，称为分层排序集抽样(SRSS)[5－9].

研究发现，辅助变量多个信息的结合可以一定程度上提高估计效率，而且基于分层排序集样本的分别比率估计的改进目前少见报道. 鉴于此，为了进一步提高总体均值的估计精度，本文利用辅助变量的多个指标信息，基于分层排序集样本讨论了一种改进的分别比率估计量.

## 1　分层排序集抽样及相关结论

排序集抽样方法最初由 McIntyre 在估计牧草产量时提出[10]，采用方法为：1 次抽取 $r^2$ 个样本，随机地划分为 $r$ 组，对每组样本进行排序，从第 $i$ 组抽取秩为 $i$ 的样本单元并具体测量，记为 $X_{(i)}$，$i=1,2,\cdots,r$，

① 收稿日期：2017－08－12
基金项目：河南省高等学校重点科研项目(17A110023)；河南省软科学研究计划项目(192400410091).
作者简介：乔松珊(1978－)，女，硕士，副教授，主要从事应用数学方面的研究.

类似过程可重复进行. 分层排序集抽样的特点是各层采用排序集抽样方式，而非随机抽样，具体过程为：1 次从二维总体的第 $h$ 层随机抽取容量为 $r_h^2$ 的独立样本，随机划分为 $r_h$ 组，接着按照排序集抽样方式，依据辅助变量 $X$ 对各层进行排序抽样，得到样本量为 $r_h$ 的分层排序集样本，$h=1,2,\cdots,L$. 若 $r=r_1+r_2+\cdots+r_h$，重复上述过程 $m$ 次，得到容量为 $mr$ 的分层排序集样本，记为 $(X_{h(1)k}, Y_{h[1]k})$，$(X_{h(2)k}, Y_{h[2]k})$，$\cdots$，$(X_{h(r_h)k}, Y_{h[r_h]k})$，$k=1,2,\cdots,m$，其中 $X_{h(i)k}$ 为变量 $X$ 在第 $h$ 层秩为 $i$ 的次序统计量，$Y_{h[i]k}$ 为伴随变量，$i=1,2,\cdots,r_h$. 基于分层排序集样本，变量 $X,Y$ 在第 $h$ 层的样本均值为 $\bar{X}_{h(r_h)}=\frac{1}{n_h}\sum_{k=1}^{m}\sum_{i=1}^{r_h}X_{h(i)k}$，$\bar{Y}_{h[r_h]}=\frac{1}{n_h}\sum_{k=1}^{m}\sum_{i=1}^{r_h}Y_{h[i]k}$，其中 $n_h=mr_h$，$h=1,2,\cdots,L$. 设 $\mu_{X_h},\mu_{Y_h},\sigma_{X_h},\sigma_{Y_h}$ 为总体在第 $h$ 层的均值与标准差，并且 $\mu_X=\sum_{h=1}^{L}W_h\mu_{X_h}$. 那么当 $\mu_{X_h}$ 已知时，总体均值 $\mu_Y$ 的分别比率估计量为：

$$\bar{y}_{\mathrm{SRSS}}=\sum_{h=1}^{L}W_h\frac{\bar{Y}_{h[r_h]}}{\bar{X}_{h(r_h)}}\cdot\mu_{X_h} \tag{1}$$

另外，根据文献[6]的研究，容易得到引理 1、引理 2 的结论.

**引理 1**　分层排序集抽样下，比率估计量 $\bar{y}_{\mathrm{SRSS}}$ 的近似均方误差为

$$MSE(\bar{y}_{\mathrm{SRSS}})\approx\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left\{(\sigma_{Y_h}^2+R_h^2\sigma_{X_h}^2-2R_h\rho_{X_hY_h}\sigma_{X_h}\sigma_{Y_h})-\frac{m}{n_h}\sum_{i=1}^{r_h}(T_{Y_{h[i]}}-R_hT_{X_{h(i)}})^2\right\}$$

其中：$\rho_{X_hY_h}$ 为变量 $X$ 与 $Y$ 在第 $h$ 层的相关系数，$T_{X_{h(i)}}=\mu_{X_{h(i)}}-\mu_{X_h}$，$T_{Y_{h[i]}}=\mu_{Y_{h[i]}}-\mu_{Y_h}$，$\mu_{Y_{h[i]}}=E(Y_{h[i]})$，$\mu_{X_{h(i)}}=E(X_{h(i)})$，$R_h=\frac{\mu_{Y_h}}{\mu_{X_h}}$，$T_{X_hY_{h(i)}}=T_{X_{h(i)}}\cdot T_{Y_{h[i]}}$.

**引理 2**　分层排序集抽样下，样本均值 $\bar{X}_{h(r_h)}$，$\bar{Y}_{h[r_h]}$ 的估计方差与协方差为

$$D(\bar{Y}_{h[r_h]})=\frac{\sigma_{Y_h}^2}{mr_h}-\frac{1}{mr_h^2}\sum_{i=1}^{r_h}(\mu_{Y_{h[i]}}-\mu_{Y_h})^2$$

$$D(\bar{X}_{h(r_h)})=\frac{\sigma_{X_h}^2}{mr_h}-\frac{1}{mr_h^2}\sum_{i=1}^{r_h}(\mu_{X_{h(i)}}-\mu_{X_h})^2$$

$$\mathrm{cov}(\bar{X}_{h(r_h)},\bar{Y}_{h[r_h]})=\frac{1}{mr_h^2}(r_h\rho_{X_hY_h}\sigma_{X_h}\sigma_{Y_h}-\sum_{i=1}^{r_h}T_{X_hY_{h(i)}})$$

其中：$\mu_{Y_{h[i]}}=E(Y_{h[i]})$，$\mu_{X_{h(i)}}=E(X_{h(i)})$，$T_{X_hY_{h(i)}}=(\mu_{X_{h(i)}}-\mu_{X_h})(\mu_{Y_{h[i]}}-\mu_{Y_h})$.

## 2　分别比率估计的改进方法及性质

在进行比率估计时，通常采用的是单一的辅助信息，为了进一步提高估计精度，一些学者在分层随机抽样下，尝试利用辅助变量多种信息的结合，记 $\beta_h$ 为变量 $X$ 在第 $h$ 层的峰度系数，$C_{X_h}$ 为变异系数，则当这些系数已知时，可以得到相应的改进分别比率估计形式[2-4]：

$$\bar{y}_{RS}^1=\sum_{h=1}^{L}W_h\bar{y}_h\left(\frac{\mu_{X_h}+C_{X_h}}{\bar{X}_h+C_{X_h}}\right),\bar{y}_{RS}^2=\sum_{h=1}^{L}W_h\bar{y}_h\left(\frac{\mu_{X_h}+\beta_h}{\bar{X}_h+\beta_h}\right),\bar{y}_{RS}^3=\sum_{h=1}^{L}W_h\bar{y}_h\left(\frac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_h+\beta_h}\right)$$

相应估计量的均方误差为：

$$MSE(\bar{y}_{RS}^1)\approx\sum_{h=1}^{L}W_h^2\frac{\mu_{Y_h}^2}{n_h}(C_{Y_h}^2+\lambda_h^2C_{X_h}^2-2\lambda_h\rho_{X_hY_h}C_{X_h}C_{Y_h}) \tag{2}$$

$$MSE(\bar{y}_{RS}^2)\approx\sum_{h=1}^{L}W_h^2\frac{\mu_{Y_h}^2}{n_h}(C_{Y_h}^2+\eta_h^2C_{X_h}^2-2\eta_h\rho_{X_hY_h}C_{X_h}C_{Y_h}) \tag{3}$$

$$MSE(\bar{y}_{RS}^3)\approx\sum_{h=1}^{L}W_h^2\frac{\mu_{Y_h}^2}{n_h}(C_{Y_h}^2+\theta_h^2C_{X_h}^2-2\theta_h\rho_{X_hY_h}C_{X_h}C_{Y_h}) \tag{4}$$

其中：$\mu_{X_h}$，$\mu_{Y_h}$，$\sigma_{X_h}$，$\sigma_{Y_h}$ 为第 $h$ 层的均值与标准差；$C_{X_h}=\dfrac{\sigma_{X_h}}{\mu_{X_h}}$，$C_{Y_h}=\dfrac{\sigma_{Y_h}}{\mu_{Y_h}}$；$\rho_{X_hY_h}$ 为相关系数；$\lambda_h=\dfrac{\mu_{X_h}}{\mu_{X_h}+C_{X_h}}$，$\eta_h=\dfrac{\mu_{X_h}}{\mu_{X_h}+\beta_h}$，$\theta_h=\dfrac{C_{X_h}\mu_{X_h}}{C_{X_h}\mu_{X_h}+\beta_h}$.

为了在抽样设计阶段进一步提高估计效率，文章受到上述比率估计改进方法的启发，尝试在分层排序集抽样下，将辅助变量变异系数和偏斜系数的线性组合做为辅助信息，重点研究如下的改进分别比率估计量：

$$\bar{y}^{p}_{\mathrm{SRSS}}=\sum_{h=1}^{L}W_h\bar{Y}_{h[r_h]}\left\{\frac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right\} \tag{5}$$

其中：$\bar{X}_{h(r_h)}$，$\bar{Y}_{h[r_h]}$ 为排序集抽样下第 $h$ 层的样本均值；$\mu_{X_h}$，$\mu_{Y_h}$ 为总体在第 $h$ 层的均值；$C_{X_h}$ 与 $\beta_h$ 为第 $h$ 层变异系数和偏斜系数；总体均值 $\mu_X=\sum_{h=1}^{L}W_h\mu_{X_h}$，$h=1,2,\cdots,L$. 其它改进形式可类似讨论.

为了比较估计量 $\bar{y}^{p}_{\mathrm{SRSS}}$ 与估计量 $\bar{y}^{3}_{\mathrm{RS}}$ 的估计效果，首先需要分析基于分层排序集样本的改进比率估计的估计无偏性和均方误差. 令 $\bar{Y}_{h[r_h]}=\mu_{Y_h}(1+\delta_0)$，$\bar{X}_{h(r_h)}=\mu_{X_h}(1+\delta_1)$，根据次序统计量的密度函数，容易验证 $E(\bar{X}_{h(r_h)})=\mu_{X_h}$，$E(\bar{Y}_{h[r_h]})=\mu_{Y_h}$，则 $E(\delta_0)=0$，$E(\delta_1)=0$，从而 $E(\delta_0^2)=D(\delta_0)=\dfrac{D(\bar{Y}_{h[r_h]})}{\mu^2_{Y_h}}$，利用引理 2 中的结论，容易得到

$$E(\delta_0^2)=\frac{1}{mr_h}\left[C^2_{Y_h}-\frac{1}{r_h}\sum_{i=1}^{r_h}\left(\frac{T_{Y_{h[i]}}}{\mu_{Y_h}}\right)^2\right] \tag{6}$$

同理

$$E(\delta_1^2)=\frac{1}{mr_h}\left[C^2_{X_h}-\frac{1}{r_h}\sum_{i=1}^{r_h}\left(\frac{T_{X_{h(i)}}}{\mu_{X_h}}\right)^2\right] \tag{7}$$

其中 $C_{X_h}=\dfrac{\sigma_{X_h}}{\mu_{X_h}}$，$T_{X_{h(i)}}=\mu_{X_{h(i)}}-\mu_{X_h}$，$C_{Y_h}=\dfrac{\sigma_{Y_h}}{\mu_{Y_h}}$，$T_{Y_{h[i]}}=\mu_{Y_{h[i]}}-\mu_{Y_h}$.

$$\begin{aligned}\mathrm{cov}(\bar{X}_{h(r_h)},\bar{Y}_{h[r_h]})&=E(\bar{X}_{h(r_h)}\cdot\bar{Y}_{h[r_h]})-E(\bar{X}_{h(r_h)})E(\bar{Y}_{h[r_h]})=\\&\mu_{X_h}\mu_{Y_h}E((1+\delta_0)(1+\delta_1))-\mu_{X_h}\mu_{Y_h}=\\&\mu_{X_h}\mu_{Y_h}E(\delta_1\delta_0)\end{aligned}$$

故 $E(\delta_1\delta_0)=\dfrac{1}{\mu_{X_h}\mu_{Y_h}}\mathrm{cov}(\bar{X}_{h(r_h)},\bar{Y}_{h[r_h]})$，由引理 2 得到

$$E(\delta_1\delta_0)=\frac{1}{mr_h}\rho_{X_hY_h}C_{X_h}C_{Y_h}-\frac{1}{mr_h^2}\sum_{i=1}^{r_h}\frac{T_{X_hY_{h(i)}}}{\mu_{X_h}\mu_{Y_h}} \tag{8}$$

其中：$C_{X_h}=\dfrac{\sigma_{X_h}}{\mu_{X_h}}$，$C_{Y_h}=\dfrac{\sigma_{Y_h}}{\mu_{Y_h}}$，$T_{X_hY_{h(i)}}=(\mu_{X_{h(i)}}-\mu_{X_h})(\mu_{Y_{h[i]}}-\mu_{Y_h})$.

**定理 1** 当 $\theta_h\cdot\delta_1<1$ 时，分层排序集抽样下，总体均值 $\mu_Y$ 的改进比率估计量 $\bar{y}^{P}_{\mathrm{SRSS}}$ 均方误差为：

$$MSE(\bar{y}^{p}_{\mathrm{SRSS}})\approx\sum_{h=1}^{L}W_h^2\frac{\mu^2_{Yh}}{mr_h}\left(C^2_{Y_h}+\theta_h^2C^2_{X_h}-2\theta_h\rho_{X_hY_h}C_{X_h}C_{Y_h}-\frac{1}{r_h}\sum_{i=1}^{r_h}\left(\frac{T_{Y_{h[i]}}}{\mu_{Y_h}}-\theta_h\frac{T_{X_{h(i)}}}{\mu_{X_h}}\right)^2\right)$$

其中：$\theta_h=\dfrac{C_{X_h}\mu_{X_h}}{C_{X_h}\mu_{X_h}+\beta_h}$，$T_{X_{h(i)}}=\mu_{X_{h(i)}}-\mu_{X_h}$，$T_{Y_{h[i]}}=\mu_{Y_{h[i]}}-\mu_{Y_h}$，$T_{X_hY_{h(i)}}=T_{X_{h(i)}}\cdot T_{Y_{h[i]}}$.

**证** $MSE(\bar{y}^{p}_{\mathrm{SRSS}})=MSE\left(\sum_{h=1}^{L}W_h\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)=\sum_{h=1}^{L}W_h^2MSE\left(\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)$

由于 $\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}=\bar{Y}_{h[r_h]}\left(\dfrac{1}{1+\dfrac{C_{X_h}(\bar{X}_{h(r_h)}-\mu_{X_h})}{C_{X_h}\mu_{X_h}+\beta_h}}\right)$

而

$$\frac{C_{X_h}(\bar{X}_{h(r_h)}-\mu_{X_h})}{C_{X_h}\mu_{X_h}+\beta_h}=\frac{C_{X_h}\mu_{X_h}}{C_{X_h}\mu_{X_h}+\beta_h}\cdot\frac{\bar{X}_{h(r_h)}-\mu_{X_h}}{\mu_{X_h}}=\theta_h\delta_1 \tag{9}$$

其中 $\theta_h=\dfrac{C_{X_h}\mu_{X_h}}{C_{X_h}\mu_{X_h}+\beta_h}$，又因为 $\bar{Y}_{h[r_h]}=\mu_{Y_h}(1+\delta_0)$，从而

$$\bar{Y}_{h[r_h]}\frac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}=\mu_{Y_h}(1+\delta_0)\cdot\frac{1}{1+\theta_h\delta_1} \tag{10}$$

当 $\theta_h\cdot\delta_1<1$ 时，$\dfrac{1}{1+\theta_h\delta_1}=1-\theta_h\delta_1+\theta_h^2\delta_1^2+O(\theta_h\delta_1)$，故

$$\mu_{Y_h}(1+\delta_0)\cdot\frac{1}{1+\theta_h\delta_1}=\mu_{Y_h}(1+\delta_0)(1-\theta_h\delta_1+\theta_h^2\delta_1^2+O(\theta_h\delta_1))\approx\mu_{Y_h}(1+\delta_0-\theta_h\delta_1)$$

从而，$MSE\left(\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)=E\left(\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}-\mu_{Y_h}\right)^2\approx\mu_{Y_h}^2E(\delta_0-\theta_h\delta_1)^2$ 即

$$MSE\left(\bar{Y}_{h[r_h]}\frac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)\approx\mu_{Y_h}^2(E(\delta_0^2)+\theta_h^2E(\delta_1^2)-2\theta_hE(\delta_0\delta_1)) \tag{11}$$

将(6),(7),(8)式代入(11)中，得到

$$\begin{aligned}MSE(\bar{y}_{\mathrm{SRSS}}^{p})&=\sum_{h=1}^{L}W_h^2MSE\left(\bar{Y}_{h[r_h]}\frac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)\approx\\&\sum_{h=1}^{L}W_h^2\frac{\mu_{Y_h}^2}{mr_h}(C_{Y_h}^2+\theta_h^2C_{X_h}^2-2\theta_h\rho_{X_hY_h}C_{X_h}C_{Y_h})-\\&\sum_{h=1}^{L}W_h^2\frac{\mu_{Y_h}^2}{mr_h^2}\left(\sum_{i=1}^{r_h}\left(\frac{T_{Y_{h[i]}}}{\mu_{Y_h}}\right)^2-2\theta_h\sum_{i=1}^{r_h}\frac{T_{X_hY_{h(i)}}}{\mu_{X_h}\mu_{Y_h}}+\theta_h^2\sum_{i=1}^{r_h}\left(\frac{T_{X_{h(i)}}}{\mu_{X_h}}\right)^2\right)=\\&\sum_{h=1}^{L}W_h^2\frac{\mu_{Y_h}^2}{mr_h}\left(C_{Y_h}^2+\theta_h^2C_{X_h}^2-2\theta_h\rho_{X_hY_h}C_{X_h}C_{Y_h}-\frac{1}{r_h}\sum_{i=1}^{r_h}\left(\frac{T_{Y_{h[i]}}}{\mu_{Y_h}}-\theta_h\frac{T_{X_{h(i)}}}{\mu_{X_h}}\right)^2\right)\end{aligned}$$

**定理 2**　在一阶泰勒近似下，分别比率估计量 $\bar{y}_{\mathrm{SRSS}}^{P}$ 的估计偏差为

$$Bias(\bar{y}_{\mathrm{SRSS}}^{p})\approx\sum_{h=1}^{L}W_h\frac{\mu_{Y_h}}{mr_h}\left((\theta_h^2C_{X_h}^2-\theta_h\rho_{X_hY_h}C_{X_h}C_{Y_h})-\frac{1}{r_h}\left(\theta_h^2\sum_{i=1}^{r_h}\left(\frac{T_{X_{h(i)}}}{\mu_{X_h}}\right)^2-\theta_h\sum_{i=1}^{r_h}\frac{T_{X_hY_{h(i)}}}{\mu_{X_h}\mu_{Y_h}}\right)\right)$$

其中：$T_{X_{h(i)}}=\mu_{X_{h(i)}}-\mu_{X_h}$，$T_{Y_{h[i]}}=\mu_{Y_{h[i]}}-\mu_{Y_h}$，$T_{X_hY_{h(i)}}=T_{X_{h(i)}}T_{Y_{h[i]}}$，$i=1,2,\cdots,r_h$.

**证**　根据定义 $\bar{y}_{\mathrm{SRSS}}^{p}=\sum_{h=1}^{L}W_h\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}$，由定理 1 中(10)式可知，$\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}=\mu_{Y_h}\dfrac{1+\delta_0}{1+\theta_h\delta_1}$，当 $\theta_h\cdot\delta_1<1$ 时，利用泰勒展开式，得到

$$\mu_{Y_h}(1+\delta_0)\cdot\frac{1}{1+\theta_h\delta_1}=\mu_{Y_h}(1+\delta_0)(1-\theta_h\delta_1+\theta_h^2\delta_1^2+O(\theta_h\delta_1))$$

故 $Bias(\bar{y}_{\mathrm{SRSS}}^{p})=E\left(\sum_{h=1}^{L}W_h\bar{Y}_{h[r_h]}\dfrac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)-\mu_Y$，由于 $\mu_Y=\sum_{h=1}^{L}W_h\mu_{Y_h}$，故

$$Bias(\bar{y}_{SRSS}^{p})=\sum_{h=1}^{L}W_h\left(E\left(\bar{Y}_{h[r_h]}\frac{C_{X_h}\mu_{X_h}+\beta_h}{C_{X_h}\bar{X}_{h(r_h)}+\beta_h}\right)-\mu_{Y_h}\right)\approx\sum_{h=1}^{L}W_h(\mu_{Y_h}(\theta_h^2E(\delta_1^2)-\theta_hE(\delta_0\delta_1)))$$

利用(7),(8)式容易得到

$$Bias(\bar{y}_{\mathrm{SRSS}}^{p})\approx\sum_{h=1}^{L}W_h\frac{\mu_{Y_h}}{mr_h}\left((\theta_h^2C_{X_h}^2-\theta_h\rho_{X_hY_h}C_{X_h}C_{Y_h})-\frac{1}{r_h}\left(\theta_h^2\sum_{i=1}^{r_h}\left(\frac{T_{X_{h(i)}}}{\mu_{X_h}}\right)^2-\theta_h\sum_{i=1}^{r_h}\frac{T_{X_hY_{h(i)}}}{\mu_{X_h}\mu_{Y_h}}\right)\right)$$

其中：$T_{X_{h(i)}}=\mu_{X_{h(i)}}-\mu_{X_h}$，$T_{Y_{h[i]}}=\mu_{Y_{h[i]}}-\mu_{Y_h}$，$T_{X_hY_{h(i)}}=T_{X_{h(i)}}T_{Y_{h[i]}}$，$i=1,2,\cdots,r_h$.

定理 2 的结论表明，当循环次数 $m$ 足够大时，$\bar{y}_{\mathrm{SRSS}}^{p}$ 仍然为总体均值 $\mu_Y$ 的近似无偏估计量.

## 3 改进比率估计量的有效性

均方误差是衡量估计量有效性的重要标准，下面比较分层随机抽样下估计量 $\bar{y}_{\mathrm{RS}}^{3}$ 与分层排序集抽样下估计量 $\bar{y}_{\mathrm{SRSS}}^{p}$ 的均方误差.

**定理 3** 如果各层样本容量相同，即 $n_h = mr_h$ 时，分层随机抽样和分层排序集抽样下，总体均值的两种比率估计量 $\bar{y}_{\mathrm{RS}}^{3}$ 与 $\bar{y}_{\mathrm{SRSS}}^{p}$ 满足关系：$MSE(\bar{y}_{\mathrm{SRSS}}^{p}) \leqslant MSE(\bar{y}_{\mathrm{RS}}^{3})$.

**证** 令 $\omega_{X_{h(i)}} = \dfrac{T_{X_{h(i)}}}{\mu_{X_h}}$，$\omega_{Y_{h[i]}} = \dfrac{T_{Y_{h[i]}}}{\mu_{Y_h}}$，因为 $T_{X_hY_{h(i)}} = T_{X_{h(i)}} \cdot T_{Y_{h[i]}}$，从而 $\dfrac{T_{X_hY_{h(i)}}}{\mu_{X_h}\mu_{Y_h}} = \dfrac{T_{X_{h(i)}}}{\mu_{X_h}} \dfrac{T_{Y_{h[i]}}}{\mu_{Y_h}} = \omega_{X_{h(i)}} \cdot \omega_{Y_{h[i]}}$，$i = 1,2,\cdots,r_h$. 故

$$MSE(\bar{y}_{\mathrm{SRSS}}^{p}) = \sum_{h=1}^{L} W_h^2 \frac{\mu_{Y_h}^2}{mr_h}(C_{Y_h}^2 + \theta_h^2 C_{X_h}^2 - 2\theta_h \rho_{X_hY_h} C_{X_h} C_{Y_h}) - \sum_{h=1}^{L} W_h^2 \frac{\mu_{Y_h}^2}{mr_h^2} \sum_{i=1}^{r_h} (\omega_{Y_{h[i]}} - \theta_h \omega_{X_{h(i)}})^2$$

当 $n_h = mr_h$ 时，利用式(4)容易得到

$$MSE(\bar{y}_{\mathrm{RS}}^{3}) - MSE(\bar{y}_{\mathrm{SRSS}}^{P}) = \sum_{h=1}^{L} W_h^2 \frac{\mu_{Y_h}^2}{mr_h^2} \sum_{i=1}^{r_h} (\omega_{Y_{h[i]}} - \theta_h \omega_{X_{h(i)}})^2$$

由于 $\sum_{h=1}^{L} W_h^2 \dfrac{\mu_{Y_h}^2}{mr_h^2} \sum_{i=1}^{r_h} (\omega_{Y_{h[i]}} - \theta_h \omega_{X_{h(i)}})^2 \geqslant 0$，从而 $MSE(\bar{y}_{\mathrm{SRSS}}^{p}) \leqslant MSE(\bar{y}_{\mathrm{RS}}^{3})$.

定理 3 表明，利用排序集样本代替随机样本的分别比率估计有效地降低了估计的均方误差，也即基于分层排序集样本的分别比率估计效率要高于分层随机抽样下的估计效率.

## 4 随机模拟与算例分析

假定分层随机抽样和分层排序集抽样方法在每一层中样本容量和抽样比均相同，为了便于比较，只需研究一层的抽样估计结果. 选取研究总体为二维正态分布 $N(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$，相关系数分 3 种情形 $\rho = 0.80,\rho = 0.90,\rho = 0.99$ 分别讨论，不妨令 $\mu_X = 2,\mu_Y = 4,\sigma_X = \sigma_Y = 1$. 首先基于 matlab 软件分别生成 5 000 个二维随机数，样本容量 $n$ 分别取为 30,60,90，进行排序集抽样时 $r_h = 3$，循环次数 $m = 10,20,30$，采用随机抽样和排序集抽样两种抽样方法进行比较，利用 R 软件进行 100 次的统计模拟，均方误差计算公式为 $MSE(\hat{\mu}) = \dfrac{1}{l}\sum_{i=1}^{l}(\hat{\mu}_i - \mu)^2$，其中：$l$ 为模拟次数，$\mu$ 为参数真值，$\hat{\mu}_i$ 为第 $i$ 次估计值. 具体计算结果如表 1 所示.

**表 1 3 种比率估计量的估计结果**

| $\rho$ | 样本量 | $\bar{y}_{\mathrm{RS}}$ | | $\bar{y}_{\mathrm{RS}}^{3}$ | | $\bar{y}_{\mathrm{SRSS}}^{p}$ | |
|---|---|---|---|---|---|---|---|
| | | 估计值 | 均方误差 | 估计值 | 均方误差 | 估计值 | 均方误差 |
| 0.80 | 30 | 4.017 | 0.072 04 | 4.088 | 0.288 3 | 3.997 | 0.001 48 |
| | 60 | 3.991 | 0.024 75 | 3.937 | 0.212 1 | 3.998 | 0.000 55 |
| | 90 | 3.974 | 0.022 61 | 3.981 | 0.027 9 | 3.972 | 0.000 93 |
| 0.90 | 30 | 4.018 | 0.057 97 | 4.013 | 0.225 8 | 4.019 | 0.000 82 |
| | 60 | 3.971 | 0.022 73 | 4.033 | 0.116 5 | 4.015 | 0.000 53 |
| | 90 | 3.981 | 0.015 71 | 3.996 | 0.025 8 | 3.988 | 0.000 23 |
| 0.99 | 30 | 3.963 | 0.033 69 | 4.039 | 0.288 4 | 4.017 | 0.000 64 |
| | 60 | 4.022 | 0.021 07 | 3.944 | 0.055 0 | 4.015 | 0.002 23 |
| | 90 | 3.989 | 0.009 69 | 3.998 | 0.017 1 | 4.012 | 0.000 31 |

由表 1 的计算结果容易看出，对于正态分布，随机抽样下利用辅助变量的变异系数和偏度系数改进估计量的效果并不显著，而以排序集样本代替随机抽样后的分别比率估计均方误差明显下降，说明估计精度的提高主要是抽样设计阶段充分利用辅助信息的结果. 其它总体分布情况可做类似研究.

随机模拟假设总体是正态分布，为了进一步说明结论的可靠性，文章以 77 棵杨树第 11 年生长数据为研究对象[11]，进行实例分析，具体数据见表 2. 选取树木胸径为辅助变量，树木高度为研究变量，基于 3 种方法估计树木高度的总体均值. 样本容量分别设定为 15,30,60. 由于排序集抽样 $n_h = mr_h$，考虑到实际应用中，排序集容量过大容易造成误差，故选取 $r_h = 3$，循环次数 $m$ 分别为 5,10,20. 为了提高计算可信度，利用 $R$ 软件重复计算 100 次，得到估计均方误差，计算结果见表 3.

**表 2　杨树的树高和胸径**

| 胸径/cm | 树高/m | 胸径/cm | 树高/m | 胸径/cm | 树高/m | 胸径/cm | 树高/m | 胸径/cm | 树高/m | 胸径/cm | 树高/m | 胸径/cm | 树高/m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16.5 | 24.2 | 15.8 | 19.6 | 10.8 | 17.2 | 18 | 31.2 | 18 | 27.7 | 14.1 | 22.6 | 18.6 | 27 |
| 19.1 | 25.8 | 11.5 | 17.9 | 22.1 | 33.1 | 14.9 | 20.7 | 17.5 | 29 | 18 | 23.2 | 18.6 | 26.7 |
| 16 | 22.2 | 16.6 | 23.2 | 20.6 | 32.5 | 14.9 | 21 | 15 | 25.5 | 16.5 | 22.6 | 16 | 20.7 |
| 21 | 28.5 | 15.8 | 22.6 | 16.5 | 27 | 14 | 17.9 | 20.5 | 32.1 | 20 | 25.1 | 15.9 | 22 |
| 18 | 24.7 | 16.8 | 26.1 | 19 | 30.7 | 17.6 | 26.1 | 16.5 | 24.5 | 19 | 38.3 | 18 | 28.3 |
| 20.2 | 30.6 | 17.2 | 25.5 | 18 | 25.1 | 14.6 | 22.9 | 17 | 19.1 | 14.6 | 20.1 | 18.9 | 23.2 |
| 17.8 | 28.4 | 21 | 27.7 | 13.8 | 22 | 20.5 | 26.1 | 19 | 20.4 | 17 | 23.9 | 16 | 22.6 |
| 18.6 | 27.7 | 17.5 | 21.2 | 13.9 | 22 | 18 | 27.1 | 16.5 | 22.5 | 13.9 | 17.5 | 18.4 | 22.9 |
| 18.5 | 25.1 | 19.5 | 25.5 | 18.3 | 25.9 | 22 | 31.8 | 17 | 29.1 | 17.3 | 29.3 | 21 | 32.5 |
| 18.6 | 22.9 | 18 | 22.9 | 20 | 30.3 | 18.6 | 23.9 | 17.5 | 27.3 | 20.5 | 28.6 | 15.3 | 22.6 |
| 20 | 28.6 | 19 | 27.3 | 17 | 22.9 | 16.8 | 22.3 | 21.5 | 30.6 | 17.1 | 27.7 | 19 | 34.7 |

**表 3　不同比率估计量的估计结果比较**

| 样本量 | $\bar{y}_{RS}$ | | $\bar{y}^{3}_{RS}$ | | $\bar{y}^{p}_{SRSS}$ | |
|---|---|---|---|---|---|---|
| | 估计值 | 均方误差 | 估计值 | 均方误差 | 估计值 | 均方误差 |
| 15 | 25.445 | 0.495 | 25.413 | 0.737 | 25.552 | 0.045 |
| 30 | 25.485 | 0.208 | 25.351 | 0.217 | 25.538 | 0.019 |
| 60 | 25.447 | 0.035 | 25.424 | 0.032 | 25.479 | 0.006 |

表 3 的计算结果表明：第一，无论采用哪一种抽样方法，均方误差都随着样本量的增加不断变小，排序集抽样方法变化更加明显；第二，与分层随机抽样下改进比率估计量相比，基于分层排序集样本和多指标线性组合的改进比率估计的精度进一步提高，特别是估计均方误差明显降低，也进一步说明改进比率估计量的有效性.

## 5　结　　论

分层排序集抽样结合了分层抽样和排序集抽样的优点，基于该抽样方法可以建立多种比率估计形式. 文章同时考虑辅助变量变异系数和偏斜系数，以两者的线性组合作为辅助信息，构造了改进的分别比率估计模型，并进一步研究了估计量的偏差和均方误差. 最后，借助实际例子做了进一步分析，验证了估计量的有效性.

**参考文献：**

[1] 杜子芳. 抽样技术及其应用 [M]. 北京：清华大学出版社，2005：133-137.

[2] KADILAR C，CINGI H. Ratio Estimators in Stratified Random Sampling [J]. Biometrical Journal，2003，45(2)：218-225.

[3] TAILOR R，CHOUHAN S. Ratio Type Estimator of Ratio of Two Population Means in Stratified Random Sampling [J]. Journal of Modern Applied Statistical Methods，2012，11(1)：279-283.

[4] YAN Z，TIAN B. Ratio Method to the Mean Estimation Using Coefficient of Skewness of Auxiliary Variable [M]// Information Computing and Applications. Berlin：Springer，2010.

[5] SAMAWI H M. Stratified Ranked Set Sample [J]. Pakistan Journal of Statistics，1996，12(1)：9-16.

[6] SAMAWI H M, SIAM M I. Ratio Estimation Using Stratified Ranked Set Sample [J]. Metron- International Journal of Statistics, 2003, LXI(1): 75-90.

[7] 张建军, 乔松珊. 基于分层排序集抽样方法的改进比率估计 [J]. 华中师范大学学报(自然科学版), 2015, 49(6): 816-821.

[8] MANDOWARRA V L , MEHTA N M. Modifiedratio Estimators Using Stratified Ranked Set Sampling [J]. Hacettepe Journal of Mathematics and Statistics, 2014, 43 (3): 461-471.

[9] KHAN L, SHABBIR J, GUPTA S. Unbiased Ratio Estimators of the Mean in Stratified Ranked Set Sampling [J]. Hacettepe Journal of Mathematics and Statistics, 2016, 46(108): 1-11.

[10] MCINTYRE G A. A Method for Unbiased Selective Sampling, Using Ranked Sets [J]. The American Statistician, 2005, 59(3): 230-232.

[11] ZHAO W, HOU W, LITTELL R C, et al. Structured Antedependence Models for Functional Mapping of Multiple Longitudinal Traits [J]. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1): 1-29.

# Modified Separate Ratio Estimators of Population Mean and its Application

QIAO Song-shan[1], ZHANG Jian-jun[2]

1. *College of Information and Business Zhongyuan Institute of Technology, Zhengzhou, Henan 450007, China;*

2. *Collage of Information and Management Science of Henan Agricultural University, Zhengzhou, Hennan 450002, China*

**Abstract**: Auxiliary information can be used in sampling design and estimation design. In this paper, with ranked set samples instead of the random samples in stratified sampling, we have proposed the new separate ratio estimators of population mean based on linear Combination of Multiple Indexes of Auxiliary Variable. We have obtained the bias and mean squared error of the proposed estimators and compared the estimated accuracy of new separate ratio estimators with traditional separate ratio estimators. The results are supported by stochastic simulation and numerical example.

**Key words**: stratified ranked set sampling; coefficient of skewness; coefficient of variation; separate ratio estimation; efficiency

**责任编辑** 张 枸