分层三阶段抽样样本大小的研究及应用*

高 歌[1] 金丕焕[2] 王旭辉[3]

【提 要】 **目的** 为调查设计中常用的分层三阶段随机抽样方法寻求其样本大小估计公式。**方法** 利用微积分学中求极小值的方法。**结果** 当采用分层三阶段随机抽样作参数估计时，在限定抽样误差的大小使调查花费达到最小及限定调查花费的大小使抽样误差达到最小两种情况下，推导出其最优样本大小的计算公式。**结论** 推导出分层三阶段随机抽样样本大小的估计公式，在中国铁路职工医疗费用的抽样调查中取得了成功的应用效果。

【关键词】 调查设计 分层抽样 三阶段抽样 样本大小

抽样调查是医学科研及卫生工作中常用的主要调查研究方法。抽样调查必须有一个科学的抽样设计，确定样本大小是抽样设计的关键之一。统计学家提出了不少的抽样方法，但对多数复杂的抽样方法并没有提供相应的样本大小估计公式。本文对重要的尤其适用于大规模调查的分层三阶段随机抽样调查方法，从数学上推导出其参数估计时的最优样本大小计算公式。

公式推导

一、分层三阶段抽样的概念

假定：被抽总体划分成若干层，第 h 层包含 N_h 个个体，由 N_{1h} 个群组成，h 层第 i 群由 N_{i2h} 个小群组成，h 层平均每一群包含 $\overline{N}_{2h}$ 个小群；h 层第 i 群第 j 小群由 N_{ij3h} 个个体组成，h 层平均每一小群包含 $\overline{N}_{3h}$ 个个体。为估计总体均数，假定：第一阶段从 h 层随机抽取 n_{1h} 群；第二阶段从 h 层第 i 个抽中群内随机抽取 n_{i2h} 个小群，平均从 h 层每个抽中群内抽取了 $\overline{n}_{2h}$ 个小群；第三阶段从 h 层第 i 个抽中群第 j 个抽中小群随机抽取 n_{ij3h} 个个体，平均从 h 层每个抽中的小群内抽取了 $\overline{n}_{3h}$ 个个体，y_{ijkh} 为 h 层第 i 群第 j 小群第 k 个个体的指标值，$\overline{X}_{ijh}$ 为 h 层第 i 群第 j 小群的样本均数，$\overline{X}_{ih}$ 为 h 层第 i 群的样本均数，$\overline{X}_h$ 为 h 层的样本均数，$\overline{X}$ 为被抽总体的样本均数。

对三阶段抽样，当各群、各小群大小相等时，Cochran W.G. 已给出样本均数的方差计算公式，根据此结果，不难推导出（各群、各小群大小不等时）分层三阶段抽样样本均数 $\overline{X}$ 的方差 $V(\overline{X})$ 为：

$$V(\overline{X})=\sum_h W_h^2\left[\frac{\sigma_{1h}^2}{n_{1h}}\left(1-\frac{n_{1h}}{N_{1h}}\right)+\frac{\sigma_{2h}^2}{n_{1h}\overline{n}_{2h}}\left(1-\frac{\overline{n}_{2h}}{\overline{N}_{2h}}\right)+\frac{\sigma_{3h}^2}{n_{1h}\overline{n}_{2h}\overline{n}_{3h}}\left(1-\frac{\overline{n}_{3h}}{\overline{N}_{3h}}\right)\right],\left(W_h=\frac{N_h}{\sum_h N_h}\right) \quad (1)$$

其中 σ_{1h}^2 的样本估计量

$$S_{1h}^2=\frac{1}{n_{1h}-1}\sum_{i=1}^{n_{1h}}(\overline{X}_{ih}-\overline{X}_h)^2 \quad (2)$$

σ_{2h}^2 的样本估计量

$$S_{2h}^2=\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{1}{n_{i2h}-1}\sum_{j=1}^{n_{i2h}}(\overline{X}_{ijh}-\overline{X}_{ih})^2 \quad (3)$$

σ_{3h}^2 的样本估计量

$$S_{3h}^2=\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{1}{n_{i2h}-1}\sum_{j=1}^{n_{i2h}}\frac{1}{n_{ij3h}-1}\sum_{k=1}^{n_{ij3h}}(y_{ijkh}-\overline{X}_{ijh})^2 \quad (4)$$

二、估计总体均数时的最优样本大小

C 为调查的总花费，C_{0h} 为调查 h 层的基本花费，C_{1h} 为 h 层每调查一个群的平均基本花费，C_{2h} 为 h 层每调查一个小群的平均基本花费，C_{3h} 为 h 层每调查一个个体的平均直接花费。得花费函数：

$$C=\sum_h C_{0h}+\sum_h C_{1h}n_{1h}+\sum_h C_{2h}n_{1h}\overline{n}_{2h}+\sum_h C_{3h}n_{1h}\overline{n}_{2h}\overline{n}_{3h} \quad (5)$$

1. 限定 $V(\overline{X})$ 使 C 达到最小

将(1)式变形得：

$$V(\overline{X})=\sum_h W_h^2\left[\frac{1}{n_{1h}}(\sigma_{1h}^2-\sigma_{2h}^2/\overline{N}_{2h})+\frac{1}{n_{1h}\overline{n}_{2h}}(\sigma_{2h}^2-\sigma_{3h}^2/\overline{N}_{3h})+\frac{\sigma_{3h}^2}{n_{1h}\overline{n}_{2h}\overline{n}_{3h}}-\frac{\sigma_{1h}^2}{N_{1h}}\right] \quad (6)$$

限定 $V(\overline{X})$ 的值为 V，使调查花费 C 达到最小的最优样本大小就是三元函数(5)在约束方程(6)之下的

* 铁道部科技基金资助项目（编号 B99(6)）

1. 同济大学医学院预防医学教研室（200070）
2. 复旦大学卫生统计教研室（200031）
3. 同济大学医学院医学二系（200065）

Chinese Journal of Health Statistics, December 2000, Vol. 17, No. 6

条件极小值点。

令 $F=\sum_h C_{0h}+\sum_h C_{1h}n_{1h}+\sum_h C_{2h}n_{1h}\bar{n}_{2h}+\sum_h C_{3h}n_{1h}\bar{n}_{2h}\bar{n}_{3h}+\lambda(V(\bar{X})-V)$

得

$$\begin{cases}\dfrac{\partial F}{\partial n_{1h}}=C_{1h}-\dfrac{\lambda W_h^2(\sigma_{1h}^2-\sigma_{2h}^2/\bar{N}_{2h})}{n_{1h}^2}=0\\ \dfrac{\partial F}{\partial n_{1h}\bar{n}_{2h}}=C_{2h}-\dfrac{\lambda W_h^2(\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h})}{n_{1h}^2\bar{n}_{2h}^2}=0\\ \dfrac{\partial F}{\partial n_{1h}\bar{n}_{2h}\bar{n}_{3h}}=C_{3h}-\dfrac{\lambda W_h^2\sigma_{3h}^2}{n_{1h}^2\bar{n}_{2h}^2\bar{n}_{3h}^2}=0\end{cases}\tag{7}$$

解得

$$\begin{cases}\bar{n}_{2h}=\sqrt{\dfrac{\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h}}{\sigma_{1h}^2-\sigma_{2h}^2/\bar{N}_{2h}}\cdot\dfrac{C_{1h}}{C_{2h}}}&(8)\\ \bar{n}_{3h}=\sqrt{\dfrac{\sigma_{3h}^2}{\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h}}\cdot\dfrac{C_{2h}}{C_{3h}}}&(9)\\ n_{1h}=\dfrac{W_h\sigma_{3h}\sqrt{\lambda}}{\bar{n}_{2h}\bar{n}_{3h}\sqrt{C_{3h}}}&(10)\end{cases}$$

将(7)式代入(6)式,且 $V(\bar{X})=V$,得

$$\sqrt{\lambda}=\frac{\sum_h W_h\sqrt{C_{3h}}/\sigma_{3h}\cdot[\bar{n}_{2h}\bar{n}_{3h}(\sigma_{1h}^2-\sigma_{2h}^2/\bar{N}_{2h})+\bar{n}_{3h}(\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h})+\sigma_{3h}^2]}{V+\sum_h W_h^2\sigma_{1h}^2)/N_{1h}}\tag{11}$$

2. 限定 C 使 $V(\bar{X})$达到最小

当限定调查的花费值 C 使 $V(\bar{X})$达到最小的最优样本大小就是函数(6)在约束方程(5)之下的条件极小值点。

令 $F=V(\bar{X})+\lambda(\sum_h C_{0h}+\sum_h C_{1h}n_{1h}+\sum_h C_{2h}n_{1h}\bar{n}_{2h}+\sum_h C_{3h}n_{1h}\bar{n}_{2h}\bar{n}_{3h}-C)$

得

$$\begin{cases}\dfrac{\partial F}{\partial n_{1h}}=-\dfrac{W_h^2(\sigma_{1h}^2-\sigma_{2h}^2/\bar{N}_{2h})}{n_{1h}^2}+\lambda C_{1h}=0&(12)\\ \dfrac{\partial F}{\partial n_{1h}\bar{n}_{2h}}=\dfrac{-W_h^2(\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h})}{n_{1h}^2\bar{n}_{2h}^2}+\lambda C_{2h}=0&(13)\\ \dfrac{\partial F}{\partial n_{1h}\bar{n}_{2h}\bar{n}_{3h}}=\dfrac{-W_h^2\sigma_{3h}^2}{n_{1h}^2\bar{n}_{2h}^2\bar{n}_{3h}^2}+\lambda C_{3h}=0&(14)\end{cases}$$

解得

$$\begin{cases}\bar{n}_{2h}=\sqrt{\dfrac{\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h}}{\sigma_{1h}^2-\sigma_{2h}^2/\bar{N}_{2h}}\cdot\dfrac{C_{1h}}{C_{2h}}}\\ \bar{n}_{3h}=\sqrt{\dfrac{\sigma_{3h}^2}{\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h}}\cdot\dfrac{C_{2h}}{C_{3h}}}\\ n_{1h}=\dfrac{W_h\sigma_{3h}}{\bar{n}_{2h}\bar{n}_{3h}\sqrt{C_{3h}}}\cdot\dfrac{1}{\sqrt{\lambda}}\end{cases}$$

将(12)～(14)式代入(5),得

$$\frac{1}{\sqrt{\lambda}}=\frac{C-\sum_h C_{0h}}{\sum_h W_h(\sqrt{C_{1h}}\cdot\sqrt{\sigma_{1h}^2-\sigma_{2h}^2/\bar{N}_{2h}}+\sqrt{C_{2h}}\cdot\sqrt{\sigma_{2h}^2-\sigma_{3h}^2/\bar{N}_{3h}}+\sqrt{C_{3h}}\cdot\sigma_{3h})}$$

三、估计总体率时最优样本大小

估计总体率时,在限定抽样误差使调查花费最小及限定调查花费使抽样误差最小的最优样本大小计算公式及推导过程,与估计总体均数时完全相同,p_h 为 h 层的样本率,p_{ih} 为 h 层第 i 个抽中群的样本率,p_{ijh} 为 h 层第 i 群第 j 小群的样本率。此时,

$$S_{1h}^2=\frac{1}{n_{1h}-1}\sum_{i=1}^{n_{1h}}(p_{ih}-p_h)^2$$

$$S_{2h}^2=\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{1}{n_{i2h}-1}\sum_{j=1}^{n_{i2h}}(p_{ijh}-p_{ih})^2$$

$$S_{3h}^2=\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{1}{n_{i2h}}\sum_{j=1}^{n_{i2h}}p_{ijh}(1-p_{ijh})$$

应用实例

一、实例分析

为了制定中国铁路系统卫生机构与企业主体分立的实施方案,需抽样调查估计中国铁路职工 1999 年的人均医疗费用。采用分层三阶段抽样,将 45 岁以下的职工(占 57%)作为第一层,将 45 岁以上(含 45 岁)的职工(占 43%)作为第二层,故 $W_1=0.57$, $W_2=0.43$。全国共有 14($N_{11}=N_{12}=14$)个铁路局(群),平均每个铁路局有下属单位(小群)293 个($\bar{N}_{21}=\bar{N}_{22}=293$),平均每个单位 45 岁以下职工 226 人($\bar{N}_{31}=226$),45 岁以上(含 45 岁)职工 171 人($\bar{N}_{32}=171$)。调查每一个铁路局的平均基本花费 $C_{11}=C_{12}=5\ 500$ 元,调查每一个单位的平均基本花费 $C_{21}=C_{22}=500$ 元,调查每一名职工的平均直接花费 $C_{31}=C_{32}=5$ 元。限定 $V(\bar{X})$的值为 17,为使调查总花费最小,试估计最优样本大小。

二、计算步骤

1. 按公式(2)～(4),利用过去的医疗费用资料,估计得(过程从略):$S_{11}^2=32$, $S_{12}^2=41$, $S_{21}^2=279$, $S_{22}^2=316$, $S_{31}^2=19\ 367$, $S_{32}^2=21\ 048$。

2. 按公式(8)、(9)计算得(σ_{ih}^2用 S_{ih}^2代替):

$$\bar{n}_{21}=\sqrt{\frac{279-19367/226}{32-279/293}\cdot\frac{5500}{500}}\approx 8$$

$$\bar{n}_{22}=\sqrt{\frac{316-21048/171}{41-316/293}\cdot\frac{5500}{500}}\approx 7$$

$$\bar{n}_{31}=\sqrt{\frac{19367}{279-19367/226}\cdot\frac{500}{5}}\approx 100$$

$$\bar{n}_{32}=\sqrt{\frac{21048}{316-21048/171}\cdot\frac{500}{5}}\approx 104$$

3. 按公式(11)、(10)计算得(σ_{ih}^2用S_{ih}^2代替):

$$\sqrt{\lambda}=\{0.57\sqrt{5}\div\sqrt{19367}[8\times 100(32-279/293)+100(279-19367/226)+19367]+0.43\sqrt{5}\div\sqrt{21048}[7\times 104(41-316/293)+104(316-21048/171)+21048]\}/[17+0.57^2\times 32/14+0.43^2\times 41/14]$$

$$\approx 48$$

$$n_{11}=\frac{0.57\sqrt{19367}\times 48}{8\times 100\sqrt{5}}\approx 2$$

$$n_{12}=\frac{0.43\sqrt{21048}\times 48}{7\times 104\sqrt{5}}\approx 2$$

4. h 层第 i 个抽中铁路局内抽取的单位个数可由公式 $n_{i2h}=N_{i2h}\cdot\bar{n}_{2h}/\bar{N}_{2h}$ 来估计;h 层第 i 个抽中铁路局第 j 个抽中单位抽取的职工人数可由公式 $n_{ij3h}=N_{ij3h}\cdot\bar{n}_{3h}/\bar{N}_{3h}$ 来估计。

讨　论

1. 本文从数学上推导出分层三阶段随机抽样样本大小的估计公式,方法科学、公式简便。并采用分层三阶段随机抽样方法,使用本文给出的公式,计算出各层各阶段的最优样本大小,对全国铁路职工 1999 年的人均医疗费用作出了科学的估计,为制定《中国铁路系统卫生机构与企业主体分立的实施方案》提供了重要的科学依据。

2. 在分层三阶段抽样中,调查一个群的基本花费要远高于调查一个小群的基本花费,调查一个小群的基本花费要远高于调查一个个体的直接花费。一般来讲,当少抽一些群,而在各抽中群内多抽一些小群,在各抽中小群多抽一些个体,会使抽样误差增大而调查花费减少;当多抽一些群,而在各抽中群少抽一些小群,在各抽中小群少抽一些个体,会使抽样误差减少而调查花费增大。本文正是为了解决这一矛盾,在限定抽样误差的大小使调查花费达到最小及限定调查花费的大小使抽样误差达到最小的两种情况下,求出了各阶段最优样本大小计算公式。

3. 为了确定 $V(\bar{X})$ 的限定值 V,当 $\bar{X}$ 或 p 呈正态分布时,可按公式 $V=(\delta/U_\alpha)^2$ 来估计,δ 代表容许误差,α 表示作$(1-\alpha)$%的区间估计,如本文所举实例中,$\delta=8$, $\alpha=0.05$,则 $V=(8/1.96)^2\approx 17$。当 $\bar{X}$ 或 p 不呈正态分布时,也可参照上述公式或凭经验来近似确定 V 值。

4. 作者对多阶段抽样方法作过系列研究,已对常用的二阶段抽样、分层二阶段抽样、三阶段抽样方法推导出其样本大小计算公式,见参考文献。

The Method to Estimate the Sample Sizes for Stratified Three-Stage Sampling and its Application *Gao Ge, Jin Pihuan, Wang Xuhui. Medical College Shanghai Tong Ji University* (200070), *Shanghai*

【Abstract】 **Objective** To deduce the formulate of the sample sizes for stratified three-stage sampling. **Methods** Method of deducing minimum value in differential and integral calculus was applied. **Results** The authors deduced the formulate of the optimum sample sizes to minimize the costs of taking the samples for specified values of the sampling errors and to minimize the sampling errors for specified costs for stratified three-stage sampling. **Conclusion** The optimum sample sizes were deduced for stratified three-stage sampling, the formulate were employed successfully in survey study of China Railway's staff about treatment cost.

【Key words】 Survey design　Stratified sampling　Three-stage sampling　Sample sizes

参 考 文 献

1. Cochran W. G. Sampling techniques. 3rd ed. New York: John Wiley & Sons, Inc., 1997, 286.
2. 菲赫金哥尔茨著,叶彦谦等译. 微积分学教程. 一卷二分册. 北京:人民教育出版社, 1980, 475.
3. 高敬,汤旦林. 二阶段抽样样本大小的估计方法. 中国卫生统计, 1984, 1(1):8.
4. 高敬,王松桂. 分层二阶段抽样样本大小的估计方法. 中国卫生统计, 1998, 15(6):51.
5. 高敬,汤旦林. 三阶段抽样样本大小的研究及应用. 中国卫生统计, 1999, 16(3):139.
6. Elbers AR. Estimating sample sizes for a two-stage sampling survey of seroprevalence of pseudorabies virus (PRV)-infected swine. Vet-Q, 1995, 17(3):92.