林业科学研究 1995,8(5):483～488
Forest Research

两阶群团抽样在森林调查中的估计效率研究*

曾伟生 骆期邦 彭长清

摘要 从两阶群团抽样的概念入手,根据两阶群团抽样当群团之间不存在差异成分($S_\beta^2=0$)时与系统抽样或简单随机抽样具有相同的抽样效率的性质,定义了误差扩大因子,提出了效率系数的概念,并推导出了效率系数等于1的临界状态的相关系数r^*的表达式,为群团抽样设计和效率评估提供了重要依据。此外,还对群团抽样估计效率与群内样地间距的关系进行了较深入的研究,并用实例进行了说明。最后,对两阶群团抽样在森林调查中的综合效率评价也进行了讨论。

关键词 两阶群团抽样、误差扩大因子、效率系数、群内相关系数

在森林资源清查中,群团(整群、群状)抽样确是一种可供选择的有效方法,国内外均已有不少应用实例[1~5]。但由于对群团抽样的概念和性质理解不当,加之在理论上尚缺乏判断效率的定量依据等原因,在应用实例中也不乏因应用不当而导致无效乃至负效应的结果。为此,有必要对这一抽样方法进行深入研究,其目的在于澄清概念并在理论上为判断效率高低提供定量依据。

对于群团抽样,首先应区分开两种情况:一种与严格意义上的整群抽样相类似,群团中的小样地只是记录单元,整个群团才是样本单元,这实质上是单阶群团抽样。欧洲国家森林清查中的群团抽样,多属于这种情况,它是由最初的大样地到样带再到小样地发展而来的。其目的是利用群状小样地(方阵)来代表整个大样地,在标准误不致于明显增加的前题下,减少工作量,从而提高效率[1]。另一种则属于两阶或多阶抽样,群团中的小样地为次阶样本单元,而群团则为初阶样本单元。本文所讨论的两阶群团抽样是指后一种情况。

1 两阶群团抽样简介

1.1 两阶群团抽样的估计公式

两阶抽样可以分成一阶单元大小相等和不等两种类型,本文只讨论前者。其总体平均数和方差的常用估计公式为[6~8]:

$$\bar{y} = \Sigma\Sigma y_{ij}/nm = \Sigma\bar{y}_i/n \tag{1}$$

$$S_{\bar{y}}^2 = S_{间}^2/n \cdot (1 - n/N) + S_{内}^2/nm \cdot n/N \cdot (1 - m/M) \tag{2}$$

式中,$S_{间}^2=\Sigma(\bar{y}_i-\bar{y})^2/(n-1)$,为群间方差;$S_{内}^2=\Sigma\Sigma(y_{ij}-\bar{y}_i)^2/[n(m-1)]$,为群内方差。

1994--12--07 收稿。

曾伟生工程师,骆期邦,彭长清(林业部中南林业调查规划设计院 长沙 410014)。

为便于进一步讨论，这里再考虑国外常用的另一个方差估计公式[1]：

$$S_{\bar{y}}^2 = S_\beta^2/n \cdot (1 - n/N) + S_\gamma^2/nm \cdot (1 - nm/NM) \tag{3}$$

式中，$S_\beta^2 = S_{间}^2 - S_\gamma^2/m$，$S_\gamma^2 = S_{内}^2$，$S_{间}^2$、$S_{内}^2$ 同(2)式，可以证明，公式(2)、(3)是互通的。

1.2 两阶群团抽样的性质

当抽样总体为无限总体时，公式(3)将系数消去可变成：

$$S_{\bar{y}}^2 = S_\beta^2/n + S_\gamma^2/nm = S_{间}^2/n \tag{4}$$

这与(2)式所导出的结果是一致的。为了进一步分析两阶群团抽样的性质，现以(3)式为基础来考虑。

方差估计公式(3)是从这样的思路提出的[1]：两阶群团抽样的每个样本单元值都含有三个成分，即：

$$y_{ij} = \mu + \beta_i + \gamma_{ij} \tag{5}$$

式中，β_i——群团 i 的差异成分；γ_{ij}——群团 i 内样地 j 的差异成分。然后再由离差公式推出，群内方差的估计值可以用 j 成分表示如下：

$$S_{内}^2{}' = \Sigma\Sigma(y_{ij} - \bar{y}_i)^2/[n(m-1)] = S_\gamma^2 \tag{6}$$

群间方差的估计值包括和两个成分，其公式为：

$$S_{间}^2{}' = m\Sigma(\bar{y}_i - \bar{\bar{y}})^2/(n-1) = mS_\beta^2 + S_\gamma^2 \tag{7}$$

总体平均数的方差必须包括 S_β^2 和 S_γ^2 这两个成分，其估计式即为(3)式。

从(6)、(7)两式可以看出，群内方差的定义式与我国林业统计书刊中常用的概念一致，即 $S_{内}^2{}' = S_{内}^2$；而群间方差则多一个 m 倍的系数，即 $S_{间}^2{}' = mS_{间}^2$，国外参考文献[9]也是如此。为保持我国传统用法，群内方差与群间方差均统一采用(2)式中的定义，从而由(7)式可得：

$$S_{间}^2 = S_\beta^2 + S_{内}^2/m \tag{8}$$

此式乃是(2)、(3)两式具有互通性的基本前提。

为分析方便，再将(3)式列如下：

$$S_{\bar{y}}^2 = S_\beta^2/n \cdot (1 - n/N) + S_\gamma^2/nm \cdot (1 - nm/NM)$$

通过分析知道，当(3)式中 $n=N$ 时即为有限制随机抽样或分层抽样的误差。因此，两阶群团抽样的误差总是要大于相应的有限制随机抽样(可视为单阶抽样，若将“随机”改为“系统”则相当于系统抽样，估计公式与简单随机抽样相同。)然而，当不存在群团差异成分，总体混合得很均匀时，会近似有 $S_\beta^2=0$ 成立，此时两阶群团抽样的误差与相应的有限制随机抽样相当。两阶抽样的误差究竟大多少，取决于总体内各群团(小区)间的差异性以及样本的成群程度[1]。

由(8)式知道，度量总体内群团差异程度的 S_β^2 值可表示为：

$$S_\beta^2 = S_{间}^2 - S_{内}^2/m \tag{9}$$

可以看出，只有当 $S_{内}^2 = mS_{间}^2$ 时，才有 $S_\beta^2=0$。此时，由公式(2)得：

$$S_{\bar{y}}^2 = S_{间}^2/n \cdot (1-n/N) + mS_{间}^2/nm \cdot n/N \cdot (1-m/M) = S_{间}^2/n \cdot (1-nm/NM)$$
$$= S_{内}^2/nm \cdot (1-nm/NM)$$

由公式(3)得：$S_{\bar{y}}^2 = S_{内}^2/nm \cdot (1-nm/NM)$

当 $S_\beta^2=0$ 时还可同时推出 $S^2 = S_{内}^2$，其中，$S^2 = \Sigma\Sigma(y_{ij} - \bar{\bar{y}})^2/(nm-1)$，为总体方差估计值。

由此可见，由(2)、(3)式所推出的结果与简单随机抽样或单阶随机抽样公式是完全一致

的。至此我们可以得出结论:只有当总体内各个群团之间不存在差异成分,亦即 $S_\beta^2=0$ 时,两阶抽样的估计误差才会与抽样比相同的简单随机抽样相等。

2　两阶群团抽样的估计效率

2.1　纯统计学意义上的估计效率

2.1.1　*估计效率的理论分析*　为了研究群团抽样的效率,这里引入如下相关系数的概念[1]:

$$\rho = \sigma_{\beta^2}/(\sigma_{\beta^2} + \sigma_{\gamma^2}) \tag{10}$$

将(10)式变形,并用样本的 S_β、S_γ 和 r 来代替 σ_β、σ_γ 和 ρ,就能用 S_γ 和 r 来表示标准误的平方(对于无限总体):

$$S_{\bar{y}}^2 = S_\beta^2/n + S_\gamma^2/nm = rS_\gamma^2/[n(1-r)] + S_\gamma^2/nm = S_\gamma^2/nm \cdot [mr/(1-r)+1] \tag{11}$$

这里的 S_γ^2/nm 与单阶有限制随机抽样的标准误平方完全相同。因此(12)式:

$$K = mr/(1-r) + 1 \tag{12}$$

即为两阶抽样的标准误平方大于有限制随机抽样或简单随机抽样(具有相同的抽样比)的标准误平方的倍数,这里将其定义为误差扩大因子。标准误的增加取决于相关系数,也取决于每个群团中所抽取的样地数。

由简单随机抽样的误差公式 $S_{\bar{y}}^2=S_y^2/n$ 知,标准误平方是与方差成正比而与样本大小成反比的。因此,当采用两阶群团抽样致使误差扩大因子 K 并不比样本增加的倍数 m 小时,两阶群团抽样与简单随机抽样相比是没有什么效率的。

根据上述分析,可将误差扩大因子 K 与群内样地数 m 相等时的状态定义为临界状态。由(12)式可导出临界状态的相关系数:

$$r^* = (m-1)/(2m-1) \tag{13}$$

由(9)、(10)、(13)式还可推出,临界状态具有一个重要特性:$S_{\text{内}}^2=S_{\text{间}}^2$。作为更普遍的形式,(12)、(13)式在有限总体条件下的表达式为(推导过程略):

误差扩大因子　$K = mr(1-n/N)/[(1-r)(1-nm/NM)] + 1$　(14)

临界相关系数 $r^*=(m-1)(1-nm/NM)/[m(1-n/N)+(m-1)(1-nm/NM)]$　(15)

这里再引入"效率系数"的新概念,并作如下定义:

$$E_c = m/K \tag{16}$$

式中,m 含义同前,K 为(14)式所定义的误差扩大因子。当相关系数 $r=r^*$ 时,效率系数 $E_c=1$,此时群团样地与单个样地相比没有效率;当 $r>r^*$ 时,$E_c<1$,群团抽样为负效率;当 $r<r^*$ 时,$E_c>1$,群团抽样为正效率。"效率系数"的直接含义是:群团抽样中的一个群团相当于简单随机抽样(相同抽样比)中单个样地的数量。很明显,只有当一个群团相当于一个以上的独立样地时才有可能谈得上效率。

2.1.2　*估计效率与群内样地间距的关系*　我国林业统计书刊在讨论整群抽样时建议用"群内相关系数"ρ_w 来分析其估计效率[8]。据作者研究发现,由(10)式定义的相关系数 ρ 与群内相关系数 ρ_w 之间存在一定的函数关系,但两者差异极小,一般情况下可近似认为两者相等。

关于估计效率与群内样地间距的关系,已经有人进行过试验[4]。根据其试验材料,群内相关系数与群内样地间距大致呈负指数相关。因为群内相关系数所反映的是群内各样地之间的

相关性，因此从理论上讲，当群内各样地之间的距离为无穷远时，其相关性应该为零。因而可以确定群内相关系数与群内样地间距的关系式如下：

$$r_w = aL^{-b} \tag{17}$$

式中，r_w 为 ρ_w 的样本估计值，L 为群内样地间距，a、b 为待定参数。

以上只是理论分析，在实际应用中群内样地间距不可能无穷远，群内相关系数也只要当样地之间达到一定距离 L_o 后就会趋近于零。因此可将(17)式改为如下形式：

$$r_w = a(L^{-b} - L_o^{-b}), L \leqslant L_o \tag{18}$$

当 $L > L_o$ 时，取 $r_w = 0$。

利用江西省德兴县的试验数据，用非线性最小二乘法拟合(17)式，可得 $a = 2.615\ 47$，$b = 0.321\ 26$，$R = 0.626\ 6$。对于(18)式的拟合，则取决于 L_o 值。根据不同 L_o 值的拟合结果可以发现，当 $L_o < 4$ km 时，R 明显减小；当 $L_o \geqslant 4$ km 时，R 都稳定在 0.6 以上。这就是说，对江西德兴县而言，当群内样地间距达到 4 km 远时，基本上可以认为群内相关系数已趋于 0。

对于省级森林资源清查体系，系统抽样的样地间距一般为 4 km 左右。如果设计为方形群团样地，并取 $m = 4$，那么，当群内样地间距为 4 km，群与群之间相距为 8 km 时，群团抽样就与间距 4 km 的系统抽样完全一样。即认为此时 $r = 0$ 或 $S_\beta^2 = 0$，一个由 4 个样地组成的群团与 4 个独立样地完全相当。根据德兴县的材料拟合 $L_o = 4$ km 时的群内相关系数模型($a = 3.477\ 68$，$b = 0.079\ 581$，$R = 0.617\ 1$)，就可以对不同群内样地间距时的群团抽样效率进行估计，详见表 1。因为只是说明问题，表 1 中的数据计算未考虑 r 与 r_w 之间的细小差异，而将其视为相等；误差扩大因子采用无限总体条件下的(12)式计算。

表 1 不同群内样地间距时的群团抽样效率($m = 4$)

群内样地间距 (m)	群内相关系数 r_w	误差扩大因子 K	效率系数 E_c
250	0.443 7	4.190 4	0.954 6
500	0.323 4	2.911 9	1.373 7
1 000	0.209 6	2.060 7	1.941 1
1 500	0.145 9	1.683 3	2.376 3
2 000	0.101 9	1.453 8	2.751 4
3 000	0.041 6	1.173 7	3.408 0
4 000	0	1	4

从表 1 可以看出，对于德兴县而言，当群内样地间距为 1 km 时，一个由 4 个样地组成的群团大致可相当于 2 个独立样地；如果群内样地间距减少至 500 m，则其效率还不如 1.5 个独立样地；距离减至 250 m 时甚至出现了“负效率”。由(13)式容易知道，$m = 4$ 时临界相关系数 $r^* = 0.428\ 6$，按群内相关系数模型可反推出临界状态时的群内样地间距约为 270 m。也就是说，只有当群内样地间距大于 270 m 时群团抽样才有可能谈得上有效率。

2.2 在森林调查中的综合效率

在此之前所讨论的群团抽样效率，都纯粹从统计学概念出发的，没有考虑野外调查时的费用问题。但是，首先必须在统计概念上有效率，即效率系数必须大于 1，才有可能在森林调查中比简单随机抽样或系统抽样合算。然而，效率系数大于 1，却并不一定会在经济上合算，这取决于从一个初阶单元(群团)转移到另一个初阶单元的费用及量测一个次阶单元(样地)的费用多少。

假设，迁移一个营帐并到达一个群团的平均费用为 C_1，量测一个样地(包括从营地到达这个样地)的平均费用为 C_2，那么，采用两阶群团抽样时野外工作的总费用就是：

$$C = nC_1 + nmC_2 = n(C_1 + mC_2) \tag{19}$$

如果采用系统抽样，并且上述群团抽样中 1 个群团相当于 k 个独立样地的效率，那么类似地可以得到其野外工作的总费用为：

$$C' = kn(C_1' + C_2') \tag{20}$$

在确定了(19)、(20)式中的各项参数后就可以对两种抽样方式的估计效率作出最后比较。

假设两种抽样方式迁移一个营地都是 1 d 时间，完成 1 个群团(含 4 个样地)需要 4 d，完成 1 个独立样地需要 1 d，1 个群团相当于 2.5 个独立样地的效率。到底哪种抽样方式要合算？

如果用工作日多少代替调查费用来作比较，则由(19)式可得 $C=5n$，由(20)式得 $C'=2.5n\times 2=5n$，因此，两种抽样方式是等效的。很明显，只要改变上述假设参数，就可能有不同的比较结果。

总之，评价抽样方式在森林调查中的效率高低，除取决于统计学上的抽样效率外，还在很大程度上受各项调查费用参数的影响。

3　结　论

通过对两阶群团抽样的性质及其在森林调查中的估计效率研究，可以得出如下一些结论：

(1)两阶群团抽样的误差总是要大于抽样比相同的系统抽样或简单随机抽样的误差。只有当总体混合得很均匀，群团之间不存在差异成分($S_\beta^2=0$)时，两者才具有相等的抽样误差。

(2)在 $S_\beta^2\neq 0$ 的情况下，两阶群团抽样与简单随机抽样的相对效率可以用误差扩大因子或效率系数表示。当效率系数等于 1(相当于 $S_{内}^2=S_{间}^2$)时，两阶群团抽样无效率；当效率系数小于 1($S_{内}^2<S_{间}^2$)时，为负效率；当效率系数大于 1($S_{内}^2>S_{间}^2$)时，为正效率。

(3)效率系数等于 1 时的状态可用临界相关系数 r^* 表示。对于有限总体，r^* 的大小取决于 N、M、n、m 值；而对于无限总体，r^* 只取决于 m 的取值。r^* 值的大小可作为群团抽样设计的重要依据。

(4)群内相关系数与群内样地间距之间存在着负指数关系。群内样地之间的距离 L 越大，群内相关系数 r_w 就越小；当距离 L 大到一定程度后，r_w 已近似等于 0，从而可将群内的每个样地当作独立样地看待。

(5)两阶群团抽样的效率系数大于 1 时为正效率是纯粹从统计学理论出发的。在森林调查中是否确实在经济上合算，还取决于考虑野外调查费用时的综合效率。

参 考 文 献

1　洛茨，哈勒，佐勒(林昌庚，沙琢等译校). 森林资源清查. 北京：中国林业出版社，1988.

2　李茂深. 群状抽样在森林资源调查中的应用. 林业资源管理，1987，(3)：42～47.

3　杨宗勋. 群状抽样在建立县级森林资源连续清查体系中的应用. 广东林业科技，1988，(1)：19～25.

4　林毓资. 群团抽样最优间距试验报告. 林业调查与设计，1985，(1)：10～21.

5　IUFRO S4. 02, Finnish Forest Research Institute, Department of Forest Resource Management of University of Helsinki. Proceedings of Ilvessalo Symposium on National Forest Inventories. Helsinki, Finland, 1992.

6　林业部调查规划院主编. 森林调查手册. 北京：中国林业出版社，1980.

7　陈华豪，丁思统，蔡贤如，等. 林业应用数理统计. 大连：大连海运学院出版社，1992.

8　北京林学院主编. 数理统计. 北京：中国林业出版社，1980.

9　卡尔·温格编(林业部华东林业调查规划设计院译). 林业手册. 北京：国际文化出版公司，1990.

Study on Efficiency of Two-Stage Cluster Sampling in Forest Inventory

Zeng Weisheng Luo Qibang Peng Changqing

Abstract According to the special feature that two-stage cluster sampling has the same efficiency as the systematic sampling or simple random sampling when there are no differences among clusters, the term of error expansion factor is presented in this paper. Besides the correlation coefficient for critical state, when efficiency coefficient equals to 1, is formulated, which provide important reference of implementing cluster sampling design and sampling efficiency assessment. In addition, the relationship between efficiency of two-stage cluster sampling and distance between plots in a cluster is further studied with an example demonstrated in detail. Finally, the efficiency assessment of two-stage cluster sampling in forest inventory is also discussed.

Key words two-stage cluster sampling, error expansion factor, efficiency coefficient, correlation coefficient within clusters

Zeng Weisheng, Engineer, Luo Qibang, Peng Changqing (South-Central Forest Inventory & Planning Institute Changsha 410014).