

中国黄牛品种基因频率 抽样估计效率的研究*

常 洪 耿社民 武 彬

(西北农业大学畜牧系)

李 写 一 陈 幼 春

(西北农业大学农经系) (中国农业科学院畜牧研究所)

摘 要

本文给出基因频率纯随机抽样、随机整群抽样、系统随机抽样、系统随机整群抽样方差无偏估计值的算式,并论述各种抽样方法的相对精确度。讨论基因频率估计值的精确度、可靠性与样本规模的一般关系,证明当类别数、抽样群数、标准离散度和标准偏差分别为 d 、 k 、 θ 和 $\lambda, \dots$

时,基因频率整群抽样估计值的可靠性为 $\beta_{1,1,\dots} = Pr \{ - (dk)^{\frac{1}{2}} (2\theta)^{-1} \leq \lambda_{1,1,\dots} \leq (dk)^{\frac{1}{2}} (2\theta)^{-1} \}$ 。研究肯定中心产区系统随机整群抽样是中国黄牛品种遗传检测的有效抽样方式。

关键词: 中国黄牛; 系统随机整群抽样; 标准离散度

在一定的样本规模下,品种或地域群基因频率估计值的精确度和可靠性标志着抽样估计的效率。近20年来虽然社会统计学和人类医学统计学在二项总体抽样偏差^[1]、规模估计^[2,3]、分层抽样参数^[4,5]等领域取得了一系列具有普遍意义的成果,但对群体基因频率各种抽样估计方法的效率尚缺乏系统的研究。为了提高基因频率抽样调查效率,将国家正在着手进行的中国黄牛品种分类^[6]置于可靠的基础上,本文在阐明各种方法抽样估计值误差的基础上,对此进行了探讨。

1 不同抽样方法的相对精确度

以相同的总头数 n ,分别用纯随机、随机整群、系统随机和系统随机整群四种方式由 d 个含量为 N_h ($h = 1, 2, \dots, d$, $N_1 = N_2 = \dots = N_h = \dots = N_d$)、包括 k_h ($k_1 = k_2 = \dots = k_h = \dots = k_d$) 个头数为 n_{h1} ($n_{11} = n_{12} = \dots = n_{21} = n_{22} = \dots = n_{h1} = \dots = n_{dkd}$) 的牛群的类别构成、总头数为 N 、基因频率为 P 的品种中抽样,并设 k , f , N_h 和 W_h 分别代表两种整群抽样的总群数、抽样率、类别抽样头数和类别权值(即 N_h/N), S_T^2 , $S_{\dots}^2$ 和 S_h^2 分别代表总均方、混类群间均方与类别中均方, SS_T 和 SS_h 代表相应的平方和, p , $p_{h\dots}$ 与 $p_{h\dots}$ 分别代表品种基

本文于1988年7月2日收到。

* “七五”期间国家科技攻关课题“中国黄牛品种分类研究”的一部分。

因频率的估计值、第 h 类以及第 h 类第 u 群的基因频率, 则四种方式的基因频率抽样方差之无偏估计量顺序为:

$$\begin{aligned} v(p_s) &= \frac{p_s q_s}{2(n-1)} \left(\frac{N-n}{N} \right) = \frac{S_T^2}{2n} \left(\frac{N-n}{N} \right) \\ v(p_{st}) &= \frac{1 - \frac{n}{N}}{\frac{n}{n_{hu}}} \frac{n/n_{hu}}{\frac{n}{n_{hu}} - 1} \frac{(r_u - p_s)}{\frac{n}{n_{hu}} - 1} = \frac{S_u^2}{2n} \left(\frac{N-n}{N} \right) \\ v(p_{st}) &= \sum \frac{w_h^2 p_h q_h}{2(n_h - 1)} (1 - f_h) = \frac{(1-f)}{2n} \sum W_h S_h^2 \\ v(p_{st \cdot c}) &= \frac{1-f}{\frac{n}{n_{hu}} \left(\frac{n}{n_{hu}} - d \right)} \sum_{h=1}^d \sum_{u=1}^{n/d_{hu}} (p_{hu} - p_{h \cdot c})^2 \\ &= \frac{n_{hu}^2 SS_h}{n(n-d_{hu})} \left(\frac{N-n}{N} \right) = \frac{n_{hu}}{n} S_h^2 \left(\frac{N-n}{N} \right) \text{ (另文详论)} \end{aligned}$$

因而, 品种基因频率的随机整群抽样、系统随机抽样和系统随机整群抽样估计值的相对精确度分别是

$$\begin{aligned} \frac{v(p_s)}{v(p_{st})} &= \frac{npq}{(n-1)S_u^2} = S_T^2/S_u^2 \\ \frac{v(p_s)}{v(p_{st})} &= \frac{pq}{\sum w_h p_h q_h} = \frac{S_T^2}{\sum W_h S_h^2} \\ \frac{v(p_s)}{v(p_{st \cdot c})} &= \frac{n(n-dn_{hu})pq}{2(n-1)n_{hu}^2 \sum_{h=1}^d \sum_{u=1}^{n/d_{hu}} (p_{hu} - p_{h \cdot c})^2} \\ &= \frac{(n-dn_{hu})SS_T}{2(n-1)n_{hu}^2 SS_h} = \frac{S_T^2}{2n_{hu} S_h^2} = \frac{kS_T^2}{2nS_h^2} \end{aligned}$$

由分类总体的标准代数式可得:

$2(N-1)S_T^2 = 2 \sum_h (N_h - 1)S_h^2 + 2 \sum_h N_h (p_h - p)^2$, 当品种规模很大, 以致 $1/2N_h$ 与 $1/2N$ 皆可忽略不计时, 则有:

$$\begin{aligned} S_T^2 &= \sum w_h S_h^2 + \sum w_h (p_h - p)^2 \\ \therefore v(p_s) &= v(p_{st}) + \frac{1-f}{2n} \sum w_h (p_h - p)^2 \\ \therefore \frac{(1-f)}{2n} \sum w_h (p_h - p)^2 &\geq 0 \\ \therefore \frac{v(p_s)}{v(p_{st})} &\geq 1 \end{aligned}$$

因此, 在抽样规模相等的条件下: (1) 基因频率随机整群抽样估计值的相对精确度取决于总均方与群间均方的比值, 地方畜禽品种后者通常大得多, 所以其值一般小于1; (2) 基因频率系统随机抽样估计值的相对精确度永远大于1; (3) 基因频率系统随机整群抽样的相对精确度取决于总均方和平均群含量乘类别中均方之积的比值, 当类别间差异较大(同类别各群相对较一致)、群多而含量较小时, 其值较高。

在上述条件下,如以 n_c , n_{st} 和 $n_{st..c}$ 分别代表相对精确度为1时各相应方法的抽样规模,并将 n/N , n_c/N , n_{st}/N 和 $n_{st..c}/N$ 皆忽略不计,可得:

$$n_c = \frac{n S_u^2}{S_T^2};$$

$$n_{st} = \frac{n \sum w_h^2 C_h^2}{S_T^2};$$

$$n_{st..c} = \frac{2 n \cdot n_{st} C_h^2}{S_T^2};$$

如设两种整群抽样方式在各类别中的抽样群数相等 ($k_1 = k_2 = \dots k_h = \dots k_d$), 则有

$$v(p_s) : v(p_{st..c}) = \frac{1}{k(k-1)} \left[\sum_{h=1}^d \sum_{u=1}^{kd} (p_{hu} - p_h)^2 + k_h \sum_{h=1}^d (p_h - P)^2 \right]$$

$$: \frac{1}{d k_h (k_h - 1)} \sum_{h=1}^d \sum_{u=1}^{kd} (p_{hu} - p_h)^2$$

$$= (k-d) SS_T : (k-1) SS_h = S_T^2 : S_h^2$$

可见, 抽样规模相等时, 系统随机整群抽样估计值的精确度远高于随机整群抽样。

2 基因频率抽样估计值的精确度及可靠性与样本规模的一般关系

1960年H'ajek, J.证明, 频率的抽样分布近似于以总体实际频率为中正值的正态分布^[2]。因而由基因频率纯随机抽样标准差

$$\sigma_{p_s} = \sqrt{\frac{PQ}{2n} \left(\frac{N-n}{N-1} \right)}$$

以及标准偏差

$$\lambda_s = \frac{p_s - P}{\sigma_{p_s}} \text{ 可得:}$$

$$n = \frac{\frac{PQ\lambda_s^2}{(p_s - P)^2}}{2 + \frac{1}{N} \left[\frac{PQ\lambda_s^2}{(p_s - P)^2} - 2 \right]}$$

如果基因频率抽样估计所要求的可靠性已由 λ_s 规定, 对于特定的总体实际频率 P , n 就是达到偏差 $(p_s - P)$ 所标志的精确度所必需的最小抽样规模。设相对偏差为 $\eta = (p_s - P)/P$, 并要求:

$$Pr \left\{ \frac{|p_s - P|}{P} \geq \eta \right\} = Pr \left\{ |p_s - P| \geq \eta P \right\} = \alpha$$

必需最小抽样规模则可示为:

$$n = \frac{Q/P \left(\frac{\lambda_s}{\eta} \right)^2}{2 + \frac{1}{N} \left[Q/P \left(\frac{\lambda_s}{\eta} \right)^2 - 2 \right]}$$

而相对偏差在限定范围内之样本的概率则是:

$$\beta_s = Pr \{ a \leq \lambda_s \leq b \} = \int_a^b \varphi(\lambda_s) d\lambda_s$$

$$= \int_0^{\lambda_s} \frac{2}{\sqrt{2\pi}} e^{-\frac{\lambda_s^2}{2}} d\lambda_s$$

如果品种规模很大, 以至 $1/N$ 可忽略不计, 基因频率抽样估计值的可靠性按统计学通常的概率界限 0.9545 给定, 则有

$$n = \frac{2(1-p)}{\eta^2 p};$$

如果相对偏差以 0.5 为限, 则有

$$\beta_s = \Pr \left\{ - (0.5p)^{\frac{1}{2}} Q^{-\frac{1}{2}} n^{\frac{1}{2}} \leq \lambda_s \leq (0.5p)^{\frac{1}{2}} Q^{-\frac{1}{2}} n^{\frac{1}{2}} \right\}.$$

据此可得表 1, 表 2 和图 1。

由表 1、表 2 可见: 对于 0.01, 0.05 和 0.10 的品种实际基因频率, 如果要求 95% 以上的样本估计值不偏离其 0.5 倍, 所必需的最小抽样规模分别是 792, 152 和 72 头; 抽样 50 头, 对于 0.01, 0.05 和 0.10 的品种实际频率, 估计值不偏离其 0.5 倍的可靠性分别只有 0.3669, 0.7485 和 0.8864。可见品种实际基因频率值越低, 抽样估计值达到同等精确度和可靠性水准所必需的样本规模越大。

表 1 估计精确度与抽样规模的关系

基因 频率	相 对 偏 差									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.01	19800.00	4950.00	2200.00	1237.50	792.00	550.00	404.08	309.37	244.44	198.00
0.02	9800.00	2450.00	1088.89	612.50	392.00	272.22	200.00	153.12	120.99	98.00
0.03	6466.67	1616.67	718.52	404.17	258.67	179.63	131.97	101.04	79.84	64.67
0.04	4800.00	1200.00	533.33	300.00	192.00	133.33	97.96	75.00	59.26	48.00
0.05	3800.00	950.00	422.22	237.50	152.00	105.56	77.55	59.38	46.91	38.00
0.06	3133.33	783.33	348.15	195.83	125.33	87.04	63.95	48.96	38.68	31.33
0.07	2657.14	664.29	295.24	166.07	106.29	73.81	54.23	41.52	32.80	26.57
0.08	2300.00	575.00	255.56	143.75	92.00	63.89	46.94	35.94	28.40	23.00
0.09	2022.22	505.56	224.69	126.39	80.89	56.17	41.27	31.60	24.97	20.22
0.10	1800.00	450.00	200.00	112.50	72.00	50.00	36.73	28.13	22.22	18.00
0.20	800.00	200.00	88.89	50.00	32.00	22.22	16.33	12.50	9.88	8.00
0.30	466.67	116.67	51.85	29.17	18.67	12.96	9.52	7.29	5.67	4.67
0.40	300.00	75.00	33.33	18.75	12.00	8.33	6.12	4.69	3.70	3.00
0.50	200.00	50.00	22.22	12.50	8.00	5.56	4.08	3.12	2.47	2.00

表2 抽样规模与估计可靠性的关系

基因 频率	样 本 规 模										
	50	60	70	80	90	100	110	120	130	140	150
0.01	0.3669	0.4180	0.4278	0.4540	0.4781	0.5227	0.5211	0.5637	0.5822	0.5995	0.5916
0.02	0.5249	0.5427	0.5779	0.6337	0.6373	0.6857	0.6857	0.7314	0.7260	0.7679	0.7600
0.03	0.6207	0.6398	0.6769	0.7338	0.7375	0.7862	0.8077	0.8042	0.8436	1.8586	0.8513
0.04	0.6676	0.7363	0.7727	0.8031	0.8289	0.8295	0.8697	0.8667	0.9000	0.8951	0.9227
0.05	0.7485	0.7909	0.8251	0.8317	0.8561	0.8765	0.8938	0.9242	0.9354	0.9449	0.9407
0.06	0.7933	0.8334	0.8442	0.8708	0.8924	0.9101	0.9388	0.9368	0.9582	0.9653	0.9712
0.07	0.8297	0.8669	0.8767	0.9005	0.9341	0.9345	0.9579	0.9663	0.9729	0.9781	0.9758
0.08	0.8595	0.8935	0.9187	0.9234	0.9396	0.9522	0.9621	0.9775	0.9824	0.9863	0.9892
0.09	0.8646	0.9148	0.9370	0.9410	0.9648	0.9737	0.9731	0.9792	0.9887	0.9914	0.9935
0.10	0.8864	0.9169	0.9388	0.9546	0.9662	0.9747	0.9810	0.9901	0.9892	0.9946	0.9938
0.20	0.5875	0.9938	0.9950	0.9973	0.9992	0.9996	0.9996	0.9998	0.9998	0.9999	1.0000
0.30	0.9989	0.9993	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.40	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
基因 频率											
	160	170	180	190	200	250	300	350	400	450	500
0.01	0.6068	0.6212	0.6596	0.8726	0.6602	0.7141	0.7815	0.7933	0.8446	0.8681	0.8687
0.02	0.7985	0.8120	0.8245	0.8360	0.8467	0.8706	0.9196	0.9410	0.9565	0.9677	0.9679
0.03	0.8841	0.8949	0.9046	0.9133	0.9212	0.9505	0.9686	0.9799	0.9370	0.9877	0.9945
0.04	0.9169	0.9400	0.9340	0.9532	0.9473	0.9637	0.9875	0.9896	0.9938	0.9928	0.9978
0.05	0.9597	0.9553	0.9703	0.9662	0.9781	0.9896	0.9923	0.9975	0.9979	0.9994	0.9987
0.06	0.9680	0.9729	0.8833	0.9861	0.9884	0.9926	0.9980	0.9991	0.9993	0.9998	0.9999
0.07	0.9857	0.9385	0.9365	0.9324	0.9938	0.9964	0.9986	0.9994	0.9998	0.9998	1.0000
0.08	0.9916	0.9934	0.9920	0.9936	0.9968	0.9990	0.9937	0.9938	0.9939	1.0000	1.0000
0.09	0.9950	0.9940	0.9971	0.9978	0.8971	0.9991	0.9999	0.9999	1.0000	1.0000	1.0000
0.10	0.9953	0.9964	0.9973	0.9988	0.9984	0.9998	0.9989	1.0000	1.0000	1.0000	1.0000
0.20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.40	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

3 整群抽样估计值的精确度和可靠性与样本规模的关系

将随机整群抽样作为系统随机整群抽样的类别数 $d = 1$ 的特例, 将类别中群间标准差与总体实际频率的比值定义为标准离散度, 并设:

$$\theta = \sqrt{\frac{\sum S_{au}^2}{P}}$$

则有

则有

$$k = \frac{\theta^2 \lambda^2}{d \eta^2}.$$

当相对偏差与标准偏差按前述标准限定时, 则分别有

$$k = \frac{16\theta^2}{d},$$

$$\beta_{s.t.c} = \Pr \left\{ - (dk)^{\frac{1}{2}} (2\theta)^{-1} \leq \lambda_{s.t.c} \leq (dk)^{\frac{1}{2}} (2\theta)^{-1} \right\}$$

据以可得表3和图2。

表3 必抽群数与标准离散度的关系

类别数	标准离散度														
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
1.00	0.16	0.64	1.44	2.55	4.00	5.76	7.84	10.24	12.96	16.00	19.36	23.04	27.04	31.36	36.00
2.00	0.08	0.32	0.72	1.28	2.00	2.88	3.92	5.12	6.48	8.00	9.68	11.52	13.52	15.68	18.00
3.00	0.05	0.21	0.48	0.85	1.33	1.92	2.61	3.42	4.32	5.33	6.45	7.68	9.01	10.45	12.00
4.00	0.04	0.16	0.36	0.64	1.00	1.44	1.96	2.56	3.24	4.00	4.84	5.76	6.76	7.84	9.00
5.00	0.03	0.13	0.29	0.51	0.80	1.15	1.57	2.05	2.59	3.20	3.87	4.61	5.41	6.27	7.20

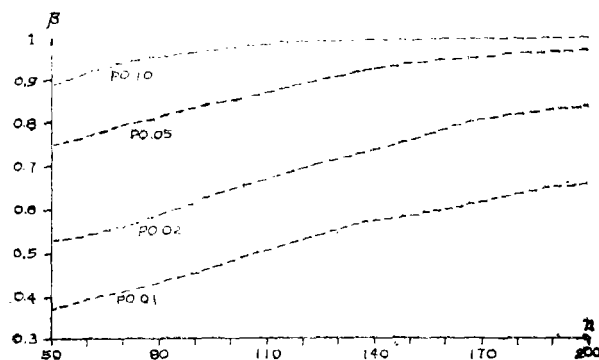


图1 样本规模与估计可靠性关系

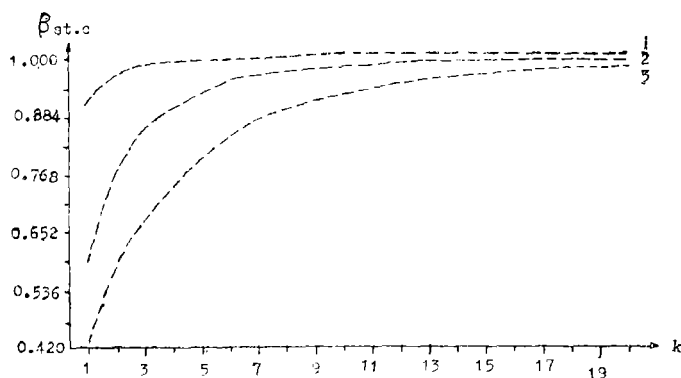


图2 抽样数与估计可靠性的关系

1. $\theta=0.5$; 2. $\theta=1.0$; 3. $\theta=1.5$

ΣS_{1i}^2 是群体特性,而 θ 则可作为确定补充抽样规模的基本参数。

4 讨 论

4.1 系统随机抽样和纯随机抽样虽有较高精确性和可靠性,但鉴于中国黄牛群体地理学特点和目前品种登记事业的发展水平^[3],目前几乎完全不可能付诸实施。

4.2 70年代后期以来在苏联和一些亚洲国家地方牛种研究中流行的大群体整群抽样方法,既不能有效地提高估计值精确度,又有降低抽样随机化程度的危险,因此不是高效抽样方式。

4.3 系统随机整群抽样可以用较小的样本规模取得精确度、可靠性较高的估计值。当 $\Delta S_T^2 > 2nS_b^2$ 时,其效率高于纯随机抽样。

5 结 语

中心产区系统随机整群抽样是中国黄牛品种遗传检测的有效抽样方式。根据对品种结构的既有认识,多分类,降低抽样群含量标准,增加抽样群数,有助于提高其效率。

参 考 文 献

- 1 David I P et al. J Amer Stat Assoc, 1974; 69: 464—466
- 2 Cochran W G. Sampling Techniques (3rd ed) . 1977: 39—79
- 3 Gross A J et al. Biometrics, 1987; 43(4): 875—884
- 4 Booth G et al. J Amer Stat Assoc, 1969; 64: 560—573
- 5 Johnson R A et al. Biometrics, 1987; 43 (3) : 641—655
- 6 陈幼森,常洪,中国黄牛,1987(3): 75—77

STUDIES OF EFFICIENCY OF SAMPLING ESTIMATION FOR GENE FREQUENCIES IN CHINESE CATTLE BREEDS

Chang Hong, Gong Shemin, Wu Bin, Li Xieyi

(Northwestern Agricultural University)

Chen Youshun

(Institute of Animal Science, Chinese Academy of Agricultural Sciences)

Abstract

Algorithms for computing the unbiased estimates of sampling variance for gene frequencies in simple random, random cluster, stratified random and stratified random cluster sampling are presented, and relative precision of various sampling methods is described in this paper. The generalized relations between precision of estimate for gene frequencies and sample sizes and between sample sizes and reliability of estimate of gene frequencies are discussed. Also reliability of estimate of gene frequencies in cluster sampling is proposed to be $\beta_{1..c} = \Pr \{ -(dk)^{\frac{1}{2}} (2\theta)^{-1} \leq \lambda_{1..c} \leq (dk)^{\frac{1}{2}} (2\theta)^{-1} \}$, where d is the number of stratum, k the number of cluster unit, θ the standard discrete degree, and $\lambda_{1..c}$ the standard normal deviate. Thus, it is established that the stratified random cluster sampling in the central area of distribution is efficient for genetic investigation of Chinese cattle breeds.

Key word: Chinese cattle, stratified random cluster sampling, standard discrete degree.