生物多样性 1997,5 (1):61~66
CHINESE BIODIVERSITY

生物群落多样性的测度方法 Ⅳ 刀切法和自助法在生物多样性测度研究中的应用*

刘灿然 马克平
(中国科学院植物研究所, 北京 100044)
周文能
(中国农业科学院, 北京 100081)
张新生
(中国科学院地理研究所, 北京 100101)

1 引言

目前已有相当数量的多样性测度公式提出,笔者在前文对此做过评述[1,2],Dennis 等讨论了其中某些公式被正确使用或错用的情况。此外,对这些公式的抽样性质更引起一些学者的关注。Kempton 指出,多样性指数的理论标准差几乎在所有的情况下都是不适当的,因为它们假设固定大小的样本单位是从一个同质总体中取出来的,这一假设在实际中很少能够满足[3]。然而,一个更严重的问题是,多样性指数大都假设生物个体是随机抽取的[4,5]。由于生物个体常常是高度集聚的,从而使随机抽样模型不能严格地应用[6],在实际调查时往往使用样方抽样,这样得到的样本是关于所研究的空间范畴的随机样本,而不是关于个体的随机样本[7]。因此,用样方抽样得到的数据计算多样性指数等就会在一定程度上存在缺陷[8]。还有一个实际问题,就是很多测度公式是样本观测值的复杂函数,对其抽样性质很难进行理论上的分析,致使大多数生态学家没能对其研究的参数进行方差估计或构造置信区间。

生物分布的聚块性、抽样的非个体随机性以及生物多样性测度公式的统计分布的复杂性,给群落多样性测度的经典统计分析带来了极大的困难。而现代非参数统计方法——"Computer intensive"技术为这一问题的解决提供了一条崭新的途径。本文拟介绍其中的两种方法——刀切法(Jackknife)和自助法(Bootstrap),评述其在物种多样性指数、丰富度和均匀度指数估计中的应用情况,着重讨论在样方抽样时,这两种非参数方法对各种指数的点估计、方差估计和置信区间估计的适用性。

2 刀切法和自助法

刀切法是由 Quenouille 提出的,旨在减少估计的偏差[7,9]。Turkey 又对其加以改进,并指出在标准方法很难应用的时候,可以用该技术得到近似的置信区间[10,11]。自助法是由 Efron 作为刀切法的一种替代方法提出的,但它的应用更广泛且可靠[7]。

2.1 刀切法

设 θ 是要估计的参数,$(X_1, X_2, \cdots, X_n)$ 是来自总体 $F(\theta, X)$ 的一个具有 n 个独立观察值的样本,假设 $\hat{\theta} = f(X_1, X_2, \cdots X_n)$ 是参数 θ 的一个估计量,经过如下步骤可以得到刀切估

收稿日期:1995-04-27 接受日期:1996-04-27

计[7,11]：

(i)去掉一个观察值，如 X_i；(ii)用 $n-1$ 个观察值($X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n$)计算 θ 的估计值，记为 $\hat{\theta}_{-1}$；(iii)计算虚拟值 $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$

对 $i = 1, 2, \cdots\cdots, n$ 重复步骤(i) ~ (iii) n 次，就可以得到参数 $\hat{\theta}$ 的刀切估计：

$$J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$$

这是参数 θ 的一阶刀切估计，可以消除阶为 $1/n$ 的偏差。

这一估计的抽样方差为：$\mathrm{Var}\{J_n(\hat{\theta})\} = \frac{1}{n(n-1)}\sum_{i=1}^{n}[J_n(\theta) - \hat{\theta}_i]^2$

Turkey 猜想，这些虚拟值 $\hat{\theta}_i (i = 1, 2, \cdots, n)$ 可以认为是独立同分布的。于是，可以构造近似的置信区间，例如 95% 的置倍区间[13]为：$J_n(\theta) \pm t_{n-1,0.05}[\mathrm{Var}\{J_n(\theta)\}]^{1/2}$

2.2 自助法

还假设($X_1, X_2, \cdots, X_n$)是来自一个总体(分布 F 未知)的具有 n 个独立观察值的一个样本，θ 是要估计的参数，$\hat{\theta} = f(X_1, X_2, \cdots, X_n)$ 是参数 θ 的一个估计量。自助估计的步骤如下[7]：

(i)构造经验概率分布：$F(X_i^*) = \frac{i}{n}$，其中(X_i^*)是($X_1, X_2, \cdots, X_n$)的从小到大的第 i 个值；(ii)通过有放回抽样，从 n 个数据点中抽出大小为 n 的一个样本，这就构成了“自助”样本；(iii)基于(ii)中的“自助”样本，计算 θ 的估计值；(iv)重复(ii)和(iii) m 次，得到参数 θ 的 m 个估计值，记为 $\hat{\theta}_{(i)}, i = 1, 2, \cdots m$。于是，得到 θ 的自助估计：$B_n(\theta) = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{(i)}$，

其抽样方差为：$\mathrm{Var}\{B_n(\theta)\} = \frac{1}{m-1}\sum_{i=1}^{m}[\hat{\theta}_{(i)} - B_n(\theta)]^2$

参数 θ 的 95% 的置信区间[13]为：$B_n(\theta) \pm t_{n-1,0.05}[\mathrm{Var}\{B_n(\theta)\}]^{1/2}$

用自助法进行置信区间估计还有其它种几方法[12,13]。

3 刀切法和自助法在生物群落多样性测度中的应用

3.1 在多样性指数研究中的应用

Zahl 最早将刀切法应用于多样性指数的研究[5]。他在样方抽样下考察了 Shannon 和 Simpson 两个多样性指数。设 C_{ij} 为第 j 个样方中第 i 种的多度，n 为样方个数，s 为样本中观察到的物种数目，则 $C_i = \sum_{j=1}^{n} C_{ij}$ 为样本中第 i 种的总多度，$C = \sum_{i=1}^{s} C_i$ 为样本中所有种的总多度。他用 $\hat{H}' = -\sum_{i=1}^{s}(C_i/C)\ln(C_i/C)$，作为 Shannon-Weiner 多样性指数 $H' = -\sum_i P_i \ln P_i$ 的原始估计量，其中，P_i 为第 i 个种的相对多度。用 $\hat{\lambda}_0 = \sum_{i=1}^{s} C_i^2/C^2$，作为 Simpson 聚集度指数 $\lambda = \sum_i P_i^2$ 的原始估计量，其中，P_i 的意义同上，$1-\lambda$ 即为 Simpson 多样性指数。并提出对这两个估计量实施一阶刀切估计，这样不仅可以消除阶为 $1/n$ 的偏差，而且可以得到抽样方差估计[14]。他还指出，对 Shannon 和 Simpson 多样性指数来说，在大多数情况下，刀切虚拟值的确是正态分布的，并且不要求对个体的随机抽样，这种抽样在实际中很难得到。Adams 和 McCune 以及 Heltshe 和 Bitz 也得出了相似的结论。前者认为刀切虚拟值的方差明显地比 Shannon 指数的其它形式估计的方差优越，并且如果是对个体随机取样的话，刀切 $D(=1-\sum_{j=1}^{s} n_j(n_j-1)/N(N-1)$，它

是 Simpson 多样性指数的极大拟然估计,其中 n_j 是第 j 种的多度,N 是所有种的总多度)会给出 Simpson 多样性指数的最小方差无偏估计。后者发现刀切估计的偏差比 Pielou 的" Pooled quadrat "方法估计的方差小得多[5]。但是,Zahl 也指出,样方之间的相关性也可以导致对正态性的偏离,尤其是在样方数很小的时候[14]。

Routledge 用 400 个 20 cm^2 的维管植物群落样方资料,并借助模拟技术进一步考察了 Shannon 和 Simpson 两个指数。对于 H',他认为一阶刀切估计对于小样本来说都不能充分地减小偏差。要使偏差不超过 5%,则样方数要超过 50。更高阶的刀切估计本应该进一步降低偏差,例如,二阶刀切估计应该消除一、二阶偏差项。但是,当样方数在 10 ~ 100 之间时,也可能出现较大的偏差。因此,他指出,除非真正找到了 H'的一个好的估计量,否则 H'的无偏估计将需要更大的样方数[14]。

对于 λ,他提出也可以用:

$$\hat{\lambda}_1 = \frac{(\sum_{i,k=1}^{S}\sum_{j=1}^{n} C_{ij}C_{kj})/(\frac{n}{2})}{(C/n)^2} = \frac{2n\sum_{i,k=1}^{S}\sum_{j=1}^{n} C_{ij}C_{kj}}{(n-1)C^2}$$

作为 λ 的一个原始估计量。对 $\hat{\lambda}_0$ 和 $\hat{\lambda}_1$ 分别实施刀切估计得到 $\hat{\lambda}_2$ 和 $\hat{\lambda}_3$ 两个刀切估计量。他认为 $\hat{\lambda}_2$ 和 $\hat{\lambda}_3$ 不仅消除了阶为 $1/n$ 的偏差,还可以给出方差估计,因此,它们比 $\hat{\lambda}_0$ 和 $\hat{\lambda}_1$ 要优越。然而,对于大的样方数量 n,$\hat{\lambda}_2$ 的偏差是 $\hat{\lambda}_3$ 的 s 倍。在他的例子中,要使 $\hat{\lambda}_2$ 的偏差不超过 5%,样方数量至少要 30,而对 $\hat{\lambda}_3$ 只需 3 个样方。因此,$\hat{\lambda}_3$ 优于 $\hat{\lambda}_2$。总的来说,Simpson 指数也要比 Shannon 指数优越[14]。

Heltshe 和 Forrester 利用模拟技术研究了在样方抽样时,应用刀切法估计 Brillouin 和 Simpson 多样性指数的情况,详细考察了样方大小、样方数量和抽样面积对估计行为的影响。发现较小的样方对 Brillouin 指数趋于给出较小偏差和较小方差的估计;对 Simpson 多样性指数来说,在样方抽样下刀切估计是无偏(或接近无偏的)。但是,如果样方数量很大,则估计的标准差有点偏高,致使估计的 95% 置信区间的的覆盖率过高。他还发现,要使刀切估计得到改进,不仅样方大小是重要的,而且样本中的总个体数也是重要的[3]。

陈华豪根据 11 个林班(共 57 块标准地)的分树种胸高断面积调查资料,用刀切法分别对每个林班进行了 Simpson 和 Shannon 多样性指数的点估计、标准差估计和置信区间估计;他还将 57 个标准地作为来自一个更大的总体的样本,检验了 57 个刀切虚拟值的正态性,并接受了正态性假设,从而验证了 Turkey 猜想的适用性[9],也为刀切法应用于森林群落多样性指数的估计开创了先例。

另外,刀切法也被 Hatton 等应用于植物群落的 α、β、γ 多样性指数及其方差的估计[15]。

自助法在多样性指数研究中应用较少,仅见 Brokaw 将其应用于与林窗动态有关的镶嵌多样性的研究中[16]。

3.2　在种的丰富度和均匀度指数研究中的应用

种的丰富度就是一个群落中种的总数。对一个小的可普查的群落来说,很容易数出种的数目。但是,对于一个大的不可普查的群落,测度种的丰富度就出现了困难。因为样本含量越大,得到的种就会越多。因此,在样本中发现的种数最多是群落中总种数的下限。Good 和 Engen 也曾对群落中的总种数做过估计,但他们都假设抽样是对个体的随机取样[17]。这种取样方式在实际操作中的困难性,前面业已述及。因此,传统的方法已经受到了挑战。

Heltshe 和 Forrester 研究了在样方取样的情况下,用刀切法估计群落中种的丰富度及其方

差和置信区间的问题。他假设一个随机样本中包含有 n 个样方，$y^0 = s$ 为在 n 个样方中发现的种数，f_j 为包含 j 个单一种（仅在一个样方中出现的种）的样方数，则有：$\sum_{j=0}^{y^0} f_j = n$，并令 $\sum_{j=0}^{y^0} jf_j = k$

其中，k 是样本中出现的单一种的总数。于是，可以得到种的丰富度 S 的刀切估计：

$$\hat{S} = y^0 + [(n-1)/n]k$$

$\hat{S}$ 的方差为：$\mathrm{Var}(\hat{S}) = \frac{n-1}{n}\left(\sum_{j=1}^{y^0} j^2 f_j - k^2/n\right)$

近似的 95% 置信区间为：$\hat{S} \pm t_{n-1,0.025}[\mathrm{Var}(\hat{S})]^{1/2}$

他还模拟了分别包含 10 个种、25 个种以及高、轻度偏斜［指各个种的多度大小差别，差别大者为高度偏斜（high skew），差别小者为轻度偏斜（low skew）］组合而成的 4 种“群落”，对每种“群落”考察了样方大小、样方数量和抽样面积对刀切估计行为的影响[17]。他认为，当“群落”中包含 10 个种时，不管“群落”是轻度还是高度偏斜，刀切法的估计行为都是好的，但是 95% 的双边置信区间的覆盖率还是相当高。当“群落”中包含 25 个种时，刀切估计对富集种与对稀疏种同样敏感，尽管标准差和平均置信区间长度都较高，但对轻度偏斜的“群落”来说，其置信区间的覆盖率与上述情况相似。但是对于高度偏斜的“群落”来说，不管样方数量和样方大小怎样，置信区间的覆盖率都很低，平均置信区间长度也很长。由于从抽样中得到的种数总是群落中种数的下限，当群落中包含大量的稀疏种时，刀切估计量的偏差可能比原始估计量 S_0（即样本中观察到的种数）的大。但是对于模拟的具有 10 个种的两个“群落”和一个具有 25 个种轻度偏斜的“群落”来说，在至少抽取 80 ~ 100 个样方时，刀切估计的偏差才比 S_0 的小；对于高度偏斜的且有 25 个种的“群落”来说，在至少抽取 2500 个样方时，刀切估计量的偏差才比 S_0 的小，但偏差还是比较大，致使置信区间不能包含真值。他相信如果增加样方数量，可以改进刀切法的估计行为，并使其偏差比 S_0 的小。使这个偏差减小的一个可能途径是用二阶刀切或者用广义的刀切方法，这正是 Adams 和 McCune 在研究 Shannon 多样性指数时所使用的方法[17]。

Smith 和 Belle 讨论了在样方抽样下，用刀切和自助两种方法估计种的丰富度的问题，并给出了估计式[7]。k-阶刀切估计为：

$$J_n^k(\hat{S}) = S_0 + \left[\sum_{j=1}^{k} r_{1(j)} \sum_{i=j}^{k} (-1)^{i+1} \binom{k}{i} (n-i)^k \binom{n-i}{i-j} / \binom{n}{i}\right] / k!$$

其中，S_0 为观察到的种数，r_{1i} 为仅在样方 i 中发现的种数，r_{1ij} 为仅在样方 i 和 j 中发现的种数，$r_{1(1)} = \sum_{i=1}^{n} r_{1i}$ 为刚好在一个样方中发现的种数，$r_{1(2)} = \sum_{i<j} r_{1ij}$ 为刚好在两个样方中发现的种数，等等。

当 $k=1$ 时，有 $J_n^1(S) = S_0 + \{r_{1(1)}(n-1)\}/n$

当 $k=2$ 时，有 $J_n^2(S) = S_0 + [\{r_{1(1)}(2n-3)\}/n - \{r_{1(2)}(n-2)^2\}/\{n(n-1)\}]$

一阶刀切估计的方差为：

$$\mathrm{Var}\{J_n^1(S)\} = \{(n-1)/n\}\left(\sum_{j=1}^{S_0} j^2 f_j - r_{1(1)^2}/n\right)$$，其中，f_j 为包含 j 个单一种的样方数。

自助估计为：$B_n(S) = S_0 + \sum_{j=1}^{S_0} (1 - Y_{.j}/n)^n$

方差为：

$$\mathrm{Var}(S_0) = \sum_{j=1}^{S_0} (1 - Y_{.j}/n)^n [1 - (1 - Y_{.j}/n)^n] + \sum_{j \neq k} \sum [(Z_{jk}/n)^n - (1 - Y_{.j}/n)^n (1 - Y_{.k}/$$

$n)^n]$,其中,S_0 为在自助样本中观察到的种数,$y_{\cdot j}$ 是出现种 j 的样方数,Z_{jk} 为种 j 和 k 都不存在的样方数。

他又在随机分布的模型下评价了上述估计行为。他认为,尽管在大量的稀疏种以及抽取较少的样方时,刀切法和自助法低估了真实的种数,但它们还是降低了偏差。当样方数较少时,刀切法趋于过高估计种的数目,这时自助法的估计行为较好[17]。

陈华豪根据云冷杉红松林的 319 个 2×2 m^2 的样方资料用 Heltshe 和 Forrester 的方法估计植物(不计乔木)的种数、方差及 95% 置信区间[8]。他还考虑了这样的情况,即当 $\mathrm{Var}(\hat{S})$ 较大时,估计的 95% 置信区间 $\hat{S} \pm t_{n-1,0.025}[\mathrm{Var}(\hat{S})]^{1/2}$ 的左端点 $\hat{S} - t_{n-1,0.025}[\mathrm{Var}(\hat{S})]^{1/2} < y^0$,这时用 $\hat{S} \pm t_{n-1,0.25}[\mathrm{Var}(\hat{S})]^{1/2}$ 来对种数做区间估计时,其中的 $(\hat{S} - t_{n-1,0.025}[\mathrm{Var}(\hat{S})]^{1/2}, y^0)$ 这一部分实际上就没有意义。因此,他建议在这种情况下用截尾分布求取一个不对称的估计区间。

Palmer 在利用外推来估计种的丰富度时,比较了四种方法:①观察的种数;②种-面积曲线的外推;③对数正态分布的综合利用;④刀切法和自助法的估计量。通过对 1200 个样方资料的计算,得出结论:非参数估计量(一阶、二阶刀切估计量和自助估计量)比其它几个估计量都好[18,19]。

Mingoti 和 Meeden 利用存在与否的二元数据构造了 Bayes 估计量来估计种的丰富度,并将其与刀切法和自助法的估计量进行了比较,认为他的估计量比非参数估计量优越[20]。

另外,Hatton 等在进行植物群落多样性研究时,也利用刀切法对丰富度和均匀度做了点估计和方差估计[15]。Troussellier 和 Legendre 在进行微生物功能均匀度指数研究时,提出用刀切法估计该指数的标准差[21,22],得到了比较好的结果。

4　结语

综上所述,非参数统计方法的确为我们研究生物多样性测度问题提供了一条很好的途径。因为对这些指数大都不能给出其抽样分布[22],而非参数方法不需要关于种的分布的任何假设,只是用计算机算法代替了传统的数学方法,得出数值解。在样方数较少的情况下,更显出它比传统的统计方法具有优越性。这使得生态学家不必为寻找合适的统计方法而烦恼[23,24]。

但是,还应该看到这些方法并不是在任何情况下都表现得很好。例如:刀切 Shannon 或 Simpson 多样性指数,有时会给出很荒唐的结果[5,9]。因此,对这些方法在生物多样性研究中的应用还需进行系统的研究,尤其是自助法,对它的研究更嫌不足。我们可以采用模拟技术,产生生物的各种分布格局(包括种数、每个种的个体数、每个种在空间中的分布等),利用不同的抽样方法(包括系统抽样、随机抽样等以及样方大小、数量和形状等),来考察刀切法(一阶或二阶)和自助法(变化再抽样次数)对各种指数的估计行为。如能再配以实际调查数据,则会更具说服力。只有经过这样系统的研究,才能提出可供遵循的原则。

尽管如此,在抽取的样方数不多或者有大量的稀疏种时,利用这些非参数方法估计多样性及其相关指数一般会得到比较好的结果。

参　考　文　献

1　马克平. 生物群落多样性的测度方法 Ⅰ α 多样性的测度方法(上),生物多样性,1994,**2**(3):162 ~ 168

2　马克平,刘玉明. 生物群落多样性的测度方法 Ⅰ α 多样性的测度方法(下),生物多样性,1994,**2**(4):231 ~ 139

3　Heltshe J F, N E Forrester. Statistical evaluation of the jackknife estimate of diversity when using quadrat samples. *Ecology*, 1985, **66**(1):107 ~ 111

4 Pielou E C. Ecological diversity. New York: John Wiley, 1975
5 Magurran A E. Ecological diversity and its measurement. Princeton: Princeton University Press, 1988
6 Heck K L Jr, G van Belle, D Simberloff. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 1975, **56**: 1459 ~ 1461.
7 Smith E P, G van Belle. Nonparametric estimation of species richness. *Biometrics*, 1984, **40**: 119 ~ 129
8 **陈华豪. 刀切法在生态学中的应用,东北林学院学报,**1984,**12(增刊)**: 134 ~ 143
9 **陈华豪. 用刀切法估计多样性指数. 东北林学院学报,**1982,**10**(4): 87 ~ 97
10 Miller R G. The jackknife——a review. *Biometrika*, 1974, **61**: 1 ~ 15
11 **施锡铨. 非参数统计中的刀切法**(Jackknife). **应用概率统计,**1987,**3**(1): 69 ~ 76
12 **施锡铨. 关于** Bootstrap **的回顾. 应用概率统计,**1987,**3**(2): 167 ~ 177
13 Meyer J S, C G Ingersoll, L L McDonald, M S Boyce. Estimating uncertainty in population growth rates: jackknife vs. bootstrap techniques. *Ecology*, 1986, **67**(5): 1156 ~ 1166
14 Routledge R D. Bias in estimating the diversity of large, unsensused communities. *Ecology*, 1980, **61**(2): 276 ~ 281
15 Hatton T J, N E West. Early seral trends in plant community diversity on a recontoured surface. *Vegetatio*, 1987,**73**(1): 21 ~ 29
16 Brokaw N V L. Species compostition in gaps and structure of a tropical forest. *Ecology*, 1989. **70**(3): 538 ~ 541
17 Heltshe J F, N E Forrester. Estimating species richness using the jackknife procedure. *Biometrics*, 1983,**39**: 1 ~ 11
18 Palmer M W. The estimation of species richness by extrapolation. *Ecology*, 1990, **71**(3): 1195 ~ 1198
19 Palmer M W. Estimating species richness: the second-order jackknife reconsidered. *Ecology*,1991,**72**(4): 1512 ~ 1513
20 Mingoti S A, G Meeden. Estimating the total number of distinct species using presence and absence data. *Biometrics*, 1992, **48**: 863 ~ 875
21 Troussellier M, P Legendre. A functional evenness index for microbial ecology. *Microbial Ecology*, 1981, **7**: 283 ~ 295