MÉTODOS COMPUTACIONALMENTE INTENSIVOS NA ESTIMAÇÃO DO NÚMERO DE ESPÉCIES.

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Srta. Luzia Aparecida Trinca e aprovada pela Comissão Julgadora.

Campinas, 01 de fevereiro de 1988.

Profa. Dra. Gabriela Stangenhaus
Orientadora

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística.

Aos meus pais

AGRADECIMENTOS

A Professora Gabriela Stangenhaus pela orientação, colaboração, sugestões, apoio e amizade sempre dedicados.

Ao Professor Miguel Petrere Junior pelo incentivo, colaboração e amizade, e por me abrir as portas deste caminho.

Aos Professores do Departamento de Estatística pelos ensinamentos e colaboração concedidos.

Aos Professores Thomas Michael Lewinsohn (Departamento de Zoologia - I.B.) e George John Shepherd (Departamento de Botânica - I.B.) pelas idéias, sugestões e discussões, e pelo oferecimento do conjunto de dados de campo por parte do professor Shepherd.

Ao pessoal do CPD-IPEN - São Paulo, em especial a Wanda, pelo auxílio na execução dos programas no computador.

A Iara Rehder pela eficiência, coloboração e amizade dedicada em todos os momentos e pelos trabalho de datilografia.

A Luiza Costa Mota Paraíba pela grandiosa colaboração na correção da redação deste trabalho.

Ao pessoal do Centro de Computação da UNICAMP pela colaboração nos problemas computacionais.

Aos funcionários do IMECC pelas facilidades concedidas.

Ao CNPq, Capes e UNICAMP pelo apoio financeiro.

Aos colegas da Pós-Graduação pelo apoio, amizade e pelos bons momentos que passamos juntos.

A aqueles que me cativaram.

ÍNDICE

RESUMO

Em diversos estudos de ecologia de comunidades tem-se como interesse a diversidade de espécies medida, geralmente, através de um índice. Vários índices de diversidade são discutidos na literatura biológica e nos últimos anos tem-se recomendado o uso de medidas mais simples e objetivas, como o número de espécies na comunidade.

Neste trabalho o método bootstrap foi aplicado à estimação do número de espécies e de sua variância quando a amostragem é por quadrats. O desempenho do método bootstrap, sob diferentes riquezas de espécies e diferentes relações espécie-abundância, foi avaliado quando o tamanho da amostra e o tamanho dos quadrats variam, usando-se comunidades simuladas. Para comunidades com poucas espécies raras boas estimativas foram obtidas para qualquer tamanho de amostra e de quadrat. Nesse caso, as faixas de confiança para a curva do número de espécies em relação ao tamanho da amostra mostraram-se eficientes na indicação do tamanho de amostra suficiente para estimar o número total de espécies em uma amostragem sequencial. Para comunidades com muitas espécies raras, o método bootstrap, em geral, subestimou ligeiramente o número de espécies. O vício das estimativas tendeu a zero com o aumento do tamanho da amostra e essa convergência foi mais rápida para quadrats maiores. Esses

resultados indicam que, talvez, alguma correção a ser determinada, venha a melhorar o desempenho do método bootstrap, eliminando o vício. Mesmo no caso de espécies muito raras, a construção de faixas de confiança para a curva do número de espécies em relação ao tamanho da amostra pode ser usada como um guia em uma amostragem sequencial. Finalmente, o método bootstrap foi aplicado a um conjunto de dados da mata de Santa Genebra, Campinas, S.P.

ABSTRACT

Several community ecological studies focuse their interest in the species diversity measured, generally, by an index. Biological literature has discussed a variety of indexes and it has been recommended, in the last few years, the use of measures that are simpler and more, objective, as the number of species in a community.

The bootstrap method was applied, in this work, to estimate the number of species and its variance, when sampling is made by quadrats. The performance of the bootstrap method was analysed through simulated communities, varying the sample and quadrat sizes. Different species richness and species-abundance relations were considered. Good results were obtained for communities with few rare species, for all sample and quadrat sizes considered. In this case it was showm that confidence bands for the number of species vs. sample size curve can be efficiently used to indicate the sample size needed to estimate the total number of species in the community, in a sequential scheme. The bootstrap method slighthy underestimates the species number in communities with many rare species. This bias goes to zero as the sample size increases and faster convergence is achieved for bigger quadrats. The results lead us to think that a correction, to be determined, will improve the bootstrap performance. Even in this case, confidence bands for the species number vs. sample size curve can be used as a guide in a

sequential sampling scheme. To conclude, the bootstrap method was applied to a set of data obtained in the Santa Genebra woods, Campinas, S.P.

INTRODUÇÃO

Uma comunidade ecológica pode ser definida como sendo qualquer conjunto de animais e ou plantas vivendo em um local onde estão estabelecidas as interações indivívio-indivíduo e indivíduo-meio ambiente, interações estas, em qualquer grau. Assim, podem-se falar da comunidade de peixes de um lago tropical ou da comunidade tropical lacustre, da comunidade de abelhas que coletam seu alimento nos jardins da UNICAMP ou da comunidade de insetos ocorrendo em ambiente urbano, da comunidade vegetal da Mata de Santa Genebra ou da comunidade de plantas inferiores vivendo no tronco de uma árvore da mata de Santa Genebra, ou seja, os limites da comunidade são definidos de acordo com os interesses do pesquisador.

Um assunto que vem recebendo muita atenção, em qualquer estudo de ecologia de comunidades, principalmente com o avanço da crise ambiental, é a diversidade de espécies na comunidade. Com relação a isto, existem várias propostas de vários pesquisadores como forma de medir essa característica. O principal objetivo em estabelecer uma medida de diversidade é que, através dela, podem-se comparar comunidades diferentes no tempo e ou no espaço e se entender melhor as suas estruturas e alterações.

A maioria das medidas propostas para representar a diversidade de espécies de uma comunidade são tratadas como uma medida descritiva que depende das seguintes propriedades: do número de espécies e da distribuição de abundância das espécies na comunidade. Em certas situações, a tentativa de juntar essas duas propriedades independentes em uma medida uni-dimensional, pode trazer problemas na interpretação. Hurlbert (1971) em uma revisão da literatura, concluiu que a maioria dos índices usados para medir a diversidade são aplicados inapropriadamente às comunidades ecológicas e sem sentido funcional quanto às descrições das propriedades biológicas da comunidade, e propõe uma família de índices que é o número de espécies esperado em uma amostra aleatória de m indivíduos. Tal conceito de medida de diversidade é, na verdade, uma medida da riqueza de espécies da comunidade e foi o conceito utilizado quando, pela primeira vez, usou-se o termo "índice de diversidade" proposto por Fisher, Corbet e Willians (1943) (Pielou, 1975). Medidas desse tipo parecem livres dos problemas de ambigüidade dos outros índices, mas são dependentes do tamanho da amostra.

Além dos problemas relacionados à interpretação e adequabilidade das várias medidas de diversidade propostas, existem outros problemas bastante sérios que são os relacionados à amostragem e estimação para a medida escolhida. Em uma comunidade natural, o modelo de amostragem aleatória simples não se aplica. Em geral, ambientes naturais apresentam

heterogeneidade espacial e temporal e podem ser divididos em áreas com características diferentes (estratos). Nessas condições o procedimento de amostragem aleatória simples não é eficiente. Estimativas derivadas com base no modelo Multinomial são inapropriadas e subestimam a variância real (Kempton, 1977). Na prática, a amostragem por área (quadrats ou parcelas) é mais viável, no entanto, este tipo de amostragem apresenta um efeito de borda ("edge-effect"), ou seja, a possibilidade do observador incluir um indivíduo que não deveria ser incluído dentro do quadrat ou vice versa. Quanto menor o quadrat, maior o comprimento da fronteira do quadrat por unidade de área e consequentemente, maior a possibilidade de efeito de borda significante. Em situações onde os indivíduos das espécies, sob consideração, têm uma área grande ou não apresentam limites bem definidos, é aconselhável que os quadrats não sejam muito pequenos (Greig-Smith, 1983).

Considerando, em particular, a estimação da riqueza de espécies na comunidade, a contagem direta do número de espécies na amostra subestima o verdadeiro número de espécies na comunidade e esse problema torna-se mais grave quando trata-se de comunidades com muitas espécies raras. Diante disso, se colocam duas questões fundamentais. Qual deve ser o tamanho da amostra para se obter estimativas razoáveis? Qual a precisão de tais estimativas? Pielou (1975) tentou resolver o problema selecionando amostras de quadrats e acumulando-as até obter um

comportamento assintótico da curva do número de espécies em função do tamanho da amostra. Mas para comunidades com muitas espécies raras, esse comportamento pode ser obtido muito antes de terem-se observado todas as espécies.

Como era de se esperar, o número de espécies na amostra depende do tamanho da amostra. Sanders (1968) propôs o método da rarefação para a correção desse vício e concluiu que seu método é mais dependente da forma da curva de abundância das espécies do que do tamanho da amostra. Outros métodos estatísticos, como o jackknife (Quenouille, 1949) e o bootstrap (Efron, 1979) estão sendo propostos e estudados, na tentativa de resolver o problema.

Nesse trabalho, realizou-se um primeiro estudo do comportamento do método bootstrap na estimação da curva do número de espécies em função do tamanho da amostra, usando a estimativa bootstrap corrigida para o vício, proposta por Smith e Van Belle (1984), enfatizando a possibilidade de obtenção de faixas de confiança para tal curva. No capítulo 1, fez-se uma revisão e discussão do problema relacionado às medidas de diversidade. No capítulo 2, os métodos jackknife e bootstrap são apresentados e uma revisão de estudos de suas aplicações na estimação da diversidade de espécies foi feita. No capítulo 3, fez-se um estudo da aplicação do método bootstrap a populações simuladas para a estimação do número de espécies e no capítulo 4, o método bootstrap foi aplicado a um conjunto de dados reais.

CAPÍTULO 1

REVISÃO SOBRE DIVERSIDADE

1.1 Curvas espécie-área e espécie-abundância.

Desde o início deste céculo, vem sendo mostrada em vários estudos, a relação existente entre o número de espécies e a área que elas ocupam (Brenner, 1921; Jaccard, 1902, 1908; Arrhenius, 1921; Gleason, 1925; Hopkins, 1955-em McGuinness, 1984).Em geral, grandes regiões menores e essa variedade, dentro de um mesmo grupo taxônico, tende a aumentar com o decréscimo da latitude (Fischer, 1960; Pianka, 1966; Preston, 1960; Willians, 1964; Mac Arthur e Wilson, 1967 e Simberloff, 1972 - em Connor e McCoy, 1979). Há muito, essas variações têm intrigado os pesquisadores e conduzido grande parte da pesquisa ecológica moderna. Princípios originados a partir dos padrões espécie-área são usados no delineamento e manejo de reservas naturais e até mesmo, na predição do número esperado de espécies perdidas em certos tipos de manejo (Abele & Connor, 1979; Simberloff & Abele, 1982; East, 1983; Miller & Harris, 1977 - em McGuinness, 1984). Curvas que relacionam o número de espécies e a área são construídas e muito tem-se debatido sobre a forma dessa relação, sua

interpretação e razões para sua existência. Uma curva espécie-área é afetada por vários fatores característicos da comunidade em questão, mas, em geral, as formas obtidas para dados de campo são similares, tanto para plantas como para animais, sugerindo que um processo simples e geral caracteriza as curvas (Kobayashi, 1979). Alguns modelos matemáticos foram propostos para simplificar e explicar a relação espécie-área. Arrhenius (1920), propôs o então chamado modelo da função potência, dado pela expressão

$$S = C\ A^Z, \qquad (1.1.1)$$

onde S é o número de espécies na comunidade, A é a área, C e Z são os parâmetros, intercepto e inclinação, respectivamente, para explicar a relação espécie-área em uma comunidade.

Dados de campo de pequenas regiões mostraram um ajuste razoável do modelo, mas Arrhenius concluiu que ainda faltavam justificativas empíricas a favor do modelo. Para grandes regiões o número de espécies estimado pelo modelo é extremamente alto (McGuinness, 1984).

Baseando-se na suposição de ocupação aleatória de espaço pelos indivíduos em uma comunidade, Arrhenius (1921) propôs um segundo modelo, para a curva espécie-área, dado pela expressão

$$S = \sum_{j=1}^{S} [1 - (1-a)^{nj}] \qquad (1.1.2)$$

Aqui, a é a área amostrada; nj é o número de indivíduos na amostra pertencentes à espécie j (j=1,2, ... s) e s é o número de espécies observado.

Sob a mesma suposição, Gleason (1922) propôs uma expressão mais realista para o primeiro modelo de Arrhenius, que é o modelo exponencial (McGuinness, 1984). O modelo exponencial é dado por

$$S = Z \log A + C \qquad (1.1.3)$$

Esse modelo tem melhor adequabilidade a dados de grandes áreas do que os modelos anteriores (Evans et.al., 1955; Hopkins, 1955 - em Kobayashi, 1979) e foi amplamente usado, principalmente, por ecologistas de plantas (Connor & McCoy, 1979; Greig-Smith, 1983).

Kobayashi (1979) apresentou três modelos matemáticos da curva espécie-área, particularizando para coleções obtidas por amostragem discreta (por quadrats), por amostragem contínua (a área amostrada expande continuamente) em um habitat homogêneo e para uma comunidade delimitada (finita), respectivamente, baseando-se nas probabilidades de cada espécie ocorrer em uma determinada área. As aplicações desses modelos a dados de campo desempenharam-se com sucesso.

Apesar da curva espécie-área ser uma representação explícita da riqueza de espécies de uma comunidade, sua forma é afetada por vários fatores relacionados à comunidade em questão. Características como as abundâncias relativas das espécies, a distribuição espacial dos indivíduos, as associações interespecíficas e também o método de amostragem utilizado exercem influências fortes sobre a forma da curva (Kobayashi, 1979; Goodall, 1952 - em Kobayashi, 1979; Greig-Smith, 1983).

Fisher et. al. (1943) e Willians (1943, 1944, 1947), demonstraram que sob a suposição de tamanho da população proporcional à área, a forma exponencial da curva espécie-área é diretamente conduzida pela distribuição logarítmica ou distribuição de logséries, quando esta se ajusta aos dados de abundância das espécies. Preston (1948, 1960, 1962) propôs o modelo lognormal para a abundância relativa das espécies e Preston (1962) e Bliss (1965) mostraram que sob suposições similares as de Fisher et. al., esse modelo conduz ao modelo da função potência para a forma da relação espécie-área (Connor e McCoy, 1979).

Em geral, comunidades com muitas espécies apresentam uma variação marcante na abundância dessas espécies. Algumas delas são muito abundantes, enquanto que, a grande maioria são muito raras e tem-se, aqui, outra questão de muito interesse e que está relacionado com vários outros fatores como por exemplo, a grande

variação nos limites de tolerância ambiental das espécies, as diferenças nos habitats e as interações inter e intra-específicas (Pielou, 1975).

Vários modelos teóricos foram propostos para explicar a relação espécie-abundância em comunidades. A maioria desses modelos são baseados na suposição de que as espécies que compõem uma comunidade estão dispersas aleatoriamente no espaço, isto é, o número de indivíduos da espécie j é uma variável de Poisson com parâmetro $\lambda_j (j=1,2, \ldots ,S)$ e a distribuição do número de espécies com exatamente r indivíduos tem portanto, a forma de uma Poisson composta. Esse é o caso dos dois modelos citados acima (logséries e lognormal) e também do modelo binomial negativa proposto por Brian (1953), quando as espécies são agregadas (Pielou, 1975).

Como espécies com $r=0$ indivíduos são não observáveis, usam-se as formas truncadas dessas distribuições.

Supondo-se que os diferentes valores de λ_j $(j=1,2,\ldots,S)$ constituem S variáveis aleatórias com distribuição gama com parâmetros k e p, a probabilidade de que qualquer espécie seja representada por r indivíduos é

$$q_r = \int_0^{\infty} \frac{\lambda^r e^{-\lambda}}{r!} \frac{p^{-k}}{\Gamma(k)} \lambda^{k-1} \exp[-\lambda/p]\, d\lambda \qquad (1.1.4)$$

$$= \frac{\Gamma(k+r)}{r!\,\Gamma(k)} \left[\frac{p}{1+p}\right]^r \left[\frac{1}{1+p}\right]^k \qquad (1.1.5)$$

para $r = 0,1,2,\ldots$, que é o termo geral da distribuição binomial negativa e

$$q'_r = \frac{q_r}{1-q_0} = \frac{\Gamma(k+r)}{r!\Gamma(k)} \left[\frac{p}{1-p}\right] \frac{1}{(1+p)^k - 1} \qquad (1.1.6)$$

($r = 1,2, \ldots$) é o termo geral da distribuição binomial negativa truncada em zero. Os parâmentros p e k podem ser estimados através do método dos momentos, tornando possível a estimação de S. Igualando as frequências observadas e esperadas das espécies não observáveis tem-se

$$E(S - s) = S(1 + p)^{-k} \qquad (1.1.7)$$

e então,

$$\tilde{S} = \frac{s}{1-(1+\tilde{p})^{-k}} \quad , \qquad (1.1.8)$$

onde s é o número de espécies observado, $\tilde{p}$ e $\tilde{k}$ são estimativas dos parâmentros p e k, respectivamente. Mas as variâncias amostrais de $\tilde{p}$, $\tilde{k}$ e $\tilde{S}$ são desconhecidas. p é função de S, k e N (o número total de indivíduos)

$$p = N/kS \qquad (1.1.9)$$

e k está relacionado à forma da distribuição espécie-abundância. Quanto mais longa a calda da distribuição, menor é o valor de k.

De (1.1.8) e (1.1.9) tem-se a forma com que s aumenta com N.

$$s = S \left[1 - (1 + N/kS)^{-k} \right] \qquad (1.1.10)$$

Essa equação é a conhecida curva de coletor que relaciona o número de espécies com o esforço de coleta.

Pielou (1975) mostrou que, quando o parâmetro k na expressão (1.1.6) tende a zero, obtem-se a expressão da distribuição de logséries ou logarítmica. A idéia de se fazer k

tender a zero foi sugerida por Fisher (em Fisher, Corbet e Willians, 1943), já que muitos conjuntos de dados analizados por ele produziam valores de k muito próximos a zero, apesar de que na expressão (1.1.6) k deve ser finito. Como k representa, inversamente, a variabilidade da distribuição, quando k tende a zero tem-se que a variabilidade entre as frequências das espécies é muito grande.

Fazendo-se $p = X/(1-X)$, o modelo de logséries é dado por

$$\pi_r = \lim_{k \to 0} q'_r = \gamma \frac{X^r}{r!} \quad , \qquad (1.1.11)$$

onde π_r é a probabilidade de uma espécie ter r indivíduos, $\gamma = -1/\ln(1-X)$ e X são os parâmetros da distribuição. como γ é função de X, a distribuição tem, na verdade, apenas um parâmetro. X é um fator de escala e γ depende da forma da distribuição. Usualmente, s e $\alpha = s\gamma$ são usados para descrever a distribuição de espécie-abundância de uma amostra. s (o número de espécies da amostra) e não S (o número de espécies na comunidade) é usado na expressão de α para descrever a população, isto porque quando se supõe que a amostra provém de uma comunidade cujas abundâncias das espécies se distribuem de acordo com o modelo de logséries, está se supondo, também, que S é infinito, o que pode ser visto através da expressão da curva de coletor,

$$s = \alpha \ln \left[1 + \frac{N}{\alpha} \right] . \qquad (1.1.12)$$

Supõe-se que o número de espécies na amostra cresce indefinidamente com o tamanho da amostra.

Um algorítmo iterativo para estimar α e X foi apresentado por Birch (1963). Anscombe (1950) mostrou que um estimador da variância amostral de $\hat{\alpha}$ é dado por

$$\hat{Var}(\hat{\alpha}) \cong \frac{\hat{\alpha}}{\ln X(1-X)} . \qquad (1.1.13)$$

Outro modelo bastante usado para explicar a relação espécie-abundância é o lognormal. Se existe uma fonte de recurso limitante para as muitas espécies que constituem a comunidade, que é dividida de forma que cada parcela é uma variável aleatória cuja função de distribuição é aproximadamente lognormal e se a abundância de cada espécie é proporcional à parcela de recurso utilizada, então a distribuição espécie-abundância é lognormal, cuja função densidade de probabilidade é

$$\phi(y) = \frac{1}{y\sqrt{2\pi V_z}} \exp\left[\frac{-(\ln y - \mu_z)^2}{2V_z}\right] \qquad 0<y<\infty , \qquad (1.1.14)$$

onde Y é a abundância ou o "tamanho" das espécies, μ_z e V_z são, respectivamente a média e a variância de Z, uma variável aleatória normalmente distribuída ($\ln Y = Z$). A média e variância de Y são

$$\mu_y = \exp\left[\mu_z + \frac{V_z}{2}\right] \quad \text{e} \quad V_y = (\exp[V_z] - 1)\exp[2\mu_z + V_z], \qquad (1.1.15)$$

Como em geral as abundâncias das espécies são medidas por contagem do número de indivíduos, sendo, portanto, variáveis discretas, considera-se as abundâncias como sendo variáveis de Poisson cujo parâmentro λ é lognormalmente distribuído. Tem-se, assim a distribuição lognormal discreta (Anscombe, 1950; Pielou, 1969) ou a Poisson lognormal (Holgate, 1969; Bulmer, 1974), onde a probabilidade de que uma espécie contenha r indivíduos é

$$\pi_r = \frac{1}{r!\sqrt{2\pi V}} \int_0^1 \frac{1}{\lambda} \exp\left[-\lambda + r\ln\lambda - \frac{(\ln\lambda - \mu)^2}{2V}\right] d\lambda \qquad (1.1.16)$$

para $r = 0,1,2,...$

A Poisson lognormal apresenta dificuldades computacionais. Bulmer (1974) obteve uma fórmula, aproximada para calcular as probabilidades π_r para $r \geq 10$, mas para $r < 10$ nenhuma técnica alternativa foi encontrada para calcular a integral em (1.1.16). Pielou (1975) apresenta um procedimento para a estimação dos parâmetros e da variância dos estimadores para a distribuição lognormal. Uma estimativa de S pode ser obtida, porém sua variância amostral é desconhecida.

Trabalhos práticos constataram a adequabilidade desses modelos teóricos a dados reais, mas será que eles são convincentes ecologicamente, assim como são matematicamente e ou estatísticamente? Em alguns casos, vários modelos diferentes se ajustam ao mesmo conjunto de dados, dificultando as conclusões.

Outros modelos foram propostos sob a suposição da existência de um recurso limitante ao tamanho da população (por exemplo, alimento) que é dividido entre as espécies coexistentes, segundo alguma regra. Tem-se, então, o modelo da ocupação do nicho ("niche preemption") ou das séries geométricas proposto por Whittaker (1972) e May (1975), que considera a partição de um recurso limitante entre as espécies como uma série geométrica. O modelo do segmento dividido ("broken stick") de MacArthur (1957), que considera o recurso limitante como um segmento de uma unidade de comprimento, que é quebrado em S partes disjuntas, cujos $(S-1)$

pontos de quebra são alocados aleatoriamente e os S segmentos são considerados os tamanhos ou as abundâncias das S espécies, respectivamente. Esse modelo é uma particularização do anterior para o caso em que a abundância das espécies é aproximadamente uniforme. Por último tem-se o modelo da sobreposição de nicho, ("overlapping niche"), também proposto por MacArthur (1975) que difere dos outros dois porque não pressupõe que existe uma fonte de recurso limitante, que é dividida entre os competidores, mas sim que as espécies são independentes e cada uma toma aquilo que necessita para sobreviver. Geralmente esses modelos se adequam a pequenos conjuntos de dados (Pielou, 1975; May, 1975).

A curva espécie-abundância, por sua vez, tem a forma dependente do número de espécies existentes na comunidade e da forma com que os indivíduos estão distribuídos entre as espécies. Quando uma distribuição teórica se ajusta aos dados de freqüência de abundância das espécies, os parâmentros do modelo caracterizam as propriedades da curva. Um problema que surge aqui está relacionado à estimação desses parâmetros. Para a maioria dos modelos não é possível obter-se boas estimativas e quando o é, nada se sabe sobre a precisão dos possíveis estimadores.

Com o avanço da crise ambiental dos últimos anos, o enfoque abordado pelos ecologistas de comunidades está ligado ao interesse em comparar comunidades diferentes no espaço e ou no tempo, com relação às curvas espécie-abundância. Tais comparações tornam-se complexas, por causa dos problemas de ajuste e

estimação dos parâmetros de modelos teóricos e, às vezes, sem sentido, quando os mesmos conjuntos de dados são ajustados por modelos diferentes. Surge, então, o interesse em construir ou estabelecer uma medida descritiva, independente do modelo teórico ajustado, que possa ser usado para qualquer comunidade e através da qual se possa comparar comunidades diferentes.

1.2 Medidas de Diversidade

Se os dados coletados de uma comunidade consistem de uma lista do número de indivíduos que pertencem a cada espécie, N_1, N_2, ... ,N_S, onde S é o número total de espécies na comunidade e se esses dados são colocados em um gráfico de barras, duas estatísticas descritivas são necessárias para caracterizá-lo. A mais óbvia delas é S que mede a amplitude do gráfico e a segunda estatística deve representar a sua forma. Se os N_j (j=1,2, ... ,S) fossem freqüências de uma variável quantitativa, a variância seria uma estatística adequada para caracterizar a forma da distribuição. Mas, espécies são variáveis qualitativas e não tem sentido se ordenar ou se associar pesos às freqüências e, portanto, não tem sentido se falar em variância. Uma propriedade da comunidade que poderia ser vista como análoga à variância, nesse caso, é a sua diversidade de espécies (Pielou, 1975).

Muitas formas de definir e medir a diversidade foram propostas, mas seu conceito e interpretação ainda são obscuros. Intuitivamente, uma medida de diversidade, com sentido análogo ao da variância, é uma função que leva em consideração a riqueza de espécie e a equitabilidade das espécies (Loyd e Ghelardi, 1964; Hurlbert, 1971; Pielou 1975). A equitabilidade é uma propriedade da comunidade que é máxima quando todas as espécies tem a mesma abundância ou seja, a distribuição dos indivíduos da comunidade está equilibrada entre as espécies, e decresce com o aumento das disparidades entre as abundâncias.

Em princípio, a função $H'(p_1, p_2, \ldots, p_S)$ onde p_j é a proporção da espécie j (j=1,2, ... ,S), satisfazendo as propriedades i-ii-iii, seria a medida de diversidade ideal.

i. Dado S, o número de espécies na comunidade, H' deve ser máxima quando p_j=1/S para todo j(j=1,2, ... ,S), isto é, a diversidade é máxima quando a equitabilidade é máxima.

ii. Dado duas comunidades com equitabilidade máxima, uma com S spécies e a outra com S+1 espécies, a última deve ter H' maior.

iii. Se os membros da comunidade estão sujeitos a classificações diferentes, por exemplo, uma classificação A com a classes e

uma classificação B com b classes, e se p_i (i=1,2, ... ,a) é a proporção dos membros da comunidade pertencentes à i-ésima classe da A e q_{ij} (i=1,2, ... ,a ; j=1,2, ... ,b) é a proporção desses membros pertencentes a j-ésima classe de B, então, $H_{ij} = p_i q_{ij}$ é a proporção de membros da comunidade que pertencem à i-ésima classe de A e à j-ésima classe de B e H' deve ser dada pela expressão

$$H' = H'(AB) = H'(A) + H'_A(B) \qquad (1.2.1),$$

onde $H'(A)$ é a diversidade sob a classificação A; $H'_A(B) = \sum_{i=1}^{a} p_i H'_i(B)$ é a diversidade de B sobre todas as classes de A; $H'_i(B)$ é a diversidade sob a classificação B dentro da i-ésima classe de A.

Se as classificações são independentes, tal que, $q_{ij} = q_j$ para todo i, então

$$H_i(B) = H'(B) \qquad (1.2.2) \text{ para todo i;}$$

$$H'_A(B) = \sum_{i=1}^{a} p_i H'_i(B) = H'(B) \qquad (1.2.3)$$

e então

$$H'(AB) = H'(A) + H'(B) = H' \qquad (1.2.4)$$

É possível mostrar que a única função que satisfaz as três propriedades i-ii-iii é dada pela expressão

$$H' = -C \sum p_i \log p_i \ , \qquad (1.2.5)$$

onde C é uma constante positiva (Pielou, 1975).

Existe uma classe geral de funções usadas em teoria da informação, onde, em alguns casos especiais são usadas como um índice de diversidade. Dado um código com s símbolos diferentes, a função em (1.2.6) é conhecida como a entropia de ordem α do código e mede o grau de "desorganização" do código, onde $p_i (i=1,2, \ldots ,s)$ é a proporção do i-ésimo símbolo (Pielou, 1975).

$$H_\alpha = \frac{\log \sum p_i^\alpha}{1 - \alpha} \qquad (1.2.6)$$

Fazendo $\alpha \longrightarrow 1$, tem-se:

$$H_1 = \lim_{\alpha \longrightarrow 1} H_\alpha = - \Sigma\, p_i \log\, p_i \qquad (1.2.7)$$

que é a expressão (1.2.5) quando C=1. Essa função é o conhecido índice de Shannon-Weaver, amplamente usado pelos ecologistas como uma medida de diversidade.

Na teoria de informação, H' mede a informação por símbolo de um código. A partir desse código muitas mensagens podem ser obtidas (ou amostras removidas) sem o código se esgotar. Fazendo uma analogia à Ecologia, o código seria a comunidade, infinitamente grande tal que, a remoção de amostras não causam mudanças perceptíveis, e os símbolos seriam as espécies.

Para comunidades pequenas, que podem ser observadas por completo, as chamadas coleções, a medida de diversidade usada, equivalente a H', é o índice de Brillouin. Na teoria da informação esse índice é usado para medir a informação por símbolo de uma mensagem particular. O índice de Brillouin é dado pela expressão

$$H = \frac{1}{N} \log \frac{N!}{\prod_{j=1}^{S} N_j!}\,, \qquad (1.2.8)$$

onde N é o número total de indivíduos na coleção, N_j é o número de indivíduos pertencentes à espécie j (j=1,2, ... ,S e $\Sigma N_j = N$) e S é o número de espécies na coleção.

Pode-se mostrar que $\lim_{\min(N_j)\longrightarrow\infty} H = H'$, ou seja o índice de Brillouin é equivalente ao índice de Shannon-Weaver quando a comunidade tende a ser infinitamente grande.

Voltando à expressão (1.2.6) e fazendo $\alpha \longrightarrow 2$ tem-se a expressão (1.2.9), o índice de Simpson.

$$H_2 = - \log \Sigma p_j^2 \qquad (1.2.9)$$

Simpson (1946) interpretou a função $\lambda = \Sigma p_j^2$ como sendo uma medida de "dominância" ou de "concentração". λ é a probabilidade de que quaisquer dois indivíduos selecionados aleatoriamente e independentemente da comunidade sejam da mesma espécie. Se λ é grande significa que a comunidade é relativamente homogênea em termos de espécies. Então, intuitivamente, dominância é o oposto de diversidade e funções decresoentes de λ foram propostas para medir a diversidade. As expressões (1.2.9) e (1.2.10) são alguns exemplos.

$$D = 1 - \lambda \quad . \qquad (1.2.10)$$

O índice H_2, da expressão (1.2.9), tem sido mais aceito, talvez por sua relação a H', da expressão (1.2.5), nas quais, ambas são medidas de entropia de ordem 2 e 1, respectivamente (Pielou, 1975).

Em certas situações, a tentativa de representar a riqueza de espécies e o vetor de abundância das espécies da comunidade em uma medida uni-dimensional pode trazer dificuldades na sua interpretação. Comunidades completamente diferentes podem ter o mesmo índice de diversidade, e a maioria delas são insensíveis às espécies raras (Kempton, 1979; Smith e Grassle, 1977; Routledge, 1979; Hurlbert, 1971), e portanto, comparações entre comunidades diferentes, com respeito a um índice pode não ter sentido. Hurlbert (1971) concluiu que a maioria dos índices de diversidade, inclusive os baseados na teoria da informação, são aplicados inapropriadamente à comunidades ecológicas e sem sentido funcional quanto às descrições das propriedades biológicas. A relação teoria de informação-teoria ecológica de comunidades ainda não é aceita e a analogia entre código e comunidade pode ser apenas superficial. Porém, a ambigüidade e outros problemas de interpretação de um índice em particular podem ser minimizados se sua aplicação for restrita a comunidades compatíveis para comparação.

Quando pela primeira vez, o termo índice de diversidade foi usado (Fisher, Corbet e Willians (1943) em Pielou, 1975) problemas de ambigüidade não existiam. O sentido da palavra diversidade estava relacionado à riqueza de espécies e a medida proposta para representá-la correspondia ao parâmetro α da distribuição de logséries ajustada aos dados de abundância das espécies. Nestas condições, α satisfaz a expressão

$$S = -\alpha \ln(1-X), \qquad (1.2.11)$$

onde S é o número total de espécies e $0<X<1$.

Outras medidas baseadas apenas na riqueza de espécies foram propostas para medir a diversidade de uma comunidade (Sanders 1968; Hurlbert 1971; Fager 1972; Peet 1974 e Raup 1975 em Smith e Grassle 1977).

Hurlbert (1971) propôs uma família de medidas de diversidade, aperfeiçoando o método da rarefação de Sanders (1968), definida pelo número de espécies esperado quando m indivíduos são selecionados aleatoriamente de uma comunidade. Para m fixo, a contribuição de cada espécie à medida de diversidade é dada pela probabilidade daquela espécie aparecer entre os m indivíduos selecionados aleatoriamente da população.

Para m pequeno, a medida é dominada pelas espécies abundantes e para m grande a medida passa a ser sensível às espécies raras. Para m=2, a medida está relacionada ao índice de Simpson. Supondo-se uma população (comunidade) finita consistindo de S espécies, cada uma com n_j indivíduos (j=1,2, ... ,S) e $\mathbf{n}$ = (n_1,n_2, ... n_S) representando a população toda, s_m é uma variável aleatória que corresponde ao número de espécies em uma amostra de m indivíduos. O número de espécies esperado em uma amostra aleatória sem reposição, dada a população $\mathbf{n}$, S_m, é

$$S_m = E[s/\mathbf{n}] = \sum_{j=1}^{S} \left[1 - \frac{\binom{n - n_j}{m}}{\binom{n}{m}} \right], \qquad (1.2.12)$$

onde n é o número total de indivíduos na comunidade, $n = \sum_{j=1}^{S} n_j$. S_m é a família de medidas de diversidade sugerida por Hurlbert (1971). Para m grande $S_m \approx S$, o número total de espécies na comunidade.

Smith e Grassle (1977), supondo uma população multinomial π = (π_1,π_2, ... ,π_K) onde π_j é a proporção de indivíduos da espécie j na população, generalizaram a expressão (1.2.12) para uma população infinita. O modelo de amostragem multinomial assume que a população é infinita e que existe independência entre os

indivíduos amostrados, ou seja, assume que não existe interação intra e inter-específicas. Sob essas condições S_m é dado por

$$S_m = E(s/\pi) = \sum_{j=1}^{S} [1 - (1 - \pi_j)^m]. \qquad (1.2.13)$$

Estes autores apresentaram um estimador não tendencioso e de variância mínima para S_m, dado por

$$\hat{S}_m = E(s/N) = \sum_{j=1}^{s} \left[1 - \frac{\binom{N - N_j}{m}}{\binom{N}{m}} \right] \qquad (1.2.14)$$

onde $N = (N_1, N_2, \ldots, N_S)$ é o vetor que descreve uma amostra aleatória de tamanho $N = \sum_{j=1}^{S} N_j$ e N_j é o número de indivíduos da espécie j, e apresentaram uma estimativa para a variância de S_m.

O próprio número de espécies (S), na comunidade, tem sido usado como uma medida de diversidade, porém com um sentido diferente daquele proposto por Pielou. Medidas desse tipo têm interpretações mais claras e são, talvez, mais objetivas.

CAPÍTULO 2

MÉTODOS COMPUTACIONALMENTE INTENSIVOS E DIVERSIDADE

2.1 Introdução

Além das dificuldades quanto à adequabilidade e interpretação de uma medida de diversidade para uma comunidade ecológica, existem os problemas relacionados à amostragem e à estimação da medida escolhida.

Geralmente, os seres vivos tem uma ocupação contínua no tempo e espaço e se dispõem segundo padrões espaciais, as chamadas manchas de comunidades ecológicas, decorrentes das leis naturais que regem a sobrevivência. Nessas condições, na prática, só é possível se obter amostras aleatórias de espaço, ou seja, a amostragem é de área, volume, grupos de indivíduos em armadilhas e não de indivíduos independentes. Portanto, o modelo de amostragem aleatória simples (multinomial) não se aplica, e estimativas de diversidade baseadas nesse modelo padrão são inapropriadas.

Para comunidades vegetais, costuma-se selecionar amostras de quadrats ou parcelas, onde porções de área são selecionadas ao acaso e as informações de interesse são

observadas, ou a amostragem é por transectos, no caso de comunidades de plantas de pequeno porte, onde seleciona-se uma direção e observa-se o que está sob a linha que passa por essa direção. Em caso de animais, em geral, a coleta é feita em grupos de indivíduos apanhados em armadilhas ou simplesmente são observados visualmente ao longo de faixas de regiões.

Na maioria dos casos, estimativas de diversidade dependem do padrão espacial dos indivíduos, assim como das características do procedimento de amostragem usado e, em geral, subestimam a diversidade da comunidade. Esse problema é mais grave quando se trata de comunidades com muitas espécies raras e pouco se sabe sobre o vício e outras medidas de precisão das estimativas. Reconhecendo esse problema, alguns autores vêm estudando métodos alternativos para a estimação da diversidade.

Pielou (1975) propôs uma técnica de estimação da diversidade e de sua variância, em particular para o índice de Shannon-Weaver. Supondo-se que foram selecionadas N unidades amostrais, sua técnica consiste em ordenar, aleatoriamente, as N unidades amostrais e calcular a estimativa de H' baseando-se na primeira unidade amostral, adicionar a segunda unidade e estimar H' novamente e assim por diante até acumular as N unidades amostrais, obtendo estimativas de H' a cada vez que se acumula a próxima unidade amostral. Com esses valores constroe-se a curva dos valores obtidos para H' com relação ao tamanho da amostra

acumulado. Se N é suficientemente grande essa curva alcança sua assíntota a partir de um certo ponto. Se para t unidades amostrais acumuladas a curva já atingiu esse comportamento, a estimativa de H' proposta por Pielou é dada pela média amostral dos valores de H', h_k, para $k = t+1, t+2, \ldots, N$, e a estimativa da variância é dada pela variância amostral da média. Se não existem fortes correlações entre os valores de h_k ($k = t+1, t+2, \ldots, N$) essas estimativas podem ser consideradas como sendo não tendenciosas. Heyer e Berven (1973) propuseram repetir esse processo para várias combinações das N unidades amostrais a fim de se obter estimativas da diversidade, com o respectivo desvio padrão.

Sanders (1968) aplicou o método da rarefação em um estudo comparativo da diversidade de animais bentônicos em vários tipos de ambientes marinhos. Seu estudo teve como objetivo definir uma nova medida de diversidade independente do tamanho da amostra, viabilizando, assim, as comparações entre comunidades de tamanhos diferentes, explicando os padrões de diversidades observados. O método da rarefação consiste em manter fixa as abundâncias das espécies observadas na amostra e simular, para amostras de tamanhos diferentes, o número de espécies esperado. Com os dados simulados constroe-se a curva do número de espécies em função do tamanho da amostra para cada ambiente. Sua medida de diversidade foi definida como sendo a taxa de incremento de espécies à curva para cada unidade amostral adicionada. Nesse

estudo, Sanders observou que cada ambiente tem sua própria taxa característica de incremento, concluindo que seu método é mais dependente da curva espécie-abundância do que do tamanho da amostra. Hurlbert (1971), Fager (1972) e Simberloff (1971) mostraram que o método de rarefação de Sanders super-estima o número de espécies esperado em uma amostra de n indivíduos. Shinozaki (1963) já havia mostrado que este valor é determinado pela distribuição hipergeométrica e Simberloff (1972) concluiu que a magnitude da super-estimação do método de rarefação é fortemente influênciada pela distribuição de abundância das espécies e pela presença de agregação espacial.

Outros métodos estatísticos como o jackknife e o bootstrap, vêm sendo estudados na estimação da diversidade quando a amostragem é por área. Esses são métodos computacionalmente intensivos que estão surgindo, graças ao avanço dos computadores, nos últimos anos, que substituem grande quantidade de cálculo pelas análises teóricas padrões. Suas principais vantagens estão no fato de possibilitarem análises estatísticas livres das suposições de modelos teóricos para a distribuição dos dados e dão margem à exploração de propriedades amostrais de interesse, independentemente de suas formas analíticas. Na seção 3.2 a metodologia geral bootstrap e o método jackknife são apresentados e na seção 3.3 uma revisão de aplicações desses métodos na estimação de diversidade é feita.

2.2 *Métodos Computacionalmente Intensivos*

2.2.1 Bootstrap

O método bootstrap é uma técnica estatística bastante recente, desenvolvida por Efron (1979), com o objetivo de resolver algumas questões que surgem em problemas práticos de estimação.

Consideremos $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de tamanho n, selecionada de uma população cuja função de distribuição de probabilidade, $F(X,\theta)$, é totalmente desconhecida. θ é um parâmetro de $F = F(X,\theta)$ e tem-se interesse em estudar algumas de suas características. Depois de escolhido o estimador adequado para estimar θ com base na amostra $\underset{\sim}{X}$, surge a questão: quão precisa é tal estimativa? Ou seja, quão confiante podem-se estar a respeito da estimativa e a partir desta se fazer inferências sobre a população? O método bootstrap é uma metodologia geral que tenta responder questões desse tipo em problemas de estimação, em situações onde métodos padrões falham ou são analiticamente difíceis de se examinar. Ele é aplicado, principalmente, na estimação da precisão estatística como desvio padrão, erro de predição, intervalo de confiança e vício para algum estimador em questão.

A idéia básica do método bootstrap está relacionada a

outros métodos mais antigos como o jackknife (Quenouille, 1949), validação cruzada e replicações balanceadas repetidas, e tem se mostrado de grande aplicação na maioria dos casos (Efron, 1979).

Voltando à amostra $\underset{\sim}{X}$, considere-se, inicialmente, uma situação bem simples. Sendo $\theta = \mu = E_F(X)$ a média populacional de F, observada a amostra $\underset{\sim}{X} = \underset{\sim}{x}$, $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, uma estimativa não tendenciosa para μ é $\bar{x}$, a média amostral dos x_i ($i=1,2, \ldots, n$). Então

$$\hat{\theta} = \hat{\mu} = \frac{\Sigma X_i}{n} \qquad (2.2.1.1)$$

é o estimador escolhido para θ. Uma medida de precisão tradicional é o desvio padrão do estimador. O desvio padrão de $\bar{x}$ é

$$\sigma_{\bar{x}}(F) = \left\{ \frac{\sigma^2(F)}{n} \right\}^{1/2} \qquad (2.2.1.2),$$

onde

$$\sigma_x^2(F) = \mu_2 = E_F(X)^2 - E_F^2(X) \qquad (2.2.1.3)$$

é o segundo momento centrado de F e E_F indica expectância sob a função de distribuição F. Porém, F é desconhecida e portanto não se conhece $\sigma^2(F)$, mas pode-se estimar $\sigma^2(F)$. Usando o procedimento padrão, uma estimativa não tendenciosa para $\sigma^2(F)$ é dada pela expressão

$$\hat{\sigma}^2(F) = \hat{\mu}_2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \qquad (2.2.1.4)$$

Logo, a estimativa padrão, não tendenciosa de $\sigma_{\bar{x}}$ é

$$\hat{\sigma}_{\bar{x}}(F) = \left\{ \frac{\Sigma(x_i - \bar{x})^2}{n(n - 1)} \right\}^{1/2} \qquad (2.2.1.5)$$

A teoria bootstrap estima tal precisão baseando-se na função de distribuição empírica, $\hat{F}$, que é uma função que coloca massa de probabilidade $1/n$ sobre cada observação da amostra, ou seja,

$$\hat{\mu} = E_{\hat{F}}(X) = \frac{\Sigma x_i}{n} \qquad (2.2.1.6)$$

é a estimativa bootstrap de θ e

$$\hat{\sigma}^{2}(F) = \sigma^{2}(\hat{F}) = E_{\hat{F}}(X)^{2} - E^{2}_{\hat{F}}(X) = \frac{\Sigma(x_i - \bar{x})^{2}}{n} \qquad (2.2.1.7)$$

é a estimativa bootstrap de $\sigma^2(F)$, e portanto,

$$\hat{\sigma}_{\bar{x}} = \sigma_{\bar{x}}(\hat{F}) = \left\{\frac{\sigma^{2}(\hat{F})}{n}\right\}^{1/2} = \left\{\frac{\Sigma(x_i - \bar{x})^{2}}{n^{2}}\right\}^{1/2} \qquad (2.2.1.8)$$

é a estimativa bootstrap do desvio padrão da média amostral. Esse exemplo simples é uma forma de mostrar a validação do método bootstrap quando métodos padrões funcionam, já que a diferença entre as expressões (2.2.1.5) e (2.2.1.8) é insignificante na maioria dos casos, permitindo, também, fácil correção de $\sigma_{\bar{x}}(\hat{F})$, dando crédito à aplicação do método em situações mais complicadas. Generalizando, o método bootstrap pode ser definido como sendo uma técnica de reamostragem que permite a estimação de distribuições amostrais para estimadores.

Considerando a amostra $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$, onde $X_i \overset{ind}{\sim} F$ $(i=1,2, \ldots, n)$ e $F = F(X,\theta)$ desconhecida, seja $R(\underset{\sim}{X},F)$ uma variável aleatória de interesse dependente da amostra $\underset{\sim}{X}$ e da função de distribuição de probabilidade F. A metodologia geral bootstrap para estimar a distribuição amostral de $R(\underset{\sim}{X},F)$ pode ser descrita como segue:

i. Constrói-se a função de distribuição de probabilidade empírica, $\hat{F}$, colocando-se massa $1/n$ sobre cada observação x_i ($i=1,2, \ldots ,n$);

ii. Com $\hat{F}$ fixa, seleciona-se uma amostra com reposição de tamanho n a partir de $\hat{F}$, ou seja,

$$X_i^* = x_i^* \quad , \quad X_i^* \sim_{ind} \hat{F} \qquad (i=1,2, \ldots ,n) \qquad (2.2.1.9)$$

$\underset{\sim}{X}^* = (X_1^*, X_2^*, \ldots , X_n^*)$ é a amostra bootstrap. Calcula-se a estimativa bootstrap de $R(\underset{\sim}{X},F)$, $R(\underset{\sim}{X}^*,\hat{F}) = R^*$;

iii. Aproxima-se a distribuição de $R(\underset{\sim}{X},F)$ através da distribuição bootstrap de R^*.

Na prática, a aplicação do método bootstrap envolve um algorítmo Monte Carlo onde seleciona-se, independentemente, um grande número de amostras bootstrap, e para cada uma dessas amostras calcula-se R^*. Com esses valores constrói-se a distribuição empírica que fornece uma aproximação para a distribuição de $R(\underset{\sim}{X},F)$. Essa metodologia é conhecida como bootstrap Monte Carlo e um algorítmo é dado nos seguintes passos:

1. Através de um gerador de números aleatórios seleciona-se, independentemente, um grande número B, de amostras bootstrap: $\underset{\sim}{x}^*(1)$, $\underset{\sim}{x}^*(2)$, ... , $\underset{\sim}{x}^*(B)$.

2. Para cada amostra $\underset{\sim}{x}^{*}(b)$ (b=1,2, ... ,B), calcula-se a estatística de interesse: $R^{*}(b) = R(\underset{\sim}{x}^{*}(b),\hat{F})$ b=1,2, ... ,B;

3. Aproxima-se a distribuição de $R(\underset{\sim}{x},F)$ pela distribuição bootstrap de $R^{*}(b)$.

Em algumas situações a forma da distribuição de $R^{*}(b)$ pode ser obtida diretamente, como no exemplo da média apresentado anteriormente, ou por expansões em séries de Taylor (Efron, 1979).

O método bootstrap tem a vantagem de evitar todas as suposições e dificuldades analíticas de qualquer tipo. Caso o pesquisador conheça qualquer característica de F, esta informação pode ser incorporada ao processo de estimação de F através de $\hat{F}$ e neste caso o processo é chamado bootstrap paramétrico.

De Efron e Tibshirani (1985), um esquema ilustrativo do bootstrap com a idéia geral do procedimento é apresentado na figura 2.2.1.1

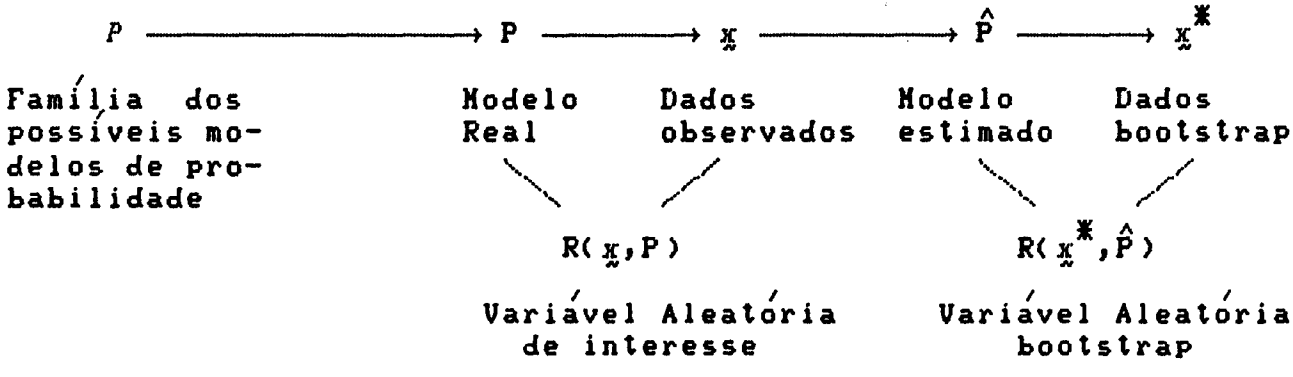

Figura 2.2.1.1 - Esquema do procedimento bootstrap.

Tem-se uma variável aleatória de interesse, $R(\underset{\sim}{X},P)$, que depende dos dados amostrais $\underset{\sim}{X}$ e do modelo desconhecido P. O objetivo é estimar algum parâmetro da distribuição de $R(\underset{\sim}{X},P)$. Através de qualquer técnica de estimação, P é estimado por $\hat{P}$ a partir dos dados $\underset{\sim}{X}$, paramétrica ou não parametricamente. Tendo estimado P por $\hat{P}$, o método Monte Carlo é aplicado para gerar os conjuntos de dados bootstrap, $\underset{\sim}{X}^*$, a partir de $\hat{P}$, da mesma forma que $\underset{\sim}{X}$ é gerado a partir de P. A variável aleatória $R(\underset{\sim}{X}^*,\hat{P})$ é observável e podemos encontrar sua distribuição, em geral, também por Monte Carlo. A estimativa bootstrap de $E_P(R(\underset{\sim}{X},P))$ é $E_{\hat{P}}(R(\underset{\sim}{X}^*,\hat{P}))$ e a idéia é a mesma para qualquer outro parâmetro da distribuição de $R(\underset{\sim}{X},P)$.

Supondo-se que o objetivo é estimar o vício do estimador de $\theta(P)$ (um parâmetro do modelo), $\hat{\theta}(\underset{\sim}{X})$, a variável aleatória de interesse nesse caso é:

$$R(\underset{\sim}{X},P) = \hat{\theta}(\underset{\sim}{X}) - \theta(P) \qquad (2.2.1.10)$$

e

$$E_P(R(\underset{\sim}{X},P)) = E_P(\hat{\theta}(\underset{\sim}{X})) - \theta(P) \qquad (2.2.1.11)$$

é o parâmetro de interesse da distribuição de $R(\underset{\sim}{X},P)$.

A estimativa bootstrap do vício é

$$\text{Vício}(\hat{\theta}) = E_{\hat{P}}(R(\underset{\sim}{X}^{*},\hat{P})) = E_{\hat{P}}\,[\hat{\theta}(\underset{\sim}{X}^{*}) - \theta(\hat{P})]$$
$$= E_{\hat{P}}\,[\hat{\theta}(\underset{\sim}{X}^{*})] - \theta(\hat{P}) \qquad (2.2.1.12)$$

Para calcular o vício($\hat{\theta}$), numericamente, segue-se o algorítimo Monte Carlo apresentado anteriormente e no passo (3) calcula-se

$$\text{Vício}_B(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} R(x^{*}(b),\hat{P}) =$$
$$= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}(x^{*}(b)) - \theta(\hat{P}) \qquad (2.2.1.13)$$

Quando $B \longrightarrow \infty$, a expressão em (2.2.1.13) aproxima-se da expressão (2.2.1.12).

2.2.2 - Jackknife

O método jackknife é uma técnica estatística introduzida por Quenouile (1949) com a finalidade de reduzir o vício de

estimadores. Tukey (1958) sugeriu que essa técnica poderia, também, ser usada para a obtenção de intervalos de confiânça aproximados em situaçoes onde os métodos padrões não são aplicáveis. Condições para a normalidade assintótica e então, para a validade dos intervalos de confiança aproximados foram dadas por Arvensen (1969) (Miller, 1974). O método foi generalizado por Schucany, Gray e Owen (1971) e uma revisão e algumas aplicações foram apresentadas por Miller (1974).

Considerando-se uma amostra aleatória de n observações, $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, com função de distribuição de probabilidade $F(\underset{\sim}{x}, \theta)$, onde θ é um parâmetro de interesse da população e $\hat{\theta}$ um estimador razoável para θ, a estimativa jackknife de θ é obtida da seguinte forma:

i - Retira-se a observação x_i da amostra;

ii - Calcula-se a estimativa de θ baseando-se na amostra $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, $\hat{\theta}_{-i}$;

iii - Calcula-se o chamado pseudo-valor:

$$\hat{\theta}_i = (n-1)\hat{\theta}_{-i} - n\hat{\theta} \qquad (2.2.2.1)$$

na qual $\hat{\theta}$ é a estimativa de θ com base nas n observações;

iv - Repete-se os passos i-ii-iii para cada x_i (i=1,2, ... ,n). A estimativa jackknife de θ é dada pela expressão (2.2.2.2).

$$J_n^1(\theta) = \frac{\sum_{i=1}^{n} \hat{\theta}_i}{n} \quad . \qquad (2.2.2.2)$$

Em geral, o vício de estimadores com base nas n observações amostrais pode ser escrito como a expansão em Série de Taylor:

$$\text{Vício}(\hat{\theta}) = \frac{a_1}{n} + \frac{a_2}{n^2} + O(1/n^3) \quad , \qquad (2.2.2.3)$$

onde os a_i (i=1,2) são constantes. Nesse caso temos, então que:

$$E(J_n^1(\theta)) = \theta + O(1/n^2) \quad , \qquad (2.2.2.4)$$

ou seja, a estimativa jackknife de primeira ordem, $J_n^1(\theta)$, tem o

vício reduzido na ordem de $1/n$. A variância estimada de $J_n^1(\theta)$ é dada pela expressão

$$\hat{\text{Var}}\,(J_n^1(\theta)) = \frac{\Sigma(\hat{\theta}_i - J_n^1(\theta))^2}{n(n-1)} \qquad (2.2.2.5)$$

Schucany, Gray e Owen (1971) generalizaram o método jackknife para a redução do vício em ordens mais altas. Retirando-se grupos de observações de tamanhos 2,3, ... ,k ($k \leq n/2$) das n observações originais (no passo i), o vício das estimativas é reduzido nas ordens $O(1/n^2)$, $O(1/n^3)$, ... ,$O(1/n^k)$, respectivamente.

Associando-se o método jackknife ao bootstrap, o primeiro pode ser visto como uma particularização do segundo. Efron (1979) mostrou a equivalência dos resultados obtidos pelo bootstrap, quando a aproximação da distribuição é através de expansões em séries de Taylor, e o jackknife infinitesimal de Jaeckel (1972). As expressões obtidas por esses dois procedimentos diferem levemente daquelas obtidas pelo jackknife usual. O jackknife infinitesimal usa derivadas em lugar das diferenças finitas do jackknife usual. O outro ponto de comparação é que no método jackknife usual, amostras aleatórias de tamanhos (n-1) são selecionadas sem reposição a partir de $\underset{\sim}{x}$ e

no bootstrap amostras de tamanho n são selecionadas com reposição a partir de $\underset{\sim}{x}$.

2.3 Revisão sobre aplicações dos métodos bootstrap e jackknife na estimação da diversidade.

Zahl (1977) introduziu a metodologia jackknife na estimação da diversidade em comunidades ecológicas. Ele aplicou o método a um exemplo particular de amostras não aleatórias de quadrats de uma comunidade vegetal. Seus dados eram constituídos de amostras repetidas dos mesmos quadrats em anos diferentes (1956 - 1960 - 1966 - 1969 - 1975) e o estudo considerou a estimação dos índices de Simpson e de Shannon-Weaver. Foram obtidas estimativas jackknife para os dois índices para cada ano, com os respectivos intervalos de confiança de 95%, estimativas jackknife para as diferenças entre os índices para os vários pares de anos e para os coeficientes angulares das retas ajustadas para cada índice de diversidade com relação ao ano, com respectivos desvios padrões estimados. Seu estudo limitou-se a comparações entre os resultados para os dois índices e a testes de normalidade para as estimativas.

Adams e McCune (1979), interessados no comportamento do vício apresentaram um estudo do jackknife generalizado aplicado à estimação do índice de Shannon-Weaver. Estimativas jackknife de

primeira, segunda e terceira ordens foram comparadas com relação ao vício médio, erro quadrático médio e coberturas de intervalos de confiança, através de simulações de Monte Carlo. As conclusões foram que a remoção de vício de alta ordem é acompanhada de um aumento da variância e portanto do erro quadrático médio, indicando que, talvez, o uso da estimativa jackknife de primeira ordem seja preferida, inclusive pela simplicidade das fórmulas.

Heltshe e Forrester (1983) desenvolveram fórmulas explícitas para a estimativa jackknife de primeira ordem para o número de espécies em uma comunidade e para a estimativa de sua variância. O estudo incorporou análises sobre o vício, porcentagem de cobertura e comprimento dos intervalos de confiança, quando o tamanho da amostra (número de quadrats) e o tamanho dos quadrats variam. Através de simulações de Monte Carlo, o método foi aplicado a comunidades com distribuição de abundância e número de espécies diferentes. Os resultados mostraram que para comunidades com poucas espécies o desempenho do método melhora com o aumento do tamanho da amostra e tamanho de quadrats fixo, independentemente da distribuição de abundância das espécies. Agora, quando a área amostrada é fixa e o tamanho dos quadrats é aumentado, qualquer ganho na porcentagem de cobertura dos intervalos de confiança é acompanhado de aumentos nos seus comprimentos (os desvios padrões aumentam). Porcentagens de cobertura dos intervalos de confiança aumentam e intervalos de confiança tornam-se mais estreitos quando o tamanho da amostra e

tamanho dos quadrats são aumentados simultaneamente. O desempenho do método é semelhante para o caso de comunidades com número de espécies um pouco maior, porém com desvios padrões bem maiores e os intervalos de confiança são subestimados. Tanto para comunidades com poucas espécies como para comunidades com mais espécies, o método mostra-se sensível às espécies raras, produzindo desvios padrões estimados maiores.

Smith e van Belle (1984) desenvolveram e compararam os métodos jackknife e bootstrap na estimação do número de espécies (S). Considerem-se uma amostra aleatória de n quadrats selecinada a partir de uma comunidade, e seja $\hat{S}$ o número de espécies observado nessa amostra. Aplicando a metodologia jackknife, quando o quadrat i é retirado da amostra, o número de espécies presentes na nova amostra de tamanho $(n-1)$ é:

$$\hat{S}_{-i} = \hat{S} - r_{1i} \qquad (2.3.1)$$

onde r_{1i} é o número de espécies encontradas apenas no quadrat i. Se 2 quadrats são retirados, i e j, o número de espécies na amostra é

$$\hat{S}_{-ij} = \hat{S} - r_{1i} - r_{1j} - r_{1ij} \quad , \qquad (2.3.2)$$

onde r_{1ij} é o número de espécies encontradas somente em ambos os quadrats i e j. Assim, pode-se chegar à fórmula geral da estimativa jackknife de k-ésima ordem que é dada pela expressão

$$J_n^k(S) = \hat{S} + \left\{ \sum_{j=1}^{k} r_{1(j)} \sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} (n-i)^k \binom{n-i}{i-1} / \binom{n}{i} \right\} / K!, \qquad (2.3.3)$$

onde $r_{1(1)} = \sum_{i=1}^{n} r_{1i}$ é o número de espécies encontradas em exatamente um quadrat, $r_{1(2)} = \sum_{i<j} r_{1ij}$ é o número de espécies encontradas em exatamente dois quadrats e assim por diante.

Para $k=1$,

$$J_n^1(S) = \hat{S} + (r_{1(1)} (n-1)) / n \qquad (2.3.4)$$

é a estimativa jackknife de S, de primeira ordem, e sua variância estimada é dada por:

$$\text{Var}_{est} \left\{ J_n^1(S) \right\} = \sum_{i=1}^{n} \frac{(r_{1i} - r_{1(1)} \cdot 1/n)}{n(n-1)} \qquad (2.3.5)$$

Desenvolvendo a metodologia bootstrap para essa mesma situação tem-se: Seja I_j uma variável indicadora que assume o valor 1 quando a espécie j da amostra das n observações originais está presente na amostra bootstrap e zero caso contrário. Sob o procedimento de reamostragem com reposição, I_j é uma variável aleatória com distribuição de Bernoulli com parâmetro $p = [1 - (1 - Y_j/n)^n]$, onde Y_j é o número de quadrats da amostra que apresentam a espécie j e p é a probabilidade da espécie j estar presente em pelo menos um quadrat da amostra bootstrap. Temos, então, que o número de espécies na amostra bootstrap é uma variável aleatória definida por:

$$S^* = \sum_{j=1}^{\hat{S}} I_j \quad . \qquad (2.3.6)$$

$$E_B(S^*) = E_B(\Sigma I_j) = \Sigma P_B(I_j = 1)$$
$$= \hat{S} - \sum_{j=1}^{\hat{S}} (1 - Y_j/n)^n \qquad (2.3.7)$$

é o valor esperado do número de espécies na amostra bootstrap. Então, o vício estimado de S^* é

$$\hat{\text{Vício}}_B(S^*) = \sum_{j=1}^{\hat{S}} (1 - Y_j/n)^n \qquad (2.3.8)$$

que é usado como uma aproximação para o vício de $\hat{S}$. Assim, é obtida uma estimativa de S corrigida para o vício, dada por

$$B_n(S) = \hat{S} + \sum_{j=1}^{\hat{S}} (1 - Y_j/n)^n \qquad (2.3.9)$$

Tem-se que

$$B_n(S) \in \left[\hat{S} \ , \ \hat{S} \left(1 + \left(\frac{n-1}{n} \right)^n \right) \right] \qquad (2.3.10)$$

ou seja, a estimativa bootstrap de S ($B_n(S)$) alcança seu valor máximo quando todas as espécies da amostra são espacialmente raras ($Y_j=1$ para $j = 1,2, \ldots ,\hat{S}$) e é mínima se e somente se todas as espécies estão presentes em todos os quadrats ($Y_j=n$ para $j = 1,2, \ldots ,\hat{S}$).

A variância de S^* é dada pela expressão:

$$\hat{\mathrm{Var}}_B(S^*) = \sum_{j=1}^{\hat{S}} \left\{ 1 - (1 - Y_j/n)^n - \left(1 - (1 - Y_j/n)^n\right)^2 \right\}$$

$$+ \sum_{\substack{j \\ j \neq k}} \sum_{k} \left\{ 1 - [(1 - Y_j/n)^n - (Y_k/n)^n - (Z_{jk}/n)^n] \right.$$

$$\left. - [1 - (1 - Y_j/n)^n] \times [1 - (1 - Y_k/n)^n] \right\} \ ,$$

onde Z_{jk} é o número de quadrats onde ambas as espécies j e k estão ausentes.

O objetivo do estudo de Smith e van Belle (1984) foi comparar o comportamento do vício das estimativas de S em relação às densidades das espécies nos quadrats, sob os procedimentos jackknife e bootstrap. Para a comparação dos resultados, as esperanças dos estimadores foram calculadas sob a suposição de que a abundância da espécie j segue a distribuçao de Poisson com parâmetro λ_j sobre a área (j = 1,2, ... ,S). Os vícios das estimativas corrigidas pelo jackknife e pelo bootstrap foram comparados com as estimativas obtidas através do modelo suposto, para tamanhos de amostra de 5 e 10 quadrats.

Os resultados mostraram que o jackknife reduz mais o vício do que o bootstrap, mas para densidades pequenas nenhum dos métodos o reduz significativamente. Com o aumento da densidade, o vício das estimativas obtidas através dos três procedimentos (jackknife, bootstrap e modelo) tende a zero e esta convergência é mais rápida para tamanho de amostra maior. Para amostras grandes o jackknife super-estima o número de espécies e o bootstrap tem melhor desempenho. O jackknife apresenta melhor comportamento quando as amostras são pequenas, compensando mais o vício.

Os autores sugeriram o uso das curvas do comportamento

do vício em relação às densidades das espécies, como uma alternativa para a curva comumente usada para detectar o tamanho da amostra suficiente para estimar S, $\hat{S}$ em relação ao tamanho da amostra. Mesmo quando esta última torna-se assintótica, espécies muito raras podem ainda não terem sido observadas. A idéia de usar um dos dois métodos propostos, é que, segundo esses autores, através de um desses métodos, é possível selecionar-se a densidade da espécie por quadrat (λ^*) e o número de quadrats correspondente que fazem o vício esperado das estimativas sempre não negativo, em lugar da área mínima amostrada.

As grandes vantagens desses dois métodos não paramétricos estão no fato de que informações sobre presença ou ausência das espécies nos quadrats são suficientes para a obtenção das estimativas e não é necessário nenhuma suposição a respeito de como as espécies coexistem dentro dos quadrats. Quando modelos paramétricos são usados, a suposição de independência entre as espécies é necessária, assim como a contagem do número de indivíduos de cada espécie o que, na prática, pode ser complicado e às vezes impossível.

CAPÍTULO 3

BOOTSTRAP MONTE CARLO NA ESTIMAÇÃO DO NÚMERO DE ESPÉCIES

3.1 - Introdução

Iniciam-se, neste capítulo, um estudo para avaliar a aplicação do método bootstrap na estimação do número de espécies em uma comunidade, incentivado pelo trabalho de Smith e van Belle (1984). O enfoque principal do estudo é a avaliação do método quando amostras de quadrats ou parcelas são selecionadas, sem se fazerem suposições sobre os modelos teóricos da distribuição da abundância das espécies, ou sobre a distribuição espacial dos indivíduos, ou sobre qualquer outra relação ecológica existente entre as espécies e ou indivíduos. Para isso, foram feitas simulações e a metodologia bootstrap Monte Carlo (não paramétrico) foi aplicada, avaliando-se o seu desempenho.

Os objetivos principais, deste trabalho, são a avaliação do bootstrap sob diferentes equitabilidades e diferentes riquezas de espécies, quando o tamanho da amostra e o tamanho das parcelas são aumentados e a obtenção de faixas de confiança para a curva do número de espécies em função do tamanho da amostra, utilizando essas mesmas faixas como uma regra para a escolha do tamanho da amostra suficiente para estimar S.

Na seção 3.2 apresentam-se a metodologia desenvolvida

para o estudo e na seção 3.3 tem-se os resultados e conclusões do estudo para as populações simuladas. Para a avaliação dos resultados, foram considerados o vício e o desvio padrão das estimativas.

3.2 - Metodologia

3.2.1. Bootstrap

Considerem-se, de início, o procedimento para a construção da curva do número de espécies em função do tamanho da amostra, quando esta vai se acumulando aos poucos, a partir de N parcelas selecionadas ao acaso de uma comunidade com S espécies. As m primeiras parcelas são observadas e o número de espécies presentes é $\hat{S}_{(m)}$; $(\hat{S}_{(m)}, m)$ é o primeiro ponto da curva. Em seguida, adicionam-se m parcelas às anteriores, $\hat{S}_{(2m)}$ é o número de espécies encontradas nessa amostra e obtem-se o ponto $(\hat{S}_{(2m)}, 2m)$. Prosseguem-se dessa forma até obterem-se o ponto $(\hat{S}_{(N)}, N)$. Se n é o tamanho de amostra acumulado, através da qual $\hat{S}_{(n)}$ foi observado, então n assume os valores m, $2m$, ... , N e tem-se N/m pontos para a construção da curva. Se N é suficientemente grande esperam-se obter um gráfico como o apresentado na figura 3.2.1.1.

Aplicando a metodologia bootstrap, onde as unidades reamostradas são os N quadrats, e usando o mesmo procedimento

para a reconstrução da curva, podem-se obter uma faixa de confiança para a curva e estimativas bootstrap da variância do número de espécies para cada tamanho de amostra acumulado. Repetindo-se esse processo R vezes obtêm-se as variâncias das estimativas bootstrap.

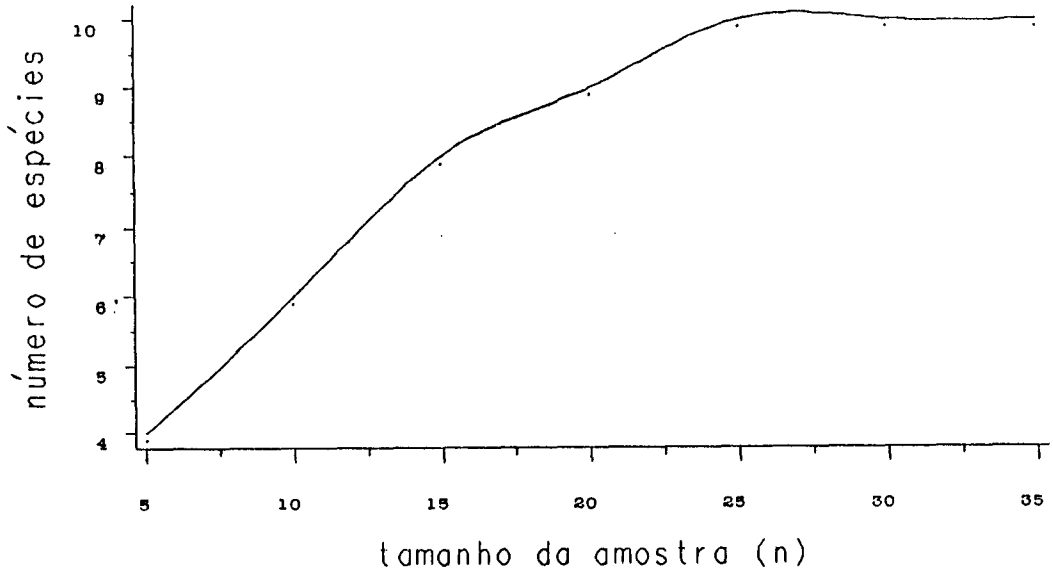

Figura 3.2.1.1 Curva do número de espécies observado em função do tamanho da amostra.

Considerem-se $S^{*}_{ij(n)}$ como sendo o número de espécies presentes na j-ésima amostra bootstrap de n unidades amostrais, da i-ésima repetição ($j = 1,2, \ldots ,B$; $i = 1,2, \ldots ,R$; $n = m,2m, \ldots , N$). Tem-se que

$$0 < S^{*}_{ij(n)} \leq \hat{S}_{(N)} \quad , \qquad (3.2.1.1)$$

e portanto os valores dos $S^{*}_{ij(n)}$ ($j = 1,2, \ldots ,B$; $i = 1,2, \ldots ,R$; $n = m,2m, \ldots ,N$) produzem uma faixa de confiança para a curva que relaciona $\hat{S}_{(n)}$ a n, como mostra a figura 3.2.1.2.

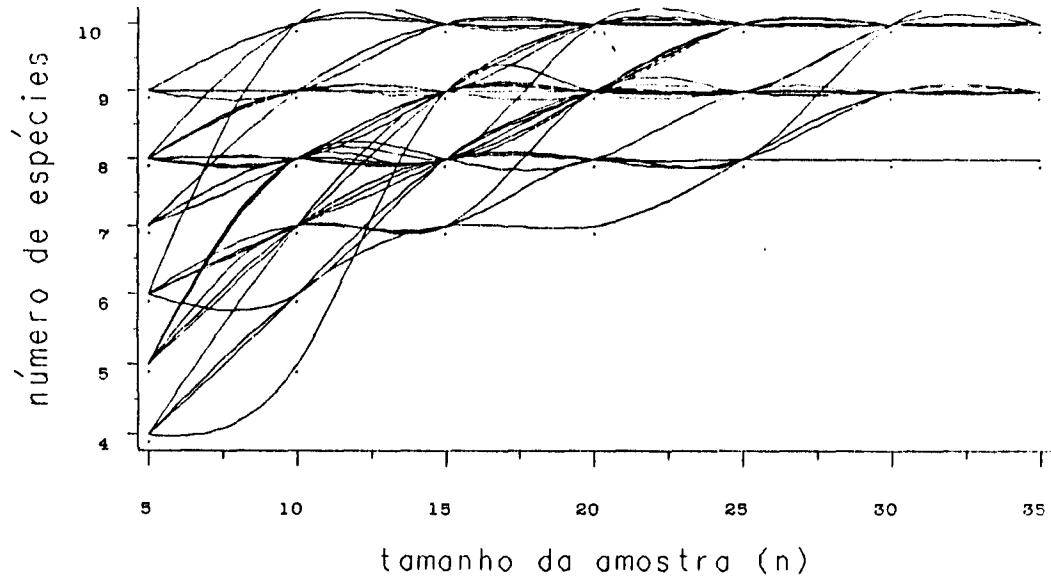

Figura 3.2.1.2 Faixa de confiança para a curva do número de espécies em função do tamanho da amostra, obtida pelo bootstrap, sem usar a correção para o vício das estimativas.

O máximo que $S^{*}_{ij}(n)$ atinge é $\hat{S}(N)$ e a faixa de confiança estará sempre abaixo desse valor.

Smith e van Belle (1984) propuseram uma estimativa

bootstrap de S, corrigida para o vício, dada por

$$B_n(S) = \hat{S}_{(n)} + \sum_{k=1}^{\hat{S}_{(n)}} (1 - Y_k/n)^n \quad , \qquad (3.2.1.2)$$

onde Y_k é o número de parcelas da amostra de tamanho n cuja espécie k está presente ($1 \leq Y_k \leq n$; $0 < \hat{S}_{(n)} \leq \hat{S}_{(N)}$) e $\sum_{k=1}^{\hat{S}_{(n)}} (1 - Y_k/n)^n$ é estimativa bootstrap do vício de $\hat{S}_{(n)}$.

Usando o mesmo raciocínio, a estimativa bootstrap do vício dos $S^*_{ij(n)}$ ($i=1,2,\ldots,R$; $j=1,2,\ldots,B$; $n=m,2m,\ldots,N$) é dado por

$$\text{Vício}_B(S^*_{ij(n)}) = \sum_{k=1}^{S^*_{ij(n)}} (1 - Y_{ijk}/n)^n \quad , \qquad (3.2.1.3)$$

onde Y_{ijk} é o número de quadrats da j-ésima amostra bootstrap de tamanho n, da i-ésima repetição, cuja espécie k está presente ($i=1,2,\ldots,R$; $j=1,2,\ldots,B$; $k=1,2,\ldots,S^*_{ij(n)}$ e $1 \leq Y_{ijk} \leq n$). Portanto,

$$B_n^{*}(S)_{ij} = S^{*}_{ij(n)} + \sum_{k=1}^{S^{*}_{ij(n)}} (1 - Y_{ijk}/n)^n \qquad (3.2.1.4)$$

são as estimativas bootstrap de $B_n(S)$. Colocando-se os valores de $B_n^{*}(S)_{ij}$ (i=1,2, ... ,R ; j=1,2, ... ,B ; n=m,2m, ... ,N) em um gráfico, em relação a n, obtem-se a figura 3.2.1.3 que pode ser vista como sendo uma faixa de confiança para $B_n(S)$ x n. Esta faixa sugere uma regra alternativa para a decisão do tamanho da amostra suficiente para estimar S, que pode ser determinado quando a faixa se torna relativamente estreita (desvio padrão pequeno), conforme a precisão desejada pelo pesquisador.

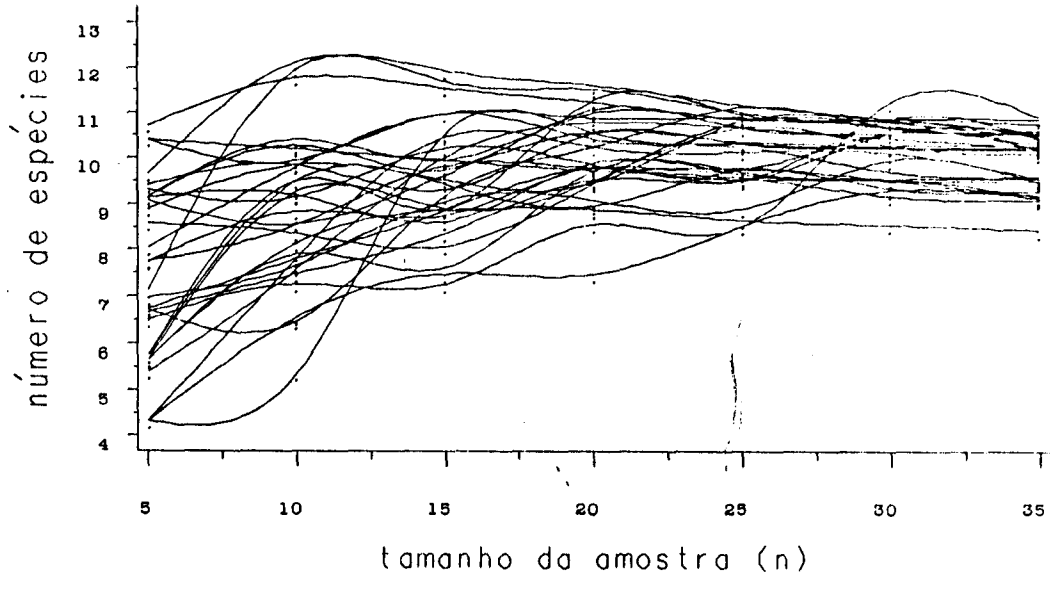

Figura 3.2.1.3 Faixa de confiança para a curva do número de espécies em função do tamanho da amostra, obtida pelo bootstrap, usando as estimativas corrigidas para o vício.

A estimativa bootstrap do valor esperado de $B_n(S)$ e de sua variância são, respectivamente

$$\hat{E}_B\,(B_n(S)) = \sum_{i=1}^{R} \sum_{j=1}^{B} \left[\frac{B_n^*(S)_{ij}/B}{R} \right] \qquad (3.2.1.5)$$

$$\hat{E}_B\,(\hat{\sigma}^2_{B_n(S)}) = \hat{E}_B\,[\,\hat{var}\,B_n(S)] = \frac{\sum \hat{var}_B(\bar{B}_n^*(S)_i)}{R}$$

$$= \sum_{i=1}^{R} \left\{ \sum_{j=1}^{B} [B_n^*(S)_{ij} - \bar{B}_n^*(S)_i]^2 / B - 1 \right\} / R \quad , \qquad (3.2.1.6)$$

onde

$$\bar{B}_n^*(S)_i = \sum_{j=1}^{B} B_n^*(S)_{ij} / B \quad . \qquad (3.2.1.7)$$

A estimativa da variância de $\hat{\sigma}^2_{B_n(S)}$ é

$$\hat{var}_B(\hat{\sigma}^2_{B_n S)}) = \frac{\sum \left\{ var_B(B_n^*(S)_{ij}) - \hat{E}_B\,(\hat{\sigma}^2_{B_n(S)}) \right\}^2}{R-1} \qquad (3.2.1.8)$$

3.2.2 - Simulações

Para analizar a sensibilidade do método bootstrap às

mudanças na relação espécie-abundância e às diferentes riquezas de espécies, quatro comunidades foram criadas variando-se esses dois fatores. Seja S o número de epécies na comunidade e cada comunidade classificada em alta equitabilidade e baixa equitabilidade, de acordo com a relação espécie-abundância das espécies. As quatro comunidades geradas seguiram o esquema da tabela 3.2.2.1

Comunidade	S	Equitabilidade
I	10	Alta
II	10	Baixa
III	30	Alta
IV	30	Baixa

Tabela 3.2.2.1 Comunidades geradas para o estudo.

As abundâncias relativas ou as proporçõoes das espécies, para cada comunidade, foram geradas seguindo os valores das probabilidades da distribuição geométrica. Por exemplo, ordenando-se as S espécies de acordo com a abundância, a proporção da j-ésima espécie é dada por

$$\pi'_j = p(1 - p)^{j-1} \quad \text{para} \quad j = 1,2, \ldots ,S \qquad (3.2.2.1)$$

onde p é o parâmetro da distribuição geométrica ($p = \pi'_1$).

Como $S(\infty, \sum_{j=1}^{S} \Pi'_j < 1$. Para que as S espécies constituam os 100% da comunidade, a proporção da j-ésima espécie é

$$\Pi_j = \Pi'_j + (1 - \Sigma \Pi'_j)/S \quad \text{para} \quad j = 1,2, \ldots ,S \qquad (3.2.2.2)$$

As curvas das distribuições das proporções das espécies para as quatro comunidades são representadas nas figuras 3.2.2.1 e 3.2.2.2 para S=10 e S=30, respectivamente.

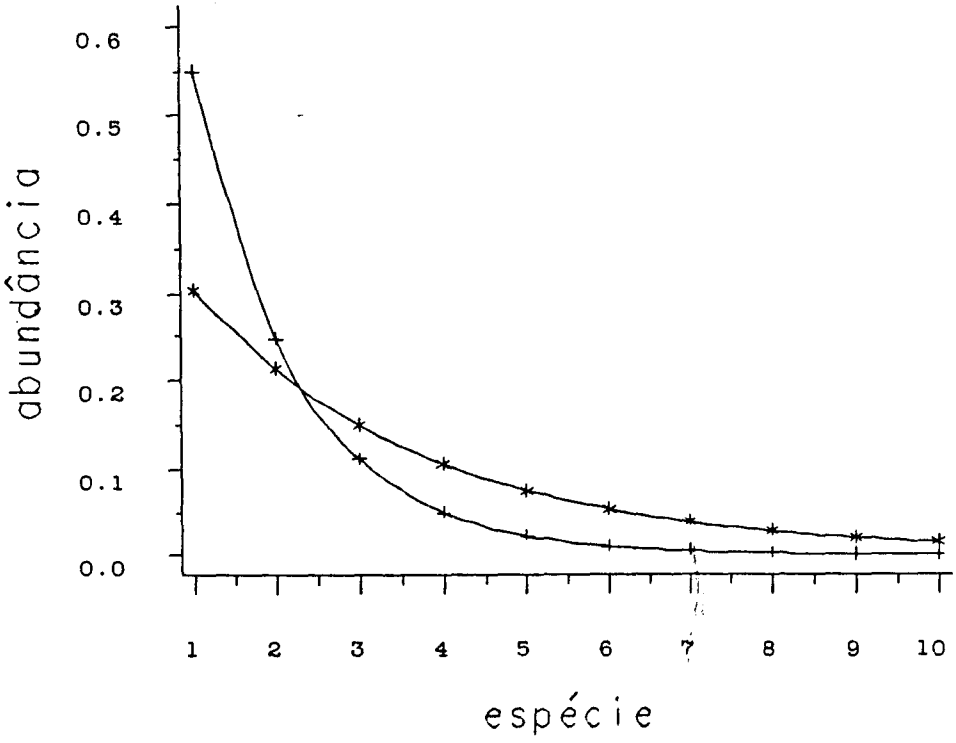

Figura 3.2.2.1 Distribuição da abundância das espécies para as comunidades simuladas. (*) comunidade I, (+) comunidade II.

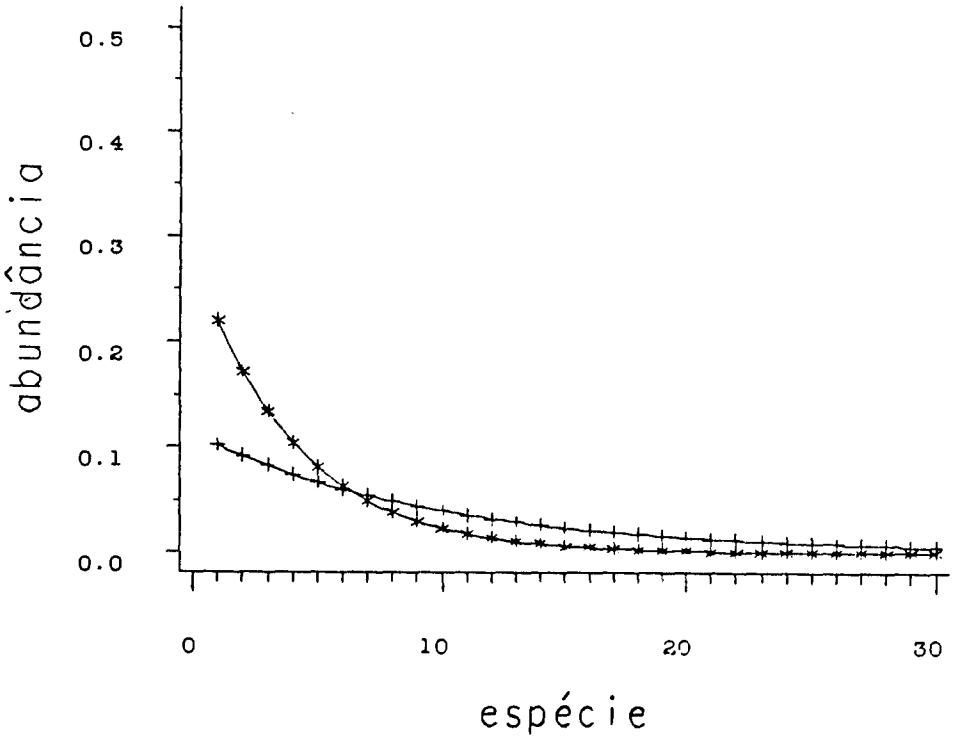

Figura 3.2.2.2. Distribuição da abundância das espécies para as comunidades simuladas. (+) comunidade III, (✱) comunidade IV.

O estudo considerou as comunidades como sendo infinitas com relação ao número total de indivíduos e nenhuma imposição foi feita quanto a forma da distribuição espacial desses indivíduos, assim como da existência ou não de interações ecológicas inter ou intra específicas.

Para cada comunidade, amostras de *N* quadrats foram geradas sequencialmente. Para as duas comunidades com muitas espécies raras 200 quadrats foram gerados e para as outras duas comunidades 100 quadrats foram suficientes. Cada quadrat foi

representado pelo número de indivíduos presentes e foi gerado através da função geradora de números pseudo-aleatórios, UNIFORM, do pacote estatístico SAS (STATISTICAL ANALISYS SYSTEM). Assim,o número de indivíduos por parcela é distribuído uniformemente dentro de um intervalo escolhido de acordo com o número médio de indivíduos desejado dentro de cada parcela. Para avaliar o efeito de tamanho de quadrat, o número médio de indivíduos por quadrat foi aumentado, o que seria equivalente a aumentar a área da parcela. Quatro tamanhos de quadrats foram estudados sendo que a variância do número de indivíduos por quadrat foi mantida constante. Tendo gerado um quadrat, cada indivíduo foi classificado entre as S espécies de acordo com a distribuição de abundância das espécies. Informações sobre a presença de cada espécie foram acumuladas de m em m parcelas (m=10) e $\hat{S}(n)$ foi observado para n=10,20, ... ,100 para as comunidades I e III e n=10,20, ... ,200 para as comunidades II e IV.

A tabela 3.2.2.2 mostra o esquema experimental montado para o estudo. As caselas na tabela, indicam o número esperado de indivíduos nas n parcelas acumuladas (n=10,20, ... ,N) onde Q é o número esperado de indivíduos dentro de cada quadrat.

Q \ n	10	20	30	.....	N
10	100	200	300	.....	10.N
20	200	400	600	.....	20.N
30	300	600	900	.....	30.N
40	400	800	1200	.....	40.N

Tabela 3.2.2.2 Esquema experimental montado para o

A metodologia bootstrap foi aplicada a cada linha da tabela usando-se R=50 e B=100. Para a avaliação dos resultados, cem repetições Monte Carlo foram realizadas sobre o experimento, usando-se o mesmo esquema de construção da curva do número de espécies em relação a n, descrito anteriormente, e os valores esperados para $B_n(S)$ foram obtidos para cada linha da tabela 3.2.2.2, com os repectivos desvios padrões.

Todos os programas para a geração das comunidades e para os estudos Monte Carlo e bootstrap foram feitos usando-se o pacote estatístico SAS. Os programas referentes às simulações de Monte Carlo foram executados no computador VAX do Centro de Computação da UNICAMP e os referentes às repetições bootstrap foram executados no computador IBM do CPD do IPEN - São Paulo.

3.3 Resultados e Conclusões

Os resultados obtidos pelas simulações Monte Carlo e bootstrap são apresentados nas tabelas 3.3.1, 3.3.2, 3.3.3 e 3.3.4 para as comunidades I, II, III e IV, respectivamente. Os valores nas tabelas são as médias do número de espécies corrigidos para o vício e respectivos desvios padrões médios. Os valores entre parênteses são as estimativas dos desvios padrões para os desvios padrões médios.

Q	n / método	10	20	30	40	50	60	70	80	90	100
10	Monte Carlo	10.06 0.78	10.11 0.30	10.04 0.17	10.03 0.05	10.01 0.02	10 0.01	10 0	10 0	10 0	10 0
	Boot	10.05 0.77 (0.06)	10.12 0.35 (0.04)	10.06 0.17 (0.04)	10.03 0.08 (0.03)	10.01 0.04 (0.02)	10 0.02 (0.01)	10 0.01 (0.01)	10 0 (0.01)	10 0 (0)	10 0 (0)
20	Monte Carlo	10.08 0.3	10.02 0.03	10 0.01	10 0	10 0	10 0	10 0	10 0	10 0	10 0
	Boot	10.11 0.27 (0.05)	10.02 0.05 (0.02)	10.00 0.01 (0.01)	10.00 0 (0)						
30	Monte Carlo	10.03 0.19	10 0.01	10 0	10 0						
	Boot	10.07 0.18 (0.04)	10.01 0.02 (0.02)	10 0 (0)							
40	Monte Carlo	10.03 0.08	10 0	10 0							
	Boot	10.03 0.08 (0.03)	10 0 (0)								

Tabela 3.3.1. Resultados obtidos pelo estudo de Monte Carlo e pelo bootstrap para a comunidade I.

Q	n / método	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
10	Monte Carlo	6.81 1.11	7.67 1.07	7.97 1.04	8.29 1.04	8.51 1.06	8.74 1.03	8.97 1.08	9.13 1.06	9.24 1.04	9.35 1.04	9.45 1.00	9.62 0.92	9.71 0.86	9.73 0.85	9.73 0.84	9.76 0.84	9.75 0.83	9.73 0.82	9.76 0.81	9.78 0.79
	Boot	6.57 1.09 (0.08)	7.21 1.05 (0.08)	7.60 1.00 (0.07)	7.84 0.96 (0.06)	8.02 0.94 (0.06)	8.16 0.91 (0.06)	8.27 0.90 (0.06)	8.37 0.88 (0.05)	8.45 0.86 (0.04)	8.51 0.84 (0.04)	8.57 0.82 (0.04)	8.62 0.80 (0.04)	8.66 0.77 (0.04)	8.70 0.76 (0.05)	8.74 0.74 (0.04)	8.77 0.73 (0.04)	8.79 0.71 (0.04)	8.81 0.70 (0.04)	8.82 0.69 (0.03)	8.85 0.67 (0.03)
20	Monte Carlo	7.59 1.14	8.39 1.19	8.89 1.12	9.23 1.02	9.41 0.97	9.50 0.87	9.65 0.84	9.71 0.82	9.75 0.80	9.85 0.79	9.90 0.74	9.93 0.72	9.95 0.65	9.97 0.64	9.96 0.62	9.94 0.61	9.95 0.60	10.03 0.53	10.06 0.50	10.07 0.48
	Boot	7.79 1.18 (0.09)	8.56 1.06 (0.08)	8.87 1.00 (0.07)	9.09 0.97 (0.05)	9.26 0.86 (0.05)	9.40 0.92 (0.05)	9.50 0.89 (0.05)	9.60 0.86 (0.05)	9.68 0.84 (0.05)	9.75 0.80 (0.06)	9.81 0.77 (0.05)	9.86 0.73 (0.05)	9.90 0.70 (0.05)	9.95 0.66 (0.04)	9.98 0.63 (0.04)	10.00 0.61 (0.04)	10.02 0.58 (0.05)	10.04 0.55 (0.05)	10.05 0.52 (0.04)	10.07 0.50 (0.05)
30	Monte Carlo	8.09 1.11	8.68 1.03	8.99 1.02	9.22 1.01	9.38 0.99	9.52 0.89	9.70 0.81	9.86 0.77	9.96 0.70	9.99 0.63	9.96 0.60	9.99 0.58	10.04 0.54	10.06 0.47	10.05 0.46	10.09 0.39	10.06 0.38	10.06 0.38	10.05 0.35	10.04 0.35
	Boot	7.65 1.04 (0.06)	8.27 1.01 (0.06)	8.62 0.99 (0.06)	8.88 0.96 (0.06)	9.08 0.93 (0.07)	9.93 0.90 (0.06)	9.36 0.87 (0.06)	9.45 0.84 (0.06)	9.51 0.81 (0.05)	9.58 0.79 (0.05)	9.62 0.77 (0.05)	9.66 0.75 (0.04)	9.70 0.74 (0.04)	9.72 0.72 (0.03)	9.75 0.70 (0.03)	9.77 0.69 (0.03)	9.79 0.68 (0.03)	9.81 0.67 (0.03)	9.82 0.66 (0.03)	9.83 0.64 (0.03)
40	Monte Carlo	8.40 0.99	9.15 0.91	9.50 0.85	9.72 0.80	9.77 0.77	9.84 0.74	9.98 0.63	9.99 0.59	10.01 0.55	10.03 0.51	10.12 0.41	10.16 0.32	10.13 0.31	10.13 0.24	10.11 0.24	10.10 0.23	10.08 0.23	10.08 0.20	10.07 0.19	10.05 0.18
	Boot	8.28 1.08 (0.07)	8.91 0.96 (0.07)	9.23 0.90 (0.07)	9.41 0.84 (0.06)	9.54 0.80 (0.06)	9.63 0.77 (0.04)	9.70 0.74 (0.03)	9.76 0.71 (0.03)	9.81 0.68 (0.03)	9.86 0.65 (0.03)	9.89 0.62 (0.03)	9.92 0.60 (0.03)	9.95 0.57 (0.02)	9.96 0.55 (0.03)	9.98 0.52 (0.03)	9.99 0.50 (0.03)	10.01 0.47 (0.03)	10.02 0.45 (0.03)	10.03 0.43 (0.04)	10.04 0.41 (0.04)

Tabela 3.3.2 Resultados obtidos pelo estudo de Monte Carlo e pelo bootstrap para a comunidade II.

Q	n / método	10	20	30	40	50	60	70	80	90	100
10	Monte Carlo	27.87 2.25	30.13 1.41	30.56 1.02	30.05 1.70	30.33 0.46	30.30 0.31	30.19 0.23	30.11 0.15	30.06 0.13	30.05 0.06
	Boot	27.67 2.25 (0.18)	29.85 1.54 (0.12)	30.29 1.13 (0.07)	30.34 0.83 (0.07)	30.30 0.65 (0.05)	30.25 0.51 (0.04)	30.19 0.41 (0.04)	30.15 0.34 (0.04)	30.11 0.28 (0.04)	30.09 0.23 (0.04)
20	Monte Carlo	30.36 1.49	30.58 0.64	30.25 0.33	30.13 0.14	30.05 0.09	30.02 0.05	30.01 0.01	30 0.01	30 0	30 0
	Boot	30.23 1.45 (0.12)	30.42 0.65 (0.05)	30.22 0.32 (0.05)	30.10 0.17 (0.04)	30.04 0.09 (0.03)	30.02 0.05 (0.02)	30.01 0.03 (0.02)	30 0.01 (0.01)	30 0.01 (0.01)	30 0 (0)
30	Monte Carlo	30.69 1.89	30.20 0.4	30.07 0.15	30.01 0.11	30 0.15	30 0	30 0	30 0	30 0	30 0
	Boot	30.43 1.12 (0.09)	30.29 0.39 (0.05)	30.10 0.17 (0.04)	30.03 0.07 (0.03)	30.01 0.03 (0.02)	30 0.01 (0.01)	30 0 (0.01)	30 0 (0)	30 0 (0)	30 0 (0)
40	Monte Carlo	30.50 0.69	30.12 0.17	30.02 0.04	30 0.01	30 0	30 0	30 0	30 0	30 0	30 0
	Boot	30.47 0.74 (0.07)	30.13 0.22 (0.05)	30.03 0.07 (0.03)	30.01 0.03 (0.02)	30 0.01 (0.01)	30 0 (0)	30 0 (0)	30 0 (0)	30 0 (0)	30 0 (0)

Tabela 3.3.3. Resultados obtidos pelo estudo de Monte Carlo e pelo bootstrap para a comunidade III

Q	n / método	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
10	Monte Carlo	17.24 2.19	19.84 2.29	21.68 2.27	22.78 2.21	23.53 2.02	24.13 1.89	24.72 1.89	24.72 1.84	25.55 1.82	25.95 1.80	26.28 1.77	26.38 1.76	26.45 1.75	26.64 1.71	26.88 1.67	27.15 1.73	27.40 1.71	27.45 1.68	27.65 1.74	27.77 1.76
	Boot	17.14 2.34 (0.16)	20.15 2.25 (0.15)	21.74 2.11 (0.18)	22.72 2.04 (0.16)	23.43 2.01 (0.17)	23.93 1.97 (0.14)	24.35 1.92 (0.14)	24.67 1.89 (0.12)	24.95 1.86 (0.15)	25.18 1.84 (0.14)	25.38 1.83 (0.12)	25.57 1.79 (0.10)	25.74 1.75 (0.11)	25.85 1.72 (0.10)	25.96 1.69 (0.09)	26.07 1.67 (0.08)	26.18 1.63 (0.09)	26.25 1.61 (0.10)	26.32 1.58 (0.09)	26.39 1.55 (0.09)
20	Monte Carlo	19.49 2.23	22.26 2.15	23.99 2.20	25.21 2.10	25.87 2.21	26.48 2.03	27.00 1.99	27.29 1.93	27.63 1.84	27.83 1.73	28.02 1.61	28.20 1.62	28.32 1.58	28.51 1.54	28.64 1.52	28.70 1.39	28.85 1.34	28.96 1.39	29.06 1.36	29.16 1.40
	Boot	19.87 2.01 (0.16)	22.62 2.01 (0.13)	24.16 1.94 (0.15)	25.16 1.86 (0.14)	25.87 1.81 (0.13)	26.40 1.76 (0.14)	26.85 1.73 (0.12)	27.20 1.70 (0.11)	27.47 1.67 (0.10)	27.72 1.61 (0.12)	27.93 1.56 (0.12)	28.11 1.52 (0.10)	28.26 1.46 (0.09)	28.38 1.42 (0.09)	28.51 1.37 (0.10)	28.61 1.31 (0.09)	28.71 1.26 (0.09)	28.79 1.21 (0.09)	28.86 1.17 (0.09)	28.92 1.14 (0.09)
30	Monte Carlo	21.58 2.32	24.09 2.28	25.66 1.95	26.77 1.93	27.54 1.90	27.92 1.83	28.32 1.70	28.60 1.65	28.98 1.43	29.13 1.43	29.24 1.38	29.50 1.37	29.64 1.31	29.72 1.18	29.86 1.13	29.91 1.14	29.91 1.15	29.99 1.14	30.04 1.11	30.12 1.06
	Boot	21.16 2.17 (0.16)	23.82 2.17 (0.14)	25.27 2.11 (0.16)	26.24 2.00 (0.12)	26.91 1.93 (0.12)	27.37 1.84 (0.12)	27.78 1.79 (0.12)	28.10 1.73 (0.12)	28.32 1.68 (0.12)	28.52 1.64 (0.10)	28.71 1.59 (0.09)	28.85 1.56 (0.08)	28.98 1.52 (0.10)	29.10 1.50 (0.09)	29.19 1.47 (0.09)	29.27 1.43 (0.09)	29.35 1.38 (0.09)	29.43 1.33 (0.09)	29.50 1.30 (0.09)	29.56 1.26 (0.09)
40	Monte Carlo	22.36 2.36	24.37 2.00	26.24 1.86	27.21 1.78	27.92 1.66	28.42 1.60	28.60 1.58	28.72 1.47	29.02 1.50	29.27 1.45	29.48 1.35	29.60 1.28	29.73 1.21	29.86 1.10	29.98 1.04	30.00 1.07	30.00 0.98	30.03 0.97	30.09 0.96	30.14 0.90
	Boot	22.56 2.28 (0.18)	25.26 2.20 (0.18)	26.79 2.12 (0.15)	27.74 1.98 (0.15)	28.39 1.85 (0.14)	28.88 1.75 (0.12)	29.24 1.64 (0.12)	29.52 1.53 (0.12)	29.73 1.42 (0.12)	29.89 1.31 (0.11)	30.01 1.23 (0.11)	30.11 1.14 (0.12)	30.17 1.06 (0.12)	30.22 0.99 (0.10)	30.26 0.92 (0.09)	30.28 0.85 (0.08)	30.29 0.79 (0.08)	30.29 0.74 (0.08)	30.30 0.68 (0.06)	30.29 0.63 (0.06)

Tabela 3.3.4. Resultados obtidos pelo estudo de Monte Carlo e pelo bootstrap para a comunidade IV.

O estudo de Monte Carlo mostra, nos quatro casos, que as diferenças entre os valores estimados de S e o verdadeiro valor de S diminuem quando o tamanho da amostra é aumentado, resultado que é esperado. Essa tendência se acentua quando os quadrats amostrados são maiores. Para o caso de comunidades com poucas espécies raras (I e III), a convergência da estimativa Monte Carlo do número de espécies a S é muito rápida, e os valores dos desvios padrões médios das estimativas são bem pequenos, tendendo a zero com o aumento do tamanho da amostra. Essa convergência foi mais rápida para quadrats maiores. Esse resultado também é esperado para as outras duas comunidades (II e IV) se a amostragem ultrapassasse além dos 200 quadrats. De modo geral, os resultados bootstrap acompanham essas tendências, seguindo o comportamento do estudo de Monte Carlo. Para as duas comunidades com poucas espécies raras, onde, portanto, não há problemas na estimação de S, a concordância entre os resultados bootstrap e Monte Carlo é quase perfeita, a menos de variações inerentes ao processo, resultando em faixas de confiança bastante estreitas para a curva do número de espécies corrigido em função do tamanho da amostra, para todos os tamanhos de quadrats. Podem-se chegar a esta conclusão apenas observando a magnitude do desvio padrão médio das estimativas, nas tabelas 3.3.1 e 3.3.3..

Para as comunidades cuja maioria das espécies são raras o bootstrap sub-estima o número de espécies esperado. Para parcelas pequenas (Q=10) esse vício cresce com n; para quadrats

maiores o vício é menor e o desempenho do método melhora com o aumento de n, o que pode ser visto mais claramente nas figuras 3.3.1 e 3.3.2. Quanto aos desvios padrões, o bootstrap tende a sub-estimá-los quando as parcelas são pequenas, passando, em geral, a super-estimá-los quando se aumentam os tamanhos das parcelas. Nos dois casos, o vício da estimativa do desvio padrão aumenta com o tamanho da amostra.

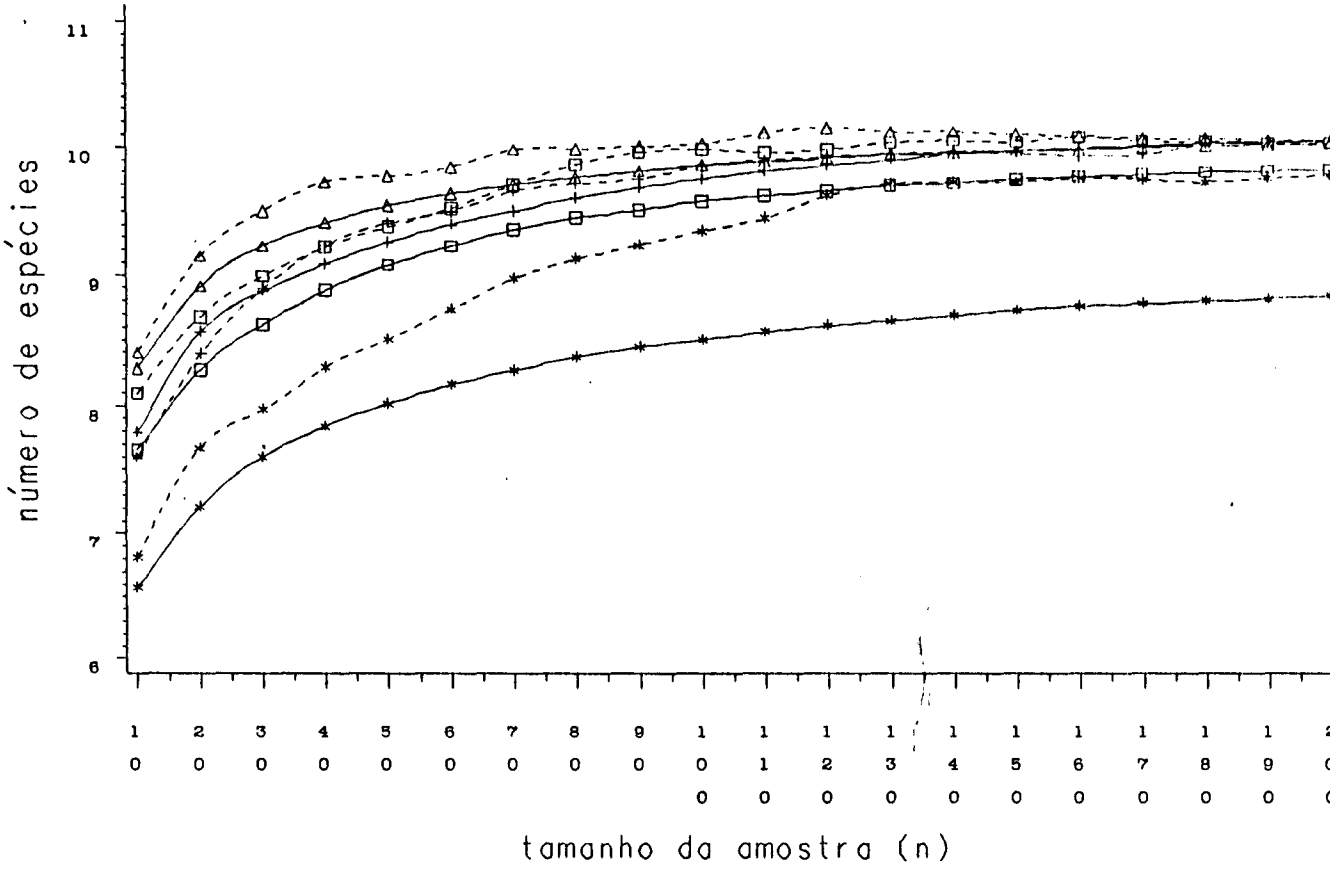

Figura 3.3.1. Curvas do número de espécie em relação ao tamanho da amostra. Resultados bootstrap (——) e Monte Carlo (---), para a comunidade II. (✱) :Q=10 (+) :Q=20 (□) :Q=30 (Δ) :Q=40.

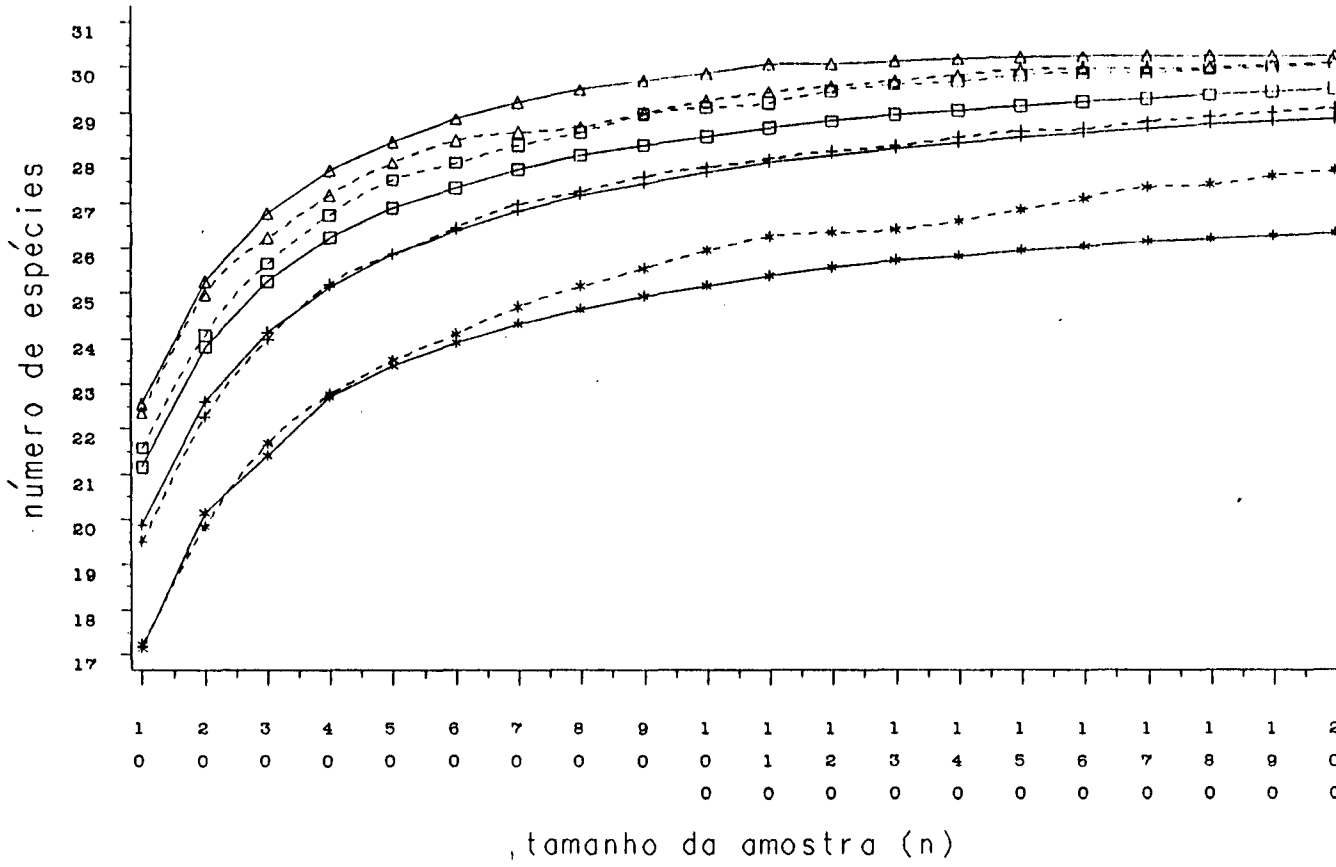

Figura 3.3.2. Curvas do número de espécie em relação ao tamanho da amostra. Resultados bootstrap (——) e Monte Carlo (---), para a comunidade IV. (✱) : Q=10 (+) : Q=20 (□) : Q=30 (Δ) : Q=40.

De modo geral, os valores das estimativas bootstrap dos desvios padrões para os desvios padrões médios são pequenos, diminuindo com o aumento de n.

Fixando-se o número de indivíduos esperados na amostra podem-se analisar o efeito do tamanho do quadrat e do tamanho da amostra sobre as estimativas bootstrap. Será que para um mesmo número de indivíduos amostrados, o bootstrap comporta-se melhor quando poucos quadrats grandes ou muitos quadrats pequenos são amostrados? Observando-se as tabelas 3.3.2 e 3.3.3 notam-se que o vício das estimativas bootstrap de S diminui em relação ao aumento de Q. Isto é, para um mesmo número de indivíduos amostrados, melhores estimativas são obtidas, com relação ao vício, para um menor número de quadrats grandes. Esse resultado é realçado para amostras com grande número de indivíduos. As estimativas dos desvios padrões parecem não serem influenciadas sob esse ponto. Podem-se verificar que o comportamento do estudo de Monte Carlo não apresenta essas características já que as comunidades foram geradas considerando-se as espécies independentes entre si.

Uma segunda forma de obtenção de faixas de confiança para a curva do número de espécies em função do tamanho da amostra é apresentada na figura 3.3.3 que é dada por um desvio padrão acima e um baixo da curva dos valores de $\hat{E}_B(B_n(S))$ em relação a n.

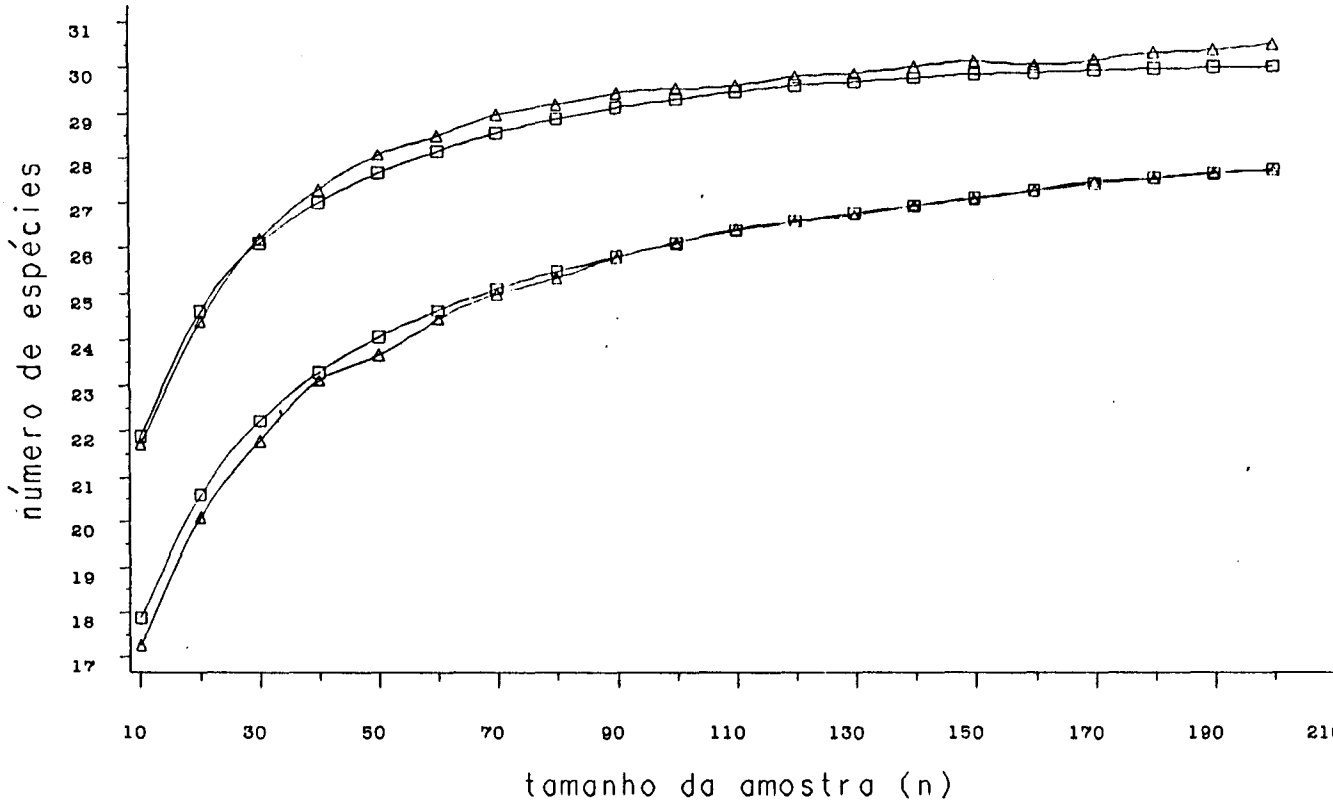

Figura 3.3.3. Faixas de confiança para a curva do número de espécies em relação ao tamanho da amostra para a comunidade IV, Q=20. (Δ) Monte Carlo, (□) bootstrap.

Fazendo-se um apanhado geral dos resultados do estudo chegam-se às seguintes conclusões:

a) O método bootstrap tem um bom desempenho para comunidades com poucas espécies raras. Os desvios padrões são adequadamente estimados e a construção das faixas de confiança

mostrou, rapidamente, o tamanho da amostra necessário para estimar o número total de espécies, fornecendo uma regra de parada em uma amostragem sequencial.

b) O problema em comunidades com muitas espécies raras é que um tamanho de amostra muito grande é necessário, como indica o estudo de Monte Carlo nas tabelas 3.3.2 e 3.3.3. Apesar do método bootstrap, em geral, sub-estimar ligeiramente o número de espécies, o seu desempenho é bom. O uso das faixas de confiança como regra de parada reflete o problema que é inerente à situação de muitas espécies raras. A vantagem aqui, é a obtenção dos desvios padrões das estimativas sem fixar modelos.

c) O método bootstrap é sensível ao número de espécies raras e não parece sensível ao número total de espécies.

d) Talvez, alguma correção, a ser determinada, venha melhorar o desempenho do método bootstrap, eliminando o vício.

e) Uma grande vantagem do método é que, na prática, é suficiente observar a presença ou ausência das espécies nas parcelas.

f) Fixado o número de indivíduos, obtêm-se melhores resultados, com o método bootstrap, quando a amostra é composta de poucos quadrats com muitos elementos.

CAPÍTULO 4

APLICAÇÃO DO MÉTODO BOOTSTRAP A UM CONJUNTO DE DADOS REAIS

A metodologia bootstrap foi aplicada a um conjunto de dados reais, gentilmente cedido e detalhado pelo Prof. Dr. George John Shepherd, Depto. de Botânica, Instituto de Biologia, UNICAMP.

Os dados foram coletados na reserva municipal da Mata de Santa Genebra, situada no distrito de Barão Geraldo, região norte do município de Campinas, São Paulo, às margens da rodovia Campinas-Paulínia, no período de dezembro de 1985 a agôsto de 1986. A área total da reserva é de 200 hectares, constituída de matas de Planalto. Esse tipo de vegetação caracteriza-se por ser descontínua, entremeada por cerrados, cerradões, campos ruprestes e matas ciliares (Leitão Filho, 1982 - em Shepherd, 1987).

Os dados foram obtidos a partir da delimitação de uma parcela de 200 metros de comprimento e 50 metros de largura, subdividida em 100 parcelas iguais de 10x10 metros cada uma, num total de 10.000 m^2. Dentro de cada parcela menor coletaram-se as informações de interesse que, para esse estudo se resumem na presença ou ausência das diferentes espécies em cada parcela.

Todo indivíduo vivo (árvore) com diâmetro na altura do peito (aproximadamente 1.30 m) igual ou superior a 4.8 cm (15 cm de perímetro) foi codificado. Como o conjunto de dados foi a partir da amostragem exaustiva da parcela maior, ele foi considerado um senso da comunidade em questão. No total, foram observados 1465 indivíduos pertencentes a 103 espécies diferentes. A tabela 4.1 apresenta a lista das 103 espécies com o número de quadrats que cada uma delas estava presente.

NOME DA ESPÉCIE	Nº DE PARCELAS QUE CADA ESPÉCIE ESTAVA PRESENTE
Aspidosperma polyneuron	89
Trichilia lagoensis	84
Psychotria wanthieri	71
Astronium graveolens	46
Actinostemon communis	31
Piptadenia gonoacantha	29
Esenbeckia febrifuga	23
Trichilia claussenii	20
Ixora gardneriana	19
Acacia polyphylla	17
Syphoneugenia densiflora	17
Trichilia pallida	17

Zanthoxylum minutiflorum	16
Chrysophyllum gonocarpum	15
Galipea jasminiflora	14
Machaerium stipitatum	13
Syagrus romanzoffiana	12
Eugenia ligustrina	10
Holocalyx balansae	9
Chorisia speciosa	9
Ixora velunosa	8
Solanum swartzianum	8
Coffea arabica	7
Cariniana estrellensis	7
Madraerium brasiliensis	6
Campomanesia mascalantha	6
Zanthoxylum pohlianum	6
Cariniana legalis	6
Chorophora tinctoria	6
Jacaranda micrantha	6
Seguieria langsdorffii	6
Urera baccifera	6
Casearia sylvestris	5
Myroeugenia campestris	5
Myrcia rostrata	5
Colubrina glandulosa	5
Pachystroma longifolium	5
Rhaminidium elaeocarpum	5

Guapira opposita	5
Qualea jundiahy	5
Machaerium aculeatum	4
Casearia gossypiospermum	4
Sweetia fruticosa	4
Myrtaceae sp1	4
Eugenia sp1	4
Guatteria nigrescens	4
Groton salutanis	4
Savia dictyocarpa	4
Sebastiania edwalliana	4
Zanthoxylum regnellianum	3
Cantarea hexandra	3
Trichilia hirta	3
Allophylus edulis	3
Celtis tala	3
Agonandra englerii	3
Duguetia lanceolata	3
Inga luschnatiana	3
Zanthoxylum cheloperone	3
Zanthoxylum hiemale	3
Myroxilon peruiferum	2
Croton floribundus	2
Gomidesia affinis	2
Laplacea semiserrata	2
Chomelia obtusa	2

Chomelia sericea	2
Cedrela fissilis	2
Guarea guidonia	2
Cupania vernalis	2
Picramnia warmingiana	2
Rollinia sylvatica	2
Sessea brasiliensis	2
Protium widgrenii	2
Vernomia diffusa	2
Amaiova gruanensis	2
Centrolobium tomentosum	2
Luehea speciosa	2
Myrtaceae sp3	2
Capaifera langsdoffii	2
Hirtella hebeclada	1
Myrtaceae sp2	1
Cuspania paniculata	1
Aegiphylla sellowiana	1
Ocotea puberula	1
Casearia obliqua	1
Trichilia elegans	1
Patagonula americana	1
Xylopia brasiliensis	1
Lafoensia pacari	1
Casearia decandra	1
Aspidosperma cylindrocarpum	1

Myxcia rostrata	1
Inga affinis	1
Cordia ecalyculata	1
Cauterea contracta	1
Trema micrantha	1
Miconia enaequidem	1
Diatenopteryx sorbifolia	1
Maytenus communis	1
Nectandra saligna	1
Cryptocarya moschata	1
Cordia trichotoma	1
Pseudobombax grandiflorum	1
Machaerium villosum	1

Tabela 4.1. Lista das espécies observadas na área estudada da reserva municipal da mata de Santa Genebra, Campinas, S.P.
Fonte: George J. Shepherd, Departamento de Botânica, I.B.-UNICAMP (1986)

A maioria das espécies nessa região são muito raras, como mostra a figura 4.1.

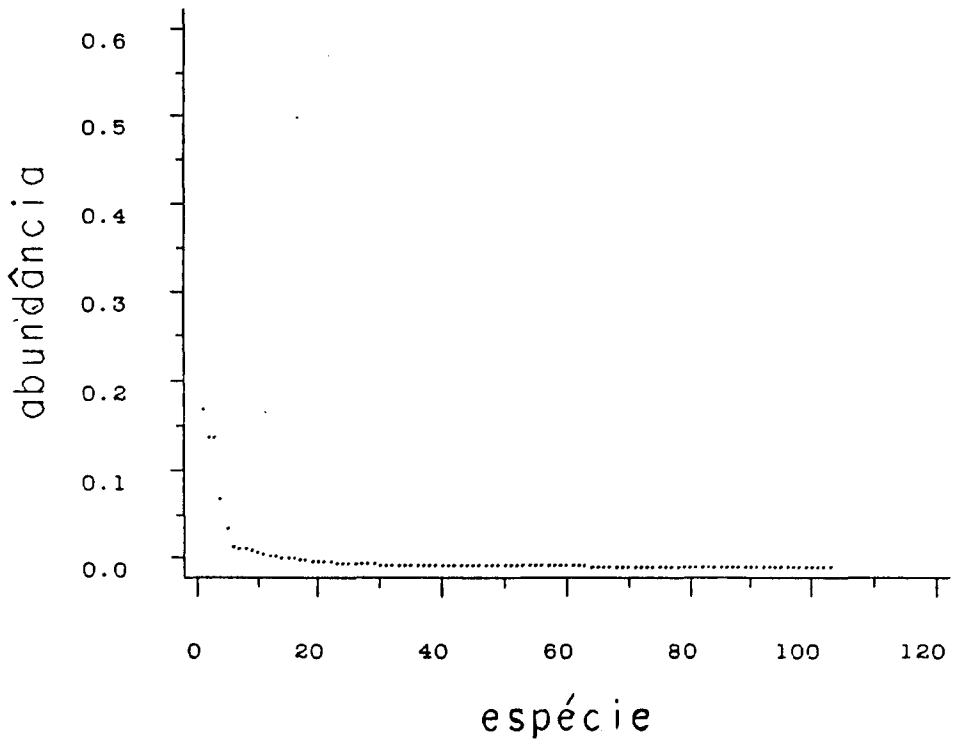

Figura 4.1. Distribuição da abundância das espécies para os dados coletados na reserva municipal da mata de Santa Genebra, Campinas, São Paulo. Fonte: George J. Shepherd, Departamento de Botânica, I.B.-UNICAMP (1986).

As três espécies mais abundantes constituem 22.05% , 18.23% e 14.88% cada uma, somando 55.16% dos indivíduos da comunidade, 81 espécies são constituídas de menos de 0.3% dos indivíduos, cada uma. Trata-se, portanto de uma comunidade rica em espécies e destas, poucas são frequentes e a maioria muito raras, indicando baixa equitabilidade.

O número médio de indivíduos por parcela foi de 14.65 com desvio padrão de 3.98. Na Figura 4.2 tem-se a distribuição do número de indivíduos por parcela. A adequabilidade do modelo de Poisson com média 14.65 à distribuição do número de indivíduos por parcela foi constatada através da estatística qui-quadrado, com probabilidade de significância aproximada de 0,75. O gráfico "Poissonness" também foi construído e indicou bom ajuste do modelo (Hoaglin, D.C. e Tukey, J. 1985)

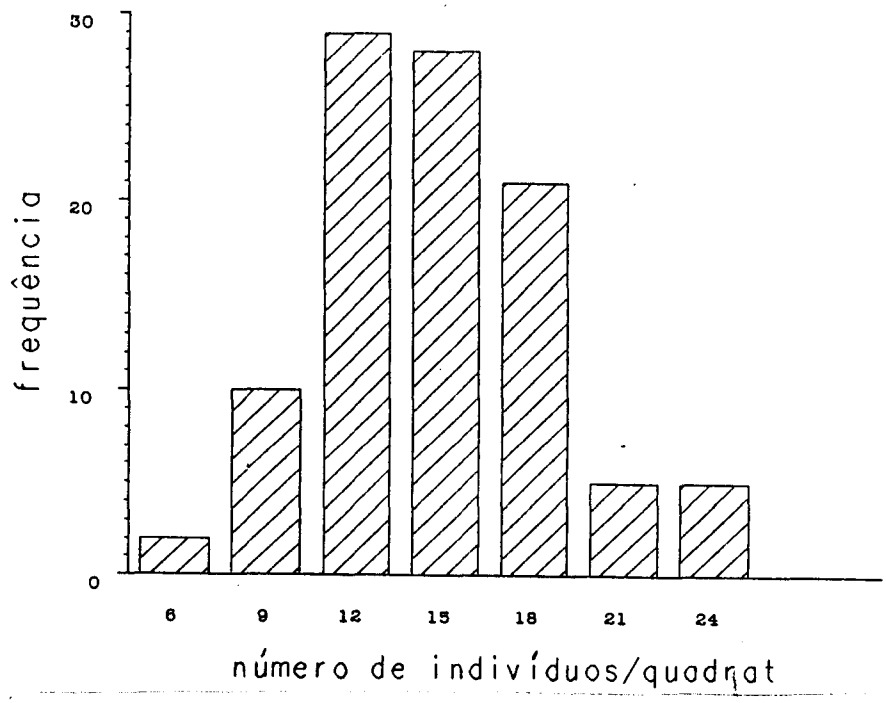

Figura 4.2. Distribuição do número de indivíduos por quadrats para os dados da mata de Santa Genebra.

Fonte: George J. Shepherd, Departamento de Botânica, I.B.-UNICAMP (1986).

Um aspecto interessante nesse conjunto de dados é que ele provém de uma situação natural bastante diferente daquela construída no estudo de simulação anterior. A pesquisa do Prof. George J. Shepherd constatou a presença de alguns padrões espaciais na floresta (algumas espécies se distribuindo em conglomerados) decorrentes das condições e interações ecológicas da região e também, a existência de clareiras indicando a interferência de fatores externos dentro da floresta. Essas características podem influenciar no comportamento do método bootstrap.

Uma amostra aleatória de N=80 parcelas foi selecionada, sequencialmente,a partir das 100 parcelas. Como no procedimento anterior, amostras foram acumuladas para n=10,20,30, ... ,80 e o método bootstrap foi aplicado (B=100 e R=50), reamostrando-se as parcelas para cada n. O número de espécies foi estimado para cada n e as curvas do número de espécies em relação ao tamanho da amostra foram construídas. Para analisar o comportamento do método bootstrap nessa situação, 100 repetições Monte Carlo foram realizadas, onde para cada repetição selecionaram-se uma amostra aleatória de tamanho 80, sem reposição e sequencialmente, dos 100 quadrats. O número de espécies corrigido e seu desvio padrão foram obtidos para as amostras acumuladas (n=10,20, ... ,80).

Os resultados Monte Carlo e bootstrap encontram-se na Tabela 4.2; os valores na tabela são as médias e desvios padrões médios para o número de espécies, para cada n (n=10,20, ... ,80).

Os números entre parênteses são os desvios padrões estimados pelo bootstrap para os desvios padrões médios do número de espécies.

Método	MONTE CARLO		BOOTSTRAP	
n	MÉDIA	DES.PAD.	MÉDIA	DES.PAD.
10	39.25	2.25	29.22	4.24 (0.31)
20	52.28	3.08	42.0	4.50 (0.33)
30	65.20	3.63	53.58	4.53 (0.36)
40	78.26	2.90	63.37	4.28 (0.36)
50	89.41	3.04	73.86	4.73 (0.33)
60	98.91	3.07	79.71	4.93 (0.34)
70	103.05	2.91	85.65	4.23 (0.32)
80	110.31	3.14	92.70	4.10 (0.29)

Tabela 4.2 Resultados Monte Carlo e Bootstrap para o número de espécies em relação ao tamanho da amostra (n), para os dados da reserva municipal da mata de Santa Genebra, Campinas, S.P.

Fonte: George J. Shepherd, Departamento de Botânica - I.B.- UNICAMP (1986)

Os resultados do estudo Monte Carlo, vistos mais claramente na figura 4.3 mostram que a curva do número de espécies com relação ao tamanho da amostra continua crescente, mesmo para amostras grandes. A partir de 60 parcelas amostradas, a curva apresenta uma diminuição da taxa de crescimento do número de espécies, mas longe de uma estabilização. Para $n>70$ o número de espécies passa a ser super-estimado devido à correção do vício. Esse comportamento é refletido pela grande quantidade de espécies espacialmente raras na comunidade. Podem-se considerar os valores dos desvios padrões do número de espécies pequenos (analisando os coeficientes de variação), porém, estes não diminuem com o aumento do tamanho da amostra, como esperavam-se. Suspeitam-se que esse comportamento é devido ao grande número de espécies raras e ao padrão espacial de comunidade.

Os resultados bootstrap mostram um acompanhamento do comportamento Monte Carlo, porém sub-estimando o número de espécies, como aconteceu com as populações simuladas. O vício da estimativa bootstrap do número de espécies, em relação ao valor obtido por Monte Carlo, aumenta com o tamanho da amostra, mas esse aumento é devido à super-estimação do Monte Carlo, para tamanhos de amostras grandes. Já as estimativas bootstrap dos desvios padrões super-estimam os desvios padrões do número de espécies obtidos por Monte Carlo. A Figura 4.3 apresenta as faixas de confiança para o número de espécies corrigido em relação ao tamanho da amostra, obtida por Monte Carlo e pelo bootstrap, mostrando a sub-estimação da curva e a super-estimação da faixa de confiança pelo bootstrap.

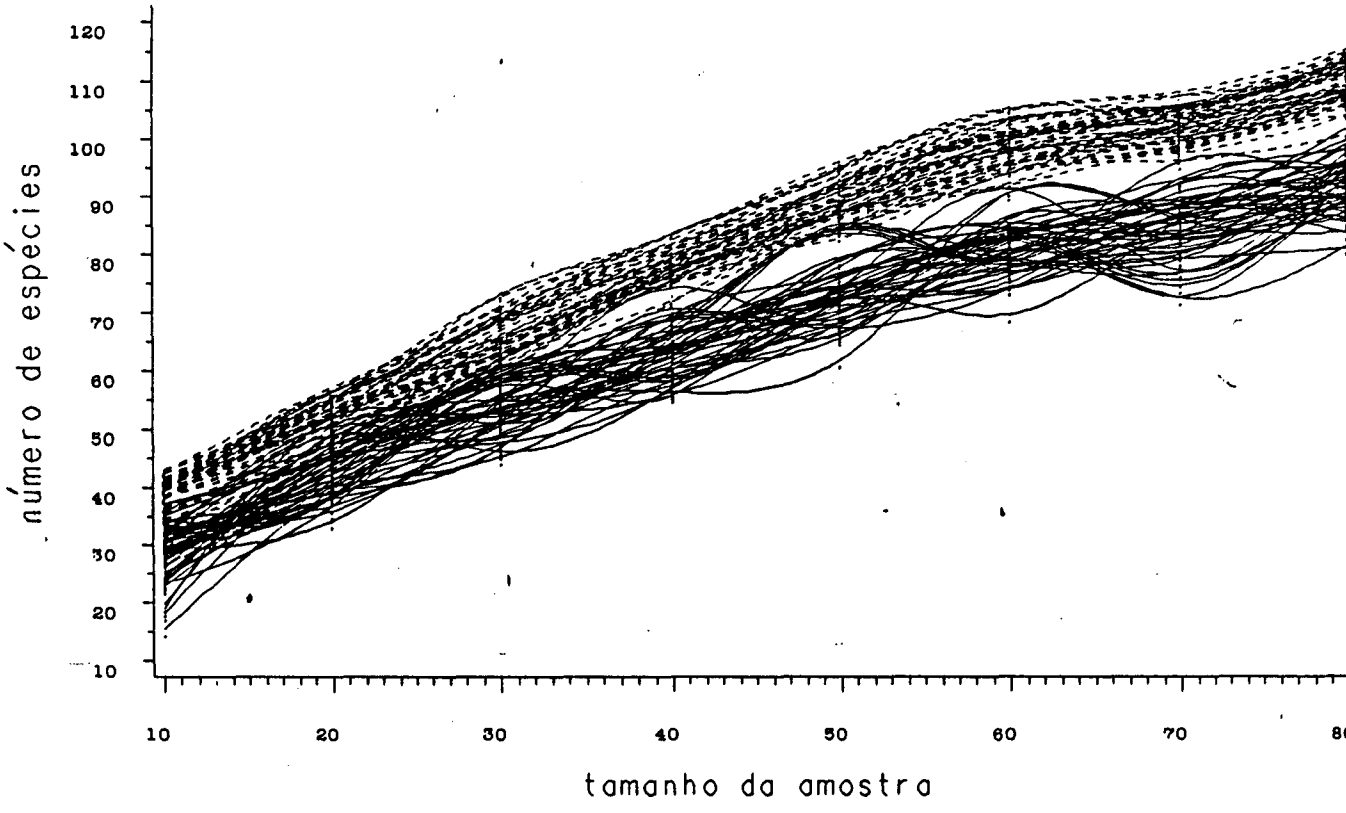

Figura 4.3. Faixas de confiança para o número de espécies em relação ao tamanho da amostra, obtidas por Monte Carlo (--) e bootstrap (—), para os dados da mata de Santa Genebra, Campinas, S.P.
Fonte: George J. Shepherd, Departamento de Botânica, I.B.-UNICAMP (1986).

As características nos resultados são semelhantes aos caso anterior de comunidades simuladas, cuja estrutura é totalmente distinta. Novamente, o comportamento do estudo de

Monte Carlo foi acompanhado, observando-se a necessidade de um fator de correção. O comportamento do bootstrap reflete que ainda existem mais espécies raras a serem observadas e que a amostragem deveria continuar. Em outra situação, só seria possível a obtenção das variâncias das estimativas através da adoção de um modelo.

O método bootstrap, parece não ter sido afetado pelo padrão geográfico (acompanhou com os problemas usuais, o Monte Carlo).

REFERÊNCIAS BIBLIOGRÁFICAS

- Abele,L.G. & Connor,E.F. (1979)1 Application of island biogeography theory to refuge design: making the right decision for the wrong reasons. Em *Proceedings of the First Conference on Scientific Research in the National Parks*. Vol 1, pags 89-94 (R.M. Linn ed.). Department of the Interior, Washington, D.C.

- Adams,E.J. e McCune,D.E. (1979). Application of the generalized jackknife to shannon's measure of information used as an index of diversity. Em *Ecological Diversity in Theory and Practice. Statistical Ecology Series* Vol.6 :349-368 (J.F. Grassle,G.P. Patil e C. Taillie eds.). International Co-operative Publishing House, Burtonsville, Maryland.

- Anscombe,F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37 :358-382.

- Arrhenius,O. (1921). Species and area. *Journal of Ecology*, 9 :95-99.

- Arvensen,J.N. (1969). Jackknifing U-Statistics. *Annals of Mathematical Statistical*, 40 :2076-2100.

- Birch, M.W. (1963). An algorithm for logarithmic series distribution. *Biometrics*, 19 :651-652.

- Bliss,C.I. (1965). An analysis of some insect trap records. Pags. 385-397. Em *Classical and Contagious Discrete Distributions*. (G. P. Patil, ed.) Statistical Publishing Society. Calcutta.

- Brenner,W. (1921). Vaxtgeografiska studien: Barosunds skargard. I. Allman del och floran. *Acta societatis pro Fauna et Flora Fennica*. 49 :1-151.

- Brian,M.V. (1953). Species frequencies in randon samples from animal population. *Biometrics*, 19 :651-652.

- Bulmer,M.G. (1974). On fitting the Poissin lognormal distribution to species abundance data. *Biometrics*, 30 :101-110.

- Coleman,B.D. (1981). On randon placement and species-area relations. *Mathematical Bioscience* 54 :191-215.

- Connor,E.F. e McCoy,E.D. (1979). The statistical and biology of species-area relationship. *The American Naturalist* 113(6) :791-832.

- Diaconis,P. e Efron,B. (1983). Computer-intensive methods in Statistics. *Scientifical American* 248(5) :96-108.

- East,R. (1983). Application of species-area curves to African Savannah Reserves. *African Journal of Ecology*, 21 :123-128.

- Efron,B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 1 :1-20.

- Efron,B. (1981). Nonparametric estimates of standart error: The jackknife, the bootstrap and other methods. *Biometrika* 68(3) :589-599.

- Efron,B. e Tibshirani,R. (1985). *The bootstrap method for assessing statistical accuracy*. Stanford, Califórnia, 1985. (Tecnical Report n° 101 - Division of Biostatistic Stanford University).

- Efron,B. and Thisted,R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* 63(3) :435-447.

- Evans,F.C., Clark,P.J., e Brand,R.H. (1955). Estimation of the number of species present on a given area. *Ecoloty*, 36 :342-343.

- Fager,E.W. (1972). Diversity: A sampling study. *The American Naturalist* 106(949).

- Fischer,A.G. (1960). Latitudinal variations in organic diversity. *Evolution*, 14 :64-81.

- Fisher,R.A., A.S. Corbet e C.B. Willians (1943). The relation between the number of species and the number of individuals in a randon sample of an animal population. *J. Anim. Ecol.*, 12 :42-58.

- Gleason,A.H.(1922). On the relation between species and area. *Ecology*, 3 :158-162.

- Gleason,A.H. (1925). Species and area. *Ecology*, 6 :66-74.

- Goodall,D.W. (1952). Quantitative aspects of plant distribution. *Biological Review of Cambridge Philosophical Society*, 27 :194-245.

- Greig-Smith,P. (1983). *Quantitative Plant Ecology*, 3ª ediçãao. Butterworts, London.

- Heck,K.L.; Belle,G. van and Simberloff,D. (1975). Explicit calculation of the rarefaction diversity measurement and determination of sufficient sample size. *Ecology* 56 :1459-1461.

- Heltshe,J.F. e Forrester,N.E. (1983). Estimating species richness using the jackknife procedure. *Biometrics* 39 :1-11.

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

- Heyer,W.R. e Berven,K.A. (1973). Species diversity of hepertofaunal samples from similar micro habitats at two tropical sites. *Ecology* 54 :642-645.

- Hoaglin,D.C. e Tukey,J.W. (1985). Checking the shape of discrete distribuitions. Em *Exploring Data Tables, Tends, and Shapes* (D.C. Hoaglin,F. Hosteller e J.W. Tukey, eds.). Jonh Wiley & Sons.

- Holgate,P. (1969). Species frequency distributions. *Biometrika*, 56 :651-660.

- Hopkins,B. (1955). The species-area relations of plant communities. *Journal of Ecology*, 43 :409-426.

- Hughes,R.G.(1976). Theories and models of species abundance. *The American Naturalist* 128(6) :879-899.

- Hurlbert,H.S. (1971). The nonconcept of species diversity: A critique and alternative paramters. *Ecology* 52(4) :577-586.

- Jaccard,P. (1902). Lois de distribution florate dans la zone alpine. *Bulletin de la Sociètè Vandoise des Sciences Naturelles*, 44 :223-270.

- Jaeckel,L. (1972). The infinitesimal jackknife. *Bell Laboratories Memorandum* # MM 72-1215-11.

- Kempton,R.A. (1979). The structure of species abundance and measurement of diversity. *Biometrics* 35 :307-321.

- Kobayashi,S. (1979). Species-area curves. Em *Statistical Distribuitions in Ecological Work* volume 4 :349-368. *Statistical Ecology Series*. International Co-operative Publishing House, Burtonsville, Maryland.

- Kobayashi,S. (1981). Diversity indices: Relations to sample size and spatial distribuition. *Japan Journal Ecology* 31 :231-236.

- Kobayashi,S. (1982). The rarefaction diversity measurement and the spatial distribuition of individuals. *Japan Journal Ecology* 32 :255-258.

- Kolata,G. (1984). The art of learning from experience. *Science* volume 225(13) :156-158.

- Leitão Filho,H.F. (1982). Aspectos taxonômicos das florestas do estado de São Paulo. Em *Anais do Congresso Nacional sobre Essências Nativas*. *Silvicutura* 16(1) :197-206.

- Lloyd,M., e R.J. Ghelard: (1964). A table for calculating the "equitability" component of species diversity. *J. Anim. Ecol.*, 33 :217-225.

- MacArthur,R.H. (1957). On the relative abundance of bird species. *Proc Nat. Acad. Sei. Wash.*, 43 :293-295.

- MacArthur,R.H. e E.O. Wilson (1967). *The theory of island biogeography*. Princeton University Press, Princeton, N.J.

- May,R.M. (1975). Patterns of species abundance and diversity. Em *Ecology and Evolution of Communities* (M.L. Cody e J.M. Diamond, eds.). Belknap Press, 81-120.

- McGuinness,A.K. (1984). Equations and explanations in the study of the species-area curves. *Biological Reviews* 59 :423-440.

- Miller,R.G. (1974). The jackknife: A review. *Biometrika* 61(1) :1-15.

- Miller,R.I. & Harris,L.D. (1977). Isolation and extirpations in wildlife reserves. *Biological Conservation*, 12 :311-315.

- Peet,R.K. (1974). The mensurement of species diversity. *Annual Review of Ecology and Systematics*, 5 :285-307.

- Pianka,E. (1966). Latitudinal gradients in species diversity: a review of concepts. *American Naturalist*, 100 :33-46.

- Pielou,E.C. (1969). *An Introduction to Mathematical Ecology*. Wiley, New York.

- Pielou,E.C. (1975). *Ecological Diversity*. John Wiley & Sons, London, Sydney, Toronto.

- Preston,F.W. (1948). The commonness, and rarity, of species. *Ecology*, 29 :254-283.

- Preston,F.W. (1960). Time and space and variation of species. *Ecology*, 41 :611-627.

- Preston,F.W. (1962). The canonical distribution of communness and rarity. *Ecology*, 43 :185-215 e 410-432.

- Quenouille,M.H. (1949). Aproximate tests of correlation in time-series. *Journal of Statistical Computation and Simulation*, 8 :75-80.

- Raup,D.M. (1975). Taxonomic diversity estimation using rarefaction. *Paleobiology*, 1 :333-342.

- Routledge,R.D. (1979). Diversity indices: Which ones are admissible? *Journal of Theoretical Biology* 76 :503-513.

- Routledge,R.D. (1980). Bias in estimating the diversity of large, uncensused communities. *Ecology* 61(2) :276-281.

- Sanders, H. L. (1968). Marine benthic diversity : A comparative study. *The American Naturalist* 102(925) : 243-282.

- Schucany,W.R.; Gray,H.L. e Owen,D.B. (1971). On bias reduction in estimation. *Journal the American Statistical Association* 66(335) :524-533.

- Sharot, T. (1976). Sharpening the jackknife. *Biometrika* 63(2) : 315-321.

- Shepherd,G.J.. *Estudo Florístico e Fitossociológico da Reserva da Mata de Santa Genebra, Campinas*. Campinas, FAPESP, 1987.66p. (Relatório do Instituto de Biologia - Departamento de Botânica, UNICAMP).

- Shinozaki,K. (1963). Note on the species-area curve. Apresentado no *10th Annual Meeting of Ecological Society of Japan*, 1963, Tokio, 5.

- Simberloff,D. (1971). Properties of the rarefaction diversity mesurement. *The American Naturalist*, 107 :414-418.

- Simberloff,D.S. (1972). Models in biogeography. Pags. 160-191. Em *Models in Paleobiology*. (T.J.M. Schof, ed.). Freeman, San Francisco.

- Simberloff,D.S. & Abele,L.G. (1982). Refuge design and island biogeographic theory: effects of fragmentation. *American Naturalist*, 120 :41-50.

- Smith,W. e Grassle,F. (1977). Sampling properties of a family of diversity measures. *Biometrics* 40 :119-129.

- Smith, P. E. e Belle, G. van (1984) . Nonparametric estimation of species richness. *Biometrics* 40 :119-129.

- Simpson,E.H. (1949). Mesurement of diversity. *Nature*, 163 :688.

- Tukey,J. (1958). Bias and confidence in not quite so large samples. *Annals of Mathematical Statistics*, 29 :614.

- Whittaker,R.H. (1972). Evolution and measurement of species diversity. *Taxon*, 21 :213-251.

- Willians,C.B. (1943). Area and number of species. *Nature*, 152 :264-267.

- Willians,C.B. (1944). Some applications of the logarithmic series and the index of diversity to ecological problems. *Jornal of Ecol.*, 32 :1-44.

- Willians,C.B. (1947). The logarithmic series and its application to biological problems. *J. Ecol.*, 34 :253-272.

- Willians,C.B. (1964). *Patterns in the Balance of Nature* Academic Press, London.

- Zahl,S. (1977). Jackkinifing an index of diversity. *Ecology* 58 :907-913.