In: Bicudo, C. & D. Bicudo. Amostragem em Limnologia (no prelo).

Suficiência amostral

Valério De Patta Pillar
Departamento de Ecologia, Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, 91540-000, Brasil
E-mail: vpillar@ecologia.ufrgs.br

Resumo. A escolha de procedimentos de amostragem deve ser guiada pelos objetivos do estudo e características do meio a ser amostrado. Em estudos limnológicos, e em ecologia em geral, o meio e os objetivos nem sempre se enquadram nas condições ideais consideradas pela estatística convencional. Este capítulo define termos, discute procedimentos de amostragem, e apresenta novos métodos para a determinação de suficiência amostral baseados na reamostragem dos próprios dados coletados. Métodos de reamostragem são descritos para avaliação de suficiência amostral quando o objetivo é a estimativa de parâmetros simples, tais como médias de uma varíavel, e quando o objetivo do levantamento é o reconhecimento de padrões e sua interpretação, com o uso de análise de agrupamentos e ordenação.

Palavras-chave: Amostragem, Análise de agrupamentos, Análise multivariada, Auto-reamostragem, "Bootstrap", Delineamento, Estimação, Intervalos de confiança, Ordenação, Reamostragem, Suficiência.

INTRODUÇÃO

A amostragem é necessária porque em geral não é possível ou não é conveniente acessar a totalidade de um dado universo amostral ou população. Assim, tomam-se informações sobre uma parte deste, uma *amostra*, para inferir atributos sobre o todo. As unidades que compõem o universo amostral e a amostra, ou seja, as *unidades amostrais*, podem ser objetos perfeitamente distinguíveis, tais como um indivíduo vegetal ou animal, ou um ponto, ou um evento (relacionado a comportamentos, por exemplo). As unidades amostrais em levantamentos de ecossistemas, porém, são comumente agregados de objetos, com limites arbitrários, tais como um volume de água, de solo ou de sedimentos, ou uma área de vegetação. O universo amostral é especificado pelo(a) pesquisador(a). Em limnologia, dependendo do contexto, o universo amostral pode ser um pequeno tanque experimental, uma porção de um rio, lago ou banhado, ou até toda uma bacia hidrográfica. Da mesma forma, procedimentos de laboratório podem envolver amostragem; e.g., contagem de organismos em uma placa de Petry, cujas unidades amostrais são campos selecionados para contagem.

Quando a única informação disponível é de uma amostra tomada de um universo amostral, não é possível saber se o estado de um atributo obtido a partir da amostra coincide exatamente com o estado verdadeiro desse atributo no universo amostral. Porém, quanto maior o *número de unidades amostrais*, i.e., o *tamanho da amostra*, maior é a probabilidade de que novas amostras tomadas do mesmo universo amostral permitirão as mesmas conclusões. A avaliação da *precisão* da estimativa indicará a amplitude de estados em que é mais provável que se encontre o estado verdadeiro do atributo no universo amostral. Portanto, em qualquer levantamento será sempre necessário avaliar se o tamanho da amostra é suficiente para uma dada precisão requerida. Deve ser também considerado que a quantidade de trabalho e materiais utilizados em um levantamento é em grande parte função do tamanho da amostra, sendo portanto a avaliação de suficiência amostral uma ferramenta importante para o uso racional desses recursos.

A ênfase deste capítulo é em métodos para a determinação de suficiência amostral. Para avaliar suficiência amostral, poderíamos seguir as orientações da teoria amostral clássica (Cochran 1977, Green 1979, Krishnaiah & Rao 1988). Entretanto, no caso de levantamentos ecológicos, as soluções clássicas não são adequadas, pois estas assumem um universo amostral "bem-comportado" e não tão complexo como em sistemas ecológicos (Pillar 1998). Tal complexidade resulta de alta diversidade, respostas não-lineares, interações complicadas e, mais importantes do ponto de vista de amostragem, arranjos não-aleatórios nos espaços geográfico e ecológico (Orlóci 1993, Kenkel et al. 1989). Ademais, os métodos tradicionais não oferecem alternativas para avaliar suficiência amostral quando o objetivo do levantamento é o reconhecimento de padrões e sua interpretação. A precisão de estimativas obtidas por amostragem tem sido geralmente avaliada com base em distribuições teóricas de freqüências, e.g., distribuição normal, as quais nas condições acima descritas têm valor limitado (Patil et al. 1988, Orlóci 1993). Existem, porém, outros métodos que utilizam intensivamente a computação, tais como o método de reamostragem "bootstrap" que gera uma distribuição empírica a partir dos próprios dados (Efron 1979, Efron & Tibshirani 1993). A suficiência da amostra pode ser avaliada com base em limites de confiança ou probabilidades obtidas de tais distribuições empíricas (Pillar 1998, 1999a, 1999b).

Ao amostrar, também é necessário decidir quanto ao método de seleção e, em alguns casos, tamanho e forma das unidades amostrais que irão compor a amostra. Discutirei mais adiante que essas decisões devem ser guiadas pelo contexto, especialmente em ecologia, em que o meio amostrado nem sempre se enquadra nas condições ideais tratadas pela estatística convencional.

OBJETIVOS DA AMOSTRAGEM

Podemos distinguir duas categorias de objetivos em levantamentos limnológicos. O objetivo de um levantamento pode se restringir a uma estimativa de quantidades, tais como variáveis limnológicas físicas e químicas, biomassa, densidade de uma ou mais espécies, em que o resultado final obtido consiste geralmente em médias de cada uma dessas variáveis. Os levantamentos, porém, freqüentemente têm como objetivo estudar a variação desses ou de outros atributos dentro do universo amostral, buscando o reconhecimento de padrões no espaço e/ou no tempo e sua interpretação. Neste caso geralmente são utilizadas técnicas de análise multivariada, tais como classificação e ordenação (Orlóci 1978, Pielou 1984, Podani 1994, Legendre & Legendre 1998).

É importante notar, entretanto, que amostragens em diferentes níveis hierárquicos e com diferentes objetivos, podem estar envolvidas num mesmo estudo. Por exemplo, digamos que o objetivo principal de um levantamento seja descrever e interpretar a variação espacial e temporal entre zonas de um lago ao longo de um ano; um delineamento amostral sistemático será adotado, sendo que em cada ponto ao longo de cada dia de amostragem serão coletadas várias unidades amostrais para determinações de variáveis limnológicas; há aqui dois níveis de amostragem: (1) em cada ponto o objetivo da amostragem é obter uma estimativa das características médias ao longo de um dia, pois decidiu-se ignorar as variações horárias, (2) a análise conjunta dos dados médios dos pontos em vários dias ao longo do ano permitirá revelar padrões de variação no espaço e no tempo, os quais serão interpretados em relação a fatores externos, tais como clima e ação antrópica.

O EFEITO DA ESCALA

As unidades amostrais em levantamentos de ecossistemas são em geral agregados de organismos e de substrato, representando subdivisões arbitrárias de um meio contínuo (Orlóci 1993). O tamanho e forma da unidade amostral é definido pelo(a) pesquisador(a), pois em geral não é possível distinguir unidades amostrais com limites naturais. Por exemplo,

cada unidade amostral pode ser definida como uma determinada área, ou um determinado volume de água ou de sedimentos coletado utilizando um determinado tipo de equipamento. Esse elemento complicador da amostragem é evidente em ecologia de comunidades, manifestando-se na profícua discussão sobre o conceito de comunidade (vide Palmer & White 1994) e no fato de que as conclusões serão dependentes da escala ou tamanho da unidade amostral (Juhász-Nagy & Podani 1983, Greig-Smith 1983, Palmer 1988, Kenkel et al. 1989, Camiz & Gergely 1990, Podani et al. 1993).

Uma das características de sistemas ecológicos é a sua variação não-aleatória, o que se manifesta na existência de padrões no espaço e no tempo. A possível estratificação vertical e horizontal em lagos e cursos d'água é um exemplo. Nessas condições, medidas comparativas entre unidades amostrais tais como similaridade, dissimilaridade, e diferenças em diversidade, serão dependentes do tamanho das unidades amostrais. Nessas condições, parâmetros tais como a variância também serão dependentes do tamanho da unidade amostral: unidades maiores tenderão a ser menos variáveis entre si do que unidades menores. Sabe-se que quanto menor a variância, menor é o número de unidades amostrais necessárias para uma mesma precisão da estimativa de uma média (ver, e.g., Cochran 1977). Logo, se o objetivo é estimar a média de uma variável, unidades amostrais maiores e mais heterogêneas internamente permitem atingir suficiência amostral com um menor número de unidades amostrais. A decisão sobre o tamanho das unidades amostrais, neste caso, deve considerar também viabilidade e custo entre usar um menor número de unidades amostrais maiores, ou um maior número de unidades menores. Entretanto, se o objetivo da amostragem é revelar e interpretar padrões de variação, o procedimento é provavelmente o oposto, pois unidades amostrais muito grandes poderão borrar aspectos importantes da variação no sistema. Portanto, as condições de amostragem que satisfazem o objetivo de estimar atributos simples podem não coincidir com as que satisfazem o objetivo de estudar padrões (Orlóci & Pillar 1989).

Se a variação em sistemas ecológicos for, ao contrário, aleatória, o que raramente parece ser o caso em sistemas naturais, o universo amostral será homogêneo, não haverá efeito de escala, e os resultados não serão afetados pelo tamanho e forma das unidades amostrais (Palmer 1988). Um sistema é homogêneo quando, ao ser subdividido, as suas partes mantêm-se semelhantes (Palmer 1988). A homogeneização artificial do universo amostral é perfeitamente aceitável quando o objetivo da amostragem é obter um estimativa de uma média com o menor número possível de unidades amostrais. Por exemplo, na determinação de teores de fósforo, a agitação do material coletado permite reduzir drasticamente a variância entre determinações de um mesmo volume de material e até eliminar a necessidade de réplicas. Podemos dizer que o efeito da homogeneização é semelhante ao de utilizar uma unidade amostral de maior tamanho. Da mesma forma, a coleta de sub-unidades amostrais as quais são misturadas em uma unidade amostral composta é equivalente a aumentar o tamanho da unidade amostral.

É comum fazer-se a distinção entre variação espacial e variação temporal. Essa distinção, porém, é ambígua em ecossistemas muito dinâmicos, como em determinados ambientes aquáticos (Legendre & Legendre 1998). Tal particularidade de alguns sistemas aquáticos tem conseqüências importantes para a amostragem. A primeira é que unidades amostrais coletadas num mesmo ponto ao longo de um dado período de tempo poderão apresentar variação semelhante ao de várias unidades amostrais coletadas simultaneamente em vários pontos. Além disso, a utilização de uma "janela" temporal mais longa, que pode ser definida, por exemplo, como o tempo decorrido entre a primeira e a última coleta dentro de uma unidade amostral composta, tem efeito semelhante ao de um aumento do tamanho da unidade amostral. Outra conseqüência é que unidades amostrais coletadas num mesmo ponto

ao longo de um dia serão provavelmente independentes, um dos requisitos exigidos para alguns tipos de análises.

É importante notar que se as unidades amostrais são agregados, o universo de amostragem é contínuo, havendo teoricamente um número infinito de possíveis unidades amostrais, com infinitas opções de tamanho, forma, e localização dentro do universo amostral. Porém, quando as unidades amostrais são naturais, distintas, reconhecíveis, tais como organismos animais ou vegetais individuais ou unidades geográficas isoladas (ilhas, lagos), o universo amostral assim definido tem um tamanho finito e um número finito de amostras possíveis. O problema de amostragem nesse caso é mais simples; é apenas uma questão de definir o número e o método de seleção das unidades amostrais; o efeito da escala não estará presente.

SELEÇÃO DAS UNIDADES AMOSTRAIS

Uma amostra de n unidades tomada de um universo amostral de N unidades será uma possibilidade entre $C = \frac{N!}{n!(N-n)!}$ diferentes amostras. Como selecionar a amostra? O uso de amostragem sistemática, estratificada ou não, ou mesmo preferencial, é freqüente em levantamentos de ecossistemas; raramente é utilizada amostragem aleatória irrestrita (Orlóci 1978, Jongman et al. 1995, Goedickemeier et al. 1997). A seleção é aleatória irrestrita quando todas as unidades amostrais têm a mesma probabilidade de serem incluídas na amostra. Amostragem aleatória irrestrita tem sido considerada pouco prática no campo pela dificuldade em localizar os pontos de amostragem, os quais devem ser previamente escolhidos ao acaso sobre o mapa da área; mas atualmente essa dificuldade pode estar superada com o uso de sistemas automatizados de determinação de coordenadas geográficas (GPS). A amostragem é sistemática quando apenas o primeiro membro da amostra, ou do estrato, é selecionado ao acaso, sendo os demais tomados a intervalos regulares. A amostragem é estratificada quando o universo amostral é dividido em estratos, ou segmentos, o que pode ser feito de forma subjetiva, e dentro de cada estrato é feita a seleção aleatória ou sistemática das unidades amostrais. Quando o objetivo é a estimativa de atributos, por exemplo, de médias, a seleção das unidades amostrais deve seguir um desses métodos, pois do contrário a estimativa do atributo será viciada.

Exemplos

1. Amostragem aleatória irrestrita: Para avaliar o grau de contaminação da água captada para abastecimento urbano numa dada região a amostra foi selecionada aleatoriamente a partir de uma lista de pontos de captação.
2. Amostragem aleatória irrestrita: O objetivo do levantamento é descrever comunidades vegetais quanto a interações de espécies entre si e com fatores de ambiente. Mapeiam-se os limites da área. Há um número infinito de pontos para localizar quadros (unidades amostrais) aleatoriamente através de coordenadas geográficas. A definição do tamanho e forma da unidade amostral é arbitrária. Poderá ocorrer sobreposição de unidades amostrais.
3. Amostragem sistemática: Em uma lagoa pretende-se estudar as relações entre composição do fitoplancton e variáveis físicas e químicas da água. No mapa da lagoa marca-se um pivot aleatoriamente, sobre o qual posiciona-se um dos nós de uma grade quadriculada. A amostra será composta por unidades amostrais localizadas em todos os nós da grade que estiverem sobre a lagoa. A densidade de amostragem é definida pela distância entre-nós. Alternativamente, marcam-se transecções localizadas sistematicamente sobre a lagoa, as quais são percorridas, sendo as unidades amostrais localizadas sistematicamente ao longo de cada transecção.
4. Amostragem estratificada sistemática: No exemplo 3, a lagoa é dividida em estratos, de forma subjetiva ou de acordo com algum critério, e.g., profundidade. Uma amostragem sistemática é então realizada dentro de cada estrato. A estratificação garante que todas as áreas de interesse sejam incluídas na amostra.

No entanto, grande parte do que se sabe a respeito de processos biológicos, organismos, populações e comunidades é resultado de pesquisas em que foi usada amostragem preferencial, em que as unidades são selecionadas porque parecem típicas ao

pesquisador (Orlóci 1991). Nesse caso, ou a propriedade considerada é uniforme na população (e.g., número de cromossomas) não sendo importante o método de seleção das unidades estudadas, ou o objetivo da amostragem é confirmar padrões mais ou menos evidentes. Por exemplo, em taxonomia têm sido usados espécimens tipo; em fitossociologia, e.g. Braun-Blanquet (1979), selecionam-se sítios homogêneos para delimitar a comunidade vegetal a ser descrita, porque padrões de vegetação são muitas vezes óbvios ao pesquisador, sendo mais eficiente descrever cada mancha onde as comunidades parecem mais típicas, homogêneas, do que descrever e analisar um sem número de unidades amostrais aleatórias. A amostragem preferencial portanto tem sido aceita em ecologia quando se objetiva estudar ou confirmar padrões percebidos subjetivamente (Pillar 1998).

SUFICIÊNCIA AMOSTRAL

Solução tradicional

A solução tradicional (ver, e.g., Cochran 1977)**,** aplicável quando o objetivo é a estimativa de médias, é baseada na variância da média

$$S_{\bar{x}}^2 = \frac{S_X^2}{n}\left(1 - \frac{n}{N}\right)$$

onde S_X^2 é a variância da variável X, n o tamanho da amostra e N o tamanho do universo amostral.

Sendo o universo amostral muito grande, logo $\frac{n}{N} \approx 0$, e usando a distribuição t de Student, o tamanho da amostra pode ser determinado por

$$t = \frac{\delta}{\sqrt{S_{\bar{x}}^2}} \therefore t = \frac{\delta}{\frac{\sqrt{S_x^2}}{\sqrt{n}}} \therefore t = \frac{\delta\sqrt{n}}{\sqrt{S_X^2}} \therefore t^2 = \frac{\delta^2 n}{S_X^2} \therefore n = \frac{t^2 S_x^2}{\delta^2}$$

onde δ é a diferença mínima a ser detectada e t o valor da distribuição de Student para n-1 graus de liberdade correspondente a uma dada probabilidade $P(t^0 \geq t) = \alpha$. Como t depende de n, o valor de n é encontrado iterativamente.

Essa solução é problemática para levantamentos de ecossistemas porque (1) assume distribuição normal da variável X; e (2) o objetivo da amostragem pode não ser estimativa de médias e variâncias.

Amostragem iterativa

A amostragem iterativa encontra suporte na relação entre precisão e estabilidade. Quanto mais precisa a estimativa de um atributo, mais estável será a medida do atributo obtida de outras amostras de maior tamanho. A interpretação da amostragem como um processo de sucessivas aproximações tem precedentes em Greig-Smith (1983) para a estimativa de atributos simples e em Orlóci & Pillar (1989) para o estudo de padrões. Nessa abordagem o estado de um dado atributo obtido a partir da amostra evolui e atinge estabilidade na medida em que se aumenta o número de unidades amostrais na amostra. O tamanho suficiente de amostra é aquele no qual o atributo simples ou complexo de interesse começa a ter estabilidade, ou seja, quando o fato de agregar-se novas unidades amostrais à amostra resulta em alterações relativamente menores no valor do atributo considerado. Assim, se por um lado o objetivo é estimar a média de alguma variável, o tamanho suficiente da amostra será aquele em que a média da amostra atinge estabilidade. Um exemplo simples é o caso em que a média na amostra é monitorado para tamanhos sucessivos de amostra (Figura 1).

Definindo mais formalmente o método utilizado na Fig. 1, a estabilidade da amostra é percebida pela magnitude relativa da alteração do atributo de interesse entre passos de

amostragem com tamanhos crescentes de amostra $n_1, n_2, ..., n_k, ... n$, onde n_1 é um tamanho inicial de amostra (no primeiro passo de amostragem). O incremento constante do tamanho de amostra é s, que é o número de unidades amostrais agregadas à amostra a cada um dos passos de amostragem seguintes. O número total de passos de amostragem é $t = 1+\text{INT}((n-n_1)/s)$, mais 1 se n_k no último passo de amostragem não coincidir com n. INT indica a porção inteira do quociente. A escolha de um valor de s pequeno produzirá um grande número de passos de amostragem e uma curva mais regular do atributo de interesse.

Outra aplicação dessa abordagem é a curva "número de espécies *versus* número de unidades amostrais", muito usada em ecologia de comunidades para, entre outros objetivos, indicar suficiência de amostragem; o atributo considerado é o número de espécies. A curva "número de espécies *versus* tamanho da unidade amostral", usada para determinar a área mínima fitossociológica, é um caso análogo; o processo nesse caso pode ser entendido como uma agregação de novas unidades amostrais sistematica e contiguamente às que já estão na amostra. Quaisquer outros atributos, simples ou complexos (e.g., medidas de diversidade), poderiam também ser considerados nessas curvas.

A limitação da utilização da amostragem iterativa é que a ordem na qual as unidades amostrais são agregadas à amostra afeta a percepção de estabilidade da curva. Também, dependendo da precisão requerida, a amostra pode ser suficiente mesmo sem que a curva tenha atingido estabilidade. O método "bootstrap", discutido a seguir, simula reamostragem da própria amostra, permitindo avaliar o grau de estabilidade quando combinado a uma amostragem iterativa.

Reamostragem "bootstrap"

O método "bootstrap", inventado por Efron (1979, Efron & Tibshirani 1993), baseia-se no princípio de que não havendo melhor informação, a distribuição de freqüências na amostra é a melhor indicação da sua distribuição no universo amostral. "Bootstrap" poderia ser literalmente traduzido como "cadarço de bota", mas o termo é usado em linguagem figurada (Efron 1979). Creio ser *auto-reamostragem* um termo que expressaria adequadamente o significado do método "bootstrap" em português, ou seja, a reamostragem dos próprios dados; entretanto, deixo ao leitor a tarefa de adotar o neologismo. A reamostragem dos dados da amostra, com reposição, simula a reamostragem do universo amostral. Cada amostra obtida por reamostragem é uma *amostra bootstrap*. A amostra sendo reamostrada define um *pseudo universo amostral*. A reamostragem permite calcular a precisão de estimativas através de limites de confiança ou probabilidades.

O método bootstrap pode ser integrado à amostragem iterativa. Descrevo aqui o método aplicado a levantamento de ecossistemas, conforme Pillar (1998): Os dados obtidos estão arranjados em uma matriz com n unidades amostrais e p variáveis. Esses dados podem representar uma amostra num dado ponto de um processo de amostragem iterativa, amostra que poderá ser expandida se os resultados da avaliação de suficiência amostral assim indicarem. Dados já existentes podem também ser o ponto de partida, caso em que será avaliado se a amostra é suficiente para o objetivo desejado. As unidades amostrais podem ser de qualquer tipo, como explicado anteriormente. As variáveis podem ser atributos do substrato ou componentes biológicos, e.g., espécies, descritos nas unidades amostrais. O conjunto de n unidades amostrais é tomado como pseudo universo amostral. O algoritmo computacional reamostra com reposição o pseudo universo amostral, gerando amostras bootstrap com um número crescente de unidades amostrais $n_k \leq n$, e calcula para cada passo k de reamostragem, com tamanho de amostra n_k, o atributo θ_k^*. Este é o atributo do universo amostral que se tem interesse em inferir a partir da amostra. O atributo de interesse pode ser

simples, como a média ou a variância de alguma variável, ou mais complexo como a correlação entre duas variáveis, a medida da nitidez da classificação da amostra em um dado número de grupos (Pillar 1999a), ou a medida do estado da ordenação das unidades amostrais (Pillar 1999b). Avalia-se se o atributo de interesse atinge o nível mínimo de precisão dentro da amplitude de tamanhos de amostra $n_k \leq n$ avaliados; sendo o resultado positivo, conclui-se que o tamanho de amostra é suficiente. O detalhamento dos métodos para diferentes atributos será apresentado a seguir.

Suficiência amostral avaliada com base em limites de confiança

O método pode ser aplicado a qualquer atributo θ da amostra para o qual a suficiência amostral possa ser avaliada pela precisão da estimativa indicada por intervalos de confiança. Nesta categoria incluem-se atributos tais como a média ou a variância de alguma variável, a correlação entre duas variáveis, e outros que possam ser derivados a partir dos dados. Atributos adequados ao uso de intervalos de confiança são aqueles cujos valores podem ser interpretados diretamente, tais como os coeficientes de correlação (se o intervalo inclui zero ou não é um indicativo de significância), ou que serão comparados entre si diretamente, tais como médias.

O intervalo de confiança para um dado tamanho de amostra $n_k \leq n$ é obtido através do seguinte algoritmo de reamostragem bootstrap (Pillar 1998):

1. Seleciona-se aleatoriamente no pseudo universo amostral uma amostra bootstrap de tamanho n_k com reposição. Sendo a seleção com reposição, a mesma unidade amostral poderá aparecer mais de uma vez na mesma amostra bootstrap.
2. Computa-se na amostra bootstrap o parâmetro θ_k^* de interesse. O valor resultante é armazenado.
3. Repetem-se os passos 1 e 2 um grande número de vezes (indica-se no mínimo 1000 vezes).
4. Ordenam-se os valores de θ_k^* do menor ao maior. Determinam-se limites de confiança para uma especificada probabilidade α. Se forem 1000 iterações e $\alpha = 0.05$, o limite inferior será o valor de θ_k^* na 25ª posição e o limite superior aquele na 976ª posição. Na verdade, nesse caso, é somente necessário armazenar os 25 valores menores e os 25 valores maiores de θ_k^*.
5. Pode-se então afirmar, com uma probabilidade α de estar errado, que o valor verdadeiro do parâmetro θ avaliado encontra-se entre os limites de confiança.

A determinação de intervalos de confiança para uma série de amostras bootstrap de tamanho $n_k \leq n$ permite examinar a estabilidade da amplitude entre limites superior e inferior. A Tabela 1 ilustra com um pequeno exemplo a obtenção de limites de confiança. A Figura 2 mostra outro exemplo e a sua interpretação.

Suficiência amostral em análise de agrupamentos

A análise de agrupamentos aplicada em ecologia objetiva classificar unidades amostrais (ecossistemas, comunidades, ou indivíduos) permitindo simplificar em tipologias a variação complexa comum em sistemas naturais. Métodos de análise de agrupamentos são discutidos no capítulo ??. A Fig. 3 mostra um exemplo. Um problema sempre presente em análise de agrupamentos é a escolha do nível de partição, e essa decisão está relacionada a suficiência amostral como veremos mais adiante. Um dado nível de classificação (número de grupos) será considerado nítido se os tipos revelados aparecerem consistentemente quando o levantamento for repetido no mesmo universo amostral. A reamostragem do universo amostral pode ser simulada através de reamostragem bootstrap.

A avaliação de suficiência amostral através de reamostragem bootstrap em análise de agrupamentos é baseada no método usado para a determinação da significância de grupos em análise de agrupamentos descrito em Pillar (1999a). Dados multivariados podem ser representados em um espaço geométrico, abstrato, multidimensional; neste as variáveis são as suas dimensões e as unidades amostrais os pontos no espaço. Quanto mais nítida for a estrutura de grupos no espaço abstrato, os grupos revelados por análise de agrupamentos de amostras bootstrap serão mais estáveis; como conseqüência, suficiência amostral será atingida com um tamanho menor de amostra. O atributo medido em cada amostra bootstrap de tamanho k para um dado nível m de partição em grupos é

$$G_k^* = 1 - \frac{S}{T}$$

onde T é a soma de quadrados total, envolvendo $(n + n_k)(n + n_k - 1)/2$ dissimilaridades ao quadrado de $n + n_k$ unidades amostrais, sendo n unidades amostrais originalmente do pseudo universo amostral e n_k unidades amostrais da amostra bootstrap. S é a soma de quadrados de contrastes aos pares entre grupos na amostra bootstrap e o grupo mais próximo no pseudo universo amostral. A determinação de S envolve um processo iterativo de análise com o objetivo de encontrar pares exclusivos formados por grupos da amostra bootstrap com grupos do pseudo universo amostral de forma a minimizar o valor de S. Para maior detalhamento do método consultar Pillar (1999a, 1999c).

Diferentemente do método anterior, em que intervalos de confiança são deterrminados, aqui o valor de G_k^* é comparado a G_k^0 gerado a cada iteração de bootstrap sob a hipótese nula (Ho) de que os grupos são nítidos. Se Ho é verdadeira, cada grupo encontrado pela análise de agrupamentos nas amostras bootstrap será uma amostra aleatória do grupo correspondente (mais próximo) no pseudo universo amostral. A probabilidade $P(G_k^0 \le G_k^*)$ é a proporção de iteracões bootstrap em que $G_k^0 \le G_k^*$. A determinação de $P(G_k^0 \le G_k^*)$ para uma série de amostras bootstrap de tamanho $n_k \le n$ permite examinar a estabilidade de $P(G_k^0 \le G_k^*)$. Se para um dado tamanho de amostra n_k a probabilidade $P(G_k^0 \le G_k^*)$ não for maior do que um limiar de probabilidade α, digamos $\alpha = 0{,}05$, Ho será rejeitada e a classificação em m grupos será considerada difusa e pouco nítida, logo instável. Neste caso, de rejeição de Ho, a amostra de tamanho n_k é suficiente, pois tamanhos de amostra maiores do que n_k tendem a determinar probabilidades $P(G_k^0 \le G_k^*)$ consistentemente menores do que α. Ou seja, as conclusões a respeito da falta de estrutura nítida de grupos nos dados não se alteraram ao se aumentar o tamanho da amostra. Caso contrário, se $P(G_k^0 \le G_k^*) > \alpha$, Ho é aceita, e duas alternativas são possíveis: (1) se as probabilidades $P(G_k^0 \le G_k^*)$ são consistentemente maiores do que α e estáveis para tamanhos de amostra maiores do que n_k, a amostra é suficiente, e a classificação será considerada nítida; (2) se a magnitude de $P(G_k^0 \le G_k^*)$ é ainda instável ou decrescente para tamanhos de amostra maiores do que n_k, a amostra é considerada insuficiente, não sendo possível nenhuma conclusão a respeito da nitidez da estrutura de grupos. Casos típicos com dados artificiais estão na Fig. 4. Um exemplo com dados limnológicos é mostrado na Fig. 5.

Suficiência amostral em ordenação

Métodos de ordenação, discutidos no capítulo ??, permitem obter uma síntese da variação observada em um espaço geométrico, abstrato, multidimensional, no qual dados ecológicos podem ser representados. A síntese obtida pode ser visualizada em diagramas de

dispersão como na Fig. 6. Qual a probabilidade de que tendências de variação observadas através da ordenação de dados obtidos de um levantamento se mantenham ao se repetir o levantamento no mesmo universo amostral? A questão está vinculada à significância dos eixos de ordenação, mas somente poderá ser respondida se a amostra for suficiente. Há antecedentes na aplicação de reamostragem bootstrap na determinação de significância de eixos de ordenação (Stauffer et al. 1985, Knox & Peet 1989, Jackson 1993).

Em Pillar (1999b) descrevo método baseado em reamostragem bootstrap para avaliar a significância de eixos de ordenação. O procedimento inicia-se pela aplicação do método de ordenação ao pseudo universo amostral, armazenando-se os escores das unidades amostrais como escores de referência. A seguir, para cada tamanho k de amostra, o seguinte procedimento é seguido e repetido um grande número de vezes (iterações): É tomada uma amostra bootstrap de tamanho n_k a qual é submetida ao método de ordenação. Os escores de ordenação da amostra bootstrap para um dado número de eixos da ordenação são armazenados em uma matriz $\boldsymbol{X}_k$*****, e os escores das unidades amostrais que estão na amostra bootstrap, mas extraídos dos escores de referência, são armazenados em uma matriz $\boldsymbol{X}_k$. Um ajuste Procrusteano (Schönemann & Carroll 1970) envolvendo os primeiros i eixos da ordenação torna os escores das duas ordenações comparáveis; tal ajuste envolve rotação, translação e dilatação do subespaço de ordenação na amostra bootstrap, de tal forma a maximizar o ajuste com a ordenação do pseudo universo amostral. Os escores no eixo de ordenação i em $\boldsymbol{X}_k$***** e $\boldsymbol{X}_k$ são comparados pelo coeficiente de correlação

$$\theta^*_{ki} = r(\boldsymbol{x}_{ki}^*, \boldsymbol{x}_{ki})$$

Quanto mais alta a correlação, melhor é a concordância entre os escores bootstrap e de referência, e mais estável são as tendências de variação observadas na ordenação da amostra de tamanho k. A cada iteração a correlação θ^*_{ki} é comparada a uma correlacão θ^0_{ki} gerada sob a hipótese nula de que os dados não têm estrutura. Ou seja, as matrizes $\boldsymbol{X}_k$***** e $\boldsymbol{X}_k$ são agora obtidas através de reamostragem bootstrap dos dados observados com as observações permutadas aleatoriamente dentro de variáveis. Se $\theta^0_{ki} \geq \theta^*_{ki}$, o algoritmo soma 1 à freqüência acumulada $F(\theta^0_{ki} \geq \theta^*_{ki})$. Após B iterações bootstrap, a probabilidade $P(\theta^0_{ki} \geq \theta^*_{ki})$ é a proporção $F(\theta^0_{ki} \geq \theta^*_{ki})/B$. Mais detalhes do método poderão ser encontrados em Pillar (1999b).

CONSIDERAÇÕES FINAIS

Apresentei neste capítulo métodos recentes de avaliação de suficiência amostral baseados em reamostragem bootstrap, computacionalmente intensivos, mas que superam limitações impostas pelos métodos oferecidos pela teoria amostral clássica. A limitação destes últimos é evidente em ecologia quando o objetivo da amostragem é freqüentemente o reconhecimento de padrões e sua interpretação. O problema computacional, presente há poucos anos atrás, está superado com a generalização de microcomputadores com processadores cada vez mais rápidos, combinados com o uso de algoritmos eficientes. Resultados com o programa SAMPLER (Pillar 1999d) podem ser obtidos em um microcomputador em questão de segundos ou poucos minutos, dependendo do tamanho da amostra. Apesar de fortes argumentos a favor da utilização desses novos métodos computacionalmente intensivos, o seu conhecimento e uso ainda não é generalizado, havendo uma evidente inércia manifestada nos livros textos básicos e softwares de estatística.

Os exemplos usando dados de levantamentos limnológicos mostraram que, com os mesmos dados, a suficiência amostral pode ser indicada com diferentes tamanhos de amostra dependendo dos objetivos. Um dado tamanho de amostra pode ser suficiente, por exemplo,

para interpretar os primeiros eixos de ordenação mas não para revelar grupos com um certo nível de partição. Quando o objetivo da análise é o reconhecimento de padrões e sua interpretação, deve-se distinguir claramente suficiência amostral de significância de partições ou de eixos de ordenação. O tamanho da amostra pode ser suficiente para avaliar a significância de um dado eixo de ordenação, mas o teste poderá indicar que tal eixo de ordenação é não-significativo por apresentar padrões inconsistentes na reamostragem. Por outro lado, uma amostra pode ser suficiente para avaliar nitidez de estrutura de grupos a um dado nível de partição, mas a estrutura de grupos pode não ser necessariamente nítida.

AGRADECIMENTOS

O autor agradece a Ronaldo Padilha por ter gentilmente cedido seus dados para serem utilizados em exemplos neste trabalho, e a Albano Schwarzbold por sugestões no texto.

BIBLIOGRAFIA

Braun-Blanquet, J. 1979. Fitosociologia; bases para el estudio de las comunidades vegetales. Madrid: Blume. 819p.

Camiz, S. & Gergely, A. 1990. An exploratory method for determining optimal plot size in plant community studies. Abstracta Botanica 14: 83-108.

Cochran, W.G. 1977. Sampling Techniques, 3 ed. New York: Wiley. 428p.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7: 1-25.

Efron, B. & Tibshirani, R. 1993. An Introduction to the Bootstrap. London: Chapman & Hall. 436p.

Goedickemeier, I., Wildi, O. & Kienast, F. 1997. Sampling for vegetation survey: Some properties of a GIS-based stratification compared to other statistical sampling methods. Coenoses 12: 43-50.

Green, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. New York: Wiley. 257p.

Greig-Smith, P. 1983. Quantitative Plant Ecology 3rd ed. Oxford: Blackwell.

Jackson, D.A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74: 2204-2214.

Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R. (eds.). 1995. Data Analysis in Community and Landscape Ecology. Cambridge: Cambridge University Press. 299p.

Juhász-Nagy, P. & Podani, J. 1983. Information theory methods for the study of spatial processes and succession. Vegetatio 51: 129-140.

Kenkel, N.C., Juhász-Nagy, P. & Podani, J. 1989. On sampling procedures in population and community ecology. Vegetatio 83: 195-207.

Knox, R.G., & Peet, R.K. 1989**.** Bootstrapped ordination: a method for estimating sampling effects in indirect gradient analysis. Vegetatio 80: 153-165.

Krishnaiah, P.R. & Rao, C.R. (eds.). 1988. Sampling. Amsterdam: North-Holland. 594p.

Legendre, L. & Legendre, P. 1998. Numerical Ecology 2nd ed. New York: Elsevier. 853p.

Orlóci, L. 1978. Multivariate Analysis in Vegetation Research. The Hague: Junk. 450p.

Orlóci, L. 1993. The complexities and scenarios of ecosystem analysis. *In*: Patil, G.P. & Rao, C.R. (eds.) Multivariate Environmental Statistics. Amsterdam: North-Holland. p.423-432.

Orlóci, L. & Pillar, V.D. 1989. On sample size optimality in ecosystem survey. Biométrie-Praximetrie 29: 173-184.

Padilha, R.S. 1997. Limnologia de pequenas lagoas e arroios da Reserva Ecológica do Morro Santana, Porto Alegre, Rio Grande do Sul. Dissertação de Bacharelado. Porto Alegre: Universidade Federal do Rio Grande do Sul, 128p.

Palmer, M.W. 1988. Fractal geometry: a tool for describing spatial patterns of plant communities. Vegetatio 75: 91-102.

Palmer, M.W. & White, P.S. 1994. On the existence of ecological communities. Journal of Vegetation Science 5: 279-282.

Patil, G.P., Babu, G.J., Hennemuth, R.C., Myers, W.L., Rajarshi, M.B. & Taillie, C. 1988. Data-based sampling and model-based estimation for environmental resources. *In*: Krishnaiah, P.R. & Rao, C.R. (eds.). Sampling. Amsterdam: North-Holland. p. 489-513.

Pielou, E.C. 1984. The interpretation of Ecological Data. A primer on Classification and Ordination. New York: Wiley-Interscience. 263p.

Pillar, V.D. 1998. Sampling sufficiency in ecological surveys. Abstracta Botanica 22: 37-48.

Pillar, V.D. 1999a. How sharp are classifications? Ecology 80: 2508-2516.

Pillar, V.D. 1999b. The bootstrapped ordination reexamined. Journal of Vegetation Science 10(6): ???-??? (no prelo).

Pillar, V.D. 1999c. Software for testing classification sharpness combined with sampling sufficiency evaluation. Ecological Archives E080-014-S1.

Pillar, V.D. 1999d. SAMPLER software for bootstrap resampling and evaluation of sampling sufficiency. Porto Alegre: Departamento de Ecologia, UFRGS.

Podani, J. 1994. Multivariate data analysis in ecology and systematics. The Hague: SPB. 316p.

Podani, J., Czárán, T. & Bartha, S. 1993. Pattern, area and diversity: the importance of spatial scale in species assemblages. Abstracta Botanica 17: 37-51.

Schönemann, P.H., & Carroll, R.M. 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika 35: 245-256.

Stauffer, D.F., Garton, E.O. & Steinhorst, R.K. 1985**.** A comparison of principal components from real and random data. Ecology 66: 1693-1698.

Tabela 1. Determinação de intervalos de confiança através de reamostragem bootstrap ilustrada através de um exemplo numérico. A amostra contém 11 unidades amostrais, descritas por uma variável apenas, cujas observações são as seguintes: 41, 29, 3, 42, 42, 42, 16, 11, 6, 42, 42. Neste exemplo, intervalos de confiança de 80% foram determinados para amostras com 3, 5, 7, 9 e 11 unidades amostrais, através de 10 iterações bootstrap. Recomenda-se que em situações reais o número de iterações seja pelo menos 1000. A cada iteração médias foram computadas com as unidades amostrais tomadas na ordem indicada. Por exemplo, na primeira iteração a média de uma amostra bootstrap com 3 unidades amostrais foi (11+42+41)/3 = 31,33, de uma amostra com 5 unidades amostrais foi (11+42+41+42+42)/5 = 35,6 e assim sucessivamente. Tendo arranjado em ordem crescente as médias para cada tamanho de amostra, os limites inferior e superior foram respectivamente os valores nas posições w e B-w+1, sendo w = Bα/2 = 10(1-0,8)/2 = 1, B o número de iterações bootstrap e α a probabilidade especificada para a zona de exclusão do intervalo de confiança. Neste exemplo os limites coincidem com os valores mínimo e máximo.

a) Resultados intermediários em 10 iterações de reamostragem bootstrap:

	Amostras bootstrap											Tamanho da amostra bootstrap 3	5	7	9	11
1	11	42	41	42	42	42	42	29	3	11	6	31,33	35,6	37,43	32,67	28,27
2	41	16	42	41	41	6	29	42	11	42	6	33	36,2	30,86	29,89	28,82
3	42	3	41	41	42	42	42	11	41	42	42	28,67	33,8	36,14	33,89	35,36
4	16	42	3	29	29	42	42	42	42	11	3	20,33	23,8	29	31,89	27,36
5	6	6	29	42	6	42	42	41	16	42	42	13,67	17,8	24,71	25,56	28,55
6	41	42	16	29	3	6	6	42	11	41	3	33	26,2	20,43	21,78	21,82
7	16	29	42	42	3	29	6	3	42	16	42	29	26,4	23,86	23,56	24,55
8	6	6	41	16	42	11	11	29	29	42	6	17,67	22,2	19	21,22	21,73
9	16	16	42	29	3	42	42	16	42	42	42	24,67	21,2	27,14	27,56	30,18
10	42	6	29	42	42	42	11	41	42	11	42	25,67	32,2	30,57	33	31,82

b) Intervalos de confiança (80%):

	Tamanho de amostra 3	5	7	9	11
Limite inferior	13,67	17,8	19	21,22	21,73
Limite superior	33	36,2	37,43	33,89	35,36
Média das médias geradas nas 10 iterações	25,7	27,54	27,91	28,1	27,85

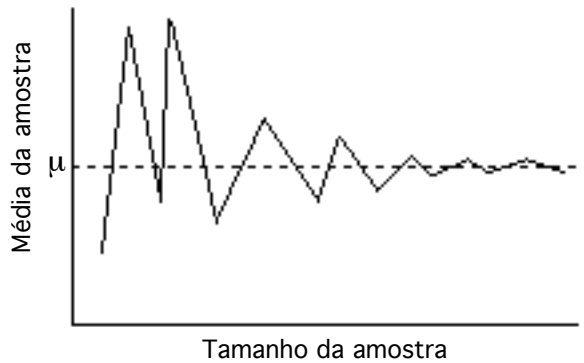

Figura 1. A estimativa do atributo será mais acurada quanto mais próximo do estado verdadeiro do universo amostral for o estado inferido via amostragem.

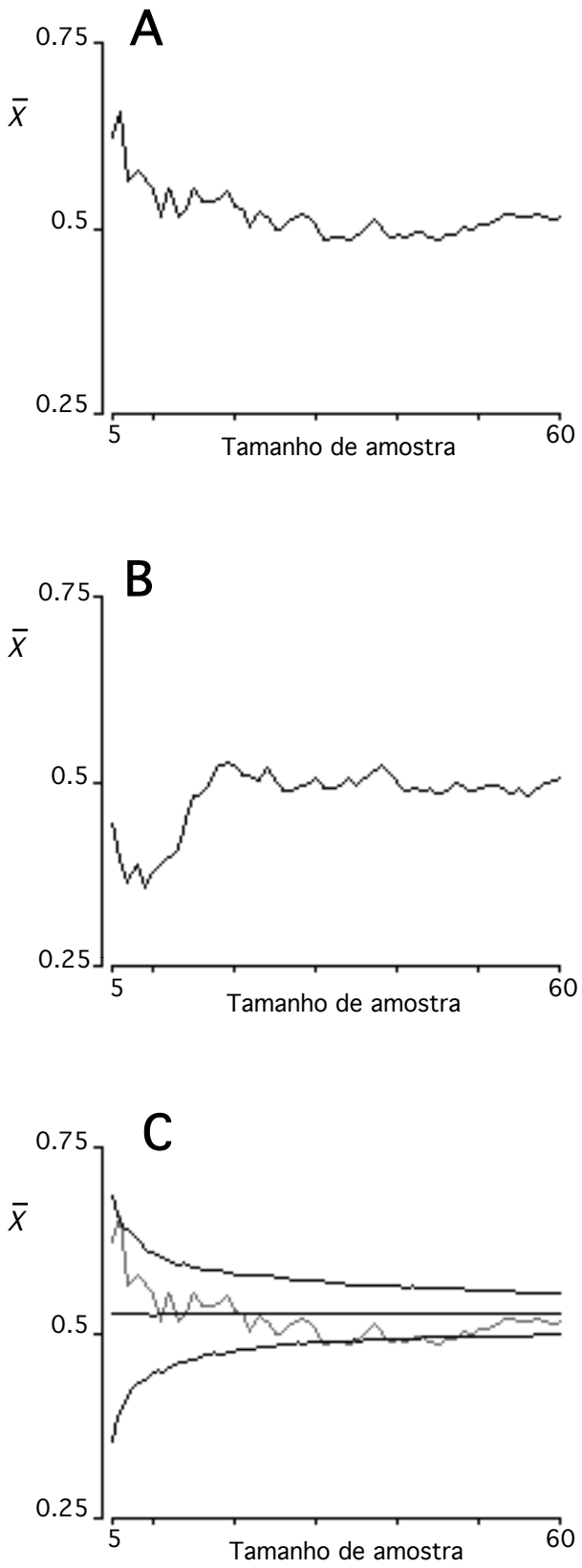

Figura 2. Valores médios de uma variável *X* obtidos por reamostragem com reposição de um conjunto de dados com 60 unidades amostrais. Os tamanhos de amostra variam de 5 a 60. Duas das muitas seqüências de possíveis médias das amostras são mostradas em A-B. Em C 90% intervalos de confiança foram definidos com base em 1000 iterações de reamostragem para cada tamanho de amostra. Para ilustrar, são mostrados com os limites em C o caso em A

e a média das 1000 médias a cada tamanho de amostra (quase uma linha reta e num valor idêntico à média da variável *X*). Para estimar a média da variável *X*, usando, e.g., uma amostra com 5 unidades amostrais a média esperada estará em 90% dos casos 0,36 e 0,67, ou seja, médias com uma diferença de até 0,31 podem não ser significativamente diferentes (assumindo que as populações têm as mesmas distribuições de freqüências). Diferenças bem menores podem ser detectadas com 30 unidades amostrais; a média estará entre 0,42 e 0,60. Há uma vantagem muito pequena em tomar 60 unidades amostrais; o intervalo de confiança estará entre 0,45 and 0,58. Adaptado de Pillar (1998).

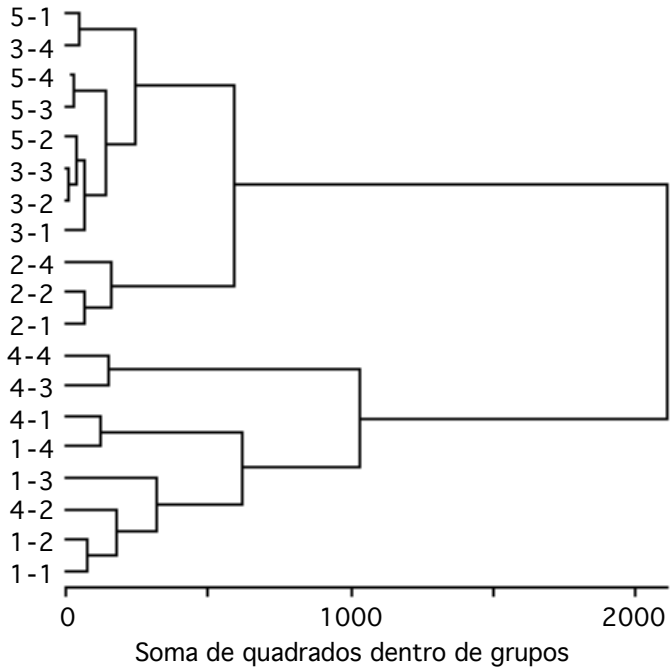

Figura 3. Dendrograma de análise de agrupamentos, obtida pelo método da variância mínima, com dados contendo 19 unidades amostrais descritas pela composição de algas (109 espécies). Dados de Padilha (1997). A análise utilizou distâncias euclidianas calculadas com os dados transformados por log (x+1). A análise de agrupamentos oferece várias possibilidades de classificação (partição em grupos).

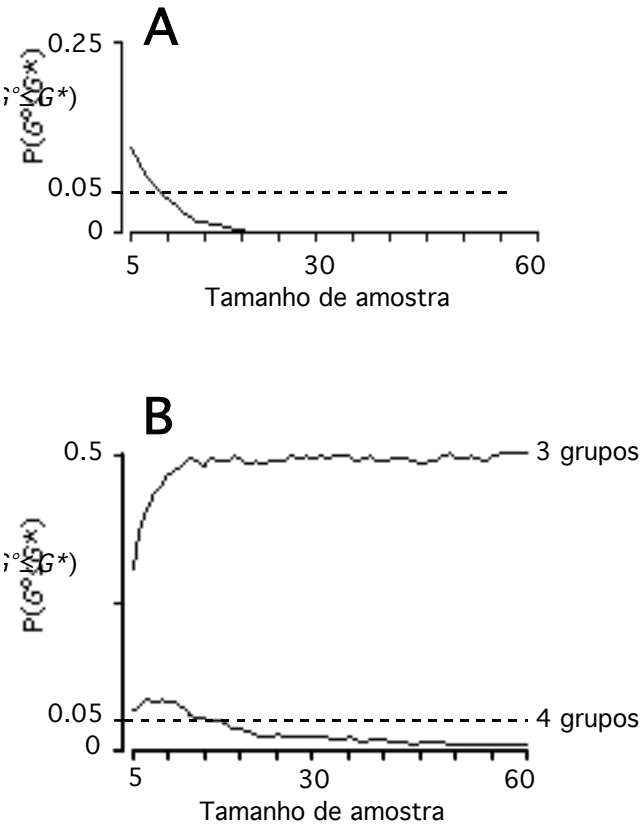

Figura 4. Avaliação de suficiência amostral e significância de níveis de partição em grupos através de probabilidades P($G_k^0 \leq G_k^*$) em diferentes dados. Probabilidades geradas em 10000 iterações de reamostragem bootstrap a cada tamanho de amostra. Dados e nível de partição são os seguintes: (A) Dados artificiais com 60 unidades amostrais descritas por 60 variáveis geradas por números aleatórios (não há grupos nítidos), nível de partição 2; (B) Dados artificiais gerados com 3 grupos bem nítidos, níveis de partição em 3 e 4 grupos. A análise de agrupamentos é pelo método de variância mínima. Adaptado de Pillar (1998). Os dados em A, sem nenhuma estrutura de grupos, foram corretamente identificados como tal (adotando um limiar α = 0,05) em amostras com 6 ou mais unidades amostrais. Os dados em B, gerados com uma estrutura nítida de 3 grupos, foram corretamente identificados como tal em amostras com 8 ou mais unidades amostrais (para um limiar α = 0,05), sendo que amostras menores indicariam 4 grupos nítidos. Adotando-se um limiar α = 0,1, conclusões corretas seriam obtidas para amostras com 5 ou mais unidades amostrais.

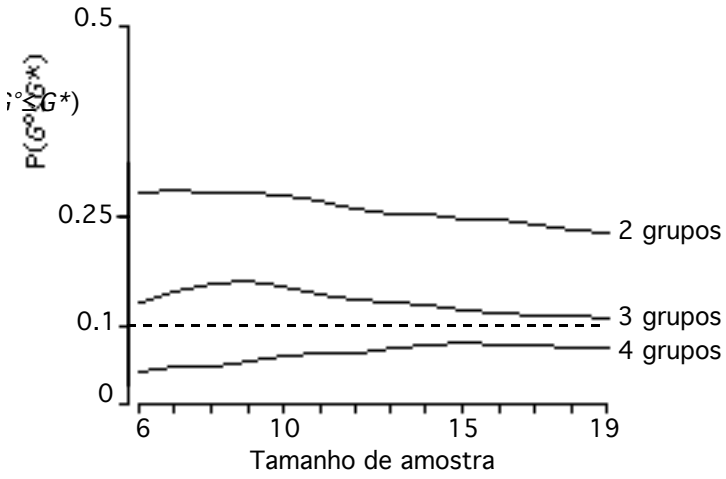

Figura 5. Avaliação de suficiência amostral e significância de níveis de partição em grupos através de probabilidades P($G_k^0 \leq G_k^*$) geradas por reamostragem bootstrap (Pillar 1999a). Os dados (Padilha 1997) foram obtidos em 19 unidades amostrais descritas pela composição de algas (109 espécies). Probabilidades foram geradas em 10000 iterações de reamostragem a cada tamanho de amostra. Os grupos foram obtidos por análise de agrupamentos pelo método de variância mínima; a Fig. 3 mostra dendrograma obtido com as 19 unidades amostrais. Considerando um limiar α = 0,1, o teste indica que partições em 2 e 3 grupos são nítidas, enquanto partições em 4 grupos ou mais (estes não mostrados) são difusas. Porém, as curvas para 2 ou 3 grupos ainda são levemente decrescentes até 19 unidades amostrais, indicando que a amostra com 19 unidades amostrais é insuficiente para conclusões definitivas a respeito da nitidez dos grupos nesses níveis de partição. A curva para 4 grupos também é levemente decrescente até 19 unidades amostrais, mas a conclusão de que os grupos são difusos não se alterará se a curva continuar decrescendo com tamanhos maiores de amostra.

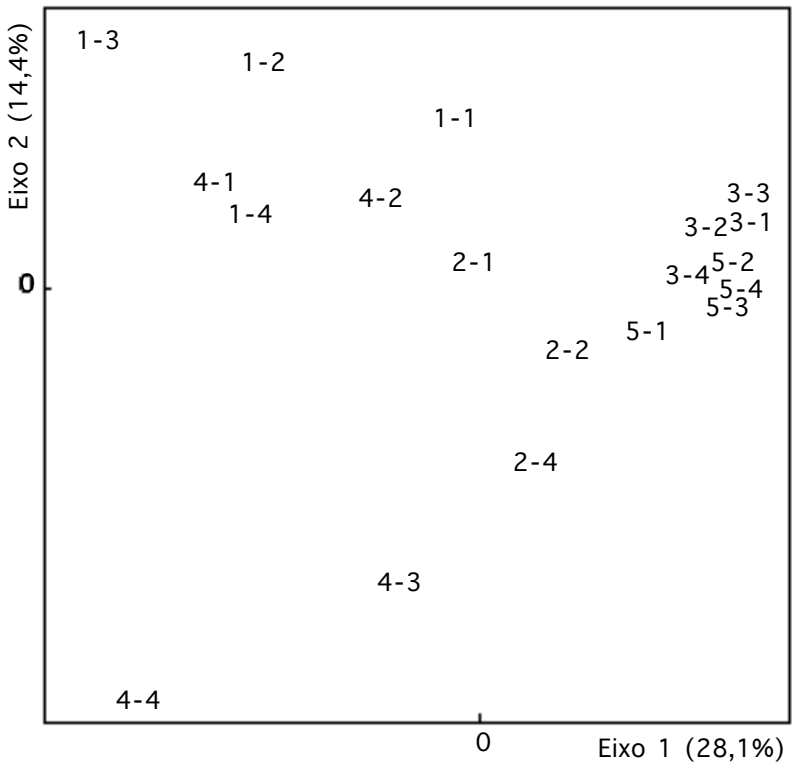

Figura 6. Diagrama de dispersão obtido por ordenação de 19 unidades amostrais. Os dados (Padilha 1997) foram obtidos em 19 unidades amostrais descritas pela composição de algas (109 espécies). Os pontos são as unidades amostrais; os eixos foram obtidos por análise de coordenadas principais a partir de distâncias euclidianas calculadas com os dados transformados por log (x+1). Os dois eixos contêm 28,1 + 14,4 = 42,5% da variância total. Os taxons cuja variação está mais correlacionada com o eixo 1 são as seguintes: *Sphaerocystis* sp. (r = -0.89), *Cymbella* sp. (r = -0.88), *Hyaloraphidium* sp. (r = -0.85), *Rhizosolenia* sp. (r = -0.85), *Micrasterias* sp. (r = -0.85) e *Stenopterobia* sp. (r = -0.82). Com o eixo 2 estão mais correlacionados os seguintes taxons: *Scenedesmus* sp. (r = -0.80), *Eunotia* sp. (r= -0.79), *Radiococcus* sp. (r = -0.79) e *Kirchneriella* sp. (r = -0.78). Qual a probabilidade de que essas tendências de variação observadas no diagrama se mantenham ao se repetir o levantamento no mesmo universo amostral?

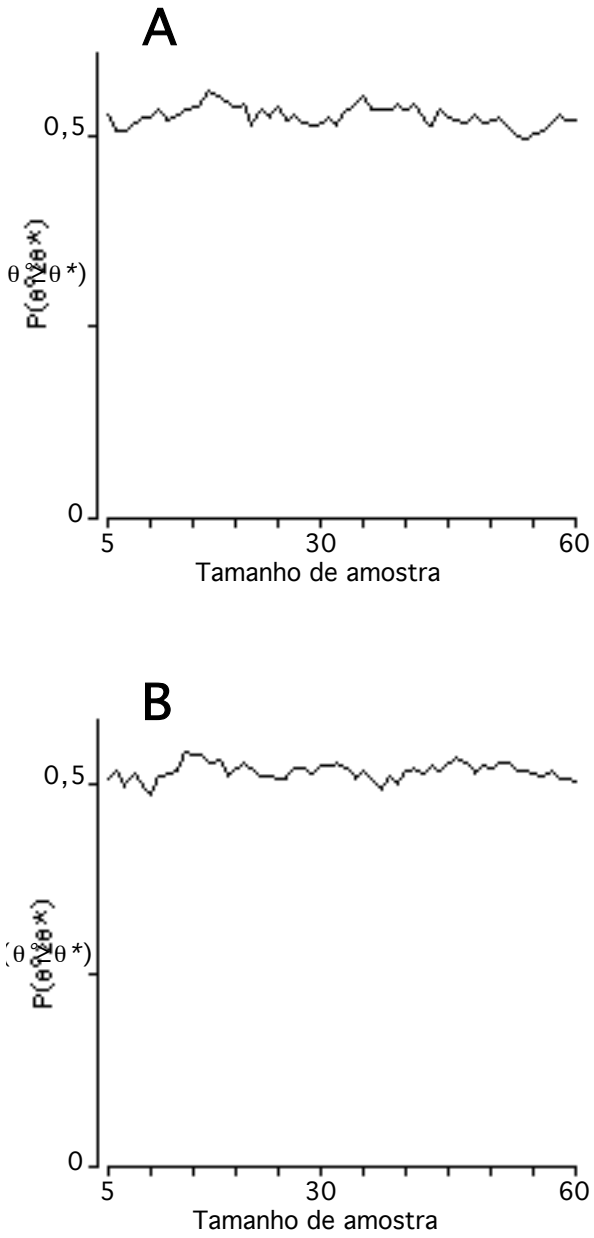

Figura 7. Avaliação de suficiência amostral e significância de eixos de ordenação em um conjunto de dados artificiais com 60 unidades amostrais e 60 variáveis geradas por números aleatórios. O método de ordenação é por análise de coordenadas principais. Em A é avaliado o eixo 1 da ordenação, e em B o eixo 2. As probabilidades $P(\theta_{ki}^{0} \geq \theta_{ki}^{*})$ foram geradas em 1000 iterações de reamostragem bootstrap (Pillar 1999b). Probabilidades próximas de 0,5 indicam que os eixos de ordenação, como esperado, não representam tendências consistentes de variação. A estabilidade das curvas indica que uma amostra com 5 ou mais unidades amostrais seria suficiente neste caso.

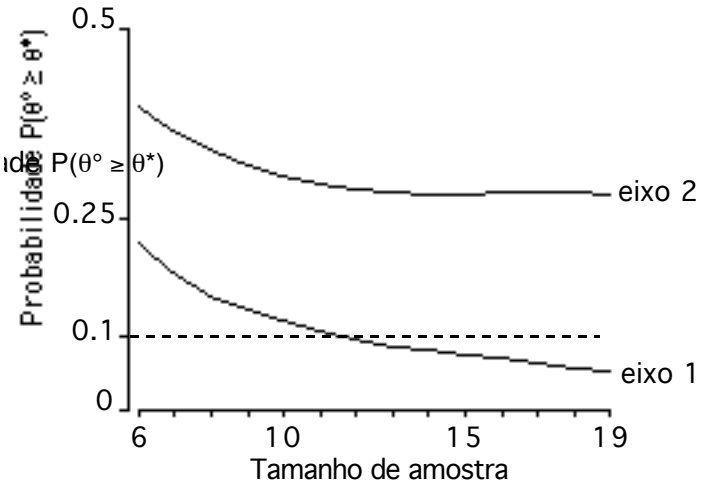

Figura 8. Efeito do tamanho da amostra na significância de eixos de ordenação, obtidos por análise de coordenadas principais. Os dados, que são os mesmos de exemplos anteriores (Padilha 1997), foram obtidos em 19 unidades amostrais descritas pela composição de algas (109 espécies). O método envolve reamostragem bootstrap e ordenação com tamanhos crescentes de amostra (Pillar 1999b). Para um limiar $\alpha = 0,1$, o teste indica que amostras com 13 ou mais unidades amostrais são suficientes para interpretar como consistentes as tendências de variação reveladas sobre o eixo 1 da ordenação. O teste indica que o eixo 2 da ordenação não é significativo; conseqüentemente interpretações deste eixo quanto a correlação com taxons ou variáveis físicas e químicas serão provavelmente inconsistentes se o levantamento for repetido. A curva para o eixo 2 estabiliza com amostras de 13 ou mais unidades amostrais, indicando que a amostra é suficiente para uma conclusão definitiva a respeito desse eixo.