

Biologia Geral e Experimental

Universidade Federal de Sergipe

São Cristóvão, SE 3 (2): 47-53

26.iii.2003

CONSIDERAÇÕES SOBRE OS MÉTODOS QUANTITATIVOS NA PESQUISA BIOLÓGICA

Jeane Carvalho Vilar¹
Celso Morato de Carvalho¹

RESUMO

Discute-se brevemente alguns aspectos sobre métodos quantitativos na pesquisa biológica.

Palavras-chave: métodos quantitativos, pesquisa biológica.

ABSTRACT

Theoretical aspects of the quantitative methods in the biological research are briefly discussed.

Key words: quantitative methods, biological research.

INTRODUÇÃO

Estas notas são dirigidas aos estudantes de graduação da área biológica que freqüentam o nosso curso optativo sobre métodos quantitativos na pesquisa biológica, oferecido semestralmente no Departamento de Biologia da Universidade Federal de Sergipe. Nosso objetivo neste exercício é tentar, resumidamente, discutir a importância de alguns procedimentos essenciais na elaboração de uma pesquisa biológica.

A prática da observação e a execução de pesquisas biológicas exigem o conhecimento de algumas regras básicas sobre a área temática de estudo e sobre métodos de análise. A elaboração de questões e o que observar dependem muito mais de leitura crítica, busca incessante de informações e vivência

científica do que qualquer outro fator.

Quando se deseja relacionar um problema biológico específico aos métodos de análise, necessariamente se interpõe entre estes dois fatores a coleta de dados, que pode ser feita de duas maneiras, dependendo da questão: i) *mensurativa*, onde os dados são coletados em natureza, sem a manipulação nos tratamentos, como em muitos estudos ecológicos, botânicos e zoológicos, ou ii) *manipulativa*, na qual diferentes unidades experimentais recebem tratamentos diferenciados induzidos, como ocorre na farmacologia, na fisiologia ou na biofísica (Hurlbert, 1984).

A coleta de dados, essência da observação e da verificação de hipóteses, apresenta algumas fases importantes, até a análise final e a interpretação biológica das variações. Estas fases são (Fearnside,

¹ Departamento de Biologia, Universidade Federal de Sergipe. Av. Marechal Rondon s/n; Jardim Rosa Elze, São Cristóvão, SE. 49100-000.

1986; Hayek, 1994): i) identificação do problema biológico; ii) formulação das questões, elaboradas sem ambigüidades; iii) elaboração de um sistema de proposições entre variáveis, as quais consistem das hipóteses de estudo (hipóteses nula e alternativa): num sistema de proposições falso/verdadeiro, a hipótese nula (H_0) é aquela a ser testada, a hipótese alternativa (H_1) é “aceita” quando a nula for rejeitada; iv) planejamento experimental; v) obtenção de dados e verificação da hipótese - análise das variações dos dados obtidos e vi) interpretação dos resultados.

Questões

É impossível executar uma pesquisa ou avaliar a literatura científica sem um conhecimento básico probabilístico e estatístico sobre as inter-relações entre os parâmetros biológicos da área de interesse e as variações dos fenômenos observados. Essencial para esta compreensão é a elaboração de questões claras, as quais fazem indagações sobre as relações entre duas ou mais variáveis. As questões devem ser precisas, de modo a permitir a elaboração de hipóteses verificáveis. Questões generalistas ou metafísicas devem ser evitadas. Por exemplo “será que a qualidade dos fatores climáticos interferem na reprodução das serpentes?” não fornece elementos para verificações. Uma pergunta mais apropriada para verificação poderia ser “o potencial reprodutivo de uma população de *Leptodeira annulata* (Serpentes) da região da caatinga do oeste de Sergipe aumenta com o aumento das chuvas?”.

Neste contexto, é importante a definição das variáveis utilizadas, para delimitar a natureza destas e dos procedimentos. Imprecisões nas definições levam a ambigüidades e impossibilidade de verificações, como aquelas nas quais “cabem tudo”, como um leito de Procusto e com base apenas nas representações sociais. Por exemplo, a afirmação “a diversidade de serpentes da região de Sergipe parece ser relativamente boa e as espécies tendem a ter razoável número de indivíduos” não diz nada. Afirmações deste tipo não definem quantitativamente as variáveis, trazem

informações imprecisas e nenhuma verificação pode ser feita. Para uma discussão mais abrangente sobre as definições operacionais e definição das variáveis sugerimos o trabalho de Hayek (1994).

Variáveis

Qualquer atributo ou estado biológico que possa diferir entre os elementos de uma população pode ser chamado de *variável* (Zar, 1994; Vanzolini, 1993). Quatro tipos de variáveis biológicas podem ser reconhecidas (Stevens, 1946):

- i) Variáveis de escala geométrica: variáveis quantitativas com ponto zero verdadeiro e que admitem todas as operações aritméticas. Podem ser *discretas*, com valores positivos inteiros, como o número de escamas ventrais de serpentes, ou *contínuas*, as quais podem tomar quaisquer valores positivos.
- ii) Variáveis de escala intervalar: variáveis contínuas com ponto zero e unidades arbitrárias, por exemplo as diversas escalas de temperatura.
- iii) Variáveis de escala ordinal: estas variáveis podem ser arranjadas por postos, numa ordenação crescente ou decrescente, como as variações de cores e formas.
- iv) variáveis de escala categórica (ou nominais): permitem a classificação sem ambigüidade e em classes mutuamente exclusivas, como o sexo.

Populações e Amostras

O objetivo da análise quantitativa na pesquisa biológica é fazer inferências de atributos presentes em um grande grupo, através da análise destas características em um pequeno grupo de dados retiradas ao acaso do grupo maior. Estas inferências são expressas em probabilidades.

O total de medidas sobre as quais se quer tirar conclusões é a população ou universo. Por exemplo, podemos estar interessados em tirar conclusões sobre o comprimento do corpo de lagartos na idade da primeira reprodução, em uma determinada espécie que ocorre na Serra de Itabaiana, Sergipe. Todos os

comprimentos de corpo destes lagartos, com as mesmas características reprodutivas, formam a população em questão.

Populações podem ser muito grandes, como o exemplo dos lagartos, mas às vezes podem ser pequenas, como a população constituída por todos os mocós (*Kerodon rupestris*) fêmeas com mais de cinco anos de idade, da região de Canindé do São Francisco. Se a população a ser estudada for pequena, praticamente todas as medidas do conjunto poderão ser obtidas. Na realidade, isso é difícil de ocorrer, devido a vários fatores limitantes, como tempo e dinheiro, mas podemos obter as medidas de parte da população, como um subconjunto. Este subconjunto constitui a amostra. Através das características da amostra podemos obter as características da população de onde esta amostra foi retirada (Zar, 1996).

Os modos de obtermos amostras biológicas podem variar, dependendo do tipo de questão e da área temática da pesquisa. Independente destes fatores, é essencial a aleatoriedade com que as amostras são obtidas. Isto requer que todos os elementos da população tenham a mesma chance de serem selecionados na amostra, i.e., a seleção de um elemento na amostra não deve influenciar a chance de outro ser selecionado (amostras independentes).

Duas perguntas frequentes que nos fazem são relacionadas ao tamanho da amostra e se um aumento no tamanho destas leva a melhores resultados. Muitos fatores interferem na obtenção de amostras, cuja essência é determinar o tamanho mínimo necessário para as verificações estatísticas. Isso vai depender diretamente da questão biológica formulada e do tipo de teste a ser utilizado, planejado antes da pesquisa ser realizada. Por exemplo, um teste muito utilizado na pesquisa biológica é o teste de t (Student). A distribuição de t é caracterizada pelo quociente da diferença entre a média amostral e a média da população (verdadeira) pelo desvio padrão da média amostral. Se aumentarmos os graus de liberdade a distribuição converge para uma normal, com N infinito, cujo valor é $t = 1,96$. Acima de 15 graus de

liberdade e um nível de significância de 5%, podemos aumentar o quanto quisermos o tamanho da amostra que não aumentará o poder do teste em discriminar possíveis diferenças (ver Vanzolini, 1993).

Distribuição normal

A distribuição normal é uma curva em forma de sino (curva de Gauss), caracterizada por ter valores preponderantes em torno da média, os quais vão decrescendo para os extremos da curva (Figura 1). É a distribuição mais importante, porque é a base dos testes de hipóteses pré-estabelecidos.

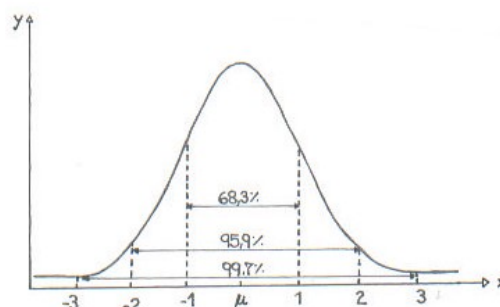


Figura 1. Distribuição normal.

68,3% das observações estão contidas entre a média ± 1 desvio padrão,
95,9% das observações estão contidas entre a média ± 2 desvios padrão,
99,7% das observações estão contidas entre a média ± 3 desvios padrão.

Na distribuição normal, para cada valor de x corresponde um valor de y , expresso pela função:

$$y = \frac{1}{\delta \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}}$$

Os parâmetros desta equação são o desvio padrão (δ) e a média (μ), de modo que para um determinado desvio padrão e médias diferentes, poderão existir infinitas curvas – neste caso a média determina a distribuição, quanto maior, mais afastada de y . Para uma determinada média e desvios padrões diferentes poderão também existir infinitas curvas – neste caso o desvio padrão determina a distribuição, quanto maior o desvio, mais dispersos serão os pontos ao redor da média.

Cada amostra retirada de uma população

normalmente distribuída terá uma média. Se retirarmos várias amostras aleatórias desta população, iremos obter várias médias, as quais terão uma distribuição normal, cuja variância irá diminuir à medida que o número de amostras aumentar. O desvio padrão de uma distribuição amostral chama-se *erro padrão*, depende da variância e dos graus de liberdade e tem muita importância nos testes de hipóteses.

Formulação de Hipóteses

Como trabalhamos com amostras na pesquisa biológica, a partir destas fazemos inferências sobre a média da população e as comparamos entre populações. Desse modo podemos verificar se as médias das amostras entre populações são significativamente diferentes entre si, se os dados (repetições) de vários procedimentos experimentais diferem entre si, dentro e entre amostras, ou se a média de uma amostra está contida no intervalo de confiança da média de uma população conhecida.

Fazemos estas inferências *antes* da coleta de dados, porque fazê-las depois dos dados coletados pode invalidar a pesquisa. Elaborada a questão e definida suas variáveis, formulamos duas hipóteses. A primeira, que vai “orientar” a pesquisa, chamamos de *hipótese nula* (H_0), na qual estabelecemos que as variáveis não têm efeito, que as amostras são provenientes da mesma população, com médias contidas no intervalo de confiança de 95% ou que as relações biológicas investigadas não estão associadas ou são homogêneas. Se concluirmos que a hipótese nula deva ser rejeitada, a *alternativa* (H_1) é então considerada verdadeira.

Erros

Para rejeitarmos ou não a hipótese nula precisamos de um critério objetivo, que determine a probabilidade de rejeição (*valor do p*). Por convenção adotamos a probabilidade de 5% (0,05) ou menos como *critério* para rejeitarmos H_0 . Esta probabilidade é chamada de *nível de significância*, representada pela letra grega α (alfa). Isto quer dizer que uma observação tem 95% de chance de estar nas áreas sob a curva normal compreendidas entre os parâmetros que estabelecemos na hipótese nula (ver Figura 1). Existe uma chance de 5% dos valores excederem aqueles especificados na H_0 , correspondentes aos extremos do nível de significância; na curva normal são conhecidas como *regiões de rejeição*, que podem estar situadas entre os menores (esquerda) ou maiores valores (direita), dependendo da questão estabelecida na hipótese nula (Figura 2).

Um teste estatístico tem um valor que corresponde ao critério (nível de significância) para rejeitarmos a hipótese nula. Este valor é chamado de *valor crítico*, que vai indicar as áreas de rejeição da hipótese nula.

Uma pergunta freqüente que também nos fazemos é sobre a probabilidade de rejeitarmos erroneamente a hipótese nula. Esta probabilidade é dada pelo nível de significância ($\alpha = 0,05$ ou menos). Isto significa que há uma probabilidade de 5% em rejeitarmos a hipótese nula e ela ser verdadeira. As tábuas estatísticas dão estas probabilidades associadas ao valor encontrado nos testes de verificação de hipóteses. Rejeitar erroneamente a hipótese nula é

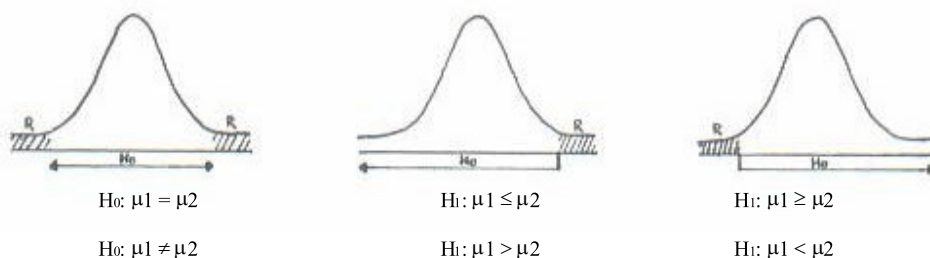


Figura 2. Região de rejeição (R).

conhecido como erro tipo I ou erro alfa. Existe associado outro tipo de erro em estatística, conhecido como erro tipo II ou erro β (letra grega beta), que consiste em não rejeitar a hipótese nula (H_0 não é “aceita”) e ela ser falsa. Este tipo de erro não podemos estabelecer, como fazemos com o nível de significância (erro alfa). Qual tipo de erro é mais grave irá depender do tipo de pesquisa. Se a pesquisa estabelecer que cometer um erro tipo I for mais grave do que cometer o erro tipo II, podemos então determinar o nível de significância em 1% ($\alpha = 0,01$). Como dissemos, estes erros têm de ser estabelecidos antes de ser executada a pesquisa, quando ainda estiver na fase de planejamento.

Associado ao erro de não rejeitarmos a hipótese nula e ela ser verdadeira, temos o poder do teste estatístico, determinado pela relação $1 - \text{erro } \beta$. Poder do teste é aquele que apresenta alta probabilidade de que a hipótese nula seja rejeitada quando esta é de fato falsa.

Testes Estatísticos

A escolha do teste a ser utilizado para verificação da hipótese nula (rejeição ou não) deve ser determinada *antes* da pesquisa ser executada. Existem vários programas de computador que facilitam os cálculos, muitas vezes enfadonhos de serem feitos a mão. Entretanto, nós sugerimos sempre aos alunos que não utilizem os programas antes de saberem alguns passos intermediários, forçosamente feitos “no braço”, para que o comportamento dos dados e a conclusão do teste possam ser corretamente interpretados.

Quem vai determinar o tipo de teste é a variável, amarrada à questão. Nos testes univariados, com variáveis de escala geométrica, podem ser aplicados testes paramétricos, com uma amostra, comparando a média amostral com a média de uma população conhecida, com áreas de rejeição situadas em ambas as caudas ou somente em uma, à esquerda ou à direita da curva normal, dependendo da questão formulada. Um teste paramétrico pode estimar os

parâmetros da população e condiciona que a distribuição seja normal, comparando médias ou variâncias. No caso de duas amostras, pareadas ou não, comparam-se as médias; se ambas não forem significativamente diferentes, não se rejeita a hipótese nula e optamos por considerá-las como tendo sido retiradas da mesma população. Em ambos os casos, uma e duas amostras, utilizamos a distribuição de t (Student) para verificar a significância das diferenças entre as médias (t = diferença pelo seu desvio padrão). Os análogos não paramétricos, cujas variáveis são do tipo ordinal e os dados ordenados por postos (ranks), permitem as mesmas conclusões; neste caso não são comparadas médias e não fazemos inferências dos parâmetros da população. Os testes não paramétricos têm maior probabilidade de cometermos erro tipo II.

Para análises com mais de duas amostras, é inválido fazer comparações duas a duas, porque a chance de cometermos um erro tipo I é muito grande, devido ao nível de significância ficar maior do que 5% (Zar, 1996). Nestes casos, utilizamos análises de variâncias (Anova), que podem ser executadas por blocos, com um, dois ou três fatores e com repetições ou sem repetições, sempre dependendo do tipo de questão. A essência da análise de variância, respeitadas as condições de variâncias iguais (homocedasticidade) e aleatoriedade das amostras, é comparar a variação dentro e entre as amostras, se a segunda for considerada significativamente maior do que a primeira, a hipótese nula será rejeitada. É um dos testes mais executados na pesquisa biológica; na forma paramétrica quando as variáveis são de escala geométrica, na forma não-paramétrica se as variáveis forem de escala ordinal. A complementação da Anova, quando significativa, é feita através de comparações múltiplas das médias (paramétricos) ou através de comparações múltiplas dos postos (não- paramétricos). Na forma paramétrica os testes de comparações múltiplas seguem a “linha” dos testes de Tukey-Duncan-Kramer (H_0 : as médias são iguais); na forma não-paramétrica há os análogos, tipo Nemenyi, Dunn e Newman-Keuls, que comparam postos ou diferenças

entre múltiplas medianas (Zar, 1996; Vanzolini, 1993; Siegel, 1977).

Muitas vezes é importante verificar na pesquisa se duas ou mais variáveis geométricas estão relacionadas entre si, situação na qual uma das variáveis é dependente da outra, não implicando em relações de causa e efeito. Colocadas em gráfico, a variável dependente é o eixo do y; a independente é o eixo do x, dizemos então regressão de y sobre x, sempre a variável dependente sobre a independente. Na sua forma simples, a regressão linear tem a relação funcional $y_i = a + bx$; a = constante da regressão, b = coeficiente da regressão (inclinação da reta). Análises de regressão, com uma ou mais retas, são extremamente úteis na pesquisa biológica, para verificar relações entre caracteres morfológicos (proporções corporais ou estruturas de plantas), várias relações entre parâmetros fisiológicos e ambientais, efeito de drogas, modelos dose-dependente e probit (DL50), permitindo predizer valores de y dentro da amplitude dos dados (Vanzolini, 1993; Fisher, 1946). A significância da regressão é verificada através de uma análise de variância; na forma simples nos dá a variação total, variação devida à regressão, com 1 grau de liberdade e a variação sobre a regressão (erro) com n-2 graus de liberdade. Se a regressão for significativa, a variância devida à regressão será significativamente maior do que a variância sobre a regressão; o quociente entre estas é um teste de F. Podemos verificar também, qual porcentagem da variação total é explicada pela regressão, através do coeficiente de determinação (r^2), uma medida de quanto uma variável participa na variância da outra. No caso de várias retas, verificamos também as diferenças entre duas retas e a homogeneidade entre mais de duas retas. Nos casos em que a regressão não conforma uma linha reta, podemos usar de anamorfoses para “ajustes”, utilizando transformações logarítmicas do tipo $\log x - \text{arit } y$, $\text{arit } x - \log y$ e $\log x - \log y$.

Questões envolvendo correlação entre duas variáveis geométricas (paramétrico) ou duas ordinais ou uma ordinal e outra geométrica (não-paramétrico)

são comuns na pesquisa biológica, embora não permitam conclusões sobre causa e efeito. O coeficiente de correlação é a raiz quadrada do coeficiente de determinação da regressão linear, a covariação entre duas variáveis. Dois ou mais coeficientes de correlação podem ser comparados; testes semelhantes ao de Tukey para comparações múltiplas podem ser utilizados para comparar pares de coeficientes entre si.

No caso de termos duas ou mais variáveis categóricas (nominais), utilizamos as tabelas de dupla entrada formando células que contêm as frequências esperadas e observadas, com o objetivo de verificar associações e homogeneidades entre as categorias (por exemplo, numa tabela 2 x2 teremos 4 categorias). Estas verificações são extremamente úteis na pesquisa biológica, o critério é a distribuição qui-quadrado, introduzida por Karl Pearson e expandida por ele e R.A.Fisher (Zar, 1996). Os testes medem o ajustamento (ou distância) entre as frequências esperadas e observada. O teste G (Wilks), para várias linhas e colunas, o teste exato de Fisher, para expectativas pequenas (< 5) e o teste de Kolmogorov-Smirnov são dos mais utilizados.

Outras distribuições, importantes em estudos ecológicos, são (Zar, 1996): a distribuição binomial, que considera proporção de indivíduos em duas categorias, p e q, de modo que $q = 1 - p$; a distribuição de Poisson, para verificar a distribuição espacial de indivíduos na população, caracterizada pela hipótese nula média igual à variância - distribuição ao acaso (Vilar *et.al*, 2000); a distribuição circular, importante nos estudos sobre orientação em aves. Para todos estes paramétricos há os análogos não paramétricos. Existem bons textos sobre estatística para estudos ecológicos, nós sugerimos Abuara & Petrere (1997), Brower & Zar (1997) e Hairston (1994).

REFERÊNCIAS

Abuara, M.A.P. & M. Petrere, Jr. 1997. **Estimativas da**

- abundância de populações animais. Introdução às técnicas de captura e recaptura.** Editora da Universidade Estadual de Maringá/Nupelia, Maringá, Pr. 161 p.
- Brower, J.E., J.H.Zar & C.N.Ende, 1997. **Field and laboratory methods for general ecology.** 4th. ed. WCB/McGraw-Hill 273 p.
- Fearnside, P.M. 1986. **Human carrying capacity of the Brazilian rainforest.** Columbia Univ.Press, New York 293 p.
- Fisher, R.A. 1946. **Statistical methods for research workers.** Oliver and Boyd, Edinburgh, London 354p.
- Hurlbert, S.H. 1984. **Pseudoreplication and the design of ecological field experiments.** Ecological Monographs 54 (2): 187-211.
- Hairston, N.G.1994. **Ecological Experiments. Purpose, design and execution.** Cambridge Univ. Press 370 p.
- Hayek, L.C.1994. Research design for quantitative amphibian studies, pp. 21 – 39. *In: Measuring and monitoring biological diversity. Standards methods for amphibians.* (W.R.Heyer, M.A.Donnely, R.W.McDiarmid, L.A. & M.S.Foster, Eds.). Smithsonian Inst. Press 364 p.
- Siegel, S. 1977. **Estatística não-paramétrica para as ciências do comportamento.** McGraw-Hill do Brasil, São Paulo 350p.
- Stevens, S.S. 1946. On the theory of scales measurement. **Science** 103:677-680.
- Zar, J.H. 1996. **Biostatistical analysis.** 3rd.ed. Prentice Hall, New Jersey 662p + Tabs.
- Vanzolini, P.E. 1993. **Métodos estatísticos elementares em sistemática zoológica.** Hucitec, São Paulo 130p.
- Vilar, J.C., N.A.C.Ziengier & C.M.Carvalho, 2000. Distribuição espacial de *Vellozia dasypus* Sembert (Velloziaceae) e *Melocactus zehntneri* (Britt et Rose) Lützelb (Cactaceae) na Serra de Itabaiana, Sergipe. **Biol. Geral Exper.** 1(1): 5 – 15.