Ronaldo Figueiró

Manual prático de Bioestatística Computacional

FOA

Ronaldo Figueiró

Manual prático de Bioestatística Computacional

1ª Edição

FOA

Volta Redonda

2014

Centro Universitário de Volta Redonda - UniFOA
Campus Três Poços
Av. Paulo Erlei Alves Abrantes, nº 1325
Três Poços, Volta Redonda /RJ / CEP. 27240-560
Tel.: (24) 3340-8400 - FAX: 3340-8404
www.unifoa.edu.br

EDITORA FOA
www.unifoa.edu.br/editorafoa

FICHA CATALOGRÁFICA
Bibliotecária: Alice Tacão Wagner - CRB 7/RJ 4316

F475m Figueiró, Ronaldo.
Manual prático de bioestatística computacional. [recurso eletrônico] / Ronaldo Figueiró. FOA, 2014. 60 p.

ISBN: 978-85-60144-60-0

1. Bioestatística. 2. Bioestatística computacional - manual. I. Fundação Oswaldo Aranha. II. Centro Universitário de Volta Redonda. III. Título.

CDD – 570.15195

Prefácio

O objetivo deste manual prático de bioestatística não é o de substituir livros-texto da área, mas sim prover uma leitura rápida e dinâmica, a qual instrumente o aluno com os conceitos e procedimentos básicos em softwares estatísticos. Neste manual são apresentados os conceitos basais de bioestatística voltada para experimentação, desde o planejamento experimental pautado pelo tratamento estatístico que o pesquisador utilizará *a posteriori*, até os testes de hipóteses, foco deste manual e ferramentas de frequente uso nas Ciências Biológicas.

Sumário

CAPÍTULO 1
Introdução à bioestatística

A Estatística pode ser conceituada como a área do conhecimento que trata da coleção, organização e análise de dados dados esses, os quais trazem informações sobre características de grupos ou sistemas. Ela pode ser interpretada fundamentalmente como uma ferramenta que permite aos pesquisadores identificarem padrões e diferenciá-los de eventos decorrentes do acaso.

Dentro desta área do conhecimento, o ramo que trata essencialmente da coleta, organização, análise e processamento dados referentes às ciências biológicas e da saúde é denominado bioestatística, e é de extrema relevância para o estudo de fenômenos biológicos.

Na área médica, por exemplo, quando um clínico geral pergunta a um paciente sobre seu histórico familiar, ele o faz devido ao conhecimento prévio de padrões identificados pela estatística, tais como a predisposição ao desenvolvimento de determinadas doenças em uma mesma família.

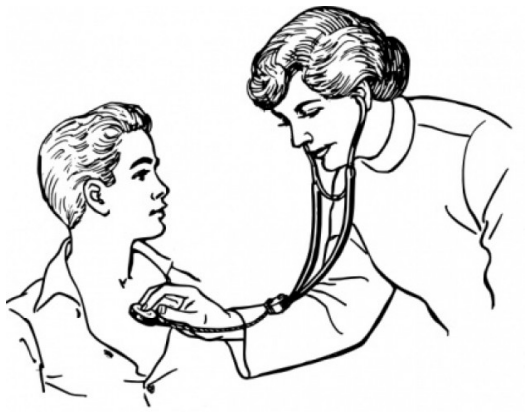


Imagem de domínio público retirada de http://www.freebievectors.com

Na área das ciências ambientais, por exemplo, o uso de determinados organismos como bioindicadores, isto é, indicadores do grau de integridade ambiental, se dá devido ao conhecimento dos fatores abióticos correlacionados significativamente com sua distribuição: ou seja, através de análises estatísticas, foi possível se determinar um padrão de resposta daqueles organismos para um ambiente impactado.

Larvas de simulídeos (Diptera: Simuliidae) fixadas em substrato vegetal (Foto: Ronaldo Figueiró)

Na figura acima, é possível se observar um conjunto de larvas de borrachudos (Diptera: Simuliidae): organismos que análises estatísticas em estudos científicos demonstram que tendem a se tornar mais abundantes em condições de poluição moderada da água, assim constituindo organismos bioindicadores.

A bioestatística pode ser ramificada em duas vertentes: a chamada estatística descritiva, a qual se dedica à organização e caracterização de conjuntos de dados, a partir de uma série de medidas descritivas, tais como a média, desvio padrão e variância, e a estatística analítica ou inferencial, ou seja, aquela na qual a partir de uma amostra retirada de uma população (conjunto universo), é possível se chegar a algumas conclusões sobre a população de origem.

Estatística descritiva	**Estatística analítica**
Levantamento	Formulação de hipóteses
Organização	Inferência estatística
Classificação	
Descrição	
Cálculo de parâmetros representativos	

Algo importante a ser ressaltado, é que em bioestatísica, alguns termos são frequentemente utilizados, e alguns apresentam significados bastante distintos de seus significados biológicos:

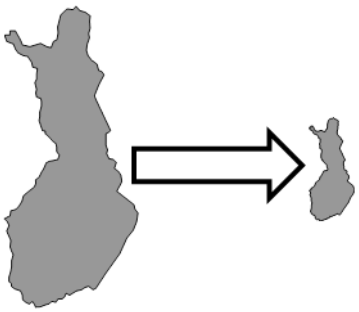
População: diferente do significado biológico que se refere a um conjunto de indivíduos de uma mesma espécie em um dado local, população no sentido estatístico se refere a qualquer que seja o conjunto universo em estudo. (Exemplos: A fauna de um determinado local, os estudantes de uma determinada universidade, as universidades que compõem o ensino superior de um determinado local, etc.)

Unidade amostral: É a unidade na qual são observadas e medidas as características quantitativas e qualitativas da população: cada amostra é composta por um conjunto de unidades amostrais, cada uma destas gerando uma única observação da variável de interesse.

Tratamentos: este termo se refere na verdade às diferentes condições as quais o pesquisador está comparando para testar sua hipótese. Enquanto podem se tratar literalmente de tratamentos médicos (comparar indivíduos que receberam a droga A com indivíduos que receberam a droga B), pode na verdade se tratar de quaisquer dois cenários sendo comparados, como por exemplo as abundâncias de determinadas espécies em uma área preservada e em uma outra área impactada.

Noções de amostragem

Uma amostra pode ser conceituada como um subconjunto de um conjunto maior, o conjunto universo (ou a população, em termos estatísticos), o qual é impossível de ser estudado em sua totalidade. Uma premissa básica da amostragem é que a amostra deve ser representativa do todo.

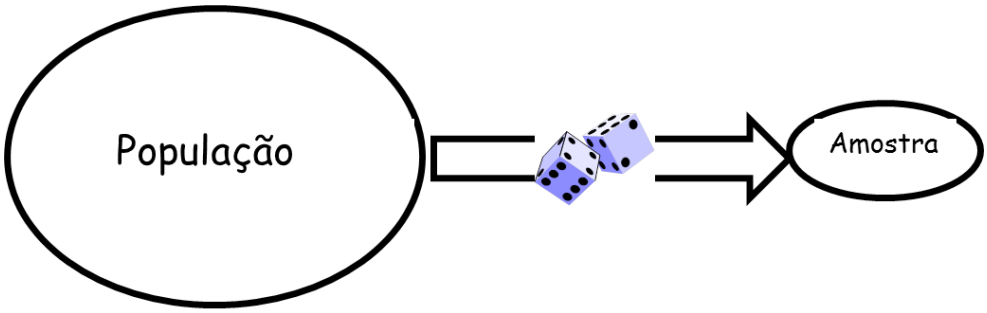


De acordo com o tipo de pergunta a qual o pesquisador pretende responder em seu estudo, diferentes técnicas de amostragem devem ser empregadas. A seguir, serão apresentadas as principais técnicas utilizadas em estudos científicos.

As amostragens podem ser de natureza probabilística, isto é, quando as amostras são retiradas do conjunto universo de forma aleatória, ou não probabilísticas, quando estas são realizadas sem aleatoriedade.

Tipos de amostragem probabilística:

Amostragem aleatória simples: Neste tipo de amostragem, os pontos são retirados de forma aleatória, isto é, ao acaso, do seu conjunto universo de estudo.



Um exemplo de amostragem aleatória simples em dados ambientais é o uso de quadrats, que são molduras quadradas as quais são distribuídas de forma aleatória pelo ambiente, sendo coletado ou quantificado todo o seu interior.

Figura ilustrativa de coleta com quadrats.

Amostragem aleatória sistemática: Neste tipo de amostragem, o ponto de partida é sorteado, e a partir deste ponto, amostras são retiradas em intervalos regulares. Estes intervalos podem ser de tempo, ou intervalos espaciais. Em estudos ambientais, os transectos são bons exemplos de amostragem sistemática, e são particularmente úteis no estudo de gradientes ambientais.

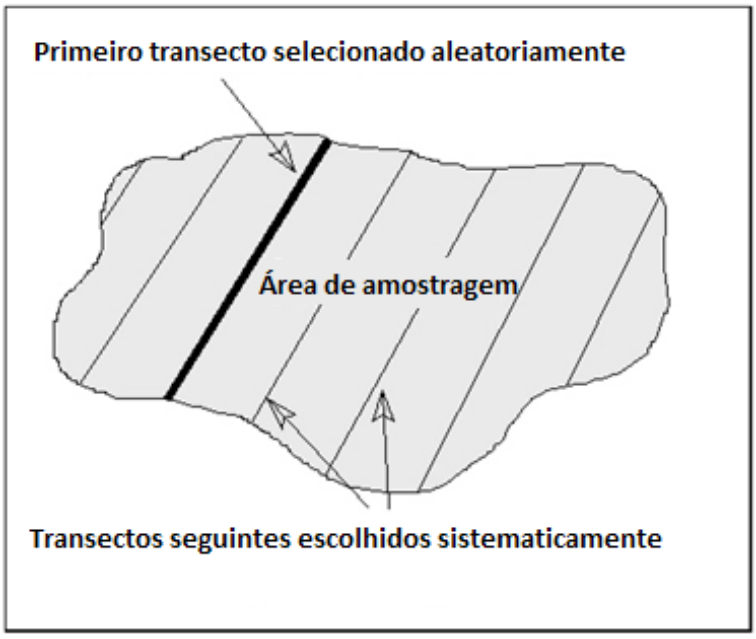


Figura ilustrativa de coleta por método de transectos.

CAPÍTULO 2
Princípios de delineamento experimental

Antes de um pesquisador iniciar qualquer experimento, ele deve ponderar sobre quais são suas perguntas, suas hipóteses, e como os dados que serão coletados no decorrer do experimento possibilitarão alcançar as respostas procuradas.

Em essência, antes que qualquer experimento tenha início, é necessário se pensar no seu desenho experimental, ou delineamento experimental: de acordo com o fenômeno a ser estudado, diferentes métodos devem ser utilizados, e de acordo com o tipo de análise estatística que se tem em mente, diferentes procedimentos de amostragem devem ser executados.

Sim! ANTES de se iniciar qualquer experimento, deve-se ter em mente exatamente o tratamento estatístico que será aplicado posteriormente, e não se pensar na estatística que será usada somente após os dados estarem coletados. Quando um experimento é planejado, o pesquisador deve ter claramente em mente quais são as premissas das análises que ele tem em mente usar, para que depois não se depare com a impossibilidade de analisar seus dados conforme desejava, e consequentemente, não ser capaz de alcançar as respostas que buscava, devido aos seus dados não atenderem às premissas da análise desejada. Todas as análises estatísticas apresentam premissas, ou seja, condições que se não forem atendidas pelos dados, ou impossibilitarão sua execução, ou pior, levarão a resultados enganosos.

Elaboração de hipóteses em um estudo

A primeira questão pertinente ao delineamento experimental é o estabelecimento da questão central do estudo, isto é: o que o pesquisador quer saber? Em estudos ecológicos, por exemplo, em geral as principais questões são se existem variações espaciais ou temporais da variável Y, se uma variável independente X tem efeito sobre a variável Y e qual é seu efeito, se as variações da variável Y são consistentes com as predições da hipótese H.

Entretanto, o que é uma hipótese? Trata-se de um postulado testável, o qual pode ser confirmado, ou refutado. Em todo o estudo científico, o primeiro passo é o estabelecimento de hipóteses.

Hipótese nula (H0): Trata-se da negação da premissa, se referindo à ausência de efeito, à ausência de diferença. Por exemplo: em um estudo para se comparar a eficiência de bactérias degradadoras do 2,4-D, a hipótese nula seria de que não há diferença na eficiência das diferentes cepas.

Hipótese alternativa (H1): A hipótese alternativa, ou hipótese científica, se refere à afirmativa, à confirmação da existência de diferença entre grupos ou condições: no caso do exemplo anteriormente citado, a hipótese alternativa seria que as cepas apresentam diferenças.

Tipos de experimentos biológicos

Uma vez estabelecidas nossas hipóteses, como testá-las? É nesse ponto que entram os experimentos. Experimentos em biologia podem basicamente classificados em duas categorias: na primeira, os estudos manipulativos, o pesquisador manipula uma variável independente de seu interesse para investigar o seu efeito sobre uma ou mais variáveis dependentes de interesse que respondem a essa manipulação: são essencialmente aqueles estudos *in vitro* realizados em laboratório, tais como o estudo da resistência de diferentes cepas de um dado microorganismo à uma determinada droga, ou aqueles estudo *in situ*, ou seja, realizados na natureza, mas que no entanto envolvem artifícios utilizados para a manipulação das variáveis de acordo com o interesse do pesquisador.

Experimentos manipulativos

Um exemplo desta última categoria seria determinar se um determinado macroinvertebrado lótico tem sua densidade controlada pela predação que sofre de peixes. Neste caso, trechos do rio poderiam ser isolados com redes de forma a termos diferentes tratamentos, neste caso, diferentes densidades de peixes por meio de manipulação, e assim verificarmos como a densidade do macroinvertebrado em questão responde à densidade de peixes. Outro exemplo seria se observar a parasitemia no sangue de camundongos expostos a diferentes condições de cativeiro.

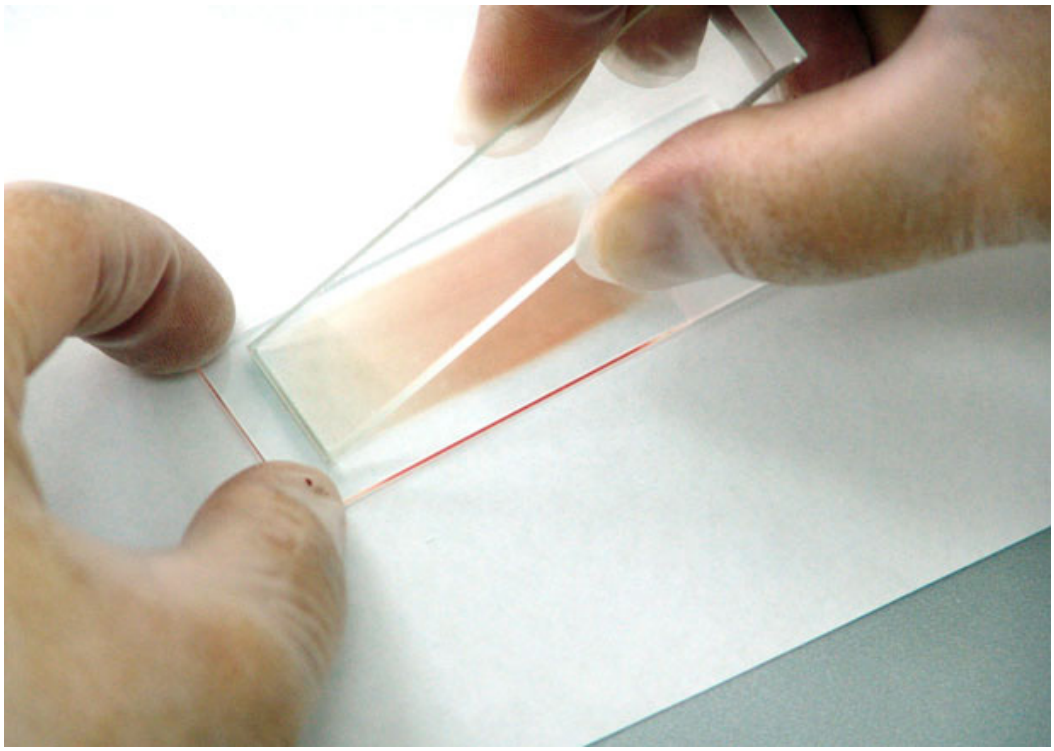


Imagem retirada de autoria de Максим Кукушкин retirada de http://www.publicdomainpictures.net

Embora este tipo de desenho experimental seja bastante difundido, ele possui limitações. Experimentos dessa natureza conduzidos em grandes escalas acabam comprometendo sua replicação, e experimentos em escalas menores, apesar de mais facilmente replicáveis, podem não produzir padrões similares aos de escalas maiores.

Além estas questões, ainda se faz necessário que seja levado em consideração que, em experimentos manipulativos de campo, dificilmente o pesquisador será capaz de controlar a variação apenas da variável de interesse, dessa forma podendo gerar outras variáveis de confusão em potencial.

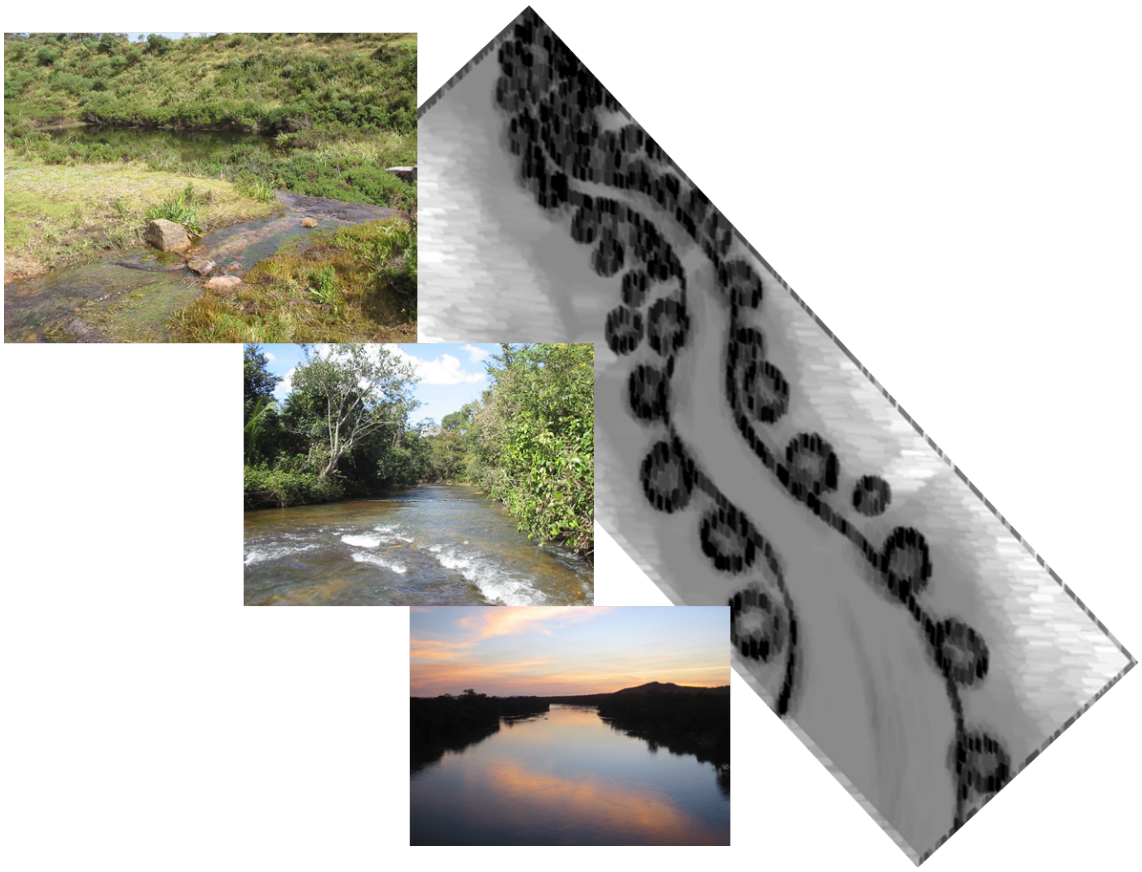
Experimentos de Pressão X Experimentos de Pulso

Em experimentos manipulativos, podemos observar duas subdivisões categóricas: os experimentos de pressão, nos quais uma determinada variável de interesse é manipulada durante todo o período de experimento de forma a se comportar como uma constante, e os experimentos de pulso, no qual a variável é manipulada em um primeiro momento, e a partir de então seu comportamento natural é acompanhado.

Experimentos Naturais

Uma alternativa aos experimentos manipulativos são os chamados experimentos naturais, que na verdade se tratam de estudos observacionais nos quais o pesquisador irá tirar vantagem da variação natural presente na variável independente de interesse.

No caso do experimento anteriormente proposto, uma alternativa na forma de um experimento natural seria amostrar diversos trechos de rios que variem naturalmente em relação às suas densidades populacionais de peixes, e assim analisar a relação entre as densidades de macroinvertebrados e de peixes.



Variação de habitats ao longo de um rio (Fotos: Ronaldo Figueiró)

Em experimentos naturais, outras variáveis além das de interesse não são controladas de forma alguma, o que faz com que não seja possível se estabelecer uma única hipótese alternativa, desta forma aumentando a complexidade da interpretação dos resultados deste tipo de estudo. Uma forma de lidar com esta limitação é buscar identificar outras variáveis potencialmente relacionadas à variável de interesse, e procurar diminuir ao máximo sua variação entre as áreas amostradas.

Neste exemplo, a presença de cobertura vegetal densa pode ser um fator que influencie tanto as populações de macroinvertebrados quanto a de peixes. Uma alternativa seria só trabalhar com trechos com a mesma densidade de cobertura vegetal.

Experimentos fotográficos X Experimentos de trajetória

Os experimentos naturais, por sua vez, podem ser categorizados como experimentos fotográficos ou experimentos de trajetória. A diferença básica entre esses dois desenhos experimentais está na forma que a replicação é feita: em experimentos fotográficos, as réplicas são espaciais, realizadas em um mesmo momento no tempo, enquanto que em experimentos de trajetória, as réplicas formam uma série temporal, são realizadas ao longo do tempo.

As vantagens de um experimento fotográfico estão no fato de ele ser mais rápido, e apresentar réplicas que são estatisticamente mais independentes umas das outras do que as réplicas temporais de um experimento de trajetória. Na verdade, grande parte dos estudos de séries temporais são de fato experimentos fotográficos, uma vez que a variação no espaço é tratada como substituta para a variação no tempo.

Um experimento de trajetória já tem como vantagem revelar as mudanças de sistemas biológicos ao longo do tempo, algo que é descrito por muitos modelos ecológicos e ambientais, e esse tipo de experimento gera dados mais facilmente comparáveis com as predições dos modelos. Entretanto, este tipo de delineamento, conforme exposto anteriormente, pode gerar amostras potencialmente dependentes entre si.

A questão da replicação é essencial em um estudo científico, e o número de réplicas necessário deve ser determinado com base na variância dos dados e no tamanho do efeito, ou seja, a diferença que o pesquisador deseja detectar entre as médias de grupos em comparação.

Um método de se estimar a variância dos dados para determinar a necessidade de réplicas é a partir de um experimento piloto, ou a partir de estudos anteriores de outros pesquisadores. O ideal é que todas as réplicas sejam realizadas de forma simultânea, gerando um efeito fotográfico, o que se torna progressivamente mais difícil à medida que a escala espacial do estudo aumenta.

Usualmente é utilizada a regra dos 10, que prevê que ao menos 10 réplicas de cada categoria ou nível de tratamento, entretanto, em estudos de larga escala e estudos de impacto ambiental o delineamento ADCI (Antes-Depois, Controle-Impacto) é mais adequado: são coletadas repetidamente amostras do local controle e impactado antes e depois do impacto.

Usina de aproveitamento hidrelétrico antes e após sua construção, exemplo de sistema para um estudo no delineamento ADCI (Fotos: Ronaldo Figueiró)

Assegurando a independência das amostras

Outra questão de extrema importância em estudos ambientais é assegurar a independência entre as amostras. Esta é uma questão que frequentemente não é abordada em artigos científicos, no entanto de vital importância para a interpretação correta dos padrões. Por independência se entende que uma amostra não tem qualquer influência sobre a outra. Uma outra questão que pode comprometer a interpretação de dados ecológicos são as chamadas variáveis de confusão. Este tipo de variável, é caracterizada por estar associada à variável resposta ou à variável preditora, o que interfere na relação de causa e efeito entre elas. Uma forma de lidar com essas duas questões é a própria replicação, e além dela, a aleatorização das amostragens.

A escala espacial de um estudo é na verdade composta por dois conceitos complementares: a extensão e o grão. Por extensão entende-se a área total englobada por todas as unidades amostrais do estudo, e por grão se entende a menor unidade amostral. No exemplo dos macroinvertebrados, supondo-se que em cada trecho amostrado os macroinvertebrados fossem coletados de quadrats aleatórios, o quadrat seria o grão.

CAPÍTULO 3
Introdução à Estatística Descritiva

Conforme abordado anteriormente, podemos subdividir a estatística em duas subcategorias: a estatística inferencial, ou estatística analítica, que trata do teste de hipóteses, e a estatística descritiva, assunto deste capítulo, que trabalha com as chamadas medidas-resumo.

As medidas-resumo não tem o objetivo de testar hipóteses: sua função é caracterizar a amostra, como seu nome já informa, resumir as características do subconjunto em estudo.

É importante lembrar que uma boa amostra deve ser representativa do todo, isto é: suas medidas resumo devem estar o mais próximas dentro do possível da população, ou seja, do conjunto universo do qual elas são oriundas.

Descritiva	**Analítica**
Tabelas e Gráficos	Inferência Estatística
Média, Mediana e Moda	Testes de Hipóteses
Amplitude de variação	
Coeficiente de variação	
Separatrizes	
Desvio Padrão	
Variância	

Desta forma, iniciaremos este capítulo definindo os dois subconjuntos de medidas-resumo existentes: existem as medidas de tendência central, que conforme o nome diz indicam qual é a tendência central da amostra, e as medidas de dispersão, as quais irão demonstrar o quanto os valores na amostra variam.

Medidas de Tendência Central

Média

Destas medidas, o exemplo mais comum á a média aritmética. Ela consiste da soma dos valores de uma variável divididas pelo número de observações ou réplicas.

Vamos imaginar que um pesquisador deseje estimar o tamanho médio dos indivíduos de um inseto, o *Simulium stellatum*. Consideremos que 10 indivíduos tivessem sido coletados, com os seguintes valores de tamanho de corpo:

1,0 mm; 0,8 mm; 0,6 mm; 1,2 mm; 1,1 mm; 0,6 mm; 1,0 mm; 0,6mm; 1,3 mm; 1,1 mm

Como o tamanho médio seria determinado? Pelo somatório dos valores observados divididos pelo número de observações, isto é: (1+0,8+0,6+1,2+1,1+0,6+1+0,6+1,3+1,1)/10 = 9,3

Média de frequências

Outra forma de aplicarmos a média aritmética para caracterizar uma amostra, é realizar a média de frequências. Para tanto, é necessário definirmos frequência: trata-se do número de vezes que um determinado evento ocorre em uma amostra.

Fazendo uso do exemplo anterior, é possível se estabelecer a frequência com que cada tamanho de corpo foi observado na amostra, desta forma:

Tamanho do corpo (mm)	Fi	XiFi
1,0	2	2,0
0,8	1	0,8
0,6	3	1,8
1,2	1	1,2
1,1	2	2,2
1,3	1	1,3

Na primeira coluna, temos o tamanho do corpo, na segunda o número de vezes (frequência) que este foi observado na amostra, e na terceira, o produto destes dois valores.

Somando-se todos os produtos, chegamos ao valor 9,3. Como nossa amostra tem 10 espécimes, a soma das frequências só pode ser 10: assim, dividindo-se 9,3/10, chegamos à frequência média de 0,93.

Média de valores em intervalo de classe

Em alguns casos, o pesquisador pode optar por trabalhar com intervalos, ao invés de valores absolutos. Nesse caso, o cálculo da média é feito da seguinte forma: o valor da frequência absoluta

(fi) é multiplicado pelo ponto médio de seu intervalo de classe (Xipm). Após este procedimento, os produtos devem ser somados e divididos pelo total da frequência absoluta, que nesse caso é: 98/11 = 8.90909...

Tamanho	fi	Ponto Médio (Xipm)	Xipmfi
4,0 – 8,0	6	6	36
8,0 - 12,0	3	10	30
12,0 – 16,0	1	14	14
16,0 – 20,0	1	18	18
Total	11		98

Mediana

A mediana é o valor que ocupa a posição central de uma distribuição de n observações, uma vez que as mesmas estejam ordenadas de forma crescente.

Quando existe um número ímpar de observações, a mediana é dada pelo valor central, enquanto que quando o número de observações é par, a mediana é dada pela média aritmética dos dois valores centrais.

Tomando como exemplo o caso anteriormente exposto das medidas do tamanho de indivíduos de *S.stellatum*, organizando em ordem crescente, teríamos: (0,6; 0,6; 0,6; 0,8;1,0; 1,0; 1,1; 1,1; 1,2; 1,3).

Desta forma, como temos um número par de observações, a mediana seria dada pela média aritmética dos valores centrais, sendo: 1+1/2=1.

Moda

A moda é o valor que mais se repete nas observações. Utilizando o exemplo anterior, temos a seguinte distribuição de observações: (0,6; 0,6; 0,6; 0,8;1,0; 1,0; 1,1; 1,1; 1,2; 1,3)

Nesse caso, o valor que mais se repete é 0,6, sendo assim, esta é a moda deste conjunto de dados.

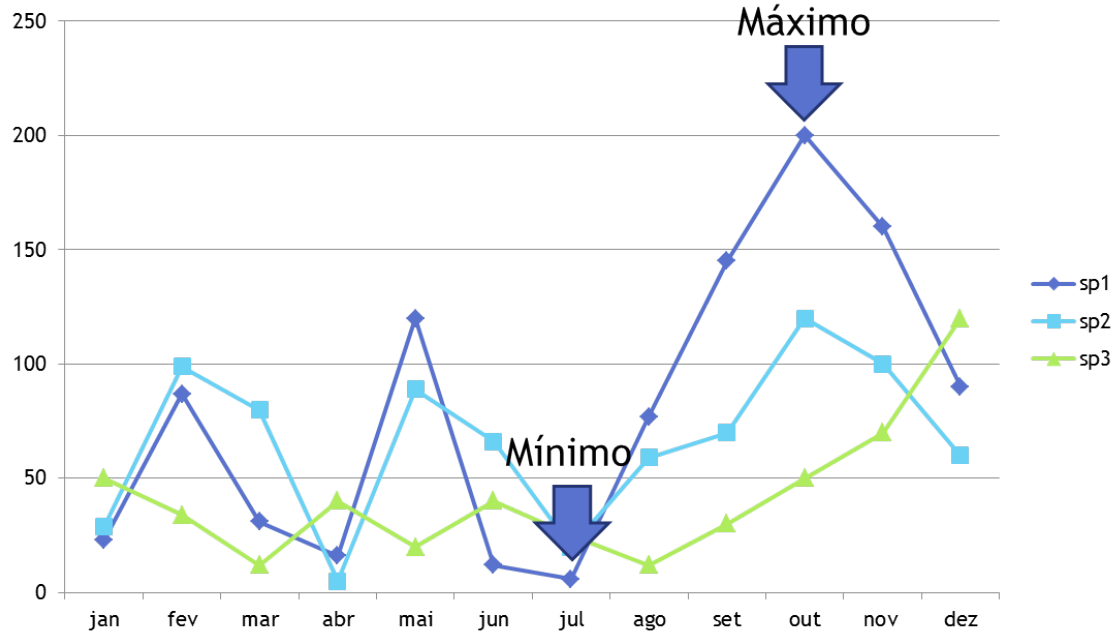
Larva de ***Simullium stellatum*** (Foto retirada de Gil-Azevedo, Figueiró & Maia-Herzog 2005)

Medidas de Dispersão

As medidas de tendência central possuem grande importância na estatística descritiva, entretanto elas não são suficientes para caracterizar um conjunto de dados: a sua validade somente pode ser determinada se conhecendo a variação da população, através das medidas de dispersão.

Máximo e mínimo

São o valor mais alto e mais baixo registrados na população, conforme a figura abaixo demonstra:



Amplitude de variação

É determinada pela diferença entre os valores mais alto e mais baixo registrados na população.

Embora essas medidas de dispersão sejam fáceis de serem determinadas, não são as mais adequadas para a determinação da variação em uma população, que são:

Variância: A variância mede a variabilidade ou o espalhamento dos dados ao redor da média das observações.

$$s^2 = \frac{\sum_{i}^{N} (X - \overline{X})^2}{N - 1}$$

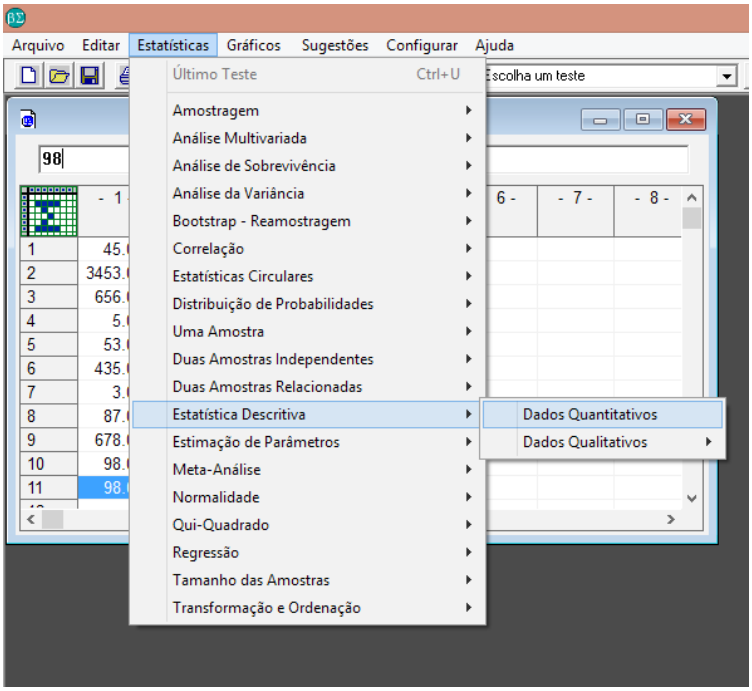
Desvio Padrão: É dado pela raiz quadrada da variância. Esta é a medida mais comum de dispersão, a qual mostra o quanto de variação existe em relação à média ou ao valor esperado. Valores baixos indicam que os dados estão dispersos próximos à média, enquanto que valores altos indicam grande dispersão dos dados.

$$s = \sqrt{\frac{\sum_{i}^{N} (X - \overline{X})^2}{N - 1}}$$

Coeficiente de variação: É uma medida de dispersão utilizada para se estimar a precisão de experimentos, que consiste do desvio-padrão expresso como porcentagem da média. Sua principal qualidade é a capacidade de comparação de distribuições diferentes. É obtido pela soma dos quadrados dos desvios em relação à média dividida pelo N.

Como obter estas medidas em um software estatístico

Existe uma série de softwares estatísticos gratuitos disponíveis para download na internet, entretanto neste livro optaremos pela utilização dos pacotes Bioestat® 5.3. , um pacote estatístico nacional desenvolvido por Manuel Ayres e disponível para download na página do Instituto Mamirauá: http://www.mamiraua.org.br/pt-br/downloads/programas/bioestat-versao-53/ e Past, desenvolvido por Øyvind Hammer, David A.T. Harper e P.D. Ryan, disponível para download em: http://folk.uio.no/ohammer/past/index_old.html



Para a obtenção das medidas de estatística descritiva de qualquer conjunto de dados, é necessário que os mesmos sejam inseridos na forma de uma planilha no programa, no qual cada série é representada em uma coluna. Uma vez inseridos os dados, deve-se acessar o menu estatísticas -> estatística descritiva -> dados quantitativos.

CAPÍTULO 4
Apresentação de dados em Gráficos e Tabelas

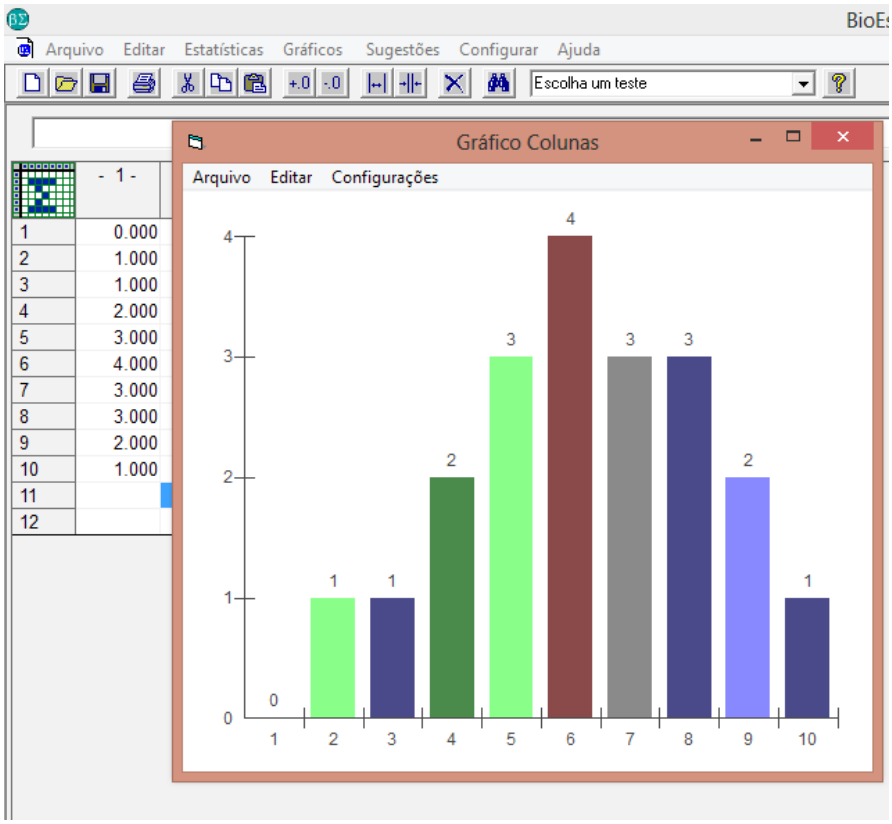
A representação gráfica de dados pode ser uma valiosa aliada na interpretação dos mesmos, entretanto, para que isto aconteça, se faz necessária a escolha do tipo de gráfico mais adequado aos dados em questão.

Por exemplo, quando queremos representar graficamente uma distribuição de frequências, tais como as notas de uma turma em uma determinada disciplina, é comum fazermos o uso de um gráfico de colunas, no qual barras representarão no eixo X a nota em questão, e no eixo Y quantas vezes a mesma se repetiu.

Imaginemos que em uma turma existiu a seguinte distribuição de notas:

Nota	1	2	3	4	5	6	7	8	9	10
Frequência	0	1	1	2	3	4	3	3	2	1

Para obtermos o gráfico de colunas desta distribuição de notas, devemos inserí-la no Bioestat® e no menu Gráficos, selecionar a opção colunas simples. Como as notas são apenas os rótulos, só é necessária a inserção da coluna frequência, sempre em colunas.

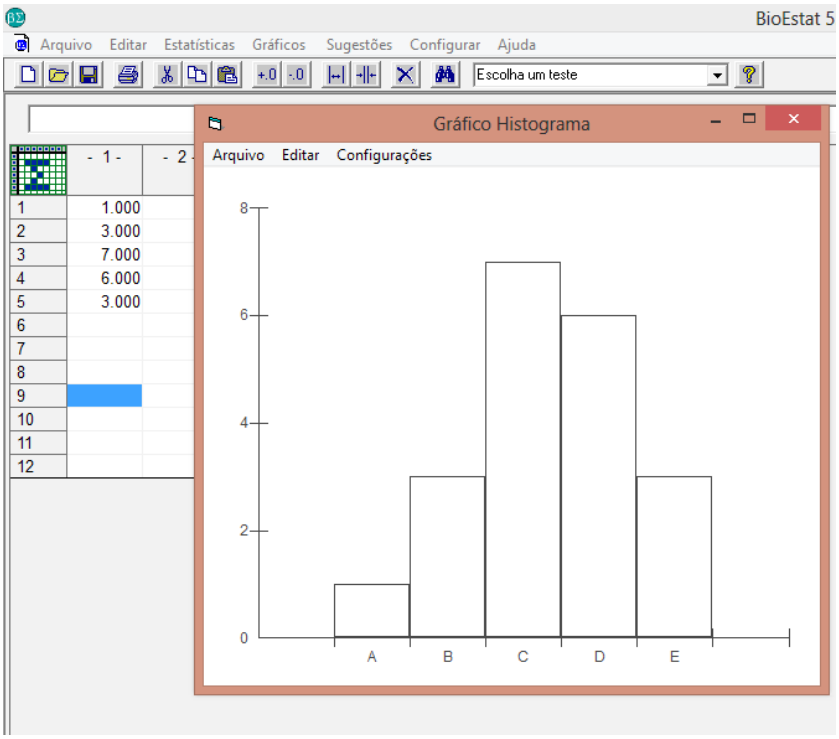


Ao observarmos o gráfico de colunas acima, podemos notar que a nota que mais se repetiu na turma, isto é, a moda, é 6 (4 alunos obtiveram esta nota), enquanto que a nota mais alta (o máximo) foi 10, obtida por um aluno, e a nota mais baixa (o mínimo desta distribuição) foi 2, obtida também por um aluno.

Aproveitando este mesmo conjunto de dados, vamos organizá-los de uma forma diferente: ao invés de trabalharmos com as notas absolutas, trabalhemos com intervalos de classe:

(1-2)	(3-4)	(5-6)	(7-8)	(9-10)
1	3	7	6	3

Neste caso, o gráfico que geraremos, apesar de ter uma aparência similar ao gráfico de colunas, recebe uma denominação especial: histograma. Repetindo o mesmo procedimento de inserção dos dados e solicitação do gráfico (desta vez histograma), temos a seguinte figura:

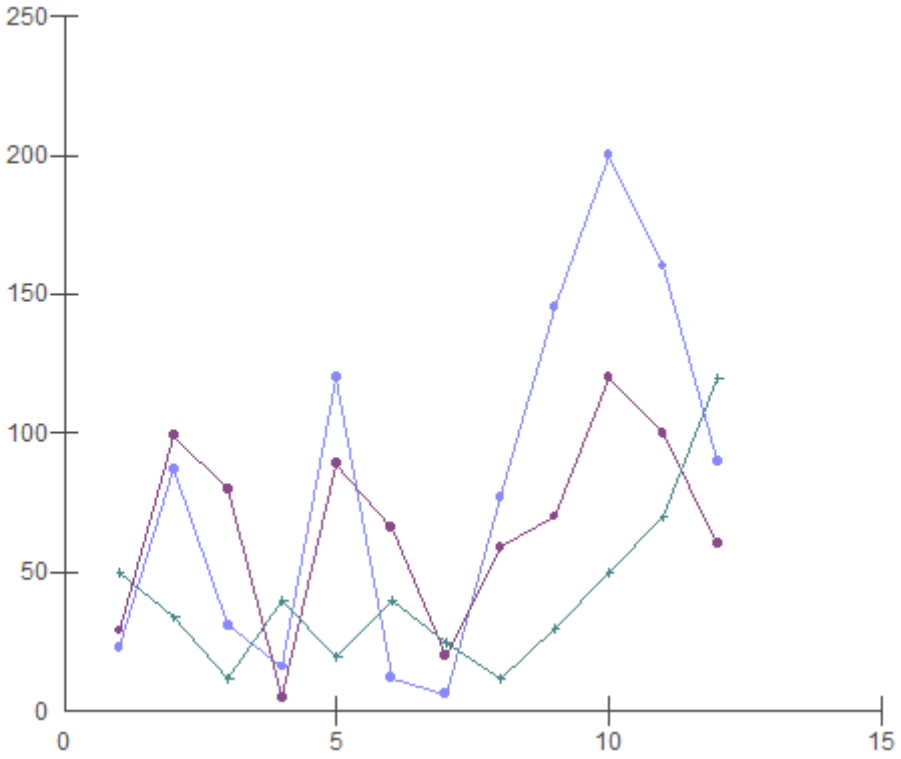


É possível se notar que no eixo X, os valores foram substituídos por letras, as quais designam as classes: A (1-2), B (3-4), C (5-6), D (7-8) e E (9-10).

Quando se está representando uma dinâmica temporal, como as flutuações populacionais ao longo de um ano, como indicado na tabela abaixo, o gráfico mais indicado é o gráfico tipo linha.

	sp1	**sp2**	**Sp3**
Jan	23	29	50
Fev	87	99	34
Mar	31	80	12
Abr	16	5	40
Mai	120	89	20
Jun	12	66	40
Jul	6	20	25
Ago	77	59	12
Set	145	70	30
Out	200	120	50
Nov	160	100	70
dez	90	60	120

Para obtermos o gráfico de linhas destas dinâmicas populacionais, devemos inserí-la no Bioestat® e no menu Gráficos, selecionar a opção linhas. Quando surgir a pergunta se os dados são pares ordenados, responder não. A primeira coluna deve ser preenchida com os números de 1 a 12, simbolizando os meses.



CAPÍTULO 5
Introdução aos Testes de Hipóteses: Qui-quadrado

Os testes de hipóteses são parte da chamada inferência estatística, e constituem um poderoso conjunto de ferramentas estatísticas que permitem que um pesquisador infira algumas características de uma determinada população.

O teste do Qui-quadrado é usualmente aplicado com o intuito de inferir se existe diferença significativa entre as frequências observadas e as esperadas. Desta forma, o teste Qui-quadrado pode se apresentar em duas diferentes variedades: o teste Qui-quadrado para aderência, no qual um conjunto de dados coletado é comparado com um conjunto de valores esperado; e o teste Qui-quadrado para independência, no qual dois conjuntos de dados coletados são comparados entre si para averiguar se existe diferença significativa entre ambos.

É importante ressaltar que neste teste somente podem ser usados valores absolutos, dados em proporções e percentuais não podem ser analisados.

Teste Qui-quadrado de independência

Considere, por exemplo, que foi realizada uma pesquisa para determinar os gêneros preferidos de filmes na população de uma cidade, com o objetivo de investigar se existe diferença entre as preferências do sexo masculino e feminino. Os pesquisadores chegaram a estes valores:

	Drama	**Comédia romântica**	**Ficção Científica**	**Ação**
Masculino	100	150	50	100
Feminino	250	300	50	5

Neste caso, as hipóteses seriam:

H0 – Não existe diferença entre as preferências dos sexos

H1 – Os sexos tem diferentes preferências de filmes

Para testar as hipóteses, deve ser aplicado o teste Qui-quadrado para independência, através da seguinte fórmula:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Onde Oi são as frequências observadas e Ei as frequências esperadas. O valor de saída desta fórmula, que para este problema é 165,3 , deve ser comparado na tabela abaixo para n-1 graus de liberdade, e p=0,05.

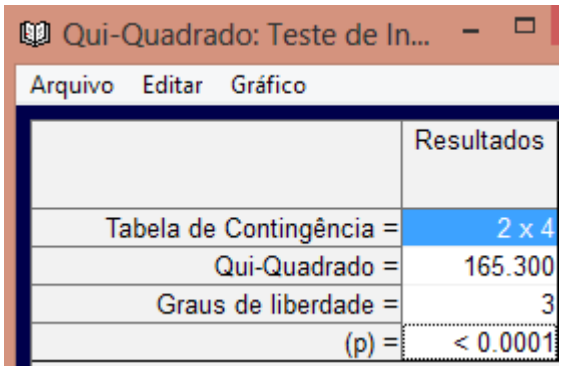
	P										
GL	0,995	0,975	0,2	0,1	0,05	0,025	0,02	0,01	0,005	0,002	0,001
1	0,0000393	0,000982	1,642	2,706	3,841	5,024	5,412	6,635	7,879	9,55	10,828
2	0,01	0,0506	3,219	4,605	5,991	7,378	7,824	9,21	10,597	12,429	13,816
3	0,0717	0,216	4,642	6,251	7,815	9,348	9,837	11,345	12,838	14,796	16,266
4	0,207	0,484	5,989	7,779	9,488	11,143	11,668	13,277	14,86	16,924	18,467
5	0,412	0,831	7,289	9,236	11,07	12,833	13,388	15,086	16,75	18,907	20,515
6	0,676	1,237	8,558	10,645	12,592	14,449	15,033	16,812	18,548	20,791	22,458
7	0,989	1,69	9,803	12,017	14,067	16,013	16,622	18,475	20,278	22,601	24,322
8	1,344	2,18	11,03	13,362	15,507	17,535	18,168	20,09	21,955	24,352	26,124
9	1,735	2,7	12,242	14,684	16,919	19,023	19,679	21,666	23,589	26,056	27,877
10	2,156	3,247	13,442	15,987	18,307	20,483	21,161	23,209	25,188	27,722	29,588
11	2,603	3,816	14,631	17,275	19,675	21,92	22,618	24,725	26,757	29,354	31,264
12	3,074	4,404	15,812	18,549	21,026	23,337	24,054	26,217	28,3	30,957	32,909
13	3,565	5,009	16,985	19,812	22,362	24,736	25,472	27,688	29,819	32,535	34,528
14	4,075	5,629	18,151	21,064	23,685	26,119	26,873	29,141	31,319	34,091	36,123
15	4,601	6,262	19,311	22,307	24,996	27,488	28,259	30,578	32,801	35,628	37,697
16	5,142	6,908	20,465	23,542	26,296	28,845	29,633	32	34,267	37,146	39,252
17	5,697	7,564	21,615	24,769	27,587	30,191	30,995	33,409	35,718	38,648	40,79
18	6,265	8,231	22,76	25,989	28,869	31,526	32,346	34,805	37,156	40,136	42,312
19	6,844	8,907	23,9	27,204	30,144	32,852	33,687	36,191	38,582	41,61	43,82
20	7,434	9,591	25,038	28,412	31,41	34,17	35,02	37,566	39,997	43,072	45,315
21	8,034	10,283	26,171	29,615	32,671	35,479	36,343	38,932	41,401	44,522	46,797
22	8,643	10,982	27,301	30,813	33,924	36,781	37,659	40,289	42,796	45,962	48,268
23	9,26	11,689	28,429	32,007	35,172	38,076	38,968	41,638	44,181	47,391	49,728
24	9,886	12,401	29,553	33,196	36,415	39,364	40,27	42,98	45,559	48,812	51,179
25	10,52	13,12	30,675	34,382	37,652	40,646	41,566	44,314	46,928	50,223	52,62
26	11,16	13,844	31,795	35,563	38,885	41,923	42,856	45,642	48,29	51,627	54,052
27	11,808	14,573	32,912	36,741	40,113	43,195	44,14	46,963	49,645	53,023	55,476
28	12,461	15,308	34,027	37,916	41,337	44,461	45,419	48,278	50,993	54,411	56,892
29	13,121	16,047	35,139	39,087	42,557	45,722	46,693	49,588	52,336	55,792	58,301
30	13,787	16,791	36,25	40,256	43,773	46,979	47,962	50,892	53,672	57,167	59,703

Como o valor encontrado é superior ao valor crítico indicado na tabela para significância de 0,05 e 3 graus de liberdade, a hipótese H0 deve ser refutada, aceitando-se H1.

Este procedimento pode ser realizado através do software Bioestat® da seguinte forma:

1. No menu estatísticas, acessar à opção Qui-quadrado

2. No menu Qui-quadrado, acessar à opção Contingência L x C



A interpretação do resultado se dá então pela observação do (p) valor, que, como neste caso é menor do que 0,05 , determina que H0 deve ser refutada em prol de H1.

Teste Qui-quadrado de aderência

O teste Qui-quadrado para aderência é utilizado quando o pesquisador precisa determinar se os dados coletados estão de acordo com o que seria esperado dentro de um determinado modelo teórico, ou distribuição hipotética. Em outras palavras, este teste é utilizado para se confrontar os dados observados com os dados que seriam esperados.

Por exemplo, um pesquisador deseja investigar a distribuição de imaturos de um inseto aquático ao longo do gradiente de correnteza de um rio, para verificar se existe alguma preferência por determinados microhabitats. Desta forma, as hipóteses seriam:

H0: Não existe preferência de microhabitat

H1: Existe preferência de microhabitat

Os valores de abundância encontrados pelo pesquisador foram:

Velocidade	Abundância
1	30
2	10
3	120
4	40
5	37
6	12
7	8

Aplicando-se estes valores à fórmula do Qui-quadrado, chegamos ao valor 248,988 , o qual é superior ao valor crítico encontrado na tabela, de forma que H0 deve ser refutada e H1 aceita.

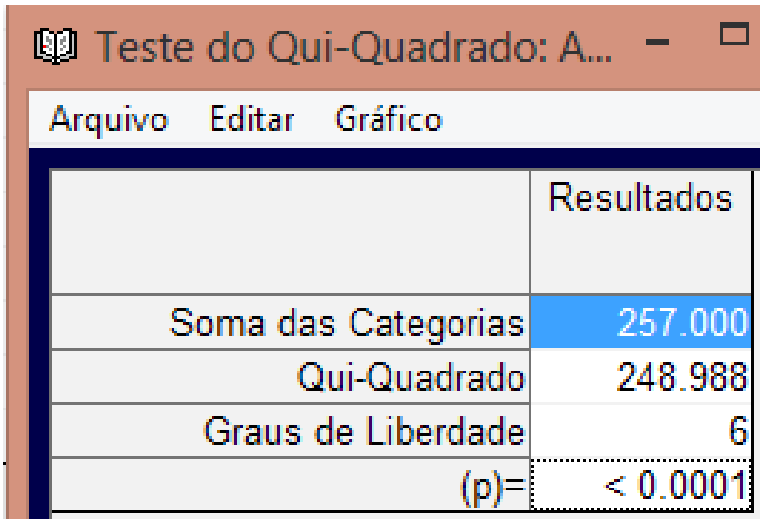
GL	0,995	0,975	0,2	0,1	0,05	0,025	0,02	0,01	0,005	0,002	0,001
1	3,93E-05	0,000982	1,642	2,706	3,841	5,024	5,412	6,635	7,879	9,55	10,828
2	0,01	0,0506	3,219	4,605	5,991	7,378	7,824	9,21	10,597	12,429	13,816
3	0,0717	0,216	4,642	6,251	7,815	9,348	9,837	11,345	12,838	14,796	16,266
4	0,207	0,484	5,989	7,779	9,488	11,143	11,668	13,277	14,86	16,924	18,467
5	0,412	0,831	7,289	9,236	11,07	12,833	13,388	15,086	16,75	18,907	20,515
6	0,676	1,237	8,558	10,645	12,592	14,449	15,033	16,812	18,548	20,791	22,458
7	0,989	1,69	9,803	12,017	14,067	16,013	16,622	18,475	20,278	22,601	24,322
8	1,344	2,18	11,03	13,362	15,507	17,535	18,168	20,09	21,955	24,352	26,124
9	1,735	2,7	12,242	14,684	16,919	19,023	19,679	21,666	23,589	26,056	27,877
10	2,156	3,247	13,442	15,987	18,307	20,483	21,161	23,209	25,188	27,722	29,588
11	2,603	3,816	14,631	17,275	19,675	21,92	22,618	24,725	26,757	29,354	31,264
12	3,074	4,404	15,812	18,549	21,026	23,337	24,054	26,217	28,3	30,957	32,909
13	3,565	5,009	16,985	19,812	22,362	24,736	25,472	27,688	29,819	32,535	34,528
14	4,075	5,629	18,151	21,064	23,685	26,119	26,873	29,141	31,319	34,091	36,123
15	4,601	6,262	19,311	22,307	24,996	27,488	28,259	30,578	32,801	35,628	37,697
16	5,142	6,908	20,465	23,542	26,296	28,845	29,633	32	34,267	37,146	39,252
17	5,697	7,564	21,615	24,769	27,587	30,191	30,995	33,409	35,718	38,648	40,79
18	6,265	8,231	22,76	25,989	28,869	31,526	32,346	34,805	37,156	40,136	42,312
19	6,844	8,907	23,9	27,204	30,144	32,852	33,687	36,191	38,582	41,61	43,82
20	7,434	9,591	25,038	28,412	31,41	34,17	35,02	37,566	39,997	43,072	45,315

Este procedimento deve ser realizado no bioestat® da seguinte forma:

1. No menu estatísticas, acessar à opção Qui-quadrado

2 No menu Qui-quadrado, acessar à opção Qui-quadrado de aderência

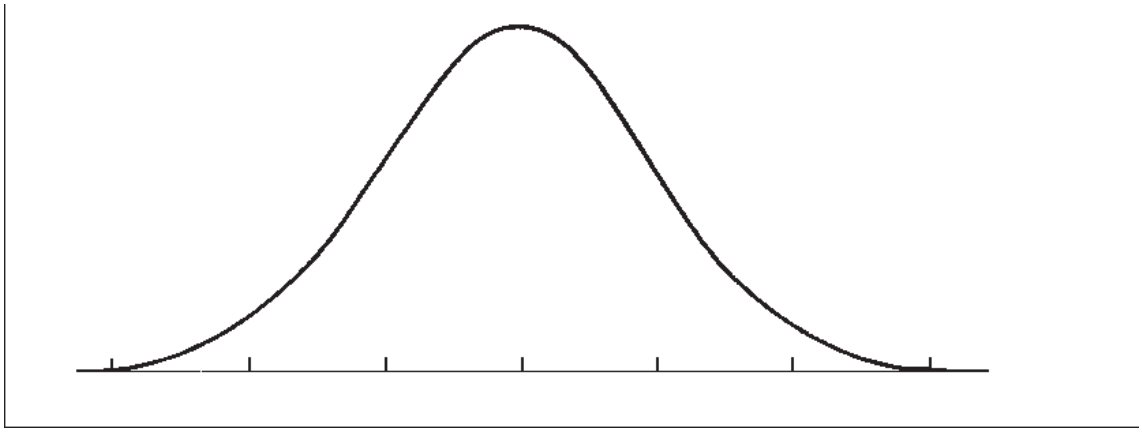
Desta forma, o (p) valor encontrado é inferior a 0,05 , indicando que H0 deve ser refutada e H1 aceita, conforme a figura a seguir demonstra:



CAPÍTULO 6
Testes de normalidade

Uma grande parte dos testes estatísticos são baseados na curva normal, e como tal, demandam normalidade dos dados para que sejam eficazes. Portanto, antes da execução de um teste que demande normalidade, se faz necessária a aplicação de um teste de normalidade para averiguar se os dados atendem à premissa do teste, e caso não atendam, o pesquisador deve optar por um equivalente não-paramétrico do teste.

Um teste de normalidade é essencialmente um teste de hipóteses, no qual, assim como no teste qui-quadrado de aderência, uma distribuição de dados é comparada com uma distribuição hipotética, no caso, a distribuição normal.



Desta forma, as hipóteses em um teste de normalidade são:

H0: Os dados não divergem significativamente da curva normal

H1: Os dados tem divergência significativa da curva normal

Embora existam vários testes de normalidade descritos na literatura, aqui serão abordados dois dos testes mais comuns e robustos, o teste de Lilliefors e o teste de Shapiro-Wilk.

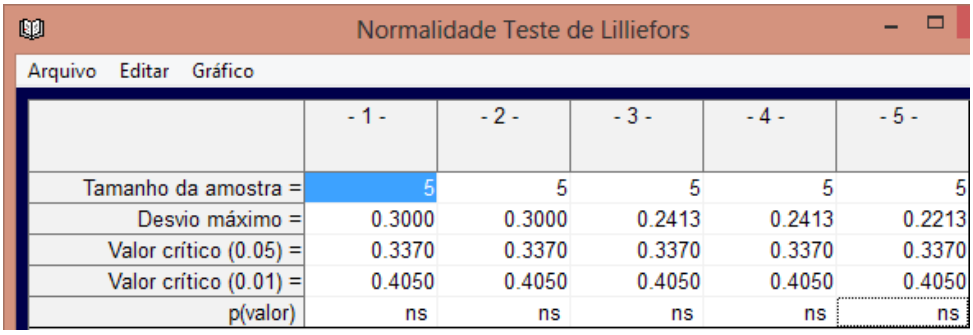
Teste de normalidade de Lilliefors

O teste de Lilliefors pode ser acessado no bioestat dentro da aba “estatísticas, em “normalidade”. Este teste permite a análise de múltiplas colunas ao mesmo tempo (k amostras).

Considere este conjunto de dados abaixo:

A	B	C	D	E
22	16	10	8	2
16	10	14	8	8
10	4	6	4	2
16	10	6	0	4
16	10	14	0	6

Pela aplicação do teste, temos que nenhum (p) valor é abaixo de 0,05 , de forma que a hipótese nula que a distribuição dos dados não difere da curva normal não pode ser refutada.



	- 1 -	- 2 -	- 3 -	- 4 -	- 5 -
Tamanho da amostra =	5	5	5	5	5
Desvio máximo =	0.3000	0.3000	0.2413	0.2413	0.2213
Valor crítico (0.05) =	0.3370	0.3370	0.3370	0.3370	0.3370
Valor crítico (0.01) =	0.4050	0.4050	0.4050	0.4050	0.4050
p(valor)	ns	ns	ns	ns	ns

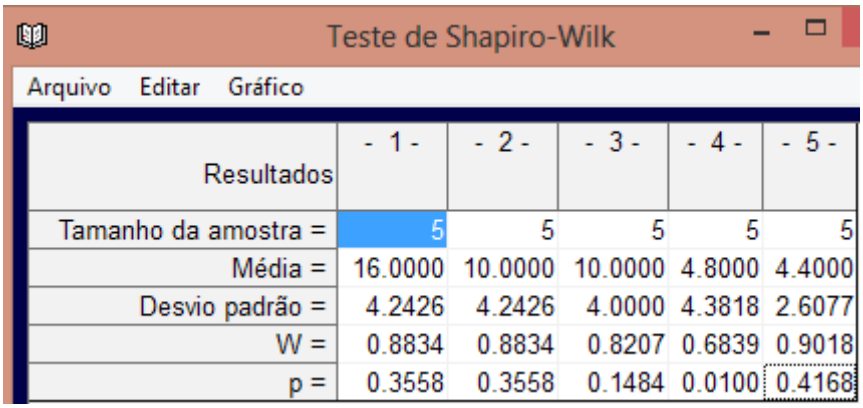
Teste de normalidade de Shapiro-Wilk

O teste de Shapiro-Wilk segue a mesma lógica do teste de Lilliefors, e assim como o teste anteriormente mencionado, é aplicável a múltiplas amostras.

Considere este conjunto de dados:

A	B	C	D	E
22	16	10	8	2
16	10	14	8	8
10	4	6	8	2
16	10	6	0	4
16	10	14	0	6

Aplicando o teste de Shapiro-Wilk, pode-se observar que o valor de (p) para a coluna D é significativo, ou seja, nesse caso a hipótese nula deve ser refutada: a distribuição de dados nesta coluna não é normal.



Resultados	- 1 -	- 2 -	- 3 -	- 4 -	- 5 -
Tamanho da amostra =	5	5	5	5	5
Média =	16.0000	10.0000	10.0000	4.8000	4.4000
Desvio padrão =	4.2426	4.2426	4.0000	4.3818	2.6077
W =	0.8834	0.8834	0.8207	0.6839	0.9018
p =	0.3558	0.3558	0.1484	0.0100	0.4168

CAPÍTULO 7
Testes de Hipóteses: Teste t de student

O teste t de student foi desenvolvido por William Sealy Gousset, então funcionário de uma cervejaria, com o intuito de monitorar a qualidade da cerveja tipo stout produzida. Como na ocasião este companhia não queria que suas concorrentes ficassem a par de seus métodos para controle de qualidade, Gousset adotou o pseudônimo "Student", o qual dá nome ao seu teste.

O teste t de student demanda normalidade dos dados, sendo assim o que é denominado um teste paramétrico. Desta forma, é necessário antes de sua aplicação a execução de um teste de normalidade. Este teste é aplicado para a comparação de médias entre duas amostras, as quais podem ser independentes ou relacionadas (pareadas). A hipótese a ser testada, por sua vez, pode ser unicaudal (quando a hipótese nula é $\bar{x} \leq \mu_0$ e a hipótese alternativa é $\bar{x} > \mu_0$) ou bicaudal (quando a hipótese nula é $\bar{x} = \mu_0$).

Assim como no teste qui-quadrado, o valor de t deve ser calculado através de uma fórmula matemática, onde X1 e X2 são as médias das amostras, n é o tamanho das amostras e comparado com o valor crítico na tabela ao nível de significância desejado.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{x_1 x_2} \cdot \sqrt{\frac{2}{n}}}$$

onde:

$$S_{x_1 x_2} = \sqrt{\frac{S_{x_1}^2 + S_{x_2}^2}{2}}$$

Teste t de student para amostras independentes

Quando as duas amostras sendo comparadas são constituídas de elementos diferentes, estas são denominadas amostras independentes.

Por exemplo, considere que um pesquisador deseja comparar a estatura média dos indivíduos de duas diferentes etnias indígenas. Neste caso, o pesquisador precisaria de uma amostra de cada etnia, assim sendo as duas amostras constituídas de diferentes indivíduos. Desta forma, as hipóteses seriam:

H0: Não existem diferenças entre as estaturas médias das duas etnias

H1: Existem diferenças entre as estaturas médias das duas etnias

Assim, foram medidos 10 indivíduos de cada diferente etnia, sendo construída a seguinte tabela:

Etnia 1	Etnia 2
1,5	1,61
1,62	1,57
1,52	1,62
1,67	1,6
1,45	1,58
1,6	1,46
1,66	1,63
1,56	1,64
1,55	1,6
1,48	1,57

Ao calcular o valor de t pela fórmula, é obtido -0,9388 , o qual é inferior ao valor crítico na tabela bicaudal para 18 graus de liberdade para significância de 0,05:

α (1 cauda)	0,05	0,025	0,01	0,005	0,0025	**0,001**	**0,0005**
α (2 caudas)	0,1	0,05	0,02	0,01	0,005	0,002	0,001
GL							
1	63.138	127.065	318.193	636.551	1.273.447	318.493	636.045
2	2.92	43.026	69.646	99.247	140.887	223.276	315.989
3	23.534	31.824	45.407	58.408	74.534	102.145	129.242
4	21.319	27.764	3.747	46.041	55.976	71.732	86.103
5	2.015	25.706	3.365	40.322	47.734	58.934	68.688
6	19.432	24.469	31.426	37.074	43.168	52.076	59.589
7	18.946	23.646	2.998	34.995	40.294	47.852	54.079
8	18.595	2.306	28.965	33.554	38.325	45.008	50.414
9	18.331	22.621	28.214	32.498	36.896	42.969	47.809
10	18.124	22.282	27.638	31.693	35.814	41.437	45.869
11	17.959	2.201	27.181	31.058	34.966	40.247	44.369
12	17.823	21.788	2.681	30.545	34.284	39.296	43.178
13	17.709	21.604	26.503	30.123	33.725	3.852	42.208
14	17.613	21.448	26.245	29.768	33.257	37.874	41.404

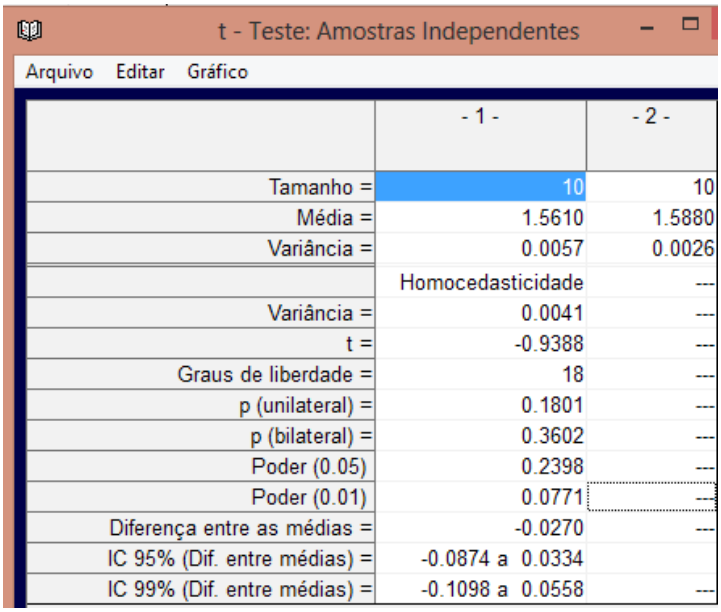
15	1.753	21.314	26.025	29.467	3.286	37.328	40.728
16	17.459	21.199	25.835	29.208	3.252	36.861	4.015
17	17.396	21.098	25.669	28.983	32.224	36.458	39.651
18	17.341	21.009	25.524	28.784	31.966	36.105	39.216
19	17.291	2.093	25.395	28.609	31.737	35.794	38.834
20	17.247	2.086	2.528	28.454	31.534	35.518	38.495

Desta forma, H0 não pode ser refutada. Pelo software Bioestat® , o procedimento seria o seguinte:

1 No menu estatísticas, acessar à opção duas amostras independentes

2 No menu Qui-quadrado, acessar à opção teste t dados amostrais

Assim, o (p) valor indicado seria 0,36 para um teste bicaudal, valor maior que 0,05 , sendo desta forma impossível se refutar H0.



	- 1 -	- 2 -
Tamanho =	10	10
Média =	1.5610	1.5880
Variância =	0.0057	0.0026
	Homocedasticidade	---
Variância =	0.0041	---
t =	-0.9388	---
Graus de liberdade =	18	---
p (unilateral) =	0.1801	---
p (bilateral) =	0.3602	---
Poder (0.05)	0.2398	---
Poder (0.01)	0.0771	---
Diferença entre as médias =	-0.0270	---
IC 95% (Dif. entre médias) =	-0.0874 a 0.0334	
IC 99% (Dif. entre médias) =	-0.1098 a 0.0558	---

Equivalente não paramétrico do Teste T de amostras independentes: Teste de Mann-Whitney

Quando alguma das amostras apresentar distribuição não-paramétrica, o teste t não pode ser aplicado, devendo ser aplicado o teste de Mann-Whitney, cuja interpretação dos resultados se dá da mesma forma que o teste t.

Teste t de student para amostras relacionadas

Quando as duas amostras sendo comparadas são constituídas dos mesmos elementos em dois momentos distintos, como por exemplo, um mesmo conjunto de indivíduos antes e depois de um determinado tratamento, estas são denominadas amostras relacionadas, ou pareadas.

A fórmula do teste t para amostras relacionadas, na qual d são as diferenças entre as amostras e n é o tamanho das amostras, é:

$$t = \frac{\Sigma d}{\sqrt{\frac{n\left(\Sigma d^2\right) - \left(\Sigma d\right)^2}{n-1}}}$$

Por exemplo, considere que um pesquisador deseja investigar o efeito de uma dieta sobre o peso de dez pacientes, e para isso, realiza sua pesagem antes do início da dieta e depois de três meses de dieta.

Neste caso, as hipóteses seriam:

H0: A dieta não influencia o peso

H1: A dieta tem efeito sobre o peso

Os pesos encontrados foram os seguintes:

Antes	Depois
60	55
52	51
55	51
56	50
50	49
38	45
52	51
54	52
52	51
50	48

Ao lançar esses valores na fórmula, é obtido o valor de 4,4736 , superior ao valor crítico na tabela para 9 graus de liberdade, de forma que H0 deve ser refutada, em prol de H1.

α (1 cauda)	0,05	0,025	0,01	0,005	**0,0025**	**0,001**	**0,0005**
α (2 caudas)	0,1	0,05	0,02	0,01	0,005	0,002	0,001
GL							
1	63.138	127.065	318.193	636.551	1.273.447	318.493	636.045
2	2.92	43.026	69.646	99.247	140.887	223.276	315.989
3	23.534	31.824	45.407	58.408	74.534	102.145	129.242
4	21.319	27.764	3.747	46.041	55.976	71.732	86.103
5	2.015	25.706	3.365	40.322	47.734	58.934	68.688
6	19.432	24.469	31.426	37.074	43.168	52.076	59.589
7	18.946	23.646	2.998	34.995	40.294	47.852	54.079
8	18.595	2.306	28.965	33.554	38.325	45.008	50.414
9	18.331	22.621	28.214	32.498	36.896	42.969	47.809
10	18.124	22.282	27.638	31.693	35.814	41.437	45.869
11	17.959	2.201	27.181	31.058	34.966	40.247	44.369
12	17.823	21.788	2.681	30.545	34.284	39.296	43.178
13	17.709	21.604	26.503	30.123	33.725	3.852	42.208
14	17.613	21.448	26.245	29.768	33.257	37.874	41.404
15	1.753	21.314	26.025	29.467	3.286	37.328	40.728
16	17.459	21.199	25.835	29.208	3.252	36.861	4.015
17	17.396	21.098	25.669	28.983	32.224	36.458	39.651
18	17.341	21.009	25.524	28.784	31.966	36.105	39.216
19	17.291	2.093	25.395	28.609	31.737	35.794	38.834
20	17.247	2.086	2.528	28.454	31.534	35.518	38.495

No software Bioestat®, o procedimento é o seguinte:

1 No menu estatísticas, acessar à opção duas amostras relacionadas

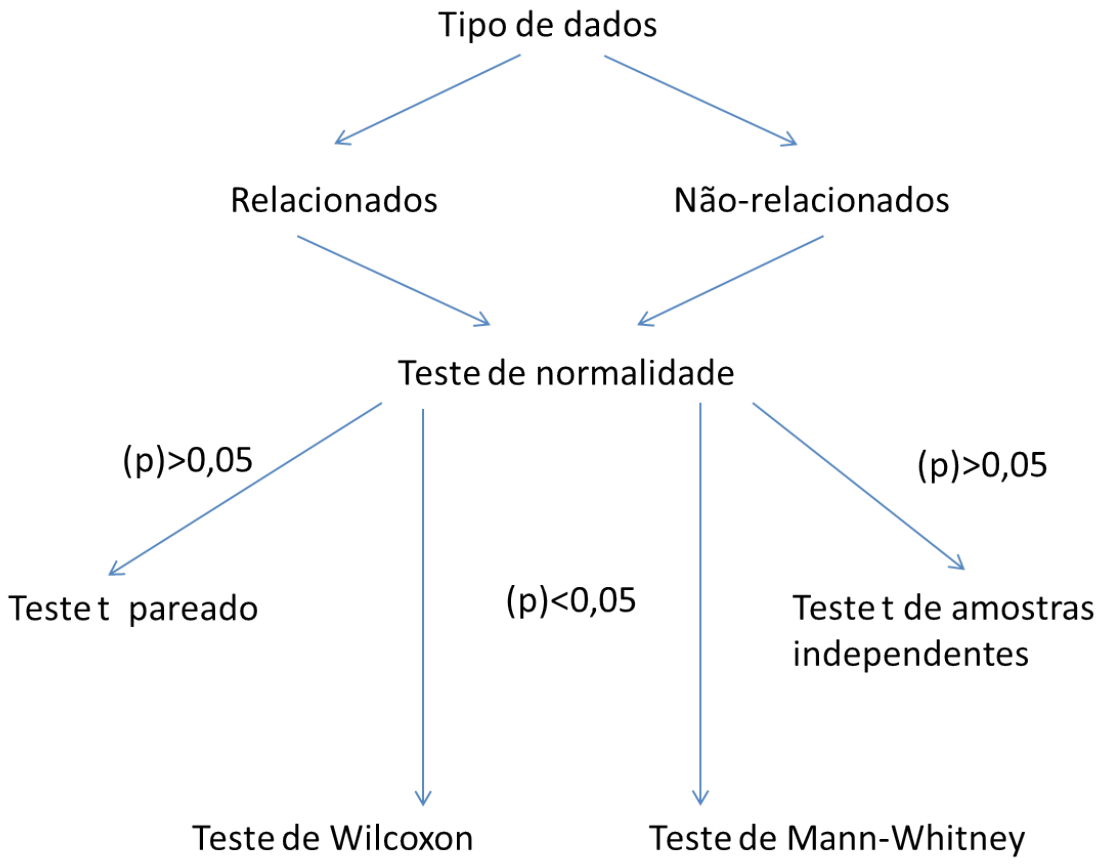
2 No menu Qui-quadrado, acessar à opção teste t dados amostrais

(t)=	4.4736	---
Graus de Liberdade	9	---
(p) unilateral =	0.0008	---
(p) bilateral =	0.0015	---
IC (95%)	1.2854 a 3.9146	---
IC (99%)	0.7111 a 4.4889	

Desta forma, o (p) valor de 0,0015 indica que H0 deve ser refutada.

Equivalente não paramétrico do Teste T de amostras relacionadas: Teste de Wilcoxon

Quando alguma das amostras apresentar distribuição não-paramétrica, o teste t não pode ser aplicado, devendo ser aplicado o teste de Wilcoxon, cuja interpretação dos resultados se dá da mesma forma que o teste t.



CAPÍTULO 8
Testes de Hipóteses: Análise de Variância

A análise de variância (ANOVA) é a técnica paramétrica mais utilizada para a comparação de médias de grupos. Existem muitos diferentes delineamentos experimentais que podem ser analisados a partir de diferentes tipos de anova, entretanto neste livro serão abordados três tipos de análise de variância: de um critério entre grupos, de um critério de medidas repetidas e dois critérios.

Em uma anova, várias observações de uma variável medida são realizadas para cada valor de uma variável nominal, como por exemplo, parasitemia no sangue de indivíduos de diferentes regiões endêmicas e hiperendêmicas em um país.

A análise de variância, assim como os testes de hipóteses anteriormente apresentados, consiste do cálculo de um valor através da fórmula abaixo (na qual MSA é o quadrado médio entre grupos e MSerror é o quadrado médio intragrupos), o qual deve ser comparado com o valor crítico em uma tabela ao nível de significância desejado para que seja determinado se a hipótese nula pode ser refutada.

$$f = \frac{\mathrm{MS_A}}{\mathrm{MS_{Error}}}$$

Uma vez que a hipótese nula tenha sido refutada, é aplicado um teste *post hoc*, ou seja, um teste aplicado *a posteriori*, cujo mais comum é o teste de Tukey, para comparação dos pares de tratamentos entre si.

$$\frac{M_1 - M_2}{\sqrt{MS_w\left(\frac{1}{n}\right)}}$$

Anova de um critério

O protocolo RCE (Riparian, Channel and Environmental Inventory) (Peterson, 1992) é uma importante ferramenta utilizada para determinar a qualidade de ambientes lóticos, isto é, de água corrente. Através deste protocolo, são atribuídos pontos para vários atributos qualitativos de um rio, o que gera uma pontuação final a qual deve ser comparada com valores em uma tabela a qual determina os diferentes estados de conservação, conforme a figura abaixo demonstra:

CLASSE	Escore	Avaliação de integridade	Ações recomendáveis
I	293-360	Excelente	Biomonitoramento e proteção do status existente
II	224-292	Muito Bom	Alterações selecionadas e monitoramento
III	154-223	Bom	Pequenas alterações necessárias
IV	86-153	Regular	Grandes alterações necessárias
V	16-85	Pobre	Reorganização estrutural completa

Considere então que um pesquisador decidiu realizar coletas de organismos aquáticos em cinco rios, cada um pertencente a uma das categorias apresentadas na tabela acima. Desta forma, o pesquisador constrói a seguinte planilha:

	I	II	III	IV	V
Sp1	11	8	5	4	1
Sp2	8	5	7	4	4
Sp3	5	2	3	2	1
Sp4	8	5	3	0	2
sp5	8	5	7	0	3

Desta forma, as hipóteses são:

H 0: Não existe influência do nível de conservação do ambiente sobre a biota

H 1: Existe influência do nível de conservação do ambiente sobre a biota

Pela fórmula, o valor de F calculado seria 8,2895. Como este valor é superior ao valor crítico na tabela para uma significância de 5% (p=0,05), a hipótese H0 deve ser refutada.

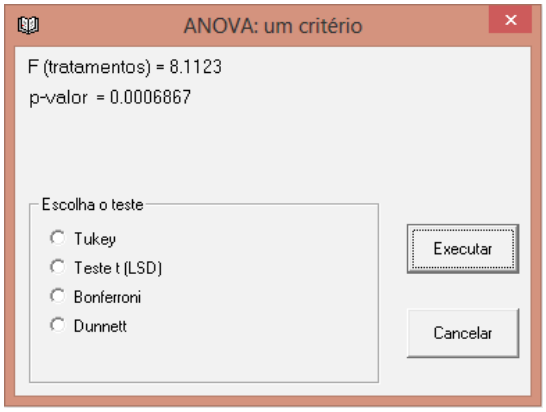
Graus de liberdade(Réplicas)	Graus de liberdade (tratamentos)								
	1	2	3	4	5	6	8	12	w
1	161	200	216	225	230	234	239	244	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.5
3	10.1	9.6	9.3	9.1	9.0	8.9	8.8	8.7	8.5
4	7.7	6.9	6.6	6.4	6.3	6.2	6.0	5.9	5.6
5	6.6	5.8	5.4	5.2	5.1	5.0	4.8	4.7	4.4
6	6.0	5.1	4.8	4.5	4.4	4.3	4.2	4.0	3.7
7	5.6	4.7	4.4	4.1	4.0	3.9	3.7	3.6	3.2
8	5.3	4.5	4.1	3.8	3.7	3.6	3.4	3.3	2.9
9	5.1	4.3	3.9	3.6	3.5	3.4	3.2	3.1	2.7
10	5.0	4.1	3.7	3.5	3.3	3.2	3.1	2.9	2.5

No software Bioestat®, o procedimento deve ser o seguinte:

1 No menu estatísticas, acessar à opção Análise de Variância

2 No menu Análise de variância, acessar à opção Anova: um critério

Ao executar a análise, como H0 é refutada devido ao valor significativo de (p), uma nova janela se abre solicitando a escolha do teste Post Hoc, na qual o teste de Tukey deve ser escolhido.



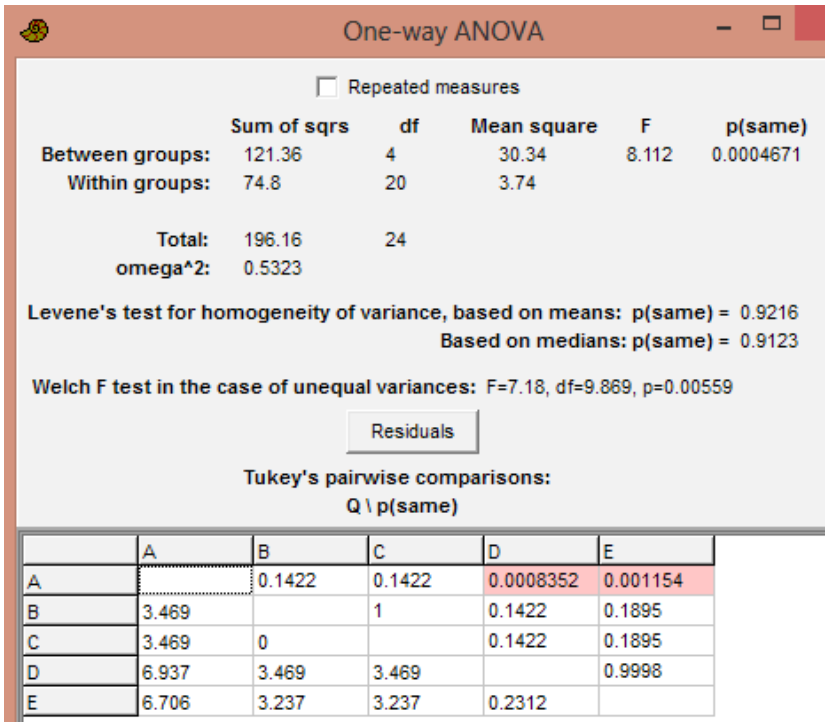
Uma vez executado o teste de Tukey, as comparações serão apresentadas, sendo que aquelas nas quais H0 for verdadeiro são indicadas como ns (não significativo). Desta forma, concluímos que as diferenças estão entre as condições IV e V e a I.

Tukey:	Diferença	Q	(p)
Médias (1 a 2) =	3.0000	3.4687	ns
Médias (1 a 3) =	3.0000	3.4687	ns
Médias (1 a 4) =	6.0000	6.9375	< 0.01
Médias (1 a 5) =	5.8000	6.7062	< 0.01
Médias (2 a 3) =	0.0000	0.0000	ns
Médias (2 a 4) =	3.0000	3.4687	ns
Médias (2 a 5) =	2.8000	3.2375	ns
Médias (3 a 4) =	3.0000	3.4687	ns
Médias (3 a 5) =	2.8000	3.2375	ns
Médias (4 a 5) =	0.2000	0.2312	ns

No software Past©, o procedimento deve ser o seguinte:

1 – No menu “edit”, acessar à opção “select all”.

2 – No menu “Statistics”, acessar à opção “one-way anova”



No Past©, o valor de p da anova pode ser observado na parte superior da janela (0,0004114), e as comparações pelo teste de Tukey podem ser observadas na tabela na parte inferior, na qual as comparações nas quais foram detectadas diferenças significativas são grifadas em vermelho (A e D, A e E, ou seja, só existe diferença significativa entre I e IV/V).

Equivalente não paramétrico da Análise de Variância de um critério simples: Teste de Kruskal-Wallis

Quando alguma das amostras apresentar distribuição não-paramétrica, a anova de um critério não pode ser aplicada, devendo ser aplicado o teste de Kruskal-Wallis, cuja interpretação dos resultados se dá da mesma forma.

Anova de um critério de medidas repetidas

A anova de medidas repetidas é desenhada para a análise de dados pareados, isto é, relacionados. Neste caso, ao invés de ser realizada a comparação entre grupos, é realizada a comparação entre diferentes momentos para um mesmo grupo.

Considere que um pesquisador deseja acompanhar ao longo de cinco anos o crescimento de cinco indivíduos arbóreos pré-estabelecidos em uma área de replantio, tendo coletado os seguintes dados:

	1 ano	2 anos	3 anos	4 anos	5 anos
Árvore 1	2	8	8	16	22
Árvore 2	8	8	10	14	16
Árvore 3	2	4	6	8	10
Árvore 4	1	3	6	10	16
Árvore 5	6	7	10	14	16

As hipóteses serão respectivamente:

H_0: Não há diferença significativa nos crescimentos das árvores entre os anos

H_1: Existe diferença significativa nos crescimentos das árvores entre os anos

Na anova de medidas repetidas, a fórmula utilizada corresponde aos quadrados médios entre as medidas (não mais entre os grupos) e os quadrados médios dos erros:

$$F = \frac{MS_M}{MS_E}$$

Por esta fórmula, com 4 graus de liberdade nas medidas e 4 graus de liberdade nos tratamentos, chega-se ao valor 28,19 , o qual é superior ao valor crítico na tabela, sendo desta forma refutada a hipótese nula.

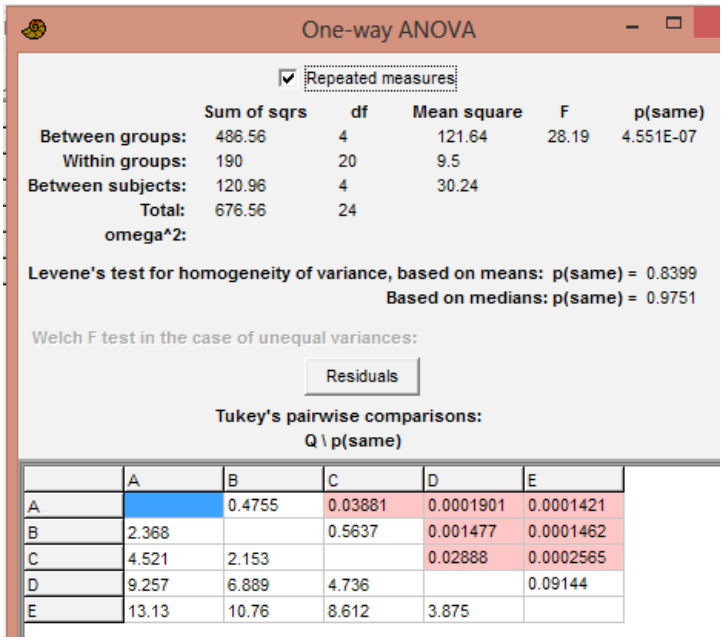
Esta análise não está disponível no pacote Bioestat®, desta forma, será indicado seu procedimento no pacote PAST©.

O procedimento deve ser o seguinte:

1. No menu “edit”, acessar à opção “select all”

2. No menu “Statistics”, selecionar a opção “one-way anova”

1. Na janela de resultado, marcar a opção “repeated measures”



Desta forma, os resultados indicam diferença significativa de crescimento entre o primeiro ano e terceiro, quarto e quinto anos; entre o segundo ano e quarto e quinto anos; entre o terceiro ano e quarto e quinto anos.

Equivalente não paramétrico da Análise de Variância de medidas repetidas: Teste de Friedmann

Quando alguma das amostras apresentar distribuição não-paramétrica, a anova de medias repetidas não pode ser aplicada, devendo ser aplicado o teste de Friedmann, cuja interpretação dos resultados se dá da mesma forma.

Anova de dois critérios

Em alguns casos, o pesquisador pode testar a influência de dois fatores, ao invés de um, na sua distribuição de dados. Neste caso, os dados apresentados na forma de colunas são denominados tratamentos, e os dados apresentados na forma de linhas são denominados blocos.

Considere que um pesquisador deseja estudar o comportamento hematofágico de mosquitos de quatro diferentes espécies ao longo do período de um ano. Para tal, o pesquisador quantificou as taxas de picadas de cada espécie em sua área de estudos em cada uma das quatro estações do ano, construindo a seguinte planilha:

	spA	spB	spC	spD

Primavera	25	25	45	40
Verão	35	40	80	50
Outono	30	25	50	40
Inverno	15	10	20	15

Desta forma, em uma Anova de dois critérios existem duas hipóteses a serem testadas, existindo assim dois H0 e dois H1.

Para os tratamentos (colunas):

H_0: Não existe diferença entre os padrões sazonais de atividade das espécies

H_1: Existe diferença nos padrões sazonais de atividade entre as espécies

Para os blocos (linhas):

H_0: Não existe diferença no padrão de atividade entre as estações

H_1: Existe diferença no padrão de atividade entre as estações

Desta forma, o valor de F encontrado para os tratamentos é 8,7009 e o valor de F para os blocos é 15,8785 , em ambos os casos, superior ao valor crítico na tabela, assim devendo ser refutadas ambas as hipóteses nulas:

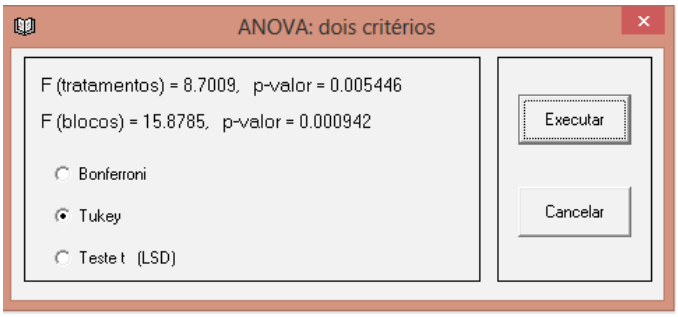
Graus de liberdade (Blocos)	Graus de Liberdade (Tratamentos							
	1	2	3	4	5	6	8	12
1	161	200	216	225	230	234	239	244
2	18,5	19	19,2	19,3	19,3	19,3	19,4	19,4
3	10,1	9,6	9,3	9,1	9	8,9	8,7	8,7
4	7,7	6,9	6,6	6,4	6,3	6,2	5,9	5,9

No software Bioestat®, o procedimento deve ser o seguinte:

1. No menu estatísticas, acessar à opção Análise de Variância

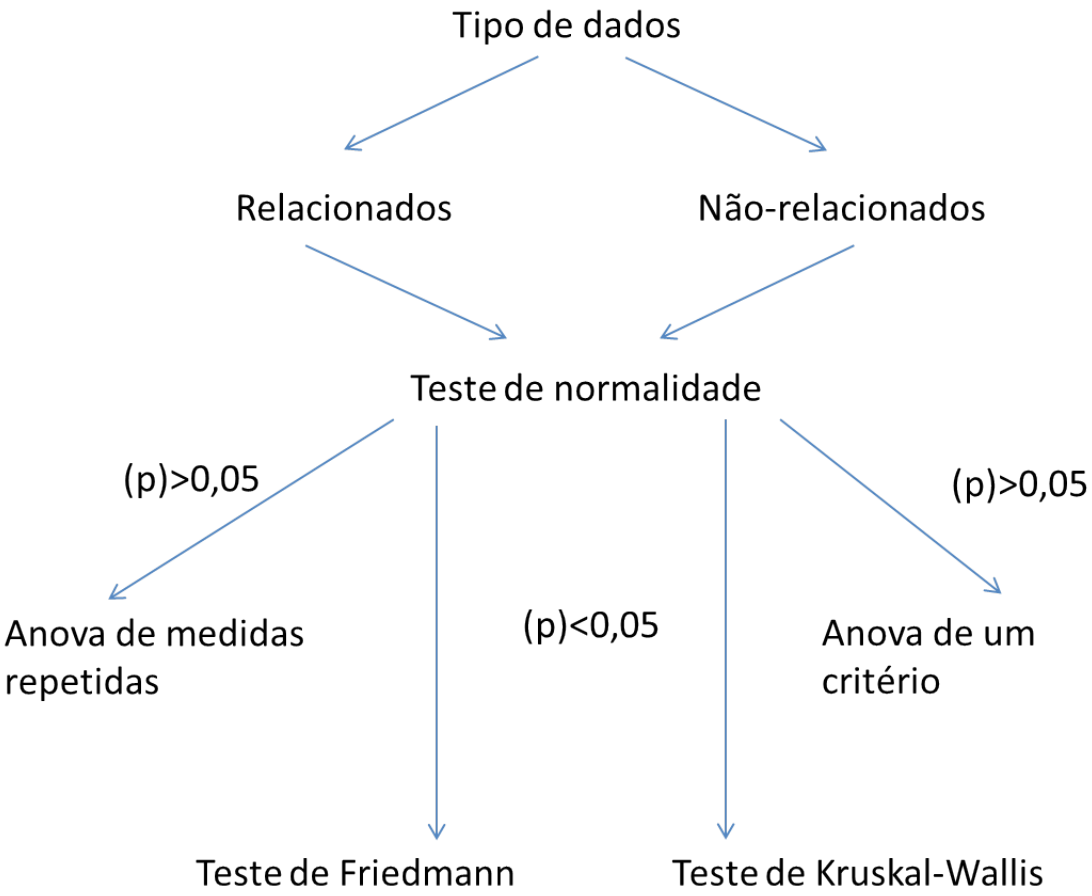
2. No menu Análise de variância, acessar à opção Anova: dois critérios

Ao executar este procedimento, o software indicará que foi detectada diferença significativa tanto entre os tratamentos quanto entre os blocos, assim solicitando a escolha de um teste *Post Hoc* para identificar entre quais pares estão essas diferenças:



Uma vez executado o teste de Tukey, é possível se estabelecer que as diferenças nos padrões sazonais estão entre a sp A e a sp C e entre a espécie A e a espécie D (tratamentos); a diferença entre as estações estão entre primavera e verão, primavera e inverno, verão e inverno e outono e inverno.

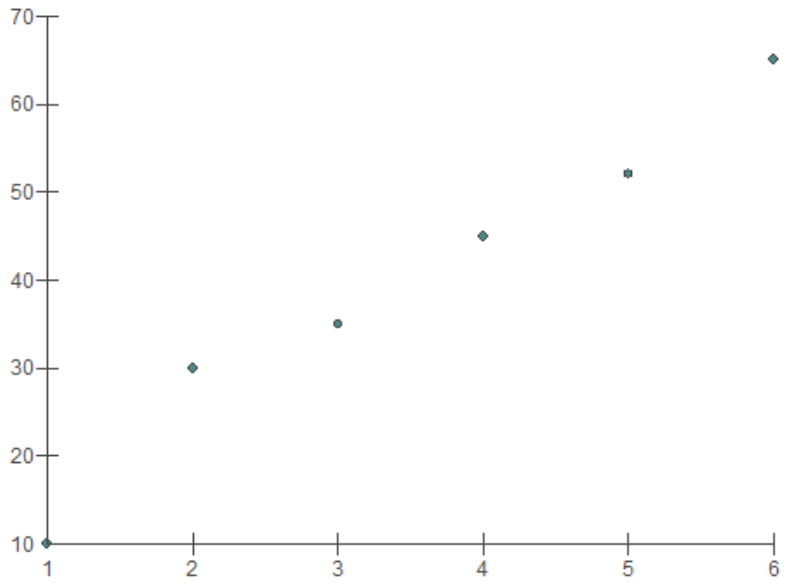
Médias (tratamentos):		
Média (Coluna 1) =	26.2500	
Média (Coluna 2) =	25.0000	
Média (Coluna 3) =	48.7500	
Média (Coluna 4) =	36.2500	
Tukey	Q	(p)
Médias (1 a 2) =	0.3349	ns
Médias (1 a 3) =	6.0280	< 0.01
Médias (1 a 4) =	2.6791	ns
Médias (2 a 3) =	6.3629	< 0.01
Médias (2 a 4) =	3.0140	ns
Médias (3 a 4) =	3.3489	ns
Médias (blocos):		
Média (Linha 1) =	33.7500	
Média (Linha 2) =	51.2500	
Média (Linha 3) =	36.2500	
Média (Linha 4) =	15.0000	
Tukey	Q	(p)
Médias (1 a 2) =	4.6884	< 0.05
Médias (1 a 3) =	0.6698	ns
Médias (1 a 4) =	5.0233	< 0.05
Médias (2 a 3) =	4.0186	ns
Médias (2 a 4) =	9.7117	< 0.01
Médias (3 a 4) =	5.6931	< 0.05


Tipo de dados
Relacionados
Não-relacionados
Teste de normalidade
(p)>0,05
(p)<0,05
(p)>0,05
Anova de medidas repetidas
Anova de um critério
Teste de Friedmann
Teste de Kruskal-Wallis

CAPÍTULO 9
Introdução à correlação linear

Muitas vezes um pesquisador pode ter como objetivo investigar a relação de covariação de duas ou mais variáveis, sendo para tanto utilizados coeficientes de correlação. Entretanto, é importante ressaltar que a correlação nem sempre implica em causalidade.

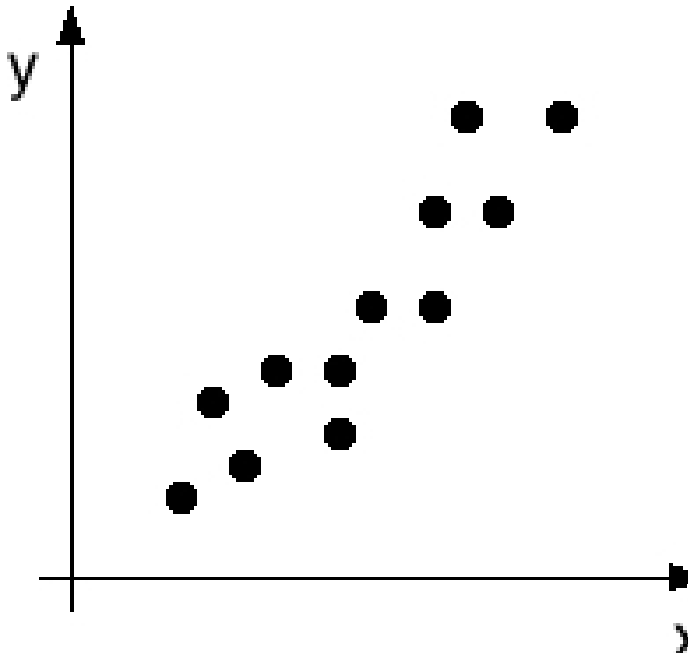
A forma de representação gráfica para uma correlação é denominada diagrama de dispersão, e é representada pelos pares ordenados (x,y), conforme a figura abaixo:



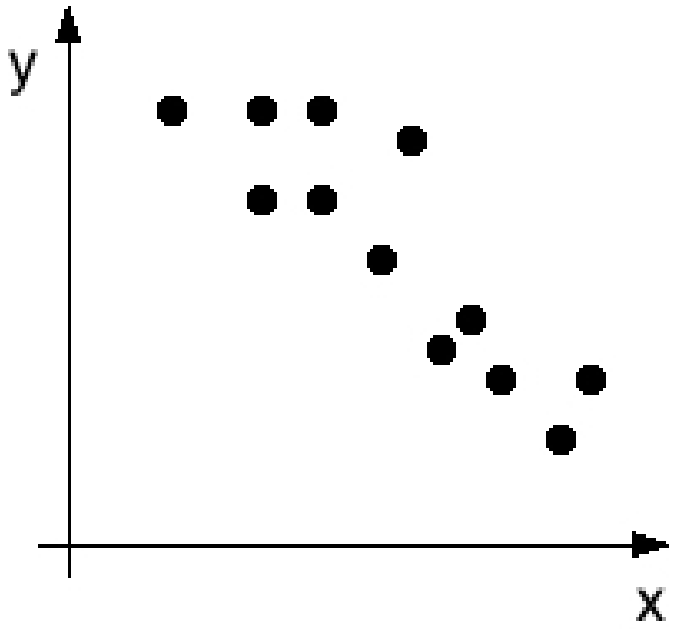
O coeficiente de correlação mais comumente utilizado é o coeficiente de correlação linear de Pearson, o qual pode assumir valores entre -1 e +1, conforme indica a fórmula abaixo:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

Esta variação entre valores positivos e negativos indica a natureza da correlação: em uma correlação positiva, as variáveis em estudo variam em um mesmo sentido, isto é, o aumento do valor de uma implicará no aumento do valor da outra, assim como a diminuição de uma também resultará na diminuição da outra, conforme a figura abaixo:



Já na correlação negativa, a variação ocorre em sentidos inversos, implicando o aumento do valor de uma variável na diminuição do valor da outra, e vice-versa, conforme pode ser observado na figura abaixo.



Por exemplo, considere que um pesquisador deseja estudar a relação entre as abundâncias de dois pequenos mamíferos. Para tanto, ele realiza coletas ao longo do período de um ano, para comparar as abundâncias de ambos, chegando a esses valores:

spA	spB
10	6
8	25
5	34
26	4
2	45
31	5
20	4
11	7
12	14
10	11
8	14

Neste caso, temos as seguintes hipóteses:

H_0: Não existe relação entre as variáveis

H_1: Existe relação entre as variáveis

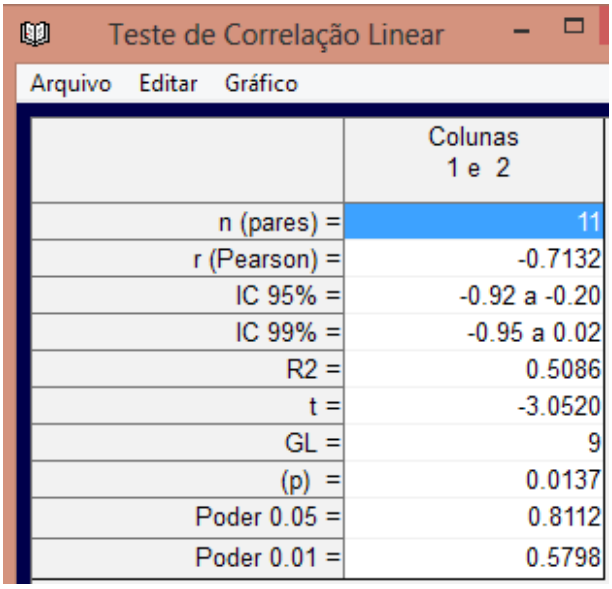
No bioestat®, o procedimento deve ser o seguinte:

1. No menu estatísticas, acessar à opção Correlação

2. No menu Correlação, acessar à opção Coef. Correlação de Pearson

Após realizado este procedimento, será aberta uma nova janela informando o (p) valor de 0,0137 , o qual como é inferior a 0,05 determina que a hipótese nula deve ser refutada, assim estabelecendo que existe relação entre as duas variáveis, e o coeficiente de Pearson no valor de -0,7132 , o qual indica que a correlação é negativa e moderadamente forte.

Estes dados podem sugerir que uma espécie seja presa da outra, entretanto, não necessariamente se trata de uma relação de causalidade: ambas as espécies podem na verdade estar sendo influenciadas por um terceiro fator, invisível neste delineamento experimental.

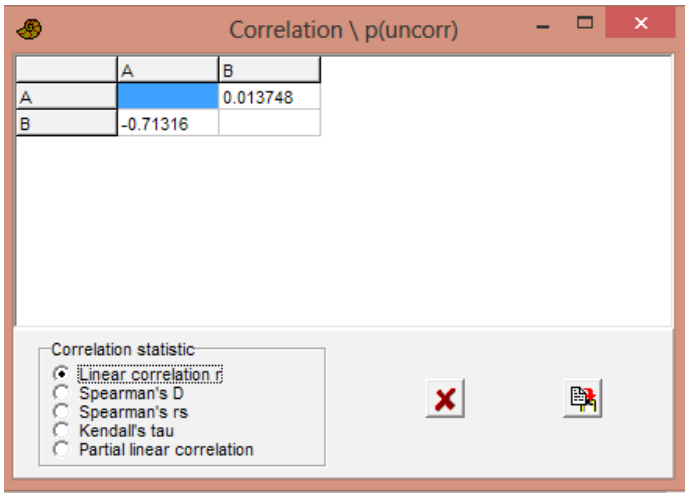


No Past©, o procedimento deve ser o seguinte:

1. No menu “edit”, acessar à opção “select all”

2. No menu “Statistics”, selecionar a opção “correlation table”

Uma vez executado este procedimento, uma nova janela surgirá, indicando o valor de Pearson no canto inferior esquerdo, e o valor de (p) no canto superior direito:



CAPÍTULO 10
Noções de regressão e ajustamento de curvas

Em muitas ocasiões, estudos prévios podem permitir ao pesquisador ter uma noção *a priori* de quais variáveis são as causas (variáveis independentes), e quais são os efeitos (variáveis dependentes). Nesses casos, pode ser interessante buscar um entendimento maior do padrão de resposta das variáveis dependentes às flutuações da variável independente, de forma a se construir um modelo preditivo.

Nestes casos, são aplicados os modelos de regressão, os quais consistem em gerar equações que permitem a partir de um valor de x inferir qual seria o valor de y esperado.

A regressão linear é o tipo mais comum de regressão, no qual os dados são ajustados à uma equação de reta.

Considere que pesquisadores desejam estudar o padrão de resposta de um inseto aquático à variações na correnteza, com o intuito de desenvolver um modelo preditivo da deriva em sistemas lóticos:

Tempo	Correnteza
5	4
8	4
8	4
7	4
9	4
7	3
9	3
8	3
9	3
10	3
10	3
11	2
10	2
12	2
9	2

Para estes dados, as hipóteses são:

H_0: Não existe relação entre a correnteza e o tempo de resposta

H_1: Existe relação entre a correnteza e o tempo de resposta

A fórmula geral da equação da reta é dada por y=a + bx , onde:

$$a = \frac{\sum y - b \sum x}{n}$$

$$b = \frac{n \sum (xy) - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

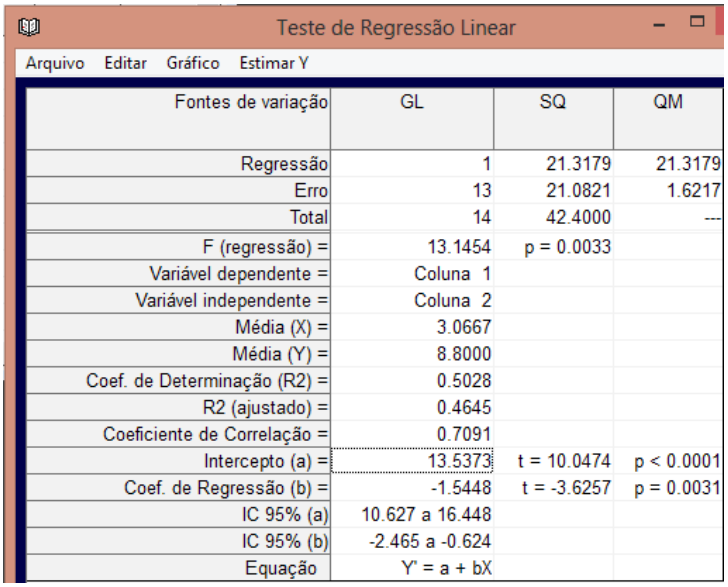
Desta forma, temos que a=13,5373 e b=-1,5448, logo Y=13,5373 -1,5488x.

Assim, se o pesquisador desejar uma estimativa de qual é o tempo de resposta para uma correnteza de 4 cm / s , o valor é dado por Y=13,5373 -1,5488 (4) = 7,3581.

No software Bioestat®, o procedimento deve ser o seguinte:

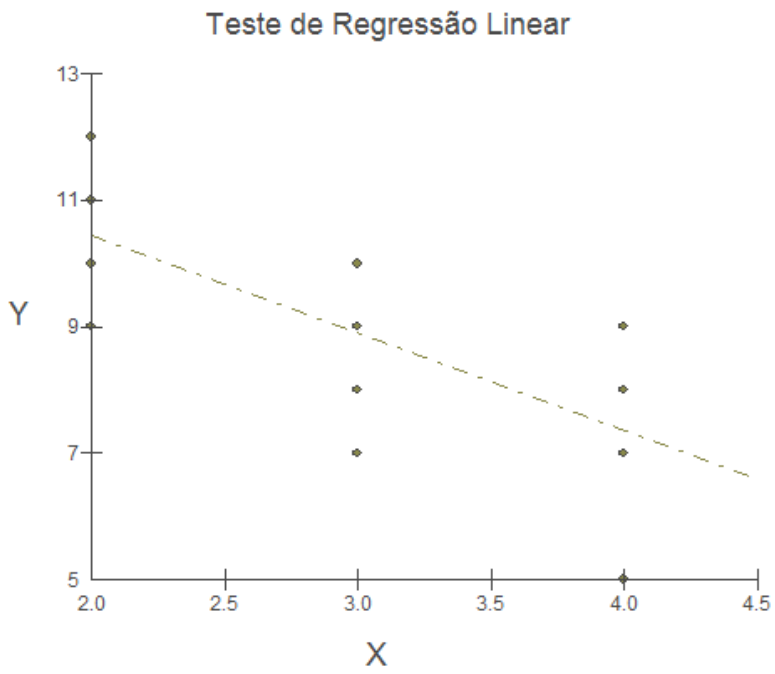
1. No menu estatísticas, acessar à opção Regressão

2. No menu Regressão, acessar à opção Regressão Linear Simples



Fontes de variação	GL	SQ	QM
Regressão	1	21.3179	21.3179
Erro	13	21.0821	1.6217
Total	14	42.4000	---
F (regressão) =	13.1454	p = 0.0033	
Variável dependente =	Coluna 1		
Variável independente =	Coluna 2		
Média (X) =	3.0667		
Média (Y) =	8.8000		
Coef. de Determinação (R2) =	0.5028		
R2 (ajustado) =	0.4645		
Coeficiente de Correlação =	0.7091		
Intercepto (a) =	13.5373	t = 10.0474	p < 0.0001
Coef. de Regressão (b) =	-1.5448	t = -3.6257	p = 0.0031
IC 95% (a)	10.627 a 16.448		
IC 95% (b)	-2.465 a -0.624		
Equação	Y' = a + bX		

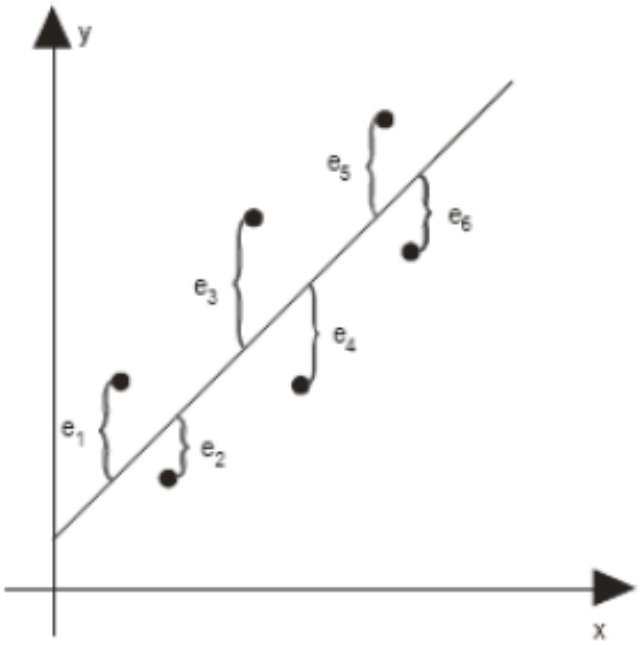
Na janela de resultados, é possível identificar que o valor de (p) é significativo (0,0033), assim sendo refutada a hipótese nula. Na mesma janela é possível se solicitar o gráfico de dispersão:



Ainda nesta janela é possível se estimar o valor de Y para um X conhecido, sendo importante destacar que quanto maior for o valor do coeficiente de determinação, maior é o poder preditivo da equação. Para a estimativa do valor Y, basta clicar em “estimar Y” e inserir o valor do X desejado.

Nem sempre o conjunto de dados permite um bom ajuste da reta, muitas vezes devido à presença de “outliers”, que são pontos de observação que fogem ao comportamento do restante dos dados. Para a identificação dos outliers, é necessário o uso de uma técnica denominada análise de resíduos, a qual identifica pontos que potencialmente foram gerados por algum artefato experimental.

O conceito de resíduo se refere à diferença entre os pontos estimados em uma regressão e os pontos de fato observados: é possível se notar, observando-se o gráfico de dispersão, que a reta apenas indica a tendência, de forma que os pontos nem sempre cairão sobre a reta, mas se o ajuste for bom, cairão próximos a ela.



A análise de resíduos identifica aqueles pontos cujos resíduos estão acima do esperado, a partir do cálculo do desvio padrão dos resíduos.

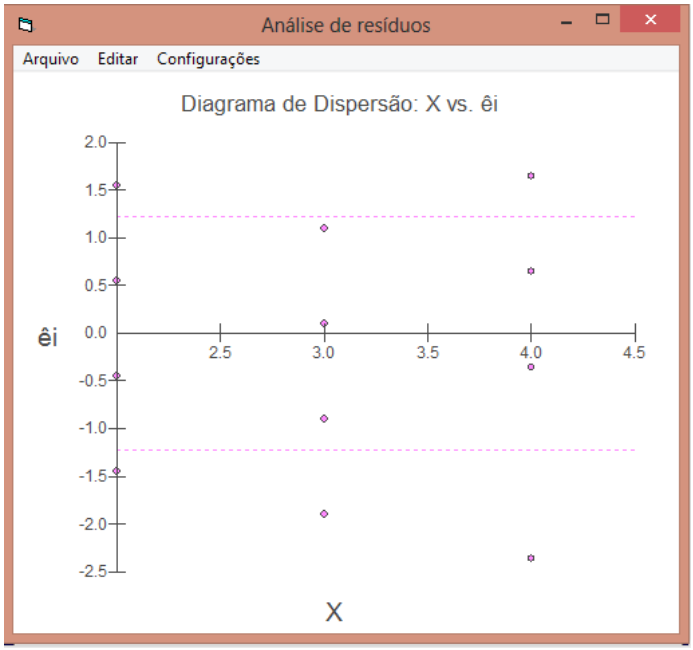
No Bioestat®, o procedimento deve ser o seguinte:

1. No menu estatísticas, acessar à opção Regressão

2. No menu Regressão, acessar à opção Ajustamento de curvas

3. A variável dependente deve ser a primeira a ser selecionada

4. Na janela de resultados, clicar na opção de criar gráfico



Neste gráfico, é possível se identificar cinco outliers, que são os pontos cujos resíduos são superiores a $\pm 1\sigma$ em torno da média residual (fora da linha tracejada). Se a regressão é novamente executada sem esses pontos, seu coeficiente de determinação irá aumentar de valor, conforme pode ser observado na figura a seguir:

Teste de Regressão Linear

Arquivo Editar Gráfico Estimar Y

Fontes de variação	GL	SQ	QM
Regressão	1	10.0000	10.0000
Erro	8	4.0000	0.5000
Total	9	14.0000	---
F (regressão) =	20.0000	p = 0.0024	
Variável dependente =	Coluna 1		
Variável independente =	Coluna 2		
Média (X) =	3.1000		
Média (Y) =	9.0000		
Coef. de Determinação (R2) =	0.7143		
R2 (ajustado) =	0.6786		
Coeficiente de Correlação =	0.8452		
Intercepto (a) =	13.4286	t = 13.2276	p < 0.0001
Coef. de Regressão (b) =	-1.4286	t = -4.4721	p = 0.0021
IC 95% (a)	11.088 a 15.770		
IC 95% (b)	-2.165 a -0.692		
Equação	Y' = a + bX		

Entretanto, nem sempre a resposta é linear. Considere que um pesquisador tem como objetivo estabelecer a relação numérica entre a precipitação e a abundância de mosquitos *Aedes aegypti*, tendo como objetivo criar um modelo que o permita prever a abundância de mosquitos esperada com base em dados de previsão meteorológica. Para tanto, ao longo de alguns meses, o pesquisador realizou coletas semanais de mosquitos, e montou a seguinte planilha comparando os dados de precipitação com seus dados de abundância:

	Precipitação	Densidade
Jan	1141	1100
Fev	1053	1050
Mar	1033	1000
Abr	1374	1200
Mai	856	80
Jun	804	55
Jul	564	30
Ago	505	20
Set	871	90
Out	882	92
Nov	956	95
dez	1690	1300

Desta forma, as hipóteses são:

H_0: A precipitação não tem influência sobre a densidade de mosquitos

H_1: A precipitação tem influência sobre a densidade de mosquitos

O primeiro passo a ser realizado é identificar a qual modelo de regressão este conjunto de dados se adequa, uma vez que a resposta da variável dependente nem sempre será linear. Para tanto, é realizada a análise denominada ajustamento de curvas, na qual vários modelos de regressão são testados para encontrar o melhor ajuste.

No bioestat®, o procedimento deve ser o seguinte:

1. No menu estatísticas, acessar à opção Regressão

2. No menu Correlação, acessar à opção Ajustamento de curvas

Ajustamento de Curvas

Arquivo Editar Gráfico

	Regressão Linear	Regressão Exponencial	Regressão Logarítmica	Regressão Geométrica
Tamanho da amostra =	12	12	12	12
Intercepto (a) =	-876.7336	3.2283	-5525.7879	0.0000
Coef. regressão (b) =	14.2282	0.0423	1332.8953	4.2037
Coef. determinação (R2) =	70.25%	74.14%	66.21%	78.71%
Média (X) =	97.4167	97.4167	92.5574	92.5574
Média (Y) =	509.3333	198.4977	509.3333	198.4977
Var. independente =	Coluna 1	Coluna 1	Coluna 1	Coluna 1
Var. dependente =	Coluna 2	Coluna 2	Coluna 2	Coluna 2
Equação =	Y' = a + bX	Y' = a * e^(bX)	Y' = a + b * ln(X)	Y' = a * X^b
Graus de liberdade =	10	10	10	10
(p) =	0.0007	0.0003	0.0013	0.0001

Após esse procedimento, uma nova janela será aberta com os parâmetros de cada regressão. Nesta janela, é importante primeiro se observar os valores de (p): neste caso, todos os valores são abaixo de 0,05 , logo, H0 deve ser refutada. O segundo passo é observar os valores do coeficiente de determinação: o maior valor de R2 indica o melhor modelo de regressão.

Desta forma, neste caso o modelo de regressão que melhor se ajusta é o modelo de regressão geométrica, seguido do modelo de regressão exponencial, regressão linear e regressão logarítmica.

A representação gráfica da regressão geométrica apresenta o formato de curva, conforme a figura abaixo: