Material de Apoio para

Análises Estatísticas

Parte 1: Estatística e interpretação de dados

Parte 2: Guia para execução das análises estatísticas

Goiânia– GO

Abril/ 2009

PARTE 1

ESTATÍSTICA E INTERPRETAÇÃO DE DADOS

Paulo De Marco Júnior

Departamento de Biologia Geral, Universidade Federal de Goiás

Adriano Pereira Paglia

Analista de Biodiversidade- Conservação International do Brasil

INTRODUÇÃO

O objetivo deste texto não é, nem de longe, ser um manual completo para guiar as suas atividades na área da análise de dados. Antes, deseja-se apresentar algumas idéias interessantes que possam desafiar a vontade de ser mais eficiente no uso destas ferramentas. A ênfase aqui é demonstrar que todos os testes estatísticos mantêm a mesma estrutura lógica e, portanto, podem ser facilmente entendidos.

POR QUE USAR ESTATÍSTICA

Considere o seguinte experimento: um pesquisador está interessado em avaliar o *status* de conservação de duas espécies filogeneticamente próximas. Tendo recursos limitados para ser gasto no manejo destas populações, ele considera a possibilidade de medir sua variabilidade populacional natural para escolher com qual delas vai gastar seus recursos. Aquela mais variável deve ser, a longo prazo, mais ameaçada de extinção por estocasticidade demográfica. O pesquisador escolhe utilizar estimativas do tamanho destas populações nos últimos 5 anos e encontra que a população A é mais variável que a B. Existe uma pergunta que gera toda a necessidade de serem utilizados métodos estatísticos: se outro pesquisador repetisse o experimento, qual a probabilidade de encontrar os mesmos resultados, a mesma conclusão?

Tratando-se de fenômenos biológicos, cuja natureza está ligada a múltiplas causas de variação, é possível que os resultados particulares observados não sejam repetidos. Isto quer dizer que suas conclusões podem ser falsas. Todo e qualquer problema para o qual a pergunta do final do parágrafo anterior possa ser formulada com significado, é um problema que exige uma solução estatística.

FILOSOFIA DE TESTES ESTATÍSTICOS

Todos os métodos de inferência estatística (testes estatísticos) são iguais no sentido de que se baseiam em uma mesma série de argumentos lógicos. Considere ainda o problema anterior e siga os passos lógicos para um teste:

Formulação de uma hipótese

Neste caso, a hipótese básica é que não há diferenças na variabilidade populacional para as duas populações estudadas que pertencem a duas espécies. Esta hipótese pode ser considerada a mais simples hipótese que pode ser formulada sobre o problema. Qualquer outra hipótese (a espécie A varia mais; a espécie B varia mais) é logicamente mais complexa, porque pressupõe a existência de no mínimo um efeito a mais (há um fator que causa a maior variabilidade da espécie A ou B). A hipótese mais simples é geralmente chamada de *Hipótese nula.*

Dedução do resultado esperado quando a hipótese nula é verdadeira

Este é um passo obviamente simples: o esperado é que a variação seja igual. Pode-se medir esta variação por uma grandeza estatística chamada variância. Este passo é importante para que se possa operacionalizar o teste, ou seja, definir claramente o que medir na natureza para testar a hipótese.

Dedução da distribuição esperada dos possíveis resultados, se a hipótese nula fosse verdadeira

Este passo é delicado. Como seria possível demonstrar que há uma alta probabilidade de que os resultados sustentem ou não a hipótese nula. Considere um exemplo: a espécie A apresentou uma variância de 17,6 e a espécie B, uma variância de 21,3. Será que isto é suficiente para assumir que a espécie B varia mais? O primeiro passo é calcular um número que represente o resultado obtido. Uma possibilidade é dividir a maior variância pela menor. Chamemos este número de F (em honra a *Ronald Fisher*). Neste exemplo, ele vale 1,21, e representa que a variação na espécie B é 1,21 vezes maior que a A. A atenção deve se voltar agora para a hipótese nula. Qual seria a distribuição esperada dos possíveis valores de F se a hipótese nula fosse verdadeira? Isto equivale a dizer: como variaria F se na verdade as duas variâncias fossem iguais ? Uma nova coleta de dados na mesma comunidade (ou mesmo amostragens em dias diferentes do estudo original) mostraria pequenas diferenças. Tais diferenças não significativas se devem ao *acaso*. O *acaso* reúne todos os outros fatores da natureza não medidos e que podem afetar os resultados do experimento, exceto os mecanismos que estão subjacentes à hipótese. Este passo é agora feito por um "estatístico-matemático" que desenvolve uma equação que representa a distribuição esperada se o fenômeno descrito fosse devido somente ao acaso. Esta equação é usualmente chamada de função de distribuição e descreve a probabilidade de ocorrer cada uma das possibilidades de resultado, quando o fenômeno é apenas dirigido pelo acaso.

A tomada de uma decisão

A decisão a ser tomada é a de aceitar ou rejeitar a hipótese nula. Isto equivale a decidir se as variâncias podem ser consideradas iguais e suas diferenças podem ser explicadas pelo acaso ou se as variâncias podem ser consideradas diferentes e é preciso invocar um outro mecanismo, fora o acaso, para explicar as diferenças. O método para testar é simples. Se a variação de F é conhecida quando a hipótese nula é verdadeira, basta calcular qual a probabilidade de encontrar um resultado como 1,21 quando a hipótese nula é verdadeira, usando a função de distribuição de F. Se esta probabilidade for alta, não há nenhuma razão

para desconfiar que a hipótese nula seja falsa. Ou seja, se as diferenças encontradas são passíveis de ocorrer mesmo quando as variâncias são iguais, deve-se aceitar o acaso para explicar as variações observadas. Se a probabilidade é baixa, então é muito raro ocorrer um resultado como o que foi observado quando a hipótese nula é verdadeira, o que mostra que ela não satisfaz como explicação para o fenômeno. Assim, faz-se necessária outra explicação, que não o acaso, para as diferenças entre as variâncias. Elas são estatisticamente diferentes.

Ao decidir pela rejeição ou não da hipótese nula (*Ho*) o pesquisador corre o risco de estar tomando uma decisão errada. Existem dois tipos de erros associados à decisão em um teste de hipóteses: o primeiro erro, dito *Erro Tipo 1*, é decidir pela rejeição da hipótese nula sendo ela verdadeira. Voltando ao exemplo, *Ho* foi rejeitada, ou seja, as populações A e B têm variâncias diferentes. Faz-se necessário estimar o grau de incerteza associado à essa decisão. A probabilidade de se cometer o *Erro Tipo 1* é o chamado nível de significância, ou α. Adotar um nível de significância de 5% quer dizer probabilisticamente que se a amostragem for repetida 100 vezes, em 95 delas a decisão tomada estará correta rejeitando-se *Ho*.

A outra decisão errada é aceitar a hipótese nula quando ela é falsa. Esse é o chamado *Erro Tipo 2,* cuja probabilidade é definida por β. O poder de um teste é definido como 1 - β, isto é, quanto menor a probabilidade de cometer o *Erro Tipo 2* mais poderoso é o teste. Ambos os erros são indesejáveis, porém o pesquisador tem controle mais efetivo sobre o *Erro Tipo 1*. Para diminuir a probabilidade de rejeitar uma hipótese nula sendo ela verdadeira, basta reduzir o nível de significância (geralmente de 5% para 1%). A mesma regra não se aplica para o valor de β. Na verdade, quanto mais se reduz o nível de significância mais se aumenta a probabilidade de cometer o *Erro Tipo 2*. A única maneira de reduzir simultaneamente ambos os tipos de erro de decisão é aumentar o tamanho da amostra. Assim, para um dado nível de significância, amostras grandes produzem um teste estatístico mais poderoso. Para concluir, é importante ressaltar que não rejeitar a hipótese nula não prova que ela é verdadeira. Pela lógica dos testes de hipóteses, quer dizer que não existem evidências suficientes para concluir que ela é falsa.

TIPOS DE VARIÁVEIS E ESCOLHA DOS TESTES

Quando procuramos testar uma hipótese, é geralmente possível identificar dois tipos de variáveis: a **independente** e a **dependente**. A variável independente ou preditora é aquela que, em teoria, causa o efeito que procuramos confirmar. A variável dependente é a que mede o efeito sofrido. No exemplo, o tamanho da população é a variável dependente e a variável independente é a espécie. Estamos investigando a possibilidade de que o tamanho populacional (e a variabilidade desta medida) seja diferente entre as espécies, como resultado de suas diferenças ecológicas.

Uma outra maneira de classificar as variáveis é quanto à natureza de suas medidas. Os dois exemplos extremos das escalas de medidas são as variáveis **categóricas** e as **quantitativas**. Variáveis categóricas apenas representam distinções de qualidade, enquanto as variáveis quantitativas representam diferenças de quantidades. No exemplo anterior, as espécies são variáveis categóricas e o tamanho da população é uma variável quantitativa. Esta divisão refere-se à forma como os dados foram coletados: uma variável categórica como a cor (preto, branco etc.) pode ser medida como quantitativa (o comprimento de onda da luz emitida). A Tabela 1 apresenta um modelo bastante simplificado para a escolha do teste estatístico apropriado.

Tabela 1. Sugestão de alguns testes estatísticos a empregar de acordo com o tipo de variável observada. Entre parênteses alguns testes não-paramétricos.

Variável Dependente	Variável Independente	Teste
Quantitativa	1 Categórica com 2 níveis	Teste t (teste U)
Quantitativa	1 Categórica com + 2 níveis	ANOVA 1-fator (Kruskall-Wallys)
Quantitativa	2 Categóricas	ANOVA 2-fatores (Friedman[1])
Quantitativa	1 Quantitativa	Regressão simples (correlação Spearman)
Quantitativa	2 ou mais quantitativas	Regressão múltipla
Quantitativa	1 categórica e 1 ou mais quantitativas	ANCOVA
Categórica	1 Categórica	Qui-quadrado[2]; Teste G[2]
Categórica	2 ou mais categóricas	Log-linear[2]

(1) No caso de amostras dependentes, (2) Esses testes eventualmente verificam não a relação de dependência entre variáveis, mas sim a associação entre elas, descaracterizando, portanto a classificação de variáveis dependentes e independentes.

A APRESENTAÇÃO DE RESULTADOS

O cientista é, em essência, um escritor. De que realmente vale o conhecimento produzido se não for exposto com clareza à comunidade que poderá utilizar este conhecimento? Assim, deve-se ter a preocupação com apresentar as idéias dando sempre ênfase ao problema biológico e ao tamanho do efeito atingido, e resguardando o resultado dos testes estatísticos ao bem delimitado espaço interno dos parênteses. Por exemplo, não se deve dizer: "As populações tiveram diferenças de variabilidade populacional estatisticamente diferentes pelo teste F". Melhor dizer: "A população A variou 2 vezes mais que a população B ($F = 2{,}31$; $P<0.05$)." Não se esqueça que é mais facilmente compreensível o que nos for apresentado por figuras, do que por longas Tabelas.

UM BREVE APANHADO DE PRESSUPOSTOS E TRANSFORMAÇÕES

Serão apresentados aqui alguns testes estatísticos mais empregados, tentando demonstrar que todos eles seguem a mesma lógica de tomada de decisão. O que um teste estatístico faz é fornecer uma medida de incerteza ou as chances de se tomar uma decisão errada. Para que tal rotina funcione, alguns pressupostos devem ser cumpridos.

Um primeiro cuidado envolve o desenho amostral. É preciso garantir que as amostras sejam tomadas ao acaso e, a menos que seja interesse explícito, que elas sejam independentes. Muitos dos problemas na análise dos dados vêm da não observância desses pontos.

Alguns testes estatísticos dependem da distribuição dos dados ou, mais precisamente, da distribuição da média amostral. Tais testes são classificados como "paramétricos" e, para empregá-los, deve-se garantir que além da distribuição ser normal as variâncias entre grupos (no caso de teste t e ANOVA) devem ser iguais. De maneira geral, os dois pressupostos: normalidade e homogeneidade de variâncias não são requisitados para os testes não-paramétricos. O problema é que nem sempre existe uma alternativa não-paramétrica para cada teste paramétrico.

As transformações dos dados geralmente são empregadas para tentar corrigir a não-normalidade ou a heterocedasticidade das variâncias. Como exemplo de transformações temos a logarítmica (para corrigir distribuições assimétricas e para remover a dependência

entre média e variância, além de homogeneizar variâncias entre grupos), a raiz-quadrada (para dados de contagens, por exemplo, número de filhotes por gestação) e a transformação arco-seno da raiz-quadrada ou angular (para dados em proporção). Independente da transformação escolhida, um problema comum é que os dados transformados perdem seu significado biológico, o que pode levar a interpretações equivocadas das possíveis relações entre as variáveis.

UMA BREVE RESENHA DOS TESTES ESTATÍSTICOS

Serão apresentados aqui alguns dos principais testes estatísticos tentando mostrar como são percorridos os passos lógicos definidos em nosso exemplo.

Comparando categorias: O teste do qui-quadrado

A Tabela 1 mostra que no estudo da dependência entre duas variáveis categóricas utiliza-se o teste de Q-quadrado. Considere a seguinte questão: existe associação entre uma determinada espécie de ave frugívora e uma determinada família de plantas? Para dar nome ao experimento considere que a ave seja *Thraupis sayaca* (o sanhaço) e a família de plantas as Melastomatáceas. Seguindo-se os passos pré-definidos observa-se:

Hipótese. A hipótese nula seria a de que não há associação entre o sanhaço e as Melastomatáceas. Como coletar dados para testar esta hipótese? Toda vez que se observar um ato de frugivoria por uma ave no campo deve-se classificar a espécie de árvore em uma das categorias: se é ou não uma *Melastomatacea*. Da mesma forma deve-se classificar a ave como sendo ou não um sanhaço. Existem agora duas variáveis *categóricas binárias*. A Tabela 2 reúne os resultados deste experimento em observações de campo no campus da Universidade Federal de Viçosa:

Tabela 2. Tabela de contingência de 99 observações de pássaros em árvores.

		É um Sanhaço?		
		Sim	Não	**Total**
É uma Melastomatácea?	Sim	13	34	47
	Não	12	40	52
	Total	25	74	99

A proporção de sanhaços encontrados em Melastomatáceas foi de 13/47=0,276 enquanto nas não Melastomatáceas esta proporção foi de 12/52=0,231.

Dedução do resultado esperado se a hipótese nula for verdadeira

Qual o valor esperado para cada célula da Tabela acima sob a hipótese de que não há associação? O esperado é que a proporção de que se encontre sanhaço em *Melastomataceae* é igual à proporção desta espécie quando não é *Melastomataceae*. Isto também quer dizer que

as diferenças encontradas nos números observados nas células internas da Tabela seriam explicadas apenas por diferenças no número de amostras (a coluna e a linha denominadas *total* na Tabela). Assim, a proporção 25 sanhaços no total de 99 aves observadas deveria se manter tanto para as 47 aves encontradas em Melastomatáceas quanto para as 52 encontradas em não Melastomatáceas. Isto é o equivalente a predizer que o resultado esperado para o número de sanhaços observados em Melastomatáceas seria obtido pela regra de três simples: 25 "está para" 99 como x "está para" 47. A Tabela 3 mostra os valores esperados.

Tabela 3. Valores esperados da Tabela 2 se *Ho* for verdadeira.

		É um Sanhaço?		
		Sim	Não	**Total**
É uma Melastomatácea?	Sim	11,9	35,1	47
	Não	13,1	38,9	52
	Total	25	74	99

A pergunta agora passa a ser: quão diferentes são os resultados observados em relação ao esperado pelo acaso? Para definir a estatística deste teste usamos o Q-quadrado cujo símbolo é χ^2. Ele seria estimado simplesmente pela diferença entre observado e esperado, elevada ao quadrado, dividida pelo esperado. Este número pode ser calculado para cada uma das células e o somatório destes números é utilizado como teste estatístico. Você pode olhar em uma Tabela de Q-quadrado com 1 grau de liberdade, calculado como: g.l. = (nº linhas-1) x (nº colunas-1), a um nível de significância de 5% e avaliar se este valor é grande comparado com o da Tabela. No entanto, mais usualmente, os programas atuais de estatística já indicam qual foi o nível de significância atingido. Neste caso, χ^2 =0,271 e o nível de significância atingido foi p=0,602

Tomada de Decisão. O que representa o valor de p acima? Ele é a probabilidade de encontrar resultados como o que se obteve quando a hipótese nula é verdadeira. **Se em um experimento delineado como o que você acaba de executar há 60,2% de chances de encontrar resultados como os que você encontrou quando a hipótese nula é verdadeira, então há fortes razões para aceitá-la.** No texto da comunicação do resultado deste estudo deve, em alguma parte, estar escrito algo como: "em torno de 27% das aves observadas em Melastomatáceas eram sanhaços e esta proporção em não Melastomatáceas foi de 23%. Tais diferenças foram consideradas como devidas ao acaso (χ^2=0,271; gl=1; p=0,602)".

O EFEITO DE UMA VARIÁVEL CATEGÓRICA COM DOIS NÍVEIS SOBRE UMA VARIÁVEL QUANTITATIVA: O TESTE T DE STUDENT

Um pesquisador quer avaliar o sucesso de duas técnicas de reintrodução de indivíduos de uma espécie de macaco em uma área. A pergunta é: será que deixá-los em um local de pré-adaptação com fornecimento apenas de complemento alimentar aumenta as chances de sobrevivência do indivíduo? Neste ponto, serão discutidos aspectos puramente estatísticos deste problema, mas ao final deste capítulo será apresentada uma análise mais completa deste problema como exemplo de questões mais amplas sobre Biologia da Conservação.

Considerando-se este como um experimento modelo, com recursos financeiros suficientemente grandes para permitir o acompanhamento deste indivíduo reintroduzido até sua morte, é pouco provável que existam muitos indivíduos que possam servir de amostra. Outro fator complicante é que, para as comparações aceitáveis, é necessário que todos os indivíduos sejam de mesmo sexo, mesma idade e sejam aceitos por grupos sociais semelhantes (mesma estrutura social com mesmo número de machos, fêmeas e filhotes). Assumindo todas estas variações, acompanhou-se a vida de indivíduos que foram reintroduzidos a partir de dois grupos, os que passaram e que não passaram pela pré-adaptação. Esta será a variável independente categórica binária. A variável resposta é a idade em que o indivíduo morreu. A Tabela 4 resume os resultados encontrados:
Tabela 4. Longevidade do primata sob duas condições experimentais.

Indivíduo	Pré-adaptação	Longevidade (anos)
1	Sim	2
2	Sim	3
3	Sim	3
4	Sim	2.5
5	Não	3
6	Não	2
7	Não	2
8	Não	1
9	Não	0.5

A hipótese nula reza que não há diferenças de longevidade dos primatas com ou sem pré-adaptação. Propositalmente foi apresentado um conjunto de dados que apresenta dois dos principais problemas que usualmente assustam quem começa a usar os testes estatísticos. Os dados parecem muito regulares para estarem apresentando "*distribuição normal*" e a longevidade na ausência do período de pré-adaptação parece variar mais que com a pré-adaptação.

Para entender melhor o significado destes dados, há necessidade de aprofundar um pouco mais a fase da construção do teste referente à *dedução da distribuição esperada caso a hipótese nula seja verdadeira.* Este passo exige uma dedução baseada em alguns pressupostos básicos que podem variar entre os testes, mas são muito semelhantes para o conjunto de testes classificados como modelos lineares gerais, do qual fazem parte o teste de t, a análise de variância e a análise de regressão.

Na dedução, parte-se do princípio de que os dados provêm de uma distribuição normal e de que a variação dos dados, em cada tratamento (a variância com e sem a fase de pré-

adaptação), é igual. Importante ressaltar que quando os pressupostos não são cumpridos, nada assegura que os resultados dos testes estejam corretos. No entanto, os estatísticos consideram que um teste é robusto quando apesar de alguns pressupostos não serem cumpridos ele permanece correto. O teste de t, por exemplo, é bastante robusto a desvios da normalidade. Quanto a diferenças de variação, há um teste de t para variâncias iguais (homogêneas) e outro para variâncias diferentes, que pode ser facilmente encontrados em qualquer dos *software* dedicados à análises estatísticas. Sendo assim, o teste t é uma ferramenta muito útil e muito robusta, podendo ser utilizado mesmo em situações como as do exemplo.

A partir dos dados da Tabela 4, observa-se que, em média, os indivíduos que receberam o tratamento de uma fase de pré-adaptação viveram 2,625 anos, enquanto os que não receberam sobreviveram 1,700 anos. Isto representa uma sobrevivência de 0,975 anos a mais com a fase de pre-adaptação, mas a pergunta persiste, qual a probabilidade disto ter ocorrido pelo acaso? Um aspecto interessante é que diferenças como estas podem ser devidas ao acaso, principalmente com poucas amostras (4 indivíduos sob a fase de pré-adaptação). Conduzindo o teste, encontra-se um valor de t=1,722, que com 7 graus de liberdade (g.l.=n-1), leva a um valor de p=0,129. A um nível de significância de 5% aceitamos a hipótese nula de que a fase de pré-adaptação não alterou a sobrevivência dos macacos.

Este pode parecer um resultado incoerente que será discutido em detalhes mais adiante neste capítulo.

O EFEITO DE UMA VARIÁVEL CATEGÓRICA COM VÁRIOS NÍVEIS SOBRE UMA VARIÁVEL QUANTITATIVA: A ANÁLISE DE VARIÂNCIA

Em algumas situações o pesquisador quer comparar não as médias de dois grupos, mas de 3 ou mais. A alternativa de comparar as médias duas a duas de cada grupo é pouco eficiente, uma vez que pode ser produzido um grande número de pares. Se existirem 6 grupos, o pesquisador necessitaria de 15 testes t para comparar as médias de todos os grupos. Para resolver essa situação, Ronald Fisher desenvolveu, na década de 20, a técnica da Análise de Variância, ou ANOVA.

Imagine uma situação na qual o pesquisador deseja comparar a densidade populacional de uma espécie de planta ao longo de um gradiente altitudinal. Para tal, ele definiu quatro cotas de altitude e em cada uma coletou em oito pontos, perfazendo um total de 32 amostras. Estimou, então, os parâmetros média e variância da densidade de plantas em cada uma das quatro cotas. A partir daí ele formulou as seguintes hipóteses:

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
Ha: Existe diferença na densidade média entre as cotas de altitude.

Observe que a hipótese nula (*Ho*) também quer dizer que não há efeito da altitude sobre a densidade da espécie, com conseqüência lógica da igualdade das médias em altitudes diferentes. Para se rejeitar a hipótese nula, basta que pelo menos, um par apresente valores médios diferentes, para um nível de significância de 5% ($\alpha = 0,05$). Os valores obtidos pelo pesquisador estão listados na Tabela 5.

Tabela **5**. Número de indivíduos coletados em cada uma das 4 cotas de altitude.

ALTITUDE	PONTOS DE COLETA							
	P1	**P2**	**P3**	**P4**	**P4**	**P6**	**P7**	**P8**
Cota 1	19	15	17	21	22	23	22	19
Cota 2	21	22	17	20	17	21	21	24
Cota 3	16	17	19	18	14	20	15	17
Cota 4	18	18	14	16	19	15	13	18

A partir dos dados coletados é possível estimar os parâmetros média e variância da densidade populacional para cada uma das quatro cotas de altitude. A variância em particular pode ser dividida em dois componentes: variância entre os grupos (ou variância devido ao tratamento) e variância dentro dos grupos (variância devido ao erro). Um quadro de ANOVA característico, resultante do conjunto de dados apresentados no exemplo está ilustrado na Tabela 6.

Tabela 6. Análise de variância testando o efeito da altitude sobre a abundância da planta.

Fonte de variação	**Soma de Quadrados**	**Graus de Liberdade**	**Quadrado médio**	**F**	**Valor p**
Efeito (Altitude)	94,25	3	31,42	5,66	0,004
Erro amostral	155,25	28	5,54		
Total	249,5	31			

Uma das maneiras de se estimar quanto um conjunto de dados varia em relação ao valor médio, é somar todas as diferenças entre cada valor e a média, tomando o cuidado de elevar a diferença ao quadrado para evitar que a soma iguale a zero. Essa é a chamada soma dos quadrados (SQ). Ao dividir esse valor pelo número de graus de liberdade temos o quadrado médio (QM), ou variância. A estatística F é calculada ao se dividir o QM do efeito (variância entre os grupos) pelo QM do erro (variância dentro dos grupos). Você deve lembrar o que foi dito no início desse texto: o valor F é uma razão entre variâncias. Compara-se o valor F calculado com o valor esperado sendo a hipótese nula verdadeira, e decide-se pela sua rejeição ou não. A maioria dos programas estatísticos calcula a probabilidade associada ao valor F calculado. No exemplo acima, o valor F calculado foi de 5,66, com um nível de significância atingido (ou valor-p estimado) de 0,004. Como o valor-p está bem abaixo do nível de significância adotado de 0,05 rejeitamos a hipótese nula, ou seja, existe efeito significativo da altitude sobre a densidade da planta. Uma boa maneira para ilustrar o resultado sem apresentar o quadro completo é fornecer o valor F com seus graus de liberdade e o valor-p. No exemplo acima, diríamos: "Existe diferença na densidade ente as cotas de altitude ($F_{3,28}$=5,66; p=0,004)". Além disso, a apresentação gráfica dos valores médios por grupo, com suas respectivas medidas de variação facilita a visualização dos resultados. Gráficos do tipo *box-plot* como o da figura 1 são bem ilustrativos.

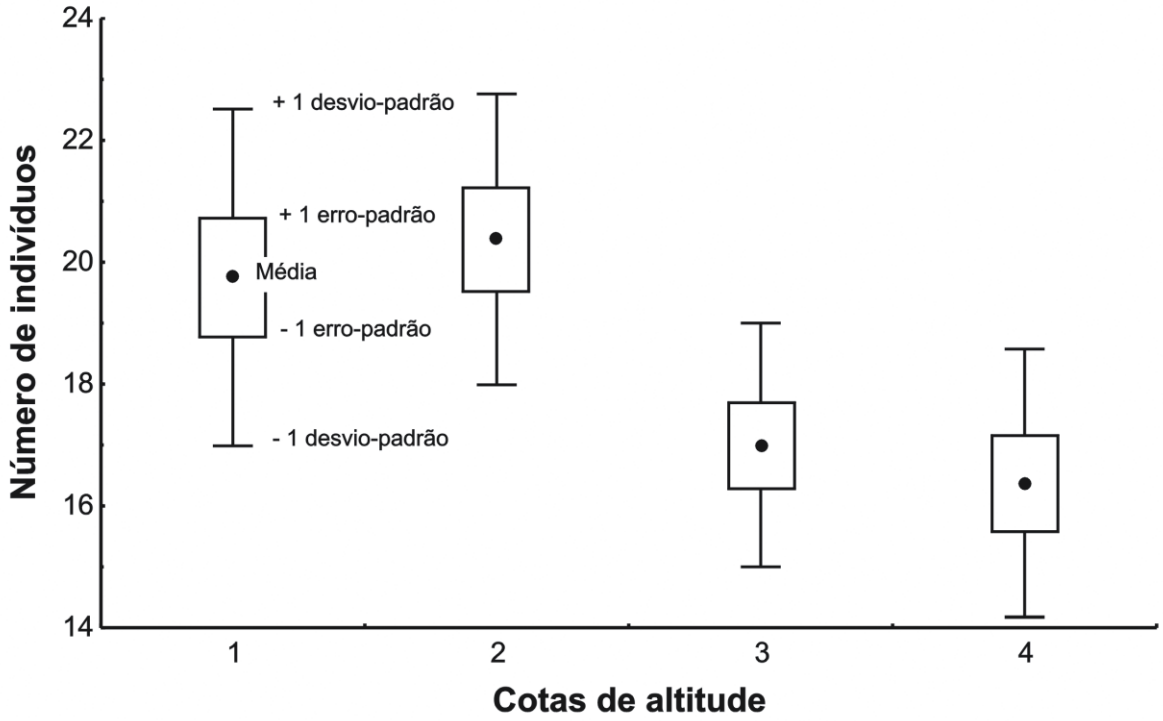


Figura 1. Representação das médias, erros-padrão e desvios-padrão do número de plantas nas quatro altitudes amostradas.

O teste ANOVA indica se existe diferença, mas não informa onde esta se encontra. Para tal, tendo rejeitado a hipótese nula pela ANOVA faz-se necessário um teste *a posteriori*. De uma maneira geral, existem dois grupos de testes a *posteriori.* Os primeiros, denominados testes de comparação múltipla, nos quais não se estabelece uma hipótese *a priori,* e os testes de comparação planejada, empregando a técnica de contrastes. Este último, mais "elegante", deve ser utilizado sempre quando o pesquisador já possuir, antes de iniciar o experimento, uma hipótese de como seus grupos devem se diferenciar.

Existem muitos testes de comparação múltipla, sendo os mais conhecidos, *Tukey, Duncan e Scheffé*. Aplicando o teste de comparação múltipla de *Tukey* no exemplo, observa-se que as diferenças se encontram entre as cotas 1 e 4; 2 e 3; 2 e 4.

A DEPENDÊNCIA ENTRE DUAS OU MAIS VARIÁVEIS QUANTITATIVAS: REGRESSÃO LINEAR

Todos os modelos estatísticos lineares apresentam a mesma formulação. Podemos escrever o modelo do exemplo acima da ANOVA como: N^o *de indivíduos* = $\alpha + \beta$*(altitude) + Erro*, ou seja, o número de indivíduos da planta é função da altitude. O que determina a associação entre a variável dependente (nº de indivíduos) com a variável independente (altitude) é o coeficiente β. Devido ao fato de que a variável independente ser, no exemplo, categórica (quatro cotas de altitude), empregamos a técnica de ANOVA (veja a Tabela 1).

Agora imagine que o pesquisador, ao invés de coletar oito amostras em cada uma das quatro cotas de altitude, fez coletas ao longo de todo o gradiente altitudinal. Além disso, o

pesquisador estimou também a riqueza de insetos polinizadores em cada ponto de coleta e obteve os seguintes resultados:

Tabela 7. Abundância da planta e riqueza de espécies de polinizadores por altitude.

Altitude (metros)	Nº de espécies de polinizadores	Número de indivídos da planta
500	27	31
550	15	32
610	12	28
680	45	29
720	20	30
770	40	20
810	10	15
890	27	15
930	29	13
990	12	12
1030	25	10
1080	8	8
1140	12	7
1200	9	9

Em primeiro lugar, cabe testar se existe associação entre a abundância de plantas e a altitude. O modelo linear seria então:

$$\text{Nº de indivíduos} = \alpha + \beta(\text{altitude}) + \varepsilon,$$

onde α e β são constantes, sendo α o intercepto, isto é o ponto onde a reta de regressão corta o eixo Y e β é o coeficiente da regressão, que indica o grau de associação entre as duas variáveis. O erro amostral é indicado por ε. O valor estimado do coeficiente da regressão indica a intensidade e a direção da regressão. A figura 2 ilustra as retas originadas a partir de diferentes valores de inclinação. O que a regressão linear faz é estimar, através do método chamado "*quadrados mínimos*", os coeficientes do modelo. Associada a essa estimativa, testa-se as seguintes hipóteses:

Ho: $\beta = 0$ (não existe associação entre as duas variáveis)
Ha: $\beta \neq 0$, (existe associação entre as duas variáveis)

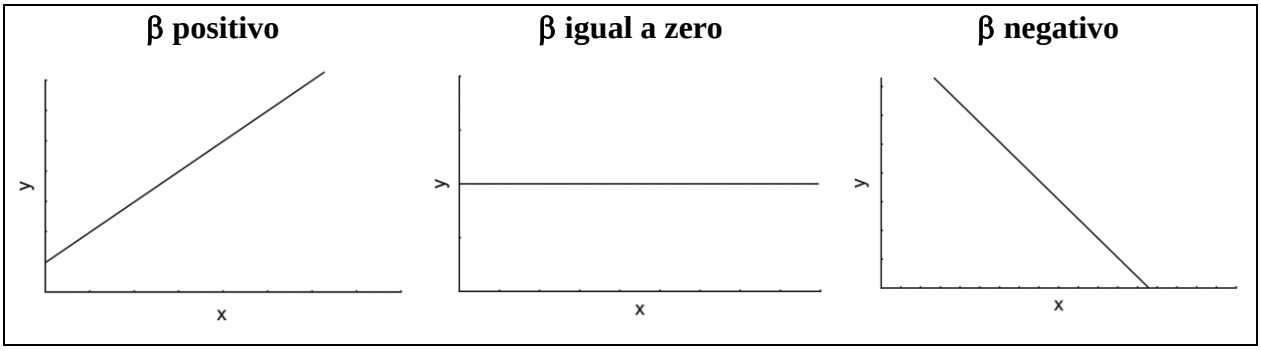


Figura 2. Retas produzidas por diferentes coeficientes de regressão. $\beta > 0$ indica associação positiva; $\beta < 0$ associação negativa e β igual a zero indica ausência de associação entre as duas variáveis.

Voltando ao modelo do exemplo, o método de quadrados mínimos estimou a seguinte equação: **Nº de indivíduos = 52,9 - 0,04 (altitude) + erro.** Isso significa que a diminuição de 0,04 unidades da variável independente leva a um aumento de uma unidade na variável dependente. Com essa equação, é possível predizer quantos indivíduos deve ter uma população dessa planta numa determinada altitude. Ainda não testamos se o coeficiente de inclinação é estatisticamente diferente de zero. Note que o valor -0,04 , indicado na equação acima, não é o valor de β. O coeficiente da regressão é calculado de tal forma que varie entre -1 (alta correlação negativa) a 1 (alta correlação positiva), passando por zero (ausência de correlação). O resultado de uma regressão pode ser visualizado na Tabela abaixo:

Tabela 8. Efeito da altitude sobre a abundância de plantas.

	Coeficientes		**Estatística**		
	β	B	g.l.	t	Valor-p
Intercepto		52,928	12	15,316	< 0,001
Altitude	-0,947	-0,0405	12	-10,275	< 0,001

O coeficiente de correlação estimado foi de -0,947, indicando uma alta correlação negativa. À medida que aumenta a altitude, diminui a abundância da planta. Essa diminuição se dá na "velocidade" de menos 1 indivíduo a cada 0,04 metros de altitude. Na Tabela 8 também está indicado o teste t utilizado para testar a hipótese nula de que o coeficiente de inclinação é igual a zero. Com o valor calculado de -10,275 para 12 graus de liberdade rejeita-se *Ho*. Uma outra maneira de testar a significância da regressão é utilizar uma análise de variância. A Tabela 9 demonstra a saída típica da maioria dos programas estatísticos para o procedimento.

Tabela 9. Análise de variância para a regressão entre altitude e abundância da planta.

Fonte de variação	**Soma de Quadrados**	**Graus de Liberdade**	**Quadrado médio**	**F**	**Valor p**
Regressão	1055,5	1	1055,5	105,57	< 0,001
Resíduo	119,9	12	9,99		
Total	1175,5				

Como foi dito no tópico sobre ANOVA, a soma dos quadrados (SQ) é uma estimativa da variância particionada entre a regressão e o resíduo, ou erro. A proporção entre a SQ_{reg} e a SQ_{tot} indica quanto da variação é explicada pela regressão. Nesse caso $\frac{1055,5}{1175,5} = 0,898$. A regressão explica 89,8% da variação dos dados. Esse valor é o chamado R^2 da regressão, e pode também ser calculado simplesmente elevando-se ao quadrado o valor do coeficiente de correlação ($R = -0,947 \rightarrow R^2 = 0,898$). O teste segue a mesma lógica de uma ANOVA comum. Calcula-se o valor da estatística F pela divisão dos quadrados médios. (QM_{Reg}/QM_{Res}). Compara-se o valor calculado com o esperado sendo verdadeira a hipótese nula e toma-se a decisão. No exemplo, o elevado valor de F indica que a regressão é altamente significativa (Figura 3).

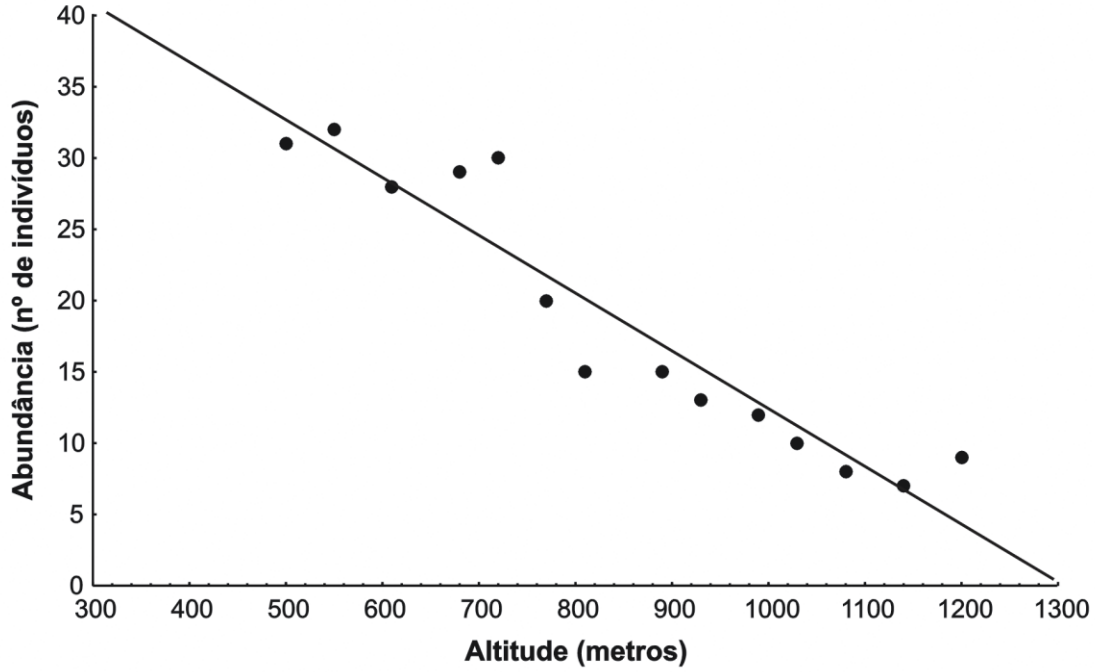


Figura 3. Regressão entre a altitude amostrada e abundância de plantas.

O pesquisador pretende testar se a altitude, assim como, também, a riqueza de espécies de polinizadores, determinam a abundância da planta. O modelo agora é:

$$\text{Nº de indivíduos} = \alpha + \beta_1(\text{altitude}) + \beta_2(\text{riqueza}) + \varepsilon,$$

Foram incorporados ao modelo o efeito da riqueza de espécies polinizadoras sobre a abundância de indivíduos. A regressão linear agora é dita regressão múltipla. Em tese, podemos tornar um modelo cada vez mais explicativo pela inclusão de novas variáveis, porém, dois pontos devem ser observados. Primeiro, o tamanho da amostra deve ser grande o suficiente para o número de variáveis. Regressões com poucos pontos em relação ao número de variáveis são altamente explicativas (apresentam altos valores de R^2), mas não são confiáveis. O outro problema com muitas variáveis independentes é que se elas estiverem correlacionadas, então a interpretação dos coeficientes de correlação de cada uma fica prejudicada.

Voltando ao modelo, a regressão múltipla testa, por meio de ANOVA, a significância do ajuste, e testa também através do teste t, os coeficientes β estimados para cada termo da regressão. A saída usual de uma análise de regressão múltipla é similar à da regressão simples, apenas incluindo-se as variáveis adicionais (Tabela 10).

Tabela 10. Efeito da altitude e da riqueza sobre a abundância de plantas.

	Coeficientes		**Estatística**		
	β	B	g.l.	t	Valor-p
Intercepto		52,495	11	10,984	< 0,001
Altitude	-0,942	-0,040	11	-9,092	< 0,001
Riqueza	0,014	0,011	11	0,138	0,892

Estima-se o coeficiente de correlação de cada variável do modelo. Nesse caso, a densidade é negativamente influenciada pela altitude e não sofre efeito da riqueza de espécies de polinizadores. Além dos coeficientes parciais de correlação, calcula-se também o coeficiente de correlação múltipla R, nesse caso de 0,947, muito próximo do coeficiente de correlação da variável altitude. A regressão explica cerca de 89,8% da variação total (R^2 = 0,898). A analise de variância da regressão múltipla também é similar à da regressão simples (Tabela 11).

Tabela 11. Análise de variância para a regressão múltipla entre altitude e riqueza com a abundância das plantas.

Fonte de variação	**Soma de Quadrados**	**Graus de Liberdade**	**Quadrado médio**	**F**	**Valor p**
Regressão	1055,7	2	527,86	48,48	<0,001
Resíduo	119,8	11	10,88		
Total	1175,5				

QUANDO A VARIÁVEL DEPENDENTE É BINÁRIA: A REGRESSÃO LOGÍSTICA

Em algumas situações práticas de campo é difícil ter boas estimativas da abundância de uma espécie. Isso é principalmente verdadeiro quando se trata de espécies raras, ou de difícil coleta e/ou visualização. A questão é que muitas vezes são essas espécies nosso foco de interesse. Imagine, então, que você está interessado em discutir a influência de fatores antrópicos sobre uma espécie rara. Imagine que tais fatores são mensuráveis como, por exemplo, área perdida ou concentração de metais pesados na água. Podemos imaginar um modelo preditivo (através da regressão linear, por exemplo) que nos forneça uma idéia de qual seria a "velocidade" com que a população perde indivíduos à medida que aumenta o nível de poluição.

Por se tratar de espécie rara, ou pelo menos inconspícua, as chances de você conseguir boas estimativas dos tamanhos populacionais é pequena. O máximo que se consegue é afirmar se a espécie está ou não presente numa determinada amostra, se não se está preocupado com a abundância, mas sim com a ocorrência da espécie. Assim, a variável resposta (dependente) é categórica, e só pode assumir dois valores (presença ou ausência). Para essa e outras situações semelhantes (morreu/sobreviveu; tem filhotes/não tem filhotes, etc...) a análise indicada é a regressão logística (veja a Tabela 1).

Uma situação mais real: algumas espécies de macro-invertebrados de água doce da família *Chironomidae* (Diptera) podem ser indicadoras de qualidade ambiental. Certas espécies só ocorrem em ambientes preservados, enquanto que outras estão presentes em sistemas aquáticos bastante eutrofizados. Os dados apresentados abaixo são de Marques *et al*. (1999). Os autores coletaram em 20 pontos da bacia do Rio Doce, no estado de Minas Gerais. Em cada ponto, foram medidas diversas variáveis físico-químicas da água, entre elas, a concentração de nitrogênio total, que é indicador de grau de eutrofização. Diversas espécies de *Chironomidae* foram coletadas. Abaixo apresentamos os dados de ocorrência de duas espécies. Observe que nos dados originais a presença das espécies está categorizada em 3 classes de abundância.

Tabela 12. Presença (1) e ausência (0) de duas espécies de *Chironomidae* concentração de nitrogênio total em 20 pontos da bacia do Rio Doce.

Ponto	***Tanitarsus* sp**	*Cryptochironomus*	Nitrogênio total **(μg/l)**
1	1	0	262,4
2	1	1	420,6
3	0	1	1889
4	1	1	718,5
5	1	1	471,3
6	0	0	1219,3
7	0	1	1587
8	1	1	482,6
9	0	1	2132
10	0	0	3112
11	0	0	5257
12	1	1	454,3
13	0	0	1221
14	0	1	837,8
15	0	0	538,9
16	1	1	136,2
17	0	0	574,5
18	0	0	775,6
19	0	0	7283
20	1	0	308,8

Podemos elaborar as seguintes hipóteses referentes à *Tanitarsus* sp.:

Ho: A ocorrência de *Tanitarsus* na bacia do Rio Doce não depende da concentração de nitrogênio total na água;
Ha: *Tanitarsus* é um organismo sensível à eutrofização, e ocorre preferencialmente em ambientes menos poluídos.

O modelo seria: Ocorrência de *Tanitarsus* $\cong \alpha + \beta_1(\text{N-tot}) + \varepsilon$, (o símbolo $\cong$ indica "é função de"). O modelo logístico é:

$$Y = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)}}$$

onde Y é a probabilidade de ocorrência da espécie; α é análogo ao intercepto na regressão linear, e βi representa o coeficiente da i-ésima variável. α e os coeficientes β representam os parâmetros que serão estimados através do método conhecido como Máxima Verossimilhança ("Maximum Likelihood", em inglês). A interpretação é análoga à regressão linear. O modelo indica a relação entre a ocorrência de *Tanitarsus* e a concentração de nitrogênio total na água. Existem duas formas para se testar essa relação em uma regressão logística: 1) O teste LR ("*Likelihood Ratio*", ou Razão de Verossimilhança) e 2) O teste de *Wald*.

O teste de razão de verossimilhança baseia-se na estatística LR. Essa estatística é calculada a partir dos valores L = -2 Ln(*Likelihood*) tanto para o modelo com a variável (chamemos de L_C) e quanto para o modelo simples, sem a variável (L_S). No exemplo de *Tanitarsus* (com valores de N-total log-transformados) temos: valor de verossimilhança para o modelo simples = -2ln(L_S) = 26,970, e valor de verossimilhança para o modelo com a variável N-tot = -2ln(L_C) = 8,695

Se fizermos L_S - L_C:

-2 Ln(L_S) - {-2 Ln(L_C)}, ou, pela propriedade de subtração de logaritmos:

-2 Ln(L_S/L_C) = LR, por isso é uma Razão de Verossimilhanças, ou LR.

A maioria dos programas fornece o valor de verossimilhança para o modelo simples e para o modelo completo e calcula o valor de LR diminuindo um do outro. LR tem uma distribuição de Qui-quadrado, com o número de graus de liberdade definido como a diferença no número de parâmetros entre o modelo completo (ou o número de variáveis + α) e o modelo simples (apenas o parâmetro α). Com o valor da estatística LR e o número de graus de liberdade calcula-se o valor-p associado ao LR.

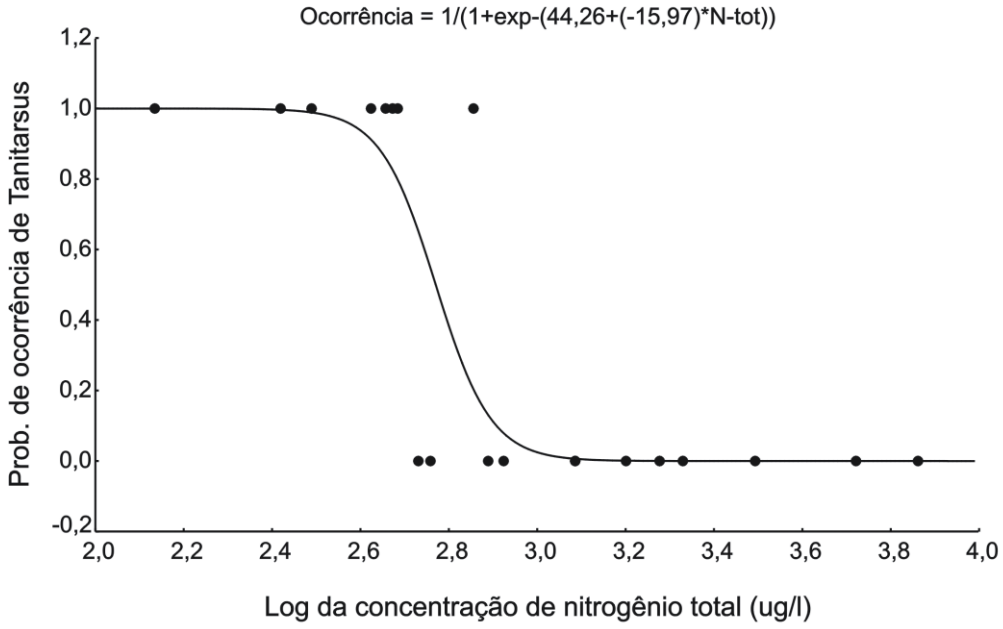
Seguindo nosso exemplo: LR = 26,920 - 8,695 = 18,225; Nº de parâmetros do modelo completo = 2 (α e β_1); Nº de parâmetros do modelo simples = 1 (α); Graus de liberdade = 1; e Valor-p < 0,001. Assim, rejeita-se H_0: A ocorrência de *Tanitarsus* sp. depende da concentração de nitrogênio total na água. Os parâmetros estimados foram α = 44,26 e β = -15,97. Sendo β negativo, a relação entre ocorrência da espécie e concentração de N-tot é inversa. A figura 4A ilustra essa relação.

A contribuição da variável N-tot é indicada pelo valor de LR. Se a variável tem pouco peso para explicar a ocorrência da espécie, então o valor de verossimilhança para o modelo com essa variável é grande, próximo ao valor de verossimilhança para o modelo simples. Ao subtrair um pelo outro, o valor de LR fica pequeno. Assim, quanto mais próximo de zero for a estatística LR, menor é o peso que a variável têm para explicar a variável dependente. Isso pode ser visto no exemplo abaixo, a análise para a relação entre N-total e a ocorrência de *Cryptochironomus* sp.:

Ho: A ocorrência de *Cryptochironomus* sp. na bacia do Rio Doce não depende da concentração de nitrogênio total na água;
Ha: *Cryptochironomus* sp. é um organismo sensível à eutrofização, e ocorre preferencialmente em ambientes menos poluídos.
Valor de verossimilhança para o modelo simples = -2ln(L_S) = 27,72; Valor de

verossimilhança para o modelo com a variável N-tot= -2ln(L_C) = 26,39; LR = 27,72 - 26,39 = 1,33; nº de parâmetros do modelo completo = 2 (α **e** β_1); nº de parâmetros do modelo simples = 1 (α); Graus de liberdade = 1; Valor-p = 0,247. Deste modo, aceita-se a hipótese nula, ou seja, a ocorrência de *Cryptochironomus* sp. não depende da concentração de nitrogênio total na água (Figura 4B).

A)



B)

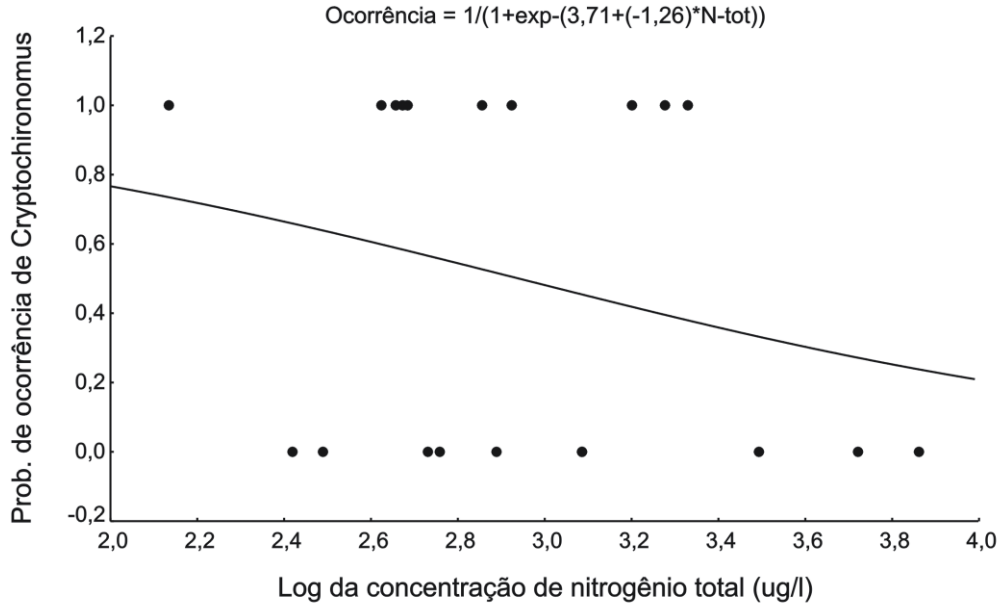


Figura 4. Relação entre a concentração de nitrogênio total e a probabilidade de ocorrência de A) *Tanitarsus* sp. e B) *Cryptochyronomus* sp. em 20 pontos da bacia do Rio Doce.

O outro teste para a relação entre as variáveis na regressão logística, o teste de *Wald*, geralmente fornece resultados semelhantes ao teste de LR. A lógica do teste de *Wald* é similar a do teste t na regressão linear usado para testar se o coeficiente de correlação R é diferente de

zero. Quando o tamanho amostral é grande os resultados de ambos os testes são iguais. Mas, se o tamanho amostral é pequeno, recomenda-se utilizar o teste LR. Um outro problema do teste de *Wald* é que sua interpretação para a situação de duas ou mais variáveis é mais complicada, e envolve a aplicação de álgebra matricial. Diversos estatísticos recomendam que se utilize preferencialmente o teste de LR para inferências estatísticas associadas à regressão logística.

De maneira geral, vale lembrar que os mesmos princípios lógicos e interpretativos da regressão linear podem ser aplicados aos modelos de regressão logística, incluindo as situações de múltiplas variáveis. Nessas situações, aplica-se à rotina de avaliação do valor de LR a medida que se adicionam variáveis no modelo.

RISCOS ASSIMÉTRICOS, PENSAMENTO "DESEJOSO" E A IMPORTÂNCIA DA ESTATÍSTICA NA BIOLOGIA DA CONSERVAÇÃO

Voltemos ao exemplo da longevidade de macacos reintroduzidos com ou sem uma fase de pré-adaptação, discutido na seção sobre o teste de t. Há muitas questões importantes a serem analisadas ali.

Em primeiro lugar vem o problema do número de amostras. É muito comum ouvir as escusas de pesquisadores na área da Ecologia e da Biologia da Conservação de que não é possível um número maior de amostras e que, portanto, deve-se trabalhar com o que se têm. Na maioria das vezes, esta observação não é aceitável e pode gerar prejuízos maiores que os custos de se aumentar o número de réplicas ou de pelo menos desenvolver um experimento bem planejado. Naquele caso, rejeitou-se uma hipótese (de que a pré-adaptação aumenta a longevidade dos animais) que pode ser verdadeira principalmente porque, para conseguir demonstrar um efeito com um número pequeno de réplicas, o tamanho deste efeito precisa ser muito grande.

Isto nos leva também ao problema dos riscos assimétricos, discutido de forma muito interessante, se bem que ligeiramente diferente, em Caughley & Gunn (1996). Considere os dois tipos de erros estatísticos que podem ocorrer neste teste. Nós poderíamos rejeitar a hipótese nula sendo ela verdadeira (*Erro tipo 1*) ou aceitá-la sendo ela falsa (*Erro tipo 2*).

Ao aceitar H_0 quando ela é falsa, está se desconsiderando uma prática de manejo que pode aumentar a sobrevivência do macaco no campo e contribuir para sua preservação. Ao rejeitá-la, sendo ela verdadeira, custos adicionais desnecessários estão sendo introduzidos, onerando o projeto. Este procedimento pode resultar em um menor número de indivíduos reintroduzidos, em razão resultado dos gastos adicionais. Isto mostra dificuldade na tomada de decisão.

É interessante notar certa assimetria entre os erros: em um caso diminui-se diretamente o sucesso do projeto por desconsiderar uma prática útil, no outro, onera-se o projeto e apenas indiretamente diminui-se o sucesso da reintrodução. Muitos conservacionistas não hesitariam em correr o primeiro risco e alguns outros fatores sustentariam esta decisão. Em uma comunidade científica eficiente, em que projetos desta natureza estão sendo continuamente avaliados, um possível erro do tipo I será facilmente detectado à medida que outros experimentos vão sendo desenvolvidos e novos dados sejam adicionados.

Há, no entanto, um problema sério no procedimento anterior. Considerar significativo a um valor-p de 0,10, aceitando um maior erro tipo I, em função de uma escolha de riscos dentro do panorama da assimetria descrita acima, só faz sentido se for uma decisão tomada *antes* do

experimento ser executado. Com uma freqüência muito maior que o esperado em uma comunidade científica madura, estas decisões são tomadas *após* os dados serem coletados, fruto do que os ingleses chamaram de *whishful thinking* -- aqui traduzido, pelo Dr Miguel Petrere Jr., como "pensamento desejoso". O "desejo" de que nossa hipótese alternativa esteja correta é o caminho mais curto para afastar a Biologia da Conservação do vacilante, mas honesto, caminho das Ciências e trazê-la para o caminho do dogmatismo. Afinal, se uma hipótese é considerada correta mesmo que os dados digam o contrário, para que, então, se coletaram os dados?

BIBLIOGRAFIA RECOMENDADA

Caughley, G. & Gunn, A. 1996. **Conservation Biology in Theory and Practice**. Blackwell Science, Inc., Cambridge, Massachusetts. 459p.

Hosmer, D. W. & lemeshow, S. 1989. **Applied Logistic Regression**. John Wiley & Sons, New York. 307 p.

Kleinbaum, D. G. 1994. **Logistic Regression: A self-learning text**. Springer-Verlag, New York. 282p.

Krebs, C. J. 1989. **Ecological Methodology**. Harper & Row, Publishers, New York. 654p.

Magurran, A. E. 1988. **Ecological Diversity and its Measurement**. Cambridge University Press, London. 179p.

Neto, P. R. P.; Valentin, J. L. & Fernandez, F. (eds.). 1995. Tópicos em tratamento de dados biológicos. Volume 2. 1ª Edição. **Oecologia Brasiliensis**, Rio de Janeiro. 161p.

Manly, B. F. J. 1991. **Randomization and Monte Carlo Methods in Biology**. Chapman and Hall, London. 281p.

Martin, P. & Bateson, P. 1986. **Measuring Behaviour**. Cambridge University Press, Cambridge. 200p.

Marques, M. M. G. S. M.; Barbosa, F. A. R. & Callisto, M. 1999. Distribution and abundance of Chironomidae (Diptera, Insecta) in an impacted watershed in south-east Brazil. **Ver. Brasil. Biol**. 59(4):553-561.

Sokal, R. R. & Rohlf, 1995. **Biometry**. W. H. Freeman and Company, New York, USA. 887p.

Tonhasca, A., Jr. 1991. The three "capital sins" of statistics used in biology. **Ciência e Cultura**, 43(6):417-422.

Young, L. J. & Young, J. H. 1998. **Statistical Ecology: a population perspective**.

Zar, J. H. 1984. **Biostatistical analysis**. Prentice-Hall, Englewood Cliffs, N.J. 218p.

PARTE 2

GUIA PARA EXECUÇÃO DAS ANÁLISES ESTATÍSTICAS

Flávia Pereira Lima; Leandro Juen; Paulo De Marco Júnior
Laboratório de Ecologia Teórica e Síntese, ICB, Universidade Federal de Goiás

A PROPOSTA DO GUIA

É freqüente encontráramos pessoas que estão muito preocupadas com as análises de dados. Foi muito esforço para coletar, geralmente o prazo para apresentação dos resultados está apertado, mas ainda faltam aquelas análises estatíticas tanto cobradas... Sentar e chorar, que nada! A estatística é uma ferramenta muitas vezes indispensável para os estudos científicos e não é um bicho de sete cabeças.

Vale a pena se dedicar às matérias e aos cursos de estatíticas e compreender as bases teóricas dos testes. Além disso, percebemos que muitas vezes as pessoas sabem escolher o teste estatístico mas tem muita dificuldade na organização das planilhas de dados e na execução. Por isso, nós elaboramos esse guia prático, com os passos das análises mais importantes que vocês podem precisar. Ele deve ser utilizado como um caderno de notas, para facilitar o uso do programa e agilizar o seu trabalho.

BANCO DE DADOS

A correta organização do banco de dados é essencial para a realização das análises estatísticas. Algumas regras auxiliam nesse processo:

1. Utilize o Excel para colocar seus dados (ou outro programa semelhante). Quando são muitos dados é mais adequado utilizar a plataforma Access.

2. Nunca utilize muitos documentos ou muitas planilhas dentro de um documento. Faça o necessário para que você tenha no máximo três planilhas: uma de dados brutos, uma de metadados (explicação do que representa cada coluna da sua tabela) e uma de resultados.

3. Planilha de dados brutos (DADOS): é essencial que você determine a unidade amostral da sua pesquisa. É importante perceber que é possível que você tenha,

dentro da mesma pesquisa, mais de uma unidade amostral. Quando for montar a planilha DADOS **coloque sempre as amostras independentes em linhas diferentes e as variáveis (as informações da mesma amostra) em colunas**. Por exemplo: Pretende-se testar se há diferença de riqueza de drosofilídeos em frutos pequenos e frutos grandes (tamanho do fruto = variável categórica/ riqueza = variável quantitativa). Se:

a. Forem observados frutos numa mata, cada um deles será uma amostra:

Tabela 1: Riqueza de drosofilídeos em frutos grandes (G) e pequenos (P).

Tamanho do fruto	S
G	10
P	4
P	6
G	12
G	9
G	8

b. Se os frutos grandes forem colocados experimentalmente ao lado de frutos pequenos, as amostras se tornam dependentes e o "ponto" passa a ser a amostra, tratando-se de um experimento pareado:

Tabela 2: Riqueza de drosofilídeos em frutos grandes (G) e pequenos (P).

Local	S do fruto pequeno	S do fruto grande
G		
P		
P		
G		

4. Planilha METADADOS: nessa planilha você deve colocar os significados dos códigos utilizados na planilha DADOS. Pode parecer desnecessário ou perda de tempo, mas esse cuidado lhe será útil caso sua planilha tenha muitos códigos, se no futuro você precisar utilizá-la (pode ser que a memória falhe) ou se uma outra pessoa necessitar.

5. Planilha RESULTADOS: nela você colocará os resultados de suas análises estatísticas.

IMPORTAR DADOS PARA O STATISTICA

Siga os seguintes passos para importar seus dados do EXCEL para o STATISTICA. No menu:

1. FILE → OPEN

2. Na janela OPEN selecione **Data files** em "Files of type" → Abrir

3. Selecionar a planilha:

→ Import all sheets to a workbook (irão todas as planilhas do documento)

→ Import selected sheets to a Spreadsheet (você seleciona apenas a planilha de dados)

1. Janela Open Excel File: nela aparecerá o número de colunas e o de linhas da sua planilha. Selecione **Get variable names from first row**, para que os nomes que você deu às variáveis (a primeira linha do Excel) não entre como um dado. Preste atenção se o número de linhas e colunas confere com os da planilha do Excel.

Pronto. A planilha estará importada. Agora é só analisar!

LEMBRETES

Variável dependente: a variável resposta
Variável independente: a que causa o efeito

Variável categórica: qualidade entre os diferentes dados
Variável quantitativa: variável contínua

Teste não paramétrico: não segue a distribuição normal.
Teste paramétrico: segue a distribuição normal.

Casas decimais: apresentar os resultados dos testes com três casas decimais.

PROCEDIMENTOS PARA AS ANÁLISES ESTATÍSTICAS

1. QUI- QUADRADO

1. Statistics → Basic Statistics/Tables → Tables and banners

2. Specify tables (select variables) → OK

3. Testar os pressupostos: i) nenhuma das freqüências esperadas pode ser menor que 1 ii) apenas 25% delas pode ser menor que 5.

→ Options → marcar Expected frequencies → Summary

4. Se os pressupostos não forem feridos:

- Marcar em Options → Statistics for two-way tables → Pearson & M-L Chi-square

- ir em Advanced → Detailed two-way tables → verificar o valor de p, o χ^2 e os graus de liberdade.

5. **Solução do Fisher:** quando a tabela de contingência for do tipo 2 X 2, pode-se utilizar o teste exato de Fisher, que não possui os pressupostos acima apresentados.

→Marcar em Options → Statistics for two-way tables → Fisher exact, Yates, McNemar (2X2)

→ ir em Advanced → Detailed two-way tables → verificar o valor de p.

6. Volta em Options → marcar percentages of row counts → Summary (apresentar uma tabela com as porcentagens).

7. Apresentação dos resultados: χ^2.....; gl.....; p.....

2. TESTE T PARA AMOSTRAS INDEPENDENTES

1. Statistics → Basic Statistics

2. t-test , independent, by groups (Test t para amostras independentes) →OK

3. Variables: selecionar a variável dependente (dependent variables) e a variável independente (independent variables)

IMPORTANTE→ Pressupostos do teste t: i) os dados devem possuir distribuição normal; ii) a variância deve ser homogênea.

4. Para testar se as variâncias são homogêneas: depois de selecionar as variáveis, retornar à janela anterior. Escolher a aba:

4.1 → Options → Levene´s test → Summary (se $p>0,05$ não rejeita a H_0 e, portanto, as variâncias serão homogêneas).

OBS: na janela do resultado do Levene já sai o resultado do teste t.

Mas se as variâncias forem heterogêneas existe uma saída: o teste t com variâncias separadas:

4.2→ Options → Test/w separate variance estimates → Summary

5. Copiar para a planilha RESULTADOS: selecionar toda a planilha (clicar no espaço branco mais à esquerda) ir ao menu em **Edit → Copy with headers**. Colar na planilha RESULTADOS.

6. Apresentação ao leitor: ao apresentar qualquer dado de uma análise estatística ao leitor lembre-se que o mais

importante é o resultado biológico por detrás dos números. No teste t você deverá apresentar o resultado do teste, os graus de liberdade e o valor de p. Analise o tamanho do efeito para apresentá-lo ao leitor. Exemplo: "Um fruto grande pode ter, em média, 2,6 espécies a mais de drosofilídeos do que os frutos pequenos. Essa diferença não pode ser explicada pelo acaso (t= ; gl= ; p=)". Se as variâncias forem separadas (t para variâncias separadas= ; gl= ; p=).

3. TESTE T PARA AMOSTRAS DEPENDENTES

1. Statistics → Basic Statistics

2. t-test , dependent samples (Test t para amostras dependents) →OK

3. Variables→ First variable list/ Second variable list

4. Summary

5. Copiar para a planilha RESULTADOS: selecionar toda a planilha (clicar no espaço branco mais à esquerda) ir ao menu em **Edit → Copy with headers**. Colar na planilha RESULTADOS.

> **OBS:** Como o teste é para amostras dependentes, as diferenças para cada amostra estão sendo controladas, por isso não há o pressuposto da homogeneidade de variância.

4. ANÁLISE DE VARIÂNCIA – ANOVA

1. Statistics → ANOVA

2. One-way ANOVA → OK

3. Variables: selecionar a variável dependente e a variável independente

4. Factor codes → all → zoom (conferir as variáveis) → OK

5. More results → Assumptions (Nesse passo verificar se os pressupostos estão sendo assumidos):

a. Variâncias homogêneas: é feito o teste de Levene para verifica se as variâncias são homogêneas (H_0= variâncias são homogêneas e Ha= variâncias são heterogêneas).

Clicar em Levene´s test (ANOVA) e conferir o valor de p (se p>0,05 as variâncias são homogêneas).

b. Testar a normalidade: em Distribution of within-cell residuals → Normal p-p

Conferir o gráfico. Se a distribuição é normal os resíduos seguem uma reta. Em casos de distribuição não normal é comum aparecer uma curva, principalmente em S.

6. Se não houver problemas com os pressupostos clicar na aba Summary → Univariate Results. Conferir o valor de p.

7. ATENÇÃO: Se o teste for significativo está indicando que há diferenças entre os grupos comparados. Para isso, há necessidade de se realizar comparações que podem ser:

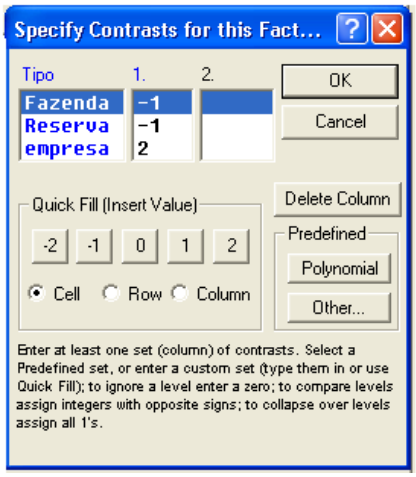
a. **Comparação a posteriori**: Anova Results → Post Hoc → Test Tukey HSD (dessa forma testa tudo contra tudo para detectar a diferença).

Para fazer o gráfico, voltar em All Effects/graphs. Colocar letras iguais para as médias iguais de acordo com o teste de Tukey.

b. **Comparação planejada**: Anova Results → Planned comps → Specify contrasts for LS means.

OBS: Como escolher o contraste?

A comparação planejada exige fundamentação teórica, pois se testa hipóteses pré-estabelecidas. Deve-se, portanto, recorrer à teoria para tomar a decisão antes de fazer o teste. Observe a figura:



* Deseja-se fazer um contraste entre Fazenda e Empresa X Reserva. Para isso selecionar em Quick Fill -1 para Fazenda; -1 para Empresa e 2 para Reserva (a soma dos contrastes deverá ser 0) → OK. Se a comparação for estatisticamente significativa (p≤ 0,05) rejeita-se a hipótese nula logo há diferença entre Fazenda e Reserva contra Empresa. Continua a análise para verificar se há diferença entre Fazenda (-1) e Reserva (+1).

3.1- Se as variâncias forem heterogêneas

Se ao testar a homogeneidade de variâncias no teste de Levene o p≤ 0,05, você deverá recorrer a algumas transformações na tentativa de homogeneizar as variâncias.

Para isso você pode transformar os dados testados em log, raiz quadrada ou arcoseno da raiz quadrada.

a. Para transformar em log:

1. Na planilha importada clique duas vezes na linha de cabeçalho do nome da variável (X, por exemplo).

2. Abaixo da janela escrever no espaço Long name (labelo r formula with Functions): = log(Variável).

b. Para transformar em raiz quadrada:

1. Na planilha importada clique duas vezes na linha de cabeçalho do nome da variável (X, por exemplo).

2. Abaixo da janela escrever no espaço Long name (labelo r formula with Functions): = Sqrt(Variável)

c. Para transformar em arco-seno da raiz quadrada:

1. Na planilha importada clique duas vezes na linha de cabeçalho do nome da variável (X, por exemplo).

2. Abaixo da janela escrever no espaço Long name (labelo r formula with Functions): = Arcsin(Sqrt(Variável)).

Depois de realizadas as transformações, repetir o teste de Levene e verificar se as variâncias se tornaram homogêneas. Caso isso não ocorra você deverá buscar outra alternativa: os testes não paramétricos.

8. Fazer o gráfico: Summary → All Effects → Graphs

9. Apresentação do resultado: F; $gl_{tratamento}$; $gl_{do\ erro}$; p

5. KRUSKAL-WALLIS

O Kruskal-Wallis é um teste de ordenamento que faz um "ranking" dos dados, para testar diferenças no somatório do "ranking" entre amostras: se a soma do "ranking" de cada tratamento é parecida entre si, os tratamentos são estatisticamente semelhantes.

H_0= a soma do ranking é estatisticamente semelhante entre os tratamentos

$H_{a=}$ a soma do ranking é estatisticamente diferente entre os tratamentos

Passos:

1. Statistics → Nonparametrics

2. Escolher o grupo de acordo com a natureza das variáveis. Por exemplo: Comparing multiple independ. samples (groups) para variáveis com mais de duas categorias → OK

3. Variables: clicar na variável dependente e na variável independente → OK

4. Summary: Kruskal-Wallis ANOVA and Median test. Aparecem duas janelas. Em uma há a soma dos ranking e o valor do teste H (gl $_{tratamento}$; N)=.....; p=..... Exemplo: Kruskal-Wallis test: H (2, N= 13) =0,231 p =0,891.

5. Fazer a comparação múltipla: Multiple comparisons of mean ranks for all group.

6. Para fazer o gráfico: voltar à janela Kruskal-Wallis → Box & whisker → seleciona a variável → seleciona o tipo Median/Quart./Range → OK.

7. Copiar o gráfico para a planilha de resultados ou para o seu documento no Word e edite-o.

8. Quando os resultados são significativos você precisa usar uma comparação a posteriori do tipo do teste de Tukey. Esse teste é o teste de Nemenyi que é explicado no Zar (1999), mas que precisará ser executado no Excel.

6. ANOVA TWO-WAY

1. Stattistics → ANOVA → Factorial ANOVA → OK

2. Variables: depedent/ independent (duas ou mais) → OK →OK

3. Testar os pressupostos:

* homogeneidade das variâncias: More Results → Assumptions → Levene´s Test (ANOVA)

* normalidade do resíduo: Normal p-p (analisar o gráfico)

4. Voltar em All Effects: aparece uma tabela e em cada linha há um valor, como no exemplo:

	SS	Degr. Of Freedom of	MS	F	p
Intercept	3110,400	1	3110,400	137,8995	0,000023
"Var1"	60,000	1	60,000	2,6601	0,154016
"Var2"	26,667	1	26,667	1,1823	0,318633
"Var1"*"Var2"	13,067	1	13,067	0,5793	0,475423
Error	135,333	6	22,556		

H_0 1: a variável 1 não afeta a germinação.

H_0 2: a variável 2 não afeta a germinação.

$H_0$3: a interação dos efeitos não afeta a germinação.

OBS: se o p da interação for significativo não precisar analisar o p dos efeitos separadamente.

5. Clicar duas vezes sobre os resultados da tabela para gerar o gráfico, aparecera uma caixa da ANOVA, clique em All effects/Graphs → OK.

OBS_1: As linhas do gráfico se cruzam quando a interação for significativa.

OBS_2: Realizar transformações dos dados se as variâncias sejam heterogêneas (logaritmo ou raiz quadrada).

OBS_3: A ANOVA two-way não tem correspondente não paramétrico.

6. Apresentação dos resultados: A melhor maneira de apresentar os resultados da ANOVA fatorial será um gráfico com média e intervalo de confiança para o efeito testado. Se a interação for significativa, apenas a interação deve ser apresentada e discutida, os efeitos individuais não poderão ser compreendidos exceto à luz do resultado da interação.

7- REGRESSÃO LINEAR

1. Statistics

2. Multiple Regression

3. Variables: dependent/independente → OK

4. Pressupostos (a distribuição dos resíduos é normal e a variância dos resíduos é homogênea)

4.1. Se a distribuição dos resíduos é normal:

→ Residuals/assumptions/prediction → Perform residuals analysis → Quick → Normal plot of residuals (análise visual)

4.2. Se a variância dos resíduos é homogênea

Residuals→ Residuals vs. independent var. → seleciona a variável independente → OK (análise visual)

OBS: se os resíduos estiverem dispostos aleatoriamente o pressuposto não foi ferido

5. Apresentação dos resultados

Graphs → Scatterplots →Variables (X=independente e Y=dependente) → Advanced → seleciona R-square e Regression equation (seleciona as variáveis X e Y)

8. REGRESSÃO MÚLTIPLA

1. Statistics→

2. Multiple regression

3. Variables: dependent variable list e predictor variables (as variáveis independentes testadas) → OK → OK

4. Para a análise dos pressupostos:

Probability plots → normal plot of residuals.

Scatteplots → predited x residual.

5. Summary → Coefficients

5.1. Verificar o valor de p das variáveis (quando for significativo observar o tamanho do efeito de acordo com os parâmetros).

5.2. Observar também o intervalo de confiança (a 95%) ao redor dos parâmetros (a inclinação da reta).

5.3. o valor de β é o R^2

OBS: As variáveis correlacionadas não podem entrar juntas na regressão múltipla

6. All effects (pegar o F e os graus de liberdade)

7. Gráfico: é importante verificar se havia co-relação entre as variáveis. Fazer o gráfico com a(s) variável(is) que for(em) significativa(s).

-Graphs → Scatterplots → Quick → selecionar as variáveis → em Graph type marcar - Multiple

- Advanced → marcar R-square e Regression equation

8. Apresentação dos resultados

9- REGRESSÃO LOGÍSTICA

1. Statistics→ Advanced Linear/ Nonlinear models → Nonlinear Estimation → Quick Logit regression → OK

2. Variables: dependent variable/ independent variable

3. Codes for dep. var: 0

and: 1 (Sempre colocar o 0 em cima e o 1 em baixo) → OK

4. Advanced → Estimation method: Quase-Newton → marcar Asymptotic standard errors → OK

5. Aparece na janela o valor de χ^2 e p.

6. Para fazer o gráfico: → Fitted 2D function & observed vals.

7. Para calcular a estimativa dos parâmetros: → Summary: Parameters & standard errors

10- REGRESSÃO LOGÍSTICA MULTIPLA

1. Statistics→ Advanced Linear/ Nonlinear models → Nonlinear Estimation → Quick Logit regression → OK

2. Variables: dependent variable/ independent variable

3. Advanced → Estimation method: Quase-Newton → marcar Asymptotic standard errors → OK

4. Marcar Difference from previous models

5. Gráfico → Graphs → Mean w/ error plots

6. Quick → seleciona variáveis → Advanced+ tudo certo ok → fazer gráficos.

7. Inserir a equação. Pegar a equação do gráfico 2D feito primeiramente → All options→ Custom function→ Add new function→ Y=(colar a equação) →ok

10- ANCOVA – ANÁLISE DE COVARIÂNCIA

1. Statistics→ Advanced Linear/ Nonlinear models → General Linear Models→ OK → OK

2. General Linear Models → OK

3. Variables: dependent variable/ independent variable categorical predit. e continuos predit. → OK → OK

4. Between effects → OK

5. Use custom effects for the betweens design → marcar a variável categorical e continuous →full factorial →OK →OK

6. All effect graphs

More results → Assumptions (Nesse passo verificar se os pressupostos estão sendo assumidos):

a. Variâncias homogêneas: é feito o teste de Levene para verifica se as variâncias são homogêneas (H_0= variâncias são homogêneas e Ha= variâncias são heterogêneas).

Clicar em Levene´s test (ANOVA) e conferir o valor de p (se p>0,05 as variâncias são homogêneas).

b. Testar a normalidade: em Distribution of within-cell residuals → Normal p-p

Conferir o gráfico. Se a distribuição é normal os resíduos seguem uma reta. Em casos de distribuição não normal é comum aparecer uma curva, principalmente em S.

7. Se não houver problemas com os pressupostos clicar na aba Summary → Univariate Results. Conferir o valor de p.

OK

Homogeneity of slopes model ok

Para fazer o gráfico

1 Graphs → Scatterplots →Variables (X=independente e Y=dependente) → Advanced → seleciona R-square →OK

2. Categorized →X-categories →ON

3. Change variable →marcar a variável categórica →OK

a. observar a inclinação da reta, se for paralela é porque não existe interação.

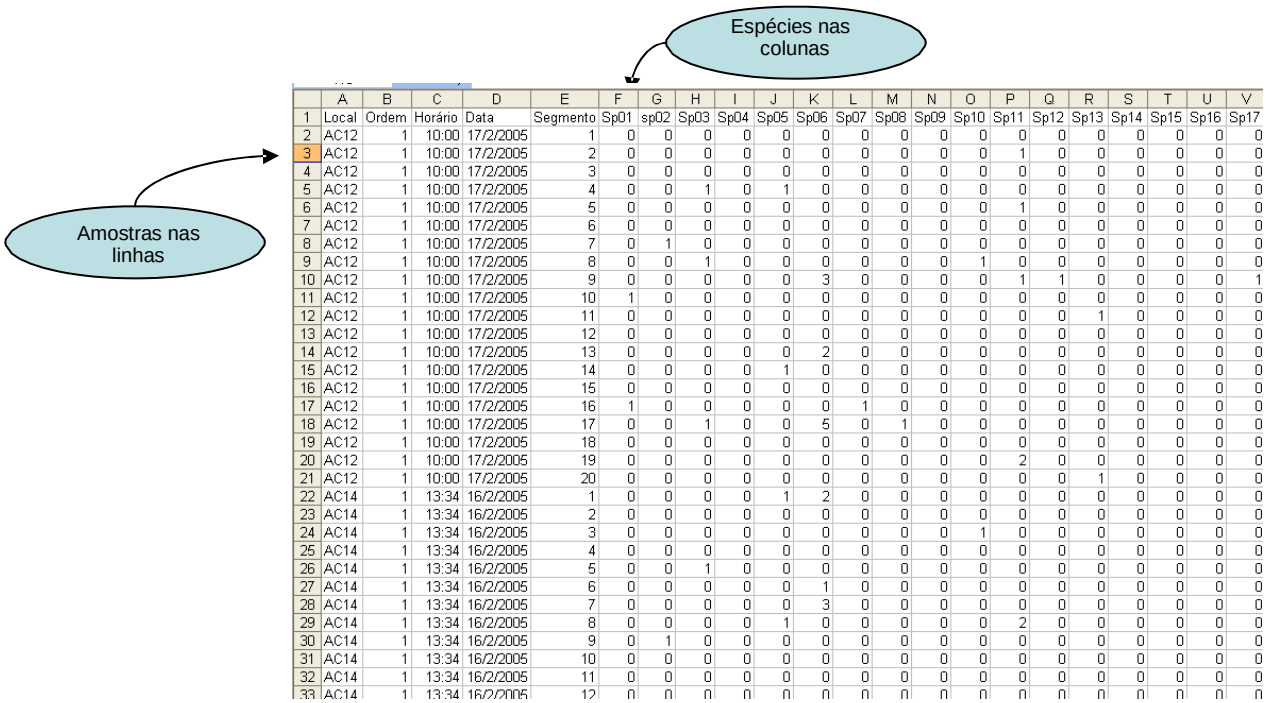
TUTORIAL PARA PREPARAÇÃO E IMPORTAÇÃO DE DADOS PARA ESTIMATIVAS DE RIQUEZA DE ESPÉCIES

Softwares utilizados: Excel, EstimateS e Statistica.

PREPARAÇÃO DOS DADOS

Os dados de suas coletas devem ser organizados em uma planilha eletrônica, pois as análises subseqüentes podem ser feitas de modo simples por meio de pequenas modificações na estrutura das mesmas. Neste caso utilizamos as planilhas do Microsoft Excel® para demonstrar como importar os dados para o programa EstimateSWin 750.

Como estaremos trabalhando com estimativas sobre espécies, devemos organizar a planilha da seguinte forma:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Local	Ordem	Horário	Data	Segmento	Sp01	sp02	Sp03	Sp04	Sp05	Sp06	Sp07	Sp08	Sp09	Sp10	Sp11	Sp12	Sp13	Sp14	Sp15	Sp16	Sp17
2	AC12	1	10:00	17/2/2005	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	AC12	1	10:00	17/2/2005	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
4	AC12	1	10:00	17/2/2005	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	AC12	1	10:00	17/2/2005	4	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	AC12	1	10:00	17/2/2005	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	AC12	1	10:00	17/2/2005	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	AC12	1	10:00	17/2/2005	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	AC12	1	10:00	17/2/2005	8	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	AC12	1	10:00	17/2/2005	9	0	0	0	0	0	3	0	0	0	0	1	1	0	0	0	0	1
11	AC12	1	10:00	17/2/2005	10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	AC12	1	10:00	17/2/2005	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
13	AC12	1	10:00	17/2/2005	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	AC12	1	10:00	17/2/2005	13	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
15	AC12	1	10:00	17/2/2005	14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
16	AC12	1	10:00	17/2/2005	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	AC12	1	10:00	17/2/2005	16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
18	AC12	1	10:00	17/2/2005	17	0	0	1	0	0	5	0	1	0	0	0	0	0	0	0	0	0
19	AC12	1	10:00	17/2/2005	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	AC12	1	10:00	17/2/2005	19	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
21	AC12	1	10:00	17/2/2005	20	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
22	AC14	1	13:34	16/2/2005	1	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
23	AC14	1	13:34	16/2/2005	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	AC14	1	13:34	16/2/2005	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
25	AC14	1	13:34	16/2/2005	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	AC14	1	13:34	16/2/2005	5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	AC14	1	13:34	16/2/2005	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
28	AC14	1	13:34	16/2/2005	7	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
29	AC14	1	13:34	16/2/2005	8	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0
30	AC14	1	13:34	16/2/2005	9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	AC14	1	13:34	16/2/2005	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	AC14	1	13:34	16/2/2005	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	AC14	1	13:34	16/2/2005	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

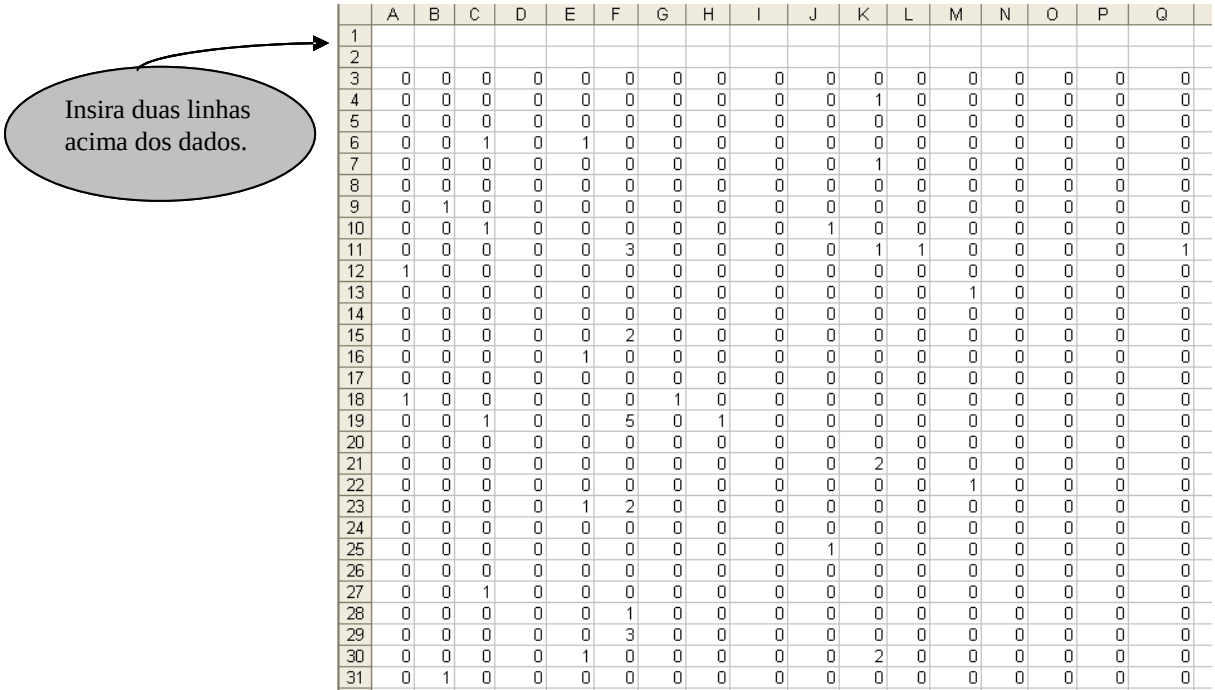
O programa EstimateS precisa que formatemos a planilha de um modo bastante específico, com a criação de um cabeçalho que o programa lerá durante a importação. Antes de criar o cabeçalho, devemos remover todo e qualquer tipo de recurso complexo do Excel, tais como comentários, acentos e os chamados caracteres diacríticos: (“ “ ? ¿ / > < @ ! ~ ` ; ‘ & % # $ * { } [ ] ( ) ç - +).

Obs: a presença desses caracteres é a causa mais freqüente de erros de importação e análise de dados nos mais diversos programas estatísticos. Eles não devem ser utilizados nas planilhas e nem em nome de arquivos.

Removidas tais características das planilhas, devemos também remover as colunas que identificam as amostras e a linha que identifica o nome de cada espécie. Isso é necessário, pois o programa irá aleatorizar indivíduos nas amostras, numa tentativa de remover ou diminuir o vício de coleta presente nas mesmas antes de calcular as estimativas de riqueza ou quaisquer índices. Como trabalhamos com riqueza, a identidade de cada espécie também não é necessária. A planilha assumirá o aspecto abaixo:

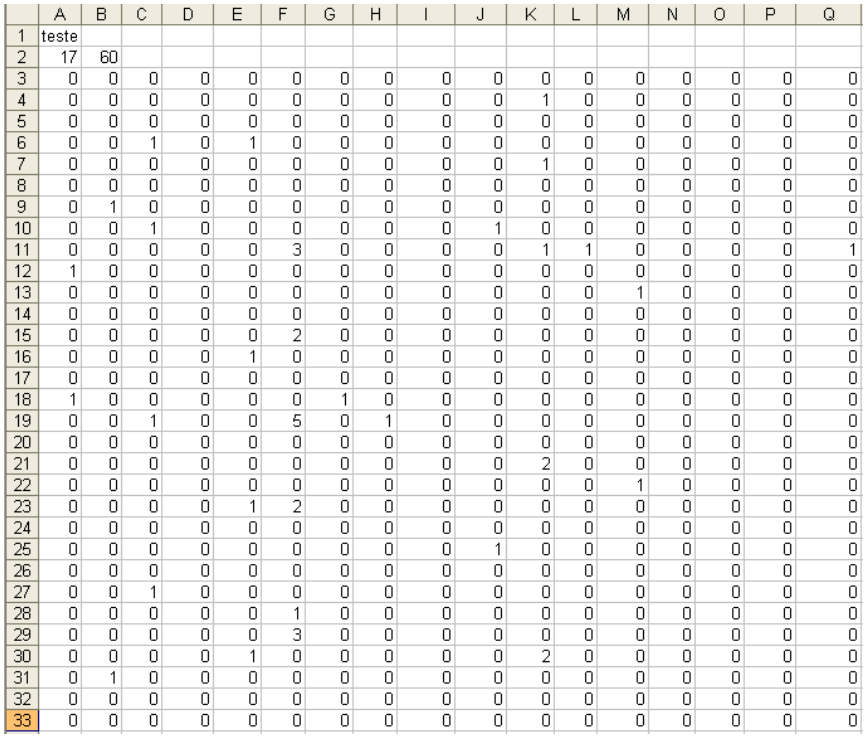
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	3	0	0	0	0	1	1	0	0	0	0	1
10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
17	0	0	1	0	0	5	0	1	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
21	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0
29	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1

Agora devemos inserir duas linhas acima dos dados. Elas servirão para o cabeçalho de legenda para o EstimateS:



Na primeira célula (A1) devemos inserir o nome que daremos para a planilha, deve ser um nome curto, com menos de seis dígitos e que não contenha diacríticos.

Na célula (A2) devemos inserir o número de espécies (que é o número de colunas) e na célula (B2)o número de amostras (linhas) respectivamente. A planilha apresentará o seguinte aspecto:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	teste																
2	17	60															
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	3	0	0	0	0	1	1	0	0	0	0	1
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
19	0	0	1	0	0	5	0	1	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
23	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0
31	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A planilha está quase pronta. É necessário salvá-la como somente texto separado por tabulações, indo em: Arquivo → Salvar como → Texto separado por tabulações.

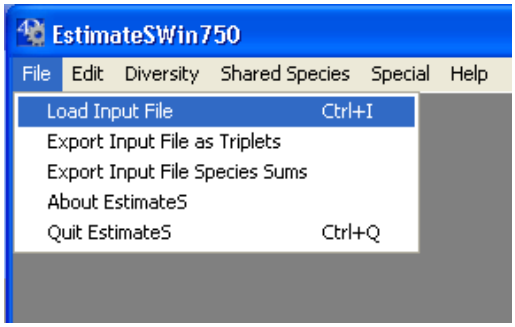
O Excel apresentará algumas mensagens de alerta antes de permitir que você salve o documento. Ignore-as e continue o processo.

Pronto: agora podemos fechar o Excel e abrir o EstimateS.

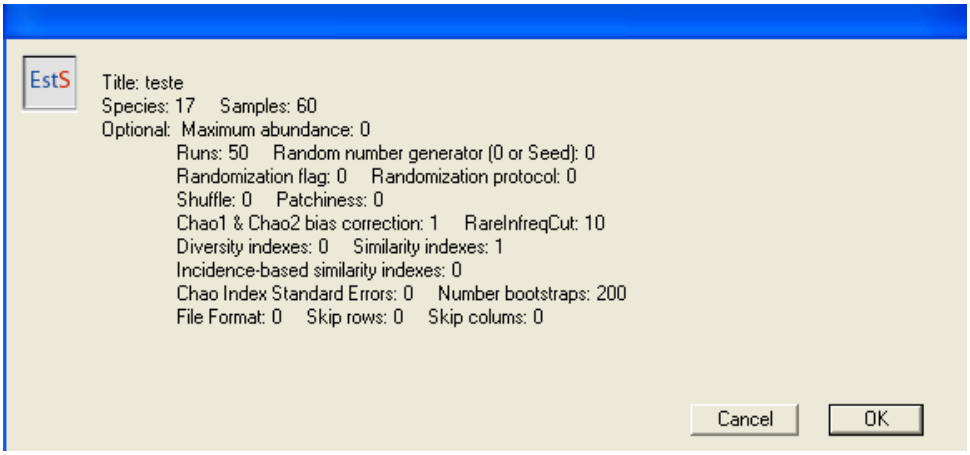
Logo que o programa é aberto, uma tela de apresentação é exibida. É só dar OK e começar a usar.

IMPORTANTE: Se o programa não abrir pode ser devido a uma configuração de seu computador. O EstimateS esta configurado no sistema Britânico cujo separador decimal é o "." (ponto), e no nosso sistema é a vírgula. Para resolver este problema, basta ir**:** Iniciar → Configurações→ Painel de controle → Opções regionais e de idioma → Personalizar → Símbolo decimal trocar vírgula por ponto → OK → OK e fecha a janela aberta. Agora é só abrir o EstimateS novamente e começar a trabalhar.

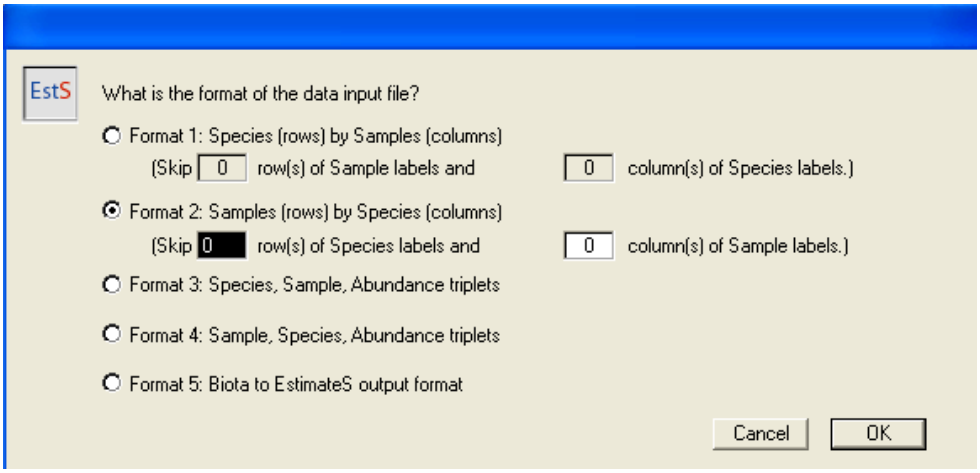
Para importar os dados que preparamos, basta ir em File → Load Input File



Uma janela do Explorer irá abrir e é só selecionarmos o arquivo de texto que preparamos antes. Ao fazer isso o programa exibirá a seguinte tela:

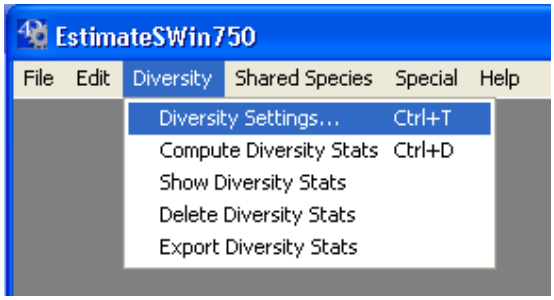


Dê OK. A seguinte tela aparecerá:

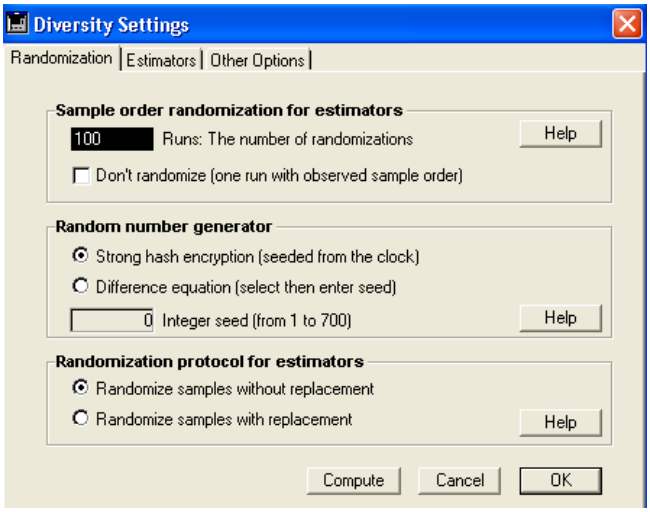


Marque a caixa com a opção Formato 2 (linhas nas amostras e espécies nas colunas) e dê OK. O programa deverá carregar a planilha na memória. Se tudo der certo não haverá nenhuma mensagem de erro.

Prossiga então clicando no menu DIVERSITY → DIVERSITY SETTINGS...



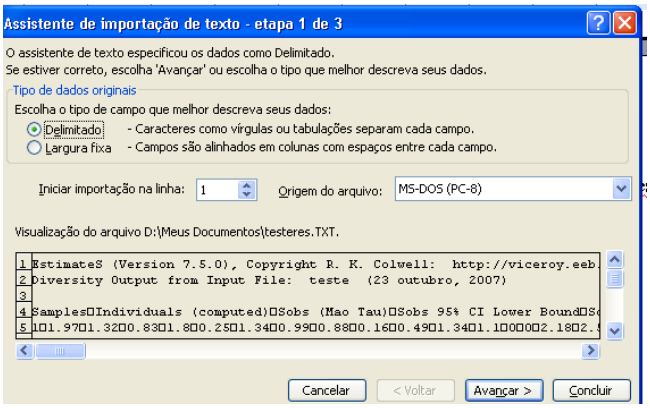
Aparecerá a seguinte tela:



O padrão para o número de “runs” (aleatorizações) é 50. Normalmente marcamos como 100 ou mais vezes, depende do tamanho do conjunto de dados que você possui. Como a re-amostragem do principal estimador de riqueza de espécies é sem reposição, devemos manter selecionada essa opção na caixa de Protocolo de Aleatorização. Clique em Compute.

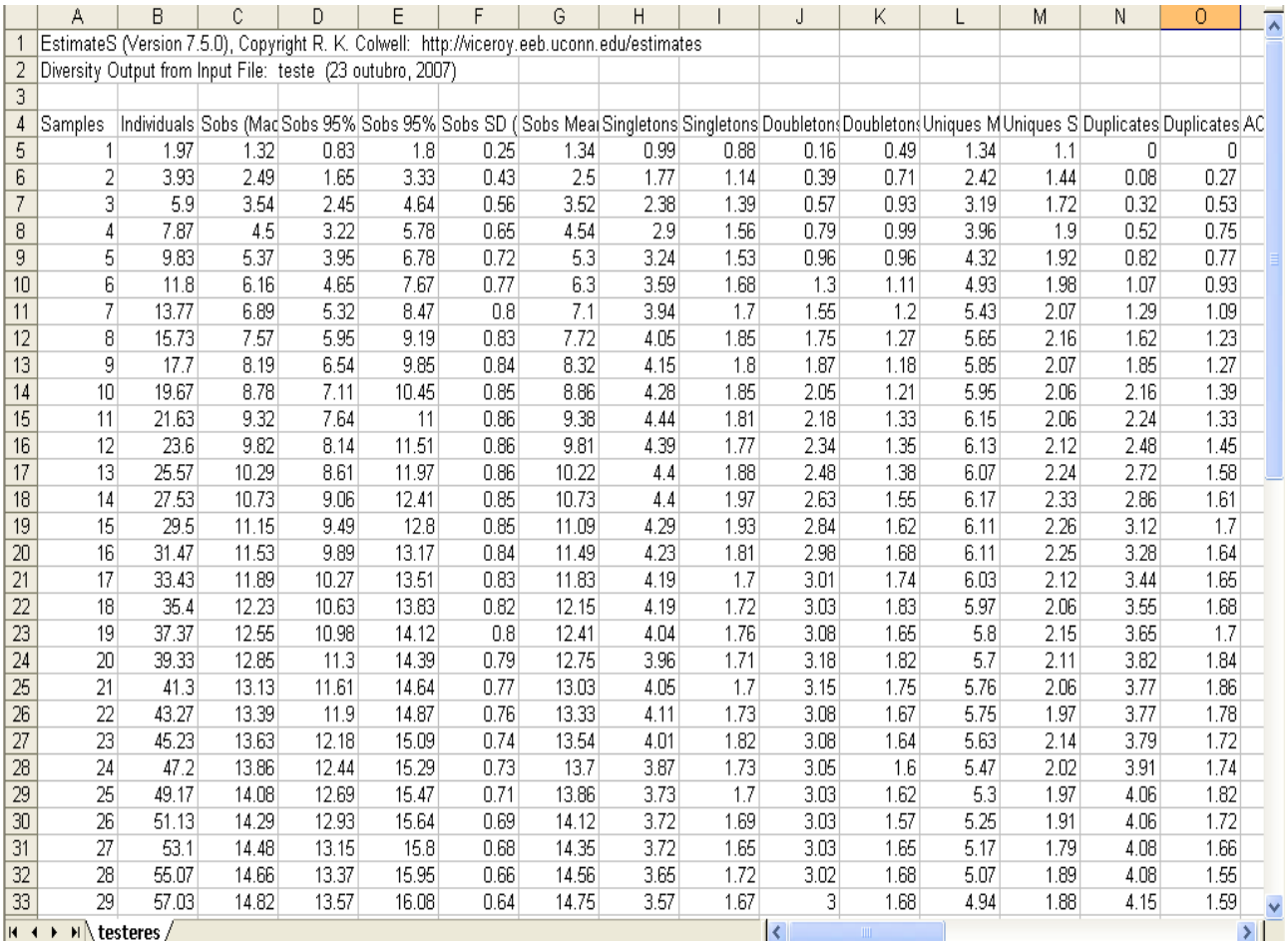
Ao fim desse tempo, você verá uma planilha com os resultados calculados. Essa planilha não é prática e é preferível trabalhar com os dados no Excel. Clique em Export aparecerá uma tela do Explorer, dê um nome para seu arquivo (sugerimos que seja dado o mesmo nome do arquivo original, adicionado com a denominação res de resultado, isso evita problemas de mistura de resultados, no nosso exemplo demos o nome de teste.txt, agora passar a ser testeres.txt e feche o Estimates. É hora de abrir o Excel.

Com o Excel, abra o arquivo de texto que foi a saída do programa Estimates. O Excel apresentará uma tela sobre definições sobre a importação de dados no formato texto.



O padrão do programa está correto, bastando clicar em concluir.

Exclua as três primeiras linhas da planilha, são apenas propaganda do programa EstimateS. Após isso, é só salvar como uma planilha do Excel e fechar. Agora vamos importar essa planilha para o programa Statistica 6.0 ou outra versão mais atualizada (você já deve estar craque nessa parte!).



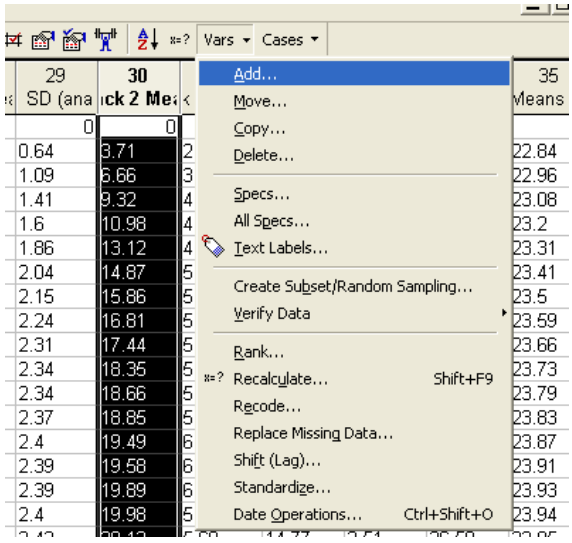
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	EstimateS (Version 7.5.0), Copyright R. K. Colwell: http://viceroy.eeb.uconn.edu/estimates														
2	Diversity Output from Input File: teste (23 outubro, 2007)														
3															
4	Samples	Individuals	Sobs (Mac	Sobs 95%	Sobs 95%	Sobs SD (	Sobs Mea	Singletons	Singletons	Doubletons	Doubletons	Uniques M	Uniques S	Duplicates	Duplicates	AC
5	1	1.97	1.32	0.83	1.8	0.25	1.34	0.99	0.88	0.16	0.49	1.34	1.1	0	0	
6	2	3.93	2.49	1.65	3.33	0.43	2.5	1.77	1.14	0.39	0.71	2.42	1.44	0.08	0.27	
7	3	5.9	3.54	2.45	4.64	0.56	3.52	2.38	1.39	0.57	0.93	3.19	1.72	0.32	0.53	
8	4	7.87	4.5	3.22	5.78	0.65	4.54	2.9	1.56	0.79	0.99	3.96	1.9	0.52	0.75	
9	5	9.83	5.37	3.95	6.78	0.72	5.3	3.24	1.53	0.96	0.96	4.32	1.92	0.82	0.77	
10	6	11.8	6.16	4.65	7.67	0.77	6.3	3.59	1.68	1.3	1.11	4.93	1.98	1.07	0.93	
11	7	13.77	6.89	5.32	8.47	0.8	7.1	3.94	1.7	1.55	1.2	5.43	2.07	1.29	1.09	
12	8	15.73	7.57	5.95	9.19	0.83	7.72	4.05	1.85	1.75	1.27	5.65	2.16	1.62	1.23	
13	9	17.7	8.19	6.54	9.85	0.84	8.32	4.15	1.8	1.87	1.18	5.85	2.07	1.85	1.27	
14	10	19.67	8.78	7.11	10.45	0.85	8.86	4.28	1.85	2.05	1.21	5.95	2.06	2.16	1.39	
15	11	21.63	9.32	7.64	11	0.86	9.38	4.44	1.81	2.18	1.33	6.15	2.06	2.24	1.33	
16	12	23.6	9.82	8.14	11.51	0.86	9.81	4.39	1.77	2.34	1.35	6.13	2.12	2.48	1.45	
17	13	25.57	10.29	8.61	11.97	0.86	10.22	4.4	1.88	2.48	1.38	6.07	2.24	2.72	1.58	
18	14	27.53	10.73	9.06	12.41	0.85	10.73	4.4	1.97	2.63	1.55	6.17	2.33	2.86	1.61	
19	15	29.5	11.15	9.49	12.8	0.85	11.09	4.29	1.93	2.84	1.62	6.11	2.26	3.12	1.7	
20	16	31.47	11.53	9.89	13.17	0.84	11.49	4.23	1.81	2.98	1.68	6.11	2.25	3.28	1.64	
21	17	33.43	11.89	10.27	13.51	0.83	11.83	4.19	1.7	3.01	1.74	6.03	2.12	3.44	1.65	
22	18	35.4	12.23	10.63	13.83	0.82	12.15	4.19	1.72	3.03	1.83	5.97	2.06	3.55	1.68	
23	19	37.37	12.55	10.98	14.12	0.8	12.41	4.04	1.76	3.08	1.65	5.8	2.15	3.65	1.7	
24	20	39.33	12.85	11.3	14.39	0.79	12.75	3.96	1.71	3.18	1.82	5.7	2.11	3.82	1.84	
25	21	41.3	13.13	11.61	14.64	0.77	13.03	4.05	1.7	3.15	1.75	5.76	2.06	3.77	1.86	
26	22	43.27	13.39	11.9	14.87	0.76	13.33	4.11	1.73	3.08	1.67	5.75	1.97	3.77	1.78	
27	23	45.23	13.63	12.18	15.09	0.74	13.54	4.01	1.82	3.08	1.64	5.63	2.14	3.79	1.72	
28	24	47.2	13.86	12.44	15.29	0.73	13.7	3.87	1.73	3.05	1.6	5.47	2.02	3.91	1.74	
29	25	49.17	14.08	12.69	15.47	0.71	13.86	3.73	1.7	3.03	1.62	5.3	1.97	4.06	1.82	
30	26	51.13	14.29	12.93	15.64	0.69	14.12	3.72	1.69	3.03	1.57	5.25	1.91	4.06	1.72	
31	27	53.1	14.48	13.15	15.8	0.68	14.35	3.72	1.65	3.03	1.65	5.17	1.79	4.08	1.66	
32	28	55.07	14.66	13.37	15.95	0.66	14.56	3.65	1.72	3.02	1.68	5.07	1.89	4.08	1.55	
33	29	57.03	14.82	13.57	16.08	0.64	14.75	3.57	1.67	3	1.68	4.94	1.88	4.15	1.59	

testeres

Após importar a planilha para o Statistica, devemos escolher o estimador de riqueza de espécies desejado. Verifique que há um valor estimado para cada uma de suas amostras, o que permite a você a criação de uma curva do coletor. Note também que para cada estimativa há também um desvio padrão. De posse desse dado, é possível construir um intervalo de confiança associado à estimativa, o que irá permitir a apresentação dos dados em um gráfico mais informativo que poderá inclusive ser utilizado na comparação de riqueza de espécies entre locais. Como construir esse intervalo e como fazer esse gráfico? Basta seguir os passos adiante.

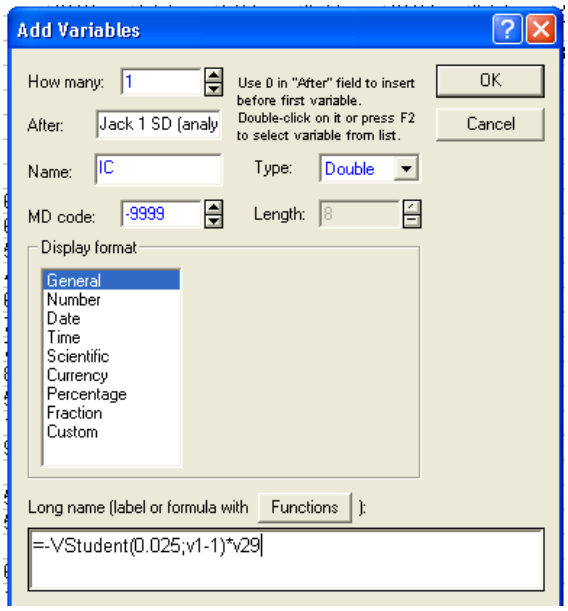
Para esse exemplo, utilizaremos o estimador não paramétrico Jackknife de primeira ordem. Esse estimador é bem interessante. Recomendamos a leitura dos artigos e livros que o discutem. Dentre os vários livros, o Ecological Methodology do Krebs é um bom início.

Para criar o intervalo de confiança precisamos primeiro inserir mais uma coluna na planilha dentro do Statistica, para isso selecione a coluna imediatamente posterior à direita da coluna do desvio padrão, no caso a coluna 30. Localize no lado direito da tela do Statistica o menu VARS, clique em adicionar.



Será aberta a seguinte tela, onde podemos configurar o conteúdo da Coluna (que o Statistica sabidamente chama de variável). Ele indica que a variável será adicionada após a coluna Jack1_SD. O nome da variável fica a seu critério. Mas IC já diz tudo.

Agora vem o importante: Vamos inserir uma fórmula no campo maior dessa tela, que será utilizada para criar o intervalo de confiança.



Como no Excel, toda a formula deve começar com o sinal de igual (=) e o que digitaremos é o seguinte =-vstudent(0,025;v1-1)*Vn

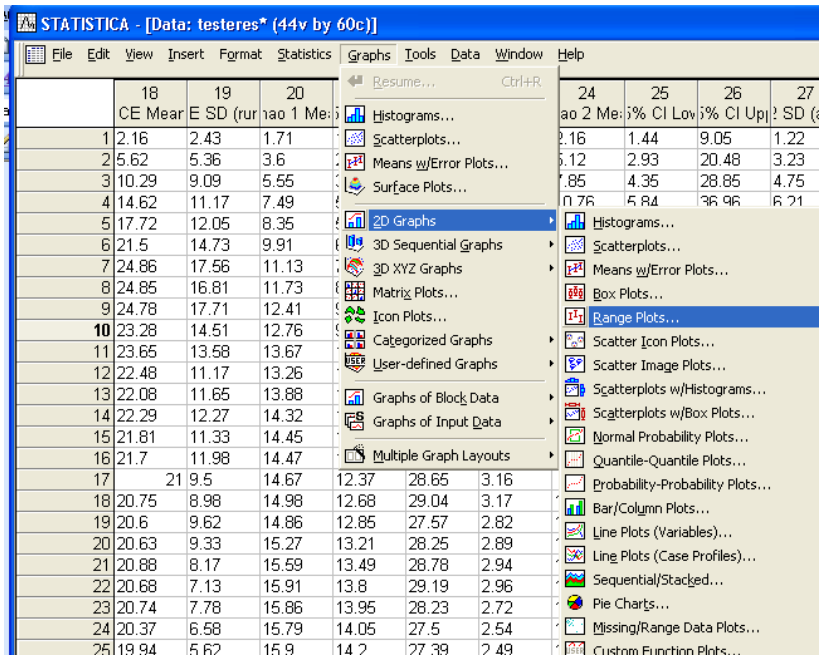
Onde –vstudent diz para utilizar a distribuição de Student (a mesma distribuição do teste t) 0,025 é o nosso alfa, já que o teste é bicaudal (0,025 + 0,025 = α = 0,05)

v1-1 é o número de amostras menos 1, ou seja, o grau de liberdade.

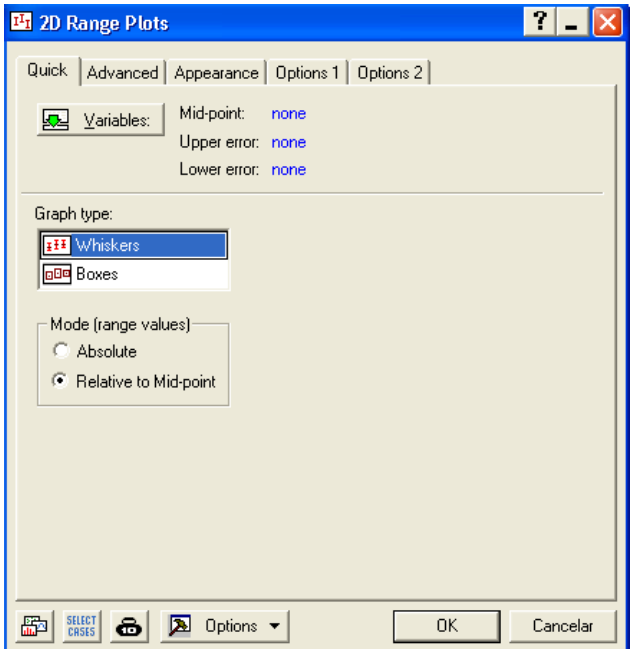
Vn deve ser substituído pelo nome da variável que contém o desvio padrão (no caso V24).

NOTA: A fórmula para cálculo do intervalo de confiança deveria ser =-vstudent(0,025;v1-1)*Vn/sqrt(v1), ou seja, deveríamos dividir o desvio padrão pela raiz quadrada de n (v1) para obter o erro padrão e aí sim multiplicar pelo resto da fórmula para conseguirmos o intervalo desejado. Mas o programa EstimateS fornece o erro padrão e o chama de desvio padrão.

Com a nova coluna podemos criar o nosso gráfico. É só ir em GRAPHS → 2D Graphs → Range plots.

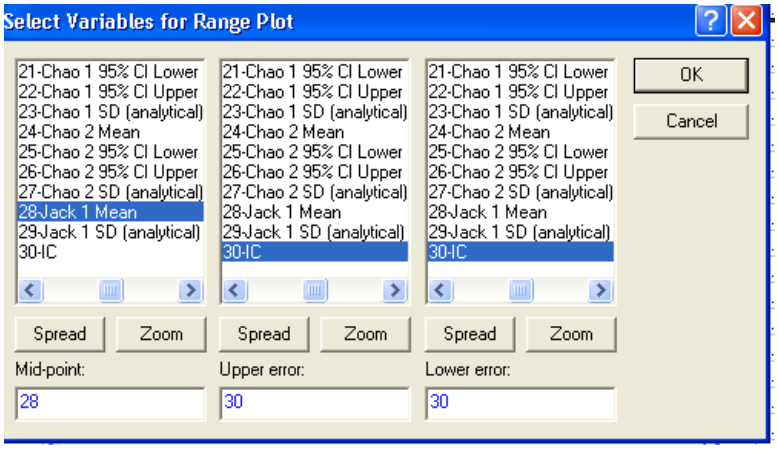


Devemos marcar a opção “relativo a um ponto central”

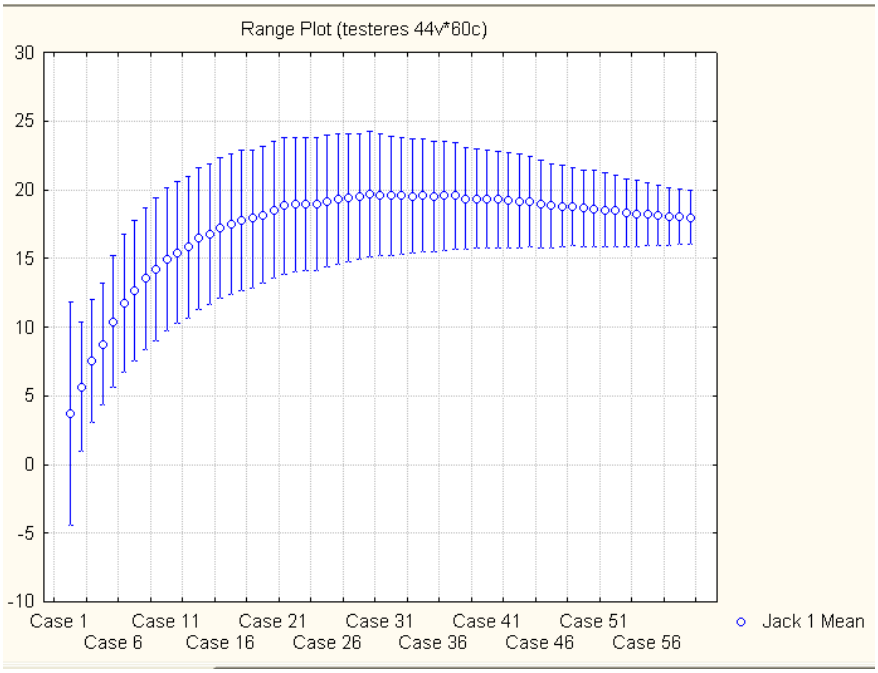


E clicar em Variables para defini-las.

Devemos selecionar a estimativa Jackknife como ponto central e o limite inferior e superior como o intervalo de confiança que criamos.

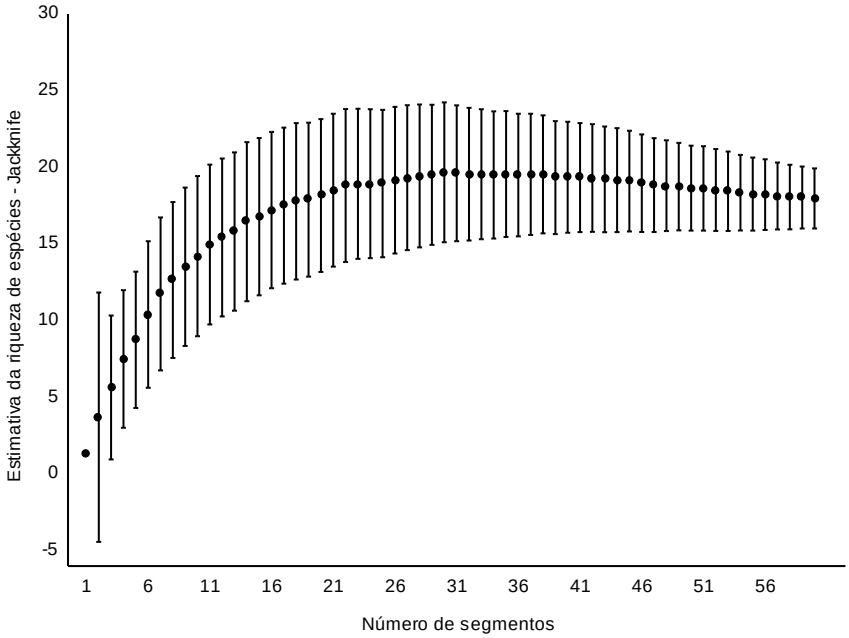


Quando clicamos em OK veremos o nosso gráfico de acumulação de espécies com o IC de 95% associado à estimativa.



Há diversas maneiras de personalizar esse gráfico para importá-lo para o Word ou qualquer outro editor de texto. Vale a pena a cada um aprender qual opção se ajusta melhor às suas necessidades ou de acordo com a regra de uma revista científica.

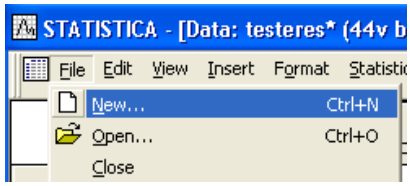
O mesmo gráfico já trabalhado pode ficar assim, por exemplo:



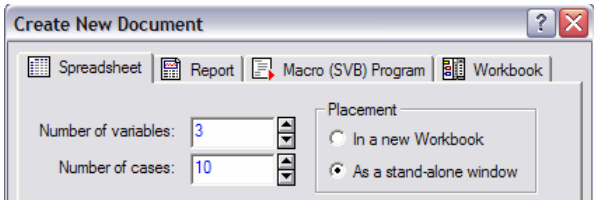
Com poucas modificações nas planilhas você pode criar um gráfico que apresente no eixo X os locais e no eixo Y as estimativas de riqueza de espécies. Com a presença do intervalo de confiança teremos um teste estatístico visual para comparação entre áreas distintas. Duas áreas serão iguais se o limite do intervalo de confiança de uma alcançar o valor central da estimativa do outro.

Para criar um gráfico que mescle as estimativas de riqueza (com IC associado) de duas ou mais áreas, é só realizar os procedimentos acima descritos para cada uma delas e reunir a última linha (último valor estimado) de três colunas na planilha já importada do Statistica.

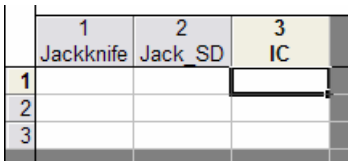
Primeiro vamos criar uma nova planilha:



O número de variáveis é 3 (só pela facilidade de copiar e colar entre planilhas, pegamos a variável com o desvio padrão). O número de "cases" ou amostras é o número de locais que você quer comparar.

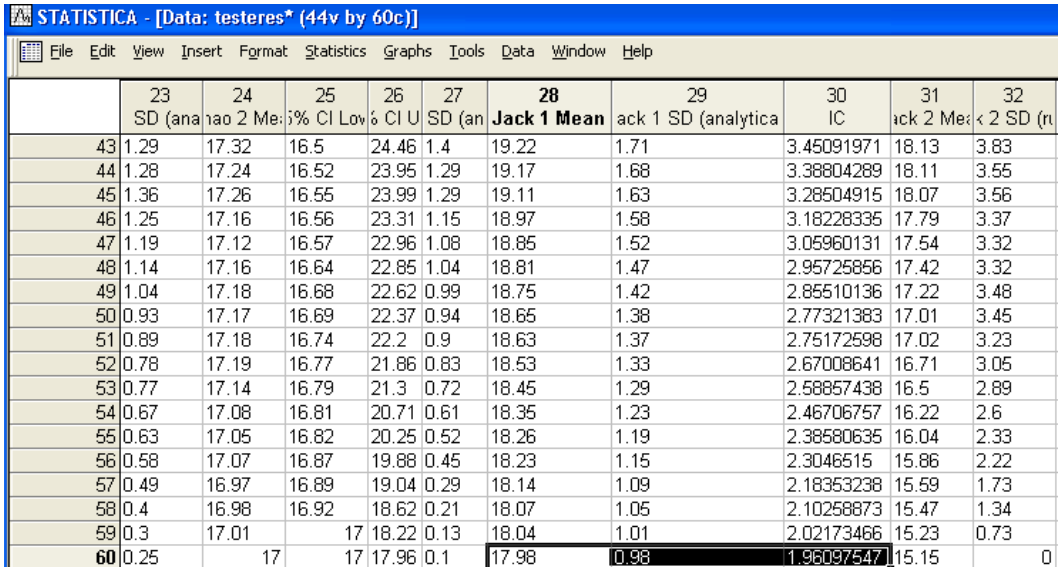


Renomeamos as variáveis:



	1 Jackknife	2 Jack_SD	3 IC
1			
2			
3			

Vamos na planilha do primeiro local e copiamos a última linha das 3 colunas que precisamos:

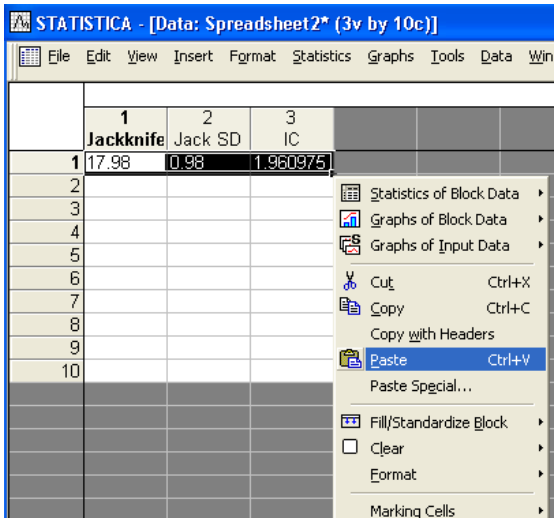


STATISTICA - [Data: testeres* (44v by 60c)]

File Edit View Insert Format Statistics Graphs Tools Data Window Help

	23 SD (ana	24 nao 2 Me	25 % CI Low	26 % CI U	27 SD (an	28 **Jack 1 Mean**	29 ack 1 SD (analytica	30 IC	31 ack 2 Mea	32 < 2 SD (n
43	1.29	17.32	16.5	24.46	1.4	19.22	1.71	3.45091971	18.13	3.83
44	1.28	17.24	16.52	23.95	1.29	19.17	1.68	3.38804289	18.11	3.55
45	1.36	17.26	16.55	23.99	1.29	19.11	1.63	3.28504915	18.07	3.56
46	1.25	17.16	16.56	23.31	1.15	18.97	1.58	3.18228335	17.79	3.37
47	1.19	17.12	16.57	22.96	1.08	18.85	1.52	3.05960131	17.54	3.32
48	1.14	17.16	16.64	22.85	1.04	18.81	1.47	2.95725856	17.42	3.32
49	1.04	17.18	16.68	22.62	0.99	18.75	1.42	2.85510136	17.22	3.48
50	0.93	17.17	16.69	22.37	0.94	18.65	1.38	2.77321383	17.01	3.45
51	0.89	17.18	16.74	22.2	0.9	18.63	1.37	2.75172598	17.02	3.23
52	0.78	17.19	16.77	21.86	0.83	18.53	1.33	2.67008641	16.71	3.05
53	0.77	17.14	16.79	21.3	0.72	18.45	1.29	2.58857438	16.5	2.89
54	0.67	17.08	16.81	20.71	0.61	18.35	1.23	2.46706757	16.22	2.6
55	0.63	17.05	16.82	20.25	0.52	18.26	1.19	2.38580635	16.04	2.33
56	0.58	17.07	16.87	19.88	0.45	18.23	1.15	2.3046515	15.86	2.22
57	0.49	16.97	16.89	19.04	0.29	18.14	1.09	2.18353238	15.59	1.73
58	0.4	16.98	16.92	18.62	0.21	18.07	1.05	2.10258873	15.47	1.34
59	0.3	17.01	17	18.22	0.13	18.04	1.01	2.02173466	15.23	0.73
60	0.25	17	17	17.96	0.1	17.98	0.98	1.96097547	15.15	0

Vamos agora para a planilha que criamos e mandamos colar na linha desejada:



Clicando duas vezes sobre a coluna externa que normalmente contém o número das linhas, podemos modificá-las e inserir o nome dos locais que desejamos comparar. Realizamos o "copiar e colar" para cada local sucessivamente até completar a planilha.

	1 Jackknife	2 Jack SD	3 IC
Local 1	17.98	0.98	1.960975
Local 2			
Local x			
4			

Depois dessa planilha estar pronta, é só criar o gráfico de “range plot” como explicado anteriormente para criação da curva do coletor e efetivamente comparar os locais.

Estudo de Caso:

Para exemplificar todos os passos do procedimento Jackknife, vamos usar a tabela planilha teste original. Conforme pode ser verificado, existem três rios onde foram coletadas espécies da Ordem Odonata na Amazônia, dois rios de primeira ordem Ac12 e Ac14, e um de segunda ordem Ac22.

Calculem a riqueza estimada de cada rio, e construa o gráfico comparando a riqueza das três áreas, para ver qual é a mais diversa. Ao final compare seus resultados com a planilha e com o gráfico abaixo.

Local	Jackknife	jack_SD	IC
AC12	17.7	2.43	5.086048
AC14	18.7	3.11	6.509305
AC22	25.55	3.23	6.760468