Aplicações Bioestatísticas

para pesquisadores e pós-graduandos

Prof. J.C. Sena Maia *
Departamento de Estatística
Universidade Federal do Paraná
(* aposentado)

Apresentação da Proposta

A proposta deste livro virtual não é ensinar Bioestatística seguindo a programação de um curso regular em nível de pós-graduação. Assumimos que os futuros usuários já detenham os conhecimentos básicos da matéria e que estejam na fase de aplicação dos processos estatísticos, analisando experimentos e interpretando resultados. Com esse objetivo selecionamos os principais testes que são usualmente utilizados em pesquisas biológicas e desenvolvemos programas em linguagem Fortran 90 para a resolução dos mesmos. Embora exista no mercado um grande número de aplicativos que podem ser utilizados com sucesso na análise de problemas bioestatísticos, optamos por desenvolver os próprios programas tendo em vista a elaboração de um manual satisfatoriamente didático e interativo que facilitasse a abordagem dos diversos temas. O tutorial está adequadamente ilustrados, clarificando cada etapa do processo analítico. Para os diversos programas foram abordados aspectos do planejamento, da coleta dos dados, e da interpretação dos resultados. A introdução das informações no computador pode ser feita seguindo as instruções reunidas no tutorial de cada programa, facilitando consideravelmente a utilização dos processos automatizados.

Organização do conteúdo

Os testes serão tratados separadamente e os respectivos tutoriais estão organizados em boxes coloridos, do seguinte modo: o box amarelo mostra o que será visto na tela do computador. No box verde apresentaremos os dados biológicos que serão analisados, faremos os comentários relativos à digitação das informações, ao tratamento que elas receberão na execução do teste e aos resultados obtidos. Complementarmente, fora dos boxes, faremos a interpretações dos resultados estatísticos bem como a extensão das conclusões para os dados biológicos com as devidas limitações e cautelas.

A automação dos cálculos

A cada teste incluído neste manual corresponde um programa desenvolvido na linguagem Fortran 90. Não obstante as linguagens mais modernas, o FORTRAN continua sendo, em minha opinião, uma linguagem que proporciona bons resultados no tratamento de fórmulas e equações matemáticas e estatísticas.

Os programas executáveis contemplando 17 aplicações foram reunidos em um único aplicativo denominado

aplicativoBioestat.f

juntamente com um tutorial denominado **"Tutorial (PDF)"** cujos downloads podem ser efetuados através de um CtrlC + CtrlV no link do Departamento de Estatística da UFPR, mostrado no quadro abaixo:

https://www.est.ufpr.br/software.html

Os PDFs deste tutorial foram dimensionados de modo a permitir a leitura amigável inclusive em tablets e smartphones. A ideia é facilitar o treinamento permitindo o uso simultâneo de vários equipamentos de acordo com as conveniências dos usuários.

Lembrete importante. O boxes coloridos deste tutorial obedecem à seguinte convenção de acordo com a respectiva cor:

Imagens da tela do computador que interagem com o usuário

Informações adicionais e comentários sobre os testes estatísticos

Considerações sobre amostragem bioestatística

Nas pesquisas biológicas é muito comum trabalharmos com populações de grande tamanho e amplamente distribuídas. Isso inviabiliza a alternativa censitária (i.e. trabalhar com todos os integrantes da população) e nos direciona para o uso da amostragem.

A amostragem está para a pesquisa biológica assim como o alicerce está para a construção de uma casa. O "pedaço" da população que utilizamos em nossas análises precisa ser representativo do universo que o originou, de modo a validar as conclusões dos testes aplicados. Em outras palavras, em cima de uma amostra de má qualidade a mais brilhante e sofisticada análise bioestatística não passa de um elegante e inútil exercício intelectual que pode nos conduzir a conclusões temerárias.

Vamos inicialmente revisar alguns conceitos e definições nessa fase introdutória.

O primeiro conceito é o de população. Os conceitos biológico e estatístico diferem sensivelmente. A população biológica normalmente tem limites naturais impostos por características intrínsecas da(s) espécie(s) estudada(s) tais como fatores genéticos, comportamentais, ambientais, geográficos e outros. Já a população estatística tem limites inteiramente arbitrários, impostos pelos pesquisadores de acordo com as necessidades, limitações e conveniências ligadas à execução da pesquisa

Assim, por exemplo, um entomologista que esteja estudando a abelha ***Apis melífera*** **pode definir a sua população alvo como sendo o conjunto formado pelas abelhas dessa espécie que ocorrem no território brasileiro.**

Seriam populações igualmente válidas se elas fossem limitadas geograficamente ao estado do Paraná, ou ao litoral do Paraná, ou ainda à pequena cidade de Antonina.

Nesse ponto vamos frisar novamente uma conclusão importantante: Os limites de uma população estatística são estabelecidos pelos pesquisadores e podem coincidir ou não com os limites naturais da população biológica. Dessa premissa fundamental derivam duas imposições igualmente fundamentais para o modelo de amostragem que estamos tratando: a amostra aleatória simples.

(1) Na população definida pelo pesquisador, todos os elementos que a integram devem ter a mesma chance de serem incluídos nas amostras que serão coletadas (equiprobabilidade).

(2) As conclusões resultantes dos testes estatísticos aplicados são válidas única e exclusivamente para a população estatística assim definida, de onde foi efetuada a coleta da amostra analisada.

A definição sobre a qualidade de uma amostra é um problema que se fundamenta muito mais na biologia do que na estatística. Somente o pesquisador, conhecendo todas as particularidades morfológicas, comportamentais, fisiológicas, ecológicas, etc. da(s) espécie(s) estudadas (s) pode avaliar se o método de inclusão das unidades amostrais assegurou a observância do princípio da equiprobabilidade.

Assim sendo, o conhecimento sobre o material estudado, juntamente com o bom-senso, são duas qualidades primordiais para a definição de um plano amostral válido e representativo em uma análise bioestatística.

RESUMINDO: Existem vários modelos amostrais que podem ser usados em um planejamento estatístico. O modelo mais simples e mais genérico e a Amostra Aleatória Simples que é o mais recomendado quando se tem poucas informações sobre a população estudada.

Quando conhecemos os segmentos, estratos ou sub-grupos que caracterizam uma população (e que sejam facilmente identificados e contabilizados) podemos reduzir consideravelmente os tamanhos das amostras usando métodos que levam em consideração os tamanhos das populações e dos subgrupos, e também as variâncias dentro deles e entre eles.

Se você não conhece as caracteristicas da população, use a amostra aleatória simples e trabalhe com o maior tamanho de amostra que seus recursos (tempo, dinheiro, mão-de-obra, etc.) permitirem.

Os Testes Paramétricos e não Paramétricos

Os testes estatísticos mais utilizados tantos em teses quanto em pesquisas biológicas, são aqueles classificados genericamente como ***testes paramétricos*** porque são baseados na distribuição normal e nos parâmetros que a definem: média aritmética e variância.

É o caso do teste 't' de Student, teste Z. análise da variância (ANOVA) e outras aplicações.

Esses testes pressupõem que a variável analisada siga a distribuição normal e, dependendo do teste, outras restrições quanto a homogeneidade das variâncias e os tamanhos das amostras também devem ser observados.

Nem sempre é fácil, na prática, verificar se esses requisitos teóricos são satisfeitos. Muitas vezes os testes são aplicados assumindo-se que as restrições são aproximadamente ou provavelmente satisfeitas.

Nos casos em que essas suposições não sejam razoavelmente satisfeitas, uma das alternativas pode ser a aplicação de testes não paramétricos (também chamados de teste de distribuição livre), para os quais as restrições são bem menores.

De um modo geral os testes não paramétricos tem uma eficiência menor que os equivalentes paramétricos, mas dependendo do tipo de pesquisa e da abrangência das conclusões essas limitações podem não ser importantes a ponto de inviabilizar o uso do teste. Em alguns casos a diferença de eficiência é tão pequena que se torna irrelevante a discussão sobre o tema.

Os testes não paramétricos são adequados às variáveis mensuradas em escalas 'nominal' e 'ordinal'. Na escala nominal trabalhamos com atributos e suas respectivas frequências observadas. Na escala ordinal as observações são hierarquizadas recebendo códigos em ordem de grandeza (também chamados de postos, por exemplo: primeiro, segundo, terceiro, ou posto 1, posto 2, posto 3 etc), sequencialmente, de acordo com o valor numérico da variável analisada.

Abordaremos aqui os seguintes testes:

a) **Teste de iteração**: para testar a aleatoriedade das amostras.

b) **Teste Qui-quadrado**: testa a aderência das frequências e a independência dos critérios de classificação.

c) **Teste de Kolmogorov-Smirnov**: análogo ao qui-quadrado porém menos restritivo.

d) **Teste de Wilcoxon**: análogo ao teste 't' para dados parelhados sem as exigências deste.

e) **Teste de Mann-Withney**: análogo ao teste 't' para dados independentes sem as restrições do teste paramétrico.

f) **Teste de Kruskal-Wallis**: análogo à análise de variância para 1 critério de classificação sem as limitações do teste F.

g) **Correlação de Spearman**: análogo ao coeficiente de correlação linear de Pearson.

Como escolher o teste mais adequado?

A escolha do teste mais adequado é o primeiro problema que um usuário da bioestística se defronta e muitas vezes tem dificuldades para resolver. Vamos ver alguns passos que podem ajudar nessa escolha, analisando a natureza da variável que será testada.

As variáveis podem inicialmente ser classificadas em dois grandes grupo:

1) Variáveis discretas
2) Variáveis contínuas.

VARIÁVEIS DISCRETAS

- Os dados são agrupados segundo critérios não-numéricos (qualitativos)
- A quantificação dentro do grupo é feita por números inteiros (contagens)

No critério qualitativo os grupos formados são identificados com palavras, como por exemplo os grupos sanguíneos (A,B,AB, O) ou as cores da concha de um hipotético molusco, que variam do amarelo claro ao marron.

No caso dos grupos sanguíneos você não consegue hierarquizar os grupos formados. Esse nivel de quantificação recebe o nome de "**Qualitativo nominal**".

No caso das cores da concha você poderia hierarquizar (ordenar) as cores criando, por exemplo, 5 grupos: amarelo claro, amarelo escuro, alaranjado, marrron claro e marron escuro e atribuir a esses grupos códigos (ou escores) em ordem crescente de 1 até 5. Esse nível de quantificação é chamado de "**Qualitativo Ordinal**".

VARIÁVEIS CONTÍNUAS

- Os dados podem ser agrupados segundo critérios numéricos ou não numéricos.
- A quantificação dentro do grupo é efetuada através de mensurações (pesagens, medidas, cronometragens, escalas termométricas etc).

Você pode formar grupos para variáveis continuas como, por exemplo,

Tipo da variável	Nivel de quantificação	Aplicações disponíveis
	Nominal (contagens)	Qui-quadrado
		Teste de Aleatoriedade
		Mann-Withney
DISCRETAS	Ordinal	Teste Kolmogorov-Smirnov
	(contagens&escores)	Teste Kruskal-Wallis
		Correlação Spearman
		Teste Wilcoxon
		Análise de Variância 1W
		Análise de Variância 2W
		Análise de Variância 2W + R
CONTÍNUAS	Mensurações	Correlação e Regressão
	(pesagens&medidas)	Dimensionamento amostral
		Teste T
		Teste T (parelhados)
		Teste Z

o agrupamento em classes em uma tabela para construir um histograma da altura de pessoas adultas (em centímetros). Nesse caso a formação dos grupos obedeceria um critério numérico: 150-160 cm, 160-170 cm, 170-180 cm etc.

Da mesma forma, você poderia agrupar os dados segundo critérios qualitativos. Por exemplo, se estiver trabalhando com amostras de localidades diferentes os dados serão agrupados de acordo com a origem do local A ou local B. O importante é que o dados sejam quantificados em uma escala contínua, como por exemplo, o comprimento total, etc.

As especificidades de cada teste estão fora do escopo deste manual e devem ser buscadas nos livros textos de Bioestatística.

No quadro abaixo relacionamos as aplicações disponibilizadas no programa "aplicativoBioestat.F", identificadas pelo respectivo código, e também as páginas deste tutorial com as orientações sobre o uso das mesmas.

LOCALIZAÇÃO DAS APLICAÇÕES E RESPECTIVOS CÓDIGOS		
APLICAÇÕES	Pag.	CÓDIGO
Aleatoriedade, Teste para uma amostra	36	8
ANOVA - 1 critério de classificação	22	4
ANOVA - 2 critérios de classificação	24	5
ANOVA - 2 critérios de classificação com repetições	28	6
Correlação & Regressão lineares	31	7
Dimensionamento de Amostras	8	1
Kolmogorov-Smirnov, Teste de	43	10
Kruskal-Wallis, Teste de	56	15
Mann-Withney, Teste de	52	13
Qui-Quadrado, Teste de aderência e independência	39	9
Regressão Curvelínea	58	16
Spearman, Correlação de	54	14
T, Teste de Student para duas amostras independentes	14	2
T, Teste de Student para dados parelhados	19	3
Wilcoxon, Teste de	49	12
Z, Teste para proporções	46	11
* Potencial Agregativo / Discriminativo *	67	17
Obs: Na aplicação 5 exemplificamos o uso do teste de Tukey (pag27)		

Aplicação:
Dimensionamento de amostras

CÓDIGO = 1

Relembrando o básico

Uma das principais dúvidas de um pesquisador ou pós-graduando quando está planejando sua pesquisa ou tese é: qual o tamanho da amostra que deve ser utilizada para analisar os dados estatisticamente? A resposta é bem simples: depende da variabilidade da população em estudo e da precisão que o pesquisador quer em suas estimativas. Felizmente existem métodos baseados na distribuição binomial e na distribuição normal que permitem responder satisfatoriamente esse tipo de dúvida.

Para iniciar nossa análise precisamos inicialmente definir qual dos dois tipos de variáveis que nós vamos trabalhar: variáveis discretas ou contínuas. As discretas são expressas em contagens enquanto que as contínuas referem-se às mensurações (medições lineares, áreas, volumes, pesagens, cronometragens, etc). As variáveis discretas são mais fáceis de serem manuseadas porque as contagens podem ser transformadas em porcentagens (quando o objetivo e testar proporções) e estas tem uma variação pré-determinada entre 0 (zero) e 100%. Através da distribuição binomial nós podemos obter o valor da variância (que alcança um valor máximo quando p e q são iguais a 0.5 ou ½) . Quando tratamos das variáveis contínuas a coisa fica um pouco mais difícil e é praticamente impossível fazer previsões sobre o tamanho de uma amostra sem termos um conhecimento prévio de algumas estimativas preliminares sobre a população alvo. Sabemos que é a heterogeneidade da população que vai nos dizer qual o tamanho que a nossa amostra deve ter. Como regra geral, quanto mais heterogênea for uma população estatística tanto maior será o tamanho da amostra para estudá-la. Como a heterogeneidade é medida através da variância, vamos ver que os dimensionamentos amostrais tomam por base o valor desta grandeza estatística.

TELA 01

```
******************************************************************
**********  DIMENSIONAMENTO DE AMOSTRAS ***********
 ******     Tipo de variavel: (1) = discreta /// (2) = contínua     ******
******************************************************************
```

A variável discreta é representada pelas contagens de determinadas características da amostra. Se as observações resultarem em mensurações (pesagens, medidas, etc) estaremos trabalhando com variáveis continuas. Nesse exemplo digitaremos 1

TELA 02

```
*** Tem um estimativa da proporcao 'p'? /// nao = 0  /sim = 1 ***
```

Informações de trabalhos anteriores ou de amostras-piloto podem nos fornecer estimativas de uma determinada proporção p. Digite 1

TELA 03

```
*** Informe o valor do p estimado ***
0.4
```

Neste exemplo, simulamos o valor de p = 0.4 (equivalente a 40%)

TELA 04

```
*** Qual o erro máximo aceito no dimensionamento? ***
.05
```

Vamos informar o erro máximo que pretendemos em nossa estimativa da proporção ‘p’. Neste exemplo usaremos 0.05 , (ou 1/20 ou 5%). Geralmente são feitas várias simulações de erros, cotejando-os com os respectivos tamanhos de amostras calculados. Escolhemos então uma margem de erro que permita a máxima precisão com um custo amostral adequado.

TELA 05

```
*************************************************************
*  Para um erro máximo = 0.05 /// Tamanho da amostra =   384.       *
*************************************************************
```

A TELA 05 nos informa que quando p = 0.4 se quisermos uma amostra com um erro aceitável de 0.05 deveremos trabalhar com 384 unidades amostrais.

Reiniciando o programa vejamos a situação na qual não temos uma estimativa confiável de 'p'. Nesse caso digitamos zero na TELA 02 e o programa assume os valores de p e q iguais a 0.5. Esta é uma atitude conservadora pois, quando p e q são iguais, nós temos os maiores dimensionamentos de amostras. Se aceitarmos a mesma margem de erro = 0.05 teremos o seguinte resultado:

TELA 05a

```
*************************************************************
*  Para um erro máximo = 0.05 /// Tamanho da amostra =   400        *
*************************************************************
```

Vamos analisar agora as variáveis contínuas (mensurações). Retornamos inicialmente à TELA 01 e digitamos '2'. O programa saltará para a TELA 06.

TELA 06

```
*** Estimativa da variância e média? // (0) não /// (1) sim  ***
```

Neste exemplo digitamos '1' simulando que temos estimativas confiáveis da variância e da média populacionais.

TELA 07

```
*** Digitar valor da variância estimada: ***
2.5
```

Nesta simulação digitamos o valor da variância igual a 2.5. Na tela seguinte o programa solicita o valor da média estimada.

TELA 08

*** Digitar o valor da média estimada: ***
10

Para efeitos demonstrativos, simulamos uma média igual a 10.

TELA 09

*** (1) Erro na unidade da média? // (2) Erro porcentual? ****

O programa nos questiona se optamos por trabalhar com erro absoluto (por exemplo, se a média estiver expressa em centímetros, o erro deverá ser informado também em centímetros) ou com erro relativo (um porcentual arbitrário de erro em relação ao valor da média). Nesta similação digitaremos 1 (erro na mesma unidade da média)

TELA 10

** Informe o erro máximo aceitável **
0.9

Arbitramos um erro aceitável igual a 0.9 unidades.

TELA 11

****
* Erro máximo aceitável = 0.9 // Tamanho da amostra = 12.*
****

A TELA 11 mostra que para trabalharmos com um erro aceitável de 0.9 cm, nossa amostra deverá ter 12 elementos.

Vamos analisar a situação mais comum quando não temos uma estimativa preliminar da média e da variância. Nesse caso precisamos coletar uma amostra-piloto (é recomendável, pelo menos, 20 unidades amostrais) para obtermos estimativas consistentes. Retornamos ao início do programa e digitamos (2) 'variáveis contínuas'; depois digitamos (0) 'não temos estimativas da variância e da média' e iremos para a TELA 12.

TELA 12

**** DIGITAR DADOS DA AMOSTRA-PILOTO ****

**** Quantos dados serão digitados? (N >=20) ****

Neste exemplo, por uma questão de espaço usaremos uma amostra-piloto com apenas 5 unidades.

TELA 13

Digitar dado número 1
5
Digitar dado número 2
4
Digitar dado número 3
1
Digitar dado número 4
2
Digitar dado número 5
3

TELA 14

*** (1) Erro na unidade da média? // (2) Erro porcentual? ****

Como já explicado na TELA 09, devemos informar se vamos trabalho com erro absoluto ou erro relativo(%)

Neste exemplo digitamos ‘2’ informando a opção de trabalhar com o erro relativo.

TELA 15

*** Qual o erro porcentual aceitável? ***

Vamos considerar que um erro de 10% em torno do valor da média seja aceitável. Digitaremos então 0.1 (= a 10% em valores decimais)

TELA 16

```
****************************************************
*    Erro máximo aceitável =   0.10 // Tamanho da amostra =    111.   *
****************************************************
```

A TELA 16 nos informa que, baseados amostra-piloto ora analisada para estimativas com um grau de precisão de 10% em torno da média, são necessárias 111 unidades amostrais.

Quando trabalhamos com amostra-piloto (ou preliminar) duas situações podem ocorrer: (a) a amostra-piloto tem um número de unidades igual ou maior que o tamanho dimensionado e (b) a amostra-piloto é menor que o valor calculado. No primeiro caso a amostra-piloto passa a ser a amostra definitiva. No segundo caso, selecionamos uma amostra complementar, seguindo rigorosamente a mesma metodologia da amostragem preliminar, de modo a atingirmos o valor de n calculado.

É importante ressaltar que o intervalo de tempo entre as coletas da amostra-piloto inicial e da amostra complementar deve ser de tal ordem que não ocorram alterações significativas em relação à variável que está sendo analisada na pesquisa.

Aplicação: Teste 't' Student

CODIGO = 2

O teste 't' de Student é talvez o teste estatístico mais conhecido no meio acadêmico, embora, algumas vezes, seja usado sem que o pesquisador observe algumas limitações e preceitos teóricos que balizam a sua aplicação. Os três requisitos fundamentais são: normalidade, independência e homogeneidade das variâncias das duas amostras analisadas. O quesito normalidade remete à distribuição normal. A quase totalidade das variáveis contínuas (mensurações, pesagens, etc) que retratam fenômenos biológicos acompanham a curva (distribuição) normal. Essa restrição, na maioria das vezes, é assumida sem a realização de teste de aderência à curva normal. A independência deve ser analisada à luz dos conhecimentos do pesquisador sobre o comportamento da variável analisada. Deve ser questionado se a seleção de uma unidade amostral influiu ou foi influenciada pela seleção de outra unidade, ou se as duas tem seus comportamentos influenciados por algum um fator comum. Nada havendo que indique esses problemas, assumimos que as duas amostras são independentes. A verificação da homogeneidade é feita através do Teste 'F', comparando a razão entre a variância maior e a menor (S^2 / s^2) das amostras. Os graus de liberdade correspondem à soma dos graus de liberdade das duas amostras (n_a -1 + n_b -1). A significância do teste F vai influir na definição dos graus de liberdade do teste 't' na comparação entre as médias. Para facilitar a interpretação todos os testes serão aplicados ao nível de 95% ($P <= 0.05$), tanto para o teste 't' quanto para o teste 'F'.

As próximas telas do tutorial serão referenciadas aos quadros 1A e 1B de tal modo que seja possível acompanhar a aplicação dos testes e comparar os resultados obtidos com os mostrados nos quadros. As amostras devem ser analisadas levando-se em conta a homogeneidade das variâncias e os respectivos tamanhos. Esses dados orientam os procedimentos a serem seguidos na aplicação do teste.

O Quadro 1A mostra duascomparações com variâncias homogêneas. No exemplo 1 as amostras tem o mesmo tamanho. No exemplo 2 , tamanhos diferentes. O programa aplica o teste F (dividindo maior variância pela menor) e mostra os valores calculados. Em 1 Fc = 2.33 e Ft = 3.18. Em 2 . Fc = 1.11 e Ft = 3.29. Nos dois exemplos $F_{calculado}$ (ou F_c) é menor que o $F_{tabelar}$, (ou F_t) então as variâncias testadas são consideradas homogêneas.

Quadro 1A

Exemplo de amostras com variâncias homogêneas

1 Exemplo de 2 amostras

$n_a = n_b$

A	B
1.0	1.0
2.0	5.0
3.0	7.0
4.0	8.0
5.0	4.0
6.0	6.0
5.0	5.0
4.0	2.0
3.0	5.0
2.0	1.0
M = 3.5	M = 4.40
S^2 = 2.50	S^2 = 5.82

Fc = 2.33 e Ft = 3.18

$t_{calculado}$ = 0.99

$t_{tabelar}$ = 2.10

GL = 18

2 Exemplo de 2 amostras

n_a /= n_b

A	B
1.0	1.0
2.0	2.0
3.0	3.0
4.0	4.0
5.0	5.0
6.0	6.0
3.0	5.0
2.0	4.0
5.0	
4.0	
M = 3.50	M = 3.75
S^2 = 2.50	S^2 = 2.79

Fc = 1.11 e Ft = 3.29

$t_{calculado}$ = 0.33

$t_{tabelar}$ = 2.12

GL = 16

No quadro 1B (página seguinte) mostramos duas comparações de amostras com variâncias heterogêneas 3 e 4 , F_c = 5.87 e F_c = 5.27 ultrapassam os limites tabelares, indicando a heterogeneidade das variâncias.

Quadro 1B

Exemplo de amostras com variâncias heterogêneas

3 Exemplo de 2 amostras
na = nb

A	B
1.0	2.0
2.0	9.0
3.0	3.0
4.0	10.0
5.0	1.0
6.0	8.0
5.0	2.0
4.0	7.0
3.0	1.0
2.0	10.0
M = 3.5	M = 5.3
S^2 = 2.50	S^2 = 14.7

Fc = 5.87 e Ft = 3.18
tc = - 1.37
t tabelar = 2.26
GL = 9

4 Exemplo de 2 amostras
na /= nb

A	B
1.0	2.0
2.0	9.0
3.0	3.0
4.0	10.0
5.0	1.0
6.0	8.0
3.0	2.0
2.0	7.0
	1.0
	10.0
M = 3.25	M = 5.3
S^2 = 2.79	S^2 = 14.7

Fc = 5.27 e Ft = 3.68
tc = 1.53
tw = 2.28
GL = ***

Na TELA 01 o usuário deve informar o tamanho de cada uma das amostras que serão analisadas pelo teste ‘t’. No exemplo que estamos utilizando 3 vamos digitar 10 tanto para a amostra A quanto para a amostra B.

TELA 01

```
*************************************
***              TESTE T          ***
*** Qual o tamanho da amostra 'A'?  na =   ***
10
*** Qual o tamanho da amostra 'B'?  nb =   ***
10
```

A TELA 2 nos mostra um trecho da digitação dos dados do exemplo 3 referentes as amostras heterogêneas com tamanhos iguais ($n_a = n_b = 10$)

TELA 2

```
 *  Digitar o dado numero   8 da amostra A  *
4
 *  Digitar o dado numero   9 da amostra A  *
3
 *  Digitar o dado numero  10 da amostra A  *
2
 ***********************************************
 *  Digitar o dado numero   1 da amostra B  *
2
 *  Digitar o dado numero   2 da amostra B  *
9
 *  Digitar o dado numero   3 da amostra B  *
3
```

Na TELA 3 vemos o resultado do teste **F** (Fc = 5.87). Consultando uma tabela do teste você encontrará Ft = 3.68 confirmando, portanto a heterogeneidade das duas variâncias (Fc > Ft).

TELA 3

```
******************************************************
*        TESTANDO A HOMOGENEIDADE DAS VARIANCIAS     *
*                                                    *
******************************************************
                  Fc =   5.87
* *     Graus de liberdade do numerador    =   9.   * *
**      Graus de liberdade do denominador =   9.     **
*****************************************************
*      Se as variancias sao homogeneas:   digitar 0   *
*      Se as variancias sao heterogeneas:  digitar 1  *
******************************************************
```

Como as variâncias são heterogêneas digitamos 1 e o teste se completa. Observe que os graus de liberdade são iguais a nove. Se fossem amostras homogêneas $GL = n_a + n_b - 2$ mas, como são heterogêneas o teste se torna mais conservador e reduz os graus de liberdade para GL = n - 1.

TELA 4

```
**************************************************************
*      t_calculado = -1.37      ////    Graus de liberdade GL = 9.    *
*                                                                    *
**************************************************************
```

Reinicie o programa e digite agora os dados do exemplo 4 referentes ao caso das amostras de tamanhos diferentes. Nesse exemplo a amostra A tem $n_a = 8$ e na amostra B, $n_b = 10$. Após digitar os dados das duas amostras você vai chegar novamente na TELA 3a (Fc = 5.27) e vai digitar 1 para prosseguir.

TELA 3a

```
**************************************************************
*        TESTANDO A HOMOGENEIDADE DAS VARIÂNCIAS
*
**************************************************************
*                         Fc =   5.27                          *
*             Graus de liberdade do numerador   =   9.         *
*             Graus de liberdade do denominador =   7.         *
**************************************************************
*           Se as variâncias são homogêneas:   digitar 0       *
*           Se as variâncias são heterogêneas:  digitar 1      *
**************************************************************
```

TELA 5

```
  *     Qual o valor de 't' para   7. graus de liberdade        *
  2.36
   *     Qual o valor de 't' para   9. graus de liberdade       *
  2.26
```

Na TELA 5 o programa solicitou informar os valores de 't' tabelar para cada uma das amostras. Para a amostra A, com GL = 7, o 't' informado é = 2.36. Para a amostra B, com GL = 9, o 't' informado é = 2.26. O programa calcula então um novo valor de 't' tabelar, o 't_w', que é igual aos 't' das amostras ponderados pelas razões entre as variâncias de cada amostra e os respectivos tamanhos. (TELA 6)

TELA 6

```
*************************************************
*          't' calculado= = -1.52   ////   't'(w) tabelar  =  2.28          *
*                                                                           *
*************************************************
```

Na comparação das amostras do conjunto 3 o valor do $t_{calculado}$ (1,37) é menor que o $t_{tabelar}$ (2.26), portanto concluimos que a diferença entre as duas médias não é significativa e aceitamos a Hipótese H_0

Na comparação das amostras do conjunto 4 o valor do $t_{calculado}$ (1.52) também é menor que o valor do $t_{tabelar}$ (tw = 2.28) indicando uma diferença não significativa e a aceitação de H_0.

Em relação aos conjuntos 1 e 2 as amostras comparadas possuem variâncias homogêneas e a aplicação do Teste 't' segue o mesmo roteiro das comparações dos conjuntos 3 e 4 já detalhados.

Aplicação:
Teste ‘t’ dados parelhados
CÓDIGO = 3

QUADRO 3

Indivíduo	Peso inicial	Peso final
A	7.2	8.2
B	8.3	9.7
C	14.5	16.0
D	10.8	12.1
E	9.3	10.7
F	8.9	8.6
G	15.8	17.9
H	8.7	9.4
I	9.3	11.0
J	11.6	13.8

Até aqui nós analisamos as aplicações do teste ‘t’ em amostras independentes. Esse teste também pode ser utilizado com amostras não independentes quando trabalhamos com dados parelhados. Essa alternativa costuma ser adotada em experimentos nos quais é possível organizar o teste formando pares tão homogêneos quanto possível e a variável considerada é a diferença, em cada par, das respostas a deterninados estímulos provocados pelo experimentador.

Essa modalidade também pode ser utilizada nos experimentos do tipo “antes e depois”, quando se avaliam as respostas de um mesmo indivíduo antes e depois de ele ser submetido a determinado tratamento / estímulo. No quadro. 3 simulamos um experimento onde medimos o ganho de peso de uma espécie qualquer que foi submetida a uma nova dieta alimentar. A variável de interesse é a diferença de peso de cada indivíduo antes e após a dieta alimentar.

Para aplicarmos essa modalidade do teste ‘t’ vamos digitar os dados de cada par separadamente. A digitação pode ser feita em uma mesma linha deixando um espaço em branco entre o “antes” e o “depois” ou, também pode digitar cada dado e apertar a tecla <enter> após a conclusão de cada informação. X_a e X_d referem-se ao “antes” e o “depois”. No exemplo a seguir usaremos os dados constantes do QUADRO 3, onde n = 10.

TELA 01

```
*****************************************************
***************Teste T para dados parelhados***************
****** Digitar Xa e Xd separados por um espaço em branco ******
*****************************************************
*****          Quandos pares serão digitados?          *****
10
```

TELA 02

```
Digitar o par (Xa e Xd) número   1
7.2  8.2
Digitar o par (Xa e Xd) número   2
8.3  9.7
Digitar o par (Xa e Xd) número   3
14.5  16
```

Após a digitação dos 10 pares de dados o programa determina as diferenças entre Xa e Xd, calcula a média e variância e aplica o teste't".

TELA 03

```
*****************************************************
**********   't' =  -5.70 /// Graus de liberdade =   9   *********
*****************************************************
```

O teste 't', analisou as diferenças de pesos nas fases 'antes' e 'depois' obtendo um resultado t_c = 5.7. Quando comparado com 't' tabelar para 9 graus de liberdade ($t_{tabelar}$ = 2.26) aceitamos a hipótese H_1, ou seja, a diferença média entre os pares analisados, antes e depois da dieta alimentar é significativa ao nível de 95%.

Como pode ser observado o teste 't' é uma das ferramentas mais úteis da bioestatística. Respeitadas as restrições e pré-requisitos que balizam a sua aplicação, ele pode ser de grande utilidade em inúmeras análises estatísticas, validando conclusões e ajudando nas decisões que envolvem questões biológicas.

Análise da Variância

Quando temos que analisar duas amostras vimos que o teste ‘t’ resolve satisfatoriamente o problema. No entanto, quando temos 3 ou mais amostras precisamos de um método mais adequado, como o desenvolvido por R.A. Fisher em 1920, denominado Análise de Variância.

A análise de variância (popularmente conhecida como ANOVA) compara a variância ‘entre as amostras’ (variância explicada) com a variância ‘dentro das amostras’, também chamada variância residual ou variância não-explicada ou variância do acaso, etc). A ‘variância dentro’ reflete a variabilidade natural dos dados, para a qual não conseguimos identificar uma causa que a origine e por isso consideramos essa variação como aleatória.

A razão F = (variância explicada) / (variância residual) nos orienta a aceitarmos uma das duas hipóteses: Hipótese H_o (as médias das amostras testadas se originam de uma mesma populaçao com média ‘Mi’ ou alternativamente, a hipótese H_1 (as médias amostrais testadas se originam de populações com médias ‘Mi’ diferentes).

Fica claro que, complementarmente ao teste F, deveremos aplicar um outro teste para a avaliar as diferenças mínimas significativas (dms) entre as médias das amostras, identificando em quais pares ocorrem as diferenças detectadas pelo teste ‘F’ (por, exemplo Teste de Tukey, de Duncan, de Schefee, etc.)

Aplicação:
Anova
1 critério de classificação.
CÓDIGO = 4

QUADRO 04

TRATAMENTOS

C1	C2	C3
4	5	4
6	9	5
11	9	6
10	12	14

O primeiro modelo que vamos analisar tem apenas um critério de classificação (one way) no qual nós testamos três amostras, cada uma delas submetida a um determinado tipo de tratamento. Neste exemplo (QUADRO 04), cada amostra de 4 indivíduos foi submetida a um tipo de alimentação. Queremos avaliar se os tratamentos produziram resultados estatisticamente iguais ou diferentes em relação à variavel peso de 12 indivíduos homogêneos utilizados no teste.

Deve ser resssaltado que os tratamentos devem ter o mesmo número de repetições (neste exemplo, 4) pois os testes complementares (Tuykey e Duncan) fornecem resultados aproximados caso as amostras não tenham o mesmo tamanho.

O teste mais simples e mais usado é o de Tukey, a partir do cálculo da dms (ou diferença mínima significativa), que leva em conta o valor tabelar do teste, o desvio padrão residual e o número de repetições das amostras. Aquelas médias cujas diferenças entre elas forem menores que a dms são consideradas estatisticamente iguais. O detalhamento da aplicação do Teste de Tukey está na **Aplicação 5**, página 27.

Para facilitar a entrada dos dados desta **Aplicação** , o programa orienta a digitação seguindo as linhas da tabela, já que estamos trabalhando com um número igual de repetições nos tratamentos. Inicialmente vamos informar (ver QUADRO 04) que trabalharemos com 3 tratamentos e 4 repetições em cada um deles.

TELA 01

**** Quantos tratamentos serão analisados? *****

TELA 02

*** Numero de repetições em cada tratamento ****

Após a digitação da última linha da tabela o programa calcula a variância total dos dados e também a variância que pode ser atribuída aos tratamentos. A diferença entre as duas tem como resultado a variância residual (variância ‘dentro’ das amostras) para a aplicação do Teste F.

TELA 03

*** Digitar repetição 1 do tratamento 1

4

*** Digitar repetição 1 do tratamento 2

6

TELA 04

```
*******************************************************
******************** RESULTADO ********************
*                                                     *
*                    Fc =  0.17                       *
*        GLCOL =  2           GLRES =   9             *
*        Desvio Padrao Residual =  3.66               *
*******************************************************

              Mediacol ( 1) =   7.75
              Mediacol ( 2) =   8.75
              Mediacol ( 3) =   7.25
*******************************************************
```

A TELA 04 nos mostra que o F calculado (Fc = 0.17) não supera o valor fornecido pela tabela do teste (Ft = 4.26) ao nível de 5%.

Com este resultado aceitamos a hipótese Ho e admitimos que a três médias mostradas na parte inferior da tela (médias das colunas, referentes aos tratamentos)) diferem por um valor menor que a dms. São, portanto, consideradas médias estatisticamente iguais, indicando que os três tratamentos se mostraram igualmente eficazes. Neste caso não há necessidade de aplicação do teste complementar para as médias

Aplicação:
Anova
2 critérios de classificação.

CÓDIGO = 5

QUADRO 05

	TRATAMENTOS		
	C1	C2	C3
L1	4	5	4
L2	6	9	5
L3	11	9	6
L4	10	12	14

Na aplicação anterior, a variavel ‘peso dos animais” foi analisada com o objetivo de avaliarmos a influência de diferentes rações no comportamento do peso dos indivíduos utilizados no experimento. Agora, vamos estudar um modelo que analisa a influência simultânea de 2 fatores, também chamados 'causas de variação'. Por exemplo, na 'causa de variação' (efeitos das rações), cada ração representa um tratamento, e eles são nominados pelas colunas C1, C2 e C3.

No QUADRO 05 vemos que existe uma outra 'causa de variação', disposta nas linhas da tabela.

Nesse exemplo essa nova causa de variação seria o 'tempo de exposição à iluminação artificial em ambiente controlado'. Os quatro níveis de iluminação foram nominados como L1, L2, L3 e L4.

A análise de variância com 2 critérios de classificação tem como objetivo analisar as influências das 'causas de variação' (neste exemplo, tipo de ração e luminosidade), investigando separadamente os efeitos de cada uma das causas sobre o peso das cobaias.

TELAS 01 e 02

```
*****************************************************
*       ANOVA 2 CRITÉRIOS DE CLASSIFICAÇÃO          *
*          Limites: 20 linhas // 20 colunas         *
*****************************************************

     *** Número de linhas da tabela de entrada ***
      4

     *** Número de colunas da tabela de entrada ***
      3
```

As telas 01 e 02 mostram as duas primeiras entradas do programa onde devemos informar o número de linhas da tabela (neste exemplo, períodos de luminosidade L1, L2, L3 e L4), quando digitamos '4'.

Em seguida o programa solicita informar o número de colunas (neste exemplo, tipos de rações C1, C2 e C3) então digitamos '3'.

TELA 03

```
VALOR DA LINHA = 1     COLUNA = 1
4

VALOR DA LINHA = 1     COLUNA = 2
```

A TELA 03 mostra como é efetuada a digitação dos dados, linha por linha. Para evitar erros na entrada das informações, o programa identifica previamente a linha e a coluna onde está o dado a ser digitado.

TELA 04

```
****************************************************
*               RESULTADOS DA ANÁLISE              *
****************************************************

F(colunas)    =   0.55          GL(colunas)  =  2
F(linhas)     =   7.49          GL(linhas)   =  3
DPadrao Res. =   2.05           GL(resíduo)  =  6

****************************************************

        Mediacol ( 1) =   7.75
        Mediacol ( 2) =   8.75
        Mediacol ( 3) =   7.25
****************************************************
        Medialin ( 1) =   4.33
        Medialin ( 2) =   6.67
        Medialin ( 3) =   8.67
        Medialin ( 4) =  12.00
****************************************************
```

A TELA 04 mostra os resultados da análise de variância para dois critérios de classificação. Nesse modelo são realizados dois testes: o primeiro testa a significância da ação das rações listadas nas colunas (C1, C2 e C3) e foi calculado um valor de Fc = 0.55. Esse valor, quando comparado com o $F_{tabelar}$ (para 2 graus de liberdade para os tratamentos e 6 para o resíduo, Ft = 5.14) nos leva a aceitar a hipótese Ho, concluindo que as 3 rações não mostraram eficiências significativamente diferentes.

Uma segunda comparação é feita em relação aos 4 tratamentos da luminosidade, dispostos nas linhas da tabela. O valor de Fc = 7.49

quando comparado com o $F_{tabelar}$ = 4.76 (para 3 graus de liberdade para os tratamentos e 6 para o resíduo) sugere que devemos aceitar a hipótese H_1. Assim, concluimos que existem diferenças significativas entre alguns (ou todos) dos quatro tratamentos relativamente ao tempo de exposição à luminosidade.

Em outras palavras, existem diferenças significativas entre as médias dos pesos das cobaias submetidas aos diferentes períodos de luminosidade.

Exemplo de aplicação do Teste de Tukey

O teste de Tukey complementa o Teste F ao identificar quais os tratamentos que são iguais ou diferentes entre si. As medias dos tratamentos listadas na TELA 04 serão testadas em relação à diferença mínima significativa (dms) que é calculada assim:

dms = q . S / raiz(r)

q = valor tabelar do teste de Tukey. Usar o número de tratamentos e os graus de liberdade do resíduo ou GL(resíduo).

S = desvio padrão residual (DPadrão Res.)

r = raiz quadrada do número de parcelas que deram origem à média testada.

dms = 4.90 x 2.05/raiz(3) = 5.80

Vamos exemplificar com as médias das linhas do QUADRO 05 e que estão listadas na TELA 04 como Medialin. As médias devem ser comparadas 'duas-a-duas' . Assim, seguindo o exemplo acima referido , temos 4 tratamentos e, consequentemente, 4 médias: Medialin(1) = 4.33; Medialin(2) = 6.67; Medialin(3) = 8.67 e Medialin(4) = 12.00. Testando as médias 'duas a duas' teremos K comparações, onde $K = n_L (n_L -1) /2$ e 'n_L' é igual ao número de tratamentos nas linhas do Quadro 05. Nesse exemplo com n_L = 4 tratamentos teremos K = 4(4-1)/2 = 6.

Comparações entre as médias

Medialin(1) - Medialin(2) = 4.33 - 6.67 = 2.34
Medialin(1) - Medialin(3) = 4.33 - 8.67 = 4.34
Medialin(1) - Medialin(4) = 4.33 - 12.00 = 7.67*
Medialin(2) - Medialin(3) = 6.67 - 8.67 = 2.00
Medialin(2) - Medialin(4) = 6.67 - 12.00 = 5.33
Medialin(3) - Medialin(4) = 8.67 - 12.00 = 3.33

CONCLUINDO: O teste F ao analisar a causa de variação 'exposição à luminosidade', indicou que existiam diferenças significativas entre os 4 tratamentos que foram testados. Complementarmente, o teste de Tukey, comparando as médias 'duas-a-duas' identificou uma diferença significativa entre as médias dos tratamentos 1 e 4. pois essa diferença (7.67) é maior que a dms calculada (5.80).

Aplicação:
Anova
2 critérios de classificação.
(com repetições)

CÓDIGO = 6

No modelo anterior, em cada casa (célula), ou seja, no cruzamento de uma coluna com uma linha, da tabela de digitação, normalmente a informação utilizada se refere a um único indivíduo ou à soma dos resultados de vários indivíduos que integram a célula, como foi exemplificado no QUADRO 05.

No QUADRO 06 as informações estão individualizadas para as duas repetições dentro de cada célula.

A análise fica muito mais rica e consistente quando nós tratarmos individualmente as informações referentes aos elementos que estão dentro das células,

A vantagem desse novo modelo é que a variação residual (variação aleatória ou variação não-explicada) fica mais depurada tornando o teste mais sensível. Por exemplo, vamos analisar comparativamente o QUADRO 05 e o QUADRO 06. Quando analisamos o valor '4' da célula L1-C1 do QUADRO 05 nós temos um conteúdo informativo menor do que quando são analisados os números '1' e '3" do QUADRO 06. No QUADRO 05 a variabilidade entre os números '1' e '3' seria incorporada à variância residual.

No QUADRO 06, além do efeito isolado de cada 'causa de variação', nós podemos medir também a ação conjunta das duas causas, ou seja, avaliamos se a interação dos tratamentos afetam o comportamento das variáveis analisadas.

QUADRO 06

	C1	C2	C3
L1	1 ; 3	2 ; 3	3 ; 1
L2	2 ; 4	4 ; 5	2 ; 3
L3	5 ; 6	3 ; 6	2 ; 4
L4	6 ; 4	7 ; 5	6 ; 8

TELA 01

```
******************************************************************
*NUMERO DE TRATAMENTOS NAS LINHAS E COLUNAS, REPETIÇÕES  *
******************************************************************
```

A TELA 01 inicia o programa informando o que deverá ser digitado. Nesse exemplo em que estamos usando os dados do QUADRO 06, vamos informar que trabalharemos com 4 tratamentos nas linhas (L1, L2, L3 e L4), 3 tratamentos nas colunas (C1, C2 e C3) duas repetições dentro de cada célula, como solicitado nas duas telas seguintes.

TELA 02 - 03

```
 NÚMERO DE TRATAMENTOS NAS LINHAS
4
 NÚMERO DE TRATAMENTOS NAS COLUNAS
3
 NÚMERO DE REPETICOES DENTRO DAS CÉLULAS
2
```

A entrada dos dados poderá ser conferida durante todo o processo já que o programa indicará para cada operação qual a linha, coluna e repetição a ser digitada.

TELA 04

```
*****************************************************
 *       DIGITAR OS VALORES A SEREM  ANALISADOS      *
 *****************************************************

 Linha = 1   Coluna = 1   Repetição = 1
1

 Linha = 1   Coluna = 1   Repetição = 2
3

 Linha = 1   Coluna = 2   Repetição = 1
```

Na TELA 05 temos os resultados do teste aplicado. O $F_{colunas}$ = 0.68 nos informa que não existe diferença entre as médias das colunas. O F_{linhas} = 9.26 detecta diferenças significativas entre as médias das linhas. O valor de $F_{linxcol}$ = 1.24 mostra que não há evidências de uma ação conjunta dos tratamentos das linhas e das colunas.

TELA 05

```
****************************************************
*              RESULTADOS DA  ANÁLISE              *
****************************************************

F(colunas)      =  0.68    GL(colunas)    =  2
F(linhas)       =  9.26    GL(linhas)     =  3
F(lin x col)    =   1.24   GL(lin x col)  =  6
DPadrao Res. =    1.31     GL(residuo)    = 12

****************************************************
          Mediacol ( 1) =   3.88
          Mediacol ( 2) =   4.38
          Mediacol ( 3) =   3.63
****************************************************
          Medialin ( 1) =   2.17
          Medialin ( 2) =   3.33
          Medialin ( 3) =   4.33
          Medialin ( 4) =   6.00
****************************************************
```

Complementarmente podemos aplicar o teste de Tukey para as médias das linhas para descobrir quais diferem significativamente.

Os tratamentos das colunas e a interação entre linhas e colunas não apresentaram diferenças significativas.

Aplicação:
Correlação & Regressão
(Linear & linearizadas)
CÓDIGO = 7

A aplicação 07 vai tratar dos modelos de correlação e regressão lineares. Em outras palavras, trataremos dos casos em que os pontos referentes às duas variáveis 'x' e 'y' quando plotados em um gráfico, se distribuem de uma forma aproximadamente linear, como se estivessem dispostos em torno de uma linha reta. Como algumas curvas podem ser linearizadas pelo uso de logarítmos, elas também serão tratadas com a mesma metodologia usada para os modelos lineares aditivos.

Nesta aplicação vamos abordar os modelos relacionados abaixo:

Linha reta	$y = a + b.x$	Modelo linear	$y = a + b.x$
Curva Hipérbole	$y = a + b/x$	Modelo linear	$y = a + b.(1/x)$
Curva logaritmica	$y = a + x^b$	Modelo linear	$y = a + b.Log(x)$
Curva potencial	$y = a.x^b$	Modelo linear	$Log(y) = Log(a) + b.Log(x)$
Curva exponencial	$y = a. e^{bx}$	Modelo linear	$Log(y) = Log(a) + b.x$

Obs.: Log = logaritmo neperiano (base ***e***)

Correlação

A correlação é medida através do 'coeficiente de correlação', 'r' cujo valor oscila entre '**- 1**' e '**+1**' nos mostrando o quanto o comportamento de duas variáveis se assemelham.

Deve ser ressaltado que a correlação não é, necessariamente, uma relação de 'causa e efeito', pois essa semelhança comportamental pode, por exemplo, dever-se a um ou mais fatores que influenciam o comportamento das duas variáveis, muitas vezes de forma independente.

Somente o pesquisador, conhecendo as particularidades biológicas das espécies e dos caracteres analisados pode inferir sobre uma possível relação de causa e efeito, que aja direta ou indiretamente.

Exemplificando: Se medirmos a correlação entre a taxa de fecundidade e o tamanho das fêmeas de uma determinada espécie, uma alto valor de ‘r’ pode não ser uma consequência do tamanho das fêmeas. Pode ser que os animais maiores tenham vantagens competitivas na disputa por alimentos e por isso se tornem aptos a produzir mais ovos.Caso não houvese competição por alimentos é possível que todas as fêmeas tivessem taxas de fecundidade semelhantes.

Regressão

A regressão é, por assim dizer, uma decorrência da correlação. Se eu tiver duas variáveis, que sejam correlacionadas, conhecendo o comportamento de uma delas eu posso fazer previsões sobre o comportamento da outra. As previsões serão tanto mais precisas quanto maior for o coeficiente de correlação entre elas.

A variável de mais fácil obtenção é genericamente denominada ‘variável independente’ e é tradicionalmente representada pela letra ‘x’.

Na regressão, a variável ‘y’ é de mais difícil obtenção e depende do valor assumido pela variável ‘x’. Por isso, ‘y’ é chamada de variável dependente: $Y = a + bX$

A título de ilustração, o quadro abaixo mostra a amplitude de variação da variável dependente **Y** para os modelos adotados, usando como a variável independente **X**, os 10 primeiros números inteiros.

Em todas as simulações usamos os mesmos valores de a = 2 e b = 0.5.

X	Valores de **Y** para os modelos linear e linearizados				
	Parâmetros utilizados **a = 2** e **b=0.5**				
	Reta	Logarítmica	Potencial	Exponencial	Hiperbólica
1	2.50	2.00	2.00	3.30	2.50
2	3.00	2.35	2.83	5.44	2.25
3	3.50	2.55	3.46	8.96	2.17
4	4.00	2.69	4.00	14.78	2.13
5	4.50	2.80	4.47	24.36	2.10
6	5.00	2.90	4.90	40.17	2.08
7	5.50	2.97	5.29	66.23	2.07
8	6.00	3.04	5.66	109.20	2.06
9	6.50	3.10	6.00	180.03	2.06
10	7.00	3.15	6.32	296.83	2.05

QUADRO 07

	X	Y
1º par	1	2
2º par	2	4
3º par	3	7
4º par	4	7
5º par	5	7

Na exemplificaçao dos cálculos usaremos os dados do QUADRO 07, ao lado. São 5 pares de dados (X e Y) que serão digitados nesta ordem. A cada passo o programa informará qual par deverá ser digitado. Na TELA 01 você deverá informar quantos pares de dados serão analizados (neste exemplo, 5) e na TELA 02 você digitará os valores correspondentes a **X** e **Y.** Caso verifiquem erros de digitação o programa deverá ser reiniciado.

TELA 01

```
*******************************************************
* * * *              CORRELAÇÃO E REGRESSÃO            ****
****                  Limite = 200 pares de dados      ****
*******************************************************

****   Quantos pares de dados serao analisados?    ****
```

TELA 02

```
5
 Digitar, nesta ordem,  X e Y do par número   1
1
2
 Digitar, nesta ordem,  X e Y do par número   2
2
4
 Digitar, nesta ordem,  X e Y do par número   3
3
7
 Digitar, nesta ordem,  X e Y do par número   4
4
7
 Digitar, nesta ordem,  X e Y do par número   5
5
7
```

Após a digitação do último par de dados, automaticamente será gerada a TELA 02a para informarmos qual o modelo a ser ajustado.

TELA 02a

Qual o código do modelo a ser ajustado ?
Obs.:(a*e = a vezes e)#(X**b = X elevado ao expoente b)

(1) Y = a + bX - (Linha reta)
(2) Y = a + b(1/X) - (Curva hiperbólica)
(3) Y = a + X**b - (Curva logaritmica)
(4) Y = a*e**(bX) - (Curva exponencial)
(5) Y = a*X**b - (Curva potencial)

Normalmente a escolha do modelo decorre da interpretação da dispersão dos valores de **X** e **Y** em um gráfico.

Muitas vezes não dispomos de um gráfico para analisar. Nesses casos podemos ajustar dois ou mais modelos e optar por aquele que se mostre mais adequado aos dados observados.

A TELA 03 nos mostra os resultados dos cálculos. O coeficiente de correlação r = 0.89 mostra que existe um correlação forte entre as variáveis ‘x’ e ‘y’.

A informação seguinte é o ‘coeficiente de determinação’ r^2 = 0.80. O coeficiente r^2 nos informa o poder explicativo da regressão, ou seja, a porcentagem de variação de ‘y’ que é estatisticamente explicada pela variação de ‘x’.

Em seguida temos as estimativas dos parâmetros da reta de regressão ‘a’ e ‘b’.

Neste exemplo a = 1.50 e b = 1.30. Com os valores de “a” e “b” nós podemos fazer a alocação de pontos em um gráfico para o traçado da reta e também fazer projeções para valores futuros ou desconhecidos.

TELA 03

```
1
**************************************************************
*  Coeficiente  de  correlação = 0.89                        *
*  Coeficiente de determinação = 0.80                        *
*  Coeficientes da Regressão: a =   1.50  b =    1.30        *
*  Erro padrão das estimativas igual a    1.20               *
**************************************************************

*** Quer ajustar outro modelo? ///   0=Não    1=Sim ***
```

Na última linha da TELA 03 nós temos uma avaliação do erro padrão da regressão quando utilizamos a equação da reta Y = 1.50 + 1.30X. O valor de +/- 1.20 significa que as nossas previsões terão, em média, um erro igual a 1.20 unidades acima ou abaixo dos valores calculados pela reta de regressão.

Comumente, costumamos considerar como 'fortes' as correlações iguais ou superiores a 0.85 ($r >= 0.85$), as quais permitem um poder explicativo da regressão acima de 70%, o que é razoável para a maioria das situações práticas. Por exemplo, um valor de $r = 0.7$ proporciona um $r^2 = 0.49$ o que equivale dizer que o comportamento da variavel 'x' conseguiria explicar menos da metade do comportamento da variável 'y'.

O erro padrão das estimativas (apresentado na mesma unidade da variável 'y') é obtido calculando-se as diferenças entre os valores de 'y'da amostra digitada e os valores correspondentes de 'y' calculados pela equação da linha reta.

Se na TELA 03 optássemos por testar outro modelo, como o 'exponencial', por exemplo, obteríamos os resultados mostrados na tela 03a

TELA 03a

```
4
**************************************************
* Coeficiente de correlação = 0.88                *
* Coeficiente de determinação = 0.77              *
* Coeficientes da Regressão: a = 1.94   b = 0.31  *
* Erro padrão das estimativas = 1.75              *
**************************************************
```

Caso calculemos as estimativas dos parâmetros "**a**" e "**b**" para vários modelos (reta, hipérbole, logarítmica, exponencial e potencial) vamos considerar como ***modelo mais bem ajustado, aquele que apresentar o menor erro médio para as estimativas***. Em outras palavras, o modelo escolhido será aquele que apresentar o maior grau de aderência entre os valores observados (valores da amostra) e os valores teóricos calculados pela equação da regressão.

Nos exemplos acima optaríamos pela linha reta (erro padrão = 1.20) pois se ajusta melhor aos dados observados que a curva exponencial (erro padrão = 1.75).

Aplicação: Teste de Iterações
(Aleatoriedade da amostra)
CÓDIGO = 8

Esse teste é utilizado quando queremos avaliar a aleatoriedade de uma amostra. Ele se baseia na ordem (sequência) de inclusão de cada elemento da amostra em relação a uma determinada característica (variável de interesse ou VI). Nessa análise os dados precisam ser agrupados em duas categorias (por exemplo: cara ou coroa, positivo ou negativo, maior ou menor, preto ou branco, verde ou não-verde, acima ou abaixo da mediana, etc. A probablidade de uma unidade amostral ser enquadrada em cada uma das categorias deve ser aproximadamente igual a 50%). O teste avalia se o número de iterações "R" entre as duas categorias indica aleatoriedade ou não. Para amostras com n < 20 os resultados de R devem ser comparados com dados tabelares encontrados em livros de Estatística Não Paramétrica. Para amostras com n >=20 o valor de R é normalmente distribuído e a significância é avaliada pelo teste Z. O programa que usaremos é apropriado para n>=20.

TELA 1

```
******************************************************
* QUANTOS VALORES SERÃO DIGITADOS?   20<= N <=200  *
******************************************************
```

Informar o valor de n (tamanho do conjunto analisado) entre 20 e 200. Nesta exemplificação usaremos n = 20.

Em seguida informe se a variável de interesse é uma contagem (número de exemplares capturados em cada coleta); se é uma mensuraçao (comprimento total de cada exemplar capturado ou se é um atributo (por exemplo, macho ou fêmea). É necessário que os dados sejam digitados respeitando a ordem de inclusão do exemplar na amostra. No exemplo a seguir digitar 1.

TELA 2

```
**************************************************
*          TIPO DA VARIÁVEL A SER ANALISADA              *
* CONTAGEM = 1 // MENSURAÇÃO = 2 // ATRIBUTO(0/1) = 3  *
**************************************************
```

Este exemplo trabalharemos com contagens, então digitaremos o código 1. O programa pedirá para digitarmos os valores (total de exemplares capturados em cada coleta), observando a ordem sequencial das coletas. Os números usados no exemplo são:
20, 25, 39, 36, 15, 18, 32, 19, 45, 33, 40, 18, 17, 10, 16, 29, 35, 13, 46, 27.

TELA 3

```
     ****************************************************
     *          DIGITAR AS CONTAGENS OU ATRIBUTOS          *
     ****************************************************
 DIGITAR A INFORMACAO REFERENTE AO ELEMENTO DE ORDEM    1
20
 DIGITAR A INFORMACAO REFERENTE AO ELEMENTO DE ORDEM    2
25
 DIGITAR A INFORMACAO REFERENTE AO ELEMENTO DE ORDEM    3
39
```

```
                 VALOR DE Z =    0.5
                 N1 =         10   N2 =         10
     MR =    11.000000     SR =    2.0000000     R =   10
  Z < 1.96       ACEITAMOS H0.   AMOSTRA ALEATÓRIA
```

Para chegarmos a esses resultados o programa inicialmente organizou os dados em ordem crescente e calculou o valor da Mediana (neste exemplo = 26). Em seguida, retorna aos dados originais e atribui valor igual a 0 para os dados que ficaram abaixo da mediana e valor 1 para os dados que ficaram acima da mediana. R = 10 informa a quantidade de iterações observadas. MR = 11 é o valor da média de interações esperado em uma sequência aleatória. O desvio padrão (SR = 2).

O valor de z = 0,50 é menor que o valor tabelar (1,96) então aceitamos a hipótese H**o**. Não constatamos diferenças significativas entre a média (MR = 11) e o número de iterações (R = 10). Assim concluimos que a sequência de valores da variável "captura total" nas 20 coletas analisadas mostrou um comportamento aleatório

H**1** é a hipótese que supunha que a sequência de dados analisados poderia estar sofrendo algum efeito ciclico ou de dependência tornando-a não aleatória.

Quando trabalhamos com variáveis contínuas seguimos os mesmos procedimentos anteriores para digitarmos os valores das mensurações, observando sempre a ordem de inclusão de cada elemento no conjunto analisado, durante a realização do experimento .

As variáveis do tipo atributo serão representadas pelos códigos 0 e 1. Para cada elemento incluído no conjuntoanalisado deverá ser digitado o código da categoria na que ele está incluído, por exemplo: 1= verde, 0 = não verde ou 1 = fêmea e 0 = macho. Será então digitada uma sucessão de 'zeros' e 'uns' respeitando sempre a ordem de inclusão de cada elemento.

Aplicação: Teste Qui-quadrado

CÓDIGO = 9

O teste do Qui-quadrado é utilizado quando trabalhamos com variáveis discretas do tipo 'contagem'. A consolidação das contagens são genericamente chamadas de 'frequências'. As contagens efetuadas pelo pesquisador são as 'frequências observadas'.

O teste consiste em comparar essas frequências observadas com as frequências esperadas, que deveriam ocorrer, de acordo com alguma teoria preestabelecida.

A comparação entre as frequências observadas e esperadas mostra que podem existir diferenças entre elas, e estas diferenças podem ser resultantes de algum fator influente ou podem simplesmente ser resultantes da variabilidade natural dos dados, ou seja, uma variação do acaso ou aleatória.

A magnitude dessas diferenças é que vai orientar a hipótese que aceitaremos:

Ho: as frequências observadas (Fo) e as frequências esperadas (Fe) são estatisticamente iguais e as diferenças entre elas podem ser atribuídas ao acaso. Portanto, os dados observados estão de acordo com os requisitos da teoria que originou as frequências esperadas.

H1: as frequências observadas (Fo) não estão de acordo com a teoria preestabelecida. Em outras palavras, as diferenças entre as 'Fo' e as 'Fe' são grandes demais para serem consideradas como aleatórias ou decorrentes da variabilidade natural dos dados analisados.

Um exemplo numérico

QUADRO 09a

	Espécies			
///////	A	B	C	D
Fo	17	10	8	5
Fe	12	8	10	10

Suponhamos que vamos analisar uma coleta de insetos efetuada em determinado nicho habitado por 4 espécies. Os resultados correspondentes às espécies A, B, C e D, num total de 40 exemplares, são as frequências observadas (Fo) do QUADRO 09a.

Trabalhos anteriores mostraram que essas espécies tem os seguintes porcentuais de participação na população:

A = 30%, B = 20%, C = 25% e D= 25%.

Com base nessa teoria podemos estimar as frequências esperadas (Fe). Por exemplo, para espécie A é 30% de 40, ou 12. De modo análogo vamos calcular as demais frequências esperadas que estão registradas no quadro.

Essa modalidade de análise do Qui-quadrado também é chamada de "teste de aderência' por avaliar o grau de aderência ou compatibilidade entre o que foi observado e o que seria esperado de acordo com algum conceito preestabelecido.

Vamos executar agora teste *teste QuiQuadrado* usando os dados do quadro 09a, digitando os pares de frequências observadas e esperadas para as 4 espécies.

TELA 01

```
****************************************************
*****************  Teste Qui-Quadrado  *****************
****************************************************
***********        Digitar o modelo de análise        ***********

**      Teste de aderência (1)  ////  Teste de Independência (2)    **
```

Nesse exemplo digitamos (1) por tratar-se de um teste de aderência.

TELA 02

```
***   Quantos pares de  Fo e Fe serão analisados?    ***
```

Como são 4 espécies, teremos 4 pares de dados

TELA 03

```
*** Informar Fo e Fe, nessa ordem ***

Digitar o par de dados número  1
17
12

Digitar o par de dados número  2
10
8
```

Vamos digitar para cada espécie as frequências observadas e esperadas, nessa ordem. Devemos apertar a tecla 'enter' após a digitação de cada informação. Concluída a digitação do último par, os resultados serão apresentados.

TELA 04

```
****************************************
*    Qui-Quadrado =  5.48   **   GL = 3     *
****************************************
```

Nesse exemplo, aceitamos H_o porque o valor do Qui-quadrado calculado (5.48) não ultrapassa o limite tabelar (7.81) num nível de confiança de 95% ($P < 0.05$). Em outras palavras, as diferenças entre as frequências observadas e esperadas podem ser atribuídas ao acaso (variabilidade natural dos dados). Podemos aceitar que os porcentuais de participação das 4 espécies sejam iguais aos estabelecidos pela teoria. Os graus de liberdade (GL) são iguais ao número de comparações menos um (4 -1 = 3).

Deve ser observada uma restrição ao teste do qui-quadrado. Todas as frequências esperadas devem ser iguais ou maiores que 5. Quando essa condição não for satisfeita é possível combinar duas ou mais casas (células) contíguas, somando-se as Fo e as Fe das casas combinadas, até satisfazer a restrição.

Tabelas de contingência

Além das comparações entre F_o e F_e (teste de aderência) o qui-quadrado também pode ser usado para testar se duas variáveis estão relacionadas entre si ou se são independentes.

Como não temos uma teoria que nos permita estimar as F_e, usaremos uma tabela de contingência, como mostrada no QUADRO 09b. Essa tabela usa os totais marginais (somas das linhas e das colunas) em relação ao total geral da amostra, para estimar as frequências esperadas.

Nesse exemplo simulado, queremos testar a hipótese que armadilhas de diferentes cores são afetadas pela luminosidade no período de coleta , contra a hipótese que 'a cor da armadilha e a 'luminosidade' agem independentemente no processo de captura'. Vamos executar o programa, utilizando as informações da tabela de contingência mostradas no QUADRO 09b, ao lado.

QUADRO 09b

//////////	LUMINOSIDADE		
COR	nublado	média	plena
amarela	26	15	20
vermelha	8	15	32
preta	10	15	56

TELA 05

Teste de aderência (1) // Teste de Independência (2)

Digitar a opção 2

TELA 06

*** Digitar as frequencias da tabela de contingencia ***

*** Quantas linhas tem a tabela? ***
3

Quantas colunas tem a tabela?
3

Informamos que a tabela do ‘QUADRO 09b’ tem 3 linhas e 3 colunas.

TELA 07

** Dado da célula: linha 1 coluna 1 **
26
** Dado da célula: linha 1 coluna 2 **
15
** Dado da célula: linha 1 coluna 3 **
20

TELA 08

** Qui-quadrado = 26.11 GL = 4 **

Após a digitação da última linha, o programa estima as ‘F_e’ para cada casa da tabela, calculando em seguida o valor do qui-quadrado. No exemplo em pauta, o qui-quadrado = 26.11 ultrapassa o valor tabelar igual a 9.49 então rejeitamos H_0.
(GL = número de linhas - 1 vezes o número de colunas - 1, ou, (3 - 1) x (3 - 1). = 4).

Esse resultado nos leva a aceitar a hipótese H_1 que supõe que a performance das armadilhas é afetada por uma associação entre a cor da armadilha e o grau de luminosidade durante o período de coleta.

Aplicação:
Teste de Kolmogorov-Smirnov
CÓDIGO = 10

Vamos exemplificar o teste K&S (Quadro 10a) medindo o grau de concordância (aderência) entre as frequências observadas para 5 classes, (denominadas Grupo-1, Grupo-2, etc), e as frequências teóricas calculadas a partir de uma teoria pré-estabelecida. A hipótese Ho supõe que os dados se distribuam igualmente pelas 5 classes, ou 20% em cada classe. Assim, se essa teoria estiver correta deveríamos ter 2 elementos em cada grupo, totalizando 10 unidades.

O teste K&S vai analisar se as diferenças entre as frequências observadas e as frequências teóricas podem ser atribuídas ao acaso (aceitando Ho) ou se são grandes demais para serem consideradas aleatórias (rejeitando Ho).

Deve ser obeservada uma clara analogia deste teste com o Teste Qui-quadrado. Porém, o teste K&S é menos restritivo permitindo comparações com frequências teóricas menores que 5, o que não é possível pelo Qui-quadrado.

O K&S trabalha com as frequências acumuladas (observadas e teóricas) e calcula a maior diferença entre elas (D_c). O valor calculado é comparado com o valor tabelar (D_t) obtido na tabela do próprio teste a partir do valor de N (tamanho da amostra), neste exemplo, 10.

QUADRO 10a

TESTE K&S - Aderência		
Classe	**%teórica**	**F observ.**
Grupo 1	0.2	0
Grupo 2	0.2	1
Grupo 3	0.2	0
Grupo 4	0.2	5
Grupo 5	0.2	4

TELA 01

```
***************************************************
*          Teste de Kolmogorov-Smirnov            *
***************************************************

* Em quantas classes os dados estão distribuidos?  *
```

Digitar 5, pois estamos trabalhando com 5 classes (Grupo-1, Grupo-2, etc.)

Digitar os nomes das classes. Nomes compostos (por letras e números) não devem ter espaços entre os caracteres. Se necessário, usar hifens.

TELA 02

```
 *** Digitar o nome da classe          1 ***
Grupo-1

 *** Digitar o nome da classe          2 ***
Grupo-2
```

TELA 03

```
** Digitar a proporção (%) teórica da classe = Grupo-1        **
0.2

 ** Digitar a proporção (%) teórica da classe = Grupo-2        **
0.2
```

Como Ho supõe que as frequências se distribuem igualmente pelas 5 classes. As proporções teóricas são iguais a 0.2 (20%).

TELA 04

```
 ** Digitar o número de observações na classe = Grupo-1
0

 ** Digitar o número de observações na classe = Grupo-2
1
```

Digitar as frequências observadas (0, 1 ,0, 5, 4)

TELA 05

```
*************************************************
*                  Dc = 0.500   **   Dt (ver tabela)                 *
*************************************************
```

O valor de Dc, ou diferença máxima entre as frequências acumuladas (observadas e teóricas), vai ser comparado com o valor tabelar para N = 10 (Dt = 0.410). Como Dc (0.500) é maior que o Dt (0.410) rejeitamos a hipótese Ho. Concluimos então que os dados observados não devem ter se originado de uma população igualmente distribuída nas 5 classes consideradas.

Os valores tabelares dos níveis de significância do teste K&S foram formatados em tabelas para valores de n <= 35. Para amostras com tamanhos superiores a 35 o valor tabelar (Dt) deve ser calculado.

Os dados do QUADRO 10b enfoca um exemplo para grandes amostras. Neste caso, considerando o nível de significância de 0.05, o cálculo do valor tabelar é efetuado pela fórmula **Dt = 1.36/(raiz de n).** O próprio programa efetua os cálculos.

O exemplo (fictício) do QUADRO 10b nos mostra os resultados de uma pesquisa realizada em uma pequena comunidade isolada, analisando a distribuição das frequências dos grupos sanguíneos em 40 indivíduos. A tipagem sanguínea originou os resultados das frequências observadas do QUADRO 10b.

QUADRO 10b

TESTE K&S - Aderência		
Classe	**%teórica**	**F observ.**
Grupo O	0.30	17
Grupo A	0.20	10
Grupo B	0.25	8
Grupo AB	0.25	5

Deve ser lembrado que nem sempre nós dispomos de uma teoria pré-estabelecida que nos permita calcular as frequências teóricas. Mas, trabalhos de pesquisas anteriores, levantamento de dados históricos confiáveis ou a dedução através de métodos adequados ao material estudado muitas vezes nos permitem obter valores satisfatórios para as proporções teóricas de diferentes classes nas populações.

Repetindo a rotina de execução

Se digitarmos os dados do QUADRO 10b, seguindo a mesma rotina do exemplo anterior, após reiniciar o aplicativo, vamos chegar aos resultados mostrados no box abaixo. As % teóricas foram obtidas através de pesquisas anteriores.

TELA 06

```
************************************************
*            Dc = 0.175    **    Dt = 0.215            *
************************************************
```

O valor de D_c = 0.175 não supera o valor tabelar (D_t) calculado como 1,36/(raiz de 40) = 0.215. Assim, devemos aceitar a hipótese H_0, concluindo que as frequências amostrais (observadas) não diferem significativamente das frequências teóricas.

Aplicação:
Teste Z
(para proporções)

CÓDIGO = 11

Quando trabalhamos com contagens um dos modelos adequados para o cálculo das probabilidades é a distribuição binomial, que é muito usada, por exemplo, na genética.

Mas, quando as amostras são grandes, os cálculos do binômio de Newton se tornam incômodos e por isso podemos usar uma aproximação da distribuição normal que dá bons resultados quando ‘p’ (proporção de casos favoráveis a um determinado evento na amostra) é igual a 1/2 ou 0.5.

Quando p **é** diferente de 0.5, uma aproximação é possível desde que seja satisfeita uma das condições básicas em relação ao valor ‘p’ e ao tamanho da amostra ‘n’:

Se p < 0.5 então np > 5

Se p > 0.5 então nq > 5

Este aplicativo verifica se as condições básicas para a aproximação binomial-normal são satisfeitas antes de aplicar o teste. Se as restrições não forem satisfeitas o teste é interrompido com a mensagem “***Amostra insuficente. Precisa ampliar a amostra e refazer o teste***”

O teste Z para proporções nos fornece, basicamente, os mesmos resultados que o teste Qui-quadrado na tabela de contingência 2x2.

O teste Z avalia a magnitude da diferença significativa entre as proporções, enquanto o qui-quadrado indica apenas que existe uma diferença significativa entre elas.

No exemplo que ilustra esta aplicação, nós vamos testar se uma amostra de n = 224, e p = 0.79 pode ter se originado de uma população com P = 0.75.

As hipóteses em teste são as seguintes: H_o: p = P e H_1: p /= P.

TELA 01

```
     ******************************************************
  **              TESTE Z PARA PROPORCÕES               **
     ******************************************************

   ***************     Tipo de Teste    ********************

 *          ‘p'da amostra   VERSUS  'P' populacional - Digitar 1   *

 *         'p1' amostra 1    VERSUS   'p2' amostra 2   -  Digitar 2   *
    *******************************************************
```

Nesse exemplo vamos digitar ‘1’ informando que será testada uma estimativa amostral ‘p’ em relação a um parâmetro populacional ‘P’ que pode ser representado como p **versus** P

TELA 02

** Qual o valor do 'P' populacional? **
0.75

Digitamos o valor de P = 0.75

TELA 03

** Qual o valor do 'p' da amostra? **
0.79

Informamos a proporção (%) amostral p= 0.79

TELA 04

** Qual o tamanho da amostra? **
224

No problema foram utilizadas 224 unidades amostrais

TELA 05

```
***************************************************
**                Z =  1.38 /// Aceitamos_Ho             **
***************************************************
```

Após verificar a possibilidade de usar a aproximação binomial-normal é aplicado o teste Z. O resultado Z = 1.38 não ultrapassa o valor tabelar igual a 1.96 e, portanto os valores de p e P não diferem significativamente. Aceitamos a hipótese Ho concluindo que a amostra pode ter se originado de uma população com P = 0.75.

Reiniciamos o programa para testar as proporções de duas amostras independentes, p1 e p2. Vamos digitar a opção 2.

TELA 06

```
****************** Tipo de Teste *******************
*  'p' da amostra  VERSUS  P' populacional  -  Digitar 1  *
*  'p' da amostra 1 VERSUS  p' da amostra 2  - Digitar 2   *
*************************************************
```

TELA 07

```
**      Qual o valor do 'p' da amostra 1    ( p1) ?
**
0.75
```

Neste exemplo simulamos que a proporção 'p_1' seja igual a 3/4 (ou 75% ou 0.75).

No passo seguinte informamos o tamanho da amostra onde foi obtida a proporção p_1, ou seja, $n_1 = 200$.

TELA 08

```
**       Qual o tamanho da amostra 1   (n1) ?    **
200
```

TELA 09

```
**     Qual o valor do 'p' da amostra 2   (p2) ?      **
0.79
```

Na TELA 09 informamos a proporção da segunda amostra. $p_2 = 0.79$

TELA 10

```
**       Qual o tamanho da amostra 2    (n2) ?    **
150
```

Na TELA 10 informamos o tamanho da segunda amostra testada, $n_2 = 150$

TELA 11

```
***************************************************
**                 Z = -0.88  ///  Aceitamos Ho             **
***************************************************
```

Nesta última tela é apresentado o resultado do teste Z e a hipótese que aceitamos. O teste conclui que as proporções (p_1 e p_2) não diferem significativamente entre si.

Aplicação:
Teste de WILCOXON
CÓDIGO = 12

O teste T de Wilcoxon tem uma aplicação análoga ao teste t de Student para dados parelhados. A diferença entre os dois modelos está nos pré-requisitos de cada teste. Enquanto o teste de Student deve ser usado com variáveis contínuas (pesagens, medidas, etc), o teste de Wilcoxon exige apenas que as mensurações sejam feitas em escala ordinal (ou superior).

Este teste pode ser aplicado quando fazemos comparações entre pares homogêneos ou quando observamos um mesmo indivíduo em dois instantes diferentes, ou seja, quando o indíviduo é utilizado como controle de si mesmo.

TELA 01

```
*****************************************************
*                      Teste de WILCOXON                    *
*****************************************************

        *****      Quantos pares de dados?      *****
```

Inicialmente informamos o números de pares que vamos trabalhar. No exemplo mostrado no QUADRO 12 temos os dados referentes às 8 cobaias submetidas a um treinamento para a obtenção de alimentos.

O QUADRO 12 mostra dados fictícios de uma pesquisa envolvendo o treinamento de 8 cobaias em um ambiente controlado de laboratório. O experimento consiste em avaliar a porcentagem de acerto em relação a uma determinada habilidade na obtenção de alimento, 'antes" e 'depois' de um mesmo treinamento para os 8 indivíduos.

Deve ser observado que neste exemplo o 'par" é formado por duas observações de um mesmo indivíduo em dois intantes diferentes. Os valores de 'dif" são as diferenças (com o respectivo sinal) entre as observações.

QUADRO 12

% de acertos			
Par	Antes	Depois	dif
1	63	82	-19
2	42	69	-27
3	74	73	1
4	37	43	-6
5	51	58	-7
6	43	56	-13
7	80	76	4
8	82	85	-3

TELA 02

```
*****        Quantos pares de dados?        *****
8
***          Digitar o par        1         ***
63
82
***          Digitar o par        2         ***
42
69
```

A TELA 02 mostra a digitação dos dois primeiros pares de dados em um total de 8 pares relacionados no QUADRO 12.

TELA 03

```
***          Digitar o par        8         ***
82
85
******************************************************
*****     T =   4.00     *****     Ver Tabela para N =   8 ****
******************************************************
```

Após a digitação do último par, a TELA 03 informa o valor de ‘T’ = 4.00, que é o resultado do Teste de Wilcoxon. Esse valor deve ser comparado com a tabela do teste. Neste exemplo o valor tabelar é igual a 4. Quando o ‘T’ calculado superar o valor tabelar (de acordo com o nível de significância escolhido) aceita-se Ho. Se o ‘T’ calculado for igual ou menor que o valor tabelar aceitamos H_1. Nesse exemplo aceitamos que o treinamento aumentou significativamente a porcentagem de acerto na busca por alimento. Os pares iguais (portanto com dif =0) não são considerados no teste podendo ocorrer então uma redução no valor de N.

Para valores de N até 25 usamos a tabela de valores críticos da prova de Wilcoxon. Para N > 25 os valores de ‘T’ tem uma distribuição aproximadamente normal com média zero e variância unitária.

O exemplo seguinte mostra a utilização de uma amostra com 30 unidades. A entrada dos dados é exatamente igual à do último exemplo, modificando apenas o processamento interno para o cálculo da média e da variância da distribuição teórica de ‘T’.

Assim sendo, mostraremos a primeira e a última tela do processamento e a interpretação do resultado.

TELA 01a

```
*****************************************************
*                  Teste de WILCOXON                *
*****************************************************

*****        Quantos pares de dados?        *****
30
```

TELA 02a

```
***          Digitar o par        30                    ***
19
20
*****************************************************
*****   Zc =  -3.11    *****   Z tabelar =1.96   N =  26   ****
*****************************************************
```

Após a digitação do 30º par é exibida a TELA 02a com o resultado do teste. O valor de Zc (Z calculado) é obtido após o cálculo da média e da variância para a distribuição de ‘T’. Como Zc > 1.96 rejeitamos Ho (P < 0.05).

Aplicação:
Teste MANN-WITHNEY
CÓDIGO = 13

O Teste de Mann-Withney é usado quando queremos testar a hipótese (Ho) que duas amostras independentes possam ter se originado de uma mesma população. Esse teste tem uma eficiência de 95% quando comparado como teste ‘t’ de Student, sem as exigências da prova paramétrica.

Nesse primeiro exemplo trabalharemos com duas amostras (amostra X e amostra Y), onde nx = 5 e ny = 4, como mostrado pela TELA 02 e TELA 03.

Não incluímos o QUADRO com os dados devido ao pequeno tamanho das duas amostras. Os dados a serem digitados no exercício podem ser obtidos neste exemplo do tutorial (telas 02 e 03).

TELA 01

```
*****************************************************
***           TESTE DE MANN-WITHNEY              ***
*****************************************************
```

TELA 02

```
 ***     Tamanho da amostra X :          nx =  ?     ***
5

 ** Para a amostra X  //// Digitar dado        1  *********
78
 ** Para a amostra X  //// Digitar dado        2  *********
64
 ** Para a amostra X  //// Digitar dado        3  *********
75
 ** Para a amostra X  //// Digitar dado        4  *********
45
 ** Para a amostra X  //// Digitar dado        5  *********
82
```

TELA 03

```
 ***        Tamanho da amostra Y:              ny =  ?        ***
4

 ** Para a amostra Y  //// Digitar dado          1  *********
110
 ** Para a amostra Y  //// Digitar dado          2  *********
70
 ** Para a amostra Y  //// Digitar dado          3  *********
53
 ** Para a amostra Y  //// Digitar dado          4  *********
51
```

TELA 04

```
*****************************************************
**      U =  9.00     ***    Ver Tabela para n1 =  4  e  n2  = 5       **
*****************************************************
```

O resultado do teste é apresentado por U = 9.00, o qual deve ser comparado com o valor tabelar do Teste Mann-Withney. A tabela nos fornece a probabilidade (P) associada ao valor de U = 9, para $n_1 = 4$ e $n_2 = 5$. Como P = 0.452 aceitamos Ho, ou seja, aceitamos que as duas amostras se originam de uma mesma população. Sempre que P > 0.05 (5%) aceitamos a hipótese Ho.

Quando as amostras têm um tamanho superior a 20, o programa calcula a média e o desvio padrão da distribuição amostral de U e compara com o valor de U' computado a partir da duas amostras. Valor do do teste Z é comparado com 1.96 caso escolhamos o nível (P<0.05) para medirmos a significância.

A TELA 04a mostra os resultados do teste quando **nx** e **ny** são maiores que 20. Como os demais procedimentos para a entrada dos dados são iguais aos do exemplo anterior, apresentamos apenas a tela do resultado final. Nesse caso rejeitaríamos a hipótese (H_o) pois Z_c = 2.83 excede o limite tabelar = 1.96. Concluimos então que as médias da duas amostras diferem significativamente e aceitamos H_1.

TELA 04a

```
*****************************************************
**                 Zc =   2.83    ***   Z tabelar = 1.96                 **
*****************************************************
```

Aplicação:
Correlação de Spearman
CÓDIGO = 14

A correlação de Spearman é a alternativa não paramétrica da correlação linear estudada na APLICAÇÃO 7 (página 31).

Enquanto a correlação linear (de Pearson) é mais recomendada para variáveis contínuas, a correlação de Spearman apresenta bons resultados quando trabalhamos com variáveis discretas, como as contagens.

As contagens muitas vezes apresentam grandes aplitudes de variação e nesses casos o uso de postos (escala ordinal) não só facilita os cálculos como contorna alguns óbices teóricos já que o cálculo da variância subentende o uso de uma escala intervalar ou de razão nas mensurações que entrarão nos cálculos.

No QUADRO 14, vamos supor que, em um trabalho experimental, 12 cobaias puderam exercitar livremente uma determinada habilidade. A situação 1 (Xi) mostra o número de tentativas de cada indivíduo e a situação 2 (Yi) refere-se ao número de acertos de cada indivíduo. A hipótese Ho supõe que o aumento do número de tentativas aumenta o número de acertos. A hipótese H1 supõe que o número de acertos independe do número de tentativas.

QUADRO 14

Avaliações / Mensurações		
Indivíduo	Situação 1	Situação 2
1	82	42
2	98	46
3	87	39
4	40	37
5	116	65
6	113	88
7	111	86
8	83	56
9	85	62
10	126	92
11	106	54
12	117	81

TELA 01

```
****************************************************
**          CORRELAÇÃO DE SPEARMAN              ***
****************************************************

***********     Quantos pares de dados?    **************
```

Usaremos os dados do QUADRO 13, portanto n = 12

TELA 02

****** Digitar o par X Y número 1 **********
82
42

Os 12 pares de dados devem ser digitados sequencialmente e, após o último par, o resultado é mostrado.

TELA 03

****
** Coeficiente de SPEARMAN = 0.818 *** t = 4.50 **
****

O coeficiente de correlação r_s igual a 0.818 mostra uma relação forte entre as duas variáveis analisadas. A significância estatística do coeficiente é avaliada pelo teste t de Student (t = 4.50).

O valor de r_s (coeficiente de correlação de Spearman) igual a 0.818 indica que, os dados da pesquisa sugerem um aumento da habilidade proporcionalmente ao número de tentativas. Aceitamos, portanto, H_1.

Para amostras com n>= 10 é possivel avaliar a significância do valor de r_s. A partir desse tamanho o coeficiente tem uma distribuição aproximadamente normal o que nos permite usar o teste 't' de Student para avaliar a significância de r_s.

O valor de GLé igual ao número de pares de dados menos 2 ou GL = n - 2. Nesse exemplo (12 - 2 = 10) e, como o valor tabelar ao nível de 95% (2.23), é superado pelo valor calculado (4.50) concluimos que o valor de r_s difere significativamente de zero.

Quando n <10 o resultado do teste 't' será igual a ***.

Aplicação:
Teste Kruskal-Wallis
CÓDIGO = 15

O teste de Kruskal-Wallis é o equivalente não-paramétrico da Análise de Variância, modalidade um critério de classificação (também chamado de inteiramente casualisado).

Através desse modelo vamos testar a hipótese Ho (as amostras se originaram de uma mesma população, contra a hipótese H1 (as amostras vieram de populações diferentes. As mensurações devem estar, no mínimo em escala ordinal

Nesse exemplo foram formados 4 grupos, cada um com 5 repetições. Cada grupo recebeu um determinado tipo de tratamento.

A variável analisada é o aumento do tamanho (em centímetros) de cada planta em relação ao comprimento inicial (antes de iniciar o experimento).

QUADRO 15

Tratamentos			
A	**B**	**C**	**D**
20	26	26	40
28	35	36	49
26	30	30	39
21	22	29	35
30	27	32	18

TELAS 01-02

```
****************************************************
***********  Teste de KRUSKAL & WALLIS  ************
****************************************************

********  Quantos tratamentos nas colunas?  *******
4

******    Quantas repeticoes nas linhas?     *******
5
```

Informamos que trabalharemos com 4 tramentos nas colunas e 5 repetições (alocadas nas linhas) dentro de cada tratamento. Digitar segundo as linhas.

Os dados da TELA 03 mostram a digitação da primeira linha da tabela.

TELA 03

```
 Digitar dado da linha   1 coluna  1
20
 Digitar dado da linha   1 coluna  2
26
 Digitar dado da linha   1 coluna  3
26
 Digitar dado da linha   1 coluna  4
40
 Digitar dado da linha   2 coluna  1
```

Após a digitação do último dado da última linha, o resultado da análise (H) é apresentado e também o número de graus de liberdade para acessarmos a tabela do teste Qui-quadrado.

TELA 04

```
****************************************************
***          H =    5.44  ***   Graus de liberdade =  3        ***
****************************************************
```

O resultado do teste, H = 5.44, deve ser comparado com a tabela do teste Qui-quadrado, para 3 graus de liberdade (GL = número de tratamentos - 1). Neste exemplo o Qui-quadrado tabelar (ao nível de 95%) é igual a 7.82 (portanto o valor calculado é menor que o valor tabelar) e, por isso, aceitamos H_0.

O teste concluiu que os tratamentos analisados não apresentaram diferenças significativas em relação ao crescimento das plantas.

Muito embora o teste de Kruskal-Wallis possa ser usado com tratamentos de diferentes tamanhos, a modalidade adotada neste aplicativo se restringe a tratamentos de mesmo tamanho.

Como vimos no QUADRO 15 cada um dos quatro tratamentos têm cinco repetições.

O uso da tabela do Qui-quadrado neste teste exige que tenhamos ***no mínimo 5 repetições*** em cada um dos tratamentos.

Para amostras menores existem tabelas de probabilidades associadas aos valores de H que podem ser encontradas em livros de estatística não paramétrica.

Aplicação: Regressão curvelínea

CÓDIGO = 16

Na aplicação 7 tratamos da correlação e regressão lineares , ou seja, quando os valores correspondentes aos pares de dados X e Y das amostras, são plotados em um gráfico e tem uma distribuição parecida com uma linha reta.

Foram abordados também alguns modelos que podem ser tratados como se fossem lineares mediante uma transformação logarítmica (curvas potencial, exponencial e logarítmica) ou, como no caso da hipérbole, quando trabalhamos com o inverso de X.

Em todos esses casos os modelos são facilmente linearizados, pois a regressão é do tipo $Y = a + bX$ (modelo linear aditivo, MLA) no qual trabalhamos com as estimativas de dois parâmetros: **a** e **b**.

Existem outros modelos de curvas que, mediante alguns cálculos adicionais, também podem ser linearizados. Os próximos modelos que trataremos aqui requerem as estimativas de 3 parâmetros (**a**,**b**, **c**) tomando como referência o modelo linear aditivo:

$$Y = a + bX + cX^2.$$

Nesta aplicação trabalharemos com 4 grupos de curvas aqui denominados Modelo 1, Modelo 2, Modelo 3 e Modelo 4.

Essas curvas podem ser observadas nas figuras correspondentes que serão mostradas nos exercícios de cada modelo.

TELA 01

```
**************************************************
*********** REGRESSOES CURVELINEAS ***************
**************************************************

**   Quantos pares de dados serão digitados ?   **
```

TELA 02

```
Digitar X e Y, nesta ordem, do par numero          1 ***

1

37
```

Na TELA 01 digitamos 7 inicialmente, informando que vamos trabalhar com sete pares de dados.

Na TELA 02 o aplicativo solicita que digitemos os valores dos pares X Y que estão ralacionados no QUADRO 16.

Esses valores foram previamente ajustados à curva do Modelo 1 de modo a obtermosvalores inteiros para as estimativas dos parâmetros **a**, **b**, **c** que definirão a equação da regressão.

Após a digitação do sétimo par de dados, o aplicativo pergunta, na TELA 03, qual o modelo que queremos ajustar.

QUADRO 16

X	Y
1	37
2	22
3	13
4	10
5	13
6	22
7	37

TELA 03

```
Digitar X e Y, nesta ordem, do par numero        7 ***
7
37

****************************************************
*          QUAL O  CÓDIGO DO MODELO A SER AJUSTADO?   *
****************************************************
*             (1)  Curvas da pagina 60 do tutorial     *
*             (2)  Curvas da pagina 61 do tutorial     *
*             (3)  Curvas da pagina 63 do tutorial     *
*             (4)  Curvas da pagina 65 do tutorial     *
****************************************************
```

A primeira etapa no estudo da regressão é a confecção de um gráfico de modo a verificar a existência ou não de alguma tendência ou comportamento que sugira uma curva na dispersão dos pontos.

O distribuição espacial dos pontos no gráfico, na maioria das vezes é suficiente para evidenciar uma tendência curvelínea que orientará a escolha do modelo a ser ajustado.

Outras vezes, no entanto, alguns trechos da distribuição dos pontos se assemelham a mais de uma das curvas relacionadas gerando dúvidas sobre qual modelo é o mais adequado.

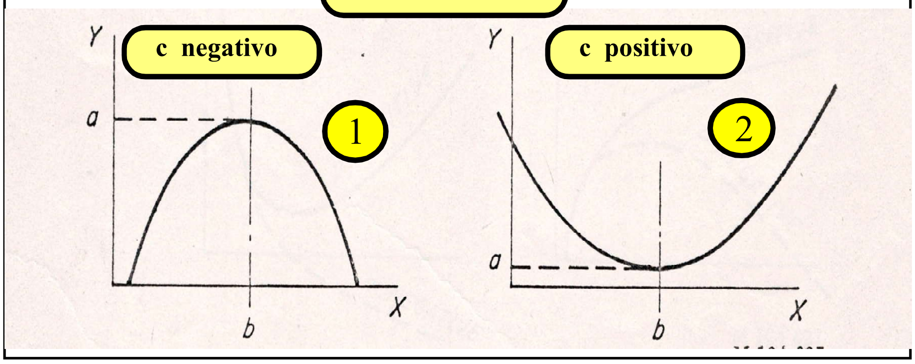
Em situações assim devemos analisar o valor do erro padrão da regressão para os diversos modelos disponíveis.

Quanto menor o valor do erro padrão (EP) do modelo testado, melhor a equação da regressão estará explicando o relacionamento funcional das variáveis X e Y.

MODELO 1

Equação da curva de regressão: $\mathbf{Y = a + c(x-b)^2}$

FIGURA 16a



Suponhamos que o nosso gráfico tenha mostrado um comportamento parecido com o desenho 2 da figura 16a.

Assim, vamos digitar o código 1 na TELA 03 e o aplicativo mostrará as estimativas de **a**, **b**, **c** e do **erro padrão**

TELA 04a

```
*****************                    *******************
*                 CURVAS D0 MODELO 1                   *
*              (Ver pagina 60 do tutorial)             *
*                                                      *
*      A =   10.00      B =   4.00      C =   3.00     *
*                                                      *
*          Erro padrão da Regressão =       0.00       *
********************************************************
```

Os valores mostrados na TELA 04a são as estimativas dos parâmetros da curva da regressão do Modelo 1: **a = 10, b = 4** e **c = 3**.

A substituição desses valores na equação da regressão ficará assim:

$$Y = 10 + 3(X - 4)^2$$

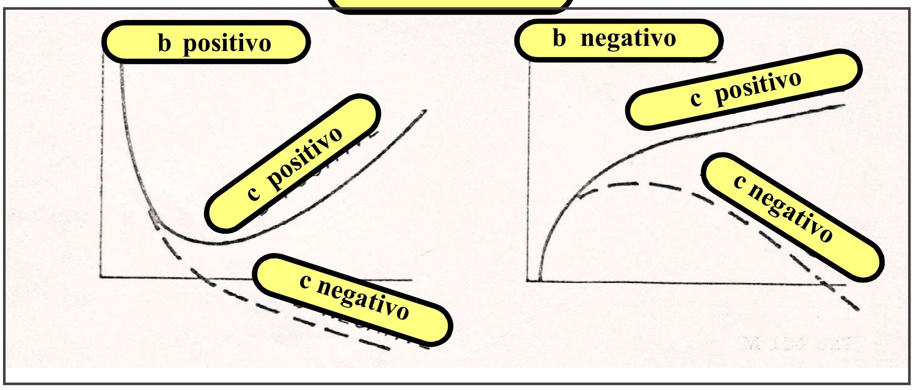
Ainda na TELA 04 o aplicativo pergunta se queremos ajustar outro modelo aos dados do QUADRO 16. A título de exercício ajustaremos outros modelos, começando com o Modelo 2. Inicialmente digitamos '1' informando que queremos ajustar outro modelo. A TELA 03 será novamente exibida e vamos digitar o código '2'

```
****************************************************
*            Quer ajustar outro modelo?            *
*            0 = Não    ////    1 = Sim            *
****************************************************
```

MODELO 2

Equação da curva de regressão: $Y = a + b(1/X) + cX$

FIGURA 16b



No MODELO 2 nós temos quatro alternativas de curvas , mostradas na FIGURA 16b, que podem ser ajustadas aos dados das nossas amostras.

Os sinais (+ e -) das estimativas de **b** e **c** nos ajudam a identificar o modelo mais adequado naqueles casos em que a visualização do gráfico não é muito esclarecedora.

Os procedimentos na aplicação do Modelo 2 são rigorosamente iguais aos do Modelo 1.

```
**************************************************
*              CURVAS DO MODELO 2                *
*         (Ver pagina 61 do tutorial)            *
*                                                *
*     A =   -32.74  B =   64.24  C =    7.74     *
*                                                *
*        Erro padrão da Regressao =     6.24     *
**************************************************
```

```
**************************************************
*            Quer ajustar outro modelo?          *
*             0 = Não    ////    1 = Sim         *
**************************************************
```

A substituição dos valores de **a, b, c** na equação da regressão ficará assim:

$$Y = -32{,}47 + 64{,}24(1/X) + 7{,}74X$$

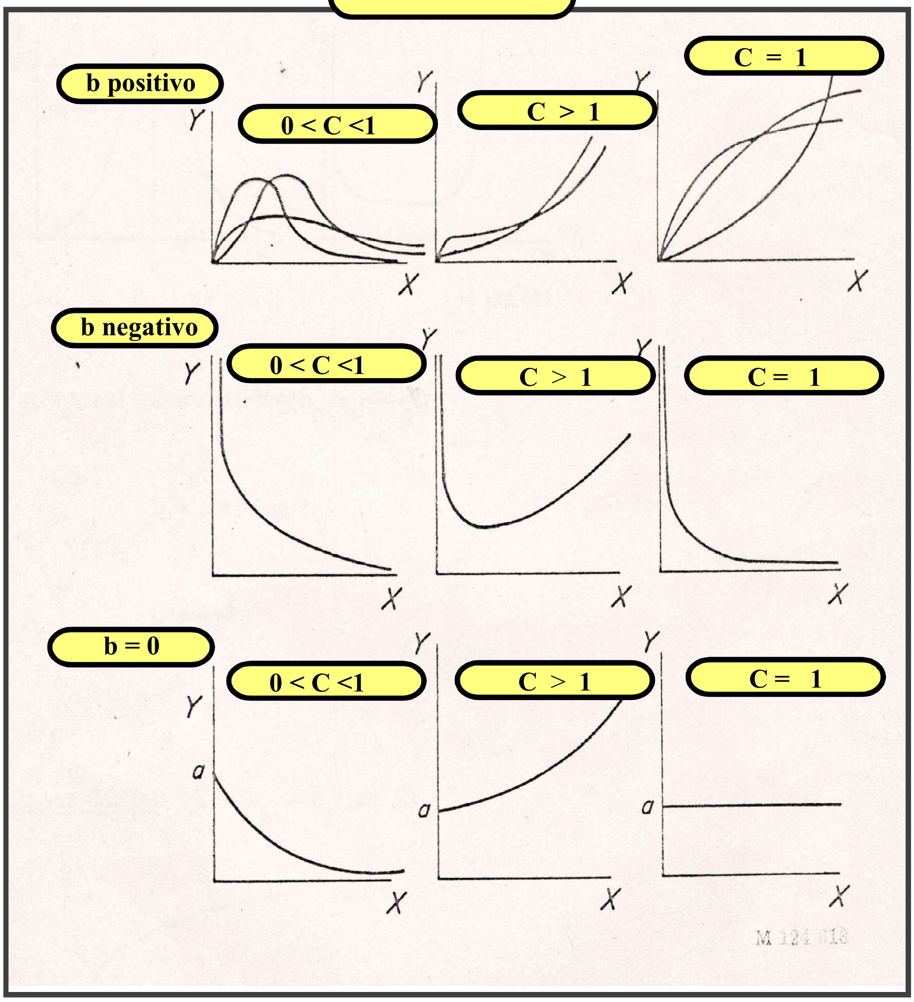
A TELA 04b mostra as estimativas de **a**, **b** ,**c** no ajuste do Modelo 2**.** O erro padrão (EP2) igual a 6.24 indica que o Modelo 2 não é um bom ajuste quando comparamos esse resultado com o erro padrão igual a zero do Modelo 1 (EP1 = 0).

Na sequência vamos ajustar o Modelo 3 e analisar o comportamento do erro padrão

MODELO 3

Equação da curva de regressão: $\mathbf{Y = a.X^b. c^X}$

FIGURA 16c



No **Modelo 3** nós temos 13 curvas diferentes que podem ser ajustadas aos nossos dados. As rotinas de uso são as mesmas dos modelos anteriores. Digitamos o código 3 na TELA03 e os resultados de **a**,**b**,**c** serão mostrados na TELA 04c

TELA 04c

```
**************************************************
*                 CURVAS DO MODELO 3                 *
*              (Ver pagina 00 do tutorial)            *
*                                                    *
*      A =   19.47      B = -2.58      C =    2.20   *
*                                                    *
*         Erro padrão da Regressão =       5.37      *
**************************************************
```

A equação da regressão é a seguinte:

$$Y = 19.47\ (X)^{-2.58}\ (\ 2.20)^{X}$$

Analisando comparativamente o valores dos erros padrões obtidos até agora concluimos que ajuste do Modelo 03 (EP3 = 5.37) também é inferior ao ajuste do Modelo 01 (EP1 = 0).

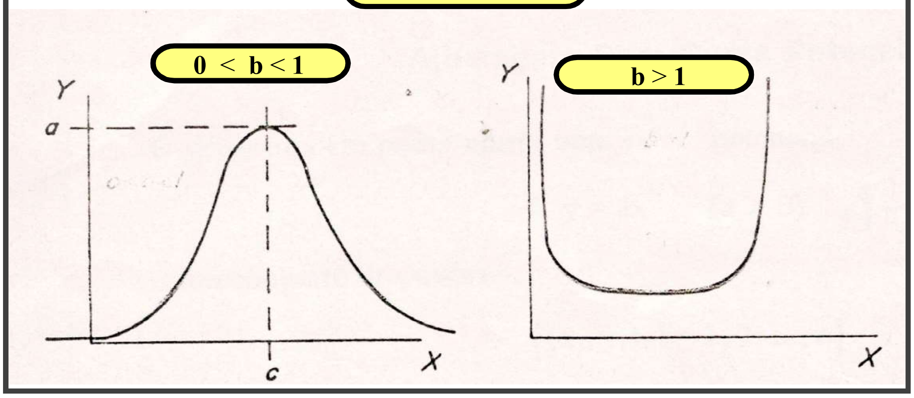
Naquelas situações em que o Modelo 3 é o mais indicado e a inspeção gráfica não permite uma identificação segura da curva correspondente devemos utilizar os sinais de **b** e a amplitude de variação de **c** para optarmos corretamente.

A seguir analisaremos o Modelo 4.

MODELO 4

Equação da curva de regressão: $Y = a.b^{(X - c)^2}$

FIGURA 16d



No **Modelo 4** nós temos 2 curvas que podem ser ajustadas neste aplicativo.

As rotinas de uso são as mesmas dos modelos anteriores. Digitamos o código **4** na TELA03 e os resultados de **a**,**b**,**c** serão mostrados na TELA 04d

TELA 04d

```
**************************************************
*              CURVAS DO MODELO 4               *
*           (Ver pagina 00 do tutorial)          *
*                                                *
*     A =  11.34      B =  1.15      C =  4.00   *
*                                                *
*     Erro padrão da Regressao =      2.27       *
**************************************************
```

Na TELA 04d mostramos as estimativas de **a**, **b** ,**c** no ajuste do Modelo 4**.** O erro padrão (EP4) igual a 2.27 indica que o Modelo 4 não é um bom ajuste quando comparamos esse resultado com o erro padrão do Modelo 1 (EP1 = 0), embora seja um ajuste superior aos dos Modelos 2 e 3 (EP2 = 8.83 e EP3 = 5.37).

A equação da regressão é a seguinte:

Equação da regressão: $$\mathbf{Y = 11.34\ (1.15)^{(X - 4)^2}}$$

Leituras recomendadas

É importante lembrar que o conjunto aplicativo/tutorial, ora apresentado, não substitui os livros textos de Bioestatística na medida em que o tutorial aborda superficialmente os conhecimentos mínimos e necessários para a aplicação correta dos testes.

O uso simultâneo do tutorial e do aplicativo permite que o usuário possa "dialogar" com os diversos testes de maneira interativa usando simultaneamente um smartphone ou tablet (para o tutorial) e um PC ou notebook (para o aplicativo). Isso facilita significativamente a exemplificação do uso prático da bioestatística, desde a digitação dos dados até a interpretação dos resultados.

Existem vários livros de Bioestatística no mercado editorial brasileiro que atendem de maneira satisfatória as diversas áreas (cursos de graduação e pós-graduação, elaboração de teses e pesquisas acadêmicas) quanto ao embasamento teórico necessário para a correta aplicação dos testes bioestatísticos. Esse fato dispensa as citações bibliográficas no tutorial, já que o objetivo principal deste aplicativo limita-se ao uso prático daqueles conhecimentos.

Aplicação:

POTENCIAL

AGREGATIVO / DISCRIMINATIVO

CÓDIGO = 17

O aplicativo número 17 foi desenvolvido para calcular o grau de agregação/discriminação (D) de um caracter taxonômico e também para calcular os pesos de um sistema de ponderação (numa escala de 1 a 6) que informa a contribuição do carácter para a formação de subgrupos em um grupo (que reúna 3 ou mais espécies) (WA) ou para individualização dessas mesmas espécies (WD).

O termo 'grupo' está sendo utilizado nos casos em que as espécies estudadas não representam a totalidade de um gênero, mas apenas parte dele.

Assim, as conclusões serão válidas apenas para o grupo estudado e não devem ser estendidas ao gênero correspondente. A base metodológica para esses cálculos pode ser encontrado na bibliografia citada no final deste capítulo.

São duas obras que tratam do tema.

Maia (2021) apresenta a metodologia para variáveis discretas (contagens e codificações), baseada no cálculo da média geométrica. Maia (2022) trata das variáveis contínuas (mensurações) utilizando análise de variância, Teste F (um critério de classificação), complementado pelo teste de Tukey.

Nos boxes amarelos, reproduziremos as telas do aplicativo numeradas na sequência em que elas aparecem durante a execução.

Os boxes verdes mostrarão comentários explicativos sobre as telas reproduzidas.

TELA 01

```
*****************************************************************
* INFORME O TIPO DE VARIÁVEL QUE SERÁ ANALISADA *
*  (1 =   Variavel Discreta - Contagens e codificações)        *
*  (2 =   Variaveis continuas  - Mensurações -  (Pesos, medidas)   *
*****************************************************************
```

Digite ‘1’ para variáveis discretas

TELA 02

```
**************************************************
*                    ATENÇÂO                     *
*     Palavras compostas devem ser grafadas com hifen     *
*          EX: Apis-melifera, amarelo-claro          *
**************************************************

**************************************************
*          Digitar o número ‘n’ de espécies          *
**************************************************
```

Informar o número de espécies que fazem parte do grupo analisado

TELA 03

```
**************************************************
*              Type the species name  1              *
**************************************************
```

Digitar o nome (ou código numérico, ou letra), para identificar as espécies. Essa rotina se repetirá para cada uma das ‘n’ espécies (1,2,....,n).

TELA 04

```
**************************************************
*        Qual carácter taxonômico será analisado?        *
**************************************************
```

Digitar o nome do caráter taxonômico para o qual será calculado o valor de **D '** (grau de discriminação / agregação).

TELA 05

```
*******************************************************
*          Quantos status esse carácter apresenta no grupo?              *
*******************************************************
```

Informar o número ‘m’ de diferentes status que o caracter apresenta dentro do grupo analisado.

TELA 06

```
*****************************************************
*             Qual status será identificado pelo código   1                *
*****************************************************
```

Nesta etapa informe o nome de cada status que será identificado por um número sequencial (código) (1,2,3...m). Exemplo: cor amarela = código 1 / cor branca = código 2 / cor preta = código 3...

TELA 07

```
******************************************************
*      Code:  Labrum-color         in          Plebeia-juliani              *
******************************************************
```

Por exemplo: após digitar o código 1 (status = cor amarela) o aplicativo pedirá a confirmação para evitar erros de digitação.

TELA 08

```
******************************************************
*                         Voce digitou     cor-amarela                          *
*                        Esta correto?  0=Nao  -  1=Sim                         *
******************************************************
```

Se a informação estiver correta digite 1 e a análise prossegue. Se não estiver, digite 0 e a digitaçáo será solicitada novamente.

- Após digitar a informação referente à ultima espécie, o aplicativo calcula o grau de discriminação/agregação **D**, em uma escala entre 0 (zero) e 5, e também os valores de WD e WA.

TELA 09

```
**************************************************
*******      Grau de Discriminacao/Agregacao      *******
*******************      D = 2.5      **********************
**************************************************

**************************************************
*                    Caracter Moderado                    *
**************************************************
*     Peso do Potencial Discriminativo:  WD = 3 .5        *
*     Peso do Potencial Agregativo:      WA =  3.5        *
**************************************************
```

- O aplicativo também informa a interpretaçáo de D (e.g. Caracter Moderado). Os pesos dos Potenciais de Discriminação (WD = 3.5) e de Agregação (WA = 3.5) são calculados em uma escala entre 1 e 6. Neste exemplo, o carácter analisado tem a mesma importância taxonômica, quer agregando quer discriminando os elementos dentro do grupo estudado.

TELA 10

```
*****************************************************
* Você quer analisar outro caracter?     0=Nao  -  1=Sim    *
*****************************************************
```

Se a análise envolver vários caracteres taxonômicos essas rotinas podem ser repetidas (digitando 1 volta-se à TELA 04) tantas vezes quantos forem os caracteres.

TELA 01a

```
************************************************** *
* INFORMAR O TIPO DAS VARIAVEIS QUE SERAO ANALISADAS *
*  (1 = Variaveis discretas  - Contagens e codificacoes)        *
*  (2 = Variaveis continuas - Mensuracoes (pesos,medidas)       *
****************************************************
```

Digite ‘2' para variáveis contínuas

TELA 02a

```
**************************************************
*           Quantas especies integram o grupo analisado?          *
**************************************************
```

Digitar o número ‘n’ de espécies que fazem parte do grupo.

Exemplificaremos com os dados de 6 especies (n = 6) do gênero Plebeia (*P.juliani, P.meridionalis, P.droryana, P.emerina, P.remota* e *P.saiqui* representadas pelas letras J,M,D,E,R e S, respectivamente*).* A variável analisada é a mensuração da ‘Largura da mandíbula’ de 5 exemplares (m = 5) de cada espécie. Ver Maia (2022).

Table 1 - Width of the jaw of six species of the genus Plebeia

J	R	S	D	E	M
0.90	1.40	1.20	1.15	1.50	1.00
1.00	1.80	1.30	1.20	1.30	0.90
1.00	1.40	1.50	1.15	1.20	0.90
0.90	1.45	1.30	1.20	1.20	0.70
0.90	1.40	1.30	1.00	1.20	0.80

O aplicativo vai solicitar, na sequência, que você digite o nome (ou um código numérico, ou código alfabético,etc) para identificar cada uma das ‘n’ espécies . Lembre-se que os nomes compostos (p.ex. *Plebeia-juliani)* devem ser escrito com hífen.

TELA 03a

```
***************************************************
*                 Digitar o nome da especie  1                  *
***************************************************
```

O aplicativo pedirá a digitaçáo dos nomes das 6 espécies (Plebeia-juliani, Plebeia-meridionalis, etc)

TELA 04a

```
***************************************************
*              Qual carater taxonomico sera analisado?              *
***************************************************
```

Digitar o nome do carácter. Neste exemplo “largura-da-mandíbula’. Nomes compostos devem ser grafados com hífen.

TELA 05a

```
***************************************************
*           Quantos exemplares de cada especie analisada?           *
***************************************************
```

Cada espécie será representada pelas mensurações originadas de amostras, todas do mesmo tamanho, com ‘m’ elementos. Neste exemplo serão 6 amostras (n = 6), com 5 elementos cada uma (m = 5).

TELA 06a

```
***************************************************
*          Informar o valor tabelar de 'q'  do Teste de Tukey          *
*              Tratamentos =         -        G.L.residual  =                  *
***************************************************
```

Precisamos informar o valor de ‘q’ para a aplicação do Teste de Tukey para avaliarmos as diferenças entres as médias das 6 amostras de 5 exemplares de cada especie. O valor de ‘q’ pode ser obtido nas tabelas anexadas ao final deste tutorial.

O cabeçalho horizontal apresenta o número de tratamentos. Neste exemplo cada espécie é considerada como um tratamento na Ánálise de Variância, portanto n = 6.

Precisamos também calcular os graus de liberdade da varîância residual (GLresidual) o que é facilmente obtido assim: GL = n(m-1). Lembrando que ’n’ é igual ao número de espécies (6) e ‘m’ é igual ao número de exemplares medidos em cada espécie (5).

Portanto GLresidual = 6(5-1) ou GL residual = 6x4 = 24

Após obter esses dois valores vou usar o cabeçalho horizontal = 6 e o cabeçalho vertical = 24. No encontro da linha com a coluna está o número q = 4.373 que será digitado.

TELA 07a

```
*********************************************************
*             Digitar exemplar  1 da especie Plebeia-juliani             *
*********************************************************
```

TELA 08a

```
*********************************************************
*             Digitar exemplar  2 da especie Plebeia-juliani             *
*********************************************************
```

A TELA 07a e TELA 08a mostram como os dados serão solicitados pelo aplicativo. Após digitar a última informação da primeira espécie, digitaremos as informações da segunda espécie, assim por diante até a última espécie.

TELA 09a

****
* Caracter analisado: Largura-da-mandíbula *
****

****
********* Grau de Discriminaçao / Agregacao *********
******************* D = 2.6667 *******************
****

****
* Caracter intermediário *
****
* Peso do Potencial Discriminativo: WD = 3.67 *
* Peso do Potencial Agregativo: WA = 3.33 *
****

- Esta tela mostra os resultados da avaliação do potencial agregativo/ discriminativo do carácter taxonômico 'Largura-do-Labro'.
- O valor numérico (em uma escala de 0 a 5) é igual a 2.67.
- Esse valor classifica o caracter como 'intermediário', ou seja, ele agrega e discrimina em proporções aproximadamente iguais.
- Os valores de WD e WA (numa escala entre 1 e 6) expressam numericamente os dois potenciais semelhantes.

Screen 10a

****
* Você quer analisar outro carácter? 0=Nao - 1=Sim *
****

Se quiser analisar outro carácter, digite '1' e o aplicativo recomeçará a partir da TELA 04a.

VALORES TABELARES PARA O TESTE DE TUKEY

Nas páginas seguintes transcrevemos duas tabelas com os valores de q" para a aplicação do Teste de Tukey. Escolhemos o nível de 95% por ser o mais utilizado nas pesquisas biológicas.

No cabeçalho horizontal (linha superior da tabela) indicaremos o número de espécies que integram o grupo analisado.

No Cabeçalho vertical (primeira coluna à esquerda) indicaremos os graus de liberdade da variância residual (ou variância dentro, ou variância do acaso, ou variância do erro amostral ou simplemente erro).

No cruzamento da linha com a coluna encontraremos o valor de 'q' para digitarmos na TELA 06a.

Essas tabelas foram publicadas originalmente por H. Leon Harter, conforme indicação bibliográfica a seguir:

HARTER, H. LEON. 1960. Tables of Range and Studentized Range. Ann. Math. Statist. 31 (4) 1122 - 1147.

TABLE 3 (*Continued*)

P = .95

$\nu \backslash n$	2	3	4	5	6	7	8	9	10
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
19	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474

$\nu \backslash n$	11	12	13	14	15	16	17	18	19
1	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83
2	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57
3	9.717	9.946	10.15	10.35	10.53	10.69	10.84	10.98	11.11
4	8.027	8.208	8.373	8.525	8.664	8.794	8.914	9.028	9.134
5	7.168	7.324	7.466	7.596	7.717	7.828	7.932	8.030	8.122
6	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508
7	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097
8	6.054	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.802
9	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579
10	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405
11	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265
12	5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089	6.151
13	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055
14	5.364	5.463	5.554	5.637	5.714	5.786	5.852	5.915	5.974
15	5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846	5.904
16	5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786	5.843
17	5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734	5.790
18	5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688	5.743
19	5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647	5.701
20	5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610	5.663
24	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545
30	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429
40	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313
60	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199
120	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.044	5.086
∞	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974

TABLE 3 (*Continued*)

P = .95

ν \ n	20	22	24	26	28	30	32	34	36
1	59.56	60.91	62.12	63.22	64.23	65.15	66.01	66.81	67.56
2	16.77	17.13	17.45	17.75	18.02	18.27	18.50	18.72	18.92
3	11.24	11.47	11.68	11.87	12.05	12.21	12.36	12.50	12.63
4	9.233	9.418	9.584	9.736	9.875	10.00	10.12	10.23	10.34
5	8.208	8.368	8.512	8.643	8.764	8.875	8.979	9.075	9.165
6	7.587	7.730	7.861	7.979	8.088	8.189	8.283	8.370	8.452
7	7.170	7.303	7.423	7.533	7.634	7.728	7.814	7.895	7.972
8	6.870	6.995	7.109	7.212	7.307	7.395	7.477	7.554	7.625
9	6.644	6.763	6.871	6.970	7.061	7.145	7.222	7.295	7.363
10	6.467	6.582	6.686	6.781	6.868	6.948	7.023	7.093	7.159
11	6.326	6.436	6.536	6.628	6.712	6.790	6.863	6.930	6.994
12	6.209	6.317	6.414	6.503	6.585	6.660	6.731	6.796	6.858
13	6.112	6.217	6.312	6.398	6.478	6.551	6.620	6.684	6.744
14	6.029	6.132	6.224	6.309	6.387	6.459	6.526	6.588	6.647
15	5.958	6.059	6.149	6.233	6.309	6.379	6.445	6.506	6.564
16	5.897	5.995	6.084	6.166	6.241	6.310	6.374	6.434	6.491
17	5.842	5.940	6.027	6.107	6.181	6.249	6.313	6.372	6.427
18	5.794	5.890	5.977	6.055	6.128	6.195	6.258	6.316	6.371
19	5.752	5.846	5.932	6.009	6.081	6.147	6.209	6.267	6.321
20	5.714	5.807	5.891	5.968	6.039	6.104	6.165	6.222	6.275
24	5.594	5.683	5.764	5.838	5.906	5.968	6.027	6.081	6.132
30	5.475	5.561	5.638	5.709	5.774	5.833	5.889	5.941	5.990
40	5.358	5.439	5.513	5.581	5.642	5.700	5.753	5.803	5.849
60	5.241	5.319	5.389	5.453	5.512	5.566	5.617	5.664	5.708
120	5.126	5.200	5.266	5.327	5.382	5.434	5.481	5.526	5.568
∞	5.012	5.081	5.144	5.201	5.253	5.301	5.346	5.388	5.427

ν \ n	38	40	50	60	70	80	90	100	
1	68.26	68.92	71.73	73.97	75.82	77.40	78.77	79.98	
2	19.11	19.28	20.05	20.66	21.16	21.59	21.96	22.29	
3	12.75	12.87	13.36	13.76	14.08	14.36	14.61	14.82	
4	10.44	10.53	10.93	11.24	11.51	11.73	11.92	12.09	
5	9.250	9.330	9.674	9.949	10.18	10.38	10.54	10.69	
6	8.529	8.601	8.913	9.163	9.370	9.548	9.702	9.839	
7	8.043	8.110	8.400	8.632	8.824	8.989	9.133	9.261	
8	7.693	7.756	8.029	8.248	8.430	8.586	8.722	8.843	
9	7.428	7.488	7.749	7.958	8.132	8.281	8.410	8.526	
10	7.220	7.279	7.529	7.730	7.897	8.041	8.166	8.276	
11	7.053	7.110	7.352	7.546	7.708	7.847	7.968	8.075	
12	6.916	6.970	7.205	7.394	7.552	7.687	7.804	7.909	
13	6.800	6.854	7.083	7.267	7.421	7.552	7.667	7.769	
14	6.702	6.754	6.979	7.159	7.309	7.438	7.550	7.650	
15	6.618	6.669	6.888	7.065	7.212	7.339	7.449	7.546	
16	6.544	6.594	6.810	6.984	7.128	7.252	7.360	7.457	
17	6.479	6.529	6.741	6.912	7.054	7.176	7.283	7.377	
18	6.422	6.471	6.680	6.848	6.989	7.109	7.213	7.307	
19	6.371	6.419	6.626	6.792	6.930	7.048	7.152	7.244	
20	6.325	6.373	6.576	6.740	6.877	6.994	7.097	7.187	
24	6.181	6.226	6.421	6.579	6.710	6.822	6.920	7.008	
30	6.037	6.080	6.267	6.417	6.543	6.650	6.744	6.827	
40	5.893	5.934	6.112	6.255	6.375	6.477	6.566	6.645	
60	5.750	5.789	5.958	6.093	6.206	6.303	6.387	6.462	
120	5.607	5.644	5.802	5.929	6.035	6.126	6.205	6.275	
∞	5.463	5.498	5.646	5.764	5.863	5.947	6.020	6.085	