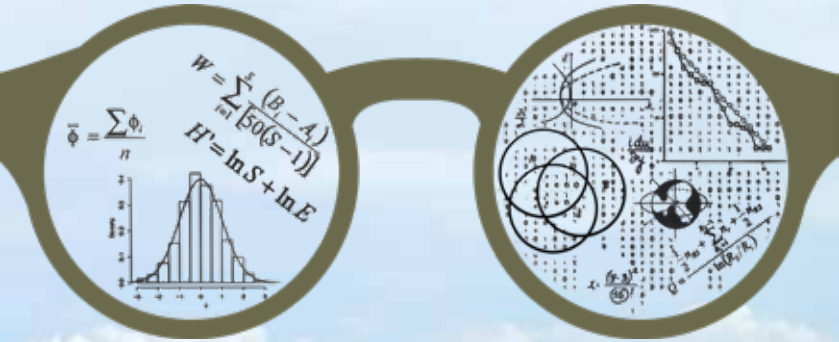


BIOESTATÍSTICA

CURSO PRÁTICO UTILIZANDO R E EXCEL

JOSÉ ROBERTO BOTELHO DE SOUZA

BIOESTATÍSTICA

CURSO PRÁTICO UTILIZANDO R E EXCEL

JOSÉ ROBERTO BOTELHO DE SOUZA

BIOESTATÍSTICA

CURSO PRÁTICO UTILIZANDO R E EXCEL

JOSÉ ROBERTO BOTELHO DE SOUZA

2019

Capa: Wesley de Oliveira Neves

Edição: José Roberto Botelho de Souza

Revisão: José Roberto Botelho de Souza

Projeto Gráfico e Diagramação: Isabella Giordano

Bibliotecária Responsável: Liliane Campos Gonzaga de Noronha (CRB4-1702)

S729b Souza, José Roberto Botelho de.
Bioestatística [recurso eletrônico] : curso prático utilizando R e Excel / José Roberto Botelho de Souza. – Recife : Ed. UFPE, 2019.

Inclui bibliografia
ISBN 978-85-415-1175-9 (online.)

1. Bioestatística. 2. Distribuição (Teoria da probabilidade). 3. Amostragem (Estatística). 4. Estatística – Programas de computador. I. Título.

570.15195 CDD (23.ed.) UFPE (BC2019-103)

Índice para catálogo sistemático:

Primeira Edição
EDITORA UFPE, Recife, 2019

SUMÁRIO

PARTE II – ESTATÍSTICA NO EXCEL

INTRODUÇÃO

A natureza apresenta vários processos e padrões, como a rotação e translação da terra, o movimento das marés, a força da gravidade, o ciclo de vida dos organismos, a distribuição espacial das espécies, época de reprodução, número de filhotes, entre outros. Estes processos e eventos são, em sua maioria, descritos e estimados utilizando técnicas estatísticas.

A estatística é uma ferramenta fundamental na busca de padrões existentes na natureza. Ela é um instrumento útil para organizar, descrever e inferir sobre os processos naturais, adequada para descrever tendências, diferenças e variações, e testar inferências.

Os principais objetivos da estatística são o planejamento, amostragem, organização, apresentação, análise e interpretação dos dados. Suas atribuições incluem: descrição, estimativa dos parâmetros de uma população e a prova de hipóteses.

A estatística pode ser dividida em três partes: **descritiva, probabilidade, e inferencial**.

- A estatística descritiva organiza e sumariza os dados
- A probabilidade está associada a análises de incertezas de fenômenos aleatórios
- A estatística inferencial pode determinar se há diferenças em uma ou mais amostras. Isto é, se as diferenças observadas são devido ao acaso ou se elas provêm de populações distintas. Muitas das inferências estatísticas comparam conjunto de dados obtidos com distribuições ou variações aleatórias de números. As técnicas de inferência iniciam a partir de hipóteses sobre a população.

CATEGORIAS DE DADOS

A estatística analisa os dados, que possuem natureza diversa: podemos dividi-los em qualitativos e quantitativos. Sexo e presença em certo tipo de

vegetação são dados qualitativos nominais, já a freqüência de dispersores é qualitativa ordinal. Os dados de abundância são quantitativos discretos e os de altura quantitativos contínuos. Definimos dado como o valor da variável para cada unidade amostral (por exemplo, biomassa de uma semente de feijão é de 263 mg) e variável como a característica que queremos analisar, como biomassa, pressão temperatura, abundância, riqueza de espécies, cor, sexo, altura, tipo de solo.

VARIÁVEIS QUALITATIVAS

- Nominal
- Ordinal

Variável nominal

Quando conseguimos apenas classificar o objeto de estudo, sem nenhuma ordenação de valor ou categoria, estamos utilizando a escala nominal. Essas categorias são mutuamente exclusivas. Por exemplo, se classificamos camisetas em relação à cor, como verde, amarela, branca, vermelha, azul e cinza, estamos usando um símbolo para representar cada categoria. Se agendarmos o pagamento do IPVA dos carros com relação ao último número da placa, os carros serão classificados numa escala de 0 a 9. O número dos jogadores de futebol ou o número de cada lote de organismos de uma coleção são exemplos de escala nominal.

Propriedades formais → a única relação existente aqui é a equivalência. Todos os indivíduos de uma classe são equivalentes. E o relacionamento dos objetos de estudo é equivalente. A relação de equivalência é reflexiva (x = x, para todo x), simétrica (se x=y, então y=x) e transitiva (se x=y e y=z, então x=z). Na escala nominal, todos os símbolos são equivalentes, por isto podem ser trocados sem que se alterem as relações entre os números. Se eu classifico cinco áreas de amostragem em 1, 2, 3, 4 e 5, depois resolvo trocar estes símbolos para 101, 102, 103 104 e 105 ou para A, B, C, D, e E, isto não altera nenhuma propriedade das áreas amostradas.

Variável ordinal

Quando os elementos de uma categoria de dada escala, além de serem diferentes das outras categorias, possuem uma relação entre as categorias. Por exemplo: extra-alto, alto, médio, pequeno, mínimo. Outro exemplo:

abundante, frequente, regular, ocasional, raro. A diferença fundamental entre uma escala nominal e uma escala ordinal é que a escala ordinal incorpora não somente a relação de equivalência (=), mas também a relação "maior do que" (>). Esta última é irreflexiva, assimétrica e transitiva.

Como qualquer transformação que preserva a ordem não altera a informação contida em uma escala ordinal, a escala se diz única a menos de uma transformação monotônica. Ou seja, não interessa que números sejam atribuídos a um par de classes ou aos membros dessas classes, desde que os membros da classe maior recebam os valores mais altos. Com o escalonamento ordinal, podemos comprovar hipóteses um grande número de estatísticas não paramétricas, também chamadas de estatísticas de ordenações ou de postos.

VARIÁVEIS QUANTITATIVAS

As variáveis quantitativas podem ser classificadas de duas maneiras: como discretas ou contínuas. O dado contínuo permite a introdução de uma infinidade de valores em um intervalo qualquer, como a altura de um animal, por exemplo. Já o dado discreto admite apenas valores inteiros, como, por exemplo, o número de ninhos de aves em uma ilha. Para estes dados, todas as estatísticas paramétricas comuns (e.g., médias, desvios-padrão, correlações de Pearson) são aplicáveis, assim como as provas paramétricas comuns.

PARTE I — FUNDAMENTOS

Capítulo 1

ESTATÍSTICA DESCRITIVA

Os dados apresentam várias características que podem ser expressas através de medidas estatísticas e gráficas. Essas medidas podem ser populacionais ou amostrais, onde população ou universo de estudo é todo o conjunto de dados de uma variável que está sendo estudado estatisticamente. Realizamos o censo, quando mensuramos toda a população. Na maioria das pesquisas, coletamos amostras da população. Amostra é uma parcela representativa da população. Isto é, um subconjunto da mesma. As amostras dever ser representativas e imparciais. Representativas por conterem em proporção tudo o que a população possui qualitativa e quantitativamente. Imparciais, pois todos os elementos da população têm a mesma chance de fazer parte da amostra.

A população é descrita através de parâmetros, são medidas que descrevem ou caracterizam a população, como a média ou desvio padrão da mesma.

Medidas de tendência central

Os parâmetros de tendência central são medidas que indicam o centro de distribuição dos dados. As mais comuns são média, mediana e moda. Estas medidas representam uma simplificação dos dados, já que a amostra será representada por apenas um valor.

Média aritmética pode ser representada pela letra μ e calculada pela equação 1.1:

$$\mu = \frac{X_1 + X_2 + \ldots + X_n}{N} \quad \text{ou} \quad \mu = \frac{\sum_{i=1}^{n} X_i}{N} \qquad \textbf{eq. 1.1}$$

Onde: 'x' são as observações e 'N' é o número total de observações

Na prática, sempre trabalhamos com amostras, trabalhando assim com a estimativa da média $\bar{x}$.

A média se localiza no centro, quando a distribuição dos dados é simétrica (e.g. Distribuição normal), Fig. 1.1. Em distribuições de dados assimétricas, Fig.1.2, A média não é uma boa medida de tendência central, por ser muito afetada pelos valores extremos.

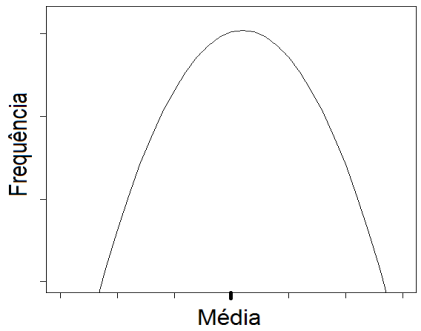


Figura 1.1 – Curva simétrica

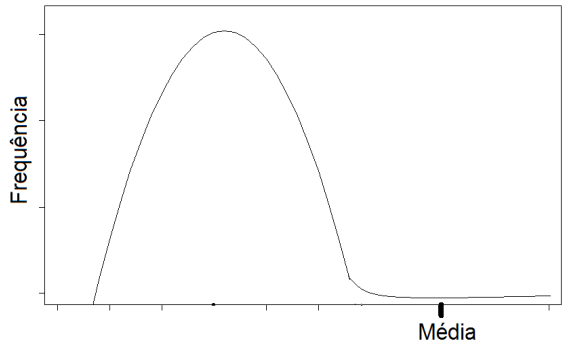


Figura 2.2 – Curva assimétrica

Para que a média aritmética seja considerada um estimador imparcial é preciso satisfazer às seguintes condições:

1. Observações realizadas de forma aleatória
2. Observações independentes entre si
3. Observações estão dentro de um universo maior que apresenta distribuição normal.

Exemplo: a altura média dos alunos da turma de análise de dados é:

1,82; 1,65; 1,73; 1,69; 1,78; 1,72; 1,59; 1,85; 1,70; 1,70; 1,68

$$\bar{x} = \frac{1{,}82 + 1{,}65 + 1{,}73 + 1{,}68}{11} = 1{,}72\ M$$

Média ponderada é uma forma de média aritmética, sendo definida como o conjunto de números $X_1, X_2 \ldots X_n$ que estão associados aos pesos $W_1\ W_2 \ldots W_n$.

$$x_p = \frac{W_1X_1 + W_2X_2 + \ldots + W_nX_n}{W_1 + W_2 + \ldots + W_n} \quad \text{ou} \quad x_p = \frac{\sum_{i=1}^{n} W_iX_i}{\sum W_i} \qquad \textbf{eq. 1.2}$$

Exemplo:

Foram capturados 5 organismos com um amostrador de 1m², 3 com um de 0,5 m² e um com um amostrador de 0,25 m². A média de organismos capturados por m² é:

$$1\text{m}^2 = \text{peso } 1,\ 0{,}5\text{m}^2 = \text{peso } 2,\ 0{,}25\text{m}^2 = \text{peso } 4$$

$$x_p = \frac{5\times1 + 3\times2 + 1\times4}{1+2+4} = 2{,}5 ind.\times m^{-2}$$

Média geométrica Entre n valores (x), é a raiz de índice n do produto desses valores. É utilizado em situações de aumentos sucessivos.

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} \qquad \textbf{eq. 1.3}$$

Exemplo: a taxa de engorda das codornas foi de 15% no primeiro mês, 12% no segundo e 30% no terceiro:

$$\sqrt[3]{1{,}15 \times 1{,}12 \times 1{,}3} = \sqrt[3]{1{,}6744} = 1{,}187$$

Se a massa de referência fosse 200:

200 × 1,15=230

230 × 1,12=257,6

257,6 × 1,3= 334,88

Ou

$200 \times 1{,}187 = 237{,}49$

$237{,}49 \times 1{,}187 = 282{,}01$

$282{,}01 \times 1{,}187 = 334{,}88$

Média Harmônica A média harmônica equivale ao inverso da média aritmética dos inversos de 'n' valores. Ela é a média adequada quando se trabalha com grandezas inversamente proporcionais (e.g.,velocidade e tempo, juros e valor total). Ela apresenta uma média menor que as médias aritméticas e geométricas (eq.1.4).

Ex: faça a média harmônica de: 1,2,3,4

$$MH = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} \qquad \textbf{eq. 1.4}$$

$$MH = \frac{4}{\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}} = \frac{4}{\frac{12+6+4+3}{12}} = \frac{4}{\frac{25}{12}} = 1{,}92$$

Média móvel simples é a média aritmética simples entre o número de elementos selecionados em sequência (eq.1.5). Ela é utilizada para identificar a tendência dos dados de uma sequência espacial ou temporal.

$$MMS = \frac{d_1 + d_2 + \ldots + d_n}{n} \qquad \textbf{eq. 1.5}$$

Exemplo: realizar a média móvel de 2 em 2 dados.

Dados	40	52	74	37	1	11	80	13	75
Média móvel $[(x_n+x_{n+1})/2]$		46	63	55,5	19	6	45,5	46,5	44

Mediana (Md) é o valor central de um conjunto de elementos, após a ordenação dos dados, de acordo com a sua ordem de grandeza. Se o número de amostras é par, a mediana será a média dos valores centrais.

A mediana não é afetada pela grandeza dos valores extremos e sim pelo número de elementos do conjunto. Se a distribuição é assimétrica, a mediana

é uma melhor medida de posição do que a média aritmética. Sua grande desvantagem é que não pode ser tratada algebricamente.

Exemplo: dado o conjunto de dados: 2, 3, 3, 5, 5, 5, 7, 7, 8, 10, 11, 15, 16. A mediana md = 7

n	1	2	3	4	5	6	7	8	9	10	11	12	13
valores	2	3	3	5	5	5	7	7	8	10	11	15	16

A mediana do conjunto de dados 2, 4, 4, 8, 10, 11 é md= (4+8)/2= 6

N	1	2	3	4	5	6
valores	2	4	4	8	10	11

Moda (Mo) é o valor que apresenta a maior frequência numa distribuição. Ela pode ou não existir.

Exemplo:

Amodal: 1, 3, 4, 5, 10, 25, 26, 38, 49, 51 (sem moda)

Unimodal: 3,3, 4, 5, 5, 6, 6, 7, 7,7, 7, 7, 9, 10, 10, 11, 13, 13,17,18 moda = 7

Bimodal: 1, 1, 2,2, 2, 3, 4, 5, 7, 9, 9, 9, 11,12, 13 moda = 2 e 9

O fato de uma distribuição apresentar mais de uma moda pode ser um indício de heterogeneidade da amostra ou população, Fig. 1.3.

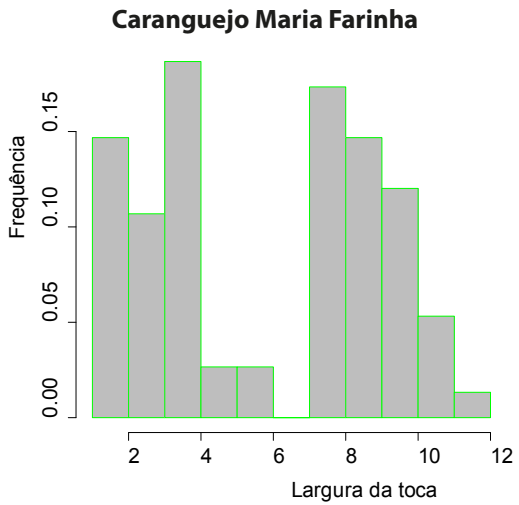


Figura 1.3: Histograma da largura de tocas do caranguejo Maria Farinha (*Ocypode quadrata*) em praias locais.

Medidas de dispersão

Medidas de dispersão e de tendência central são importantes na descrição de uma população, contribuindo para definir o padrão de distribuição dos dados.

Amplitude total ou amplitude de variação: é a diferença entre o maior e o menor valor encontrado. É a medida mais simples de dispersão.

Limitações:

- Não detalha a variação do conjunto de dados, pois mostra apenas os extremos, não informando sobre os dados intermediários.
- Dificilmente a amostra representa a amplitude populacional dos dados. Portanto, há uma sub-estimativa da amplitude dos mesmos.

É importante em trabalhos taxonômicos ou de ecologia descritiva.

Variância é a soma dos quadrados dos desvios de uma média, é representada por σ^2 para a população (eq.1.6):

$$\sigma^2 = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_n - \mu)^2}{N}$$

$$\text{ou} \quad \sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \qquad \textbf{eq. 1.6}$$

E por s^2 para as amostras (eq. 1.7):

$$s^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \ldots + (X_n - \bar{x})^2}{n-1}$$

$$\text{ou} \quad s^2 = \frac{\sum (X_i - \bar{x})^2}{n-1} \qquad \textbf{eq. 1.7}$$

Exemplo:

n	**1**	**2**	**3**	**4**	**5**	**Soma**	**Média**
Valores	3	13	8	4	7	**35**	**7**
Desvios²	16	36	1	9	0	**62**	

$$s^2 = \frac{\sum (X_i - \overline{x})^2}{n-1} = \frac{62}{4} = 15,5$$

A equação abaixo (eq.1.8) É mais precisa, pois não envolve arredondamentos:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$ **eq. 1.8**

Exemplo:

n	**1**	**2**	**3**	**4**	**5**	**Soma**	**Média**
Valores	3	13	8	4	7	**35**	7
Valores²	9	169	64	16	49	**307**	

$$s^2 = \frac{307 - \left(\frac{35^2}{5}\right)}{4} = 15,5$$

Desvio padrão é definido como sendo a raiz quadrada da variância, e expressa o desvio de cada um dos elementos em relação à média:

$$s = \sqrt{s^2}$$ **eq. 1.9**

O desvio padrão (D.P.) Pode ter um valor numérico maior do que a média. Isto geralmente é uma indicação de que a distribuição é assimétrica.

Erro padrão ou erro da média expressa a variação existente entre o conjunto de médias, correspondendo ao desvio padrão das médias.

$$s_x = \frac{s}{\sqrt{n}} \quad \text{ou} \quad s_x = \sqrt{\frac{s^2}{n}}$$ **eq. 1.10**

O erro padrão (E.P.) Sempre é menor que o desvio padrão amostral. Por isto, muitos pesquisadores preferem utilizar o primeiro. Mas, a decisão para

escolher qual medida de dispersão utilizar depende da representatividade dos dados. Se os dados são representativos da população inteira, utilize o erro padrão. Mas se são limitados ao conjunto amostral utilizado, utilize desvio padrão. Deve-se preferir o D.P. Que mostra claramente a variabilidade dos dados existentes, e dá menos importância à generalidade.

Coeficiente de variação é uma medida abstrata que independe das unidades em que foram medidos os dados. Ele expressa o desvio padrão que obteríamos se a média representasse o índice 100 (eq.1.11).

$$\begin{array}{cc} \overline{x} & s \\ 100 & CV \end{array} \qquad CV = \frac{s}{\overline{x}} \times 100 \qquad \textbf{eq. 1.11}$$

O coeficiente de variação é importante para comparar a variação entre duas séries de dados medidos em unidades diferentes, ou entre variáveis diferentes. Por exemplo, o número de poliquetas em 2 cm^2, média = 50 e s = 35, comparados com o volume de poliquetas em 1cm^3, média = 80 e s = 50.

$$CV_{área} = \frac{35}{50} = 0,7 \qquad CV_{vol.} = \frac{50}{80} = 0,62$$

Pode-se verificar que os dados de área são mais variáveis que os dados de volume.

Quantis, quartis, decis percentis. Geralmente, quando utilizamos a mediana, como medida de tendência central, utilizamos os quantis, como medidas de dispersão, pois a amplitude dos dados contém desvios, sendo menos informativa. Podemos dividir os dados em 4 partes iguais, os quartis.

- O primeiro quartil (25%)
- O segundo quartil seria a mediana (50%)
- O terceiro quartil (75%)
- Intervalo interquartil é a diferença entre o valor do 3º quartil (percentil 75) e o valor do 1º quartil (percentil 25), Fig. 1.7.

  Ex.:

			Q1			2		Q3			
n	1	2	3	4	5	6	7	8	9	10	11
dados	2	3	5	8	12	14	21	28	33	44	45
			25%			50%		75%			

$Q_1 = (5+8)/2 \quad = 6,5$
$Q_2 \quad = 14$
$Q_3 = (28+33)/2 \quad = 30,5$

Dados marginais são observações consideradas 'fora do intervalo', isto é, observações acima e abaixo de 90 e 10 percentis, respectivamente. São considerados extremos valores acima e abaixo de 95 e 5 percentis, respectivamente.

ORGANIZAÇÃO DOS DADOS

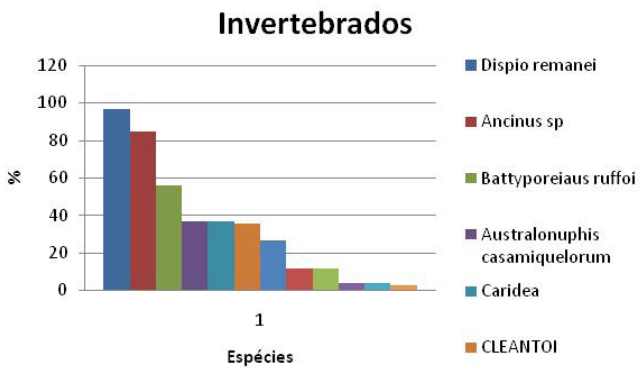
Organizando os dados em tabelas e gráficos

Os dados, geralmente são numerosos e estão dispostos aleatoriamente na planilha de dados. Há a necessidade de organizá-los para se ter uma noção do conjunto dos mesmos. Esta informação pode ser organizada em tabelas e gráficos. Recomenda-se que todas as tabelas e gráficos de um estudo fiquem em apenas um arquivo. Há vários modos de organizar os dados em planilhas, por exemplo, optamos por colocar as variáveis nas linhas e amostras nas colunas, e repetimos isto para todas as planilhas. Medidas de tendência central, de dispersão e tabelas e gráficos possibilitam conhecer a distribuição de dados.

Gráficos

Os gráficos **são** uma forma de apresentação dos resultados. Eles devem ser simples (objetivos), e capazes de mostrar de modo claro o resultado, como no ditado "uma imagem vale mais que mil palavras". Existem muitos tipos de gráficos, apresentamos a seguir os mais comuns:

Gráficos em Barra – em forma de barras verticais ou horizontais, contínuas ou descontínuas. Mostra a variação de uma ou mais variáveis, Fig. 1.4.



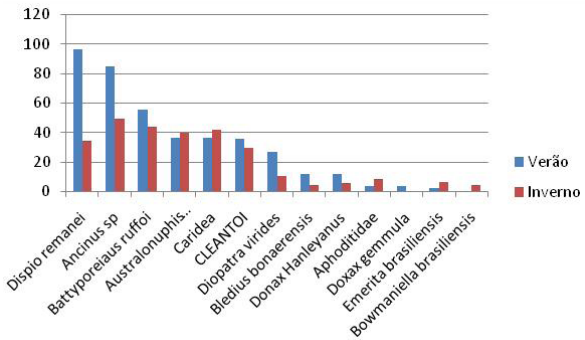


Figura 1.4: Gráficos em barras

Gráfico em setores – são gráficos circulares univariados, onde cada variável é uma fatia do mesmo, Fig. 1.5.

1.5: Gráfico em setores

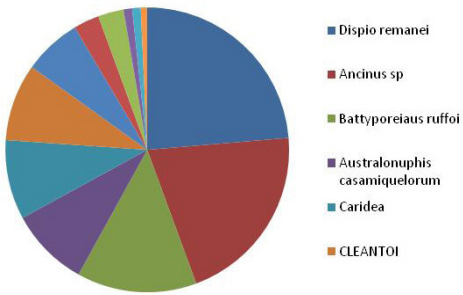


Figura 1.5: Gráfico em setores

Gráfico em linha – quando representam apenas uma variável, são equivalentes aos gráficos em barra e em setor, Fig. 1.6.

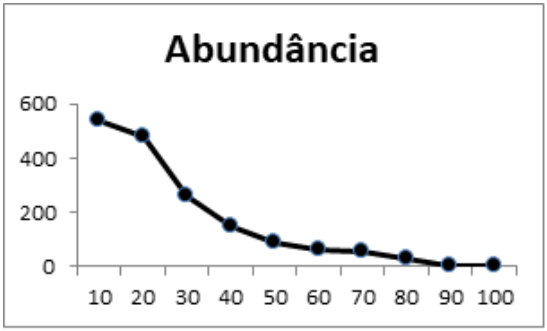


Figura 1.6: Gráfico em linha

Gráficos box-plot – propiciam a representação de medidas de tendência central (e.g., média, mediana) com medidas de dispersão (e.g., desvio padrão, quartis), Fig 1.7.

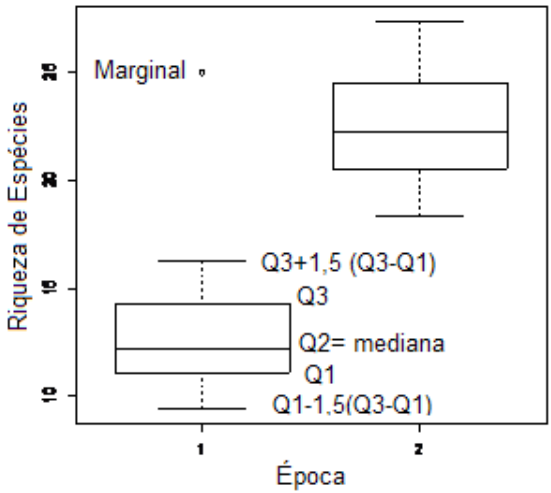


Figura 1.7: Box-plot da Riqueza de espécies no Porto de Suape,1=Seco, 2= Chuvoso.

Gráficos de dispersão – Mostram a relação entre duas variáveis, Fig. 1.8.

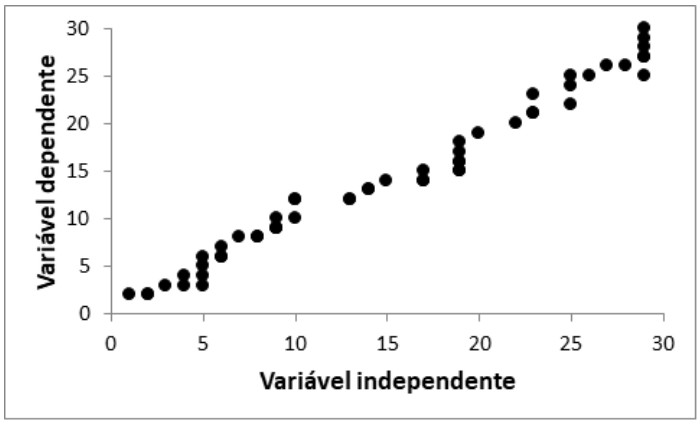


Figura 1.8: Gráfico de dispersão

Histogramas – Uma variável composta por um elevado número de dados pode ser agrupada em tabelas de frequência de ocorrência (F.O.) e, depois, visualizadas em histogramas que possibilitam visualizar a distribuição dos dados.

Para construir a tabela de frequência, precisamos definir o número de classes e a amplitude de cada classe. Há vários métodos para definir o número de classes mas, deve sempre prevalecer o bom senso. Isto é, o número de classes não deve ser tão pequeno que mascare a distribuição de dados nem tão grande que dificulte a visualização dos mesmos, algo entre 10 e vinte classes. Pode-se determinar o número de classes seguindo alguma das várias regras existentes. Dois exemplos de regras simples são: definir o número de classes através da raiz do número total de dados, ou através da equação de Sturges:

- Número de classes $\approx \sqrt{n}$
- Número de classes $\approx 1 + 3{,}222 \times \log n$ (equação de Sturges)

Também existem várias regras para estabelecer a amplitude de cada classe, a regra mais simples seria dividir a amplitude dos dados pelo número de classes:

$$\text{Amplitude da classe} = \frac{\text{Valor máximo} - \text{valor mínimo}}{\text{Número de classes}}$$

Exemplo:

Comprimento do cefalotórax (mm) de 100 crustáceos:

5,3	8,4	8,5	10,5	10,5	10,5	11,2	8,9	9,6	9,7	7,7	8,3	9,5	9,1	10,6
6,6	8,3	8,7	10,7	10,3	10,8	11,8	8,1	9,7	9,6	8,7	8,5	9,6	9,6	10,6
7,5	9,0	8,8	10,2	10,9	10,3	10,6	8,1	9,1	9,5	8,6	8,2	9,8	9,1	
8,0	8,9	9,1	9,7	9,2	9,4	11,9	8,7	10,4	10,5	8,8	8,7	9,6	9,9	
7,4	8,2	9,8	10,0	9,1	10,0	8,2	10,2	10,8	10,6	9,2	9,8	8,4	9,7	
7,4	8,1	9,4	9,9	9,4	9,4	8,1	10,9	10,9	10,7	9,8	9,9	8,6	9,8	
9,0	11,4	9,0	9,3	10,6	11,1	9,7	10,2	9,7	10,2	9,3	9,3	10,6	9,4	

Número de classes = $\sqrt{100} = 10$

Número de classes = $1 + 3{,}222 \times \log 100 = 6{,}222$

$$\text{Amplitude da classe} = \frac{11{,}9 - 5{,}3}{10} = 0{,}65$$

Para facilitar a comparação com outros dados, escolhemos intervalos de 0,5 mm

Limites de classe (mm)			F.O.
	⟜●	5,5	1
5,5	⟜●	6,0	0
6,0	⟜●	6,5	0
6,5	⟜●	7,0	1
7,0	⟜●	7,5	3
7,5	⟜●	8,0	2
8,0	⟜●	8,5	12
8,5	⟜●	9,0	12
9,0	⟜●	9,5	18
9,5	⟜●	10,0	22
10,0	⟜●	10,5	11
10,5	⟜●	11,0	13
11,0	⟜●	11,5	3
11,5	⟜●	12,0	2

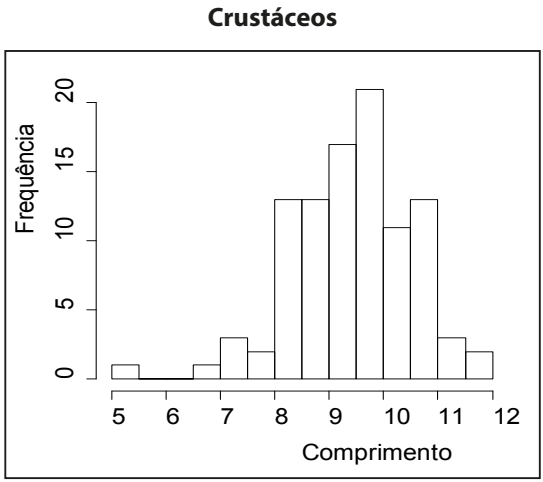


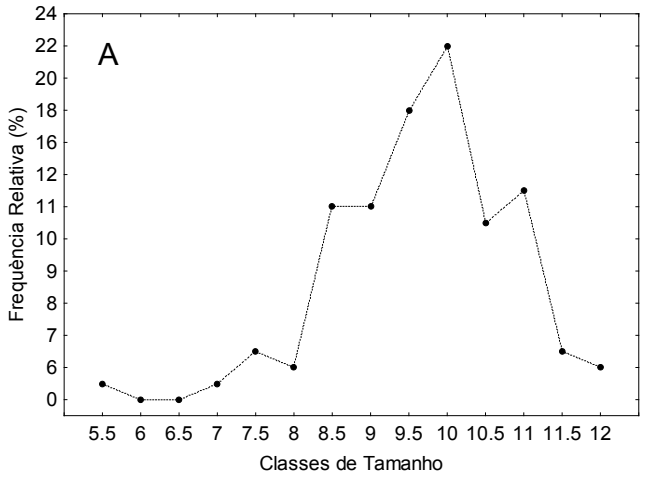
Figura 1.9: Histograma de Frequência-comprimento (mm) de crustáceos.

Frequência relativa e acumulada

A frequência relativa (F.R.) É importante para ver a representatividade de um dado em relação ao total de dados e a frequência acumulada (F.Ac.) Permite observar a tendência deste acúmulo.

A frequência relativa indica a probabilidade de ocorrência de certo evento. Por exemplo, na tabela abaixo, a probabilidade de determinado indivíduo ter entre 7,5 e 8 mm é de 2%.

Limites de classe (mm)			F.O.	F.R.	F.Ac.
	⊶•	5,5	1	1	1
5,5	⊶•	6,0	0	0	1
6,0	⊶•	6,5	0	0	1
6,5	⊶•	7,0	1	1	2
7,0	⊶•	7,5	3	3	5
7,5	⊶•	8,0	2	2	7
8,0	⊶•	8,5	12	12	19
8,5	⊶•	9,0	12	12	31
9,0	⊶•	9,5	18	18	49
9,5	⊶•	10,0	22	22	71
10,0	⊶•	10,5	11	11	82
10,5	⊶•	11,0	13	13	95
11,0	⊶•	11,5	3	3	98
11,5	⊶•	12,0	2	2	100
Soma			100		



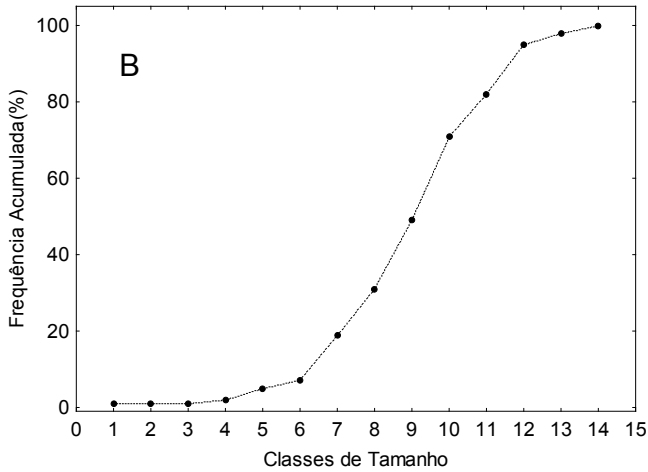


Figura 1.10: Frequência Relativa (A) e Acumulada (B) das classes de comprimento de crustáceos (mm).

EXERCÍCIOS CAPITULO 1

1. Um restaurante recebeu as seguintes notas de seus clientes: 4,5, 8, 9, 5, 8,5, 6, 5, 9,9,6, 8, 6,3, 9, 4, 7, 7,5. Descreva os resultados obtidos utilizando medidas de tendência central e de dispersão adequadas, se possível.

2. Um técnico visitou doze residências em dois bairros do Recife, anotando o número de larvas de pernilongos existentes nas armadilhas. Descreva as duas populações de dados obtidos utilizando medidas de tendência central e de dispersão adequadas, se possível.

	1	2	3	4	5	6	7	8	9	10	11	12
Bairro 1	59	7	52	89	13	7	99	17	79	25	16	16
Bairro 2	38	1	73	59	25	93	64	44	27	8	38	96

3. Um pesquisador quer saber qual variável de uma população de peixes apresenta menor variação, a massa ou o comprimento total. Faça a análise dos resultados obtidos e apresente sua conclusão.

Massa	957	52	744	37	242	660	659	728	307	773
Comprimento	18	29	31	25	9	42	15	51	40	16

4. Os recém nascidos de um hospital foram medidos com fita métrica e com paquímetro digital. Qual das duas medidas se mostrou mais confiável?

Fita	51	45	50	50	52	46	47	49	53
Paquímetro	51,0	46,6	53,0	49,0	49,9	47,5	47,6	50,1	52,5

5. O crescimento de uma espécie de camarão foi de 53% no estágio náuplio, 32% em zoea e 14% na forma bêntica. Qual o tamanho de um juvenil, que iniciou a fase larval com 13 mm?

6. Faça uma tabela de frequência e um histograma do número de filhotes nascidos a partir do registro de partos de cães de uma clínica veterinária, abaixo.

4	6	4	6	1	6	8	4	5	7	10	5	5	7	1	7	2	9	7	3	3
9	4	4	5	1	4	2	4	6	6	2	5	7	6	8	4	2	3	4	4	5

Capítulo 2

PROBABILIDADE E MODELOS DE DISTRIBUIÇÃO DE PROBABILIDADES

PROBABILIDADE

Existem vários conceitos de probabilidade, mas podemos dizer que é uma medida da incerteza de ocorrer determinado evento, eq. 2.1.

$$P_{(Evento)} = \frac{n_E}{N} \qquad \textbf{eq. 2.1}$$

onde, n_e = número de eventos, N = número de resultados do evento. n_e = um ingresso, N = cinco amigos.

Por exemplo, se quisermos sortear quem vai ficar com um ingresso para o cinema entre cinco amigos. Deduzimos que a probabilidade de qualquer amigo ficar com o ingresso é de 20%.

$$P_{(Evento)} = \frac{1}{5} = 0{,}2 \qquad \text{ou } 20\%$$

A frequência de determinado evento também indica sua probabilidade de ocorrência:

$$F_i = \frac{n_i}{N}$$

onde: n_i = quantas vezes ocorre determinado evento, N = número total de ocorrência de todos os eventos.

A probabilidade de um evento acontecer é p, e varia de zero (0) a um (1).

A probabilidade de um evento não acontecer é:1-p, que podemos chamar de q.

$$q = 1 - p \quad \text{e} \quad p + q = 1$$

A proporção de uma população em uma amostra indica a probabilidade de ela ser amostrada. Por exemplo, se num laboratório de pesquisa há 10 biólogos, 3 farmacêuticos, e 12 bioquímicos. A probabilidade de entrar um biólogo no laboratório, logo após o almoço é de 40%.

$$P_i = \frac{10}{25} = 0,40$$

Soma de probabilidades – quando dois eventos podem ocorrer em determinada situação, mas não simultaneamente, isto é, são mutuamente exclusivos ocorre um **ou** outro. Nós somamos suas probabilidades.

Por exemplo, a probabilidade de encontrarmos uma concha de determinada espécie de molusco com tons amarelos é de 50%, e com tons verdes é de 30 %. A probabilidade de eu coletar **uma** concha e ela ter uma das colorações acima é de 80%

$$P_a + p_b = \frac{x}{n} + \frac{y}{n} = \frac{x+y}{n}$$

$$0,5+0,30 = 0,80 \text{ ou } = \frac{50}{100} + \frac{30}{100} = \frac{80}{100}$$

Olhando a Teoria dos Conjuntos

Exemplo 2: um casal tem seis filhos, três homens e três mulheres: H1, H2, H3, M1, M2, M3, sendo que um menino e duas meninas foram adotados. A probabilidade de um deles ter comido a última fatia de bolo é 1/6 (16,66%) e a probabilidade de ter sido um dos homens é 1/2 (50%).

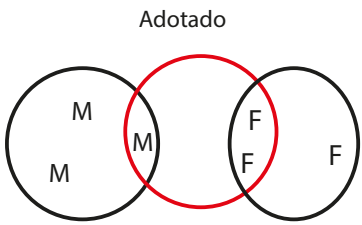
$$P_a + p_b + p_c = \frac{x}{n} + \frac{y}{n} + \frac{z}{n} = \frac{x+y+z}{n} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Soma de eventos com sobreposição

Qual a probabilidade da fatia de bolo ter sido comida por um menino ou por um dos filhos adotados?(Neste caso há intersecção entre os eventos)

Probabilidade de ser menino (3/6) mais probabilidade de ser adotado(3/6) menos a probabilidade de ser menino e adotado (1/6)

$$= \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{3+3-1}{6} = \frac{5}{6}$$



Probabilidade de dois acontecimentos independentes – quando eu quiser coletar mais de uma concha (dois acontecimentos ou mais) a pergunta poderia ser:

Qual a probabilidade de, ao pegar duas conchas, uma ser amarela **e** outra verde? Neste caso, a probabilidade de ocorrência de 'amarelo' **e** 'verde' seria o produto da probabilidade deles.

$$P_a \times p_b = \frac{x}{n} \times \frac{y}{n} = \frac{x \times y}{n^2} \qquad 0{,}50 \times 0{,}30 = 0{,}15 \text{ ou } 15\ \%$$

Am e Am	**0.5 × 0.5=0.25**
Am e V	**0.5 × 0.3=0.15**
Am e Az	**0.5 × 0.2=0.1**

V e Az	**0.3 × 0,2=0,06**
V e V	**0,3 × 0,3=0,09**
Az e Az	**0,2 × 0,2=0,04**

Número de resultados possíveis

O sexo de um filhote de onça pode ser macho ou fêmea. Suponhamos que a onça possua probabilidades iguais de ter olhos negros, castanhos, amarelos e verdes. Para sabermos o número de resultados possíveis de dois ou mais eventos simultâneos, basta multiplicar as possibilidades dos eventos. No exemplo acima são oito resultados possíveis:

Mn, Mc, Ma, Mv, Fn, Fc, Fa e Fv

$2 \times 4 = 8$

Os testes de significância estatística têm por base cálculos de probabilidade. Isto é, coletado um conjunto de dados (amostra), estima-se qual a probabilidade de obter este mesmo conjunto de dados se a hipótese nula é verdadeira.

MODELOS DE DISTRIBUIÇÃO DE PROBABILIDADE DE DADOS

A distribuição dos dados de uma amostra tem dois aspectos: representatividade e aleatoriedade. Quando a distribuição dos dados da amostra é representativa da população, ela indica a probabilidade de ocorrência de cada dado. Essa distribuição dos dados é aleatória se a variável não é influenciada por nenhum fator. Há vários modelos matemáticos de distribuição de probabilidade aleatória dos dados. Os modelos mais comuns para dados discretos são o Binomial e de Poisson. Para dados contínuos, há a distribuição normal.

Distribuição Binomial

A distribuição Binomial foi a primeira distribuição de frequência introduzida na estatística, sendo a mais importante das distribuições de variável aleatória discreta. É uma distribuição resultante do desenvolvimento do binômio

$$(p+q)^n,$$

onde 'n' é o número de eventos considerados, 'p' a probabilidade de um evento ocorrer e 'q' a probabilidade do evento alternativo ocorrer, sendo p + q = 1.

A distribuição Binomial é adequada a eventos binários, isto é, com apenas dois resultados possíveis (macho ou fêmea, sim ou não, ...), Onde as repetições são independentes e a probabilidade dos acontecimentos é constante (eq. 2.2).

$$f_{(x)} = \frac{n!}{x! \times (n-x)!} \times p^x \times q^{n-x} \qquad \textbf{eq. 2.2}$$

Onde: x= número de sucessos variando de 1 a n, p= probabilidade de sucesso em cada repetição, q= probabilidade complementar a p, n = número de repetições.

Lembrando que: 0! = 1 e $x^0 = 1$

A média e a variância são calculadas por: $\mu = n \times p$ e $\sigma^2 = n \times p\ (1-p)$, respectivamente. Entretanto, quando o desvio padrão é tomado em porcentagem, tem-se:

$$\sigma = \sqrt{\frac{pq}{n}} \qquad \textbf{eq. 2.3}$$

A distribuição Binomial é um caso particular da distribuição mais geral, a distribuição polinomial, como casos nos quais somente duas simples alternativas são consideradas, mas, nas quais o evento pode acontecer de várias maneiras com probabilidades p_1, p_2, ...,p_s.

A distribuição Binomial converge para a distribuição normal para grandes valores de 'n'. Para p=1/2 e n=20 há convergência entre as duas distribuições. Ela é usada para encontrar a probabilidade de 'x' números de ocorrências ou sucessos de um evento. Por exemplo: uma cadela está grávida de cinco filhotes. Qual a probabilidade de ela dar à luz a cinco filhotes fêmeas?

Resolução: sabemos que n=5 e p=0,5 (a chance de nascer macho ou fêmea é igual). Para obter a probabilidade de x assumir o valor 5, aplica-se a equação para obter a distribuição Binomial do problema acima devemos calcular todos os valores possíveis de x:

$$P_{(x)} = \frac{n!}{x! \times (n-x)!} \times p^x \times q^{n-x} \rightarrow p_{(x)} = \frac{5!}{5! \times (5-5)!} \times 0,5^5 \times 0,5^0 = 0,03125\ ou\ 3,1\%$$

Para obter a distribuição Binomial do problema acima devemos calcular todos os valores possíveis de x:

Número de fêmeas (x)	Cálculo	P(x)	Possibilidades
0	$p_{(x)} = \frac{5!}{0! \times (5-0)!} \times 0,5^{0} \times 0,5^{5-0}$	0,03125	MMMMM
1	$p_{(x)} = \frac{5!}{1! \times (5-1)!} \times 0,5^{1} \times 0,5^{5-1}$	0,15625	FMMMM MFMMM MMFMM MMMFM MMMMF
2	$p_{(x)} = \frac{5!}{2! \times (5-2)!} \times 0,5^{2} \times 0,5^{5-2}$	0,3125	FFMMM FMFMM MFFMM FMMFM MMFFM FMMMF MMMFF MFMFM MFMMF MMFMF
3	$p_{(x)} = \frac{5!}{3! \times (5-3)!} \times 0,5^{3} \times 0,5^{5-3}$	0,3125	FFFMM FMFFM MFFFM FMMFF MMFFF FMFMF MFMFF MFFMF FFMFM MFFMF
4	$p_{(x)} = \frac{5!}{4! \times (5-4)!} \times 0,5^{4} \times 0,5^{5-4}$	0,15625	FFFFM FFFMF FFMFF FMFFF MFFFF
5	$p_{(x)} = \frac{5!}{5! \times (5-5)!} \times 0,5^{5} \times 0,5^{0}$	0,03125	FFFFF

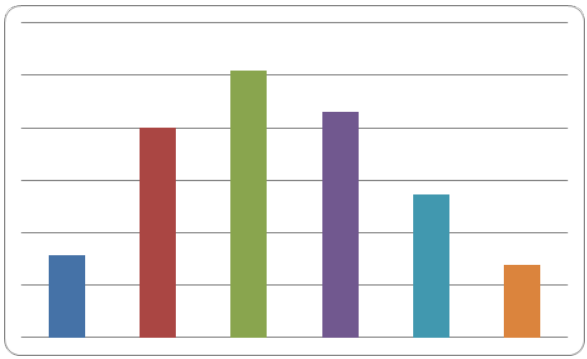


Figura 2.1: Distribuição Binomial, representando a frequência absoluta da probabilidade de nascimento de fêmeas da espécie *Canis familiaris*.

Os coeficientes binomiais podem ser representados na forma de um Triangulo de Pascal, com inúmeras propriedades que auxiliam no cálculo da distribuição Binomial.

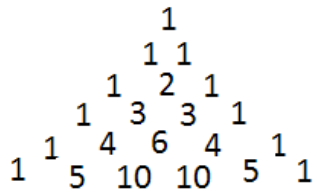


Figura 2.2: Triangulo de Pascal.

Quando utilizamos a distribuição Binomial para resolver um problema?

Quando:

- Existirem somente dois resultados mutuamente exclusivos (ou sim ou não, ou branco ou preto, ...)
- As n tentativas são independentes, e
- A probabilidade de ocorrência ou sucesso, p, permanece constante em cada tentativa.

Distribuição de Poisson

A distribuição de Poisson pode ser considerada como um caso particular de distribuição Binomial, na qual a probabilidade de ocorrência de um acontecimento é muito pequena. Ela é adequada para calcular a probabilidade de ocorrência de uma variável discreta em um intervalo de eventos contínuos ou não, como o número de espécies por área, ou número de mortes por ano.

A distribuição de Poisson é caracterizada pela média de ocorrência por unidade de medida (μ). Neste modelo a variância é igual à média: $\sigma^2 = \mu$

A equação para determinar a probabilidade de x ocorrências é:

$$P_{(x)} = \frac{\left(e^{-u} u^{x}\right)}{x!} \qquad \textbf{eq. 2.4}$$

- A probabilidade de observarmos zero indivíduos na amostra é e^{-u}
- A probabilidade de observarmos um indivíduo na amostra é e^{-u} u

- A probabilidade de observarmos dois indivíduos na amostra é $e^{-u}\ u^2\ /2!$
- A probabilidade de observarmos três indivíduos na amostra é $e^{-u}\ u^3/3!$
- E assim por diante.

A diferença dentre a distribuição Binomial e a de Poisson é que a curva Binomial é determinada pelo tamanho da amostra (n) e pela probabilidade de sucesso do evento (p) enquanto que a de Poisson depende apenas da média de ocorrências (μ).

Exemplo: a densidade média da amazônia é 2,54 habitantes/km^2. Qual a probabilidade de encontrar um habitante/km^2?

$$P_{(x)} = \frac{\left(e^{-u}u^{x}\right)}{x!} = P_{(1)} = \frac{\left(e^{-2,54}2,54^{1}\right)}{1!} \quad = \quad 0,20 \text{ ou } 20\%$$

Para traçarmos a curva de distribuição de habitantes da amazônia, precisamos construir a tabela de frequência:

N	Probabilidade
0	0,079
1	0,200
2	0,254
3	0,215
4	0,137
5	0,069

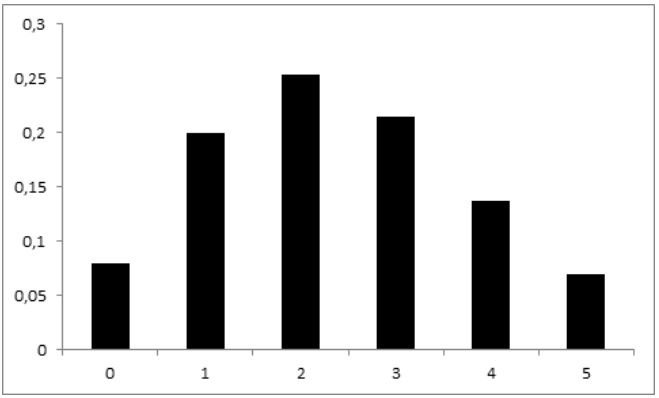


Figura 2.3: Distribuição de Poisson para média=2,54.

Distribuição normal

As estimativas estatísticas partem do pressuposto da aleatoriedade. Isto é, uma amostra aleatória é aquela em que qualquer elemento da população tem a probabilidade de ser selecionado para a amostra. Pressupõe-se que variáveis (e.g., Temperatura, altura, biomassa, oxigênio dissolvido) que não estão sofrendo influência de nenhuma outra variável qualquer, apresentam distribuição aleatória em torno da média. Existem vários modelos de distribuição aleatória, o modelo de distribuição normal é adequado para variáveis contínuas.

A distribuição normal forma uma curva simétrica, podendo ser representada pela curva de Gauss, que tem a forma de sino. A distribuição normal é caracterizada pela média (μ) e desvio padrão (σ), sendo definida pela seguinte equação:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{onde} \quad -\infty < x < \infty \qquad \textbf{eq. 2.5}$$

Onde f(x) é Y, a ordenada, x é a variável independente, 'e' é a constante de Euler (2,718...).

Características da curva normal

- A distribuição normal tem forma de sino, com caudas assintóticas ao eixo x. Isto é, jamais toca o eixo x.
- A distribuição é simétrica, com a média no centro da distribuição.
- Valores de x vão de -∞ a +∞.
- A média, a moda e a mediana são coincidentes. Assim os valores da variável x à esquerda da distribuição normal ocorrem com frequência igual aos que ocorrem à direita, pois a curva é simétrica.
- A curva tem dois pontos de inflexão, que correspondem a valores de um desvio padrão de x, isto é: (μ – σ) e (μ + σ).
- A área total da distribuição soma 1 ou 100%.
- Aproximadamente 68% dos valores da curva situam-se entre (μ – σ) e (μ + σ).
- Aproximadamente 95% dos valores situam-se entre (μ – 1,96σ) e (μ + 1,96σ).

- Aproximadamente 99,7% dos valores situam-se entre (μ – 3σ) e (μ + 3σ).

Utilidades da curva normal

Conhecendo as características da curva normal e a média e desvio da população, podemos calcular a probabilidade de valores de interesse, como saber a probabilidade de certo valor ocorrer, qual proporção da população é menor do que certa medida. Para calcular esta probabilidade utilizamos a curva normal padronizada.

Curva normal padronizada

A distribuição normal (variável x) pode ser transformada na curva normal padrão ou reduzida, que possui média igual a zero e desvio padrão igual a 1, representada pela variável z. A transformação é dada pela seguinte equação:

$$z = \frac{x - \mu}{\sigma} \qquad \textbf{eq. 2.6}$$

O valor de z pode ser convertido em probabilidade, ou área correspondente, na curva normal padrão, apresentada em uma tabela, o valor correspondente à variável 'x' na curva padrão (z). A figura abaixo (Fig. 2.4) Mostra a distribuição de z (número de desvios-padrão), também chamada de distribuição normal padrão.

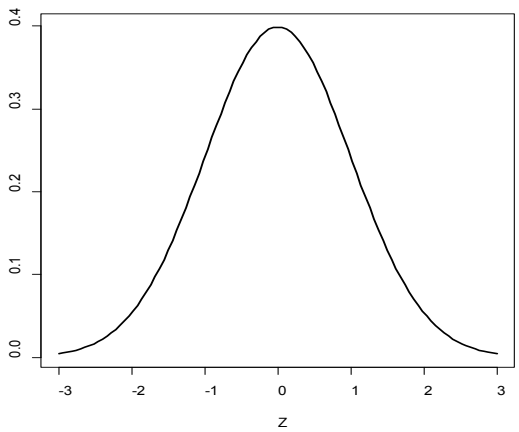


Figura 2.4: Curva normal padrão.

A área correspondente a 95% da área total sob a curva normal está compreendida no intervalo μ ± 1,96σ (Fig. 2.5)

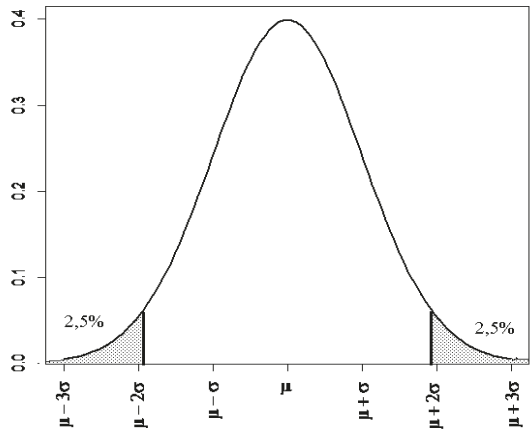


Figura 2.5: Curva normal

A equação da curva z é:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z^2}{2}} \qquad \textbf{eq. 2.7}$$

A tabela z determina a área a partir do número de desvios-padrão, sendo adimensional. Os valores de z negativos podem ser obtidos por simetria. Os valores de z possibilitam determinar a área sob a curva. Podemos saber, por exemplo, que numa população de guaiamum, com média de 65 mm de comprimento e desvio padrão de 5 mm, a probabilidade de um animal ter entre 65 e 70 mm é 0,34.

Para x = 70

$$z = \frac{70 - 65}{5} = 1$$

O valor de z=1 equivale à área de 0,3413, que pode ser visualizado na tabela z. Isto é 34%

Outro exemplo:

Encontrei um peixe no supermercado muito grande para a espécie, 60 cm. Sabendo-se que a média populacional é de 45 cm e o desvio padrão de 5 cm. Qual proporção da população tem comprimento igual ou superior a este?

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 45}{5} = 3$$

Na tabela 1 encontrei o valor de 0,4987. Este valor, entretanto, corresponde à área igual e maior que a média, devemos somar mais 50% encontrando o resultado de 0,9987. A área complementar corresponde à % buscada, ou seja 0,13%.

Teorema do limite central

À medida que o tamanho da amostra aumenta, a distribuição amostral da sua média aproxima-se cada vez mais de uma distribuição normal.

Distribuição amostral das médias

Quando trabalhamos com dados quantitativos, a média e o desvio padrão são medidas importantes para analisarmos os dados amostrais. A distribuição amostral das médias, quando representativa, apresenta características não significativamente diferentes da população.

Se a variável tem distribuição aleatória, podemos analisá-la como uma distribuição normal. Isto é, analisar a amostra em relação a uma população 'normal'. O erro padrão da média amostral ($\sigma(\bar{x})$) pode ser obtido por:

$$\sigma(\bar{x}) = \sqrt{\frac{\sum f(\bar{x} - \mu)^2}{\sum f}} \qquad \textbf{eq. 2.8}$$

Onde *f* é a frequência em que a média ocorreu, ou

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}} \qquad \textbf{eq. 2.9}$$

onde 'n' é o tamanho da amostra, $\sigma_{\bar{x}}$ é o desvio padrão da média ou erro padrão. Podemos utilizar esta medida para calcular a probabilidade de uma amostra aleatória ter uma média diferente da média da população. Para isto utilizamos a relação com z da equação acima.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad \textbf{eq. 2.10}$$

A distribuição amostral das médias é utilizada para calcular a probabilidade de uma amostra com determinada média pertencer à população conhecida

Exemplo:

Qual a probabilidade de uma amostra com 13 homens ter a altura média de 1,70 m, sabendo-se que a altura média dos brasileiros é 1,73 m (D.P. = 0,15M) para homens?

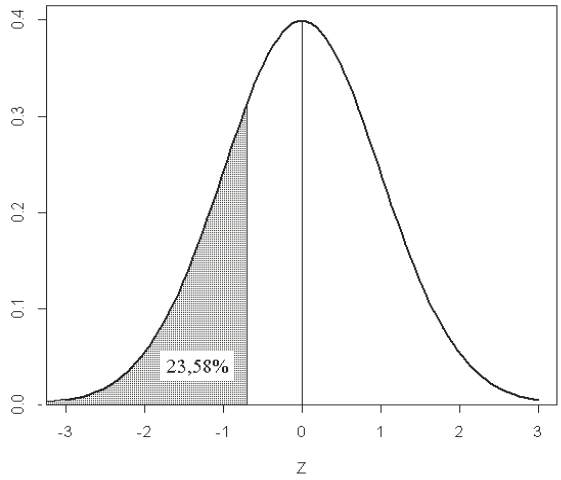
1	2	3	4	5	6	7	8	9	10	11	12	13	Média
1,65	1,68	1,73	1,87	1,63	1,75	1,67	1,68	1,67	1,7	1,64	1,71	1,72	1,7

$$z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = z_{calc} = \frac{1,70 - 1,73}{\frac{0,15}{\sqrt{13}}} = -0,7211$$

O valor de z=0,72 corresponde à probabilidade de 0,2642 (Tabela 1), o que corresponde à área entre zero (média) e o valor analisado. Entretanto, buscamos a área complementar

$$P(= 1,70) = P(Z = 0,72) = 0,5 - 0,2642 = 0,2358$$

A probabilidade de uma amostra com 13 homens ter altura média de 1,70 m é de 23,6%.



Exemplo 2: utilizando os dados populacionais acima, sabemos que no nosso curso há 200 homens, quantos deles teriam altura igual ou superior a 1,80m?

$$z_{calc} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow z_{calc} = \frac{1,80 - 1,72}{\frac{0,15}{\sqrt{13}}} = 1,923$$

$$P(\overline{x} > 1,80) = P(Z > 1,923) = 0,5 - 0,4726 = 0,0274$$

O número de homens com mais de 1,80m no curso é = 0,0274 × 200 = 5,48 homens

Como é difícil determinar $\sigma_{\overline{x}}$, que é um parâmetro populacional o melhor é fazer a estimativa a partir de amostras aleatórias:

$$s_{\overline{x}}^2 = \frac{s^2}{n}$$

EXERCÍCIOS CAPÍTULO 2

1) Considerando que a altura média da mulher brasileira é de 1,61 m e o desvio padrão (D.P.) de 0,05 m:

A. mulheres entre 1,55 e 1,6 m representam qual proporção da população?
B. qual a proporção da população feminina é mais baixa que você?
C. qual proporção da população feminina tem 1,78 m e acima?

2) O Pastor Alemão tem em média cinco filhotes por gravidez, com desvio padrão de um filhote. Qual é a probabilidade de ele ter dois filhotes? E qual a probabilidade de ter sete?

3) Uma variedade de gatos pode gerar filhotes dos dois sexos: machos(M) ou fêmeas (F) e três variedades de cores: branco (B), preto (P) ou malhado (ML). Quais os resultados possíveis para um filhote? Quais são eles?

4) Um determinado grupo apresenta quatro mulheres com sangue A+, Três com sangue AB+, oito com O-, 20 com O+, duas com B- e uma com A-, quinze homens com o+, sete com O-, três com B+, dois com B-, 2 com A+ e um com A-.

A. qual a probabilidade de selecionar aleatoriamente uma mulher com sangue b-?
B. uma mulher se o sangue for o?
C. alguém com sangue a?

5) Suponha que a probabilidade de certo casal de ter filhos de olhos azuis é 25%. Qual a probabilidade deles terem três filhos com olhos azuis?

6) Uma espécie de Tardigrada ocorre em 3% das amostras de determinada praia. Dentre 70 amostras, qual a probabilidade de encontrar tardigradas em 3 amostras?

7) Os bombeiros recebem, em média, 5 chamadas para socorro de animais silvestres por hora. Qual a probabilidade de eles receberem 3 chamadas em uma hora?

Capítulo 3

TESTES DE HIPÓTESES

Pesquisas científicas são realizadas com base em hipóteses, que serão analisadas e testadas através do método científico. Em estatística, a hipótese está relacionada ao parâmetro que estamos analisando (e.g.,.testar a média entre duas amostras), ou conjunto de parâmetros, conforme o caso. Ela é uma afirmação sobre o parâmetro estudado. A inferência estatística envolve representatividade e confiabilidade, pois geralmente trabalhamos com um conjunto pequeno da população estudada.

As hipóteses correspondem ao ramo da estatística inferencial, que fornecem métodos para o pesquisador formular a idéia a ser testada, tomar decisões e saber o erro associado a cada uma. O teste de hipóteses é uma regra de decisão, ele é binário e excludente: ou H_0 ou H_1.

As hipóteses estatísticas sempre comparam dois ou mais conjuntos de dados, sejam elas uma amostra com dados da população, uma amostra com um referencial aleatório ou de igualdade, duas amostras, várias amostras :

A. Hipótese nula ou de nulidade (H_0): estabelece ausência de diferenças entre os parâmetros (em outros termos é o mesmo que dizer que os parâmetros são iguais), hipótese que remete a aleatoriedade dos eventos (e.g.: o tamanho médio da população de peixes é aleatoriamente determinado em reservatórios preservados ou poluídos).
B. Hipótese alternativa (HA ou H_1): é a hipótese contrária à hipótese nula, ou seja, que os parâmetros são diferentes entre si, remete a eventos biológicos ou naturais, não aleatórios (e.g.: o tamanho médio da população de peixes de reservatórios preservados é diferente do tamanho médio de reservatórios poluídos).

O teste de hipóteses pode ser bilateral: H_0: $\bar{x} = \mu$ e H_1: $\bar{x} \neq \mu$, Fig. 3.1A
Ou unilateral: H_0: $\bar{x} \leq \mu$ e H_1: $\bar{x} > \mu$, Fig 3.1 B e C.

A escolha do teste, bilateral ou unilateral, é decidida na hora de definir a pergunta do teste. Por exemplo, se você está testando se um remédio para depressão afeta a pressão arterial dos pacientes. Você não sabe se esta possível influencia aumenta ou diminui a pressão. Sua hipótese, neste caso, é bilateral: afeta (≠) ou não afeta (=). Em outro exemplo, você está testando uma droga para diminuir o colesterol no sangue. Neste caso, a droga afetará significativamente, se diminuir o colesterol, a hipótese é unilateral, '≥' como H_0 ou '<' como H_1.

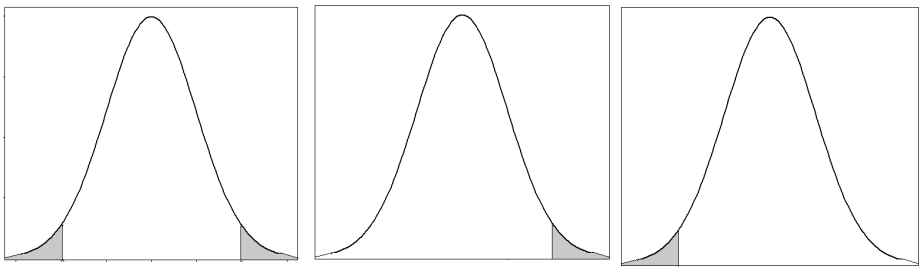


Figura 3.1: Curvas com a representação do nível de significância para hipóteses bilateral, A, e unilateral, B, e C.

Erros: tipo I e tipo II

A escolha da hipótese H_0 ou H_1 implica em um erro, relacionado à decisão em si. O erro do tipo I, ou erro alfa (α), é a rejeição da hipótese nula (H_0) quando ela for verdadeira, chamamos de nível de significância do teste (α). Por outro lado, quando não rejeitamos H_0 quando ela é falsa, chama-se erro do tipo II ou erro beta (β). A força de um teste estatístico é definida por 1-β. Não há relação matemática entre erros alfa e beta.

	Conclusão do teste	
Verdade	**Não se rejeita H_o**	**Rejeita-se H_o**
H_o é verdadeira	Decisão correta P = 1-α	Decisão errada, Erro do tipo I P=α
H_o é falsa	Decisão errada Erro tipo II p = β	Decisão correta p =1-β (Força)

Como os resultados estatísticos têm erros implícitos, é necessário informar o nível de significância. Costuma-se informar, também, o tamanho

amostral, o teste utilizado e a probabilidade da estatística (probabilidade do resultado ter ocorrido aleatoriamente, probabilidade do evento ocorrer se$H_0$for verdadeiro). Não esqueça que o alfa (α) é definido antes de realizarmos o teste.

Os testes estatísticos apresentam sensibilidade e especificidade. A sensibilidade está relacionada à capacidade de se detectar um efeito, está relacionado com o positivo verdadeiro, quando o teste dá positivo e a hipótese nula é rejeitada. Especificidade está relacionada com o negativo verdadeiro, quando o teste dá negativo e a hipótese nula não é rejeitada

		Teste	
		+	**–**
Verdade	**Ocorre o efeito** **+**	++ Rejeita-se Ho Positivo verdadeiro	+– Falso negativo
	Não ocorre **–**	–+ Falso positivo	**––** Não rejeita Ho Negativo verdadeiro

O poder estatístico (1-ß) é a probabilidade de rejeitar H_0 quando H_1 é verdadeira. Isto é, qual a probabilidade de detectar um efeito, supondo que ele existe. Quanto maior este poder, menor a chance do erro do tipo ii. De modo geral, uma probabilidade de 0,2 de identificar um efeito verdadeiro. Neste caso o poder de análise seria 0,8.

Ao realizar um estudo uma ação importante é estimar o número de amostras necessárias para o mesmo. Para evitar esforço e custos desnecessários. Existem vários métodos para estimar o número de amostras necessários a determinado estudo. O poder de análise deve ser realizado antes da coleta dos dados para estimar o número de amostras necessário para detectar o efeito se quer observar com certo grau de confiança. Isto é, para calcular qual o menor número de amostras necessário para detectar determinado efeito. Por exemplo, quantas amostras são necessárias para verificar que a média de uma característica (*e.g.*, Comprimento, temperatura, biomassa, riqueza, pressão) é diferente. Esta análise também pode ser usada para saber qual o efeito detectado se usarmos determinado tamanho amostral. Também serve para comparar o poder entre dois testes estatísticos.

AMOSTRAGEM

Os resultados de um estudo dependem, fundamentalmente, de um desenho amostral ou experimental bem feito. Caso contrário, teremos dificuldades em analisar e interpretar os resultados, podendo perder todo o esforço realizado, devido aos pressupostos das análises estatísticas. Todo estudo começa pelo planejamento amostral, quando surgem perguntas como:

- Qual a hipótese ou pergunta principal do estudo?
- Os dados virão de experimentos manipulativos ou naturais?
- Quais variáveis explicativas e preditoras vou utilizar no estudo?
- A hipótese envolve variações espaciais e ou temporais? Em que escala?
- Os dados serão coletados de forma independente?
- Qual será o esforço amostral (escala, réplicas e replicatas, número de amostras).

Qual hipótese?

A definição de uma hipótese é imprescindível para definirmos quais dados serão coletados e qual a resposta esperada. Se quisermos saber, por exemplo, se uma planta floresce uma vez ou mais por ano, a hipótese poderia ser: a planta floresce apenas uma vez no ano. Ao tentar responder à pergunta traçamos um plano de observação da realidade e escolhemos um método para processamento e análise de dados. Este plano estratégico de observação da realidade pode ser chamado de desenho amostral.

Nosso estudo será realizado no ambiente natural ou em forma de experimento?

Na forma de experimento teremos a vantagem de isolar os fatores (*e.g.* Luminosidade, temperatura, umidade) que podem influenciar a variável em estudo, que é a floração, no exemplo acima. Isto é, analisaremos os mecanismos de causa e efeito. Em ambientes naturais, os dados obtidos, foram influenciados pelo conjunto de fatores que atuou sobre a variável em estudo. Nestes estudos, a influência de cada fator é mais difícil de avaliar. Contudo, é possível escolher ambientes que possibilitem a análise de certas variáveis independentes (fatores) que queremos analisar, como variação de luminosidade, umidade ou de certos nutrientes. Vários estudos de impactos antrópicos têm sido realizados como experimentos naturais em campo. Os estudos de impacto ambiental, com coletas antes e depois do impacto (BACI em inglês) são um exemplo, assim como casos de

eutrofização, poluição de córregos, entre outros. Assim, podemos elaborar ambientes comparativos e produzir robustas conclusões sobre a intensidade e a natureza das relações entre as variáveis. A poderosa estrutura do plano de observação desse estudo aproxima-o da lógica do delineamento experimental. Se decidirmos pelo ambiente natural, precisaremos definir a população estudada. As populações não formam distribuições discretas no ambiente, isso ocorre principalmente devido as complexas interações ecológicas e aos diferentes nichos de cada espécie. Isto deve ser levado em conta, quando planejar o desenho amostral.

Quais variáveis explicativas ou preditoras vou utilizar no estudo?

A variável explicativa ou preditora é aquela que afeta a variável que estou testando (variável resposta ou dependente). As variáveis preditoras ajudam a explicar as causas do evento. No caso da produção de flores em uma determinada planta, as variáveis preditoras poderiam ser: temperatura, umidade, nutrientes no solo, presença de outras plantas ou polinizadores, luminosidade.

A hipótese envolve variações espaciais e ou temporais? Em que escala?

Existem várias escalas espaciais: local, regional, global, assim como várias escalas temporais: instantânea, diária, semanal, mensal, anual, decenial. Experimentos espaciais indicam a intensidade de resposta. Isto é, haverá respostas com pequenas variações ou não. Por outro lado, as escalas temporais indicam o tempo do processo (reprodução no caso) ou a resiliência do sistema (se for uma perturbação, por exemplo). A definição da escala (área e ou período de estudo) também precisa da definição da unidade amostral e do tamanho da amostra

As amostras serão independentes?

Os dados que estou amostrando, ocorrem em pares (esquerdo direito), resultam em mudanças na amostra (antes e depois).

Qual será o esforço amostral (réplicas e replicatas, número de amostras)?

O número de réplicas é necessário para diferenciarmos um evento ao acaso de um padrão. Quantas plantas serão necessárias para saber se este é o padrão ou foi ao acaso? Quantos anos serão necessários?

Quantas replicações?

MÉTODOS DE AMOSTRAGEM

A amostragem precisa ser representativa e imparcial. Ela consiste em obter um número de amostras, através de um método adequado e que os dados sejam obtidos de modo aleatório, isto é, toda a população tem a mesma chance de ser coletada e a coleta de um dado não influencia a coleta do próximo. Ela faz parte da inferência estatística, que é o processo de obter informação sobre uma população a partir de resultados observados na amostra.

Tipos de amostragem

A amostragem pode ser não probabilística ou probabilística. Os métodos de amostragem não probabilística podem ser aleatórios ou não. Os métodos não aleatórios têm por base informações a priori da população. Há três categorias para estes métodos:

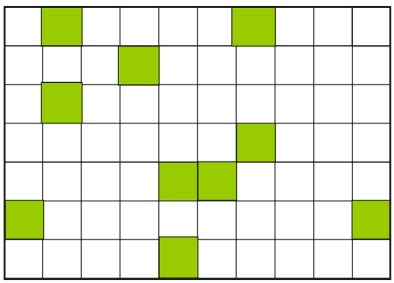
- Esmo
- Intencional
- Cotas

A amostragem probabilística pode ser classificada em aleatória, sistemática e estratificada.

- Amostragem aleatória: cada elemento tem a mesma probabilidade de ser amostrado. A amostragem aleatória pode ser usada se o objetivo é estimar a densidade média populacional e o número de amostras não é tão grande (<100). A amostragem aleatória pode ser: -com reposição, quando a mesma amostra pode ser analisada mais de uma vez e, -sem reposição, quando cada amostra é coletada apenas uma vez.

Exemplo: preciso sortear dez amostras de um universo de 70:

Números aleatórios: 2, 7, 14, 60, 45, 46, 51, 22, 37, 65



- Amostragem sistemática: é uma amostragem através de intervalos constantes, que são definidos *a priori* por sorteio (número aleatórios) ou por equações. A amostragem sistemática tem uma vantagem sobre a amostragem aleatória quando o número de amostras é grande, porque cobre de modo mais abrangente a área total de amostragem. Isto é especialmente importante para fazer mapas populacionais.

Sendo N, o tamanho da população e n, o tamanho da amostra desejado, define-se a quantidade N/n=K, chamado intervalo de amostragem.

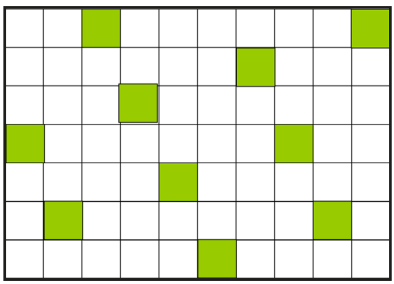
Faz-se um sorteio dos números de 1 a K, obtendo-se o número da primeira amostra, as demais serão calculadas através da progressão geométrica (eq.3.1):

$$a_n = a_1 + (i-1)K \qquad \textbf{eq. 3.1}$$

Exemplo: tenho 70 pés de alface e pretendo coletar dez amostras de modo sistemático:

N/n= 7

Sorteio entre 1 e 7 → 3, $a_2 = 3 + (2\text{-}1)\ 7 = 10$, $a_3 = 3 + (3\text{-}1)\ 7 = 17$.



- Amostragem estratificada: quando a população é heterogênea (*e.g.*, idade, estágio de desenvolvimento, grau de contaminação) ou o ambiente tem um gradiente forte (*e.g.*, umidade, salinidade). Faz-se a amostragem aleatória dentro de cada estrato.

Exemplo: quero entrevistar 10% dos alunos de bacharelado em biologia de determinada universidade. São 197 alunos

Ano do curso	Número de alunos	Tamanho da amostra
1	60	6
2	38	4
3	37	4
4	31	3
>4	31	3
Total	**197**	**20**

Fatores que determinam o tamanho da amostra

1. Nível de confiança (quanto maior o tamanho da amostra, maior o nível de confiança),
2. Erro máximo permitido (quanto menor o erro permitido, maior o tamanho da amostra),
3. Variabilidade do fenômeno que está sendo investigado (quanto maior a variabilidade, maior o tamanho da amostra).

Toda amostragem inclui um erro, denominado de **erro amostral,** que consiste na diferença entre o resultado amostral e o verdadeiro resultado populacional, sendo inversamente proporcional ao tamanho da amostra.

ESCOLHA DO TESTE ESTATÍSTICO ADEQUADO

Para decidir pelo teste estatístico mais adequado, alguns critérios devem ser levados em conta, como:

A. O poder do teste
B. A aplicabilidade do modelo estatístico, sobre o qual se baseia o teste, aos dados da pesquisa.
C. Nível de mensuração atingido na pesquisa.

O poder de uma análise estatística é, em parte, função do teste estatístico empregado na análise. Um teste estatístico pode ser considerado bom se tem pequena probabilidade de rejeitar H_0 quando este é verdadeiro, e grande probabilidade de rejeitar H_0 quando este é falso.

Existem outros fatores, além do poder, a serem levados em conta na escolha do teste estatístico: maneira como a amostra de valores foi extraída,

a natureza da população da qual se extraiu a amostra e o tipo de mensuração ou escala das variáveis envolvidas.

Modelo estatístico

Todo teste inferencial usa um modelo estatístico, determinado pelo tipo de dados e pelo modo de amostragem.

As condições do modelo estatístico de um teste são frequentemente chamadas de "suposições" do teste. É óbvio que quanto menos suposições ou mais fracas estas forem para definir um modelo, menos qualificações precisaremos impor à decisão a que tivermos chegado pelo teste estatístico. Isto é, quanto mais fracas forem as suposições, mais gerais serão as conclusões.

Todavia, os testes de maior poder são precisamente aqueles que comportam as suposições mais fortes ou mais amplas.

Entretanto, os dados da pesquisa devem ser adequados ao teste. Por exemplo, as condições que devem ser satisfeitas (pressupostos) para o teste t são as seguintes:

1. As observações devem ser independentes. A escolha de um elemento não afeta a escolha de outro.
2. As observações devem ser extraídas de populações com distribuição normal.
3. As populações devem ter a mesma variância (homogeneidade da variância)
4. As variáveis devem ser medidas pelo menos em escala intervalar, para que seja possível utilizar as operações aritméticas.

No caso da análise de variância há ainda outra condição:

5. Os efeitos devem ser aditivos.

Todas as condições acima (exceto a 4, que define as condições de mensuração) são elementos do modelo estatístico paramétrico.

Teste estatístico paramétrico

É o teste que utiliza os parâmetros da distribuição normal, isto é: a média e desvio padrão. Estes modelos especificam certas condições sobre os parâmetros da população da qual se extraiu a amostra para pesquisa, como normalidade dos dados e homogeneidade da variância.

Teste estatístico não paramétrico

É uma prova cujo modelo não especifica condições sobre os parâmetros da população da qual se extraiu a amostra. Há certas suposições básicas associadas à maioria das provas não-paramétricas, a maior parte das provas não-paramétricas se aplica a dados em escala ordinal e alguns aceitam dados em escala nominal.

Afirma-se que o teste estatístico paramétrico é mais poderoso quando todas as suposições de seu modelo estatístico são satisfeitas e quando as variáveis em estudo foram medidas pelo menos em uma escala de intervalos. Todavia, mesmo quando satisfeitas todas as suposições da prova paramétrica e as exigências sobre o nível de mensuração, sabemos que, segundo o conceito de poder-eficiência, aumentando convenientemente o tamanho da amostra, podemos utilizar uma prova não-paramétrica em substituição a uma paramétrica, mantendo ainda assim o mesmo poder na rejeição de H_0.

Como o poder de qualquer prova não-paramétrica pode ser elevado simplesmente aumentando-se o tamanho N da amostra, as provas estatísticas não-paramétricas desempenham papel cada vez mais destacado na análise dos dados.

EXERCÍCIOS CAPÍTULO 3

1) O que é uma regra de decisão? Quais erros estão envolvidos na decisão de um teste de hipóteses?

2) Porque é conveniente um erro alfa baixo (α=0,05, por exemplo)?

3) O que devo decidir em relação a Ho se o teste apresentar p=0,12?

Capítulo 4

TESTES PARA UMA AMOSTRA

TESTES PARA UMA AMOSTRA

Dados quantitativos

Os testes de hipóteses para uma amostra de dados quantitativos **são utilizados para** testar se há diferenças significativas entre a amostra e a população conhecida.

Teste Z

"Para realizar o teste Z, os parâmetros média (μ) e desvio padrão (σ) populacionais devem ser conhecidos."

Exemplo: o comprimento dos bicos de uma amostra de beija-flores, apresenta média de 58,65 mm. Para saber se o comprimento médio observado é similar ao comprimento médio populacional, já conhecido, de 65 mm, e desvio padrão de 10 mm, realiza-se o teste Z.

H_0: o comprimento médio do bico do beija-flor é similar ao comprimento médio populacional $\bar{x} = \mu$

H_A: o comprimento é diferente $\bar{x} \neq \mu$

1. Primeiro tenho que conhecer as características dos meus dados: O parâmetro bico possui distribuição normal para a população? Caso afirmativo (posso simplesmente supor isto ou inferir a partir do conjunto da população), a minha amostra também deve ter distribuição normal (entretanto, geralmente são necessárias amostras grandes para se observar isto)
2. A diferença entre o tamanho do bico da amostra e da população é aceitável? Isto é, o desvio padrão das médias é aceitável? O desvio padrão das médias ou erro das médias é calculado pela equação 2.9:

Para responder a pergunta acima devemos estabelecer a significância deste desvio. Isto é, qual a probabilidade da hipótese nula não ser verdadeira? Esta probabilidade é chamada de nível de significância (α). Um critério razoável é supor que os dados discrepantes são raros (digamos menor que 5%), assim o intervalo de desvios não significativos ou intervalo de confiança será de 95%.

3. Escolher o teste: já sabemos que a população de beija-flores possui o comprimento dos bicos em distribuição normal e conhecemos o desvio padrão populacional. Neste caso, posso escolher o teste Z (curva normal). A hipótese deste exemplo é bilateral, consultando a tabela Z, verificamos que Z=1,96, ou seja:(0,5-0,475)= 0,025.

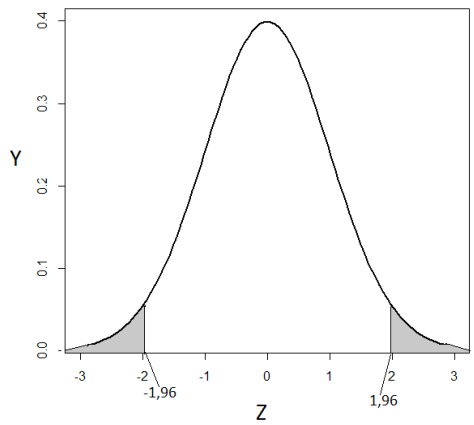


Figura 4.1: Curvas de distribuição normal padronizada, com o valor de z indicando 95% da distribuição dos dados (bilateral).

Dados:

59	56	46	61	57	65	49	60	66	59
69	56	64	60	59	66	58	59	51	53

População: μ = 65 mm, σ= 10, amostra: $\bar{x}$ = 58,65 mm, n = 20 indivíduos

Erro padrão:

$$\sigma(\bar{x}) = \frac{10}{\sqrt{20}} = 2,236$$

Nível de significância α=0,05, $z_{0,05}$= 1,96

$$z_{calc} = \frac{\bar{x} - \mu}{\sigma(\bar{x})} = \frac{58,65 - 65}{2,23} = -2,84$$

Como o $|z_{calc}|$ é maior que o $z_{0,05}$, rejeitamos H_0.

Testes de hipóteses unilaterais

A maioria dos testes de hipóteses envolvendo médias é bilateral, com "H_0=" e "H_1≠". Em algumas situações, entretanto, será necessário saber apenas se é menor ou maior que a média: "H_0≤", "H_1>" ou "H_0≥", "H_1<"

Por exemplo:

Foi acrescentada uma porção reforçada de ração para peixes em determinado cultivo. Em uma população normal, peixes com 5 meses de idade têm 350 g de peso e desvio padrão de 55 g. Uma amostra com 30 peixes apresentou uma média de 385 g. Pode-se afirmar que a dieta causa um aumento significativo de peso, ou este aumento é atribuído ao acaso?

H_0: $\mu_{dietacom\ ração} \leq \mu$

H_1 $\mu_{dietacom\ ração} > \mu$

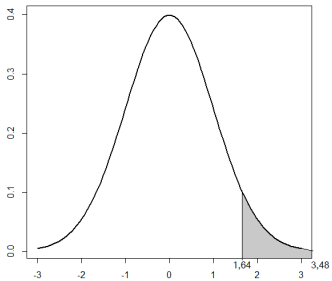
Nível de significância: α= 0,05

$Z_{0,05,\ unilateral}$= 1,64

Z calculado:

$$z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{385 - 350}{\frac{55}{\sqrt{30}}} = 3,48$$

Decisão: a ração reforçada apresenta aumento de massa significativo



Limites de confiança

Para testes bilaterais é:

$$\overline{x} \pm Z_{\alpha/2}\sigma_{\overline{x}}$$

Para testes unilaterais

$$\overline{x} + Z_{\alpha}\sigma_{\overline{x}} \text{ ou } \overline{x} - Z_{\alpha}\sigma_{\overline{x}}$$

TESTES PARA UMA AMOSTRA SEM CONHECIMENTO DO DESVIO PADRÃO POPULACIONAL

Teste t de Student

O '*t* de student' é um teste de hipóteses que utiliza a distribuição *t* de student. O valor de *t* é a medida do desvio entre a média x, estimada a partir de uma amostra aleatória de tamanho n e a média μ da população, usando o erro da média (s_x) ou erro padrão (E.P.) No lugar do desvio padrão:

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad \textbf{eq. 4.1}$$

$$s_{\overline{x}} = EP = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

O teste 't' difere de 'z' por envolver o desvio padrão amostral (s) e não o populacional (σ). A discordância entre as curvas 'z' e 't' é decorrente da diferença entre usar o desvio padrão populacional (σ) e o amostral (s), no cálculo do erro padrão. A diferença entre 'σ' e 's' depende do tamanho da amostra (n), em amostras de grande tamanho a distribuição 't' é praticamente igual à distribuição normal, com média zero (μ=o) e desvio padrão igual a 1 (σ=1). Em amostras com 'n' maior que 120, pode-se considerar a distribuição t igual à distribuição normal.

A distribuição 't' tem diferentes formas, conforme o grau de liberdade da mesma

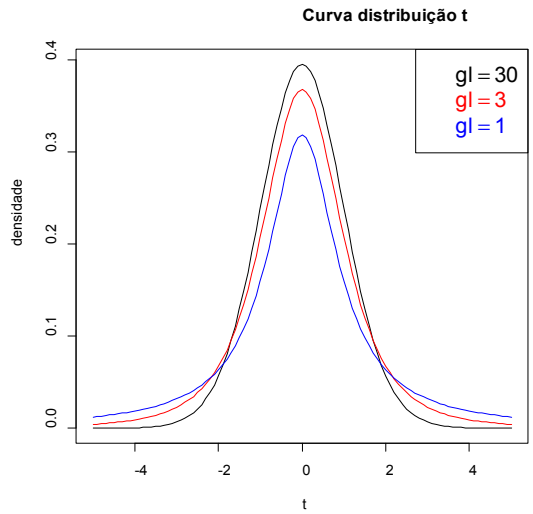


Figura 4.2: Curvas da distribuição t de Student, com os graus de liberdade 1,3 e 30.

Teste t para uma amostra (média com σ desconhecido)

Exemplo:

No mercado de peixes encontrei várias tilápias, sabendo que o tamanho mínimo para venda deste peixe é 38 cm, o tamanho médio encontrado foi de 34 cm, com desvio padrão de 8 cm para 15 peixes. Posso considerar esta amostra dentro do padrão esperado pela norma de pesca?

Tilápias	23	43	22	23	40	39	26	37	42	26	39	37	30	44	39

Solução:

A. $H_0 \rightarrow \mu_0 = 38$

$H_1 \rightarrow \mu_1 \neq 38$

B. $\alpha = 0{,}05$

C. $t_{crit,\ 0{,}05,\ 14} = 2{,}145$ (Tabela teste t)

D.

$$\overline{x} = \frac{23+43+22+\ldots+39}{15} = 34 \quad , \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = 64,\ s = 8$$

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{34 - 38}{\frac{8}{\sqrt{15}}} = -1{,}94$$

Resposta: não rejeito H_0, pois $t_{calc} = |-1{,}94| < t_{crit} = 2{,}145$ a amostra não apresenta média significativamente diferente do que o estipulado pela empresa.

Estimando a média populacional (μ) quando se desconhece o σ.

Pode-se estimar a média populacional através do intervalo de confiança para uma média populacional, que é expresso como:

$$\mu \pm t \times s_{\overline{x}}$$

$$\overline{x} \pm t \times s_{\overline{x}} \rightarrow s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{15}} = 2{,}066$$

No exemplo acima, o valor de 't' é para um intervalo de confiança de 95%, resultando nos valores abaixo:

$$38 \pm 2{,}145 \times 1{,}032\ 38 \pm 3{,}178$$

Se o autor deseja apenas mostrar a variabilidade da população amostrada, Pode apresentar os dados com o desvio padrão e 'n' amostral local,

$$\overline{x} \pm \text{D.P} \rightarrow 38 \pm 8,\ n = 15$$

Se, por outro lado, o autor quiser mostrar a precisão da estimativa da média populacional, ele deve utilizar o erro padrão local,

$$\bar{x} \pm \text{E.P.} \rightarrow \quad 38 \pm 2{,}066,\ n = 15$$

Pressupostos para a realização do teste t:

O teste t é um teste paramétrico, pois utiliza a média e a variância na sua análise. As amostras devem ser aleatórias e independentes. Para isto, devemos supor que as médias possuam distribuição normal. O teste t é robusto em relação à normalidade, isto é, mesmo com desvios consideráveis da normalidade os resultados são confiáveis, desde que as amostras sejam de tamanho igual, e o teste seja bilateral. Entretanto, pode-se transformar os dados para que eles apresentem distribuição normal ($\sqrt[4]{}$, $\sqrt[2]{}$, ln(x+1)).

A outra premissa é a homogeneidade das variâncias. Se não houver homogeneidade entre as variâncias a serem testadas, o valor do nível de significância do teste se altera, tornando imprescindível testar a homogeneidade antes do teste t.

Tamanho de uma amostra para teste t e estimativa de média populacional.

O tamanho da amostra está relacionado com a precisão que queremos, ou com o erro que estamos dispostos a aceitar. Podemos responder a essas questões através do conceito de intervalo de confiança (IC), onde o dado de IC representa metade do intervalo de confiança. Quanto maior o 'n' amostral menor o erro e menor o intervalo de confiança. O tamanho amostral depende:

1. Do erro aceitável
2. Da variabilidade da população

$$n = \frac{s^2 t^2_{\alpha(2);g.l.}}{IC^2} \qquad \textbf{eq. 4.2}$$

Por exemplo: utilizando o exemplo de peixes acima, qual o 'n' amostral necessário para estimar a média populacional, com 95% de intervalo de confiançade confiança não maior que vinte cm.

- Para estimar o $t_{critico}$ preciso de um 'n' inicial.

  n=20

  s^2= 64

IC = 20 cm

$t_{0,05(2),\,19} = 2,093$

$$n = \frac{64 \times 2,093^2}{10} = 28,04$$

- Estimando novamente, utilizando n = 28, $t_{0,05(2),27} = 2,052$

$$n = \frac{64 \times 2,052^2}{10} = 26,95$$

- Estimando com n= 27

$$n = \frac{64 \times 2,056^2}{10} = 27,05$$

- Esta convergência de resultados permite inferir que mais de 27 peixes são necessários para estimar a média, com o intervalo de confiança acima.

Hipótese unilateral

Por exemplo: uma variedade de sementes germina em média, em 10 dias, espera-se que um lote de sementes germine mais rapidamente, pois foi induzida a quebra de dormência. Dados em dias após a quebra.

5	2	1	8	3	4	7	6	5	3	7	5	4	8	3	4

A. $H_0 \rightarrow \mu_0 \geq 10$

$H_1 \rightarrow \mu_1 < 10$

B. $\alpha = 0,05$

C. $t_{crit,15,\,0,05} = 1,753$

D. $t = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}} = \dfrac{4,69 - 10}{\dfrac{2,09}{\sqrt{16}}} = -10.17$

Resposta: rejeita-se H_0, pois $t_{calc} = |-10,17| > t_{crit} = 1,753$. O lote com quebra de dormência germina significativamente antes.

Limites de confiança unicaudal

A. $H_0 \rightarrow \mu_0 \geq 10$

$H_1 \rightarrow \mu_1 < 10$

O limite para H_0 será 4,69 + 1,753 × 0,53 e o limite 2 será $=\infty$.

Teste Binomial

Teste não paramétrico utilizado para dados binários (0 ou 1, presença ou ausência, positivo ou negativo). Os dados amostrais são comparados com os populacionais. A proporção p de indivíduos é definida por p=x/n, onde 'n' é o tamanho da amostra e 'p' a proporção de indivíduos com essa característica na população.

As hipóteses bi e unilaterais podem ser:

H_0: $p=p_0$ e H_1: $p \neq p_0$.

H_0: $p \geq p_0$ e H_1: $p<p_0$. Ou H_0: $p \leq p_0$ e H_1: $p>p_0$.

Método:

1. Hipótese
2. Determinar a frequência de cada classe
3. Estimamos a probabilidade observada
4. Decisão

Exemplo: para saber se determinada espécie de papagaio prefere uma espécie de sementes, nós oferecemos duas sementes em 10 tentativas,verificamos que ele escolheu a semente 'A' em oito vezes.

H_0: o papagaio pega as sementes aleatoriamente $p_0 = p_1$.

H_1: o papagaio tem preferência por uma das sementes $p_0 \neq p_1$.

Nível de significância α=0,05

x	0	1	2	3	4	5	6	7	8	9	10
p	0,00098	0,00977	0,04395	0,11719	0,20508	0,24609	0,20508	0,11719	0,04395	0,00977	0,00098

Se H_0 é verdadeiro, a frequencia de sementes deveria ficar entre 5±3. Isto é os valores acima de oito e abaixo de 3, deveriam somar um p>0,05.

$$p = 0,00098 + 0,00977 + 0,00977 + 0,00098 = 0,021$$

Como $p_{calculado}$=0,021 é menor que o $p_{crítico}$= 0,05, concluímos que o papagaio prefere uma das sementes.

Hipótese unilateral

Utilizando os mesmos dados mas, com outra pergunta.

H_0: o papagaio não escolhe sementes grandes $p_0 \leq p_1$.

H_1: o papagaio escolhe sementes grandes $p_{0>}p_1$.

Resultado: o papagaio escolheu 8 sementes grandes.

x	0	1	2	3	4	5	6	7	8	9	10
p	0,00098	0,00977	0,04395	0,11719	0,20508	0,24609	0,20508	0,11719	0,04395	0,00977	0,00098

$$p = 0{,}04395+0{,}00977+0{,}00098 = 0{,}05469$$

Como $p_{calculado}$= 0,0547 é maior que o $p_{crítico}$= 0,05, concluímos que o papagaio não prefere sementes grandes.

TESTE QUI-QUADRADO

O teste de hipóteses Qui-quadrado analisa as diferenças entre as frequências observadas e as esperadas em cada categoria, adequado para análises de dados na escala nominal. Este teste trabalha com variáveis politômicas (várias categorias), ao contrário da distribuição Binomial, que é dicotômica. É um teste não paramétrico, não utiliza os parâmetros de média e variância nos seus cálculos.

Neste teste comparamos os valores numéricos de uma amostra com distribuições teóricas esperadas, recebendo assim a denominação de teste de aderência.

Para realizar o teste os dados necessitam:

- Ser independentes.
- Devem ter sido obtidos aleatoriamente.
- As observações devem ser frequências ou contagem
- Cada observação pertence apenas a uma categoria.
- Cada amostra deve ser relativamente grande (maior que5).

Pode-se dividir o teste Qui-quadrado em três finalidades de uso:

- Teste de aderência: aplica-se para verificar se k-eventos concordam com a frequencia esperada (FE). Isto é, se uma amostra apresenta a distribuição definida em H_0.
- Teste de homogeneidade: usado para testar o ajuste dos dados a um modelo de distribuição (Poisson, Binomial, Normal).
- Teste de independência: é utilizado para analisar o relacionamento de duas ou mais variáveis (a independência entre as mesmas). Isto é, se há independência entre elas.

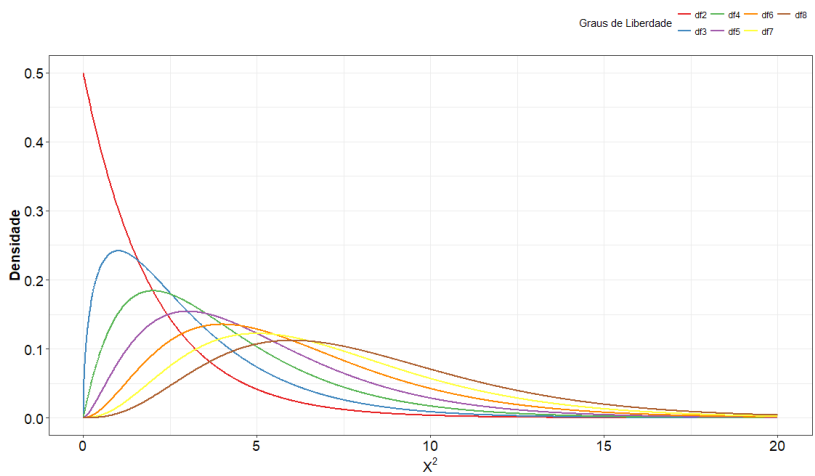
A técnica utilizada é do tipo prova de aderência, no sentido de que pode ser empregada para comprovar se existe diferença significativa entre o número observado de indivíduos, ou de respostas, em determinada categoria, e o respectivo número esperado, baseado na hipótese de nulidade. A técnica x^2 testa se as freqüências observadas estão suficientemente próximas das esperadas para justificar sua ocorrência sob H_0 (eq.4.3).

$$x^2 = \sum_{i=1}^{k} \frac{(fo-fe)^2}{fe} \qquad \textbf{eq. 4.3}$$

Onde "*fo*" é a freqüência observada e "*fe*" é a freqüência esperada.

Não é valido converter os dados em porcentagens e tentar submeter os mesmos à equação acima.

A forma da distribuição x^2 é assimétrica, e esta assimetria diminui com o aumento do número de categorias, como abaixo



A prova x^2 de uma amostra

Teste de Hipóteses:

H_o: FO=FE

H_1: FO≠FE

Nível de significância

α = 0,05

Graus de liberdade (gl)

gl = k – 1, onde k representa o número de categorias na classificação.

Pequenas frequências esperadas

Quando gl=1, isto é, quando k=2, cada freqüência esperada não deve ser inferior a 5. Quando gl>1, isto é, quando k>2, a prova x^2 para o caso de uma amostra, este teste não deve ser usado se mais de 20% das freqüências esperadas são inferiores a 5 ou se qualquer freqüência esperada é inferior a 1.

Se o pesquisador trabalha com duas categorias e tem uma freqüência esperada menor do que 5, então ele deve utilizar a prova Binomial.

Exemplo com frequências esperadas distintas:

Sabe-se que a distribuição por tipo sanguíneo da população brasileira é O= 44%, A= 39%, B= 13% E AB= 4%. Uma amostra de alunos do ensino médio apresentou os seguintes valores: O=62, A=21, B=12 E AB= 5. A amostra é representativa da população brasileira?

H_0: FO = FE

H_1: FO ≠ FE

Valor de significância: α= 0,05

Valor crítico:

Gl = (4-1) = 3 – $x^2_{0,05}$ = 7,81 (que está na tabela de Qui-quadrado)

Cálculo

	FO	*FE*	*(FO-FE)²/FE*
O	62	44	7,4
A	21	39	8,3
B	13	12	0,1
AB	4	5	0,2
Total	**100**	**100**	**16**

$X^2_{cal}(16)<X^2_{crit}(7,81)$ → rejeito H_0, os dados observados diferem significativamente dos esperados.

Exemplo para frequências esperadas iguais:

Com o objetivo de analisar a proporção entre os sexos de uma comunidade. Foram selecionados 90 casais com filho único e verificado o sexo das crianças. Com 37 filhos homens e 53 filas mulheres. A proporção esperada deveria ser 1:1.

	FO	FE	$(FO-FE)^2/FE$
H	37	45	1,42
M	53	45	1,42
total	90	90	2,84

$$X^2 = 2,84 < X^2_{0,5}=3,84$$

Não rejeito H_0.

EXERCÍCIOS CAPÍTULO 4

1) O tempo geracional médio de uma espécie de pernilongo é de 450 horas, d.p.=50. Uma amostra (n=42) vinda de área com alta incidência de dengue, o tempo geracional foi de 430 horas. Os dados indicam que esta população tem tempo geracional menor. Apresente o teste de hipóteses. Realize um teste com 5% de significância. Apresente o valor crítico e a conclusão. Suponha que o tempo geracional tem distribuição normal.

2) Um vendedor de frangos afirma que o peso médio do frango resfriado é de 2,3 kg. Sabe-se que o desvio padrão da fábrica é de 0,15 kg, e que tem distribuição normal. O comprador pesa os frangos de duas caixas (30 ind.) e acha um peso médio de 2,1 kg. O peso médio fornecido pelo vendedor está correto?

3) Na minha rua há vários ipês que começaram a florir. Para verificar se a floração é simultânea, registrei o dia de inicio de floração das árvores. Qual a sua conclusão?

Floração	0	0	1	0	0	2	3	0	1	0	0

4) Um criador de galinhas colocou som no galinheiro para que as galinhas ficassem mais calmas e produzissem mais ovos. O agricultor acha que elas estão perdendo peso, devido a este método. A perda de peso média de 40 galinhas amostradas foi de 30 g em relação ao peso padrão do período, com desvio padrão de 50g e distribuição normal. Apresente as hipóteses. Realize um teste com 5% de significância. Apresente o valor crítico e a conclusão.

5) Foram medidos os ovos de diversos ninhos em uma colônia de mergulhões, para saber se o turismo está afetando o tamanho dos mesmos, cuja média de comprimento do ovo era de 46,3 mm, em período anterior à abertura da ilha à visitação. Qual a conclusão estatística?

22,5	56,2	39,5	33,4	50,9	40,6	45,2	58,9	58,0	53,0	31,8
26,8	34,5	21,1	51,9	39,5	25,3	52	42,0	34,2	33,1	59,2

6) Um pesquisador quer saber se os assaltantes atacam mais homens ou mulheres. Analisando as fichas de ocorrência em uma delegacia. Ele verificou que entre 18 pessoas, apenas duas eram mulheres.

7) Em 20 nascimentos de uma clínica pediátrica nasceram 12 meninos e 8 meninas. Para esta situação o valor do Qui-quadrado é? Esta diferença é significativamente diferente do esperado?

8) Analisando as faltas dos alunos durante a semana nas aulas de anatomia do primeiro período durante vários anos, considerando que as turmas iniciais tem 60 alunos, obtivemos os seguintes resultados. Há alguma tendência ou as faltas são aleatórias em relação ao dia da semana?

Dias	**Seg**	**Ter**	**Qua**	**Qui**	**Sex**
Faltosos	36	22	17	5	40

9) A estrutura populacional de papagaios em uma área degradada está abaixo. A proporção de uma população natural é: 0=60%, 1=23%, 2=14%, 3=3%, 4=0. O que você conclui a partir destes dados?

Idade	0	1	2	3
Número de indivíduos	122	45	13	2

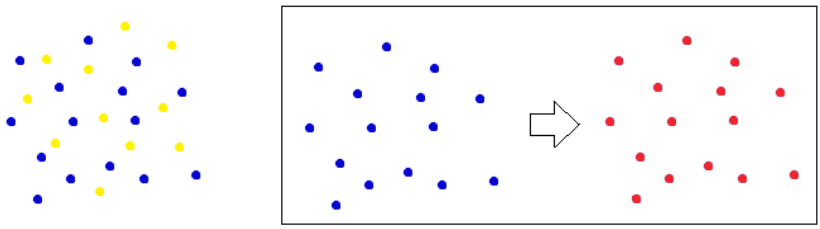
10) A estrutura populacional de uma espécie de ave foi de: indivíduos com menos de um ano= 20%, 1ano=15%, dois anos =12.5% e três anos =52,5%. Dez anos depois, foi realizado um novo censo, obtendo-se os resultados abaixo. A estrutura populacional mudou ?

Idade	0	1	2	3
N	35	25	30	30

11) Três agricultores cortaram respectivamente, 1000, 750 e 1100 kg de cana, em um dia de trabalho e receberam a mesma quantia. Você acha que o patrão foi justo?

Capítulo 5

TESTES PARA DUAS AMOSTRAS



Teste t de student para duas amostras independentes

Geralmente, este teste é utilizado para comparar as médias de duas amostras independentes, sem conhecer a média e desvio padrão populacionais. Os passos necessários são:

1. **Teste de hipóteses:**
   $H_0: \overline{x}_1 = \overline{x}_2$ ou $\overline{x}_1 - \overline{x}_2 = 0$
   $H_1: \overline{x}_1 \neq \overline{x}_2$ ou $\overline{x}_1 - \overline{x}_2 \neq 0$
2. **Escolher o nível de significância:**
   $\alpha = 0{,}05$
3. **Calcular a homogeneidade da variância.**

Para comparar duas variâncias, seguimos os passos do teste de hipóteses, que deve ser bilateral, como já mostrado nos testes Z e t de student.

A. $H_0 \rightarrow s_1^2 = s_2^2$
   $H_1 \rightarrow s_1^2 \neq s_2^2$

B. Escolher o nível de significância:

α= 0,05

C. Calcular a Homogeneidade da variância

Para testar as variâncias, utilizamos o teste 'F'. O valor esperado é 1 ($s_1^2/s_2^2=1$). Para facilitar o teste coloca-se a variância maior no numerador, de modo que o valor de F será sempre igual o maior que 1. O f calculado é obtido pela razão das duas variâncias,

$$F_{calc} = \frac{s^2_{maior}}{s^2_{menor}}$$

E o cálculo das variâncias por esta equação já detalhada anteriormente,

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

O valor crítico de F depende do nível de significância e dos graus de liberdade:

$$F_{\alpha,\ g.l.n,\ g.l.\ D}$$

D. Identifique o 'F' critico na tabela:

Para isto, utiliza a tabela de Fisher bilateral.

E. Decisão:

Para variâncias **homogêneas**, siga adiante

4. Calcular o valor da distribuição t para as amostras:

Supomos que $\mu_1 - \mu_2=0$, onde s^2= variância agrupada, s^2_1= variância da amostra 1 e s^2_2= variância da amostra 2.

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \qquad \textbf{eq. 5.1}$$

$$t_{calc} = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \rightarrow t_{calc} = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad \textbf{eq. 5.2}$$

O teste t pode ser calculado de forma análoga à computação de uma amostra de teste t, quando as amostras apresentam distribuição normal e homogeneidade das variâncias:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{x}_1 - \overline{x}_2}}$$

Onde $s_{\overline{x}_1 - \overline{x}_2}$ é o erro padrão da diferença entre as médias amostrais, que mede a variabilidade de dados dentro das duas amostras. O $s_{\overline{x}_1 - \overline{x}_2}$ é obtido a partir da variância agrupada

$$s^2 = \frac{SQ_1 + SQ_2}{gl_1 + gl_2} \rightarrow s_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$$

Assim, a equação do teste t fica:

$$t_{calc} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$, igual à equação acima (eq. 5.2)

5. **Calculamos o valor crítico do teste:**

$$\text{g.l.} = n_1 + n_2 - 2$$

6. **Decisão:**

Exemplo:
Considere as duas amostras abaixo:

N	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
#1	1	35	9	16	24	21	8		
#2	24	16	40	20	34	19	40	28	49

1. H_0: 16,3 = 30
   H_1:16,3 ≠ 30
2. α= 0,05
3. Cálculo da homogeneidade da variância

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \rightarrow s_1^2 = \frac{2644 - \frac{114^2}{7}}{7-1} = 131{,}24$$

$$s_2^2 = \frac{9134 - \frac{270^2}{9}}{9-1} = 129{,}25$$

F=131,24/129,25 → F= 1,015

$F_{0,05,\,6,\,8} = 4{,}65$

Como $F_{calculado}$(1,015) é menor que $F_{crítico}$ (4,65), não rejeito H_0 → a variância é homogênea.

4. Calcular o valor da distribuição t para as amostras:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \rightarrow s^2 = \frac{(7-1)\times 131{,}24 + (9-1)\times 129{,}25}{7+9-2} = 130{,}10$$

$$t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \rightarrow t_{calc} = \frac{(16{,}3-30)}{\sqrt{130{,}10\left(\frac{1}{7}+\frac{1}{9}\right)}} = -2{,}38$$

5. tcri(0,05, 14)t = 2,145
6. Decisão: tcrit < tcalc, |2,145|<t|-2,38|. Rejeitamos H_0, a diferença entre as duas médias é significativa.

Teste unicaudal para duas amostras independentes

Quando o pesquisador quer analisar as diferenças em apenas uma direção. Por exemplo, frutos de ameixa em um pomar sem limpeza constante têm apresentado tamanhos menores, que de um pomar 'bem tratado'. Estas diferenças são significativas? A hipótese nula será que o tamanho das frutas será igual, ou até maior do que no pomar cuidado e a hipótese alternativa será que elas serão menores.

1. $H_0 \rightarrow \mu 1 \geq \mu 2$

   $H_1 \rightarrow \mu_1 < \mu_2$
2. Escolher o nível de significância:

   $\alpha = 0{,}05$

3. Calcular o valor da distribuição t para as amostras:

Dados:

n	Sem limpeza	Com limpeza	SQ1	SQ2
1	44,60	41,51	3,42	31,00
2	42,79	48,23	0,00	1,33
3	44,62	50,43	3,49	11,24
4	46,50	45,42	14,06	2,75
5	40,54	46,91	4,89	0,03
6	39,17	52,70	12,82	31,61
7	47,54	49,99	22,94	8,48
8	35,26	45,71	56,11	1,87
9	46,69	40,12	15,52	48,41
10	43,19	48,99	0,19	3,66
11	47,61	42,93	23,61	17,20
12	37,47	51,99	27,89	24,13
13	41,01		3,03	
14	41,52		1,51	
SOMA	**598,51**	**564,93**	**189,48**	**181,70**
MÉDIA	**42,75**	**47,08**		
Variância	**14,58**	**16,52**		

F = 16,52/14,58 → F = 1,13, $F_{0,05,\ 11,\ 13} = 2,67$ → variância homogênea.

$$SQ = \left(X - \overline{X}\right)^2 \rightarrow S^2 = \frac{SQ_1 + SQ_2}{gl_1 + gl_2} \rightarrow S^2 = \frac{189,48 + 181,70}{13 + 11} = 15,47$$

$$t_{calc} = \frac{\left(\overline{x}_1 - \overline{x}_2\right)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightarrow t_{calc} = \frac{42,75 - 47,08}{\sqrt{15,47\left(\frac{1}{14} + \frac{1}{12}\right)}} = -2,80$$

4. Calculamos o valor crítico do teste:

   g.l.= $n_1 + n_2 - 2$ $t_{crit} = 1,711$

5. Decisão

$1,71<|-2,80|$

$T_{crit} < |t_{calc}| \rightarrow$ rejeitamos H_0.

Teste t para duas amostras com variâncias desiguais

Quando a variância entre duas amostras é significativamente diferente, uma das alternativas é utilizar a equação abaixo:

$$t'_{calc} = \frac{\overline{x}_a - \overline{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \qquad \textbf{eq. 5.3}$$

E o grau de liberdade deve ser calculado por:

$$gl' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad \textbf{eq. 5.4}$$

Exemplo:

Comparando duas amostras de peixes "Galo", uma capturada no estuário e outra em praias, os dados foram n=10 e média 4,50 cm ± desvio padrão de 1,138 cm, e n=10, média 8,51 ± desvio padrão de 4,68 cm, respectivamente.

- Primeiro, comparamos as variâncias entre as duas amostras:

$$F_{calc} = \frac{\sigma^2_{maior}}{\sigma^2_{menor}} = F_{calc} = \frac{324}{49} = 6,61$$

$F_{0,05,48,53}=1,75$

Rejeito H_0, as variâncias são diferentes

- Realizo o teste t para as variâncias desiguais

$$t'_{calc} = \frac{\overline{x}_a - \overline{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \rightarrow t'_{calc} = \frac{56,2 - 43,2}{\sqrt{\frac{49}{53} + \frac{324}{48}}} = 4,69$$

$$gl' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \rightarrow gl' = \frac{\left(\frac{49}{53} + \frac{324}{48}\right)^2}{\frac{\left(\frac{49}{53}\right)^2}{52} + \frac{\left(\frac{324}{48}\right)^2}{48}} \cong 61$$

$$t'_{0,05,61} = 2,00$$

Decisão: rejeito H_0, as duas populações apresentam médias significativamente diferentes.

Teste t de student para duas amostras pareadas (antes e depois)

Quando duas amostras são dependentes, isto é, ou ocorrem aos pares ou representam situações "antes" e "depois", dizemos que os dados são pareados. Por exemplo, um pesquisador está avaliando se um antidepressivo afeta a pressão arterial diastólica dos pacientes:

Paciente	Antes	Depois
1	133	132
2	134	135
3	135	136
4	142	138
5	148	140
6	150	143
7	164	144
8	170	150
9	175	151
10	179	153
11	184	155
12	185	155
13	188	159

1. Teste de hipóteses:
   H_0: $\mu_1 - \mu_2 = 0$
   H_1: $\mu_1 - \mu_2 \neq 0$
2. Escolher o nível de significância:
   $\alpha = 0{,}05$
3. Calcular o valor da distribuição t para as amostras:
   O teste estatístico para testar a hipótese nula é:

$$\bar{d} = \frac{\sum d}{n} \qquad \textbf{eq. 5.5}$$

$$\bar{d} = \frac{\sum d}{n} \qquad \textbf{eq. 5.6}$$

$$EP_{\bar{d}} = \frac{s_d}{\sqrt{n}} \quad s_{\bar{d}} = \frac{s_d}{\sqrt{n}} \quad s_d = \sqrt{\frac{\sum d^2 - \frac{\left(\sum d\right)^2}{n}}{n-1}} \quad s_d = \sqrt{\frac{\sum d^2 - n * \bar{d}^2}{n-1}} \qquad \textbf{eq. 5.7}$$

Paciente	Antes	Depois	d	d^2
1	133	132	1	1
2	134	135	-1	1
3	135	136	-1	1
4	142	138	4	16
5	148	140	8	64
6	150	143	7	49
7	164	144	20	400
8	170	150	20	400
9	175	151	24	576
10	179	153	26	676
11	184	155	29	841
12	185	155	30	900
13	188	159	29	841

Paciente	Antes	Depois	d	d^2
SOMA			**196**	**4766**
MÉDIA			**15,08**	

$$s_d = \sqrt{\frac{\sum d^2 - n \times \overline{d}^2}{n-1}} = s_d = \sqrt{\frac{4766 - 13 \times 15,08^2}{13-1}} = 12,28$$

$$\rightarrow EP_{\overline{d}} = \frac{s_d}{\sqrt{n}} = EP_{\overline{d}} = \frac{12,28}{\sqrt{13}} = 3,41$$

$$t = \frac{\overline{d}}{EP_{\overline{d}}} \rightarrow t = \frac{15,08}{3,41} = 4,42$$

4. Calcular o $t_{crítico}$:
   $t_{(0,05(2),12)}$=2,179

5. Decisão:
   Como $t_{calc} > t_{crítico}$, rejeito H_0, as médias são significativamente diferentes.

TESTE QUI-QUADRADO

A prova x^2 para duas amostras independentes (tabelas de contingência)

Pode-se aplicar a prova c^2 para determinar a significância de diferenças entre dois grupos independentes, quando estes se apresentam sob forma de categorias discretas. Neste caso, estaremos testando a homogeneidade ou independência entre os grupos.

A hipótese a ser comprovada é a de que os dois grupos diferem em relação a determinada característica e, consequentemente, com respeito à frequência relativa com que os componentes dos grupos se enquadram nas diversas categorias. *E.g.*: Se dois grupos de estudantes diferem em relação à determinada opinião (se os tubarões estão em extinção ou não), ou se há diferenças relativas ao sexo, na escolha de atividades de lazer.

$$X^2 = \sum_{i=1}^{r} \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}$$

Onde r=linhas e k=colunas

gl=(r-1)(k-1)

Exemplo: imagine que um fornecedor de alimentos quer testar se há diferença no consumo de alcaparras grandes e pequenas entre os tipos de consumidores.

F.O.	Pequena	Grande	Total
Varejista	31	44	**75**
Atacadista	60	25	**85**
Avulso	12	8	**20**
Total	**103**	**77**	**180**

H_0: as variáveis são independentes. As proporções de alcaparras não variam entre os tipos de compradores.

H_1:as variáveis são dependentes, há relação entre tamanho da alcaparra e tipo de comprador.

Frequência esperada

A frequência esperada é calculada por: (soma linha x soma coluna)/ soma total

Ex.: $(103 \times 75)/180 = 43$

F.e.	Pequena	Grande	Total
Varejista	43	32	75
Atacadista	48.6	36.4	85
Avulso	11.4	8.6	20
Total	103	77	180

O cálculo do Qui-quadrado, faz-se para as células correspondentes:

$$X^2 = \frac{(31-43)^2}{43} + \frac{(44-32)^2}{32} + \frac{(60-48,6)^2}{48,6} + \frac{(25-36,4)^2}{36,4} + \frac{(12-11,4)^2}{11,4} + \frac{(8-8,6)^2}{8,6}$$

$$= 3,3 + 4,5 + 2,7 + 3,5 + 0 + 0 = 14,1$$

$$gl = (3-1)(2-1)=2,\ x^2 = 10,67 > x^2 = 5,99 \rightarrow \text{Rejeito } H_o$$

A prova x^2 para K amostras independentes

O teste x^2 é o mesmo, tanto para 2 quanto para K amostras independentes. Os dados são apresentados em tabelas (linhas X colunas ou LXC).

A hipótese de nulidade é que as K amostras de freqüências ou proporções sejam provenientes da mesma população.

$$X^2 = \sum_{i=1}^{r} \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}$$

Exemplo:

Freqüência de utilização de serviços de saúde por 667 idosos das áreas central, intermediária e periférica do município de fortaleza.

Número de visitas	Área			Total
	Central	Intermediária	Periférica	
Nenhuma	80	90	87	257
Uma	79	48	35	162
2 a 5	58	74	63	195
Mais de 5	11	21	21	53
Total	228	233	206	667

Dados esperados:

Número de visitas	Área			Total
	Central	Intermediária	Periférica	
Nenhuma	87,85	89,78	79,37	257
Uma	55,38	56,59	50,03	162
2 a 5	66,66	68,12	60,22	195
Mais de 5	18,12	18,51	16,37	53
Total	228	233	206	667

$$X^2 = \frac{(80-87,85)^2}{87,85} + \frac{(79-55,38)^2}{55,38} + \ldots + \frac{(21-16,37)^2}{16,37} = 23,53$$

Qui-quadrado calculado:

Número de visitas	Área			Total
	Central	Intermediária	Periférica	
Nenhuma	0,701	0,001	0,733	
Uma	10,074	1,304	4,515	
2 a 5	1,125	0,508	0,128	
Mais de 5	2,798	0,335	1,3109	
Total	14,698	2,147	6,687	**23,532**

gl = (4-1)(3-1) = 6, X^2 =23,53 > $X^2_{0,05}$=12,59 → rejeito H_0.

O grau de liberdade do Qui-quadrado é calculado por:

$$G.L. = (num.linhas - 1) \times (num.colunas - 1)$$

Correção de Yates (ou correção de continuidade).

O teste x^2 assume número de amostras grande (>5) e números inteiros. Entretanto, quando a amostra for pequena, a equação Qui-quadrado poderá produzir um valor maior que o real. Quando a frequência observada for igual ou menor que cinco, deve-se utilizar a correção de YATES.

$$X^2 = \frac{(|o-e|-0,5)^2}{e}$$

Utiliza-se esta correção quando:

- O x^2 calculado for maior que 5.
- O valor de n é menor que 40 ou
- O valor de pelo menos uma classe de esperado for menor que 5.
- O valor do X^2 obtido é maior do que o crítico

Quando as exigências do x^2 não forem aceitas, deve-se utilizar o teste exato de Fisher.

Se o teste for não significativo. Isto é, quando o $x^2_{calc} < x^2_{crit}$, não precisa fazer a correção de Yates.

Exemplo:

Um senhor teve 20 netos, 16 homens e 4 mulheres. Ele quer saber se esta proporção foi significativamente maior que o esperado:

	FO	FE	FO-FE	|FO-FE|-0,5	(|FO-FE|-0,5)²/FE
Homens	16	10	6	5,5	3,025
Mulheres	4	10	-6	5,5	3,025
Totais	20	20			6,05

Como o $x^2_{calc} > x^2_{crit}$, rejeito H_0.

TESTE EXATO DE FISHER

Este teste a significância da associação (contingência) entre duas variáveis. É a alternativa para tabelas 2x2 quando não se pode utilizar o Qui-quadrado (quando o valor esperado é menor que 5, ou o numero total de indivíduos é menor que 25). Ele calcula a probabilidade de associação das características que estão em análise.

Ele é usado quando:

- O valor de n é menor que 20,
- $20 < N < 40$ e a freqüência esperada menor que 5

A probabilidade será calculada pelo produto dos fatoriais dos totais marginais dividido pelo produto do fatorial do total geral multiplicado e dos fatoriais dos valores observados em cada classe

	Variável 1A	Variável 1B	Total
Variável 2A	a	b	G
Variável 2B	c	d	H
Total	E	F	I

$$P = \frac{E!\,F!\,G!\,H!}{a!\,b!\,c!\,d!\,I!}$$

Exemplo: foi testada a reação das pessoas à abordagem de estranhos. Um senhor, bem vestido, cumprimentou as pessoas em uma praça.

	Reação +	Reação –	Total
Masculino	8	3	11
Feminino	7	9	16
Totais	15	12	27

$$P = \frac{15!12!11!16!}{8!3!7!9!27!} = 0,108583479$$

A probabilidade das reações serem significativamente diferentes é 10,9%. Por outro lado, a probabilidade de não apresentarem diferenças significativas é de 89,1%. Considerando um alfa de 5%, as duas populações não diferem significativamente.

EXERCÍCIOS CAPÍTULO 5

1) Duas variedades de banana tiveram a produção por hectare mostrada abaixo. A produção média das duas variedades é significativamente diferente ao nível de $\alpha = 0,05$? Não esqueça de testar a homogeneidade da variância.

Variedade1	24	34	19	18	21	10	22	8	22	34	9	23	28
Variedade2	11	14	19	10	2	6	2	12	9	12	19	13	17

2) Foram coletados os fungos associados aos troncos em decomposição em duas áreas de Recife, com as riquezas de espécies abaixo. Posso considerar as duas áreas semelhantes em relação à diversidade macro fungica?

Área1	17	15	3	4	5	9	31
Área2	15	9	23	18	14	7	12

3) Avaliando a pressão arterial sistólica de oito operários antes e depois de implantação do exercício laboral em uma empresa. Sabendo-se que os exercícios combatem a hipertensão.

Antes	115	120	121	123	150	165	167	140
Depois	124	130	140	143	160	162	162	177

4) Um pesquisador mediu as asas de uma espécie de sabiá de um fragmento de mata. Ele quer saber se há assimetria significativa para esta população. Qual a conclusão?

Direita	12	11	19	15	14	16	18	19	17	13	11
Esquerda	13	10	20	16	12	19	17	17	19	11	14

5) O treinador acompanhou o desempenho de atletas, na corrida de 100m rasos, antes e depois da aplicação do seu programa de treinamento. Qual sua conclusão?

Antes	12,3	14,1	12,2	13,1	14,3	13,8	12	14,6	15
Depois	10,4	10,1	10	10,2	10,4	9,9	9,7	9,9	10,9

Capítulo 6

ANÁLISE DE VARIÂNCIA (ANOVA)

A análise de variância (ANOVA) foi criada por Fisher em 1924. Ela compara simultaneamente as médias obtidas em várias amostras, ou tratamentos (três ou mais), de variáveis contínuas com distribuição normal. Isto é, compara a variáveis preditoras categóricas ("várias amostras, vários locais"), também chamadas de variáveis qualitativas, com a variável resposta contínua (dados em si, *e.g.* Temperatura).

Os testes Z e t referem-se a hipóteses sobre medidas de posição, já na análise de variância os testes de hipóteses são sobre medidas de variabilidade. A ANOVA usa a variância entre as médias dos grupos para quantificar a divergência entre os grupos.

A ANOVA, assim como a regressão linear, é uma particularização do modelo linear. Analisar a ANOVA através das semelhanças com a regressão linear permite aproveitar melhor esta ferramenta de análise.

Existem vários tipos de teste ANOVA para os mais variados tipos de desenho amostral. Quando testamos apenas uma variável, estamos analisando apenas um tratamento ou fator, é a variável independente. Os tratamentos podem ser quantitativos (altura, massa, riqueza de espécies, dosagem de nutrientes) ou qualitativos (dados nominais).

A ANOVA tem por base a divisão da 'soma dos quadrados'. A variação total de um conjunto pode ser expressa como a soma dos quadrados: as diferenças entre cada observação (Y_i) e a média geral dos dados ($\overline{Y}$) é elevada ao quadrado e somada. Esta variação total pode ser dividida em diferentes componentes.

ANOVA – 1 CRITÉRIO

Na ANOVA – one way, onde analisamos um único fator. A variação total é dividida em duas frações: a variação referente às médias dos vários grupos,

comparada com e média geral de todos os indivíduos e a variação observada entre as unidades de cada grupo. Isto é:

Variação total = variação entre tratamentos + variação dentro dos tratamentos

Cada observação ou amostra aleatória é:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad \textbf{eq. 6.1}$$

Onde: y_{ij}= valor da variável na observação 'i' no tratamento 'j', μ= média total da variável, τ_i= efeito fixo do grupo ou tratamento 'i', ε_{ij}= erro aleatório com média 0 e variância σ^2.

A variável independente τ, frequentemente chamada fator, representa o efeito dos diferentes tratamentos. O fator influencia os valores da variável dependente y.

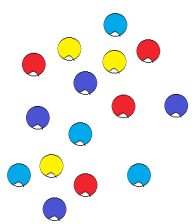
Objetivo: testar a significância das diferenças entre as médias

A base da ANOVA é que as variâncias podem ser divididas entre as diversas fontes de variação. Lembre que a variância é a soma do quadrado dos desvios da média, dividida por n-1.

Se não houver diferenças entre os tratamentos (populações), a razão entre as variâncias 'entre' e 'dentro' deverá ser igual a 1. Mas, podem ocorrer diferenças aleatórias, fazendo com que F (razão entre as duas variâncias) flutue ao acaso. Para testar a significância de F_{calc}, compara-se com o limite para uma diferença aleatória entre as variâncias, que pode ser encontrado em uma tabela.

Exemplo:

Imagine o incremento em biomassa de quatro variedades de feijão. As plântulas foram coletadas aleatoriamente para cada variedade, todas sob as mesmas condições ambientais (temperatura, umidade, tipo de solo, dia da germinação).



	Observações	V1	V2	V3	V4
Repetições	1	22	29	27	35
	2	18	25	19	30
	3	21	27	25	29
	4	23		22	31

A hipótese a ser testada é:

H_0: M1 = M2=M3=M4

H_1: M1 ≠ M2≠M3≠M4 ou M1= M2≠M3≠M4 ou M1≠ M2=M3 → onde pelo menos uma das médias é diferente.

No nosso exemplo:

H_0: 21= 27= 23,25= 31,25'

As médias das observações podem ser visualizadas no gráfico abaixo:

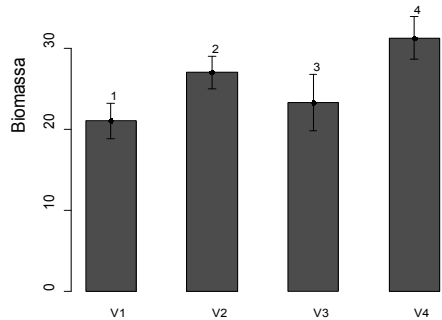


Figura 6.1: Média (±Desvio padrão) dos quatro tratamentos analisados

Os passos necessários para calcular o F de Fisher, estão resumidos na tabela abaixo, onde G.L.= grau de liberdade, S.Q.= soma dos quadrados, Q.M.= quadrado médio.

	G.L.	S.Q.	Q.M.	F
Grupos	k-1	$SQ_{grupos} = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{x}\right)^2$	$QM_{grupos} = \frac{SQ_{grupos}}{G.L._{grupos}}$	$F = \frac{QM_{grupos}}{QM_e}$
Erro	n-k	$SQ_e = \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}_i\right)^2\right]$	$QM_e = \frac{SQ_e}{G.L._e}$	
Total	n-1	$SQ_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}\right)^2$		

Nós temos 'k' grupos e 'n_i' amostras em cada grupo, O total de amostras é a soma das amostras de todos os grupos:

$$n = \sum_{i=1}^{k} n_i$$

No caso 'n' é: 4+3+4+4=15

Nós temos 4 variedades de feijão (4 tratamentos), então K=4.

Conhecendo 'n' e 'k', podemos calcular os graus de liberdade:

- O grau de liberdade dos grupos é: k-1=3
- O grau de liberdade do erro é: n-k=15-4=11
- O grau de liberdade total é: n-1= 15-1=14

Em seguida, calculamos as somas dos quadrados (S.Q.) de cada fonte:

A soma dos quadrados de todos os dados (SQ_{tot} é):

$$SQ_{tot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x} \right)^2$$

Onde $\overline{x}$ é a média geral e x_{ij} representa cada dado.

A média geral é a soma de todos os dados divididos por n:

$$\overline{x} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_j}{n}$$

$$\overline{x} = \frac{(22+18+...+31)}{15} = 25,53$$

$$SQ_{tot} = (22-25,53)^2 + (18-25,53)^2 + ... + (31-25,53)^2 = 319,73$$

$$SQ_{tot} = 10099 - \frac{383}{15} = 319,73$$

A variabilidade atribuída às diferenças entre as médias dos k grupos é:

$$SQ_{grupos} = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{x}\right)^2$$

Observações	V1	V2	V3	V4
1	22	29	27	35
2	18	25	19	30
3	21	27	25	29
4	23		22	31
Média	21	27	23.25	31.25
Soma	84	81	93	125

$$SQ_{grupos} = 4*(21-25,53)^2 + 3*(27-25,53)^2 + 4*(23-25,53)^2 + 4*(31,25-25,53)^2 = 240,23$$

A soma dos quadrados interna dos grupos (SQ_e) é calculada por:

$$SQ_e = \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}_i\right)^2\right]$$

$$SQ_e = (22-21)^2 + (18-21)^2 + \ldots + (31-31,25)^2 = 79,5$$

Resumindo as notações:

	G.L.	S.Q.
Grupos	3	240,2
Erro	11	79,5
Total	14	319,7

Para testar as hipóteses calculamos os quadrados médios (QM). Assim:

$$QM_{grupos} = \frac{240,2}{3} = 80,08$$

$$QM_e = \frac{79,5}{11} = 7,23$$

$$F = \frac{80,08}{7,23}$$

Os resultados do exemplo foram:

	G.L.	S.Q.	Q.M.	F
Grupos	3	240,2	80,08	11,08
Erro	11	79,5	7,23	
Total	14	319,7		

O valor crítico para este teste, F_{crit}, é $F_{\alpha(1),(k-1),(n-k)}$, encontrado na tabela, onde o nível de significância α é unicaudal:

Decisão: F_{calc}=11,08>F_{crit}= 2,59: pelo menos, uma das médias é significativamente diferente das outras ao nível de 5%.

Pré-requisitos da análise de variância

- Normalidade: os resíduos têm distribuição normal, com média igual a zero.
- Independência das amostras: as amostras representam um conjunto aleatório de todos os dados existentes. As observações são todas independentes entre si.
- Homocedasticidade: a variância de cada grupo é semelhante à dos outros grupos. Assim, cada grupo contribui de modo equivalente para a soma dos quadrados dentro dos grupos.

  Entretanto, a ANOVA é um teste robusto, fornecendo resultados confiáveis mesmo com considerável heterocedasticidade, desde que os tamanhos amostrais sejam iguais ou aproximadamente iguais.

Comparações múltiplas entre médias

O valor de f significativo indica apenas que há diferenças significativas entre os tratamentos mas, não diz quais médias são diferentes entre si. Existem vários métodos para identificar quais médias, par a par, diferem significativamente entre si.

Teste de Tukey:
Para realizar este teste, primeiro ordenamos as médias:

Tratamento	V1	V2	V3	V4
Média	21	27	23.25	31.25
N	4	3	4	4

Depois calculamos a diferença entre as médias

Diferenças	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
$\bar{x}_1$	0			
$\bar{x}_2$	6	0		
$\bar{x}_3$	2,25	3,75	0	
$\bar{x}_4$	10,25	4,25	8	0

Depois calculamos o erro padrão de cada diferença entre médias:

$$EP = \sqrt{\frac{QM_{res}}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$EP = \sqrt{\frac{7,23}{2}\left(\frac{1}{4}+\frac{1}{3}\right)} = 1,45$$

Erro Padrão	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$
$\overline{x}_1$	0			
$\overline{x}_2$	1,45	0		
$\overline{x}_3$	1,34	1,45	0	
$\overline{x}_4$	1,34	1,45	1,34	0

Calcule a estatística do teste q:

$$q_{calc} = \frac{\overline{x}_1 - \overline{x}_2}{EP} = q_{calc} = \frac{21-27}{1,45} = 4,14$$

$\mathbf{q_{calc}}$	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$
$\overline{x}_1$	0			
$\overline{x}_2$	4,14	0		
$\overline{x}_3$	1,68	2,59	0	
$\overline{x}_4$	7,65	2,93	5,97	0

Calcular o valor de q_{crit}, na tabela. $q_{\alpha,k,\ g.l.Res} \rightarrow \quad q_{0,05,\ 4,\ 11} = 4,2567$

$\mathbf{q_{calc}}$	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$
$\overline{x}_1$	0			
$\overline{x}_2$	4,14	0		
$\overline{x}_3$	1,68	2,59	0	
$\overline{x}_4$	7,65	2,59	5,97	0

Correção de Bonferroni

Este método corrige o valor de α:

$$\alpha_{Bonf} = \frac{\alpha}{m}$$

Onde α é o nível de significância global do experimento e 'm' é o número de comparações a serem realizadas. A escolha do número de comparações deve ser feita a priori.

O teste estatístico utilizando a correção de Bonferroni é:

$$t_{Bonf} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{QM_{\mathrm{Re}s}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

O t_{crit} é escolhido com base no alfa corrigido de Bonferroni e nos graus de liberdade do resíduo da ANOVA.

$$\alpha_{Bonf} = \frac{0,05}{4} = 0,0125$$

Na tabela *t* de student o valor mais próximo de α é 0,01, então $t_{0,01,\,11} = 3,106$

Diferenças	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$
$\overline{x}_1$	0			
$\overline{x}_2$	6	0		
$\overline{x}_3$	2,25	3,75	0	
$\overline{x}_4$	10,25	4,25	8	0

E.P.	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
$\bar{x}_1$	0			
$\bar{x}_2$	2,05	0		
$\bar{x}_3$	1,90	2,05	0	
$\bar{x}_4$	1,90	2,05	1,90	0

$\mathbf{t_{calc}}$	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
$\bar{x}_1$	0			
$\bar{x}_2$	2,92	0		
$\bar{x}_3$	1,18	1,83	0	
$\bar{x}_4$	5,39	2,07	4,21	0

Decisão: a média $\bar{x}_4$ é significativamente diferente das médias $\bar{x}_1$ e $\bar{x}_3$.

Teste de Dunnet: Este teste a posteriori compara o 'controle' com os demais tratamentos apenas

Teste de Scheffé: É o teste mais conservador.

Análise de variância segundo dois critérios ou análise de blocos aleatorizados (ANOVA two way) sem interação (blocos casualizados), sem repetição

Este teste compara a variação entre os tratamentos e entre os blocos. Os blocos são causas de variação, formando subamostras, ou blocos. Uma unidade de cada tratamento mais o controle formam um bloco. Assim, cada bloco terá uma unidade de cada um dos tratamentos. As observações serão classificadas segundo dois critérios: tratamento e bloco. O teste é semelhante à análise fatorial, mas não existe interação entre os fatores e temos apenas uma observação por célula. Este teste segue os pressupostos de normalidade, homogeneidade das variâncias, independência entre as observações, sem interação entre os fatores.

Ocorre a aleatorização dos blocos. Isto é as amostras de cada bloco são escolhidas por sorteio.

Nesta análise teremos duas hipóteses:

H_0: não há diferenças entre os tratamentos analisados.

H_1: há diferenças entre os tratamentos analisados.

H_0: não há diferenças entre os blocos analisados.

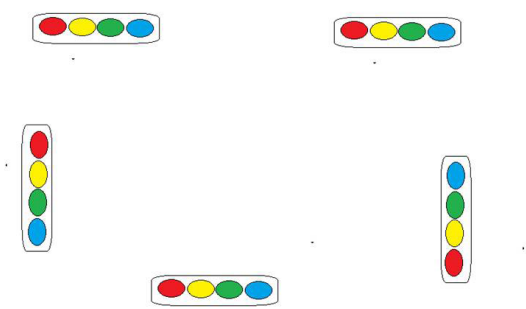
H_1: há diferenças entre os blocos analisados.

O resumo da análise é:

Fontes de variação	***G.L.***	***S.Q.***	***Q.m.***	F_{calc}
Tratamentos	k-1	$b\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2$	$QM_{grupos}=\frac{SQ_{grupos}}{G.L._{grupos}}$	$F_{calc}=\frac{QM_{grupos}}{QM_e}$
Blocos	b-1	$k\sum_{j=1}^{b}(\overline{y}_j-\overline{y})^2$	$QM_b=\frac{SQ_b}{G.L._b}$	$F_{calc}=\frac{QM_{blocos}}{QM_e}$
Resíduos	(k-1)(b-1)	$\sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij}-\overline{y}_i-\overline{y}_j+\overline{y})^2$	$QM_e=\frac{SQ_e}{G.L._e}$	
Total	kb-1	$\sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij}-\overline{y})^2$		

Onde k= número de tratamentos, b= número de blocos.

Exemplo: foram utilizados quatro meios de cultura de bactérias. Cinco blocos (uma placa de cada meio) foram colocados, cada um em um setor distinto de uma fábrica (refeitório, escritório, ferramentaria,...) Qual a diferença entre eles?



		Tratamentos					
Blocos	***Locais\Placas***	***M1***	***M2***	***M3***	***M4***	***Total (Blocos)***	$\bar{y}_j$
	#1	3	2	7	1	13	3,25
	#2	2	6	0	4	12	3
	#3	0	6	3	9	18	4,5
	#4	2	1	7	3	13	3,25
	#5	1	0	0	3	4	1
	Total (trat)	8	15	17	20	**60**	Soma Geral
	$\bar{y}_i$	1,6	3	3,4	4	**3**	Média geral

$$SQ_{grupos} = \frac{\sum\left(T_i^2\right)}{b} - \frac{\left(\sum x\right)^2}{kb} \rightarrow \frac{\sum\left(8^2+15^2+17^2+20^2\right)}{5} - \frac{(60)^2}{4\times5} = 195,6 - 180 = 15,6$$

$$SQ_{bloco\,s} = \frac{\sum\left(B_i^2\right)}{k} - \frac{\left(\sum x\right)^2}{kb} \rightarrow \frac{\sum\left(13^2+12^2+18^2+13^2+4^2\right)}{4} - 180 = 205,5 - 180 = 25,5$$

$$SQ_{tot} = \sum x^2 - \left(\frac{\left(\sum x\right)^2}{kb}\right) \rightarrow = 318 - 180 = 138$$

$$= SQ_{tot} - SQ_b - SQ_g \rightarrow = 138 - 15,6 - 25,5 = 96,9$$

Fontes de variação	***G.L.***	***S.Q.***	***Q.m.***	F_{calc}	F_{crit}	***Decisão***
Tratamentos	3	15,6	5,2	0,64	3,49	Não diferem
Blocos	4	25,5	6,375	0,79	3,26	Não diferem
Resíduos	12	96,9	8,075			
Total	**19**	**138**				

É importante realçar que a partir do momento que testamos dois fatores, temos duas hipóteses para serem testadas.

ANOVA com blocos aleatorizados com repetição

O modelo de ANOVA com blocos aleatorizados pode ser descrito como:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_j$$

Onde: μ= média geral, τ_i= efeito do tratamento i, β_j= efeito do bloco j, ε_{ij}= erro (aleatório e independente) de cada observação

Fontes de variação	*G.L.*	*S.Q.*	*Q.m.*	F_{calc}
Tratamentos	k-1	$b\sum_{i=1}^{k}\left(\overline{y}_i - \overline{y}\right)^2$	$QM_{grupos} = \frac{SQ_{grupos}}{G.L._{grupos}}$	$F_{calc} = \frac{QM_{grupos}}{QM_e}$
Blocos	b-1	$k\sum_{j=1}^{b}\left(\overline{y}_j - \overline{y}\right)^2$	$QM_b = \frac{SQ_b}{G.L._b}$	$F_{calc} = \frac{QM_{blocos}}{QM_e}$
Resíduos	(k-1)(b-1)	$\sum_{i=1}^{k}\sum_{j=1}^{b}\left(y_{ij} - \overline{y}_i - \overline{y}_j + \overline{y}\right)^2$	$QM_e = \frac{SQ_e}{G.L._e}$	
Total	kb-1	$\sum_{i=1}^{k}\sum_{j=1}^{b}\left(y_{ij} - \overline{y}\right)^2$		

Análise de variância com dois fatores com interação

$$Y_{ij} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_j$$

Onde: μ= média geral, τ_i= efeito do tratamento i, β_j= efeito do bloco j, $\tau\beta_{ij}$= efeito da interação entre tratamentos e blocos, ε_{ij}= erro (aleatório e independente) de cada observação

Num experimento podemos testar dois fatores. Por exemplo, podemos testar três temperaturas de cultivo e três tipos de ração para uma espécie de peixe. Esta análise é fatorial, pois há um cruzamento, onde cada tratamento de um fator pode ser analisado em relação a cada tratamento do outro fator.

Nesta análise teremos três hipóteses:

H_0: Não há diferenças na variação de massa nas temperaturas analisadas.

H_1: Há diferenças na variação de massa nas temperaturas analisadas.

H_0: Não há diferenças na variação de massa entre as rações utilizadas.

H_1: Há diferenças na variação de massa entre as rações utilizadas.

H_0: Não há diferenças na variação de massa considerando as interações entre temperatura e tipo de rações utilizadas.

H_1: Há diferenças na variação de massa considerando as interações entre temperatura e tipo de rações utilizadas.

Temperatura	*A*	*A*	*A*	*B*	*B*	*B*	*C*	*C*	*C*
Ração	I	II	III	I	II	III	I	II	III
	7,0	9,0	7,2	9,5	7,0	7,3	11,5	8,5	10,4
	7,0	5,2	7,0	7,2	7,3	7,8	10,1	8,7	10,4
	7,1	7,8	6,3	7,2	8,9	9,2	11,4	10,9	11,6
	7,2	8,2	6,0	6,6	8,8	7,1	7,2	10,8	8,2
Soma	28,3	30,1	26,6	30,5	32,0	31,5	40,1	38,8	40,6
Média	7,1	7,5	6,7	7,6	8,0	7,9	10,0	9,7	10,2

Neste experimento temos dois fatores, com três tratamentos cada, e cada tratamento com 4 repetições. Podemos representar o número de tratamentos do Fator A, como 'k', o número de tratamentos do Fator B, como 'b' e o número de repetições por 'n'.

Fontes de variação	***G.L.***	***S.Q.***	***Q.m.***	F_{calc}
Tratamentos	(kb-1)	$\sum_{i=1}^{k}\sum_{j=1}^{b} n\left(\overline{y}_{ij}-\overline{y}\right)^2$		
Fator A	k-1	$bn\sum_{i=1}^{k}\left(\overline{y}_i-\overline{y}\right)^2$	$QM_{fatorA}=\frac{SQ_A}{G.L._A}$	$F_{fatorA}=\frac{QM_A}{QM_e}$
Fator B	b-1	$kn\sum_{j=1}^{b}\left(\overline{y}_j-\overline{y}\right)^2$	$QM_{fatorB}=\frac{SQ_B}{G.L._B}$	$F_{fatorB}=\frac{QM_B}{QM_e}$
Interação (AxB)	(k-1)(b-1)	$SQ_{trat}-SQ_A-SQ_B$	$QM_{AxB}=\frac{SQ_{AxB}}{G.L._{AxB}}$	$F_{AxB}=\frac{QM_{AxB}}{QM_e}$

Fontes de variação	***G.L.***	***S.Q.***	***Q.m.***	F_{calc}
Resíduos	kb(n-1)	$\sum_{i=1}^{k}\sum_{j=1}^{b}\left(\sum_{l=1}^{n}\left(y_{ijl}-\overline{y}_{ij}\right)^2\right)$	$QM_e=\frac{SQ_e}{G.L._e}$	
Total	**N-1**	$\sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{l=1}^{n}\left(y_{ij}-\overline{y}\right)^2$		

Calculando o grau de liberdade:
Tratamentos: g.l.=(Kb-1) → (3×3)-1=8
Fator A: g.l.=K-1=3-1=2
Fator B: g.l.=B-1=3-1=2
Interação: g.l.=2×2=4
Resíduos: g.l.=3×3×(4-1)=27
Total: g.l.=36-1=35

Calculando a soma dos quadrados (SQ)
Tratamento:

Temperatura	*A*	*A*	*A*	*B*	*B*	*B*	*C*	*C*	*C*
Ração	I	II	III	I	II	III	I	II	III
Média	7,1	7,5	6,7	7,6	8,0	7,9	10,0	9,7	10,2

Média geral = 8,3

$$\sum_{i=1}^{k}\sum_{j=1}^{b}n\left(\overline{y}_{ij}-\overline{y}\right)^2=4\times((7{,}1-8{,}3)^2+(7{,}5-8{,}3)^2+\ldots+(10{,}2-8{,}3)^2)=55{,}52$$

Fator A:
Dados: médias de massa em cada temperatura: A=7,1, B= 7,8, C= 10
B=3, n=4

$$bn\sum_{i=1}^{k}\left(\overline{y}_i-\overline{y}\right)^2=3\times4\times\left(\left(7{,}1-8{,}3\right)^2+\left(7{,}8-8{,}3\right)^2+\left(10-8{,}3\right)^2\right)=\ 53{,}27$$

Fator B:

Dados: médias de massa para cada ração: I=8,2, II=8,4, III= 8,2, k=3,n=4

$$kn\sum_{j=1}^{b}\left(\overline{y}_j-\overline{y}\right)^2=3\times4\times\left(\left(8,2-8,3\right)^2+\left(8,4-8,3\right)^2+\left(8,2-8,3\right)^2\right)=0,24$$

SQinteração:

SQ_{AxB}=55,52-53,27-0,24 = 2

$$SQ_{AxB}=55,52-53,27-0,24\ =\ 2$$

SQresíduos:

$$\sum_{i=1}^{k}\sum_{j=1}^{b}\left(\sum_{l=1}^{n}\left(y_{ijl}-\overline{y}_{ij}\right)^2\right)=\ \left(7-7,1\right)^2+\left(7-7,1\right)^2+...+\left(8,2-10,2\right)^2=\ 42,86$$

SQtotal:

$$\sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{l=1}^{n}\ \left(y_{ij}-\overline{y}\right)^2=\left(7-8,3\right)^2+\ \left(7-8,3\right)^2+...+\left(8,2-8,3\right)^2=\ 98,37$$

Quadrados médios:

$$QM_{fatorA}=\frac{SQ_{fatorA}}{G.L._{fatorA}}=\frac{53,27}{2}=26,64$$

$$QM_{fatorB}=\frac{SQ_{fatorB}}{G.L._{fatorB}}=\frac{0,24}{2}=0,12$$

$$QM_{AxB}=\frac{SQ_{AxB}}{G.L._{AxB}}=\frac{2}{4}=0,5$$

$$QM_{e}=\frac{SQ_{e}}{G.L._{e}}=\frac{42,86}{27}=\ 1,59$$

Revisitando as hipóteses:

H_0: não há diferenças na variação de massa nas temperaturas analisadas.

$$F_{fatorA} = \frac{QM_{fatorA}}{QM_e} = \frac{26,64}{1,59} = 16,78$$

$F_{0,05,\ 2,\ 27}$=3,35 → rejeitamos H_0.

H_0: não há diferenças na variação de massa nas rações analisadas.

$$F_{fatorB} = \frac{QM_{fatorB}}{QM_e} = \frac{0,12}{1,59} = 0,076$$

$F_{0,05,\ 2,\ 27}$=3,35 → não rejeitamos H_0.

H_0: não há diferenças na variação de massa considerando as interações entre temperatura e tipo de rações utilizadas.

$$F_{AxB} = \frac{QM_{AxB}}{QM_e} = \frac{0,5}{1,59} = 0,315$$

$F_{0,05,\ 4,\ 27}$= 2,73 → não rejeitamos H_0.

Fontes de variação	***G.L.***	***S.Q.***	***Q.m.***	F_{calc}	F_{crit}	***Decisão***
Tratamentos	8	55,52				
Fator A	2	53,27	26,64	16,78	3,35	Rejeitamos H_0.
Fator B	2	0,24	0,12	0,08	3,35	Não rejeitamos H_0.
Interação (AxB)	4	2	0,5	0,32	2,73	Não rejeitamos H_0.
Resíduos	27	42,86	1,59			
Total	**35**	**98,37**				

Na categoria de ANOVA acima, estão sendo testadas 3 hipóteses: fator1, fator 2 e interação entre os fatores.

EXERCÍCIOS DO CAPÍTULO 6

1. Foi analisado o extrato de uma substancia alopática existente em quatro espécies de um mesmo gênero. As plantas apresentam a mesma concentração da substância? Construa a tabela resumo da análise de variância.

sp1	sp2	sp3	sp4
22	35	45	33
34	43	54	35
26	44	39	36
24	29	55	37
30	38	49	42

2. Foi registrada a concentração de CO_2 de cinco grandes cidades..teste se a variância é significativa

n	N	NE	S	SE	CE
1	370	340	396	415	402
2	389	380	386	400	377
3	360	350	370	396	382
4		356		392	394
média	373	356,5	384	400,7	388,7

3. Elabore uma questão dentro da área de seu interesse. Elabore um desenho amostral (número de amostras, tratamentos). Faça o teste e interprete os resultados.

4. Suponha que foi analisada a mortalidade do peixe paulistinha *Brachydanio rerio* em várias concentrações de TBT, após cinco dias expostos ao produto, com 10 réplicas e 20 animais em cada aquário. Quais as suas conclusões?

n	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	01	0	0	0	0
25 ng	2	0	1	2	0	0	1	0	0	1
50 ng	2	0	0	2	3	1	2	2	3	1
100ng	5	8	11	4	9	13	5	6	4	7
200 ng	20	12	15	18	19	20	16	14	20	19

5. A ANOVA fator único resultou em F=12,98. Isto é, a diferença entre os tratamentos é significativa..teste quais tratamentos são significativamente diferentes. QMerro=29

sp1	sp2	sp3	sp4
22	35	45	33
34	43	54	35
26	44	39	36
24	29	55	37
30	38	49	42

Capítulo 7

TESTES COM DUAS VARIÁVEIS

Quando nos deparamos com duas variáveis, a primeira pergunta que nos vem à mente é se elas têm alguma relação. Isto é, elas covariam quando uma variável aumenta, a outra aumenta ou diminui proporcionalmente. Para quantificar esta relação entre duas variáveis, podemos utilizar duas técnicas de análise: análises de correlação e de regressão (modelos).

O primeiro passo para verificar se há covariância entre as variáveis é fazer um gráfico de dispersão, para 'visualizar' o comportamento das mesmas.

DIAGRAMAS DE DISPERSÃO

O diagrama de dispersão consiste em colocar os pontos das variáveis num diagrama cartesiano "x" e "y", realizado antes da análise numérica dos dados. Esse diagrama permite visualizar as tendências dos pontos, tanto lineares (positivos, negativos), quanto não lineares (potencial, exponencial, polinomial), assim como a inexistência de relação. Além disso, pode-se identificar a presença de valores aberrantes.

Por exemplo, observando a figura abaixo, percebe-se que há uma relação entre os anos e o aumento da concentração de CO_2 na atmosfera. Entretanto, esta associação não é perfeita.

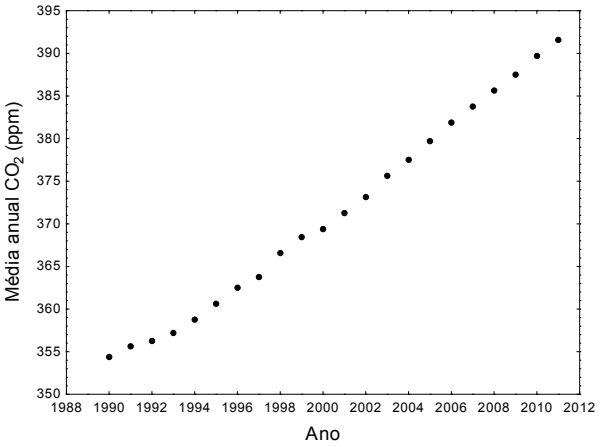


Figura 7.1: Gráfico de dispersão

A forma do gráfico indica a relação entre duas variáveis (Fig.7.1), Que neste caso é positiva, indicando que a concentração de CO_2 aumenta com o passar dos anos.

MODELOS

São descrições matemáticas, geralmente uma equação, de um fenômeno do mundo real:

A equação de crescimento de um animal, a equação de peso comprimento, relação entre metabolismo e temperatura, são alguns exemplos.

Modelos lineares

Relações lineares são aquelas em que as variáveis têm uma relação aritmética e pode ser representada graficamente por uma reta.

Exemplo: variação da concentração média anual de CO_2 atmosférico na estação de Mauna Loa Havaí (http://www.esrl.noaa.gov/gmd/ccgg/trends/).

Ano	CO_2	Ano	CO_2
1990	354,35	2001	371,13
1991	355,57	2002	373,22
1992	356,38	2003	375,77
1993	357,07	2004	377,49
1994	358,82	2005	379,8
1995	360,8	2006	381,9
1996	362,59	2007	383,77
1997	363,71	2008	385,59
1998	366,65	2009	387,38
1999	368,33	2010	389,78
2000	369,52	2011	391,57

Há várias possibilidades de calcular a inclinação da reta, uma delas é calcular a reta que passa pelo primeiro e pelo último ponto, que provavelmente não é o método mais acurado. Neste caso, a inclinação da reta seria:

$$b = \frac{\Delta CO_2}{\Delta t} = \frac{391,57 - 354,35}{2011 - 1990} = 1,772$$

E a equação seria:

$$CO_2 - 354,35 = 1,772(ano - 1990) \rightarrow CO_2 = 354,35 + 1,772(ano - 1990)$$

O modelo do nosso exemplo é linear ($f(x)= \beta_0 + \beta_1 x$), entretanto, existem muitas outras funções não lineares:

Modelos não lineares

Função	Equação
Polinomial	$P(x) = \beta_n x^n + \beta_{n-1} x^{n-1} + ... + \beta_1 x^1 + \beta_0$
Potencial	$f(x) = \beta_0 x^{\beta_1}$
Racional	$f(x) = \frac{P(x)}{Q(x)} f(x) = 3x^2 - \left[\left(5x^2 + 1\right) / 2x\right] + 5$
Algébrica	$x^2 + 1,\ x^5 + 8x^3 + \sqrt{3}x$
Trigonométricas	Sem (x)= cateto oposto/ hipotenusa
Exponencial	$f(x) = a^x$
Logaritmica	$f(x) = log_a x$

Polinômios

Um polinômio pode ser descrito pela equação:

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x^1 + a_0$$

Onde n é um número inteiro não negativo e os número a_0, a_1, ..., A_n são as constantes do polinômio. Assim, o polinômio de grau 1 é uma função linear: $P(x)= a_0+a_1x$

A função quadrática é um polinômio de grau 2: $P(x)=ax^2+bx+c$

A função cúbica, grau 3: $P(x)=ax^3+bx^2+cx+d$

Os polinômios servem de modelos para descrever vários processos naturais e sociais.

Funções potenciais

Têm a forma de $f(x)= x^n$

A forma geral do gráfico potencial depende se n é par ou impar, o gráfico par será semelhante a uma parábola e o impar será semelhante à função cúbica.

Se n for um número racional (1/a), a função será uma raiz, cujo domínio é $[0,\infty)$

Se n=-1, o gráfico será uma hipérbole

Funções racionais

Uma função racional é a razão de dois polinômios

$$f(x) = \frac{P(x)}{Q(x)}$$

Função algébrica

É aquela função que contêm operações algébricas. Toda função racional é automaticamente uma função algébrica.

Ex.: X^2+1, $x^5+8x^3+\sqrt{3}x$

Funções trigonométricas

Geralmente usa-se a medida em radianos com domínio $(-\infty,\infty)$ e variação $[-1,1]$

São funções adequadas à descrição de fenômenos repetitivos, como ondas, variações sazonais.

Funções exponenciais

São funções da forma $f(x)= a^x$, onde a base 'a' é uma constante positiva. Note que todos os gráficos passam pelo mesmo ponto (0,1), pois $a^0=1$, para $a\neq0$

"e" → é a base da função exponencial, cuja reta tangente a esta curva tem inclinação igual a 1.

Funções logarítmicas

São as funções $f(x)=\log_a x$, onde a base 'a' é uma constante positiva. Elas são as inversas das funções exponenciais.

CORRELAÇÃO

Quando queremos verificar com qual intensidade duas variáveis variam uma em relação à outra, estimamos o coeficiente de correlação, que avalia o grau de associação entre elas. Ele tem a vantagem de ser um número adimensional.

O coeficiente mais utilizado é o coeficiente de correlação linear de Pearson (r). Este coeficiente quantifica a linearidade entre duas variáveis, descrevendo o quanto uma linha reta se ajusta através da nuvem de pontos. Se os pontos caem exatamente sobre uma linha crescente então r=1 e se eles caem exatamente sobre uma linha decrescente, r= –1.

o teste não paramétrico de correlação mais utilizado é o teste de correlação de postos de Spearman.

Os dois testes são representados por "r":

- "r"varia entre -1 e +1
- "r"=0 corresponde a ausência de correlação
- Quanto maior o valor de R mais forte a associação
- "r">0 significa variáveis diretamente proporcionais,
- "r"<0 corresponde a variáveis inversamente proporcionais.

Coeficiente de Pearson

O coeficiente de correlação de Pearson pode ser visto como a razão entre a covariância de duas variáveis (cov(x,y)) pelo produto dos desvios-padrão de cada uma delas. Ou seja:

$$r = \frac{\mathrm{cov}(x, y)}{s_x \times s_y}$$

Covariância: é o valor médio do produto dos desvios de x e y, em relação às suas respectivas médias. Isto é, mede a variação concomitante das duas variáveis. $\sum(X-\bar{x})(y-\bar{y})$. A covariância pode ser positiva ou negativa, indicando o sentido da correlação.

A equação utilizada para estimar a correlação de Pearson é:

$$r = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sqrt{\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]}}$$

Suposições da análise de correlação

Não há suposições estatísticas necessárias para realizar a análise de correlação. Mas, há suposições implícitas ao teste de hipóteses e determinação dos intervalos de confiança para os coeficientes de correlação.

Para o teste de correlação, assumimos que X e Y possuem distribuição normal, e que os valores de X e Y foram obtidos aleatoriamente da população.

Resumindo, para fazer um teste de significância da correlação corretamente, deve-se verificar que:

1. Tanto a variável x quanto a y têm distribuição normal.
2. Que as variáveis sejam homocedásticas

Exemplo 1: qual a relação entre comprimento e largura nas conchas de *anomalocardia brasiliana*.

n	Comprimento	Largura (mm)	n	Comprimento	Largura (mm)
1	23,2	27,69	**7**	4,43	5,5
2	23,17	27,99	**8**	4,69	5,86
3	22,51	23,87	**9**	4,18	5,5
4	10,63	12,62	**10**	19,79	22,98
5	6,53	8,72	**11**	10,47	13,17
6	9,17	11,91	**12**	5,75	7,65

N	*Comprimento X*	*Largura Y*	X^2	Y^2	*XY*
1	23,2	27,69	538,24	766,74	642,41
2	23,17	27,99	536,85	783,44	648,53
3	22,51	23,87	506,70	569,78	537,32
4	10,63	12,62	113,00	159,26	134,15

N	*Comprimento X*	*Largura Y*	X^2	Y^2	*XY*
5	6,53	8,72	42,64	76,04	56,94
6	9,17	11,91	84,09	141,85	109,22
7	4,43	5,5	19,62	30,25	24,36
8	4,69	5,86	22,00	34,34	27,48
9	4,18	5,5	17,47	30,25	22,99
10	19,79	22,98	391,64	528,08	454,77
11	10,47	13,17	109,62	173,45	137,89
12	5,75	7,65	33,06	58,52	43,99
Soma	**144,52**	**173,46**	**2414,94**	**3352,00**	**2840,05**

$$r = \frac{2840,05 - \frac{144,52 \times 173,46}{12}}{\sqrt{\left(2414,94 - \frac{(144,52)^2}{12}\right)\left(3352,00 - \frac{(173,46)^2}{12}\right)}} = 0,995$$

Significância da análise de correlação

Para saber se o "r" tem significância estatística, podemos utilizar o teste 't', cuja equação geral é:

$$t = \frac{parâmetro\ estimado - parametro\ hipotetico}{desvio\ \ padrão\ estimativa}$$

Queremos saber se o 'r' estimado é significativamente diferente de zero (ausência de correlação). Se não houver correlação entre x e y , o parâmetro hipotético (ρ) será igual a zero. Isto é, se a relação entre x e y for aleatória, não tiver nenhuma tendência, os valores de "r", na maioria, serão próximos de zero. Por outro lado, se houver correlação entre elas, os valores serão significativamente diferente de zero. Para avaliar a significância do coeficiente de correlação, pode-se testar a hipótese nula (H_0) de que ρ=0, utilizando-se a distribuição t.

O teste de significância da correlação, segue a mesma sequencia de passos que os testes dos capítulos anteriores:

1. Elaboração das hipóteses

   H_0: ρ=0

   H_A: ρ≠ 0
2. Escolha do nível de significância → α = 0,05
3. Determinação do valor crítico do teste:

   $T_{\alpha, gl} = t_{0,05,4} = 2,776$ (gl= n-2, onde n é o número de pares de valores x e y)
4. Determinação do valor calculado de t:

$$t_{calc} = \frac{r - \rho}{EP_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

ρ desaparece na equação porque se supõe que ρ=0

$$t_{calc} = \frac{0,995}{\sqrt{\frac{1 - 0,995^2}{12 - 2}}} = \frac{0,995}{0,032} = 31,50$$

5. Decisão: Como t_{calc}= 31,50> t0,05,10=2,228, rejeita-se H_0.

Há correlação significativa entre largura e comprimento.

Avaliação qualitativa de R quanto à intensidade

|r|	A correlação é dita:
0	Nula
0—0,3	Fraca
0,3|— 0,6	Regular
0,6|— 0,9	Forte
0,9|—1	Muito forte
1	Plena ou perfeita

Coeficiente de determinação

O coeficiente de determinação é o quadrado do coeficiente de correlação e informa que fração da variabilidade de uma característica é explicada estatisticamente pela outra variável.

Para os dados de batimentos cardíacos, o coeficiente de determinação é:

$$r^2 = 0,995^2 = 0,99$$

Isto significa que 99% da variação observada no comprimento é explicada pela largura.

Considerações sobre o uso do coeficiente de correlação

O coeficiente de correlação de Pearson NÃO É NEM ROBUSTO NEM RESISTENTE. Não é robusto porque relações fortes que não forem lineares entre as duas variáveis x e y podem não ser reconhecidas. Não é resistente, uma vez que é EXTREMAMENTE SENSÍVEL a um ou a poucos pares de dados aberrantes. Não há necessidade de satisfazer nenhum pressuposto para calcular o coeficiente de correlação.

Mas, nem sempre o coeficiente de correlação de Pearson é o parâmetro mais adequado para medir a associação entre duas variáveis. Por isto a importância de se examinar o gráfico de dispersão dos pontos antes de se efetuar uma análise de correlação.

Vale lembrar que a correlação indica uma associação, não uma relação de causa e efeito. As relações entre duas variáveis, sem significado real são chamadas de relações espúrias; elas são comuns, como relações entre redução no consumo de margarina e aumento de divórcios, diminuição do tamanhos das saias e crescimento econômico

A significância de uma correlação está relacionada com o tamanho da amostra. Em uma amostra grande (n=1000), uma correlação baixa (r=0,20) pode ser estatisticamente significativa, embora seja uma associação muito fraca entre as variáveis.

Comparando dois coeficientes de correlação

Podemos testar se dois coeficientes de correlação são significativamente diferentes através da equação:

$$Z = \frac{z_1 - z_2}{\sigma_{z1-z2}} \quad \text{onde} \quad \sigma_{z1-z2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

Para calcular z → $z = 0,5 \ln\left(\frac{1+r}{1-r}\right)$

Exemplo: a relação entre o comprimento da quela e do cefalotórax em caranguejos do gênero *Uca* de duas populações foram r = 0,788 e n=39 e r = 0,849 e n = 30, respectivamente. As duas correlações são significativamente diferentes?

$$z_1 = 0,5 \ln\left(\frac{1+0,788}{1-0,788}\right) = 1,066, \quad z_2 = 0,5 \ln\left(\frac{1+0,849}{1-0,849}\right) = 1,253$$

$$\sigma_{z1-z2} = \sqrt{\frac{1}{39-3} + \frac{1}{30-3}} = 0,255 \rightarrow Z = \frac{1,066 - 1,253}{0,255} = 0,73333$$

$Z_{0,05} = t_{0,05} = 1,96$
Como $z_{calc} < z_{crit}$ não rejeitamos H_0.

COEFICIENTE DE CORRELAÇÃO DE POSTOS DE SPEARMAN

O coeficiente de Spearman, foi desenvolvido em 1904 e exige que as variáveis tenham sido medidas, pelo menos, em escala ordinal. Ele também pode ser empregado quando as variáveis quantitativas não satisfazem as exigências para o teste do R de Pearson, como linearidade, distribuição bivariada normal e homocedasticidade.

O coeficiente de Spearman (r_s) varia de -1 a +1, similar ao R de Pearson.

Realizando o teste de Spearman

1. Ordenam-se os valores de x e y em escala ascendente. Se houver empate, faça a média dos postos empatados.
2. Calcula-se d diferença entre os postos determinados para cada variável em cada linha da tabela.

3. Elevam-se ao quadrado as diferenças e as somas.
4. Calcula-se r_s:

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

5. Quando houver um número elevado de empates, o valor de rs pode ser afetado. Deve-se usar, então a seguinte equação para corrigir os empates:

$$r_s = \frac{A_x + A_y - \sum d^2}{2\sqrt{A_x A_y}}$$

onde: $A = \dfrac{\left(n^3 - n\right) - \sum\left(t^3 - t\right)}{12}$

E t é o número de empates em cada posto.

Teste de significância de r_s

A hipótese nula do teste para r_s estabelece que a correlação de Spearman é zero na população. Para amostras pequenas (até 100) pode-se comparar o r_s calculado com o da tabela.

Para amostras maiores do que 100, realize o teste t, do mesmo modo que se testa o R de Pearson.

$$t_{calc} = \frac{r_s}{\sqrt{\dfrac{1 - r_s^2}{n - 2}}}$$

Exemplo:

Um produtor quer saber se há relação entre a produção de cada caixa de mel com a distância do rio de sua propriedade.

Distância (m) X	Y Mel (ml/ abelha)	Posto de X	Posto de Y	D	D^2
23	27	3	4	-1	1
43	45	8	8	0	0
36	48	6,5	9	-2,5	6,25
17	27	1	4	-3	9
25	16	4	2	2	4
47	38	9	7	2	4
36	37	6,5	6	0,5	0,25
26	27	5	4	1	1
20	10	2	1	1	1
63	51	10	10	0	0
			Σ	0	26,5

1. Atribua postos às variáveis X e Y.
2. Calcule a diferença, D, e o quadrado da diferença, D^2.
3. Calcule r_s:

$$r_s = 1 - \left[\frac{6\sum d^2}{\left(n^3 - n\right)}\right] \rightarrow r_s = 1 - \frac{6 \times 26,5}{1000 - 10} \rightarrow r_s = 0,8394$$

4. faça o teste $H_0 = r_s \leq r_{tabela(\alpha, 2, n)}$

$H_1 = r_s > r_{tabela}$

$R_{(\alpha=0,05, 2, 10)} = 0,648$

Decisão: $r_s > r_{tabela}$, rejeito H_0, há uma correlação significativa entre quantidade de mel e distância do rio.

EXERCÍCIOS DO CAPÍTULO 7

1. No site da agência de meteorologia dos oceanos de Mauna Loa no Havaí há dados de CO_2 (http://www.esrl.noaa.gov/gmd/ccgg/trends/). Utilize dados do site para fazer um gráfico de dispersão da média anual de CO_2. Se o gráfico apresentar tendencia linear, faça a correlação.

2. Verifique se há correlação entre horas de exposição ao ar e concentração de ácido lático em ostras vendidas em um mercado. Este resultado é significativo?

Horas (X)	8	7	6	3	3	6	5	2
Ácido lático (Y)	10	8	4	8	6	9	7	4

3. Um pesquisador quer saber se há relação entre o número de árvores mortas por hectare e o número de ninhos de pica-pau. Determine e interprete os coeficientes de correlação e de determinação.

Árvores	9,68	9,81	9,59	9,68	9,84	9,59	9,61	9,55	9,25	9,08	9,2
Ninhos	6,53	6,71	6,7	6,69	6,7	6,62	6,59	6,55	6,35	6,25	6,22

4. Pergunte a pressão sistólica e idade de conhecidos seus e verifique se há correlação entre idade e pressão.

5. Um estudo visa determinar se há relação entre a abundância de *Capitella capitata* e a concentração de matéria orgânica no sedimento. Foram identificados 12532 poliquetas em 8 estações ao longo de um gradiente de poluição. A matéria orgânica foi medida em mgC/g de sedimento. A análise de relação entre essas duas variáveis mostrou um coeficiente de correlação de Pearson de R = – 0,63. Essa correlação é significativa ao nível de 5%?

6. Dois pesquisadores entrevistaram os mesmos alunos para uma seleção do mestrado. Há correlação entre as avaliações?

A	147	158	131	142	183	151	196	129	155	158
B	122	128	125	123	115	120	108	143	124	123

7. Um pesquisador testou dois bafômetros (de empresas diferentes que usam biosensores distintos) em 10 pessoas abordadas aleatoriamente na estrada. Qual a relação entre os resultados obtidos pelos dois aparelhos.

A (g/l)	0,53	1,25	0,33	0,08	2,05	1,88	0,06	0,88	0,7	0,3
B (g/l)	0,35	0,9	0,03	0,1	2,8	1,3	0,01	0,9	0,5	0,4

8. Um gerente selecionou os novos funcionários da empresa. Há uma suspeita que os mais jovens foram favorecidos. Verifique se isto ocorreu.

Idade	18	25	41	19	37	22	52	35	44	25
Classificação	2	6	9	1	8	3	9	7	9	4

9. Foi analisada a relação entre o número de batimentos cardíacos e o consumo de oxigênio para duas populações de ratos, com os resultados r=0,73 (n=36) e r=0,59 (n=19). Os resultados podem ser considerados similares?

10. Na sequência, a tabela mostra dados de intensidade de irrigação e crescimento das plantas..teste a significância da correlação.

Irrigação	6	30	29	65	70	99	140	135	303	304	300	368
Biomassa	265	273	275	285	298	303	308	315	238	359	383	400

Capítulo 8

MODELOS DE REGRESSÃO

Quando existe uma relação funcional entre duas variáveis, podemos aplicar um modelo de regressão. Assim, uma das variáveis será a dependente (Y, que responde à outra, também chamada variável resposta) e outra será independente (X, fator, variável explicativa ou variável preditiva). Isto é, uma (Y) está em função da outra (X). A função dependente pode ser claramente identificada como no caso de tempo de desenvolvimento e comprimento de um peixe. Outras vezes esta relação não é tão clara, como a relação entre o comprimento e a largura de uma espécie de bivalve.

Os modelos são extremamente úteis em todas as áreas da ciência, pois possibilitam a comparação e predição. Por exemplo, podemos **comparar** a relação peso comprimento de recém nascidos entre duas populações e testar se as duas populações apresentam o mesmo padrão. Também podemos **predizer** qual será a altura daquele recém nascido após seis meses da última medição.

O termo 'regressão' foi utilizado, pela primeira vez, por Francis Galton, em 1886, tentando explicar porque pais de estatura mais alta tinham, em média, filhos de estatura mais baixa que a deles e vice-versa.

As relações entre duas variáveis podem ser aritméticas, geométricas, exponenciais, quadráticas, entre outras. Estas relações podem ser observadas através de gráficos de dispersão e expressas matematicamente:

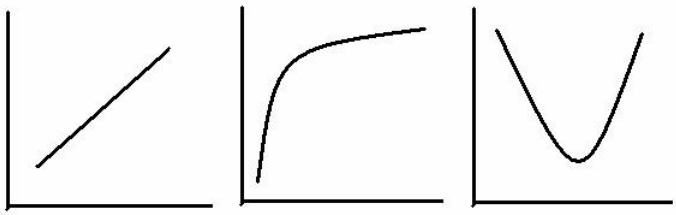


Figura 8.1: linhas de tendência, exemplificando relações lineares, potenciais e quadráticas, respectivamente.

REGRESSÃO LINEAR SIMPLES

Os objetivos da regressão são

- Avaliar a possível dependência linear de y em relação a x.
- Expressar matematicamente esta relação por meio de uma equação

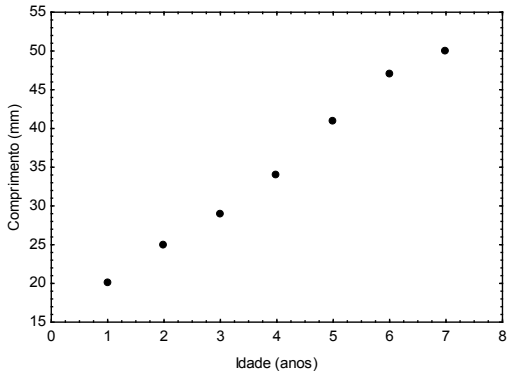
Gráfico de dispersão

O primeiro passo é saber como os pontos estão dispostos, uns em relação aos outros.

Regressão linear simples

Exemplo 1: foi analisada a idade e comprimento de uma lagosta, criada em cativeiro.

Idade X (anos)	*Comprimento Y*
1	20
2	25
3	29
4	34
5	41
6	47
7	50



Observando-se o diagrama de dispersão, podemos assumir que a relação entre idade e comprimento da lagosta é **linear**. Neste caso, podemos ajustar uma regressão linear simples.

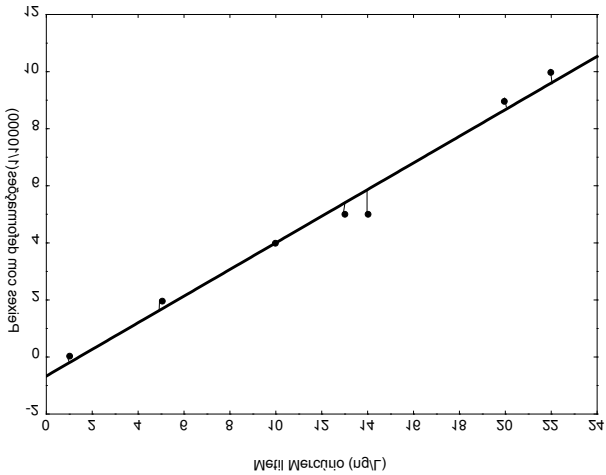
Se a reta ajustada representar satisfatoriamente a relação entre as duas variáveis, podemos dizer que o comprimento da lagosta é função linear da idade. Também podemos predizer a idade da lagosta a partir de determinado comprimento.

AJUSTE DA RETA

Para determinar a melhor reta que passa pelos pontos do diagrama de dispersão, fazemos o ajuste da mesma pelo método dos mínimos quadrados. Isto é, escolheremos a reta com a menor soma dos quadrados dos resíduos $(y-\hat{y})^2$. Ou seja:

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2,$$

Onde y_i é a variável dependente observada e $\hat{y}$ é a variável estimada a partir do modelo utilizado (no caso, a regressão linear).



ESTIMANDO OS PARÂMETROS DA RETA

$$Y=\beta_0+\beta_1 X \qquad \textbf{eq. 8.1}$$

β_0= é o intercepto, isto é, o valor de "Y" quando X=0, sua unidade de medida é a mesma que Y.

β_1 = é a inclinação da reta, ou tangente θ, é uma taxa, sua unidade de medida é $\Delta x/\Delta y$

O coeficiente de regressão, β_1, varia de $-\infty \leq \beta_1 \leq \infty$. Ele expressa a magnitude de mudança de y associado com uma unidade de X.

Ao calcular "β_0" e "β_1", estimamos parâmetros populacionais. Entretanto, estimativas incluem erros. Todo modelo tem um erro associado, pois nenhum modelo é completo o suficiente para incluir todos os fatores que influenciam a variável, sejam elas variáveis desconhecidas, ou não

estudadas. Devemos levar em conta também que todo modelo é inferido a partir de resultados conhecidos, o que não é conhecido, não faz parte do modelo. O erro não pode ser removido, compõe a parte não explicada do modelo.

Assim, estamos analisando o comportamento de uma variável (Y) em relação a outra (X), que é a variável explicativa ou resposta. As outras variáveis, não mensuradas e qualquer outra variação dos parâmetros que afetam a variável dependente, são denominadas erro. Por isso uma forma mais correta de representar a equação da reta seria:

$$Y = \beta_0 + \beta_1 X + \varepsilon_k \qquad \textbf{eq. 8.2}$$

Onde ϵ seria o erro ou resíduo, que é um desvio dos valores de y.

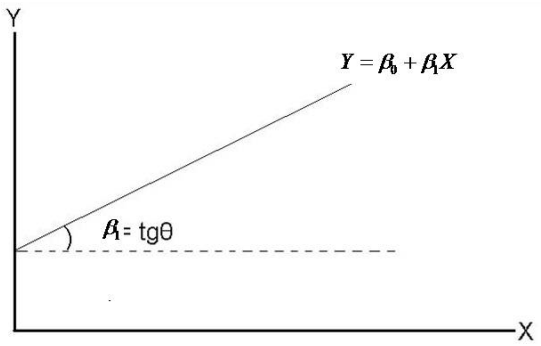


Figura 8.2: modelo de regressão linear.

Estes valores podem ser obtidos pelo método dos mínimos quadrados:

$$\beta_1 = \frac{SQ_{xy}}{SQ_x} = \frac{\sum xy}{\sum x^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \qquad \textbf{eq. 8.3}$$

O método dos mínimos quadrados é sensível a dados discrepantes (marginais), porque dá maior peso aos resíduos maiores. Por exemplo, o resíduo 1 contribui com 1^2=1 para a soma dos quadrados dos resíduos (SQR) e 2 contribui com 2^2=4 para a SQR. Por isto, a equação mais utilizada é:

$$\beta_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

eq. 8.4

Cálculo da inclinação da reta (β_1).
Exemplo:

n	Idade (x)	Comprimento(y)	xy	x^2
1	1	20	20	1
2	2	25	50	4
3	3	29	87	9
4	4	34	136	16
5	5	41	205	25
6	6	47	282	36
7	7	50	350	49
Soma	**28**	**246**	**1130**	**140**

$$\beta_1 = \frac{1130 - \frac{28 \times 246}{7}}{140 - \frac{28^2}{7}} = 5,21$$

Cálculo do intercepto (β_0)

Existem muitas linhas paralelas com a mesma inclinação, mas apenas uma linha cruza o eixo y, isto ocorre quando "x=0", que é o ponto de interseção. Assim denominamos "β_0" de intercepto:

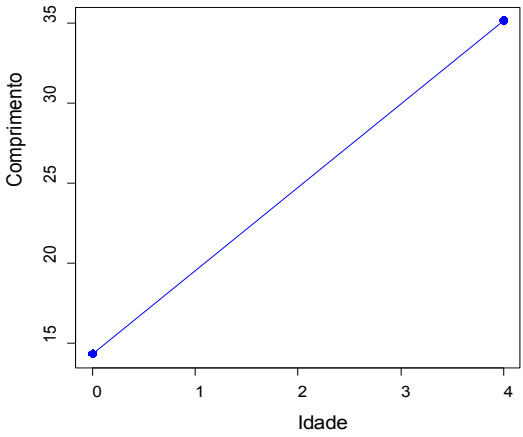
$$\beta_0 = \overline{y} - b\overline{x} \qquad \beta_0 = 35,14 - 5,21 \times 4 \qquad \beta_0 = 14,29$$

Onde $\overline{x}$ e $\overline{y}$, são as médias de x e y respectivamente.

Traçando a reta

Depois de calculados "β_0" e "β_1" , podemos traçar a reta. Para isto, precisamos de apenas dois pontos:

Temos o ponto do intercepto (β_0): (0, β_0) e o ponto médio $(\overline{x}, \overline{y})$: (0; 14,29); (4; 35,14)



Podemos calcular vários outros pontos calculando o y previsto $(\hat{y})$.

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\hat{y} = 14,29 + 5,21 \times 1 \rightarrow 19,5$$

n	X	y	ŷ
1	1	20	19,5
2	2	25	24,7
3	3	29	29,9
4	4	34	35,1
5	5	41	40,4
6	6	47	45,6
7	7	50	50,8
Soma	**28**	**246**	

Estimando o erro

Quanto menor o erro maior o ajuste dos pontos à reta. Se a variância for zero ($\sigma^2=0$) todos os pontos estarão sobre a linha de regressão. Esta descrição é similar à explicação da soma dos quadrados dos resíduos. Entretanto, a variância mede o desvio médio de cada observação em relação à média. Assim, a variância do erro da regressão será o desvio médio de cada observação em relação ao valor ajustado, isto é, o **erro padrão da regressão.**

$$\sigma^2 = \frac{SQR}{n-2} = \frac{\sum(y-\hat{y})^2}{n-2} \text{ ou } \sigma^2 = \frac{\sum\left[y-(\beta_0+\beta_1 x)\right]^2}{n-2} \qquad \textbf{eq. 8.5}$$

n	Idade (x)	Comprimento(y)	$\bar{y}$	$(y-\bar{y})^2$
1	1	20	19,5	0,25
2	2	25	24,7	0,08
3	3	29	29,9	0,9
4	4	34	35,1	1,3
5	5	41	40,4	0,4
6	6	47	45,6	2,0
7	7	50	50,8	0,6
Soma	**28**	**246**		**5,57**

$$\sigma^2 = \frac{5,57}{7-2} = 1,11$$

Teste de hipóteses para o intercepto da reta (β_0)

$t = \frac{(a-a_0)}{s_a}$, O a_0 **é determinado pelo pesquisador**

$$s_a = \sqrt{\frac{\sum(y_i - a - bx_i)^2}{n-2} \times \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)} \qquad \textbf{eq. 8.6}$$

Teste de hipóteses para a inclinação da reta (β_1)

Os programas existentes para análise de regressão apresentam, nos resultados, teste de variância e teste t. A análise de variância, testa se a regressão é significativa, isto é fundamental em análise de regressão múltipla (onde há mais de uma variável explicativa), para testar se, pelo menos uma das variáveis independentes apresenta regressão significativa com a variável dependente. O teste t, verifica se a inclinação de cada variável é significativamente diferente de zero.

Se as variáveis x e y não tiverem qualquer relação, a inclinação da reta é, evidentemente, zero. Para testar isto, faz-se o teste:

1. Elaboração da hipótese:
   $H_0 = \beta_1 = 0$
   $H_1 = \beta_1 \neq 0$
2. Nível De Significância → $\alpha = 0{,}05$
3. Estimar o valor calculado do teste:
4. Estimar O Valor Crítico Do.teste
5. Decisão

Os passos 1 e 2 são iguais para os dois testes

Análise de variância da regressão

3. Estimar o valor calculado do teste:

Testar a variabilidade total da variável dependente, computando a soma do quadrado dos desvios. A análise de variância da regressão é realizada pelos programas estatísticos para testar se, pelo menos, uma das variáveis independentes possui relação linear com a variável dependente.

A tabela a seguir organiza os dados da análise de variância. Na primeira coluna tem as fontes de variação, que são apenas duas a regressão e o erro. Na segunda coluna estão os graus de liberdade, o primeiro associado ao número de variáveis independentes e 'n-2' associado ao erro. As somas dos quadrados de cada variação estão na terceira coluna. Na quarta estão os quadrados médios. A quinta coluna mostra o F de Fisher que é a razão entre os quadrados médios.

Causas Variação	*Graus de Liberdade (g.l.)*	*SQ (soma dos Quadrados)*	*QM (quadrado médio)*	*F*
Regressão	1	$SQ_{Reg} = \sum \left(y - \overline{y} \right)^2$	$QM_{Reg} = \frac{SQ_{Reg}}{g.l._{Reg}}$	$F = \frac{QM_{Reg}}{QM_{Res}}$
Desvio da regressão	n-2	$SQ_{Res} = \sum \left(y - \hat{y} \right)^2$	$QM_{Res} = \frac{SQ_{Res}}{g.l._{Res}}$	
Total	n-1	$SQ_{total} = \sum \left(y - \overline{y} \right)^2$		

$$SQ_{total} = \sum \left(y - \overline{y} \right)^2 \rightarrow SQ_{total} = \sum y^2 - \frac{\left(\sum y \right)^2}{n}$$

$$SQ_{Reg} = \sum \left(\hat{y} - \overline{y} \right)^2 \rightarrow SQ_{Reg} = \frac{\left(\sum xy - \frac{\sum x \sum y}{n} \right)^2}{\sum x^2 - \frac{\left(\sum x \right)^2}{n}} \rightarrow SQ_{Reg} = \beta_1 \sum xy$$

$$SQ_{Res} = \sum \left(y - \hat{y} \right)^2 \rightarrow SQ_{Res} = \sum y^2 - \beta_0 - \beta_1 \sum xy$$

Dados do exemplo acima:

n	Idade (x)	Comprimento (y)	ŷ	$(\hat{y}-\overline{y})^2$	$(y-\hat{y})^2$	$(y-\overline{y})^2$
1	1	20	19,5	244,7	0,25	229,3
2	2	25	24,7	108,7	0,08	102,9
3	3	29	29,9	27,2	0,9	37,7
4	4	34	35,1	0	1,3	1,3
5	5	41	40,4	27,2	0,4	34,3
6	6	47	45,6	108,8	2,0	140,6
7	7	50	50,8	244,7	0,6	220,7
SOMA	**28**	**246**		**761,3**	**5,57**	**766,9**
Média	**4**	**35,14**				

	gl	*SQ*	*QM*	*F*	*Valor de p*
Regressão	1	761,3	761,3	683,20	$1{,}53*10^{-6}$
Resíduo	5	5,57	1,11		
Total	6	766,9			

4. Estimar o valor crítico do teste, isto é: o f crítico

Numerador =1 (para regressões simples)

Denominador → g.l.= N-2= 5

$F_{(1,5)}$= 10,0

5. Decisão

Analise de variância

$$F = \frac{QMregressão}{QMdesviosregressão} = 683,20$$

Com isto $F_{calc} > F_{crit.}$ Rejeitamos H_0, o b é significativamente maior que zero.

Teste t

3. Estimar o valor calculado do teste:

$$t = \frac{parâmetro\ estimado - parametro\ hipotetico}{desvio\ \ padrão\ estimativa} \rightarrow t = \frac{\beta_1}{s(\beta_1)}$$

Na equação acima o parâmetro hipotético é zero, isto é se deseja saber se β_1 é diferente de zero, quando não haveria relação entre x e y.

Estimativa da variância de β_1:

$$s(\beta_1) = \sqrt{\frac{\sum (y-\hat{y})^2}{(n-2)\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)}} \rightarrow \text{como } s^2 = \frac{S.Q.\text{Re}siduos}{n-2} \text{ ou } s^2 = \frac{\sum (y-\hat{y})^2}{n-2}$$

A equação resultante é:

$$s(\beta_1) = \sqrt{\frac{s^2}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$$

Outra forma da equação é:

$$s(\beta_1) = \sqrt{\frac{\sum y^2 - \beta_0 \sum y - \beta_1 \sum xy}{(n-2)\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)}}$$

Sabendo-se que idade(x) e comprimento (y):

X	Y	X²	Y²	XY	ŷ	(y-ŷ)	(y-ŷ)²
1	*20*	1	400	20	*19,5*	*0,5*	*0,25*
2	*25*	4	625	50	*24,7*	*0,28*	*0,08*
3	*29*	9	841	87	*29,9*	*-0,93*	*0,9*
4	*34*	16	1156	136	*35,1*	*-1,14*	*1,3*
5	*41*	25	1681	205	*40,4*	*0,64*	*0,4*
6	*47*	36	2209	282	*45,6*	*1,43*	*2,0*
7	*50*	49	2500	350	*50,8*	*-0,79*	*0,6*
28	***246***	140	9412	1130			***5,57***

S^2=5,57/5= 1,11

Esta variância é a medida da variabilidade de y não explicada pelo modelo.

Para se ter uma idéia da variabilidade de b, precisamos calcular o desvio padrão para b.

$$s(\beta_1) = \sqrt{\frac{s^2}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}} \qquad s(\beta_1) = \sqrt{\frac{1{,}11}{140 - \frac{(28)^2}{7}}} \rightarrow s(b) = 0{,}199$$

Ou:

$$s(\beta_1) = \sqrt{\frac{\sum y^2 - \beta_0 \sum y - \beta_1 \sum xy}{(n-2)\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)}} \rightarrow s(\beta_1) = \sqrt{\frac{9412 - 14,29 \times 246 - 5,21 \times 1130}{5\left(140 - \frac{28^2}{7}\right)}} = 0,199$$

4. Estimar o valor de t crítico:

O valor crítico de t, tirado da tabela com α = 0,05 e seis graus de liberdade, encontramos o valor de 2,57.

5. Decisão:

$$t = \frac{\beta_1}{s(\beta_1)}$$

Com n-2 graus de liberdade:

$$t = \frac{5,21}{0,199} \rightarrow t = 26,19$$

Com isto $T_{calc} > T_{crit.}$ Rejeitamos H_0, o β_1 é significativamente maior que zero.

Coeficiente de determinação

O coeficiente de determinação não é uma medida da adequação do modelo, embora seja uma medida *da qualidade do ajuste*. Ele dá a proporção da variação total explicada pelo modelo, sendo adimensional possibilita comparar com outros resultados, não necessariamente similares.

Para saber qual a contribuição de X para a previsão de Y, podemos utilizar o R^2.

$$R^2 = \frac{SQ_{Reg}}{SQ_{Total}} = \frac{SQ_{Reg}}{SQ_{Reg} + SQ_{Res}} \quad \text{ou} \quad R^2 = \frac{\sum (\hat{y} - \overline{y})^2}{\sum (y - \overline{y})^2}$$

O menor valor que R^2 pode assumir é zero. Isso acontece quando SQregressão= zero, ou seja, não existe relação linear entre x e y. O maior valor que R^2 pode assumir é 1.

Assim:

$$r^2 = \frac{761,3}{766,9} = 0,99$$

O coeficiente de determinação ajustado é corrigido pelos graus de liberdade.

$$R^2_{aj} = 1 - \left[\frac{n-1}{n-K-1} \left(1 - R^2\right) \right]$$

Onde n é o número de dados e k o número de variáveis, sendo 1 no caso de correlações simples →

$$R^2_{aj} = 1 - \left[\frac{n-1}{n-2} \left(1 - R^2\right) \right] \rightarrow R^2_{aj} = 1 - \left[\frac{7-1}{7-2} (1 - 0,99) \right] = 0,988$$

Intervalos de confiança na regressão

intervalo confiança = estatística ± (t) (desvio padrão da estatística)

Intervalo de confiança para o β_1:

$$\beta_1 \pm t_{(\alpha, n-2)} (s_{\beta_1})$$

$$5,21 \pm 2,57 \times 1.05 \rightarrow 5,21 \pm 2,71$$

Suposições do modelo de regressão linear

1. A relação entre as duas variáveis é linear.
2. A variância de Y é constante, homocedástica.
3. As amostras são independentes, e os dados foram obtidos ao acaso na população.
4. Pra qualquer valor de X os valores de y têm distribuição normal. Isto significa que os desvios (y-ŷ) têm distribuição normal, que pode ser observado na análise de resíduos.

Análise dos resíduos

Para saber se a regressão segue os pressupostos acima, devemos fazer um diagnóstico da mesma. A análise de resíduos é a ferramenta mais importante

para fazer tais verificações. O resíduo de cada observação é a diferença entre o valor observado na amostra e o valor previsto pelo modelo.

$$e_i = y_i - \hat{y}_i$$

Estes resíduos devem ter distribuição normal, ter média zero, variância constante, serem independentes.

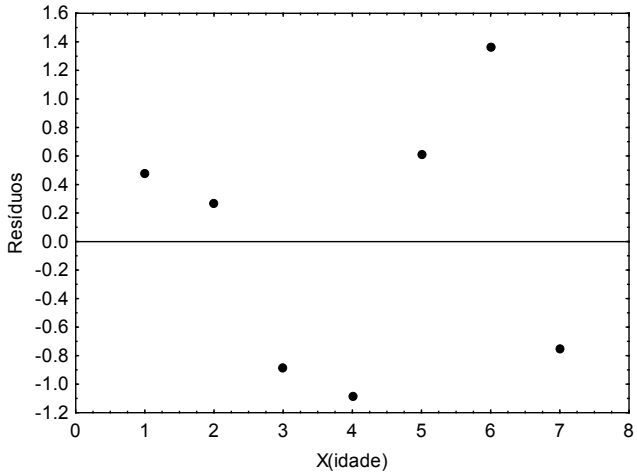
A análise de resíduos tem o objetivo de identificar se o modelo é adequado (**linear**, no caso), se os erros possuem variação **homocedástica**, se os erros são **independentes**, se há pontos marginais (outliers) e se os erros têm distribuição **normal**.

A análise de resíduos é feita por meio de gráficos.

Cálculo do resíduo:

Resíduo = (y-ŷ), resíduo padronizado = (y-ŷ)/s

Idade (x)	**Comprimento (y)**	**Ŷ**	**(y-ŷ)**	**(y-ŷ)/1,05**
1	*20*	*19,5*	*0,5*	*0.48*
2	*25*	*24,7*	*0,28*	*0.27*
3	*29*	*29,9*	*-0,93*	*-0.88*
4	*34*	*35,1*	*-1,14*	*-1.09*
5	*41*	*40,4*	*0,64*	*0.61*
6	*47*	*45,6*	*1,43*	*1.36*
7	*50*	*50,8*	*-0,79*	*-0.75*
28	***246***			



- Independência dos erros- as variáveis respostas Y são independentes.
- Normalidade dos erros- as variáveis respostas Y são normalmente distribuídas.
- Homocedasticidade- as variáveis y têm a mesma variância
- Linearidade- a relação entre a média da variável resposta Y e as variáveis explanatórias X é uma linha reta

Normalidade dos erros

A normalidade dos erros pode ser testada através de testes de ajustamento tais como o teste Kolmogorov-Smirnov ou o teste da normalidade de Lilliefors.

Ela também pode ser observada através de dois tipos de gráficos de probabilidade normal: o primeiro representa a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade observada acumulada dos erros (normal p-p plot), que é feita pela equação:

$$\frac{i-0,5}{n}$$

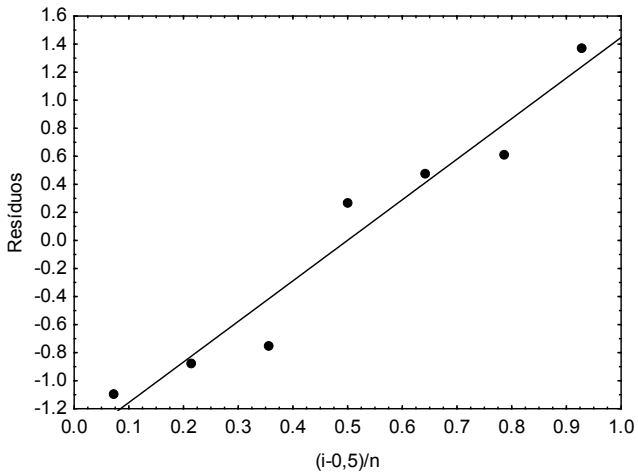
Para a construção deste gráfico, ordena-se os resíduos em ordem crescente (eixo y). Em seguida, plota-se os resíduos versus a freqüência cumulativa ($\frac{i-0,5}{n}$, eixo x).

O segundo representa o quantil de probabilidade esperada se a distribuição fosse normal em função dos resíduos (normal q-q plot):

$$P(Z < z_t) = \frac{i-0,5}{n}$$

Se os erros possuírem distribuição normal, todos os pontos dos gráficos devem se posicionar aproximadamente sobre a reta.

			Residuo	Residuo padronizado	Residuos ordenados	Distribuição observada $\frac{i-0,5}{n}$
Idade (x)	**Comprimento (y)**	**ŷ**	**(y-ŷ)**	**(y-ŷ)/1,05**		**p-p plot abscissa**
1	20	19,5	0,5	0,48	-1,09	0,071
2	25	24,7	0,28	0,27	-0,88	0,214
3	29	29,9	-0,93	-0,88	-0,75	0,357
4	34	35,1	-1,14	-1,09	0,27	0,5
5	41	40,4	0,64	0,61	0,48	0,643
6	47	45,6	1,43	1,36	0,61	0,786
7	50	50,8	-0,79	-0,75	1,36	0,929
28	246					



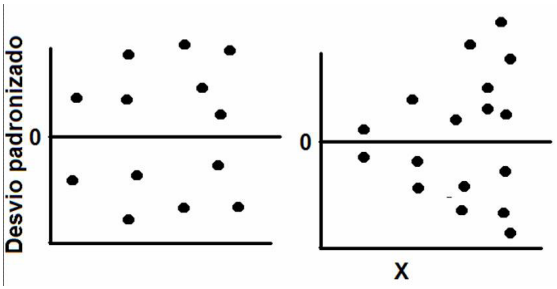
Homocedasticidade e independência dos erros

Os erros são variáveis aleatórias de variância constante (hipótese de homocedasticidade).

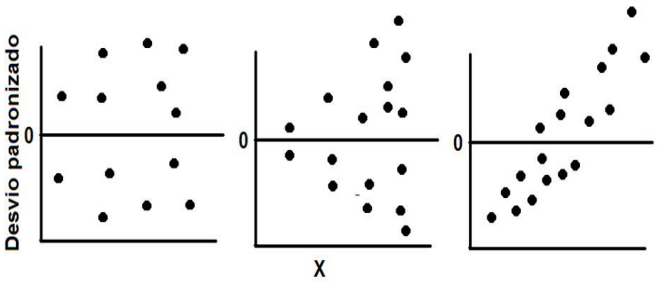
Para verificar se a média dos erros é nula, se a variância é constante e se há independência dos erros, você pode usar um diagrama com os resíduos, ou com os resíduos padronizados, que é o resíduo dividido pelo desvio padrão do mesmo (veja o cálculo da variância acima)

$$s^2 = \frac{\sum (y - \hat{y})^2}{n-2}$$

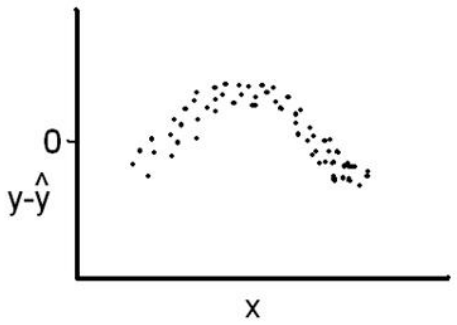
Ao examinar a distribuição dos resíduos em torno de 0, tome cuidado antes de supor que os dados não têm distribuição normal. A aparente não normalidade pode ser devido a modelo inadequado, heterocedastia ou porque são poucos dados. O próximo passo é fazer um gráfico dos resíduos padronizados contra a variável X,



Os pontos devem ficar distribuídos de forma equilibrada acima e abaixo da linha indicando resíduo zero, como no diagrama da esquerda acima. Se houver heterocedasticidade (gráfico da esquerda), deve se logaritmizar os dados.



O gráfico acima mostra que os resíduos não se comportam de modo aleatório, eles seguem um padrão, o modelo não é linear, neste caso devemos suspeitar de um erro de cálculo ou que devemos adicionar outra variável ao modelo de regressão.



Este gráfico também mostra que os resíduos não são independentes, indica que a regressão linear é imprópria para descrever os dados, talvez uma regressão quadrática seja mais adequada.

Finalmente a análise de resíduos pode envolver a busca de dados discrepantes (marginais). Estes valores podem estar, por exemplo, fora do intervalo de –3 a +3. O valor discrepante pode ser descartado se houver erro de medida ou de avaliação.

Apresentação dos resultados

Além da equação da reta, devemos apresentar o desvio-padrão residual (que em alguns programas é apresentado como "erro padrão da estimativa"), o desvio padrão da inclinação, o valor do teste "t" e o p-valor, assim como o n ou graus de liberdade.

TRANSFORMAÇÕES EM REGRESSÕES

Para testar as hipóteses de regressão é necessário que os dados tenham distribuição normal e homocedástica. Algumas vezes é necessário transformar os dados para que eles preencham estes quesitos.

Transformação da variável independente não afetará a distribuição de "y", assim transformações de "x", geralmente podem ser feitas, sem problemas. Entretanto, as transformações em "y" afetam as considerações dos mínimos quadrados e, por isto, devem ser discutidas.

Se os valores de "y" seguem a distribuição de Poisson (dados discretos e poucos dados), então a transformação em raiz quadrada é apropriada:

$$y' = \sqrt{y + 0{,}5}$$

Se os valores de y seguem uma distribuição Binomial (*e.g.*, Se eles são proporções ou porcentagens), então usamos a transformação em arcoseno, com os dados estando em razão. O resultado será um ângulo entre 0 e 90º.

$$y' = arcsen\sqrt{y}$$

Para converter ao valor de proporção: p=(seno(y)^2). Os programas R e Excel utilizam a unidade radianos, como automático, para converter em graus multiplica-se por π/180.

A transformação mais comum em regressão é calculando o logaritmo de y, apropriado quando estes dados são heterocedásticos.

$$y' = \log y \quad \text{ou} \quad y' = \log(y+1)$$

Geralmente usa-se o log(x+1) para que os valores resultantes não sejam menores do que 0 ou indefinidos (como log de zero).

RELAÇÕES NÃO LINEARES ENTRE DUAS VARIÁVEIS

Duas variáveis podem apresentar uma relação linear, mas esta não é sempre a regra. Existem muitos outros modelos que podem ser mais adequados para representar a relação entre as duas variáveis.

Uma maneira de saber qual o modelo mais adequado é através do coeficiente de determinação (r^2). O modelo que apresentar o maior coeficiente de determinação será o melhor modelo.

Modelo	Equação	Obs
Linear	$y = \beta_0 + \beta_1 x$	
Exponencial	$y = \beta_0 \beta_1^x$	y>0
Potencial	$y = \beta_0 x^{\beta_1}$	x>0,y>0
Logaritmico	$y = \beta_0 + \beta_1 \ln(x)$	x>0
Logístico	$y = \frac{c}{1 + \beta_0 e^{-\beta_1 x}}$	

Modelo	Equação	Obs
Quadrática	$y = \beta_0 x^2 + \beta_1 x + c$	
Cúbica	$y = \beta_0 x^3 + \beta_1 x^2 + cx + d$	
Quartica	$y = \beta_0 x^4 + \beta_1 x^3 + cx^2 + dx + e$	

Função exponencial

A função exponencial é descrita pela equação:

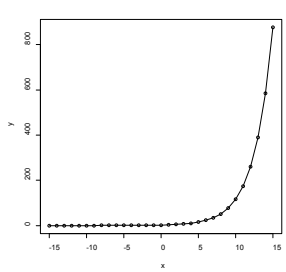
$$Y=\beta_0\ \beta_1{}^{x} \text{ ou } y=\beta_0\ e^{\beta 1\, x}$$

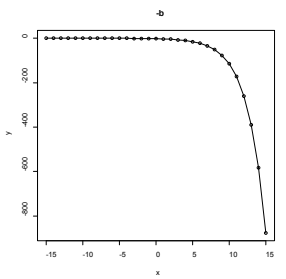
Para estimar os parâmetros precisamos transformar os dados:

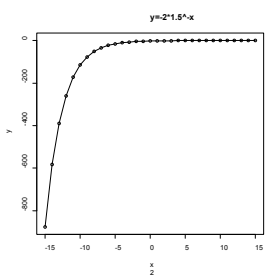
$$\log Y = \log \beta_0 + X \log \beta_1 \quad \text{ou} \quad \ln Y = \ln \beta_0 + \beta_1 X$$

Dados no R: (> x<--15:15

>y=2 × 1.5^x
>plot(x,y)
>lines(x,y))

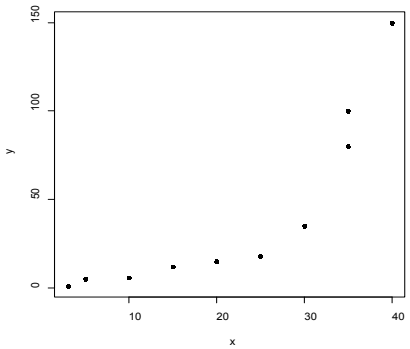






Exemplo:

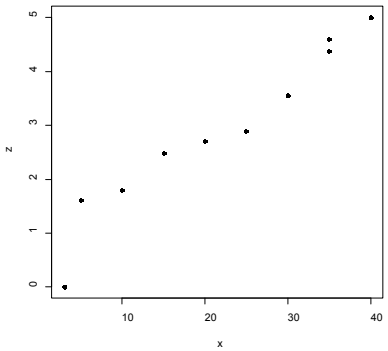
X	3	5	10	15	20	25	30	35	35	40
Y	1	5	6	12	15	18	35	80	100	150



- Primeiro passo: logaritmize o Y.

X	3	5	10	15	20	25	30	35	35	40
Z(lnY)	0	1,61	1,79	2,48	2,71	2,89	3,56	4,38	4,61	5,01

- Segundo, faça o gráfico de dispersão:



- Terceiro, calcule os parâmetros da reta.

$\beta_0 = 0{,}4465$, $\beta_1 = 0{,}1127$ → $\ln Y = \ln \beta_0 + \ln \beta_1 X$ → **z = 0,4465 + 0,1127X**

Para representar a equação na forma exponencial faça o anti-logaritmo de β_0 → $e^{0{,}4465}$, e de β_1 → $e^{0{,}1127}$,

$Y = \beta_0 \beta_1^{x}$ → **y= 1,563 × 1,12^x**

Função potencial:

$$Y = \beta_0 X^{\beta_1}$$

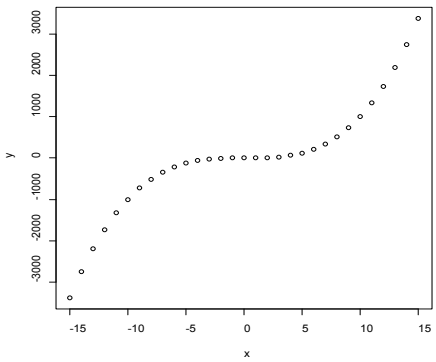
$$\log Y = \log \beta_0 + \beta_1 \log X$$

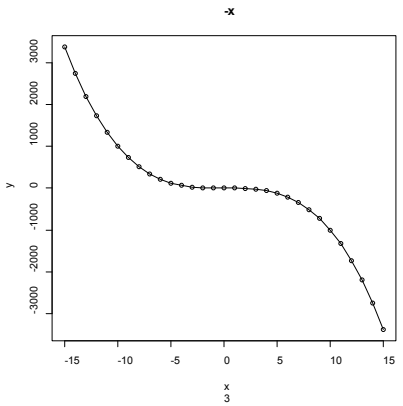
Dados no R: (> x<--15:15

>Y=-x^3

>plot(x,y)

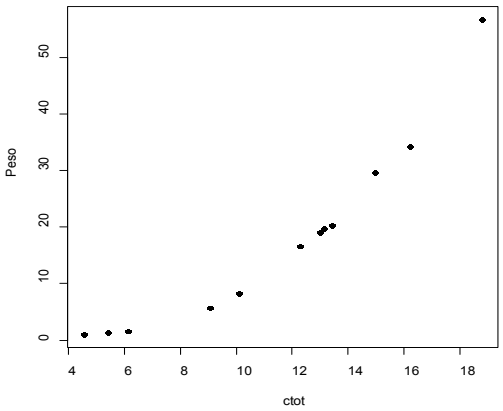
>lines(x,y))





Exemplo:

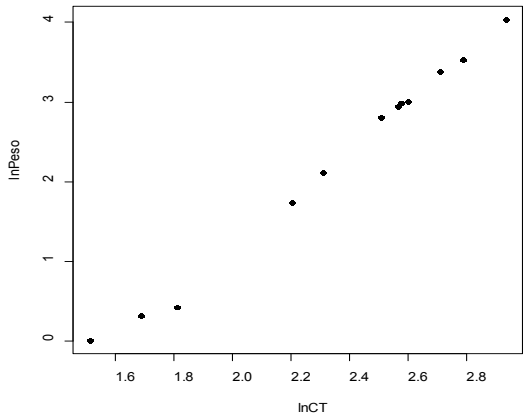
C.Total	4.54	5.40	6.12	9.05	10.09	12.27	13.01	13.14	13.44	14.97	16.21	18.80
Peso	1.01	1.38	1.53	5.70	8.28	16.58	19.07	19.77	20.30	29.55	34.22	56.73



- Primeiro, logaritmize Y e X

Ln(CT)	1,51	1,69	1,81	2,20	2,31	2,51	2,57	2,58	2,60	2,71	2,79	2,93
Ln Peso	0,01	0,32	0,43	1,74	2,11	2,81	2,95	2,98	3,01	3,39	3,53	4,04

- Segundo, faça o gráfico de dispersão:



- Terceiro, calcule os parâmetros da reta.
  β_1=2,963, α= – 4,690
  Calcule os parâmetros da relação potencial:
  $\beta_0 = e^{\alpha} = e^{-4,69} = 0,0092$
  Massa= $0,0092 * CT^{2,963}$

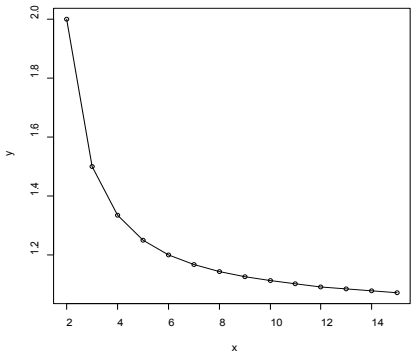
Função hiperbólica

$$y = \frac{x}{(ax - b)}$$

Os dados linearizados são:

$$\frac{1}{y} = a - b\left(\frac{1}{x}\right)$$

Dados no R: (> x<-2:15 >y=x/(x-1) >plot(x,y) >lines(x,y))



Função logaritmica

$$y = a + b \ \ln x$$

Os dados linearizados são: $y = a + b \ \ln x$

Dados no R: (> x<--15:15

>y=1+(log(x))

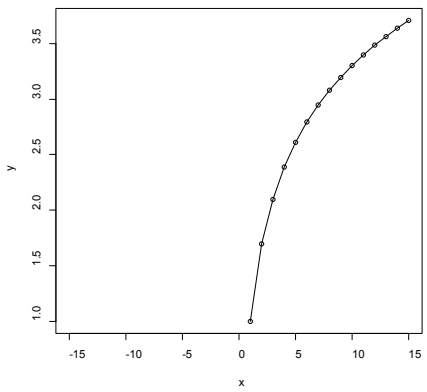
>plot(x,y)

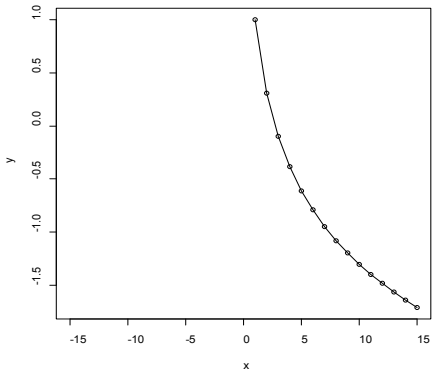
>lines(x,y) e

>y = 1-(log(x))

>plot(x,y)

>lines(x,y))

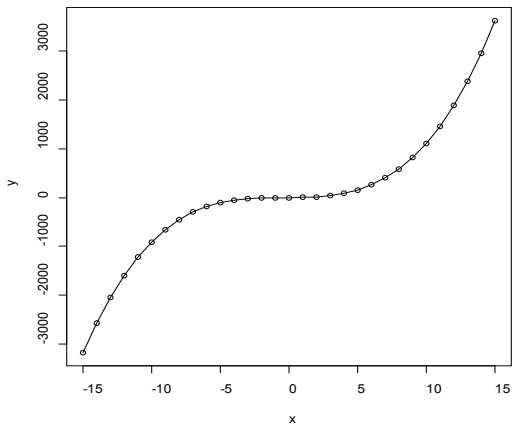




Modelos polinomiais

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Forma linear: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$



RESUMO

Tipo	*Equação*	*Transformação*	*Variável x*	*Variável y*
Linear	$Y = \beta_0 + \beta_1 x$	$Y = \beta_0 + \beta_1 x$	x	y
Exponencial	$Y = a.e^{bx}$ ou $Y = a \times b^x$	$\ln y = \ln a + bx$ $\ln Y = \ln a + x \ln b$	x	ln Y
Logarítmica	$Y = a + b.\ln x$	$Y = a + b.\ln x$	ln x	y
Potencial	$Y = a.x^b$	$\ln y = \ln a + b.\ln x$	ln x	ln y
Hiperbólica	$y = \frac{x}{(ax - b)}$	$\frac{1}{y} = a - b\left(\frac{1}{x}\right)$	1/x	1/y
Polinomial	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$		

COMPARANDO RETAS DE REGRESSÃO

É comum calcularmos regressões para mais que um conjunto de dados. Nós podemos inferir se as inclinações destas retas são significativamente diferentes.

Comparando as inclinações de duas retas.

Nós podemos utilizar o teste t para testar a diferença entre as inclinações. O teste é:

$$t = \frac{\beta_{11} - \beta_{12}}{s_{\beta_{11}-\beta_{12}}}$$

O erro padrão da diferença entre os dois coeficientes de regressão é:

$$s_{\beta_{11}-\beta_{12}} = \sqrt{\frac{\left(s^2_{YX}\right)_p}{\left(\sum x^2\right)_1} + \frac{\left(s^2_{YX}\right)_p}{\left(\sum x^2\right)_2}}$$

E o quadrado médio dos resíduos é calculado como:

$$\left(s^2_{YX}\right)_p = \frac{(SQres)_1 + (SQres)_2}{(glres)_1 + (glres)_2}$$

$$SQres = \sum y^2 - \frac{\left(\sum xy\right)^2}{\sum x^2}$$

$$\text{g.l.res} = n_1 + n_2 - 4$$

Exemplo

Foi estimada a regresão entre a concentração de cobre e o comprimento de *tagelus plebeius*, em dois estuários de pernambuco. As regressões são iguais?

H_0: $\beta_{11} = \beta_{12}$

H_1: $\beta_{11} \neq \beta_{12}$

	Estuário 1	**Estuário 2**
$\sum x^2$	14608.02	8440.97
$\sum xy$	5523.19	3227.91
$\sum y^2$	2091.69	1235.73
n	15	15
β_1	0.356	0.371
SQ resíduos	2.731	0.772

$$\left(s^2_{YX}\right)_p = \frac{2,731+0,772}{13+13} = 0,135$$

$$S_{b1-b2} = \sqrt{\frac{0,135}{14608,02} + \frac{0,135}{8440,97}} = 0,005$$

$$t = \frac{0,371-0,356}{0,005} = 3$$

Gl=13+13=26

Rejeito H_0 se $|t| \geq t_{\alpha(2),gl}$

$t_{0,05(2),26}$=2,056

Rejeito H_0.

EXERCÍCIOS CAPÍTULO 8

1. Os dados abaixo referem-se ao peso (gramas) e a idade (semanas), de peixes. Estime a equação de regressão linear, faça o teste de significância e interprete o coeficiente de determinação.

Idade	1	2	3	4	5	6	7	8	9	10
Peso	30	50	60	70	90	110	160	170	190	210

2. Foram estudadas 9 crianças com o objetivo de verificar qual é a relação entre a capacidade pulmonar e a idade..teste a significância da regressão e interprete o resultado obtido.

Idade (anos)	4	5	6	7	8	9	10	11	12
Capacidade	0,7	0,9	1,2	1,3	1,3	1,5	1,7	1,9	2,1

3..teste ergométrico foi de r2=0,64 para o modelo linear, r2=0,84 para o modelo exponencial e r2= 0,97 para o modelo potencial. Construa a equação deste modelo.

Esforço (Joule)	10	20	25	30	35	40	45	50
Batimento (Num./min)	80	121	140	149	163	167	173	175

4. O modelo de regressão obtido a partir dos dados abaixo é: Cons. O2= 3,4714 – 0,08776 temperatura. Calcule os resíduos da regressão.

Consumo O_2	5,2	4,7	4,5	3,6	3,4	3,1	2,7	1,8
Temperatura (°C)	-18	-15	-10	-5	0	5	10	19

PARTE II – ESTATÍSTICA NO EXCEL

JOSÉ ROBERTO BOTELHO DE SOUZA

Capítulo 9

INTRODUÇÃO AO EXCEL

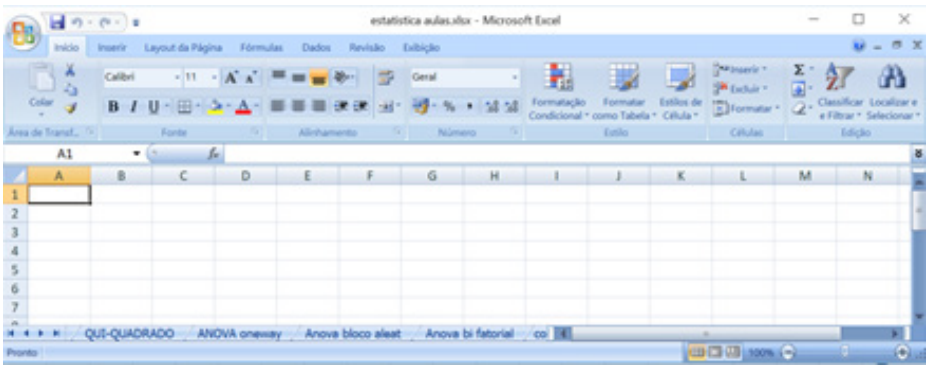
Os programas estatísticos facilitam **a análise de grandes conjuntos de dados** e possiblitam análises complexas dos mesmos. Existem vários programas estatísticos de de armazenagem de dados. O programa Excel foi desenvolvido para armazenar e administrar dados. Além disso, ele possui um conjunto de funções e rotinas para a realização de cálculos, gráficos e estatística básica. Há várias versões do Excel, mas as funções básicas não sofrem muita variação, e quase certamente estão na versão instalada em seu computador.

CONHECENDO O EXCEL

Planilha

A planilha do Excel possui vários componentes, facilmente identificáveis: a célula selecionada fica realçada em negrito.

Ela é identificada por uma coluna (letra) e uma linha (número), que estão em negrito. Cada planilha possui 1048576 linhas e 16384 colunas (Excel 2013 a 2017).



A letra e o número correspondente à célula estão escritos, à esquerda acima da planilha, e a descrição do conteúdo da mesma, fica logo ao lado. Na parte inferior da pasta de trabalho está a planilha selecionada e a barra de navegação da mesma.

Barra de menu

Existem 9 categorias diferentes. As barras de ferramentas estão associadas ao menu. Elas contêm as tarefas mais executadas pelo usuário, e podem ser personalizadas

Tipos de dados

Há três tipos de dados que podem ser inseridos no Excel: números, textos e fórmulas. Por exemplo, a planilha de mudas de plantas abaixo apresenta os três tipos de dados:

B10 =SOMA(B2:B9)

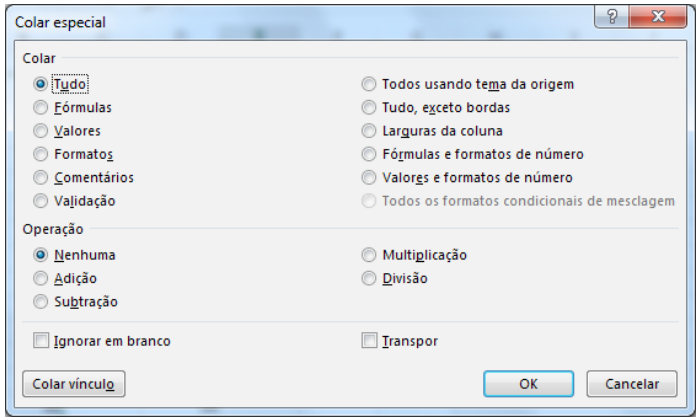
	A	B	C	D
1	Árvore	Número de mudas		
2	Pau Brasil	12		
3	Ipê Rosa	21		
4	Jatobá	36		
5	Jacarandá	44		
6	Ipê Amare	3		
7	Pau Ferro	32		
8	Cacau	27		
9	Goiabeira	11		
10	Soma	186		

Editando a planilha:

As ferramentas de edição comuns aos aplicativos da microsoft, como copiar (ctrl + "c") e colar (ctrl + "v"), funcionam bem aqui. Além disso, o Excel possui várias peculiaridades que facilitam a manipulação dos dados:

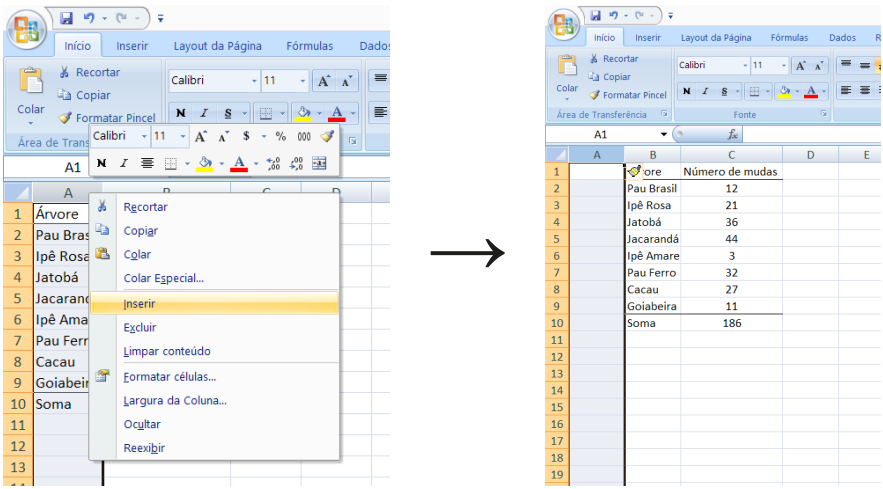
Colar especial

A função colar especial, pode ser encontrada no menu 'início' ou apertando o botão direito do mouse. Ao acionar a função aparece a caixa de diálogo abaixo:



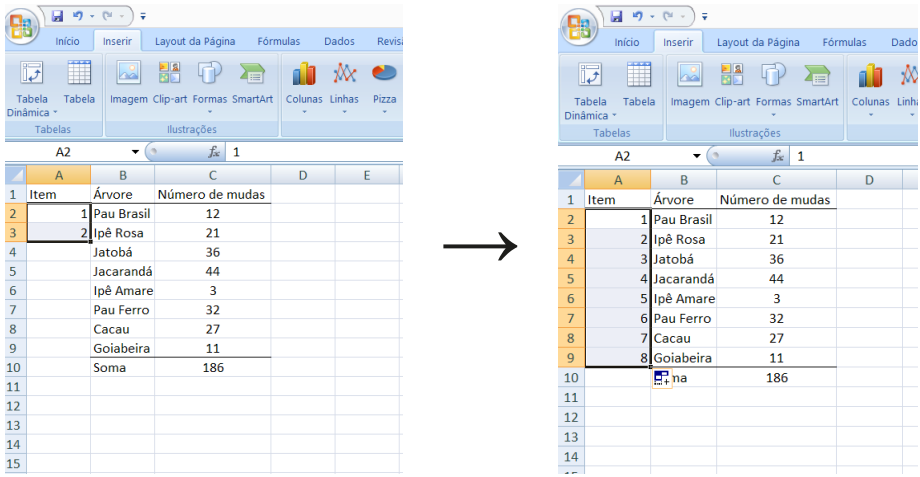
Inserir

Há um menu inserir, com várias opções, como gráficos, imagens, objetos, símbolos, entre outros. Com o botão direito do mouse, você pode inserir colunas ou células, de acordo com o que foi selecionado



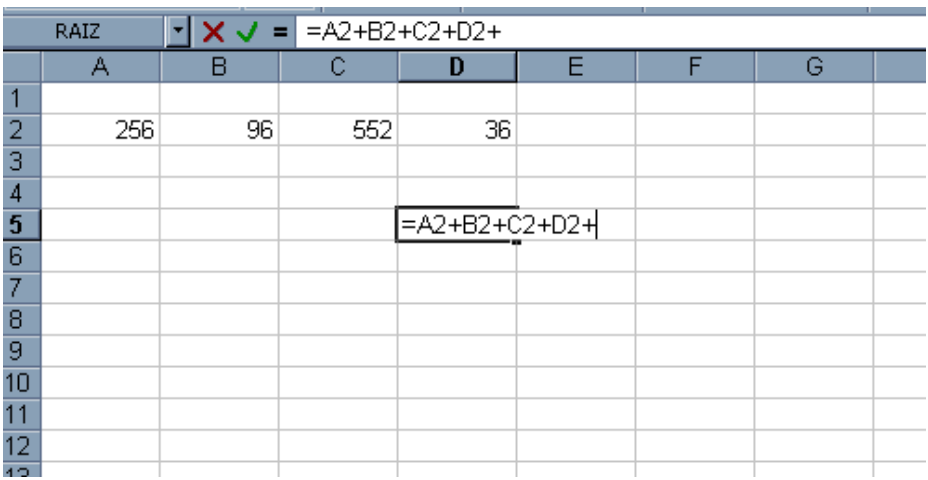
Arrastar

A seta do mouse, no Excel tem a forma de mais (✚), que muda de forma quando está no canto inferior esquerdo da região selecionada (+), possibilitando copiar ou inserir mais dados.



Barra de fórmulas

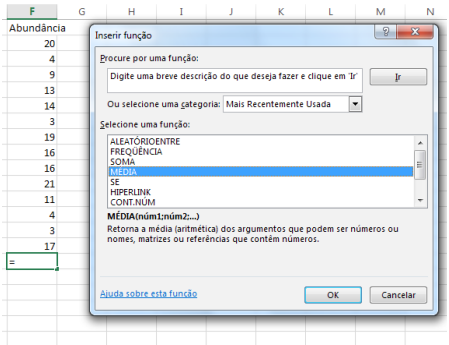
A barra de fórmulas, além de exibir o conteúdo da célula, permite inserir fórmulas e funções. Para inserir fórmulas e funções, deve-se colocar inicialmente o sinal de igual (‘=’) .

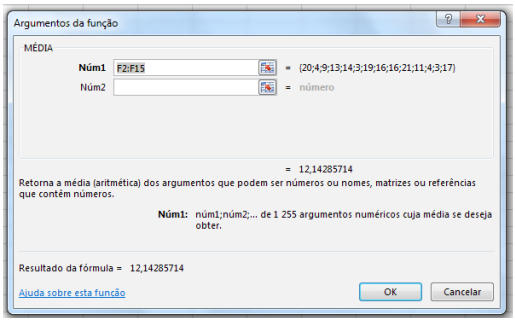


FUNÇÕES

Funções são equações de uso geral que o Excel disponibiliza ao usuário. Elas podem ser aplicadas a um conjunto de dados ou a apenas uma célula. O acesso à caixa de diálogo das funções é feito clicando na barra de fórmulas.

Exemplo: média





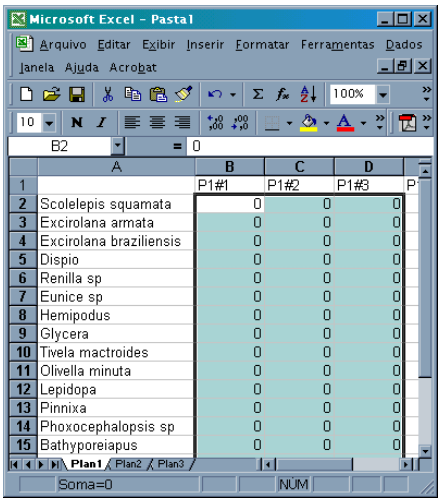
Todas as funções possuem um recurso de 'ajuda', com exemplos para a maioria das funções utilizadas

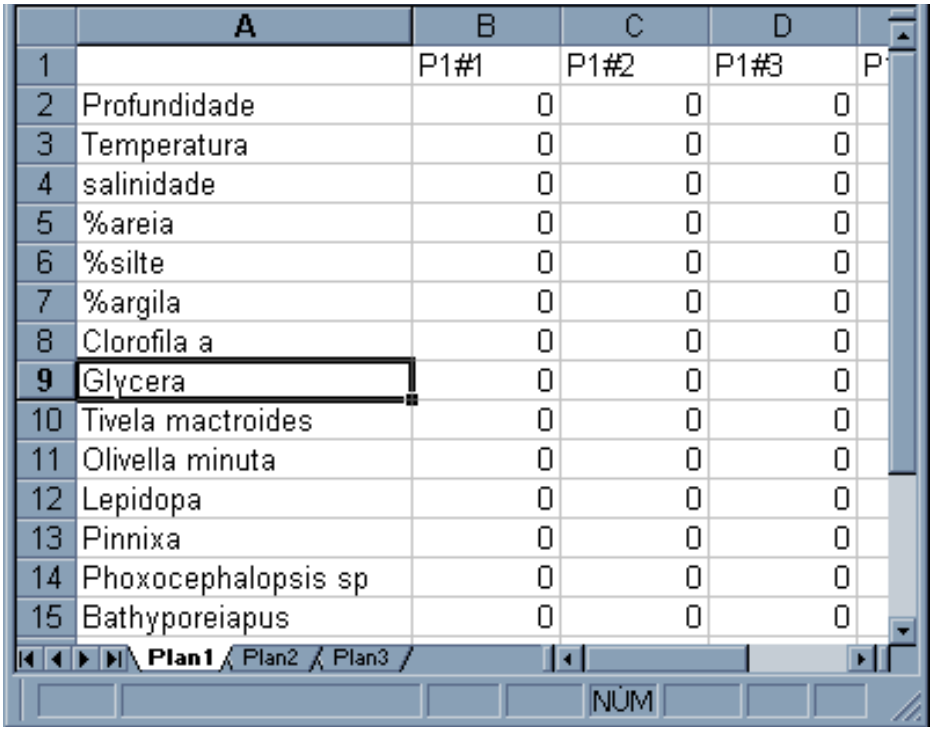
Algumas dicas:

- Coloque os dados bióticos e abióticos em planilhas distintas
- Defina os descritores e objetos ou as variáveis dependentes e independentes

Exemplo:

Descritor X objeto





	A	B	C	D
1		P1#1	P1#2	P1#3
2	Profundidade	0	0	0
3	Temperatura	0	0	0
4	salinidade	0	0	0
5	%areia	0	0	0
6	%silte	0	0	0
7	%argila	0	0	0
8	Clorofila a	0	0	0
9	Glycera	0	0	0
10	Tivela mactroides	0	0	0
11	Olivella minuta	0	0	0
12	Lepidopa	0	0	0
13	Pinnixa	0	0	0
14	Phoxocephalopsis sp	0	0	0
15	Bathyporeiapus	0	0	0

Variável dependente x variável independente

	A	B	C	D	E	F	G	H	I
1	estação	Densidade	Num.Spp	Shannon	equitativ.	Temperatu	Salinidade	Profundida	%areia
2	P1#1	0	0	0	0	0	0	0	0
3	P1#2	0	0	0	0	0	0	0	0
4	P1#3	0	0	0	0	0	0	0	0
5	P1#4	0	0	0	0	0	0	0	0
6	P1#5	0	0	0	0	0	0	0	0

- Todas as informações de uma estação, devem caber em uma célula.

#12B04outR1

- Todos os dados abióticos ou bióticos devem estar em apenas uma planilha.
- Cada hipótese ou inferência deve estar numa planilha separada.
- Organize os dados de modo que você possa rastreá-los a partir da planilha básica.

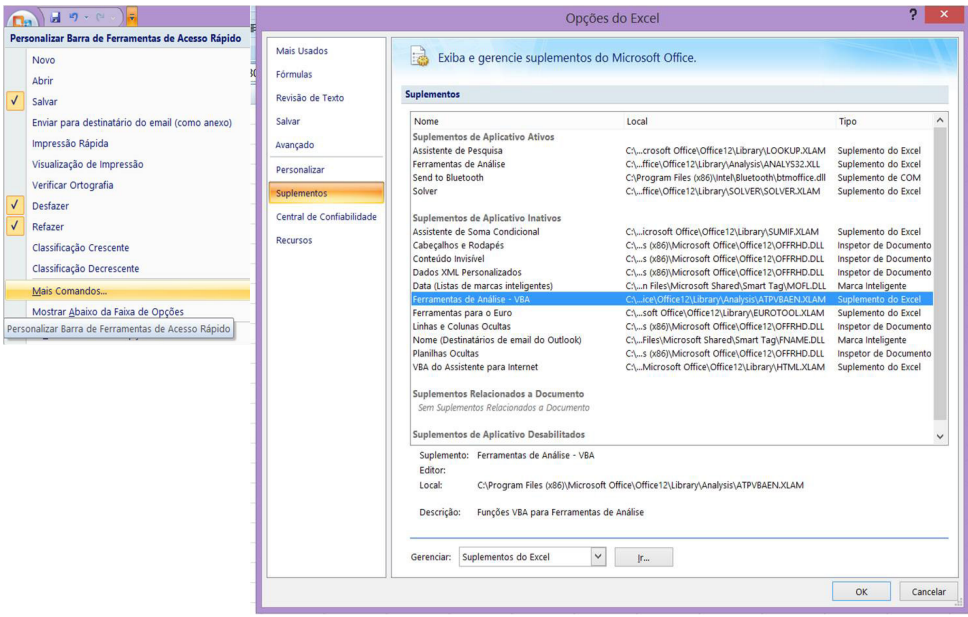
- As unidades usadas devem estar bem visíveis

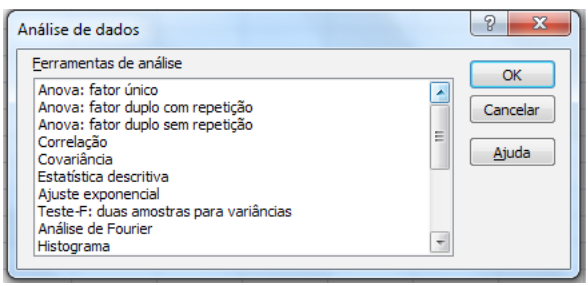
$$Ind\ m^{-2}$$

CAIXA DE DIÁLOGO: ANÁLISE DE DADOS

O Excel possui um conjunto de rotinas estatísticas, que estão no suplemento 'análise de dados'. A instalação deste suplemento varia conforme a versão do Excel.

Sequência de instalação :-personalizar barra de ferramentas de acesso rápido/mais comandos/suplementos/ ferramentas de análise.





EXERCÍCIOS DO CAPÍTULO 9

1. Nomeie a planilhas existentes na aba de planilhas , na parte inferior da pasta. Exercite a criação e exclusão de planilhas

2. Insira variáveis de um projeto seu ou que você tenha interesse. Caso você não tenha dados, utilize as funções 'ALEATÓRIOENTRE', com intervalos bem diferentes, à sua escolha, assim como a função 'ALEATÓRIO'. Com estas duas funções você pode criar dados quantitativos contínuos e discretos. Arrume as variáveis em colunas ou linhas, à sua escolha.

Capítulo 10

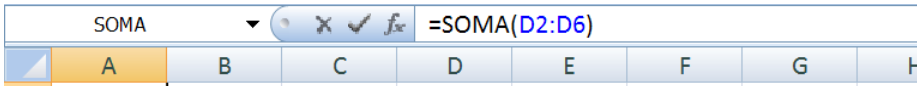
ESTATÍSTICA DESCRITIVA

Obtenção de parâmetros estatísticos, cálculos de porcentagens e gráficos

ESTIMATIVA DE PARÂMETROS ESTATÍSTICOS

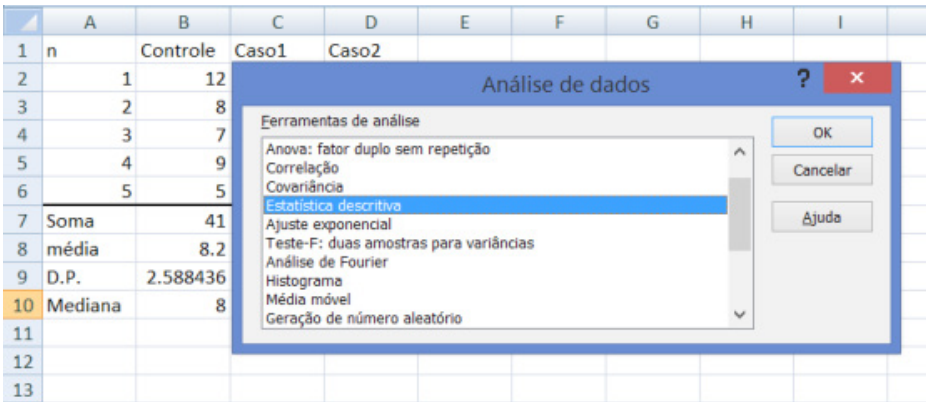
As estimativas e apresentações descritivas dos dados podem ser plenamente realizadas no Excel, através des funções, equações, gráficos e através da caixa de diálogo "análise de dados":

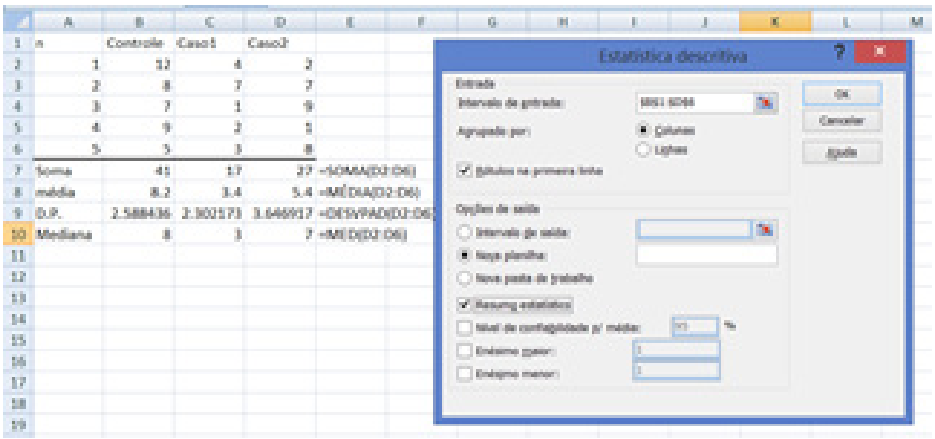
- Fórmulas e funções, são visualizadas na barra de fórmulas:



	A	B	C	D	E	F
1	n	Controle	Caso1	Caso2		
2	1	12	4	2		
3	2	8	7	7		
4	3	7	1	9		
5	4	9	2	1		
6	5	5	3	8		
7	Soma	41	17	27	=SOMA(D2:D6)	
8	média	8.2	3.4	5.4	=MÉDIA(D2:D6)	
9	D.P.	2.588436	2.302173	3.646917	=DESVPAD(D2:D6)	
10	Mediana	8	3	7	=MED(D2:D6)	
11						

- A caixa de diálogos 'análise de dados' (Menu: Dados) possui a rotina 'Estatística Descritiva':





As opções específicas desta caixa de diálogo são:

- Resumo estatístico – gera um campo para cada uma das seguintes estatísticas na tabela de saída: média, erro padrão (da média), mediana, modo, desvio padrão, variância, curtose, distorção, intervalo, mínimo, máximo, soma, contagem, maior (n), menor (n) e nível de confiança.
- Nível de confiança da média – insere o nível de confiança a ser utilizado

- Enésimo maior- mostrará o maior número dentro da ordem selecionada: 1- máximo, 2- 2º maior,...

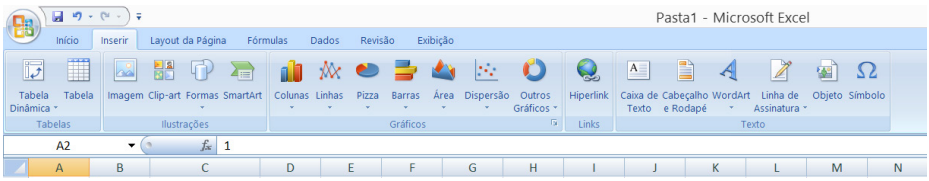
O resultado aparece como abaixo:

A1 Controle

	A	B	C	D	E	F
1	*Controle*		*Caso1*		*Caso2*	
2						
3	Média	8.2	Média	3.4	Média	5.4
4	Erro padrã	1.157584	Erro padrã	1.029563	Erro padrã	1.630951
5	Mediana	8	Mediana	3	Mediana	7
6	Modo	#N/D	Modo	#N/D	Modo	#N/D
7	Desvio pac	2.588436	Desvio pac	2.302173	Desvio pac	3.646917
8	Variância c	6.7	Variância c	5.3	Variância c	13.3
9	Curtose	0.795277	Curtose	1.128515	Curtose	-2.85092
10	Assimetria	0.501657	Assimetria	1.032659	Assimetria	-0.48243
11	Intervalo	7	Intervalo	6	Intervalo	8
12	Mínimo	5	Mínimo	1	Mínimo	1
13	Máximo	12	Máximo	7	Máximo	9
14	Soma	41	Soma	17	Soma	27
15	Contagem	5	Contagem	5	Contagem	5

GRÁFICOS

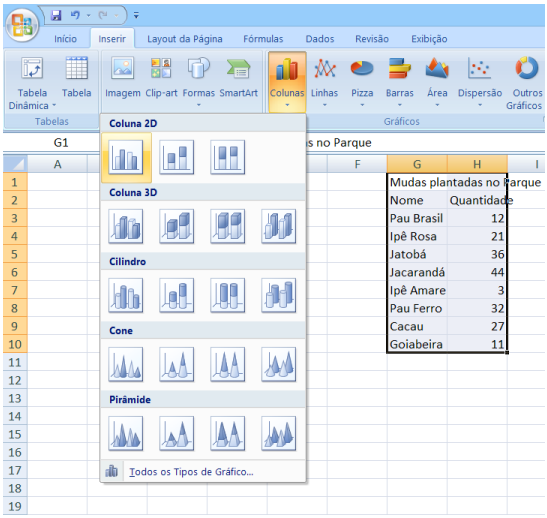
Os gráficos do Excel estão no menu "inserir", com a escolha feita a partir de ícones.

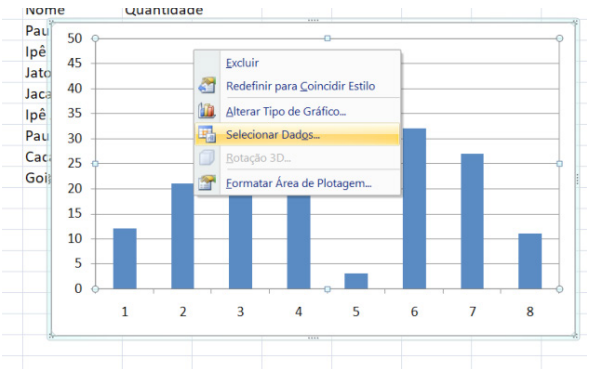


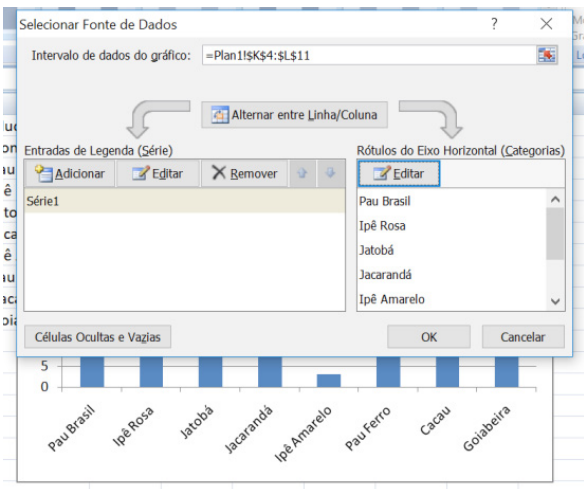
Gráficos em barra

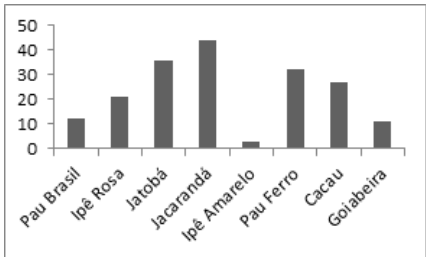
Selecione as duas colunas. Abra o menu 'inserir' e selecione o gráfico. Depois use o menu 'ferramentas de gráfico', para colocar títulos e personalizar o mesmo

Exemplo


Início
Inserir
Layout da Página
Fórmulas
Dados
Revisão
Exibição
Tabela Dinâmica
Tabela
Imagem
Clip-art
Formas
SmartArt
Colunas
Linhas
Pizza
Barras
Área
Dispersão
Outros Gráficos
Tabelas
Gráficos
Coluna 2D
Coluna 3D
Cilindro
Cone
Pirâmide
Todos os Tipos de Gráfico...
G1
s no Parque
Mudas plantadas no Parque
Nome
Quantidade
Pau Brasil 12
Ipê Rosa 21
Jatobá 36
Jacarandá 44
Ipê Amare 3
Pau Ferro 32
Cacau 27
Goiabeira 11


Excluir
Redefinir para Coincidir Estilo
Alterar Tipo de Gráfico...
Selecionar Dados...
Rotação 3D...
Formatar Área de Plotagem...
50
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8


Selecionar Fonte de Dados
Intervalo de dados do gráfico: =Plan1!$K$4:$L$11
Alternar entre Linha/Coluna
Entradas de Legenda (Série)
Adicionar
Editar
Remover
Série1
Rótulos do Eixo Horizontal (Categorias)
Editar
Pau Brasil
Ipê Rosa
Jatobá
Jacarandá
Ipê Amarelo
Células Ocultas e Vazias
OK
Cancelar
5
0
Pau Brasil
Ipê Rosa
Jatobá
Jacarandá
Ipê Amarelo
Pau Ferro
Cacau
Goiabeira

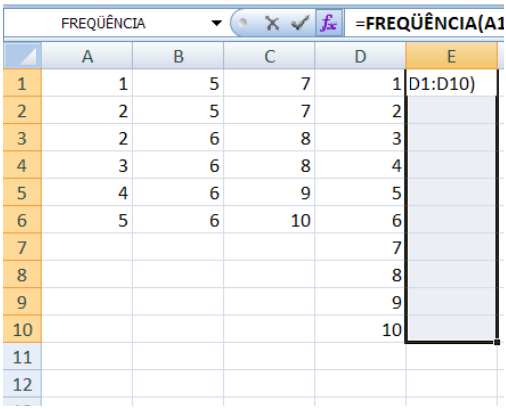


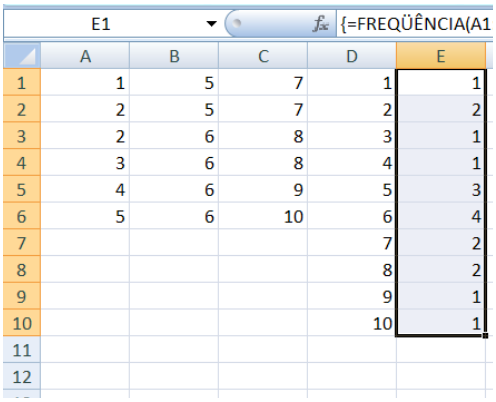
HISTOGRAMAS

O histograma pode ser feito utilizando-se a função "freqüência" do Excel:

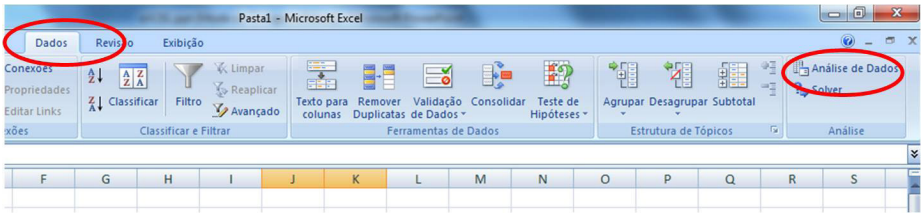
FREQÜÊNCIA (matriz_dados,matriz_bin), 'dados' é a matriz de dados que se quer contar, 'bin' são os intervalos desejados. A fórmula precisa ser inserida como uma fórmula matricial. Isto é, seleciona-se o espaço adjacente aos intervalos escolhidos, preenche os dados da função 'frequência' e em seguida, pressione CTRL+SHIFT+ENTER (não pressione o 'ok' da caixa de diálogo).

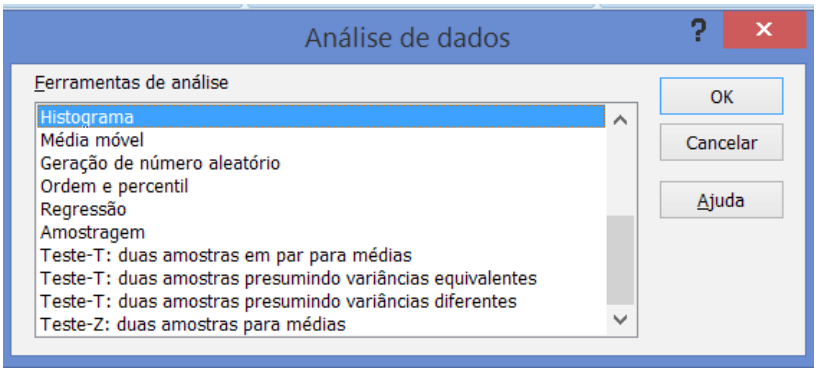
Ex:



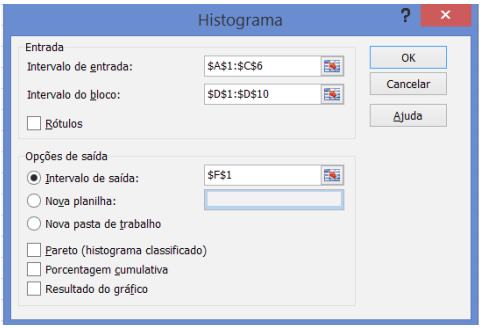


Outra maneira é utilizar a rotina 'histograma' que está no suplemento "análise de dados".





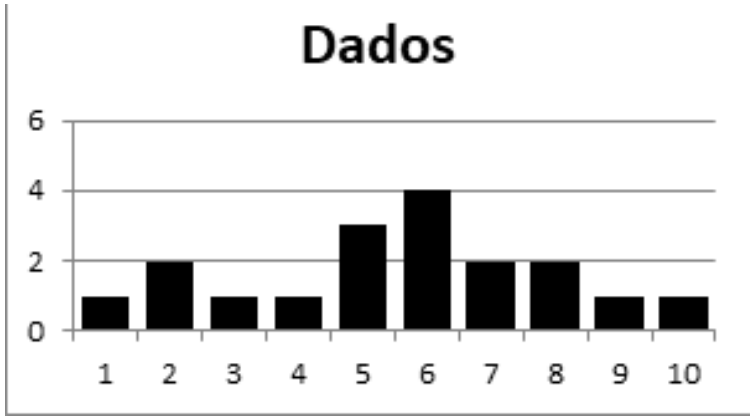
Dentro do menu histograma será necessário preencher três campos:



- O intervalo de entrada, que corresponde à linha ou coluna onde estão os dados.
- O intervalo do bloco, que corresponde às classes de tamanho que você definiu.
- O intervalo de saída ou nova planilha, onde irá aparecer a **freqüência absoluta**.

O resultado do histograma será uma tabela de freqüência

	A	B	C	D	E	F	G
1	1	5	7	1	1	*Bloco*	*Freqüência*
2	2	5	7	2	2	1	1
3	2	6	8	3	1	2	2
4	3	6	8	4	1	3	1
5	4	6	9	5	3	4	1
6	5	6	10	6	4	5	3
7				7	2	6	4
8				8	2	7	2
9				9	1	8	2
10				10	1	9	1
11						10	1
12						Mais	0



CÁLCULO DE PORCENTAGENS

Cálculos de abundância ou freqüência relativa podem ser obtidos através de tabelas ou histogramas. Para o cálculo de porcentagens simples e acumulada, primeiro calcule a soma dos dados, utilizando a função "=SOMA(B2:B14)". Depois, calcule a porcentagem usando a formula "=(B2/B$15)×100", não esquecendo de fixar a célula do total. O total da % é cem, como na imagem abaixo.

C2 f_x =(B2/B$15)*100

	A	B	C	D
1	Espécies /Abund.	Amostra 1	%	
2	Ancinus sp	85	20.73171	
3	Aphoditidae	4	0.97561	
4	Australonuphis casamiqu	37	9.02439	
5	Battyporeiaus ruffoi	56	13.65854	
6	Bledius bonaerensis	12	2.926829	
7	Bowmaniella brasiliensis	0	0	
8	Caridea	37	9.02439	
9	CLEANTOI	36	8.780488	
10	Diopatra virides	27	6.585366	
11	Dispio remanei	97	23.65854	
12	Doxax gemmula	4	0.97561	
13	Donax Hanleyanus	12	2.926829	
14	Emerita brasiliensis	3	0.731707	
15	Soma	410	100	

Para a porcentagem acumulada, ordene as porcentagensem ordem decrescente. Some as porcentagens uma a uma.

	A	B	C	D
1	Espécies /Abund.	Amostra 1	%	% Acum
2	Dispio remanei	97	23.65854	23.65854
3	Ancinus sp	85	20.73171	44.39024
4	Battyporeiaus ruffoi	56	13.65854	58.04878
5	Australonuphis casamiqı	37	9.02439	67.07317
6	Caridea	37	9.02439	76.09756
7	CLEANTOI	36	8.780488	84.87805
8	Diopatra virides	27	6.585366	91.46341
9	Bledius bonaerensis	12	2.926829	94.39024
10	Donax Hanleyanus	12	2.926829	97.31707
11	Aphoditidae	4	0.97561	98.29268
12	Doxax gemmula	4	0.97561	99.26829
13	Emerita brasiliensis	3	0.731707	100
14	Bowmaniella brasiliensis	0	0	100
15	Soma	410	100	

Nesta fórmula é necessário fixar a célula que soma os dados, para isto utilizamos o sinal '$'. Podemos fixar totalmente a célula ($A$1), fixar apenas a coluna ($A1), ou apenas a linha (A$1).

MODELOS DE DISTRIBUIÇÃO DE PROBABILIDADE

Os modelos de distribuição são a base da estatística inferencial, permitem a estimativa de probabilidades dos dados. O Excel possui diversas funções que permitem obter parâmetros destes modelos.

Distribuição Binomial

Essa distribuição é adequada para encontrarmos a probabilidade de ocorrência de determinado evento binário. A distribuição Binomial pode ser resolvida no Excel a partir da aplicação direta da equação Binomial:

$$P_{(x)} = \frac{n!}{x! \times (n-x)!} \times p^x \times q^{n-x} \rightarrow \text{COMBIN}(n,x) \times (p^x) \times ((1-p)^{(n-x)})$$

Utilizando o mesmo exemplo do capitulo 2: uma cadela está grávida de cinco filhotes. Qual a probabilidade dela dar a luz a cinco filhotes fêmeas?

B4 =COMBIN(B1,B2)*(B3^B2)*((1-B3)^(B1-B2))

	A	B	C	D	E	F	G	H
1	n=	5						
2	x=	5						
3	p=	0.5						
4	Resultado=	0.03125	B4=COMBIN(B1,B2)*(B3^B2)*((1-B3)^(B1-B2))					
5								

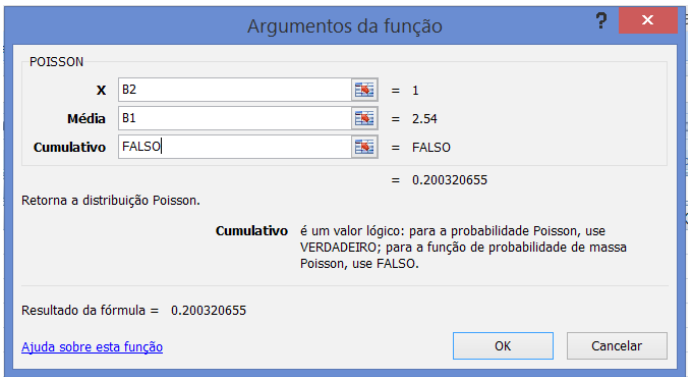
A probabilidade de nascerem cinco fêmeas é de 3,125%.

Distribuição de Poisson

A distribuição de Poisson é utilizada para determinar o número de eventos de uma variável contínua. A descrição desta distribuição e suas premissas estão no capítulo 2.

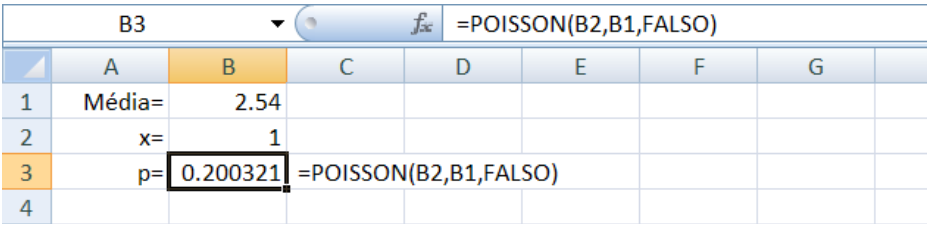
Para solucionar questões com esta distribuição no Excel, podemos utilizar a função:

POISSON(x, média, cumulativo), onde: x=número de eventos, média= o valor numérico esperado e cumulativo= valor lógico, quando 'VERDADEIRO', resultará na probabilidade de eventos entre 0 e x (cumulativo), quando 'FALSO', resultará na probabilidade de x.



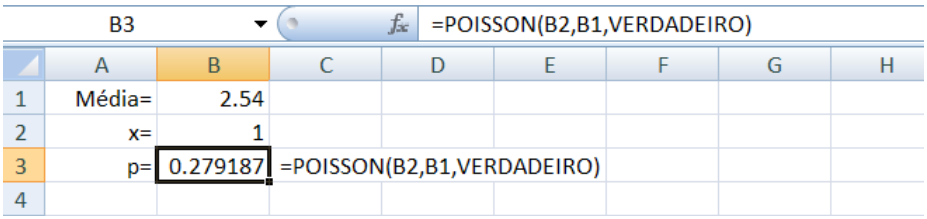
Utilizando o mesmo exemplo do capítulo 2:

A densidade média da amazônia é 2,54 habitantes/km^2. Qual a probabilidade de encontrar um habitante/km^2?



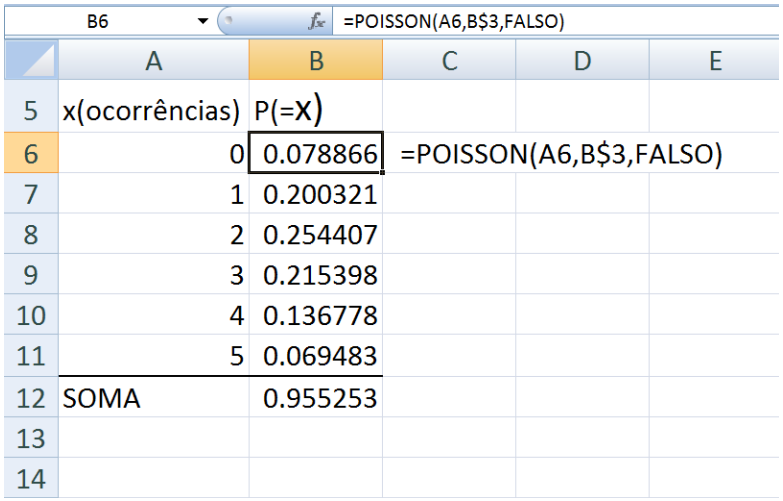
Resposta: a probabilidade de encontrarmos uma densidade 1 habitantes/km^2 é de 20%.

Por outro lado, se quisermos saber qual a probabilidade de encontrarmos um habitante ou menos. O argumento 'cumulativo' seria verdadeiro:



Neste caso, a resposta é 27,92%.

Podemos construir a curva de distribuição de habitantes na amazônia:



DISTRIBUIÇÃO NORMAL

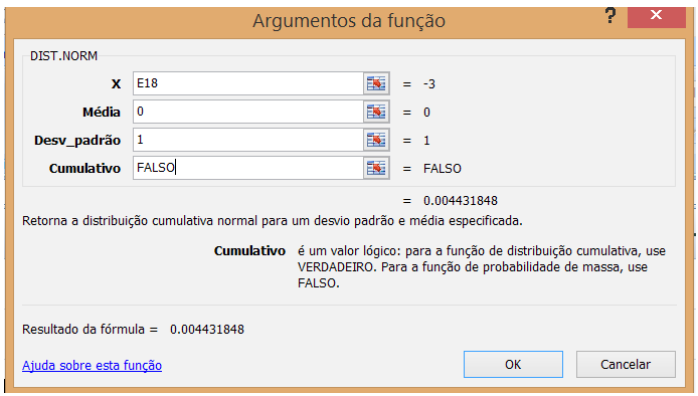
As funções 'dist.norm', 'dist.normp', 'inv.norm' e 'inv.normp' facilitam o cálculos com a distribuição normal.

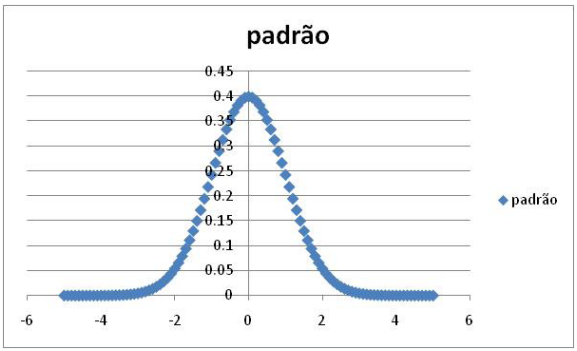
a função: "dist.norm(x,média,desv_padrão, cumulativo)" fornece a distribuição cumulativa normal para a média de desvio padrão dados, em valor de área.

Esta função possui quatro argumentos:

- X- valor da variável aleatória para a qual será determinada a densidade
- Média - média da distribuição normal
- Desv_padrão- desvio padrão da distribuição normal
- Cumulativo - colocando a palavra falso obterá o valor de densidade. Se colocar verdadeiro, terá a área ou probabilidade cumulativa.

Isto é, a palavra falso, dá a densidade e com isto podemos fazer o gráfico da curva normal padrão, como abaixo:





E o "verdadeiro" dá a área ou probabilidade da curva.

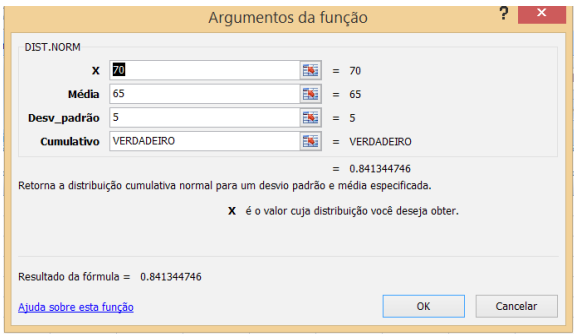
Exemplo:

Para sabermos, por exemplo, numa população de guaiamum (*cardisoma guanhumi*), com média de 65 mm de comprimento e desvio padrão de 5 mm, qual a probabilidade de um animal ter entre 65 e 70 mm podemos utilizar a função "DIST.NORM".

Para x=70

$$z = \frac{70 - 65}{5} = 1$$

Área= 0,3413. Isto é 34%



O Excel soma a curva inteira. O correto é diminuir 0,5. Assim, =0,841345-0,5 = 0,341345

Para determinar a probabilidade para intervalos de z, usa'-se, preferencialmente a função: "dist.normp", onde você insere o valor de z e a função calcula a probabilidade, ou área. Por exemplo:

z	p	Função
0	0,5	=DIST.nORMP(0)
1	0,8413	=DIST.nORMP(1)
1,65	0,9505	=DIST.nORMP(1,65)
1,96	0,9750	=DIST.nORMP(1,96)

Resumindo: a partir do cálculo de 'z' você pode estimar a probabilidade ou área correspondente.

Função "INV.NORM"

Esta função calcula o valor da média amostral (x) a partir dos valores de área (0 a 1), média e desvio padrão populacionais.

Exemplo: a média obtida por um aluno ficou 13% menor que a média populacional, no teste seletivo, sendo que a média de acertos foi de 58 questões e desvio padrão de 6. Qual foi a média do aluno?

	A	B	C	D
1	p=	0.37		
2	μ=	58		
3	σ=	6		
4	média=	56.00888	=INV.NORM(B1,B2,B3)	
5				

Foi utilizado o valor de p=0.37, Porque a média está a 50% da distribuição, ou seja 0,5, diminuindo 13 % chega-se a este valor

Função "INV.NORMP": esta função calcula o valor de 'z' a partir do valor de probabilidade ou área.

	C	D	E	F
6	p	z		
7	0.025	-1.95996	=INV.NORMP(C7)	
8	0.05	-1.64485		
9	0.5	-1.4E-16		
10	0.9	1.281552		
11	0.95	1.644854		
12	0.975	1.959964		

O p=0,5 corresponde a zero. Mas, não existe probabilidade = 0 por isto o valor $-1{,}4\times10^{-16}$.

EXERCÍCIOS DO CAPÍTULO 10

1. Imagine uma pergunta; escolha uma variável (e.g., densidade de uma planta ou animal). Obtenha 30 medidas desta variável (invente, se precisar). Faça uma tabela de frequência e o histograma. Cole a imagem do resultado obtido, como resposta.

2. Escolha uma variável qualquer (por exemplo: alguma medida, biomassa, temperatura, comprimento, inflação, desmatamento, poluição) e um fator (tempo, área, situação [antes, depois]) e:

   A. Faça um gráfico com eles. Não se esqueça de colocar título e nome dos eixos
   B. Estime a % simples e acumulada
   C. Faça gráfico em linha para elas.

3. Resolva o exercício 1 do capítulo 2

4. Resolva o exercícios 2 do capítulo 2.

Capítulo 11

TESTES PARA UMA AMOSTRA

TESTE Z PARA UMA AMOSTRA

O teste Z é utilizado em casos onde se conhece a média e desvio-padrão populacionais. Para este teste utilizamos a função TESTEZ (matriz,μ_o,sigma), onde 'matriz' é a amostra,intervalo de dados, ' μ_o' é a média populacional e 'sigma' é o desvio padrão populacional. Ela fornece o valor de probabilidade uni-caudal de um teste-z, para a média populacional, μ. Isto é, a probabilidade de μ ser maior que a média observada.

Utilizando o exemplo do bico do beija-flor do capítulo 4 referente ao teste Z para uma amostra, podemos ver o uso das funções do Excel na resolução do problema no quadro abaixo:

B1

	A	B	C	D	E	F	G	H	I	J
1	dados		Teste z para uma amostra - bicaudal							
2	59	69	Ho:	$\bar{x} = \mu$	H1:	$\bar{x} \neq \mu$				
3	56	56	μ=	65						
4	46	64	$\bar{x}$	58.65	=MÉDIA(A2:B11)					
5	61	60	σ=	10						
6	57	59	alfa=	0.05						
7	65	66	n=	20	=CONT.NÚM(A2:B11)					
8	49	58	Zcalculado	-2.8398063	=-INV.NORMP(TESTEZ(A2:B11,D3,D5))					
9	60	59	Zcrit=	1.95996398	=-INV.NORMP(D5/2)					
10	66	51	p=	0.00451409	=2*DIST.NORMP(D8)					
11	59	53	Decisão	**Rejeitar Ho**	=SE(ABS(D8)>D9,"Rejeitar Ho","Não rejeitar Ho")					

A função testez(A2:B11, D3, D5) resulta no valor de p= 0,983053. Entretanto, este valor é o complementar, já que a média de campo é bem menor que a média populacional → 1-0,983053= 0,016947.

Para obter o $z_{calculado}$ utilizamos a função 'INV.NORMP', z=-2.8398.

Para o teste unilateral à esquerda, haveria mudança apenas no $z_{crítico}$, que não é multiplicado por 2 (célula D21), nem o valor de 'p' é multiplicado por 2.

	A	B	C	D	E	F	G	H	I	J	K	L
13	dados		**Teste z para uma amostra - unicaudal (esquerda)**									
14	59	69	Ho:	$\bar{x} \geq \mu$		H1:	$\bar{x} < \mu$					
15	56	56	μ=	65								
16	46	64	σ=	10								
17	61	60	$\bar{x}$	58.65	=MÉDIA(A14:B23)							
18	57	59	n=	20	=CONT.NÚM(A14:B23)							
19	65	66	alfa=	0.05								
20	49	58	Zcalculado	-2.8398063	=-INV.NORMP(TESTEZ(A14:B23,D15,D16))							
21	60	59	Zcrit=	1.64485363	=-INV.NORMP(D19)							
22	66	51	p=	0.00225705	=DIST.NORMP(D20)							
23	59	53	Decisão	**Rejeitar Ho**	=SE(ABS(D20)>D21,"Rejeitar Ho","Não rejeitar Ho")							

Para o teste unilateral à direita, vãos supor que a média populacional é 55, cuja resolução está no quadro abaixo:

	A	B	C	D	E	F	G	H	I	J	K
25	dados		**Teste z para uma amostra - unicaudal (direita)**								
26	59	69	Ho:	$\bar{x} \leq \mu$		H1:	$\bar{x} > \mu$				
27	56	56	μ=	55							
28	46	64	σ=	10							
29	61	60	$\bar{x}$	58.65	=MÉDIA(A26:B35)						
30	57	59	n=	20	=CONT.NÚM(A26:B35)						
31	65	66	alfa=	0.05							
32	49	58	Zcalculado	1.63232962	=-INV.NORMP(TESTEZ(A26:B35,D27,D28))						
33	60	59	Zcrit=	1.64485363	=-INV.NORMP(D31)						
34	66	51	p=	0.05130503	=1-DIST.NORMP(D32)						
35	59	53	Decisão	**Não rejeitar Ho**	=SE(ABS(D32)>D33,"Rejeitar Ho","Não rejeitar Ho")						

TESTE T STUDENT PARA UMA AMOSTRA

Este teste é utilizado quando o desvio padrão populacional não é conhecido

Para a realização deste teste, devemos conhecer as funções:

Função 'DISTT': calcula o valor de 'p' para o teste t.

Distt(x,graus_liberdade,caudas)

- x - é o 't calc'
- Grau de liberdade-
- Caudas: coloque 1 para unicaudal e 2 para bicaudal

Exemplo: qual é o valor de 'p' para o tcalc = 5,43, com n = 50?

	M	N	O	P	Q	R	S	T	U	V
11	p=	1.78162E-06	=DISTT(5.42391261,49,2)							

Função 'INVT'- fornece o valor de t calculado

Esta função precisa de dois dados: INVT (probabilidade,graus_liberdade)

- Probabilidade do valor obtido ser aleatório
- Graus de liberdade

Exemplo: qual o valor do teste t para um p=0,001 e g.l.=12

	M	N	O	P	Q	R	S	T	U	V
11	p=	4.317791282	=INVT(0.001,12)							

Vamos utilizar o exemplo do teste t para uma amostra do capítulo 4. Para o calculo do t, utilizamos a equação 4.1, Como mostrado no quadro abaixo.

	A	B	C	D	E	F	G	H	I
1	Tilápias		**Teste t para uma amostra - bicaudal**						
2	23	42	Ho: $\bar{x} = \mu$			H1:	$\bar{x} \neq \mu$		
3	43	26	μ=	38					
4	22	39	$\bar{x}$	34	=MÉDIA(A2:B9)				
5	23	37	s=	8.03563492	=DESVPAD(A2:B9)				
6	40	30	n=	15	=CONT.NÚM(A2:B9)				
7	39	44	alfa=	0.05					
8	26	39	g.l.=	14	=D6-1				
9	37		t calc=	-1.92790408	=(D4-D3)/(D5/RAIZ(D6))				
10	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$		tcrit=	2.144786681	=INVT(D7,D8)				
11			p=	0.074401552	=DISTT(ABS(D9),D8,2)				
12			Decisão	**Não rejeitar Ho**	=SE(ABS(D9)>D10,"Rejeitar Ho","Não rejeitar Ho")				
13									

Os testes t unicaudais à esquerda e à direita estão nas imagens abaixo:

	A	B	C	D	E	F	G	H	I	J
15	Tilápias		**Teste t para uma amostra - unicaudal (esquerda)**							
16	23	42	Ho: $\bar{x} \geq \mu$			H1:	$\bar{x} < \mu$			
17	43	26	μ=	38						
18	22	39	$\bar{x}$	34	=MÉDIA(A16:B23)					
19	23	37	s=	8.03563492	=DESVPAD(A16:B23)					
20	40	30	n=	15	=CONT.NÚM(A16:B23)					
21	39	44	alfa=	0.05						
22	26	39	g.l.=	14	=D20-1					
23	37		t calc=	-1.92790408	=(D18-D17)/(D19/RAIZ(D20))					
24	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$		tcrit=	1.761310115	=INVT(2*D21,D22)					
25			p=	0.037200776	=DISTT(ABS(D23),D22,1)					
26			Decisão	**Rejeitar Ho**	=SE(ABS(D23)>D24,"Rejeitar Ho","Não rejeitar Ho")					
27										
28										

	A	B	C	D	E	F	G	H	I
29	Tilápias		**Teste t para uma amostra - unicaudal (direita)**						
30	23	42	Ho: $\bar{x} \leq \mu$			H1:	$\bar{x} > \mu$		
31	43	26	μ=	30					
32	22	39	$\bar{x}$	34	=MÉDIA(A30:B37)				
33	23	37	s=	8.03563492	=DESVPAD(A30:B37)				
34	40	30	n=	15	=CONT.NÚM(A30:B37)				
35	39	44	alfa=	0.05					
36	26	39	g.l.=	14	=D34-1				
37	37		t calc=	1.927904085	=(D32-D31)/(D33/RAIZ(D34))				
38			tcrit=	1.761310115	=INVT(2*D35,D36)				
39	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$		p=	0.037200776	=DISTT(ABS(D37),D36,1)				
40			Decisão	**Rejeitar Ho**	=SE(ABS(D36)>D37,"Rejeitar Ho","Não rejeitar Ho")				
41									

TESTE QUI-QUADRADO

O teste Qui-quadrado testa a relação entre duas variáveis qualitativas. O Excel possui três funções relativas ao Qui-quadrado: 'TESTE.QUI', 'DIST.QUI' e 'INV.QUI'

- A função **'TESTE.QUI'** calcula a probabilidade unicaudal da distribuição Qui-quadrado (p), a partir dos valores observados e esperados.

TESTE.QUI(intervalo_real,intervalo_esperado)

- A função "dist.Qui" calcula a probabilidade a partir do valor de Qui-quadrado. Exige o grau de liberdade também:

DIST.QUI(x,graus_liberdade)

Onde: x- é o valor no qual a distribuição será avaliada, graus_liberdade- é o número de graus de liberdade.

- A função 'inv.Qui' calcula o valor de Qui-quadrado a partir de 'p'.

INV.QUI(probabilidade,graus_liberdade)

Exemplo:

Foram obtidos os seguintes números de sementes: 350 lisas e amarelas, 91 rugosas e amarelas, 103 lisas e verdes e 52 rugosas e verdes. A freqüência esperada é 9/16, 3/16, 3/16, 1/16

	A	B	C	D	E	F	G
1	**Teste Qui-quadrado- uma amostra**						
2		Fo	Fe				
3	Lisa e ama	345	312.75	=B7*9/16			
4	Rugosa e a	86	104.25	=B7*3/16			
5	Lisa e verd	96	104.25	=B7*3/16			
6	Rugosa e v	29	34.75	=B7*1/16			
7	SOMA	556	556				
8	H0: O=E		H1: O≠E				
9	α=	0.05					
10	g.l.=	3					
11	p=	0.043504	=TESTE.QUI(B3:B6,C3:C6)			1	
12	χ^2 calc=	8.1247	=INV.QUI(B11,B10)				
13	X^2crit=	7.814728	=INV.QUI(B9,B10)				
14	Decisão	Rejeita Ho	=SE(B12>B13,"Rejeita Ho","Não rejeita Ho")				

EXERCÍCIOS DO CAPITULO 11

1. Faça o exercício 6 do capítulo 4, utilizando as funções do Excel.

2. Faça o mesmo para o exercício 7.

3. Faça o mesmo para o exercício 8.

4. Faça o mesmo para o exercício 9

Capítulo 12

TESTES PARA DUAS AMOSTRAS

Teste Qui-quadrado para duas amostras independentes- tabelas de contingência

A descrição do teste e do exemplo está no capítulo 5. As funções do Qui-quadrado para a tabela de contingência são as mesmas, já explicadas no capítulo anterior:

Exemplo:

	A	B	C	D	E	F	G	H
16	**tabela contingencia**							
17		Pequena	Grande	Total				
18	Varejista	31	44	75				
19	Atacadista	60	25	85				
20	Avulso	12	8	20				
21	Total	103	77	180				
22								
23	**ESPERADO**	Pequena	Grande	Total				
24	Varejista	43.0	32.0	75	43		32	
25	Atacadista	48.6	36.4	85	=B21*D19/D21		=C21*D19/D21	
26	Avulso	11.4	8.6	20	=B21*D20/D21		=C21*D20/D27	
27	Total	103.0	77	180				
35	H0: O=E		H1: O≠E					
36	α=	0.05						
37	g.l.=	2						
38	p=	0.000861	=TESTE.QUI(B18:C20,B24:C26)					
39	χ^2 calc=	14.11542	=INV.QUI(B38,B37)					
40	X^2crit=	5.991465	=INV.QUI(B36,B37)					
41	Decisão	Rejeita Ho	=SE(B39>B40,"Rejeita Ho","Não rejeita Ho")					

Teste Z: duas amostras para médias

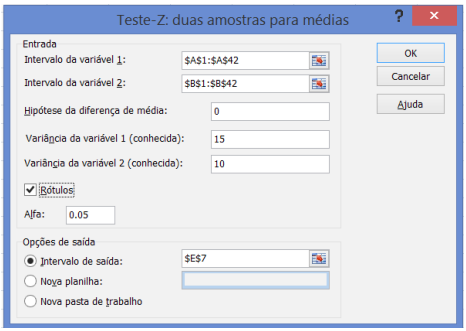
Este teste também está incluído no menu 'Análise de Dados'

Clicando nesta opção, aparecerá a caixa de diálogo

Exemplo: há diferença nos dias de germinação dentre dois lotes de

sementes de uma planta, com variância 15 e 10 respectivamente

Lote 1	25	34	29	29	35	30	28	31	31	34	32	28	35	34	33	30	31	32	28	31	34
	29	34	28	29	31	35	29	32	27	25	30	33	32	31	35	26	34	33	35	25	
Lote 2	38	32	37	36	39	37	37	34	33	35	38	32	33	36	33	35	32	33	37	37	
	39	39	39	38	38	38	38	38	38	38	38	34	32	37	36	36	38	39	37	36	35



Teste-z: duas amostras para médias		
	Lote 1	Lote 2
Média	30.90244	36.21951
Variância conhecida	15	10
Observações	41	41
Hipótese da diferença de média	0	
z	-6.80918	
P(Z<=z) uni-caudal	4.91E-12	
z crítico uni-caudal	1.644854	
P(Z<=z) bi-caudal	9.82E-12	
z crítico bi-caudal	1.959964	

Teste t para duas amostras

Para este teste, utilizamos as funções da distribuição t (DISTT e INVT), explicadas no capítulo anterior. A outra função necessária é 'TESTET', dá a probabilidade do teste t de student para duas amostras.

Teste t (matriz1,matriz2,caudas,tipo)

Onde: matriz 1 e matriz2 são os conjuntos de dados, caudas especifica se o teste é unicaudal (1) ou bicaudal (2), tipo designa qual teste vai ser

aplicado:1-teste t pareado, 2- teste t para variâncias homogêneas e 3- teste t para variâncias desiguais

Teste t para duas amostras com variância homogênea

O teste t para duas amostras com variâncias homogêneas está exemplificado no quadro abaixo, com as equações mostradas nas células adjacentes aos resultados. O exemplo utilizado é o mesmo do capítulo 5, que detalha as características do teste.

	A	B	C	D	E	F	G	H	I	J
1	Amostra 1	Amostra 2	**Teste t para duas amostras independentes- variancias homogêneas**							
2	1	24	Ho:	$\bar{x}_1 = \bar{x}_2$		H1:	$\bar{x}_1 \neq \bar{x}_2$			
3	35	16		Amostra1	Amostra2					
4	9	40	$\bar{x}$	16.28571	30	=MÉDIA(B2:B10)				
5	16	20	s^2=	131.2381	129.25	=VAR(B2:B10)				
6	24	34	n=	7	9	=CONT.NÚM(B2:B10)				
7	21	19	alfa=	0.05						
8	8	40	Fcalc=	1.015382	=E5/F5	Fcrit=	3.58058032	=INVF(E7,E6-1,F6-1)		
9		28	Decisão:	**Não rejeit**	=SE(ABS(E8)>H8,"Rejeitar Ho","Não rejeitar Ho")					
10		49	g.l.=	14	=E6+F6-2					
11			t calc=	2.385841	=INVT(TESTET(A2:A8,B2:B10,2,2),E10)					
12			tcrit=	2.144787	=INVT(E7,E10)					
13			p=	0.031717	=DISTT(E11,E10,2)					
14			Decisão:	**Rejeitar Hc**	=SE(ABS(E9)>E10,"Rejeitar Ho","Não rejeitar Ho")					

Teste t para duas amostras com variâncias heterogêneas

Também utilizamos o exemplo aplicado no capitulo 5, com peixe galo.

D31

	A	B	C	D	E	F	G	H	I	J	K	L	M
16	amostra1	amostra2	**Teste t para duas amostras indepentendes- variâncias heterogêneas**										
17	5.83	4.83	Ho:	$\bar{x}_1 = \bar{x}_2$		H1:	$\bar{x}_1 \neq \bar{x}_2$						
18	5.95	12.95		amostra1	amostra2								
19	4.24	1.24	$\bar{x}$	4.50	8.41	=MÉDIA(A17:A26)							
20	3.41	7.37	s^2=	1.295194	21.90406	=VAR(A17:A26)							
21	5.29	1.49	Fcalc=	16.9118	=E20/D2	Fcrit=	2.97823702	=INVF(0.05,10,10)					
22	2.93	12.93	Decisão:	**Rejeito Hc**	=SE(ABS(D21)>G21,"Rejeito Ho- as variâncias são distintas","Não rejeito Ho")								
23	4.06	13.06	s=	1.138066	4.680178	=DESVPAD(A17:A26)							
24	3.27	7.27	n=	10	10	=CONT.NÚM(A17:A26)							
25	5.78	10.78	alfa=	0.05									
26	4.22	12.22	g.l.=	10.06064	=(((D20/D24)+(E20/E24))^2)/((((D20/D24)^2)/(D24-1))+(((E20/10)^2)/(E24-1)))								
27			t calc=	2.573901	=INVT(TESTET(A17:A26,B17:B26,2,3),D26)								
28	$g.l.=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$		tcrit=	2.228139	=INVT(D25,D26)								
29			p=	0.027705	=DISTT(D27,D26,2)								
30			Decisão:	**Rejeitar Ho**	=SE(ABS(D27)>D28,"Rejeitar Ho","Não rejeitar Ho")								
31													
32												Teste-t: duas am	

Teste t para duas amostras pareadas

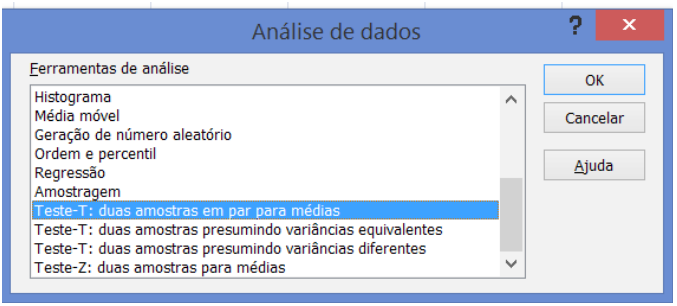
Este teste é utilizado quando as amostras estão emparelhadas, como em casos onde as amostras são mensuradas duas vezes- antes e depois de um experimento. Este teste determina se as observações feitas antes e após um tratamento têm probabilidade de serem provenientes de distribuições com médias de população iguais. Esta forma de teste-t não presume que as variâncias das duas populações sejam iguais. O teste t pareado utiliza as mesmas funções que os testes t independentes, como pode ser observado no exemplo abaixo:

	A	B	C	D	E	F	G	H	I	J	K	L
33						**Teste t para duas amostras pareadas**						
34	n	Antes	Depois	d	d^2	Ho:	μ1 - μ2 =0		H1:	μ1 - μ2 ≠0		
35	1	133	132	1	1	α=	0.05					
36	2	134	135	-1	1	n=	13					
37	3	135	136	-1	1	sd=	12.2845536	=RAIZ((E48-(G36*(D49^2)))/(G36-1))				
38	4	142	138	4	16	EPd=	3.40712214	=G37/RAIZ(G36)				
39	5	148	140	8	64	tcalc=	4.42511964	=D49/G38				
40	6	150	143	7	49	tcrit=	2.17881283	=INVT(G35,G36-1)				
41	7	164	144	20	400	Decisão	**Rejeitar Ho**	=SE(ABS(G39)>G40,"Rejeitar Ho","Não rejeitar Ho")				
42	8	170	150	20	400							
43	9	175	151	24	576							
44	10	179	153	26	676							
45	11	184	155	29	841							
46	12	185	155	30	900							
47	13	188	159	29	841							
48	SOMA				4766							
49	Média			15.07692								

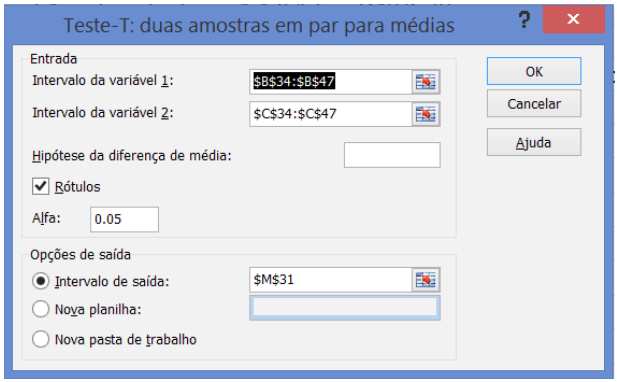
$$s_{\bar{d}} = \sqrt{\frac{\sum d^2 - n * \bar{d}^2}{n-1}} \qquad t = \frac{\bar{d}}{EP_{\bar{d}}} \qquad EP_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

O Excel disponibiliza dentro da ferramenta análise de dados, três rotinas de teste t.

- Teste t para variâncias iguais
- Para variâncias desiguais
- Para amostras pareadas



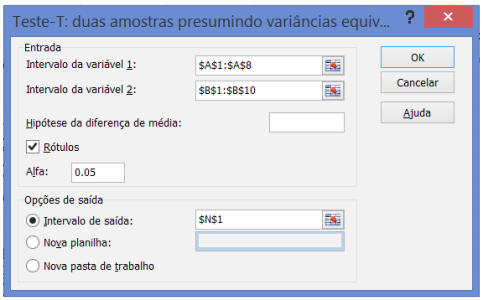
Teste-t: amostra dupla em par para médias ou teste t pareado



Teste-t: duas amostras em par para médias		
	Antes	Depois
Média	160.5384615	145.4615385
Variância	440.1025641	78.93589744
Observações	13	13
Correlação de Pearson	0.987540403	
Hipótese da diferença de média	0	
gl	12	
Stat t	4.425119637	
P(T<=t) uni-caudal	0.000414066	
t crítico uni-caudal	1.782287548	
P(T<=t) bi-caudal	0.000828132	
t crítico bi-caudal	2.178812827	

Teste-t: amostra dupla presumindo variâncias equivalentes

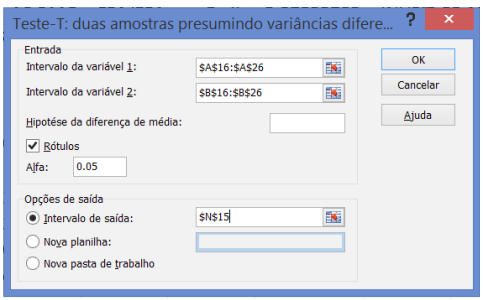
Fazendo o mesmo teste utilizando a ferramenta de análise do Excel:



Teste-t: duas amostras presumindo variâncias equivalentes		
	Amostra 1	Amostra 2
Média	16.28571429	30
Variância	131.2380952	129.25
Observações	7	9
Variância agrupada	130.1020408	
Hipótese da diferença de média	0	
gl	14	
Stat t	-2.385840584	
P(T<=t) uni-caudal	0.015858472	
t crítico uni-caudal	1.761310115	
P(T<=t) bi-caudal	0.031716944	
t crítico bi-caudal	2.144786681	

Teste-t: amostra dupla presumindo variâncias diferentes

Utilizando a ferramenta de análises do Excel:



Teste-t: duas amostras presumindo variâncias diferentes		
	Amostra1	Amostra2
Média	4.497830029	8.414282536
Variância	1.295194272	21.90406473
Observações	10	10
Hipótese da diferença de média	0	
gl	10	
Stat t	-2.571318104	
P(T<=t) uni-caudal	0.013913839	
t crítico uni-caudal	1.812461102	
P(T<=t) bi-caudal	0.027827678	
t crítico bi-caudal	2.228138842	

EXERCÍCIOS DO CAPÍTULO 12

1. Faça o exercício 1 do capítulo 5, utilizado as funções do Excel.
2. Faça o exercício 3 do capítulo 5, utilizado as funções do Excel.

Capítulo 13

ANÁLISE DE VARIÂNCIA

O Excel possui funções para a distribuição de Fisher, similares às distribuições normal e t-Student:

Distf- estima a probabilidade f. Isto é, você fornece o valor de f e obtém o 'p'. Esta função pode ser utilizada para verificar se dois conjuntos de dados possuem variabilidades distintas.

DISTF(x,graus_liberdade1,graus_liberdade2)

- X é o valor no qual se avalia a função.
- Graus_liberdade1 é o grau de liberdade do numerador.
- Graus_liberdade2 é o grau de liberdade do denominador.

INVF – fornece o inverso da distribuição de probabilidades f. Isto é o valor de F.

Invf(probabilidade,graus_liberdade1,graus_liberdade2)

- Probabilidade é uma probabilidade associada à distribuição cumulativa F.
- Graus_liberdade1 é o grau de liberdade do numerador.
- Graus_liberdade2 é o grau de liberdade do denominador

TESTEF – estima o 'p' entre a variância de duas amostras

Testef(matriz1,matriz2)

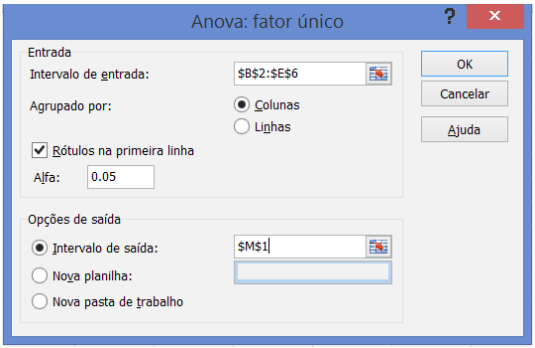
- Matriz1 é a primeira matriz ou intervalo de dados.
- Matriz2 é a segunda matriz ou intervalo de dados.

A ANOVA pode ser realizada no Excel pela aplicação direta das equações:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1							**Teste ANOVA one way**							
2	Observaçõe	V1	V2	V3	V4		Ho:	$\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4$			H1:	$\bar{x}_1 \neq \ldots$		
3	1	22	29	27	35		α	0.05						
4	2	18	25	19	30		SQ_{grupos}=	240.2333		=SOMA(B11:E11)				
5	3	21	27	25	29		SQ_{total}=	319.7333		=SOMA(B13:E16)				
6	4	23		22	31		SQ_{dentro}=	79.5		=H5-H4				
7	Média	21	27	23.25	31.25		$G.L._{grupos}$=	3		=CONT.VALORES(B2:E2)-1				
8	Soma	84	81	93	125		$G.L._{dentro}$=	11		=CONT.NÚM(B3:E6)-CONT.VALORES(B2:E2)				
9	Média Geral				25.53333		QM_{grupos}=	80.07778		=H4/H7				
10	n=	4	3	4	4		QM_{dentro}=	7.227273		=H6/H8				
11	SQ_{grupos}=	82.20444	6.453333	20.85444	130.7211		F_{calc}=	11.07994		=H9/H10				
12		=B10*(B7-$E9)^2					F_{crit}=	3.587434		=INVF(H3,H7,H8)				
13	SQ_{total}=	12.48444	12.01778	2.151111	89.61778		p=	0.001188		=DISTF(H11,H7,H8)				
14		56.75111	0.284444	42.68444	19.95111		Decisão:	**Rejeitar Ho**		=SE(ABS(H11)>H12,"Rejeitar Ho","Não rejeitar Ho")				
15		20.55111	2.151111	0.284444	12.01778									
16		6.417778		12.48444	29.88444									
17		=(B6-$E$9)^2												
18														

O Excel possui dentro do suplemento 'análise de dados' a rotina "ANOVA: fator único", que realiza a ANOVA – unifatorial (oneway).

Caixa de diálogo da ANOVA fator único



As variáveis respostas devem estar em linhas ou colunas contíguas (intervalo de entrada). Elas podem ter rótulos. Também se pode determinar o alfa utilizado. E há três opções de saída.

Ex.:

Observações	V1	V2	V3	V4
1	22	29	27	35
2	18	25	19	30
3	21	27	25	29
4	23		22	31

Anova: fator único

RESUMO

Grupo	*Contagem*	*Soma*	*Média*	*Variância*
V1	4	84	21	4.666667
V2	3	81	27	4
V3	4	93	23.25	12.25
V4	4	125	31.25	6.916667

ANOVA

te da varia	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F crítico*
Entre grup	240.2333	3	80.07778	11.07994	0.001188	3.587434
Dentro do:	79.5	11	7.227273			
Total	319.7333	14				

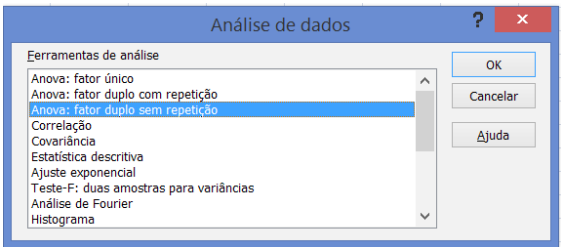
Anova: fator duplo sem replicação

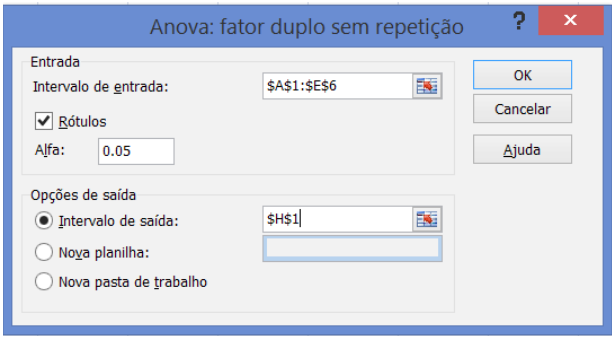
Este teste é utilizado para dados com uma observação para cada par de fatores.

Exemplo:

Locais\Placas	*M1*	*M2*	*M3*	*M4*
#1	3	2	7	1
#2	2	6	0	4
#3	0	6	3	9
#4	2	1	7	3
#5	1	0	0	3

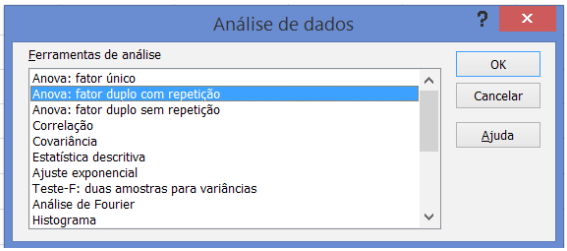
Esta análise abre a caixa de diálogo e produz os resultados, mostrados abaixo:

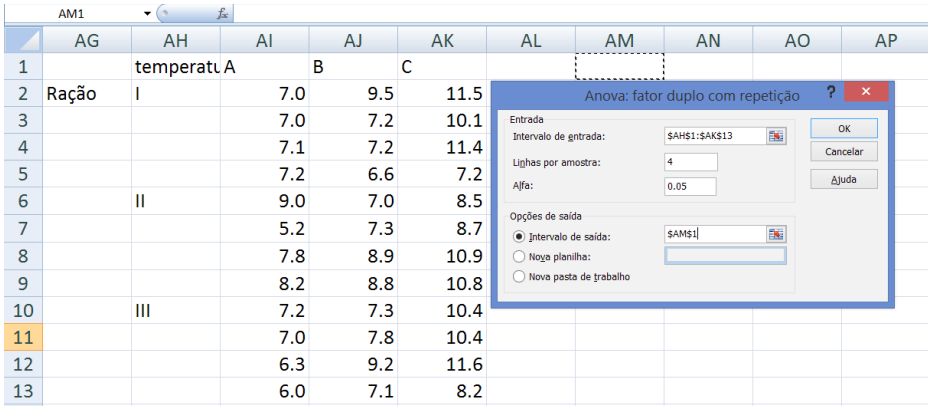




	H	I	J	K	L	M	N
1	Anova: fator duplo sem repetição						
2							
3	*RESUMO*	*Contagem*	*Soma*	*Média*	*Variância*		
4	#1	4	13	3.25	6.916667		
5	#2	4	12	3	6.666667		
6	#3	4	18	4.5	15		
7	#4	4	13	3.25	6.916667		
8	#5	4	4	1	2		
9							
10	M1	5	8	1.6	1.3		
11	M2	5	15	3	8		
12	M3	5	17	3.4	12.3		
13	M4	5	20	4	9		
14							
15							
16	ANOVA						
17	*te da varia*	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F crítico*
18	Linhas	25.5	4	6.375	0.789474	0.553907	3.259167
19	Colunas	15.6	3	5.2	0.643963	0.601427	3.490295
20	Erro	96.9	12	8.075			
21							
22	Total	138	19				

ANOVA FATORIAL – análise de variância com dois fatores com interação, pode se acompanhada pelas três imagens abaixo:





	AG	AH	AI	AJ	AK
1		temperatu	A	B	C
2	Ração	I	7.0	9.5	11.5
3			7.0	7.2	10.1
4			7.1	7.2	11.4
5			7.2	6.6	7.2
6		II	9.0	7.0	8.5
7			5.2	7.3	8.7
8			7.8	8.9	10.9
9			8.2	8.8	10.8
10		III	7.2	7.3	10.4
11			7.0	7.8	10.4
12			6.3	9.2	11.6
13			6.0	7.1	8.2

	AN	AO	AP	AQ	AR
1	Anova: fator duplo com repetição				
2					
3	RESUMO	A	B	C	Total
4	*I*				
5	Contagem	4	4	4	12
6	Soma	28.33692	30.53702	40.11823	98.99217
7	Média	7.08423	7.634256	10.02956	8.249348
8	Variância	0.008951	1.717876	4.014097	3.349338
9					
10	*II*				
11	Contagem	4	4	4	12
12	Soma	30.11608	32.00227	38.79653	100.9149
13	Média	7.529021	8.000567	9.699133	8.409574
14	Variância	2.626093	0.94092	1.657141	2.372268
15					
16	*III*				
17	Contagem	4	4	4	12
18	Soma	26.61614	31.47492	40.60787	98.69893
19	Média	6.654036	7.868729	10.15197	8.224911
20	Variância	0.332135	0.917265	2.07052	3.199274
21					
22	*Total*				
23	Contagem	12	12	12	
24	Soma	85.06914	94.01421	119.5226	
25	Média	7.089095	7.834518	9.96022	
26	Variância	0.948443	1.000325	2.151295	
27					
28					
29	ANOVA				
30	*te da varia*	*SQ*	*gl*	*MQ*	*F*
31	Amostra	0.241479	2	0.120739	0.07607
32	Colunas	53.27047	2	26.63524	16.78104
33	Interação	2.004219	4	0.501055	0.31568
34	Dentro	42.85499	27	1.587222	
35					
36	Total	98.37117	35		
37					

	AS	AT
30	*valor-P*	*F crítico*
31	0.92695	3.354131
32	1.84E-05	3.354131
33	0.864962	2.727765

	AM
30	*F crítico*
31	3.354131
32	3.354131
33	2.727765

EXERCÍCIOS DO CAPÍTULO 13

1. Sugestão: Resolva os exercícios do capítulo 6, utilizando as funções do Excel e compare os resultados

Capítulo 14

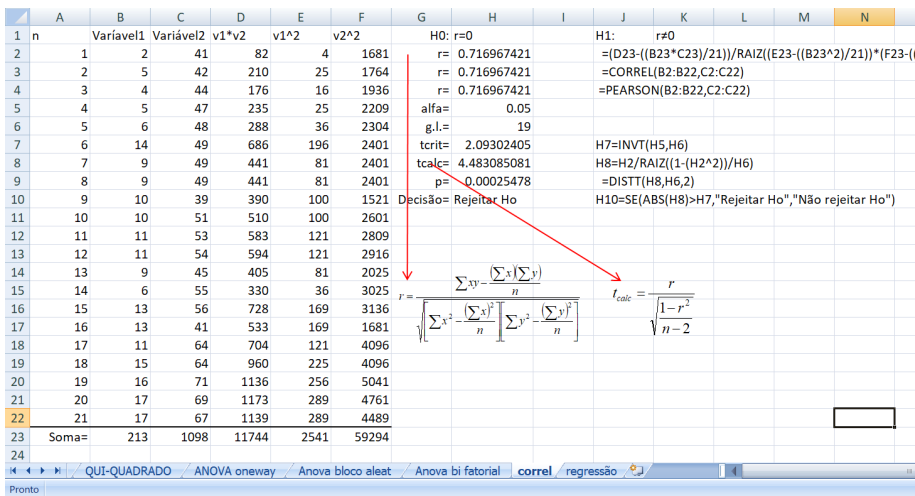
TESTES COM DUAS OU MAIS VARIÁVEIS

CORRELAÇÃO

Os cálculos de correlação podem ser feitos diretamente na planilha do Excel, utilizando a equação do r-Pearson. Outra maneira é utilizando as funções : "CORREL" e "PEARSON"

"=CORREL(matriz1,matriz2)" ou "=PEARSON(matriz1,matriz2)"

Estas três opções estão no exemplo a seguir:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	C
1	n	Varíavel1	Variável2	v1*v2	v1^2	v2^2		H0: r=0		H1: r≠0					
2	1	2	41	82	4	1681		r= 0.716967421		=(D23-((B23*C23)/21))/RAIZ((E23-((B23^2)/21))*(F23-((C:					
3	2	5	42	210	25	1764		r= 0.716967421		=CORREL(B2:B22,C2:C22)					
4	3	4	44	176	16	1936		r= 0.716967421		=PEARSON(B2:B22,C2:C22)					
5	4	5	47	235	25	2209		alfa= 0.05							
6	5	6	48	288	36	2304		g.l.= 19							
7	6	14	49	686	196	2401		tcrit= 2.09302405		H7=INVT(H5,H6)					
8	7	9	49	441	81	2401		tcalc= 4.483085081		H8=H2/RAIZ((1-(H2^2))/H6)					
9	8	9	49	441	81	2401		p= 0.00025478		=DISTT(H8,H6,2)					
10	9	10	39	390	100	1521		Decisão= Rejeitar Ho		H10=SE(ABS(H8)>H7,"Rejeitar Ho","Não rejeitar Ho")					
11	10	10	51	510	100	2601									
12	11	11	53	583	121	2809									
13	12	11	54	594	121	2916									
14	13	9	45	405	81	2025									
15	14	6	55	330	36	3025									
16	15	13	56	728	169	3136									
17	16	13	41	533	169	1681									
18	17	11	64	704	121	4096									
19	18	15	64	960	225	4096									
20	19	16	71	1136	256	5041									
21	20	17	69	1173	289	4761									
22	21	17	67	1139	289	4489									
23	Soma=	213	1098	11744	2541	59294									
24															

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} \qquad t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

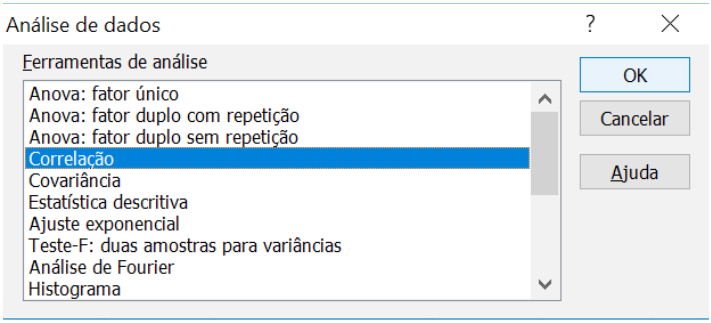
QUI-QUADRADO | ANOVA oneway | Anova bloco aleat | Anova bi fatorial | correl | regressão

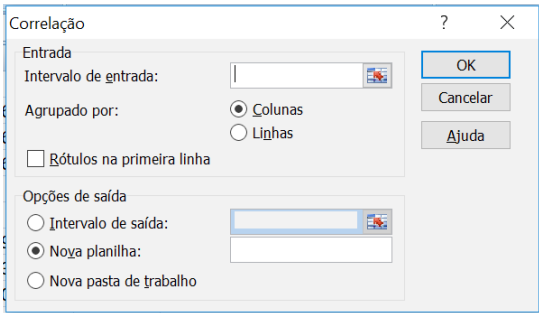
Pronto

O Excel possui ferramentas para analisar a correlação duas ou mais variáveis. Ele possui funções, gráficos e ferramentas de análise para isto

Outra maneira de testar a correlação de Pearson é através da rotina 'análise de dados':

No menu → <Ferramentas><Análise de dados>





	Varíavel1	*Variável2*
Varíavel1	1	
Variável2	0.716967	1

REGRESSÃO

Os parâmetros de regressão podem ser feitos diretamente na planilha do Excel, ou através das funções e rotinas disponíveis no programa.

Exemplo:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	n	X	Y	X*Y	X^2	y^	(y-y^)^2	β1=	1.8984	=(D23-((B23*C23)/A22))/(E23-((B23^2)/A22))				
2	1	2	41	82	4	38.5989	5.76513	βo=	34.8021	=(C23/A22)-(C24*(B23/A22))				
3	2	3	42	126	9	40.4973	2.25803	Ho:	β1=0	H1:	β1≠0			
4	3	4	44	176	16	42.3957	2.57371	alfa=	0.05					
5	4	5	47	235	25	44.2941	7.3218	g.l.=	19					
6	5	6	48	288	36	46.1925	3.26701	s^2=	7.04025	=G23/I5				
7	6	8	49	392	64	49.9893	0.97872	s(B1)=	0.13405	=RAIZ(I6/(E23-((B23^2)/A22)))				
8	7	9	49	441	81	51.8877	8.33881	tcrit=	2.09302	=INVT(I4,I5)				
9	8	9	49	441	81	51.8877	8.33881	tcalc=	14.1622	=I1/I7				
10	9	10	51	510	100	53.7861	7.76233	Decisão:	**Rejeitar**	=SE(ABS(I5)>I8,"Rejeitar","Não rejeitar")				
11	10	10	51	510	100	53.7861	7.76233							
12	11	11	53	583	121	55.6845	7.2065							
13	12	11	54	594	121	55.6845	2.83751							
14	13	11	54	594	121	55.6845	2.83751							
15	14	11	55	605	121	55.6845	0.46853							
16	15	13	56	728	169	59.4813	12.1193							
17	16	13	62	806	169	59.4813	6.34393							
18	17	14	64	896	196	61.3797	6.86608							
19	18	15	64	960	225	63.2781	0.52118							
20	19	16	64	1024	256	65.1765	1.38408							
21	20	17	69	1173	289	67.0749	3.70614							
22	21	17	73	1241	289	67.0749	35.1072							
23	Soma=	215	1139	12405	2593	1139	133.765							
24	Ho:													

$$s^2 = \frac{\sum\left(y-\hat{y}\right)^2}{n-2}$$

$$s(\beta_1) = \sqrt{\frac{s^2}{\sum x^2 - \frac{1}{n}(\sum x)^2}}$$

$$t = \frac{\beta_1}{s(\beta_1)}$$

$$\beta_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

O Excel tem funções para calcular a inclinação (β_1) e o intercepto (β_0) da reta:

=INCLINAÇÃO(val_conhecidos_y,val_conhecidos_x)

=INTERCEPÇÃO(val_conhecidos_y,val_conhecidos_x)

β1=	1.898396	=INCLINAÇÃO(C2:C22,B2:B22)
βo=	34.80214	=INTERCEPÇÃO(C2:C22,B2:B22)

A função 'PROJ.LIN' calcula os parâmetros de uma reta utilizando o método dos "mínimos quadrados". Esta função pode ser combinada com outras para calcular os parâmetros de outros modelos como polinomial, logaritmo, exponencial e série de potência. Uma vez que a função retorna uma matriz de valores, ela deve ser inserida como uma fórmula de matriz.

PROJ.LIN(val_conhecidos_y,val_conhecidos_x,constante,estatística)

Val_conhecidos_y- é o conjunto de valores de y.

Val_conhecidos_x- é um conjunto de valores x. Esta matriz pode incluir mais de uma variável

Constante- é um valor lógico que força ou não a constante β0 a se igualar a 0. Se VERDADEIRO, o intercepto será calculado normalmente, se FALSO, os valores serão ajustados para que o intercepto seja igual a zero.

Estatística – valor lógico, se VERDADEIRO, apresentará os dados estatísticos da regressão, se falso, retornará apenas os dados de inclinação e intercepto.

Esta é uma função matricial, voce deve selecionar o espaço da matriz antes de inserir a função. Depois de preencher os dados da função, aperte F2 e crtl, shift, enter.

O resultado sairá em duas colunas de 5 linhas, para uma regressão simples. A sequência de dados é

B1	B0
EP1	EP0
r2	EPy
F	g.l.
SQreg	SQerr

EP1 = erro padrão da inclinação

EP0 = erro padrão do inteercepto

EPy = erro padrão da estimativa de y.

F = valor de Fisher da ANOVA

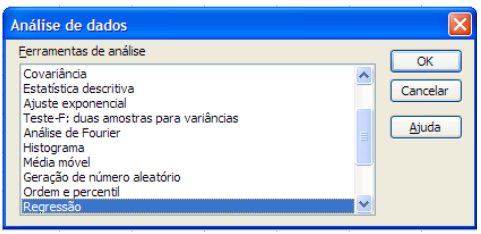
Utilizando o meso exemplo acima, obtem-se:

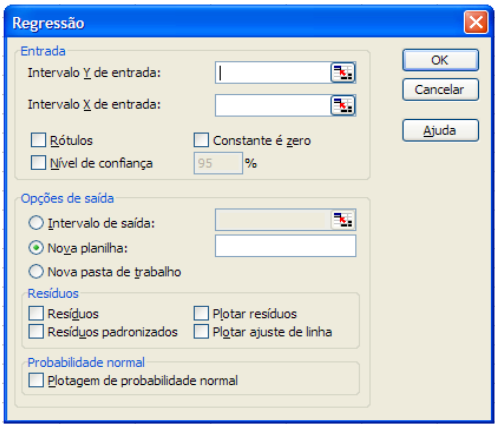
	A	B	C	P	Q	R	S	T	U	V	W
1	n	X	Y								
2	1	2	41	1.8984	34.8021	=PROJ.LIN(C2:C22,B2:B22,VERDADEIRO,VERDADEIRO)					
3	2	3	42	0.13405	1.48953						
4	3	4	44	0.91347	2.65335						
5	4	5	47	200.567	19						
6	5	6	48	1412.04	133.765						
7	6	8	49								

Use a função 'DISTF' para calcular o 'p' da regressão.

O Excel também tem uma rotina de análise de regressão:

<Ferramentas><análise de dados>





Entrada

- Intervalo de y: variável dependente
- Intervalo de x: variável independente
- Rótulos: ative quando a primeira linha das colunas tiver o título.
- Nível de confiança: ative quando quiser que apareça a coluna com o nível de confiança desejável, além do 95% que o programa fornece.
- Constante é zero: ative esta opção para forçar a reta a passar pela origem.

Opções de saída

Você tem três opções de saída: intervalo, planilha e pasta de trabalho.

Resíduos:

Há quatro opções

- Resíduos- o Excel produzirá uma planilha mostrando os valores previstos ($\hat{y}$) e o valor do resíduo (ou erro), que é $y - \hat{y}$.
- Resíduos padronizados- produzirá, além da tabela de resíduos acima, uma coluna com os resíduos padronizados. Isto é, os resíduos divididos pelo desvio padrão do mesmo.
- Plotar resíduos – produz um gráfico de dispersão com os valores dos resíduos.
- Plotar ajuste de linha- produz um gráfico de dispersão mostrando tanto os pontos dos dados reais quanto os pontos dos dados previstos, estes, aliás sobre a reta de regressão.

Probabilidade normal

- Plotagem da probabilidade normal- os dados que estão distribuídos normalmente parecem encontrar-se sobre uma reta.

Resultado

RESUMO DOS RESULTADOS

statística de regressão	
R múltiplo	0.95575
R-Quadrac	0.91347
R-quadrac	0.90891
Erro padrã	2.65335
Observaçõ	21

ANOVA

	gl	*SQ*	*MQ*	*F*	*e significação*
Regressão	1	1412.04	1412.04	200.567	1.5E-11
Resíduo	19	133.765	7.04025		
Total	20	1545.81			

	Coeficiente:	*rro padrão*	*Stat t*	*valor-P*	*i% inferior*	*% superior*	*ferior 95.0*	*perior 95.0%*
Interseção	34.8021	1.48953	23.3646	1.9E-15	31.6845	37.9198	31.6845	37.9198
X	1.8984	0.13405	14.1622	1.5E-11	1.61783	2.17896	1.61783	2.17896

REGRESSÃO MÚLTIPLA

A rotina da regressão múltipla é a mesma que para a regressão simples. Entretanto, as variáveis independentes devem estar em colunas adjacentes, para serem selecionadas ao mesmo tempo, quando colocar o intervalo de dados de X.

O Excel coloca nos resultados todas as variáveis selecionadas. Isto implica que será necessário fazer a matriz de correlação entre as variáveis e retirar aquelas que apresentam multicolinearidade.

EXERCÍCIOS CAPÍTULO 14

1. Faça o exercício 3 do capítulo 7, utilizando as funções do Excel.

PARTE III – ANÁLISE ESTATÍSTICA NO R

JOSÉ ROBERTO BOTELHO DE SOUZA
PABLO DAMIAN BORGES GUILHERME

Capítulo 15

INTRODUÇÃO AO R

O R é um programa estatístico e gráfico livre, que trabalha nas plataformas Windows, Unix, MacOs. Ele é adequado para manipulação de dados, cálculos e apresentação de gráficos (vernables et al. 2011). Com este programa você pode realizar várias operações de cálculo, em particular matrizes. Ele possui várias ferramentas para análise de dados, incluindo interface gráfica. A fonte do programa é aberta, e ele vem se aperfeiçoando ao longo do tempo, além de muitos programas estatísticos vêm sendo adicionados, cobrindo todas as áreas da ciência.

Uma das principais vantagens da utilização do R é garantir a reprodutibilidade das análises estatísticas, já que qualquer pessoa que possua as matrizes e o script (lista de comandos) pode reproduzir integralmente as an*á*lises.

BAIXANDO O R E INSTALANDO NO COMPUTADOR

O programa pode ser baixado da página do projeto R,

http://www.r-project.org/

Porque o R se chama R?

O nome se origina parcialmente das iniciais de seus autores (Robert Gentleman e Ross Ihaka) e parcialmente da linguagem 'S'. Já a linguagem 'S' é uma linguagem programável de alto nível e um ambiente para analise de dados e gráficos.

AJUDA E MANUAIS DO R

Há funções de ajuda no R:

```
demo() # coloque em parênteses as demonstrações que quer visualizar
help() # para ajuda online. Coloque no parêntesis a função que quer
entender.
help.start() # para ajuda no seu navegador
? #abre ajuda do R. Por exemplo ?lm()- ajuda sobre modelos lineares
?rnorm # distribuição normal
?t.test # teste de Student
```

Para saber algo peculiar, você deve colocar a palavra ou caractere entre aspas:

```
help ("[[")
help ("str")
help ("sum")
```

Para saber os argumentos de uma função digite 'args(função)'

```
args(lm)
function (formula, data, subset, weights, na.action, method = "qr",
    model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
    contrasts = NULL, offset, ...)
NULL
```

Existem vários manuais disponíveis sobre o R, os quais você pode encontrar na página do projeto:

http://cran.r-project.org/manuals.html

Venables W. N., Smith d. M. and the R Core Team 2018. An introduction to R: Notes on R: A Programming Environment for Data Analysis and GraphicsVersion 3.4.4 (2018-03-15)

Pacotes do R:

Pacotes (packages) são funções que você utiliza no R, alguns vêm com o programa e outros podem ser encontrados na internet. Qualquer pessoa pode criar seu próprio 'package'.

Para instalar um pacote você pode utilizar o menu do R, ou utilizar a função "install.packages()", colocando como argumento o "package" a instalar. Você vai escolher o repositório de onde fará o dowload, e o R fará a instalação.

Os pacotes ficam depositados na biblioteca – library-

```
library()
```

Citando o R:

Como é um programa livre, é vantajoso citar que utilizou o mesmo. Para citar a rotina do R que você utilizou digite

```
citation()

To cite R in publications use:

  R Core Team (2017). R: A language and environment for statistical
computing. R
  Foundation for Statistical Computing, Vienna, Austria. URL
  https://www.R-project.org/.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {R: A Language and Environment for Statistical Computing},
    author = {{R Core Team}},
    organization = {R Foundation for Statistical Computing},
    address = {Vienna, Austria},
    year = {2017},
    url = {https://www.R-project.org/},
  }

We have invested a lot of time and effort in creating R, please cite
it when using it
for data analysis. See also 'citation("pkgname")' for citing R
packages.
```

Diretório de trabalho

- Sempre iniciar o R no seu diretório de trabalho. Pois o R armazena seus dados em um arquivo chamado '.RData' que permanece gravado em disco entre sessões do R. Você pode ter múltiplos arquivos .RData em diferentes diretórios. Basta rodar o R em um determinado diretório para que o arquivo seja criado

  A função 'setwd' designa o diretório onde serão armazenados os dados.

```
setwd("C:\Users\seu_nome\Documents\R")
```

  Quando quiser saber onde seus dados estão sendo gravados, digite

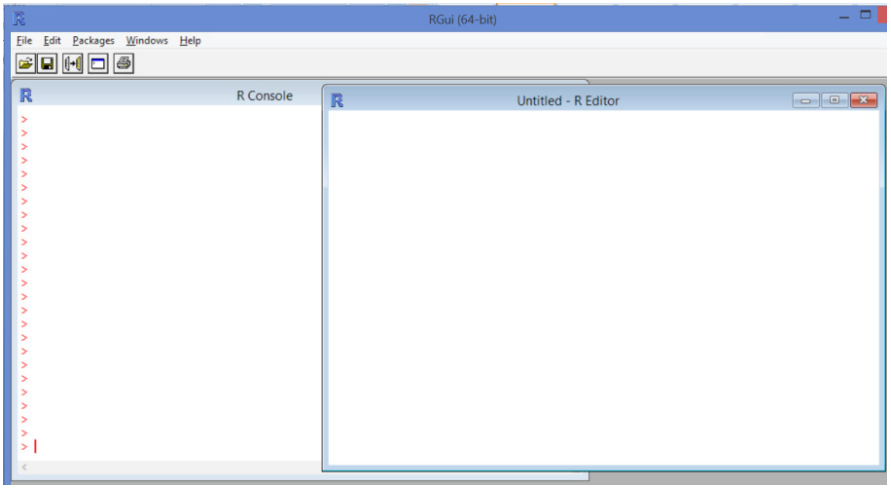
  >getwd()

  O diretório também pode ser definido através do menu do r.

Dicas

- Há distinção entre minúsculas e MAIÚSCULAS.

- Você pode ver o histórico de comandos colocados por você durante a sua sessão pressionando a tecla da seta para cima (↑). Isso é muito útil para verificar novamente os comandos anteriores ou reeditá-los.
- Os comandos elementares consistem de expressões ou designações. Os comandos são separados ou por ponto e vírgula (,) ou por uma nova linha.
- Comandos elementares podem ser agrupados juntos em uma nova expressão, limitada por parêntesis ('{' e '}').
- Você pode colocar um comentário iniciando com o símbolo do jogo da velha (#).
- Se o comando não acaba no final da linha coloque um sinal de +.
- Use 'del' para deletar.
- >q() para sair do sistema
- Crie um script para ir selecionando os scripts de cada análise



Cuidados

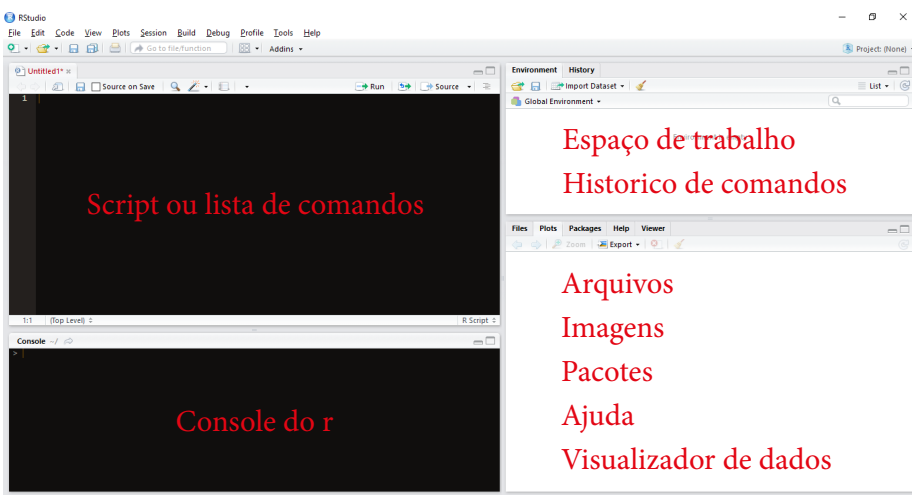
- Nomes de colunas podem ser modificadas em R
  - Espaços em branco se transformarão em "."
  - Caracteres não usuais se tornarão "."
  - Caracteres não usuais no inicio (por exemplo: %) se transformarão em "X".
- Espaços em branco na planilha serão considerados 'missing data'.

- 'NA' na planilha também será interpretado como missing data, exceto quando a coluna é formada por caracteres.
- Tudo após o símbolo "#" é ignorado e pode ser substituído.
- Lendo um arquivo ".csv" pode levar a resultados inesperados se conter vírgulas "," nos campos de caracteres, delimite os dados por tabulação.

Mesmo com uma interface gráfica simpática o programa R não é completo, a utilização de Ambientes de Desenvolvimento Integrado adicionam varias funcionalidades de forma gratuita. O mais conhecido deles é chamado de R-Studio (https://www.rstudio.com/), que trás inúmeras vantagens como:

- Realce automático de código
- Autocomplete
- Fechamento automático de (), [], {}, '' e "".
- Interface intuitiva de objetos, gráficos e script
- Ferramentas que facilitam a criação de funções e pacotes.
- Interação com html, latex entre outras linguagens.

Interface gráfica do R-Studio



INSERINDO DADOS NO R

O R trabalha com objetos, que são números, vetores, tabelas, matrizes, texto,... . Todos os objetos têm um nome associado a eles. Para dar nome

ao objeto no R, utilizamos o operador de atribuição "<-". Textos podem ser digitados após o prompt de comando(>), indicando que o R está pronto para receber e executar os comandos

Digitando

Você pode colocar os dados diretamente. Por exemplo, o conjunto de dados 'x', no caso o vetor x. Pois cada variável é um vetor.

```
> x <- c(3, 4.4, 5, 8.3, 9, 10.1)
> x
[1]  3.0  4.4  5.0  8.3  9.0 10.1
```

O dígito 1 entre colchetes indica que o conteúdo exibido inicia-se com o primeiro elemento de x.

Você também pode inserir apenas um número:

```
> y<-8
> y
[1] 8
```

Importando tabela

A função para importar tabelas é 'read.table()'.

A partir de uma tabela do Excel, você pode salvá-la como '*.txt', e depois abri-la no R.

```
Obj=read.table("arquivo.escolhido", header=TRUE)
```

Exemplo:

```
dados<-read.table("C:\Users\fulano\Documents\R\Animal1.txt", header=T)
```

O argumento header=F informa que o arquivo não possui cabeçalho. Se tiver, use o argumento header=T.

Colando dados no R

Para colar dados no R, você usa o comando scan(). Copia os dados (pode ser do Excel), e cola (ctrl-v) no R. Os dados devem estar separados por espaço.

```
exemplo_biomassa<-scan()
```

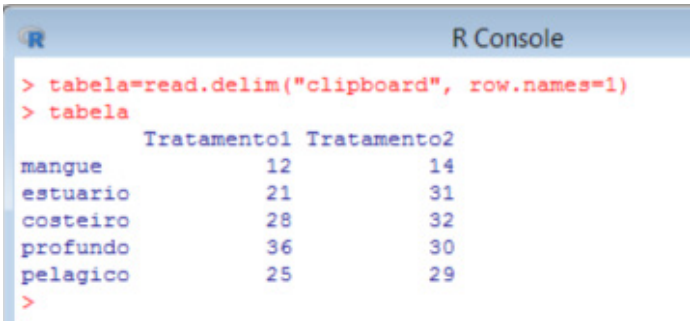
100 110 111 113 122 125 127 154 156 164 165 166 177 179 179 190 191 192 197 201 203 203 204 204 207 213 220 220 239 245 251 257 257 260 261 262 266 283 286 290 293 293 298 298

Outro modo de fazer é utilizando a função "read.delim():

> tabela=read.Delim("clipboard", row.names=1)

#Digita a linha acima no R, copia a tabela (clipboard), <enter>

>tabela



Para saber quais variáveis estão na tabela importada, utilize a função 'names':

>names(dados)

```
[1] "Comprimento" "Biomassa"
```

Como ler um arquivo Excel no R

Para isto é necessário instalar o pacote flipAPI

>install.package(devtools)

>devtools::install_github("Displayr/flipAPI")

#Para ler use os commandos abaixo

>library(flipAPI)

>downloadXLSX("nome do arquivo")

#Voce pode importar uma planilha especifica dentro do arquivo Excel.

>downloadXLSX("nome do arquivo", sheet="nome da planilha")

Salvando o histórico de comandos

O menu tem a opção de carregar e salvar toda a história de comandos utilizados. Ela também salva a área de trabalho.

Exportando dados

Voce pode exportar teus dados a partir do R criando um arquivo *.Csv

```
setwd("C:\\R")
attach(arquivo)
summary(arquivo)
write.csv(arquivo, file = "C:\\R\\arquivo.csv",row.names=TRUE)
```

OPERAÇÃO DO R

Nos exemplos acima x e y, um vetor e um número foram armazenados. Para ver o objeto basta digitar o nome dele no prompt

```
> x
[1]  3.0  4.4  5.0  8.3  9.0 10.1
```

Para ver os objetos que estão no R, basta listá-los "ls()" ou digitar "objects()"

```
> ls()
[1] "x" "y"
```

> objects()

```
[1] "x" "y"
```

Para saber se determinado arquivo está no diretório em uso utilizamos a função 'file.exists(nome do arquivo)

>file.exists("plantas.Txt")

```
[1] TRUE
```

Outros argumentos utilizados:

- Cabeçalho e separadores decimais e de cédulas:

```
Animal1<-read.table("nome do arquivo", header=T, dec=".",
```

- A opção header=T é usada quando a primeira linha dos dados refere-se ao nome da variável utilizada,
- Dec="." Quer dizer foi usado "." Na parte decimal dos números originalmente digitados,
- Sep='\t'"" quer dizer que quando o arquivo foi salvo com a extensão .Txt, optou-se pelo separador de colunas com tabulação. Outra opção é utilizar espaço ' '.

Ler arquivos em CSV

```
mytilopsis<-read.csv("nome do arquivo", header=T, dec=".",
```

- Podemos ler arquivos exportados pelo Excel com os dados separados por vírgula (*.Csv, comma separated value)
- Para remover objetos use 'remove(objeto)' ou 'rm(objeto)'
- Para remover tudo use:

```
remove(list=ls())
```

"OBJETOS" DO R

- Vetores, conjunto de dados de um parâmetro, são os objetos mais importantes em R
- Matrizes, generalização multidimensional de vetores
- Fatores, maneira compacta de manusear os dados.
- Listas, na forma de um vetor com vários elementos que não são necessariamente do mesmo tipo.
- Tabelas, matrizes em que as colunas podem ser de diferentes tipos
- Funções,

Vetores

Os vetores devem ter seus valores do mesmo tipo ou modo (mode) (numérico, complexo, lógico, caracteres).

Os vetores numéricos podem ser um valor, uma sequência, ou um conjunto de números:

```
q<- c(-7)
z<- c(4,5,6)
t<- c(2,6,7,11)
```

Os argumentos de c() podem ser escalares ou vetores:

```
> x<-c(z,7, 8, 9)
> x
[1] 4 5 6 7 8 9
```
>x

Sequências são tipos de vetores:

```
> G<-1:20
> G
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

Vetores lógicos

Vetores lógicos podem ter os valores TRUE, FALSE, e NA para não disponível.

```
> notas<-c(8, 5, 7, 7.2, 9, 3)
> aprovados<-notas>=7
> aprovados
[1]  TRUE FALSE  TRUE  TRUE  TRUE FALSE
```

Operações com vetores:

```
> x<-c(1,8,40,36,9.3, 2.2)
> x*2
[1]  2.0 16.0 80.0 72.0 18.6  4.4
> x+3
[1]  4.0 11.0 43.0 39.0 12.3  5.2
```

Fatores

Fator é um vetor usado para criar variáveis categóricas. Necessário para gráficos e análises estatísticas.

Suponha um exemplo de 10 taxa

```
taxa<-c("Aprio","diop","disp","eun","eus","Ner","prio","scol","spio")
```

O fator é criado utilizando a função “factor”()

```
taxa <-factor(taxa)
```

#.testa a atribuição do fator:

>is.factor(taxa)

```
[1] TRUE
```

Se no caso for um vetor de caracteres, 'sort' significa ordenar em ordem alfabética.

```
> sort(taxa)
 [1] Aprio diop  disp  eun   eus   Ner   prio  scol  spio  syl
Levels: Aprio diop disp eun eus Ner prio scol spio syl
```

Matrizes

São formadas utilizando a função matrix. Todos os vetores da matriz devem ser do mesmo tipo (numérico ou caracteres).

```
> xy <- matrix(1:9, nrow=3)
> xy
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> M<- matrix(1:16, ncol=4)
> M
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16

> matrix(letters[1:8], ncol=2)
     [,1] [,2]
[1,] "a"  "e"
[2,] "b"  "f"
[3,] "c"  "g"
[4,] "d"  "h"
```

```
#Os dados também podem ser ordenados por linha
M<- matrix(1:16, ncol=4, byrow=TRUE)
M

     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
[4,]   13   14   15   16

#Uma matriz de autovalores pode ser escrita com 1 na diagonal
A <-diag(1, nrow=2)
A

     [,1] [,2]
[1,]    1    0
[2,]    0    1

#Podemos extrair elementos das matrizes assim como se fossem vetores
M[3,1]

[1] 9
```

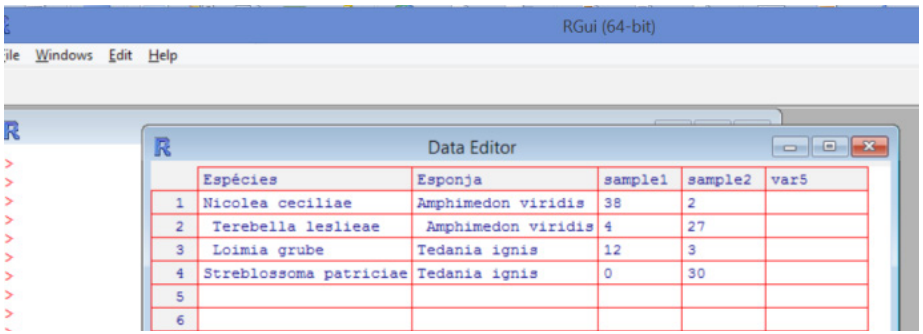
Data.frame

É uma matriz que aceita vetores numéricos e categóricos.

```
tbt<-data.frame(
  Espécies = c("Nicolea ceciliae", "Terebella leslieae", "Loimia
grube", "Streblossoma patriciae"),
  Esponja = factor(c("Amphimedon viridis", "Amphimedon viridis",
"Tedania ignis", "Tedania ignis")),
  sample1 = c(38, 4, 12, 0),
  sample2 = c(2, 27, 3, 30))

> tbt
                Espécies             Esponja sample1 sample2
1       Nicolea ceciliae  Amphimedon viridis      38       2
2     Terebella leslieae  Amphimedon viridis       4      27
3           Loimia grube       Tedania ignis      12       3
4 Streblossoma patriciae       Tedania ignis       0      30
```

Você pode editar o data.frame usando uma planilha

```
tbt<-edit(tbt)
```



Podemos extrair dados do 'data.frame', do mesmo modo que em matrizes

```
> tbt[1, 3]
[1] 38
```

Também podemos criar o 'data.frame' de modo direto:

```
levantamento <- data.frame(espécies= c("Nicolea ceciliae", "Terebella
leslieae", "Loimia grube", "Streblossoma patriciae"),
                           abundancia= c(2, 28,36,14),
                           biomassa= c(13, 5, 28.4, 37))


> levantamento
                espécies abundancia biomassa
1       Nicolea ceciliae          2      13.0
2     Terebella leslieae         28       5.0
3           Loimia grube         36      28.4
4 Streblossoma patriciae         14      37.0
```

Você pode verificar cada coluna isoladamente:

```
> levantamento$espécies
[1] Nicolea ceciliae       Terebella leslieae     Loimia grube
Streblossoma patriciae
Levels: Loimia grube Nicolea ceciliae Streblossoma patriciae Terebella
leslieae

> levantamento$abundancia
[1]  2 28 36 14

> levantamento$biomassa
[1] 13.0  5.0 28.4 37.0
```

Mostrando os dados em forma de proporções

```
> prop.table(levantamento$abundancia)
[1] 0.025 0.350 0.450 0.175
> prop.table(levantamento$biomassa)
[1] 0.15587530 0.05995204 0.34052758 0.44364508
```

FUNÇÕES

Denomina-se função ao conjunto de instruções que executam uma tarefa, que é usada com frequência. As funções podem ser acompanhadas ou não de argumentos.

- Funções em R sempre são acompanhadas com parênteses ()

  Sinais + – × /

  = = → Igual a

  "!=" → Não igual a

  Log, exp, sin, cos, tan, SQrt, min, max, length, sum, var

Exemplos das funções
Operações básicas

```
> x<-c(1,8,40,36,9.3, 2.2)

> 2*3
[1] 6

> x*2#repare que você multiplicou todo o conjunto x por 2.
[1]  2.0 16.0 80.0 72.0 18.6  4.4

> 2^3
[1] 8
> 3**3
[1] 27
```

Você também pode inserir duas funções ao mesmo tempo, basta separá-las por ponto e vírgula (,).

```
> 8+5; 3*24
[1] 13
[1] 72
```

Outras funções

Função	R name	Exemplo
Ordenar	>sort()	>x<-c(4.4,8.3,9,5,3,10.1) >sort(x) [1] 3.0 4.4 5.0 8.3 9.0 10.1
Diz qual a posição crescente dos números	>rank()	>rank(x) [1] 2 4 5 3 1 6
Soma	>sum() >colSums(nomeArquivo) >rowSums(nomeArquivo)	>sum(x) [1]39.8
Número de elementos (n amostral)	>length()	>length(x) [1]6

Função	R name	Exemplo
Máximo	>max()	>max(x) [1]10.1
Mínimo	>min()	>min(x) [1]3
Qual valor (which)	which.max(x) which.min(x) which()	>which.max(x) [1]6 #o sexto valor é o maior >which.min(x) [1]5# o quinto valor é o menor >which(x>5) [1] 2 3 6 # os valores nas posições 2,3 e 6 do conjunto, são maiores que 5.
Amplitude dos dados	range()	>range(x) [1]3 10.1
Mediana	median()	>median(x) [1]6.65
Média	mean (x) colMeans() rowMeans()	>mean(x) [1]6.6333
Quartis de x	quantile()	>quantile(x) 0% 25% 50% 75% 100% 3.000 4.550 6.650 8.825 10.100
Desvio padrão	sd()	>sd(x) [1]2.872397
Variância	>var()	>var(x) [1]8.250667
Normalizar os dados (subtrair a média e dividir pelo desvio-padrão).	scale()	>scale(x) [,1] [1,] -0.7775154 [2,] 0.5802354 [3,] 0.8239343 [4,] -0.5686307 [5,] -1.2649132 [6,] 1.2068897

Função	R name	Exemplo
		attr(,"scaled:center") [1] 6.633333 attr(,"scaled:scale") **[1] 2.872397**
Ln	log ()	>log(x) [1]1.481605 2.116256 2.197225 1.609438 1.098612 2.31255
$\log_{10}$	log10 ()	> log10 (x) [1] 0.6434527 0.9190781 0.9542425 0.6989700 0.4771213 1.0043214
Log de qualquer base	>log(x,base)	>log(64,base=2) **[1]** 6
Raiz quadrada	>SQrt ()	>SQrt(x) [1] 2.097618 2.880972 3.000000 2.236068 1.732051 3.178050
Seno	>sin( ){em radianos}	>sin(x)#{em radianos} [1]-0.9516021 0.9021718 0.4121185 -0.9589243 0.1411200 -0.6250706 >sin(pi)#próximo a zero, como esperado [1]1.224606e-16
Funções trigonométricas	cos, tan, asin, acos, atan,	>cos(1), tan(1), asin(1), acos(1), atan(1) [1]0.5403023 [1]1.557408 [1]1.570796 [1]0 [1]0.7853982
Euler elevado a um exponte	exp()	>exp(x) [1] 81.45087 4023.87239 8103.08393 148.41316 20.08554 24343.00942

Função	R name	Exemplo
função para criar textos	cat(...)	>cat("A média do vetor x,",-mean(x),"\t ,indica uma distribuição regular dos dados \n") A média do vetor x, 6.633333 ,indica uma distribuição regular dos dados
Função se	ifelse	>ifelse(x>5, "satisfatório","insuficiente") [1] "insuficiente" "satisfatório" "satisfatório" "insuficiente" "insuficiente" [6] "satisfatório"
Transpor tabela	t()	>t(nome do arquivo)

Sequência regular de dados

Há várias rotinas para gerar sequências de dados no r.

Por exemplo:

- Sequência de 1 a 10

```
> 1:10
 [1]  1  2  3  4  5  6  7  8  9 10
```

- Sequencia de 21 a 50

```
> 21:50
 [1] 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
43 44 45 46 47 48 49 50
```

Utilizando a função sequência seq()

```
> seq(1,10)
 [1]  1  2  3  4  5  6  7  8  9 10
> seq(1,10,2) # sequência de 2 em 2
[1] 1 3 5 7 9
> seq(from=5,to=15) #   Também pode nomear os parâmetros
 [1]  5  6  7  8  9 10 11 12 13 14 15
> seq(from=4,to=16,by=4)
[1]  4  8 12 16
> seq (-15, 15, by=3) #Fazendo sequencias em intervalos diferentes de
1    :
 [1] -15 -12  -9  -6  -3   0   3   6   9  12  15
> seq (-1, 1, by=.2)
 [1] -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
```

Utilizando a função repetição – rep():

```
> rep(1,4) # número 1 repetido quatro vezes
[1] 1 1 1 1

> rep(c(1,2,3), 3) #repetindo a numeração 1, 2 e 3 três vezes.
[1] 1 2 3 1 2 3 1 2 3

> c(rep(0,3),rep(1,4)) #combinação de repetições
[1] 0 0 0 1 1 1 1

> rep("a",5) #repetição de fatores
[1] "a" "a" "a" "a" "a"
```

PARA GERAR NÚMEROS ALEATÓRIOS

Um conjunto de dados aleatórios pode ser obtido pela função runif(numero de dado, min=,max)

```
> teste<-runif(10, min=30, max=65)
> teste
 [1] 53.18000 40.04635 31.43122 62.27546 37.21345 34.28024 59.21876
62.44482 50.77807 64.95353
```

'Rnorm()' gera números aleatórios com distribuição normal:

>rnorm(n,mean=, sd=)

>exemplo<-rnorm(100,mean=50,sd=5)

Como selecionar amostras aleatórias de um conjunto de dados:
Sample(x,size, replace)

- x: vetor com o conjjunto de amostras
- size: número de amostras a serem selecionadas
- replace: lógico, TRUE= com reposição, FALSE= sem reposição.

>sample(1:30, 5, replace=FALSE)

```
[1] 5 23 11 26 24
```

> lado<-c("frente","verso")

> sample(lado,10, replace=TRUE)

[1] "Verso" "frente" "verso" "frente" "frente" "verso" "verso" "frente"
[9] "Verso" "frente"

EXERCÍCIOS DO CAPÍTULO 15

1. Salve um arquivo de dados do Excel em '*.txt' e abra no R. Comente as possíveis dificuldades encontradas. Apresente as linhas de comando utilizadas.

2. Realize operações básicas com vetores e números no R. salve o script utilizado.

3. A partir de um dos vetores acima, estabeleça um vetor lógico.

4. Utilize funções básicas, como média, máximo, mínimo, raiz quadrada, com os vetores que você inseriu no R. Apresente as linhas de comando utilizadas

5. Crie uma matriz e realize operações com a mesma.

6. Crie um vetor que seja uma sequência. Faça operações com o vetor. Coloque as linhas de comando utilizadas e resultados obtidos.

7. Crie um fator e ordene em ordens crescente e decrescente.

8. Um restaurante recebeu as seguintes notas de seus clientes: 4,5; 8; 9; 5; 8,5; 6; 5; 9;9;6.; 8; 6,3; 9; 4; 7; 7,5. Descreva os resultados obtidos utilizando medidas de tendência central e de dispersão adequadas, se possível.

Capítulo 16

ESTATÍSTICA DESCRITIVA: PARÂMETROS E GRÁFICOS

EXAMINANDO A DISTRIBUIÇÃO DE UM CONJUNTO DE DADOS

Dado um conjunto (univariado) de dados, nós podemos examinar sua distribuição de várias maneiras. Os parâmetros dos dados podem ser obtidos através de funções simples, como: mean(arq), sd(arquivo), median(arquivo), entre outros. Outra maneira é fazer um sumário dos dados.

Exemplo:

> x <- c(3, 4.4, 5, 8.3, 9, 10.1)

> summary(x)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.550   6.650   6.633   8.825  10.100
```

Construindo uma tabela com frequencia relativa e absoluta:

>arq<-c(1,8,7,6,5,5,4,6,7,3,2,4,5,5,6,6,6,6,5,5,3,4)

>fo=table(arq)

>frel<-100×(fo/sum(fo))

>facum<-cumsum(frel)

>tabela<-cbind(fo,frel,facum)

>tabela

```
  fo      frel      facum
1  1  4.545455   4.545455
2  1  4.545455   9.090909
3  2  9.090909  18.181818
4  3 13.636364  31.818182
5  6 27.272727  59.090909
6  6 27.272727  86.363636
7  2  9.090909  95.454545
8  1  4.545455 100.000000
```

```
#A tabela pode ser transformada em data.frame:
data<-as.data.frame(tabela)
> is.data.frame(data)

[1] TRUE
```

GRÁFICOS

A interface gráfica do R chama-se GUI (graphical user interface), que funciona através de comandos e argumentos:

>tipo do gráfico (nome do vetor, ou conjunto de dados, ou equação,)

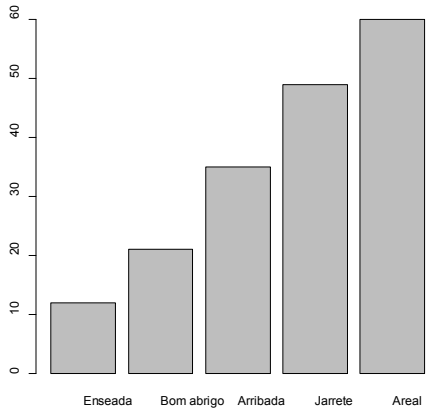
- Tipos de gráficos: >barplot(),>pie(), hist(), plot(),
- Rótulos:main="nome",xlab="nome",ylab="nome"
- Cores: col='blue'
- Tamanho dos caracteres: 'cex' (character expansion)
- Simbolo do plot (circulo, quadrado,...) Usa-se o 'pch'.
- Molduras do gráfico: bty="L"- retira as molduras direita e superior, box()-adiciona moldura.
- Limite das escalas dos eixos: 'xlim' e 'ylim'.
- Linhas, pontos: arrows()-adiciona seta, lines()-adiciona linha, points()-adiciona pontos
- Legenda: legend()

GRÁFICO EM BARRAS

>barplot()

Para fazer o gráfico em barras precisamos no mínimo de uma variável (vetor ou matriz) e de um fator (local, tempo,...)

```
>densidade<-c(12, 21, 35, 49, 60)
>locais<-c("Enseada", "Bom Abrigo", "Arribada", "Jarrete", "Areal")
>barplot(densidade, names.arg=locais)
```



Pode-se nomear as posições do vetor densidade através do comando names(). Assim teremos um nome em cada posição

>names(densidade)<-c("Enseada", "Bom Abrigo", "Arribada", "Jarrete", "Areal")

>densidade

Enseada	Bom abrigo	Arribada	Jarrete	Areal
12	21	35	49	60

#Agora basta apenas pedir o gráfico de barras do vetor:

>barplot(densidade)

#Rótulos:

Título no gráfico: use o argumento 'main'.

- Nomes nos eixos x e y use o argumento xlab=nome e ylab=nome
- Para inserir um texto dentro do gráfico, utilize o comando "text"

>text(posição eixo x, posição eixo y, "texto a inserir", outros argumentos)

- Para inserir um texto fora da área do gráfico, utilize o comando "mtext"

>mtext("texto", side=posição,line=posição, adj=posição entre esquerda(0) e direita (1)

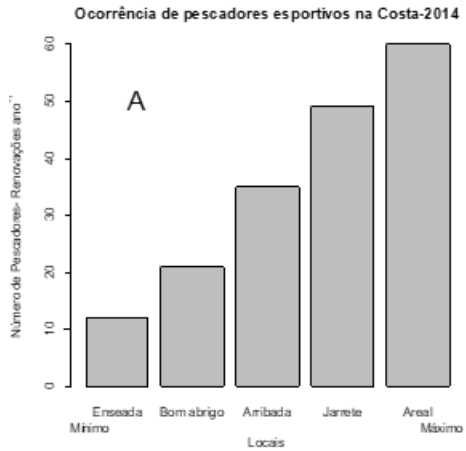
- Para inserir sobrescrito, pode-se utilizar o argumento"^".

>barplot(densidade, main="ocorrência de pescadores esportivos na costa-2014", xlab="Locais", ylab=expression(paste("Número De Pescadores- Renovações ano "^"-1")))

>text(1, 50,"a",cex=2)

>mtext("mínimo", side=1,line=2,adj=0)

>mtext("máximo", side=1,line=2,adj=1)



Tamanho dos componentes do gráfico

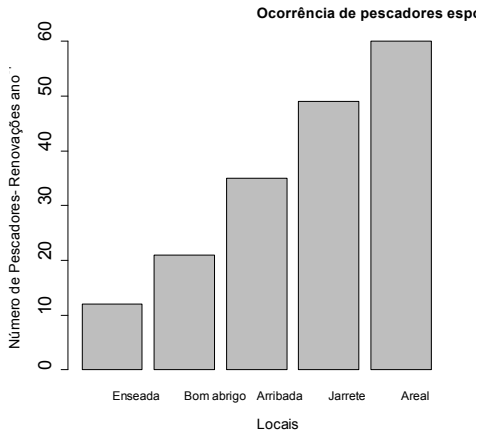
Para aumentar algo em 20%, por exemplo, utilizo o valor 1.2. A função 'cex' (character expansion) regula o tamanho dos componentes do gráfico. No R o padrão é 1, para aumentar use valores maiores, para diminuir, faça menor que 1.

Titulo: use 'cex.main=valor.

Para os eixos, use o argumento 'cex.axis=valor'.

Para o nome dos eixos, use 'cex.lab=valor'.

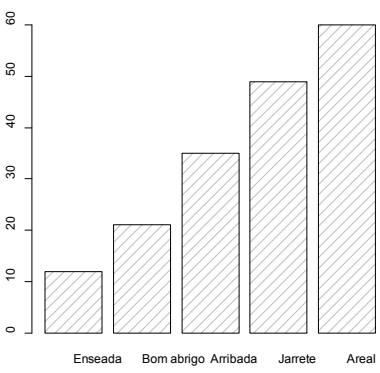
>barplot(densidade, main="Ocorrência de pescadores esportivos na Costa",xlab="locais", ylab=expression(paste("número de Pescadores-Renovações ano "^"-1")), cex.main=1.2, Cex.axis=1.5, Cex.lab=1.2)

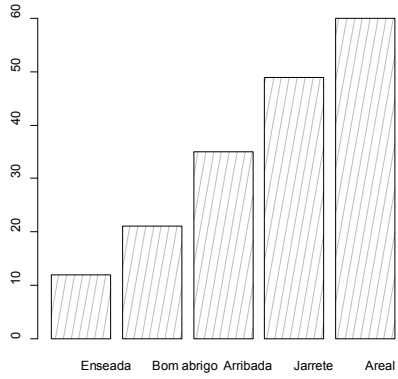


Os gráficos em barras podem ser preenchidos por linhas, o número de linhas é dado pelo argumento 'density', que significa números de linha por polegada, eles são positivos. O ângulo das linhas é dados pelo argumento 'angle', que varia de 0 a 360º no sentido anti horário.

>barplot(densidade, density=10)

> barplot(densidade, density=10, angle=80)



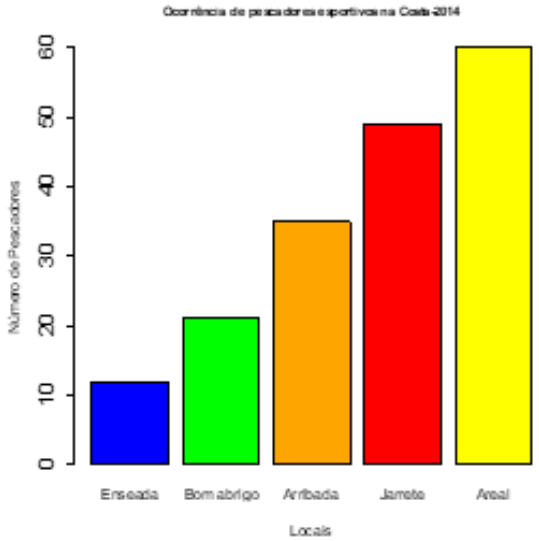


Para fazer gráficos coloridos usa-se o comando 'col':

- Cores: col='pink","blue","red","yellowl', "Green""orange", "sienna", "palevioletred1","royalblue2"
- Pode inserir como cores do arco-íris col=rainbow(número de cores)
- Ou pode inserir números 1=preto, 2=vermelho,3=verde,4=azul marinho,5=azul claro, 6=rosa,7=amarelo, 8=cinza,

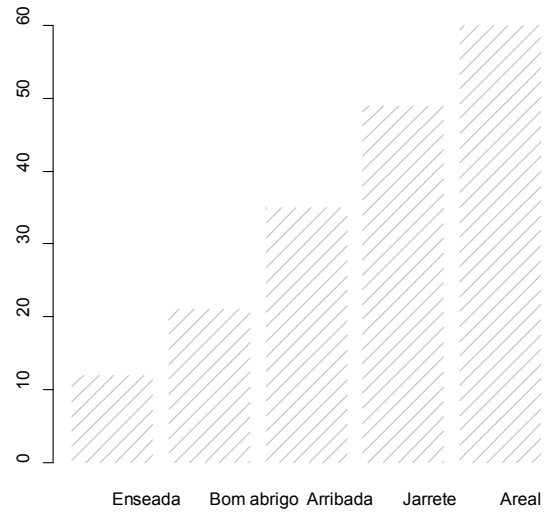
- Col=”NULL’ indica barra sem preenchimento.

>barplot(densidade, col=c(“blue”,”green”,”orange”,”red”, “yellow”), main=”Ocorrência de Pescadores Esportivos na Costa-2014”,xlab=”Locais”, ylab=”Número de Pescadores”, cex.main=0.8, Cex.axis=1.5)



Podemos tirar a borda das barras

>barplot(densidade,density=10, border=FALSE)



Ou manter

>barplot(densidade,density=10, border=TRUE)

GRÁFICOS DE BARRAS COM DUAS VARIÁVEIS.

Para isto precisamos inserir uma matriz:

>diversidade<-matrix(c(235,321,289,100,98,140), nrow=3, ncol=2)

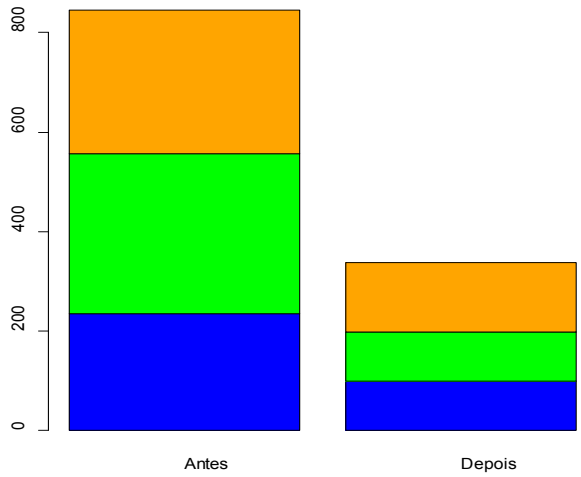
> diversidade

```
     [,1] [,2]
[1,]  235  100
[2,]  321   98
[3,]  289  140
```

Os nomes são dados pelo argumento 'dimnames'

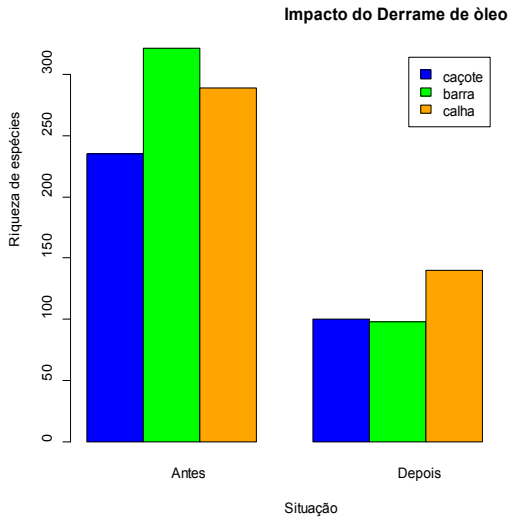
>diversidade<- matrix(c(235,321,289,100,98,140), nrow=3, ncol=2, dimnames=list(c("caçote", "barra", "calha"), c("Antes", "Depois")))

>barplot(diversidade, col=c("blue","green","orange"))



Se quisermos o mesmo gráfico, mas com as barras dispostas lado ao lado, no lugar de sobrepostas, como acima, utilizamos o argumento 'beside':

>barplot(diversidade, beside=TRUE, legend. Text=rownames(diversidade), col=c("blue","green","orange"),main="Impacto do Derrame de òleo", ylab="riqueza de espécies", xlab="Situação")



GRÁFICOS DE BARRAS, COM MÉDIA E DESVIO PADRÃO

Copie os dados das tabelas para o R

>amostras=read.delim("clipboard", row.names=1)

V1	V2	V3	V4
22	29	27	35
18	25	19	30
21	27	25	29
23	-	22	31

Media →

V1	V2	V3	V4
21	27	23.25	31.25

Desvio Padrão →

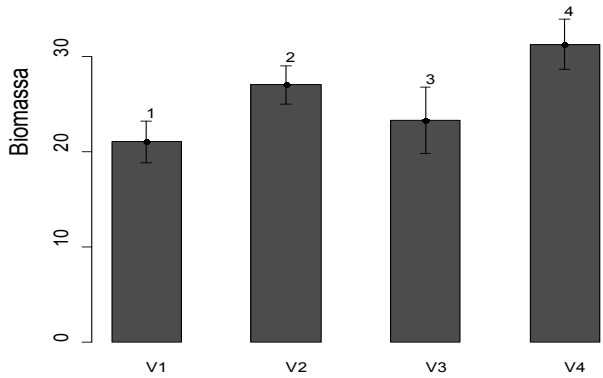
V1	V2	V3	V4
2.160246899	2	3.5	2.62995564

Número de Dados →

V1	V2	V3	V4
4	3	4	4

media <- as.matrix(read.delim("clipboard",row.names=1))

```
dp <- as.matrix(read.delim("clipboard",row.names=1))
n <- as.matrix(read.delim("clipboard",row.names=1))

# Calcule a amplitude da variação do erro
dp.sup<-media+dp
dp.inf<-media-dp

# Fazendo o gráfico
bp<-barplot(media,beside=T,ylim=c(0,max(dp.sup ×
1.15)),Ylab="Biomassa", cex.lab=1.5, Cex.axis=1.2)
points(bp,media,pch=16)
arrows(bp,dp.sup,bp,dp.inf, code=3,angle=90,length=0.05)
text(bp,dp.sup+1)
```



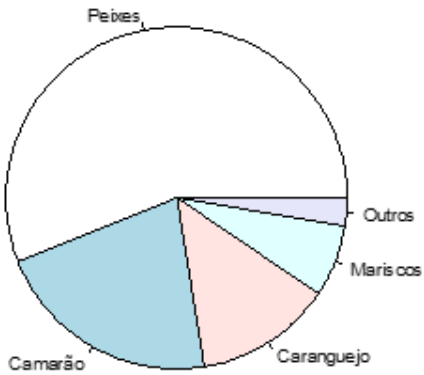
GRÁFICOS DE SETORES OU PIZZA

São gráficos em círculos, onde cada setor é representado pela proporção daquela variável em relação ao total (círculo):

Exemplo: venda de pescados no mercado da madalena:

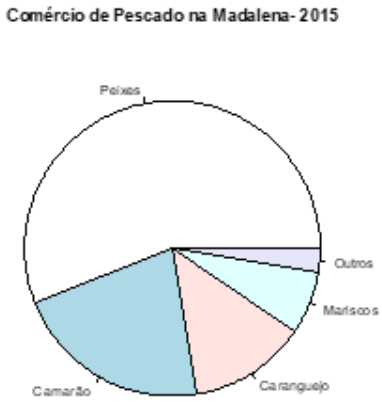
```
>pescados<-c(650, 250, 150, 80,30)
>names(pescados)<-c("Peixes", "Camarão", "Caranguejo", "Mariscos",
"Outros")
```

>pie(pescados)



Você pode adicionar o título com o gráfico aberto, basta utilizar o comando "title"

> title("Comércio de Pescado na Madalena – 2015")



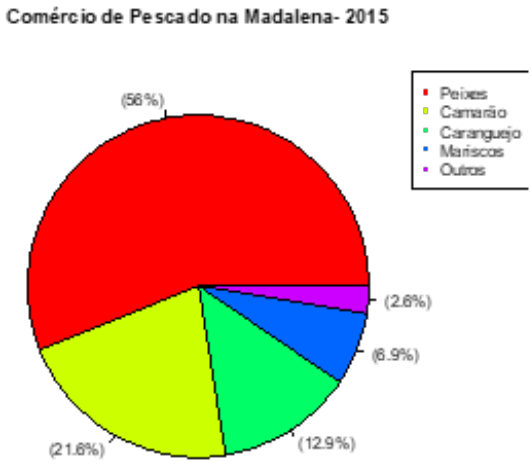
Para colocar a porcentagem usamos o argumento "labels"

>porc<-round(pescados*100/sum(pescados),1) # este último número indica decimais

>legenda<-paste("(",porc,"%)",sep="")

>pie(pescados, main="Comércio de Pescado na Madalena – 2015", labels=legenda, col=rainbow(5))

>legend(1,1, names(pescados), col=rainbow(5),pch=rep(20,6))



Podem-se colocar os nomes, com as porcentagens inseridas manualmente, o que é trabalhoso, quando há muitos setores.

>pie(pescados, main="Comércio de Pescado na Madalena – 2015",col=rainbow(5), init.angle=180)

text(locator(length(names(pescados))),legenda)

HISTOGRAMA

Um conjunto de dados poder ser agrupados em classes, formando uma tabela de frequencia. A representação gráfica desta tabela é um histograma.

Para fazer um histograma precisamos de um conjunto de dados. Podemos utilizar um vetor, no caso, o vetor z:

> z <- c (6.6, 4.5, 5.55,4.5, 5.1, 3.5, 6.8, 2.9, 7.2, 5.8, 7.3, 4.5, 4.8, 29.13, 28.89, 29.33, 25.8, 29.04, 22.27, 24.67, 25.41, 26.08, 13.64, 12.42, 5.9, 3.8, 6.45, 5.2, 6.2, 4.1, 6.5, 4.5, 9.2, 9.8, 8.5, 5.7, 6.9, 33.94, 3.8, 19.71, 21.06, 20.79, 21.65, 16.65, 22.78, 22.5, 28.03, 25.24, 23.72, 22.64, 21.93, 18.39, 22.6, 21.89, 21.16, 21.91, 19.46, 17.73, 22.43, 23.99, 20.82, 24.46, 21.27, 21.6, 22.22, 17.3, 15.8, 14.3, 13.62, 12.52, 13.82, 9.08, 7.83, 12.24, 13.44, 12.57, 11.25, 12.37, 13.72, 11.61, 12.6, 11.33, 10.23, 12.52, 12.76, 9.42, 12.8, 10.16, 14, 13.15, 12.39, 9.87, 12.75, 9.13, 8.58, 8.8, 4.1, 4.2, 5.1)

A função histograma tem 53 modos:

>hist()

Os argumentos utilizados:

Hist(z, breaks = “Sturges”,
Freq = NULL, probability = !Freq,
Include.Lowest = TRUE, right = TRUE,
Density = NULL, angle = 45, col = NULL, border = NULL,
Main = paste(“histogram of” , xname),
Xlim = range(breaks), ylim = NULL,
Xlab = xname, ylab,
Axes = TRUE, plot = TRUE, labels = FALSE,
Nclass = NULL, ...)

Z = é o vetor que se quer fazer o histograma

breaks = pode ser:

- Um vetor que dará os intervalos entre as células do histograma
- Um número que dará as células do histograma
- Um caractere e letras dando nome a um algoritmo para computar o número de células

- Um argumento para computar o número de células

  Nos últimos três casos o número é apenas uma sugestão.

  Freq= lógico, FALSE plotará em probabilidade, TRUE, frequencia absoluta

  Right= lógico, se 'TRUE, as células do histograma, são fechadas à direita, e abertas à esquerda.

  Col= cor

  Border= cor da borda das barras. O automático é a com padrão da cor de fundo.

  Labels= lógico ou character.

  Exemplo:

  >hist(z, breaks = 10, right = TRUE, col='green', border='blue3')

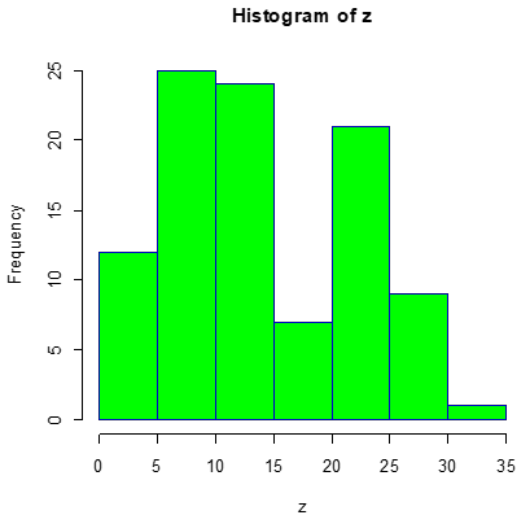


TABELA DE FREQUENCIA

>range(z)

```
[1]  2.90 33.94
```

>classes<-seq(2,37,by=5)

> classes

```
[1]  2  7 12 17 22 27 32 37
```

>intervalo = cut(z,classes)

>tabelafreq= table(intervalo)

> tabelafreq

Intervalo

```
(2,7]  (7,12] (12,17] (17,22] (22,27] (27,32] (32,37]
   25     17     21     15     15      5      1
```

>cbind(tabelafreq)

```
tabelafreq
(2,7]         25
(7,12]        17
(12,17]       21
(17,22]       15
(22,27]       15
(27,32]        5
(32,37]        1
```

>hist(z, breaks = 10, right = TRUE, freq=f, col='pink', border='blue',main= "mariscos vendidos na praia-2005", xlab="comprimento", ylab="frequência", cex.lab=1.5, Cex.axis=1.5)

>curve(dnorm(x,mean=mean(z),sd=sd(z)),lwd=2, add=t)

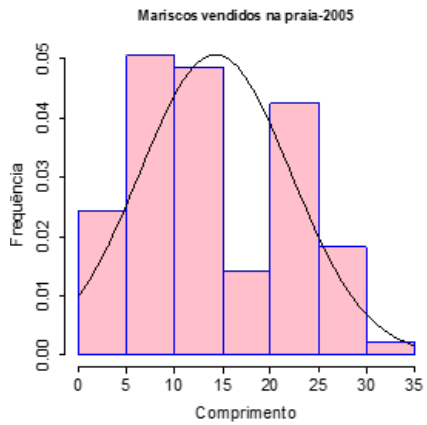


Gráfico de densidades de kernel

>d <- density(z)

<plot(d, main="mariscos vendidos na praia-2005")

Polygon(d, col="red", border="blue")

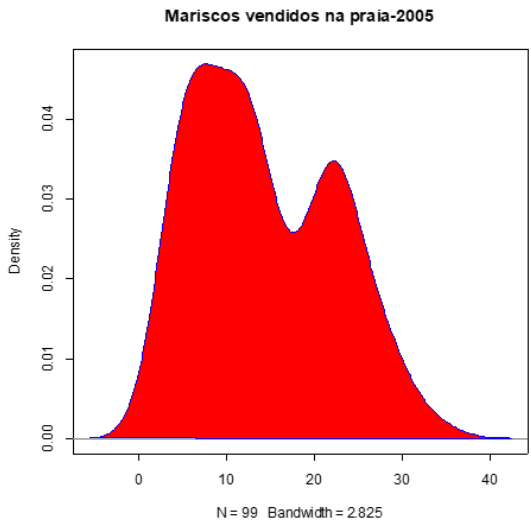


GRÁFICO DE DISPERSÃO

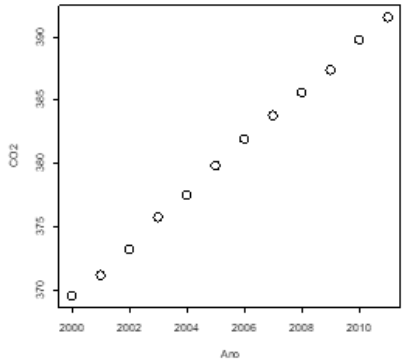
Primeiro precisamos das variáveis, que podem ser inseridas como vetores ou matrizes:

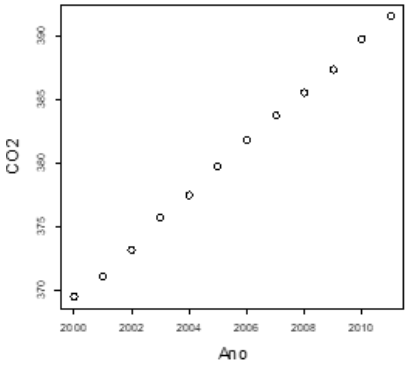
>ano<- c(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011)

>CO2<- c(369.52, 371.13, 373.22, 375.77, 377.49, 379.8, 381.9, 383.77, 385.59, 387.38, 389.78, 391.57)

>plot(ano,CO2, cex=2)

> plot(ano,CO2, cex.lab=2,cex=1.5)





Outros exemplos de `cex`: 'cex.axis' altera o tamanho dos eixos, 'cex.main' altera o tamanho do título.

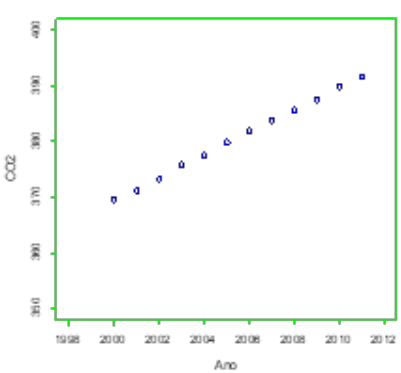
>plot(ano, CO2, xlab="Anos", ylab="Concentração CO2",main="CO2 atmosférico", cex=1.4,Cex.lab=1.4,Cex.axis=1.2, Col='blue')

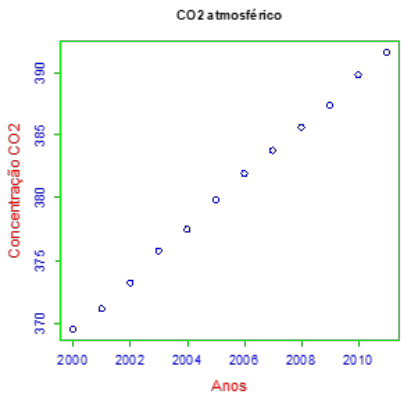
A função 'col' altera a cor. 'Col.axis' altera a cor dos eixos, 'col.lab' altera a cor do nome dos eixos. A função 'fg' dá a cor da borda (foreground):

Para dimensionar a amplitude dos eixos utilize os argumentos 'xlim' e 'ylim'

> plot(Ano,CO2, xlim=c(1998,2012), ylim=c(350,400),cex=1.2, Cex.lab=1.2, Col='blue3',fg='green')

>plot(Ano, CO2, xlab="anos", ylab="Concentração CO2",main="CO2 atmosférico", cex=1.4,Cex.lab=1.4,Cex.axis=1.2, Col='blue', col.axis='blue',col.lab='red',fg='green')





O comando 'font' designa a fonte utilizada no texto. O número 1 corresponde ao texto normal (default), 2 ao negrito, 3 ao itálico e 4 ao itálico negrito. Ainda, 'font 5' designa a fonte em si. Você também pode escolher a posição da fonte.

Font.axis- é a fonte utilizada no eixo.

Font.lab- fonte para os eixos 'x' e 'y'.

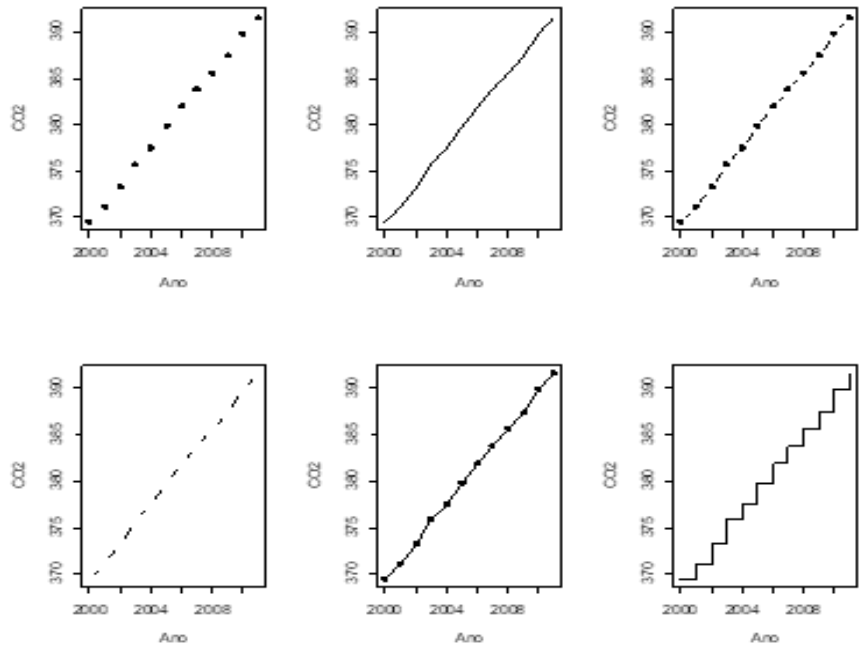
Font.main- fonte do título principal

Font.sub- fonte dos sub-títulod

Linhas ligando os pontos:

'Type'- este comando designa o tipo de linha: "p" para pontos, "l" para linhas, "b" para pontos e linhas, "c" para linhas descontínuas nos pontos, "o" para pontos sobre as linhas, "n" para nenhum gráfico, apenas a janela, "s" para linha quebrada ou em escada, "h" cada ponto formará uma linha vertical

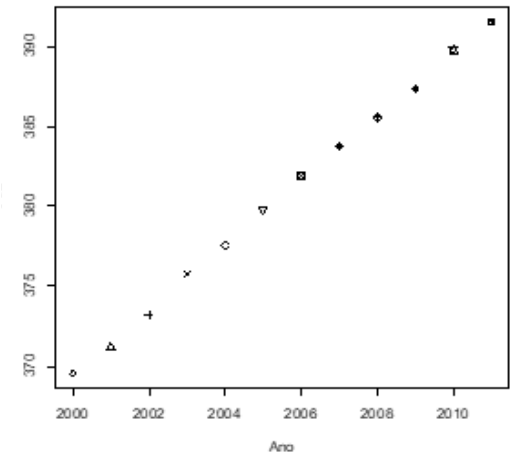
```
par(mfrow=c(2,3),pch=16)
>plot(ano,CO2, type='p')
>plot(ano,CO2, type='l')
>plot(ano,CO2, type='b')
>plot(ano,CO2, type='c')
>plot(ano,CO2, type='o')
>plot(ano,CO2, type='s')
```



Pontos, argumento 'pch=':

Pode-se mudar o padrão dos pontos utilizando o argumento 'pch=', onde 0=□, 1=○, 2= Δ, 3 = +, 4 = ×, 5= ◇, 6= ∇ . Números 7 a 14 são composições de símbolos obtidos por sobreposição dos símbolos básicos. Os números 15 to 18 são versões sólidas dos símbolos 0 a 4

>plot(Ano,CO2, pch=1:12)



Acrescentando outros pontos ao gráfico:

#Utiliza-se a função 'points()' para isto. Exemplo:

A<-c(2,4,8,3,16,9,14,9,6,31)

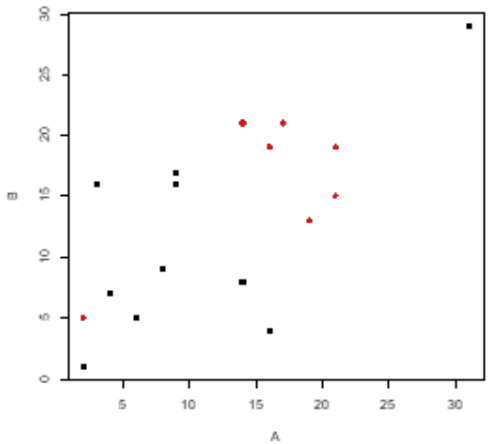
B<-c(1,7,9,16,4,16,8,17,5,29)

C<-c(21,2,17,16,14,19,21)

D<-c(19,5,21,19,21,13,15)

>plot(A,B,pch=15)

>points(C,D, pch=16,col=2)



PLOTANDO FUNÇÕES

>x<--15:15

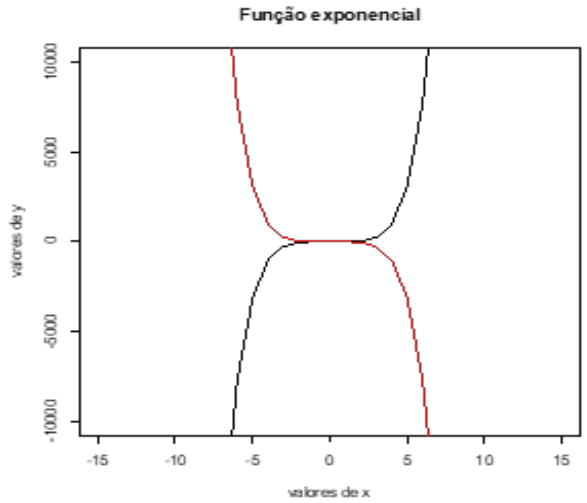
>y<-x^5

plot(c(-15,15),c(-10000,10000),type='n',xlab=na,ylab=na)

lines(x,y)

lines(x,-y, col='red')

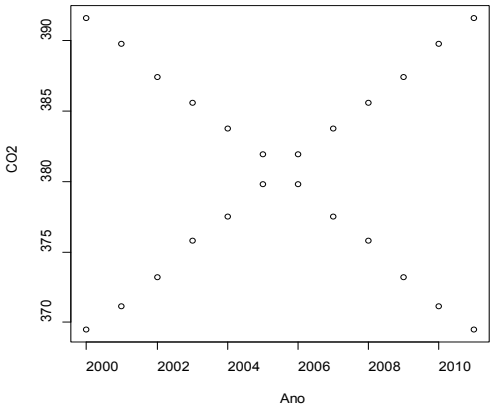
title("função exponencial",xlab="Valores de x", ylab="Valores de y")



#Fazer o plot reverso:

>plot(Ano,CO2)

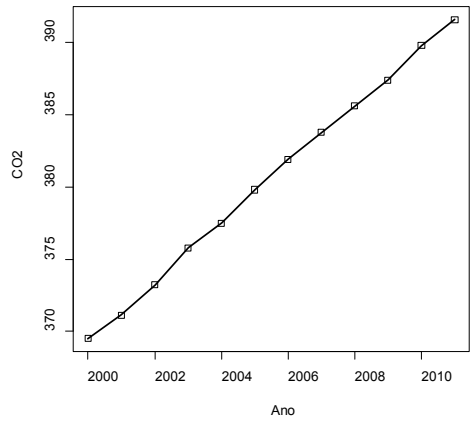
>points(rev(Ano),CO2)



A largura das linhas ou dos pontos podem ser mudados com o argumento 'lwd=', enquanto os estilos das linhas podem ser modificados com o argumento 'lty=':

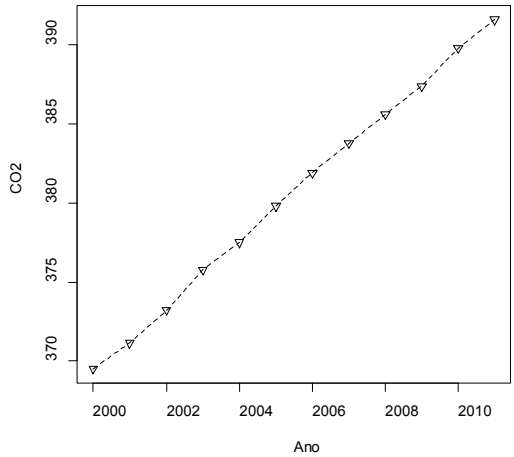
>plot(ano,CO2, pch=0)

>lines(ano, CO2, lwd=2)



>plot(ano,CO2, pch=6)

>lines(ano, CO2, lty=2)



Gráficos múltiplos:

Para mostrar vários gráficos junto utilize a função 'par(mfrow=c(nLinhas,nColunas))':

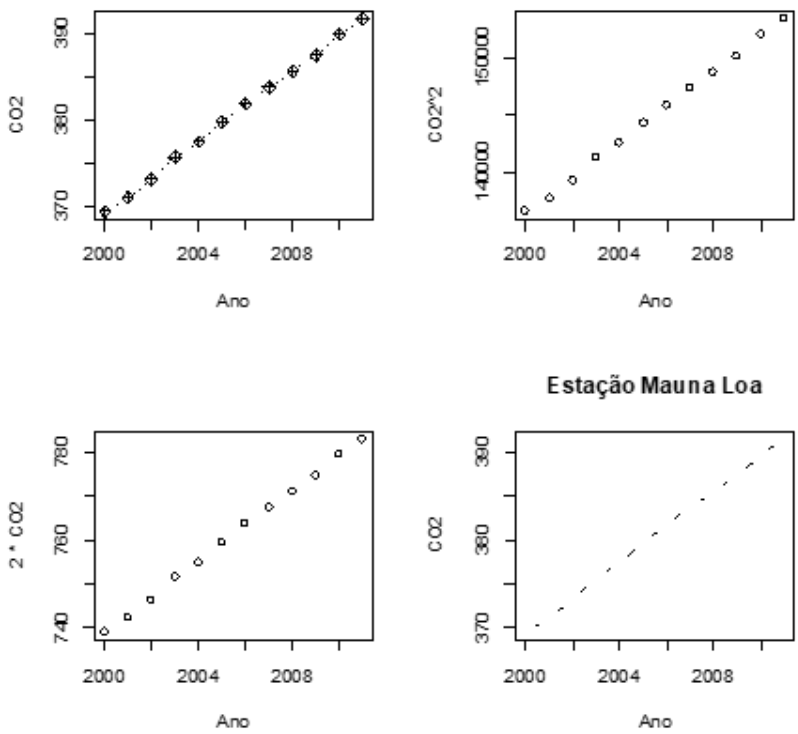
>par(mfrow=c(2,2)) → arranjo 2x2

>plot(ano,CO2, pch=9)

>plot(ano, CO2, lty=3)

> plot(ano,CO2^2)

```
>plot(ano,CO2, type='c')
>title("Estação Mauna Loa")
```

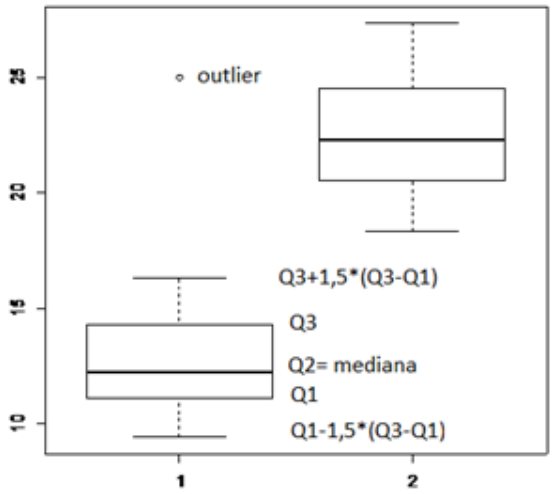


BOXPLOT

Os gráficos de boxplot propiciam a representação de medidas de tendência central (e.g., Média, mediana) com medidas de dispersão (e.g., Desvio padrão, quartis).

Por exemplo: sejam duas amostras:

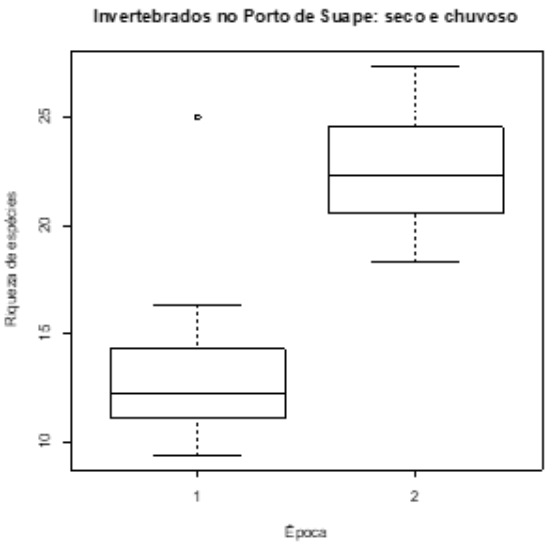
```
>p <- scan()
12.21 14.3 9.44 16.32 15.2 11.22 12.24 10.3 13.7 14.1 11.1 11.9 10.5 25
>q <- scan()
21.22 24.55 27.33 20.56 26.33 18.44 18.33 21.22 23.33 24.33
>boxplot(p,q)
```



#O significado de cada medida está realçado no gráfico.

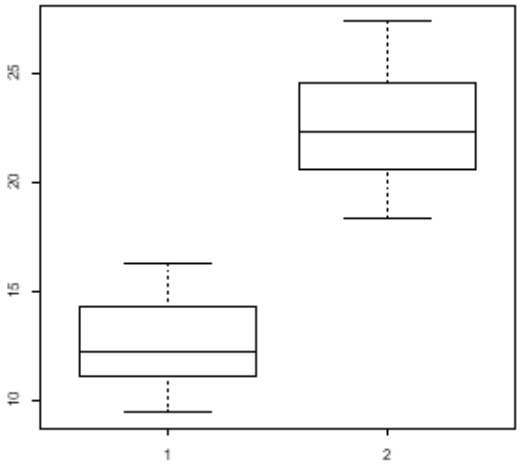
#Com o gráfico aberto, pode-se colocar o título

>title("Invertebrados no Porto de Suape: seco e chuvoso", xlab = "Época", ylab = "Riqueza de espécies")



#Para retirar o outlier usa-se o argumento 'outline' igual a FALSE.

>boxplot(P,Q, outline=FALSE)

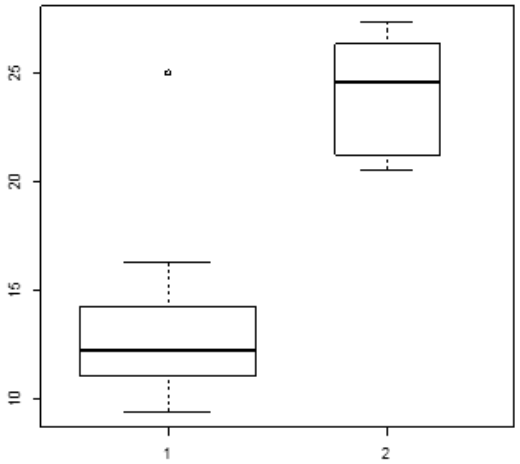


#A largura das caixas pode indicar o número de amostras quando usada a função 'varwidth'. O exemplo fica mais claro utilizando o vetor Q com cinco dados apenas:

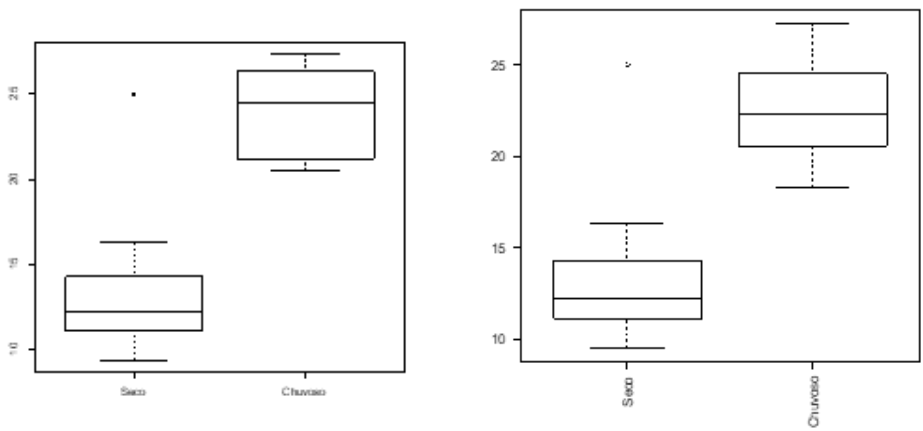
>Q <- scan()

21.22 24.55 27.33 20.56 26.33

>boxplot(P,Q, varwidth=TRUE)

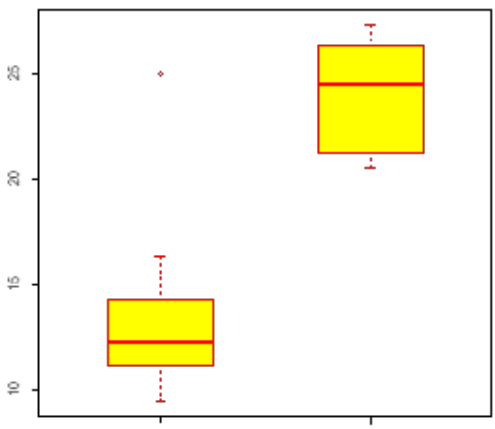


#Para dar nomes aos tratamentos:
boxplot(P,Q, names=c("Seco","Chuvoso"))
#O nome pode ficar na vertical 'las=2':
>boxplot(P,Q, names=c("Seco","Chuvoso"),las=2)
abline(h=mean(P,Q))



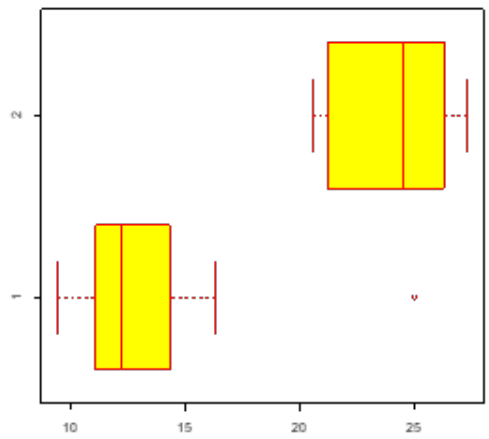
Para controlar a largura das caixas, 'boxwex'. Para mudar o tamanho das linhas limites 'staplewex'. Para colocar cor nas bordas 'border', para colocar cor dentro das caixas 'col':

>boxplot(P,Q, boxwex=0.5, Staplewex=0.1, Border="red", col="yellow")



Para mudar o sentido do gráfico para a horizontal 'horizontal', igual a TRUE.

>boxplot(P,Q, horizontal=TRUE, border="red", col="yellow")



Também podemos obter as estatísticas dos dados:

>boxplot.stats(Q)

```
$stats
[1] 18.330 20.560 22.275 24.550 27.330
minimo, Q1, mediana, Q3, máximo
$n
[1] 10
Número de amostras
$conf
[1] 20.28144 24.26856

$out
numeric(0)
pontos mariginais
```

EXERCÍCIOS DO CAPÍTULO 16

1. Os vetores a seguir correspondem ao número de larvas de pernilongos existentes em armadilhas colocadas em três bairros da cidades:

b1<-c(59,7,52,89,13,7 ,99, 17, 79, 25, 16,16), b2<-c(38,1,73,59, 25,93, 64,44,27,8, 38,96) e b3<-c(8, 18, 46, 52, 9, 45, 60, 19, 13, 9, 32,44). Faça gráficos em barra, de dispersão e Box-plot com estes dados personalize os mesmos com argumentos como 'cex' (Character expansion), 'col', 'xlab', 'ylab', 'main', 'pch",type',quando possível.

CAPÍTULO 17

MODELOS DE DISTRIBUIÇÃO DE PROBABILIDADES

DISTRIBUIÇÃO BINOMIAL

O R disponibiliza várias funções para a distribuição Binomial:

>dbinom(x, n, p, argumentos): calcula a densidade

>pbinom(q, n, p, argumentos): calcula distribuição acumulada

>qbinom(prob, n, p, argumentos): calcula o valor de x para a função 'p' acumulada.

Onde: 'x' é o número de sucessos, 'n' é o número de repetições, 'p' a probabilidade de sucesso, 'q' é o complemento de pl, 'prob' vetor contendo a probabilidade de 'n' repetições.

Exemplo: qual a probabilidade de uma cadela dar à luz, cinco fêmeas

>dbinom(5,5,0.5)

```
[1] 0.03125
```

Por outro lado, se eu quiser saber a probabilidade da fêmea dar à luz até três fêmeas:

>pbinom(3,5,0.5)

```
 [1] 0.8125
```

Isto é, a chance de nascer de zero a três fêmeas é de 81,25%

Para saber quantas fêmeas correspondem a uma probabilade de 30%, escrevo:

>qbinom(0.3,5,0.5)

```
 [1] 2
```

DISTRIBUIÇÃO POISSON

O R disponibiliza várias funções para a distribuição Poisson:

>dpois(x, lambda, opções) #calcula a densidade, isto é, a proporção daquele evento

>ppois(q, lambda, opções) #calcula distribuição acumulada

>qpois(prob, lambda, opções): calcula o valor de x para a função 'p' acumulada.

>rpois(n,lambda): calcula números aleatórios gerados pela distribuição de Poisson.

Onde 'x' é o número de ocorrências, 'lambda' é o número de eventos no intervalo considerado, 'q' é o vetor contendo os quantis, 'prob' vetor contendo a probabilidade, , n= números gerados.

Exemplo: a densidade média da amazônia é 2,54 habitantes/km^2. Qual a probabilidade de encontrar um habitante/km^2?

>dpois(1,2.54)

```
[1] 0.2003207
```

A densidade média da amazônia é 2,54 habitantes/km^2. Qual a probabilidade de encontrar um habitante/km^2 ou menos?

> ppois(1,lambda=2.54)

```
[1] 0.2791871
```

A densidade média da amazônia é 2,54 habitantes/km^2. Qual a probabilidade de encontrar um habitante/km^2 ou mais?

> ppois(1,lambda=2.54, Lower=FALSE)

```
[1] 0.7208129
```

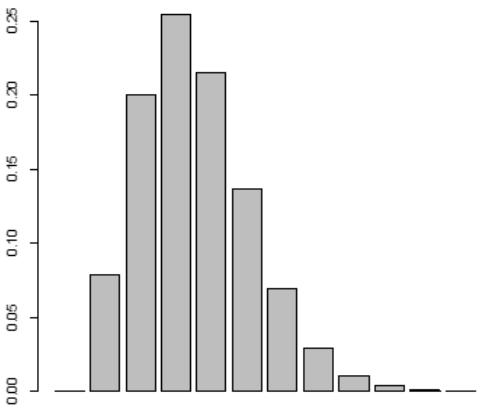
Qual a densidade que representa o acumulado de 60% das áreas da amazônia, cuja média é de 2,54 habitantes/km^2?

>qpois(0.6, 2.54)

```
[1] 3
```

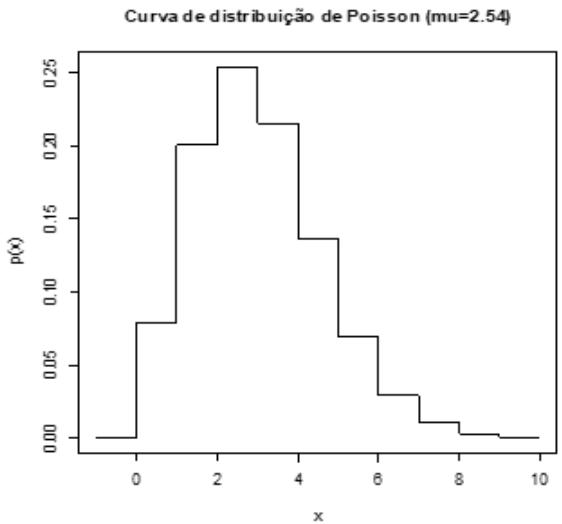
Distribuição de Poisson em um gráfico de barras

>barplot(dpois(-1:10, 2.54))



Qual a curva de distribuição de Poisson, para a densidade média acima (2,54 habitantes/km²)?

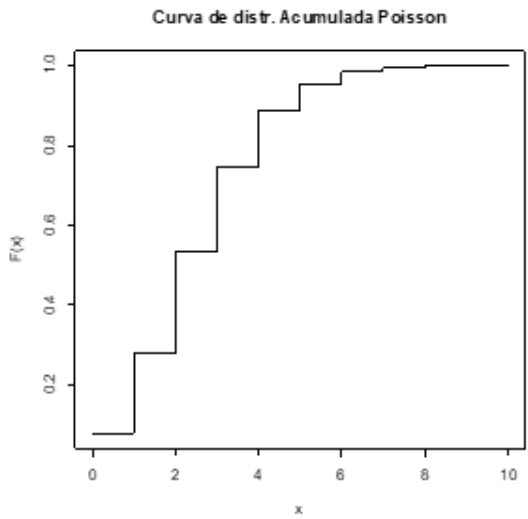
plot(-1:10,dpois(-1:10,2.54),Type="s",xlab="x",ylab="p(x)"**,main="curva de distribuição de Poisson (mu=2.54)")**



Qual a curva de Poisson acumulada, para a densidade média acima (2,54 habitantes/km²)?

x <- seq(**0, 1**0,**0.1**)

plot(x, ppois(x, 2.54**), Type = "s", ylab = "f(x)", main =** "Curva de distr. Acumulada **Poisson")**



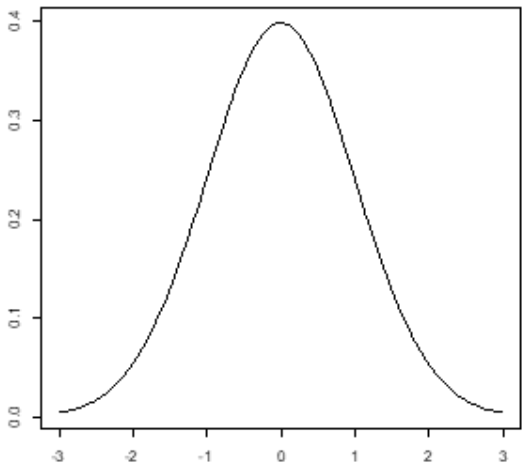
DISTRIBUIÇÃO NORMAL

Existem várias funções para a distribuição normal: dnorm, pnorm e rnorm.

A função dnorm() pode ser utilizada para construir o gráfico da distribuição normal:

>curve(dnorm(x,mean=0,sd=SQrt(1)),lwd=2,from=-3,to=3, xlab=na, ylab=na)

>title("Curva Normal Padrão")



Também podemos fazer um histograma com a curva normal associada:

> z <- c (6.6, 4.5, 5.55,4.5, 5.1, 3.5, 6.8, 2.9, 7.2, 5.8, 7.3, 4.5, 4.8, 29.13, 28.89, 29.33, 25.8, 29.04, 22.27, 24.67, 25.41, 26.08, 13.64, 12.42, 5.9, 3.8, 6.45, 5.2, 6.2, 4.1, 6.5, 4.5, 9.2, 9.8, 8.5, 5.7, 6.9, 33.94, 3.8, 19.71, 21.06, 20.79, 21.65, 16.65, 22.78, 22.5, 28.03, 25.24, 23.72, 22.64, 21.93, 18.39, 22.6, 21.89, 21.16, 21.91, 19.46, 17.73, 22.43, 23.99, 20.82, 24.46, 21.27, 21.6, 22.22, 17.3, 15.8, 14.3, 13.62, 12.52, 13.82, 9.08, 7.83, 12.24, 13.44, 12.57, 11.25, 12.37, 13.72, 11.61, 12.6, 11.33, 10.23, 12.52, 12.76, 9.42, 12.8, 10.16, 14, 13.15, 12.39, 9.87, 12.75, 9.13, 8.58, 8.8, 4.1, 4.2, 5.1)

> mean(z)

[1] 14.22434

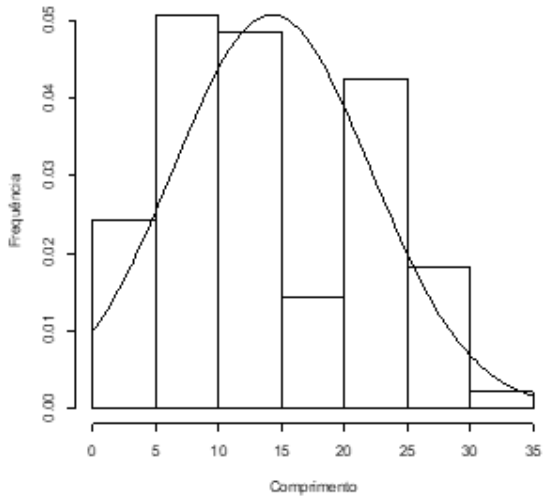
> sd(z)

[1] 7.869675

>hist(z, xlab=”comprimento”, ylab=”frequência”, freq=f, main=null)

>**curve(dnorm(x,mean=14.22434,Sd=7.869675),Lwd=2, add=t)**

Ou > curve(dnorm(x,mean=mean(z),sd=sd(z)),lwd=2, add=t)



A função 'pnorm' calcula a probabilidade de determinada observação um valor igual ou maior que x (a função pnorm() corresponde à integral da função dnorm()).

> pnorm(1.645)

[1] 0.9500151

> pnorm(1.96)

[1] 0.9750021

> pnorm(-1.96)

```
[1] 0.0249979
```

> pnorm(1.96)-Pnorm(-1.96)

```
[1] 0.9500042
```

Podemos utilizar o pnorm() para calcular a probabilidade de se obter um valor diferente da média:

Exemplo: numa população de guaiamum, com média de 65 mm de comprimento e desvio padrão de 5 mm, a probabilidade de um animal ter entre 65 e 70 mm é

> pnorm(70, mean=65, sd=5)

```
[1] 0.8413447
```

0,8413447-0,5= 0,3413447. Ou seja a probabilidade do animal ter entre 65 e 70mm é 34%

Por outro lado, utilizamos o qnorm() para calcular o valor da variável associado a determinado percentil

Exemplo: qual o comprimento do caranguejo que representa 84.12447% Da população¿

> qnorm(0.8413447, Mean=65, sd=5)

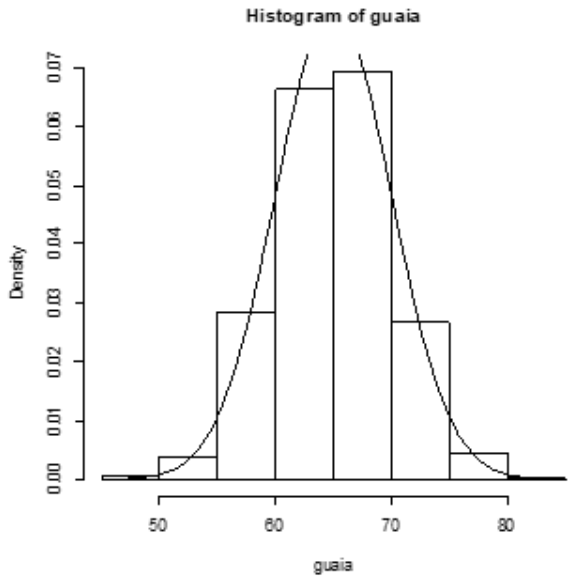
```
[1] 70
```

Para gerar números aleatórios com base na distribuição normal, utilizamos a função rnorm().

>guaia<-rnorm(1000, mean=65, sd=5)

>hist(guaia,freq=F)

>curve(dnorm(x,mean=65,sd=5),lwd=2, add=T)



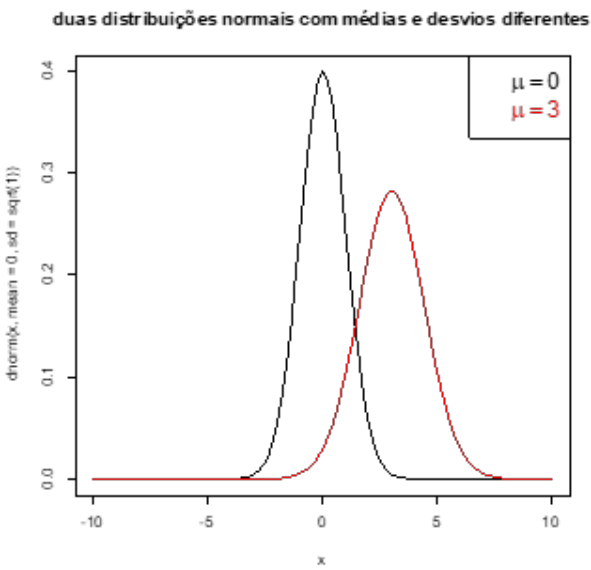
Comparando duas curvas

curve(dnorm(x,mean=0,sd=SQrt(1)),lwd=2,from=-10,to=10)

curve(dnorm(x,mean=3,sd=SQrt(2)),col=2,lwd=2,add=TRUE)

legend('topright',legend=c(expression(mu==0),expression(mu==3)),text.col=c(1,2),cex=1.5)

>title("duas distribuições normais com médias e desvios diferentes")



DISTRIBUIÇÃO T-STUDENT

Podemos usar três funções para a distribuição t-Student no r.

dt(x,df)- dá a probabilidade da função densidade da distribuição t no valor x

pt(x,df) – dá a função cumulativa da distribuição t para o valor x

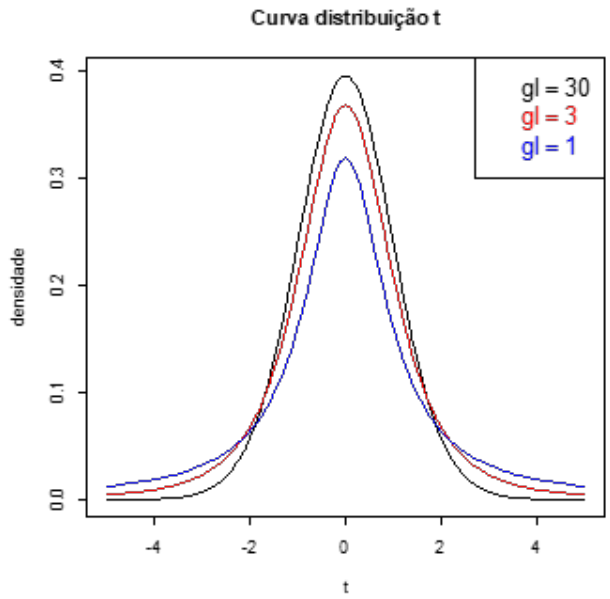
qt(x,df) – dá o valor do quantil da distribuição t para o valor x, que é o inverso da função pt().

Plotando a curva da distribuição t

>curve(dt(x,30), -5,5, col=1, xlab="t", ylab="densidade")

>curve(dt(x,3), col=2, add=TRUE)

```
>curve(dt(x,1),col=4, add=TRUE)
>legend('topright',legend=c(expression(gl==30),expression(gl==3),
expression(gl==1)), text.Col=c(1,2,4),cex=1.5)
>title("curva distribuição t")
```



EXERCÍCIOS DO CAPÍTULO 17

1. Responda os exercício 1 do capitulo 2, utilizando as funções do R.
2. Responda os exercício 2 do capitulo 2, utilizando as funções do R.
3. Responda os exercício 7 do capitulo 2, utilizando as funções do R.

Capítulo 18

TESTES PARA UMA E DUAS AMOSTRAS

TESTES PARA UMA AMOSTRA

Teste Qui-quadrado de aderência

Este teste é executado pela função 'chiSQ.test()":

chiSQ.test(x, y = null, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE),

Argumentos

- x = vetor numérico, matriz ou fator,
- y = vetor numérico, ignorado se x for matriz, também pode ser um fator do mesmo tamanho que x.
- correct = lógico, correção de Yates. A correção não é feita se simulatE.p. Value=TRUE,
- p = vetor de probabilidades do mesmo comprimento que x,
- rescale.p= lógico escalar, se TRUE p é reescalado para somar 1. Se rescale. p=FALSE e p não soma 1, um erro é dado,

Exemplo para probabilidades distintas:

Mendel obteve os seguintes números de sementes: 315 lisas e amarelas, 101 rugosas e amarelas, 108 lisas e verdes e 32 rugosas e verdes. A freqüência esperada é 9/16, 3/16, 3/16, 1/16, ou 0,5625; 0,1875; 0,1875; 0,0625.

```
>ervilha<-c(lisam=315,rugam=101,lisverd=108, rugverd=32)
>prop<-c(0.5625, 0.1875, 0.1875, 0.0625)
>chiSQ.test(x=ervilha, p=prop)
```

```
        Chi-squared test for given probabilities
data:  ervilha
X-squared = 0.47002, df = 3, p-value = 0.9254
```

O resultado mostra que a hipótese nula (H_0) não foi rejeitada, a FO não é significativamente diferente da FE.

Exemplo para probabilidades iguais:

Um bivalve apresenta 3 variedades de coloração: amarela, verde e cinza, que ocorrem em proporções iguais. Analisando o conteúdo estomacal de vários indivíduos que estavam forrageando sobre o banco de moluscos, encontramos 56 conchas amarelas, 32 verdes e 45 cinzas. Qual a sua conclusão?

>conchas<-c(56,32,44)

>chiSQ.test(conchas)

```
Chi-squared test for given probabilities

data:  conchas
X-squared = 6.5455, df = 2, p-value = 0.0379
```

Neste exemplo, rejeitamosH_0 , a fo é significativamente diferente da FE.

#Verifique a frequência esperada

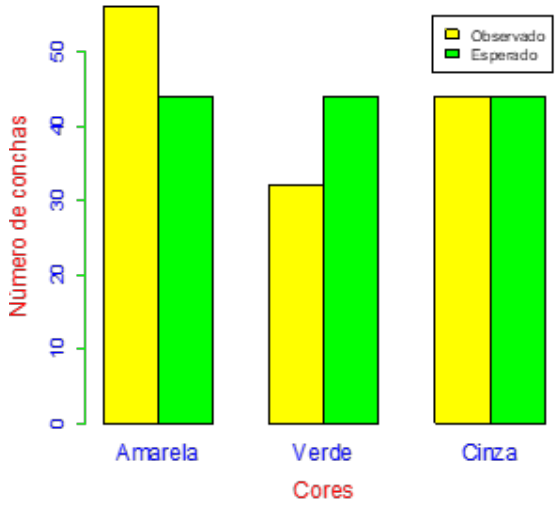
>chiSQ.test(conchas) $expected

```
[1] 44 44 44
```

Apresentando o resultado em um gráfico de barras

>matriz<-matrix(c(56,44,32,44,44,44),nrow=2,ncol=3, dimnames=list(c("observado","esperado"),c("Amarela", "Verde", "Cinza")))

>barplot(matriz, beside=TRUE, legend=TRUE, xlab="Cores", ylab= "Número de conchas", col=c("yellow","green"), cex=1.4,Cex.lab=1.4,Cex. axis=1.2, Col.axis='blue', col.lab='red',fg='green')



TESTE DE NORMALIDADE

Insira os dados que você quer analisar:

>z<-c(21, 18.3, 44, 43.3, 25, 26, 29.4, 30,31.3, 33, 23, 29)

Teste do Shapiro – Wilk

>shapiro.test(z)

```
Shapiro-Wilk normality test

data:  z
W = 0.92256, p-value = 0.3078
```

A amostra 'z' não apresenta distribuição significativamente diferente da normal pelo teste Shapiro

Kolmogorov-Smirnov

>ks.test(z,"pnorm",mean(z),sd(z))

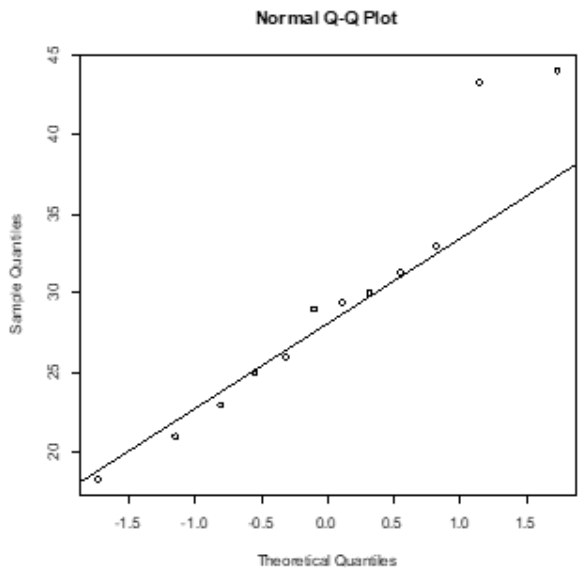
```
One-sample Kolmogorov-Smirnov test

data:  z
D = 0.15984, p-value = 0.8723
alternative hypothesis: two-sided
```

A amostra 'z' não apresenta distribuição significativamente diferente da normal pelo teste Kolmorov-Smirnov.

Utilizando a distribuição dos quartis

>qqnorm(z)

>qqline(z)



TESTE Z

Esta função não existe no pacote básico do R.

Mas, você pode utilizar a função do pacote 'TeachingDemos'.

>library(TeachingDemos)

>x<-c(59,56, 46,61,57,65,49,60,66,59,69,56,64,60,59,66,58,59,51,53)

>z.test(x, mu=65, stdev=10)

```
        One Sample z-test
z = -2.8398, n = 20.0000, Std. Dev. = 10.0000, Std. Dev. of the sample
mean = 2.2361, p-value = 0.004514
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 54.26739 63.03261
sample estimates:
mean of x   58.65
```

O resultado indica que a média amostral difere significativamente da média populacional

TESTE T

O comando para realizar o teste t no R é: 't.test', sendo utilizado para uma ou duas amostras.

>t.test(amostra1,amostra2,argumentos)

Argumentos

- mu=valor verdadeiro da média, ou média populacional.
- paired= argumento lógico indicando se você quer um teste pareado: TRUE ou FALSE.
- var.equal= argumento lógico indicando se as duas variâncias são iguais: TRUE ou FALSE.
- conf.Level= intervalo de confiança
- alternative= indica se o teste é unilateral ou não: "two-sided", "less" ou "greater".

Teste t para uma amostra

Ex:

Teste t para uma amostra, utilizando o exemplo das tilápias (capítulo 4).

>tilapias<-c(23,43,22,23,40,39,26,37,42,26,39,37,30,44,39)

>resultado<-t.test(tilapias, alternative="two.sided", mu=38, conf. Int=0.95)

>resultado

```
        One Sample t-test

data:  tilapia
t = -1.9279, df = 14, p-value = 0.0744
alternative hypothesis: true mean is not equal to 38
95 percent confidence interval:
 29.55001 38.44999
sample estimates:
mean of x
       34
```

O resultado indica que a média amostral não difere significativamente da média populacional (p=0,074).

#Plotando os pontos em relação à m**édia:**

>plot(1:15,tilapias, pch=15,col=2)

>Media=mean(tilapias)

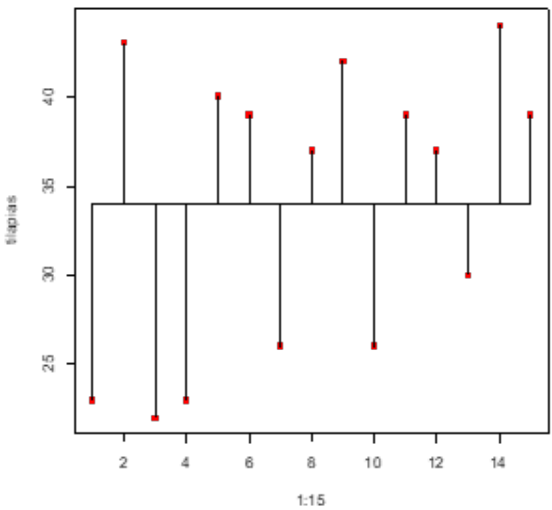
For(i in 1:15)

{

Lines(c(i,i),c(tilapias[i],media),col=1)

}

Lines(c(1,15),c(media,media),col=1)



TESTES PARA DUAS AMOSTRAS

Teste Qui-quadrado para independência entre duas variáveis, utilizando tabelas de contigência.

	Pequena	Grande	Total
Varejista	31	44	**75**
Atacadista	60	25	**85**
Avulso	12	8	**20**
Total	**103**	**77**	**180**

Alcaparra<-read.delim("clipboard", row.names=1)

> alcaparra

```
           Pequena Grande
Varejista       31     44
Atacadista      60     25
Avulso          12      8
```

Para ver os valores esperados:

>chiSQ.test(alcaparra) $expected

```
            Pequena    Grande
Varejista  42.91667 32.083333
Atacadista 48.63889 36.361111
Avulso     11.44444  8.555556
```

>chiSQ.test(alcaparra)

```
Pearson's Chi-squared test
data:  alcaparra
X-squared = 14.002, df = 2, p-value = 0.0009111
```

Rejeito H_0, há relação entre o tamanho de alcaparra e o tipo de comprador (eles compram o produto em proporções distintas).

Teste t para duas amostras independentes

>t.test(amostra1,amostra2,argumentos)

>p <-c(12.21, 14.3, 9.44, 16.32, 15.2, 11.22, 12.24, 10.3, 13.7, 14.1, 11.1, 11.9, 10.5, 25)

>q <-c(21.22, 24.55, 27.33, 20.56, 26.33, 18.44, 18.33, 21.22, 23.33, 24.33)

Verificando a normalidade dos dados:

>shapiro.test(P)

```
Shapiro-Wilk normality test
data:  P
W = 0.97055, p-value = 0.8839
```

>shapiro.test(Q)

```
      Shapiro-Wilk normality test
data:  Q
W = 0.94661, p-value = 0.6286
```

Testando a homogeneidade de variâncias:

>var.test(P,Q)

```
     F test to compare two variances
data:  P and Q
F = 0.40879, num df = 13, denom df = 9, p-value = 0.1392
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1067166 1.3539200
sample estimates:
ratio of variances
         0.4087883
```

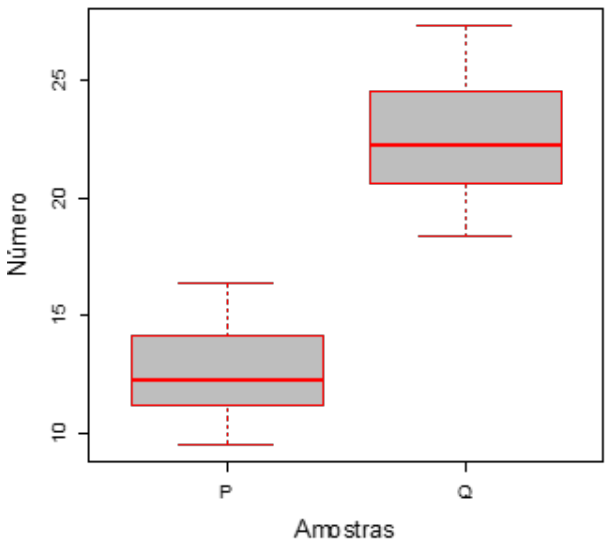
As variâncias não são significativamete diferentes; faremos o teste t para variâncias homogêneas.

>t.test(P,Q, var.equal=TRUE,alternative="two.sided")

```
   Two Sample t-test
data:  P and Q
t = -9.6839, df = 22, p-value = 2.159e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.216670  -7.907044
sample estimates:
mean of x mean of y
 12.50214  22.56400
```

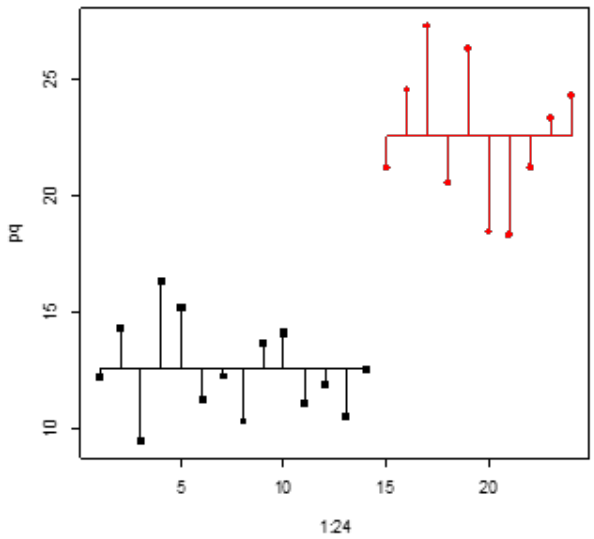
Fazendo o gráfico para as duas amostras

```
>boxplot(P,Q, names=c("P","Q"), border="red",
col="grey",ylab="Número",xlab="Amostras", cex.lab=1.4,cex.axis=1.2)
```



Plotando as amostras **e** posicionando os pontos em relação à média:

```
>mediap=mean(P)
>mediaq=mean(Q)
>pq<-c(P,Q)
>
T(1:24,pq,pch=rep(c(15,16),each=c(14,10)),col=rep(1:2,each=c(14,10)))
For(i in 1:14)
{
Lines(c(i,i),c(pq[i],mediap),col=1)
}
For(j in 15:24)
{
Lines(c(j,j),c(pq[j],mediaq),col=2)
}
Lines(c(1,14),c(mediap,mediap),col=1)
Lines(c(15,24),c(mediaq,mediaq),col=2)
```



Teste t para duas amostras pareadas

Exemplo: um pesquisador está avaliando se um antidepressivo afeta a pressão arterial diastólica dos pacientes:

>antes<-c(133,134,135,142,148,150,164,170,175,179,184,185,188)

>depois<-c(132,135,136,138,140,143,144,150,151,153,155,155,159)

>t.test(antes,depois,paired=TRUE,alternative="two.sided")

Paired t-test

```
data:  antes and depois
t = 4.4251, df = 12, p-value = 0.0008281
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.653442 22.500405
sample estimates:
mean of the differences
               15.07692
```

EXERCÍCIOS DO CAPÍTULO 18

1. Resolva o exercício 1 do capítulo 4, utilizando as funções do R.

2. Resolva o exercício 3 do capítulo 4, utilizando as funções do R.

3. Resolva o exercício 10 do capítulo 4, utilizando as funções do R.

4. Resolva o exercício 1 do capítulo 5.

Capítulo 19

ANOVA

A função de análise de variância no R é 'aov'.

Aov(formula, data = null, projections = FALSE, qr = TRUE, ...)

Argumentos:

- Formula= especifica o modelo a ser utilizado. Existem vários tipos de ANOVA, que são expressos na forma de modelos:

Modelo	Descrição
Y~x1	Fator único, y é explicado por x1
Y~x1+x2	Dois fatores, y é explicado por x1 e x2
Y~x1×x2	Dois fatores, y é explicado por x1 e x2 e pela interação de ambos
Y~x1+x2+x3	Três fatores,

- data = data.frame no qual as variáveis especificadas no modelo vão ser encontradas.
- projections = bandeira lógica (lógical flag), para dados não numéricos quando TRUE, o automático é FALSE.
- qr= bandeira lógica também.

Exemplo: vamos utilizar o exemplo da ANOVA, com as quatro variedades de feijões do capítulo 6.

Antes da ANOVA, pode-se analisar a homogeneidade das variâncias:

1. Inserindo os dados:

```
>feijao<-c(22,18,21,23,29,25,27,27,19,25,22,35,30,29,31)
>tratamentos<-c(rep(1,4),rep(2,3),rep(3,4),rep(4,4))
```

2. definir os fatores:

>tratamentos<-factor(tratamentos)

3. Calcule a variância:

>variancia<-tapply(contagem, tratamento, var)

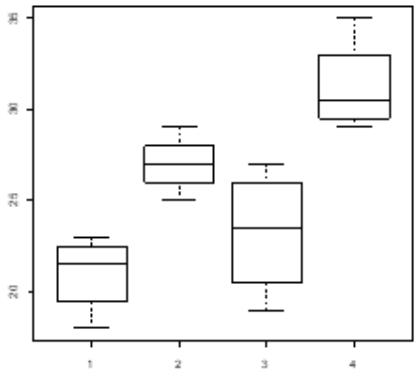
>variancia<-tapply(feijao, tratamentos, var)

> variancia

```
1         2         3         4
 4.666667  4.000000 12.250000  6.916667
```

Ou faça o gráfico:

>boxplot(feijao~tratamentos)



4. Faça o teste de homogeneidade da variância

<Bartlett.test(modelo, data.frame)>

>testeaov<-data.frame(feijao=c(22,18,21,23,29,25,27,27,19,25,22,35, 30,29,31),Tratamentos=c(rep(1,4),rep(2,3),rep(3,4),rep(4,4)))

>bartlett.test(feijao~tratamentos, testeaov)

```
Bartlett test of homogeneity of variances

data:  Contagem by tratamento
Bartlett's K-squared = 0.90282, df = 3, p-value = 0.8247
```

5. Teste a normalidade das amostras

> with(testeaov, tapply(feijao, tratamentos, Shapiro.test))

```
$`1`
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.92708, p-value = 0.5774
$`2`
        Shapiro-Wilk normality test
data:  X[[i]]
W = 1, p-value = 1
$`3`
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.97865, p-value = 0.8941
$`4`
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.88691, p-value = 0.369
```

Todas as amostras apresentaram distribuição normal

6. Rode a ANOVA:

>resultado.aov=aov(feijao~tratamentos)

>resultado.aov

```
Call:
   aov(formula = feijao ~ tratamentos)

Terms:
                tratamentos Residuals
Sum of Squares     240.2333   79.5000
Deg. of Freedom           3        11

Residual standard error: 2.688359
Estimated effects may be unbalanced
```

O sumário é mais informativo:

>summary(resultado.aov)

```
      Df Sum Sq Mean Sq F value  Pr(>F)
tratamento   3  240.2   80.08   11.08 0.00119 **
Residuals   11   79.5    7.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7. Teste aposteriori:

Tukey-honest significant difference

TukeyHSD(x, which, ordered = FALSE, conf.level = 0.95, …)

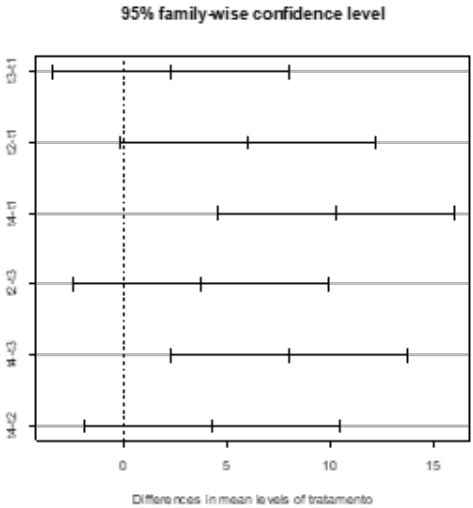
- X = modelo ajustado, geralmente `aov`
- Which = vetor listando termos no modelo ajustado, para os quais os intervalos devem ser calculados.
- Ordered = lógico, indicando se os níveis de fator deve ser ordenados com a média. Se ordenados, as diferenças serão sempre positivas.
- Conf.Level = nível de confiança.

> TukeyHSD(resultado.aov)

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Contagem ~ tratamento)

$tratamento
       diff       lwr       upr     p adj
t2-t1  6.00 -0.179408 12.179408 0.0577999
t3-t1  2.25 -3.471020  7.971020 0.6488039
t4-t1 10.25  4.528980 15.971020 0.0010675
t3-t2 -3.75 -9.929408  2.429408 0.3123804
t4-t2  4.25 -1.929408 10.429408 0.2219276
t4-t3  8.00  2.278980 13.721020 0.0068127
```

Podemos plotar o resultado acima:

>plot(TukeyHSD(resultado.aov,ordered=TRUE))



Diferenças significativas do teste de Tukey HSD. Os pares de amostras com diferenças significativas são: t1-t4 e t3-t4.

Análise de variância segundo dois critérios ou análise de blocos aleatorizados (ANOVA two way) sem interação (blocos casualizados), sem repetição

Y~x1+x2	Dois fatores, y é explicado por x1 e x2

Exemplo: o crescimento de quatro variedades de feijão são testados em 5 tipos de solo.

1. Inserir o arquivo:

>densidade<-c(3,2,0,2,1,2,6,6,1,0,5,1,3,7,4,5,8,9,7,5)

>dados=data.
frame(trat=factor(rep(1:4,each=5)),bloco=factor(rep(1:5,4)),densidade)

>attach(dados)

2. fazer o teste de homogeneidade das variâncias

>bartlett.test(densidade,trat)

```
      Bartlett test of homogeneity of variances

data:  densidade and trat
Bartlett's K-squared = 2.8591, df = 3, p-value = 0.4139
```

> bartlett.test(densidade,bloco)

```
        Bartlett test of homogeneity of variances

data:  densidade and bloco
Bartlett's K-squared = 2.4143, df = 4, p-value = 0.66
```

3. Fazer a ANOVA

>resultado=aov(densidade~trat+bloco)

> summary(resultado)

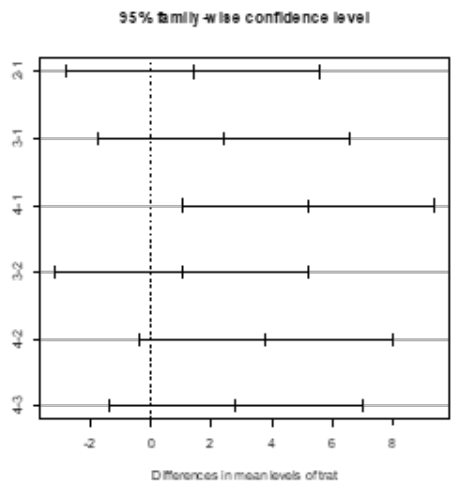
```
            Df Sum Sq Mean Sq F value Pr(>F)
trat         3  72.55  24.183   4.861 0.0194 *
bloco        4  10.30   2.575   0.518 0.7246
Residuals   12  59.70   4.975
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Teste Tukey

>TukeyHSD(resultado,ordered=TRUE)

```
Fit: aov(formula = densidade ~ trat + bloco)
$trat
    diff        lwr     upr     p adj
2-1  1.4 -2.7881504 5.58815 0.7564438
3-1  2.4 -1.7881504 6.58815 0.3646350
...
```

>plot(TukeyHSD(resultado,which=trat,ordered=TRUE))



ANOVA DOIS FATORES

Y~x1 × x2	Dois fatores, y é explicado por x1 e x2 e pela interação de ambos

Exemplo, testar o crescimento de uma especie de peixe a tres temperaturas e tres rações.

1. Digite o arquivo

Remento<-c(7,7,7.1,7.2,9,5.2,7.8,8.2,7.2,7,6.3,6,9.5,7.2,7.2,6.6,7,7.3,8.9, 8.8,7.3,7.8,9.2,7.1,11.5,10.1,11.4,7.2,8.5,8.7,10.9, 10.8,10.4,10.4,11.6,8.2)

>anovabi<-data.frame(incremento, temperatura= factor(c(rep("a",12), Rep("b",12), rep("c",12))) , racao= factor(rep(c(rep("i",4),rep("ii",4),

Rep("iii",4)),3)))

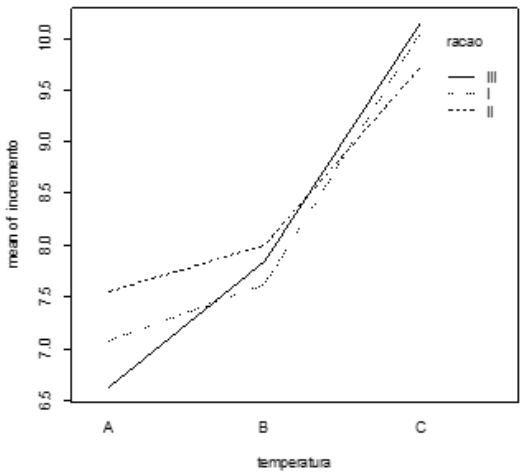
2. Possibilite que o R acesse as variáveis da planilha
   >attach(ANOVAbi)
3. Definimos e verificamos os fatores do teste

> temperatura<-factor(temperatura), racao<-factor(racao)
>is.factor(temperatura), is.factor(racao)

```
[1] TRUE
[1] TRUE
```

4. Fazemos o gráfico com as médias dos tratamentos:

>interaction.plot (temperatura,racao,incremento)



5. Rodamos a ANOVA
   >resultadobi=aov(incremento~temperatura*racao)
   > summary(resultadobi)

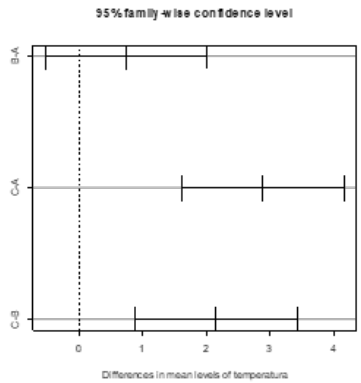
```
                 Df Sum Sq Mean Sq F value   Pr(>F)
temperatura       2  54.14  27.069  17.064 1.62e-05 ***
racao             2   0.32   0.159   0.100    0.905
temperatura:racao 4   2.07   0.519   0.327    0.857
Residuals        27  42.83   1.586
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Fazemos o teste a posteriori

>TukeyHSD(resultadobi,ordered=t)

```
Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = incremento ~ temperatura * racao)
$temperatura
         diff        lwr      upr     p adj
B-A 0.7416667 -0.5332043 2.016538 0.3340485
C-A 2.8916667  1.6167957 4.166538 0.0000167
C-B 2.1500000  0.8751290 3.424871 0.0007753
$racao
             diff       lwr      upr     p adj
I-III  0.04166667 -1.233204 1.316538 0.9963867
II-III 0.21666667 -1.058204 1.491538 0.9070994
II-I   0.17500000 -1.099871 1.449871 0.9382898
$`temperatura:racao`
             diff        lwr      upr     p adj
A:I-A:III   0.450 -2.5465611 3.446561 0.9998486
A:II-A:III  0.925 -2.0715611 3.921561 0.9781290
B:I-A:III   1.000 -1.9965611 3.996561 0.9653424
B:III-A:III 1.225 -1.7715611 4.221561 0.8973573
B:II-A:III  1.375 -1.6215611 4.371561 0.8250601
C:II-A:III  3.100  0.1034389 6.096561 0.0384636
C:I-A:III   3.425  0.4284389 6.421561 0.0163013
C:III-A:III 3.525  0.5284389 6.521561 0.0124092
A:II-A:I    0.475 -2.5215611 3.471561 0.9997735
B:I-A:I     0.550 -2.4465611 3.546561 0.9993362
...
```

7. Plotear o gráfico com as diferenças significativas de temperatura:
>plot(TukeyHSD(resultadobi, “temperatura”,ordered=T))



Teste TukeyHSD, diferenças entre os pares. Nota-se que ‘C’ é significativamente diferente dos outros dois.

Capítulo 20

CORRELAÇÃO E REGRESSÃO

Antes de realizar as rotinas de correlação e regressão dos dados, não se esqueça de plotar os gráficos de dispersão dos mesmos.

A função de correlação é cor(,)

Cor(x, y = NULL, method = c("Pearson", "kendall", "spearman"))

Argumentos:

- X= vetor, matriz ou data.frame
- Y= nulo ou vetor.
- method= "Pearson"(automático), "kendall" ou "spearman".

Exemplo:

O CO_2 apresenta alguma tendência ao longo dos anos:

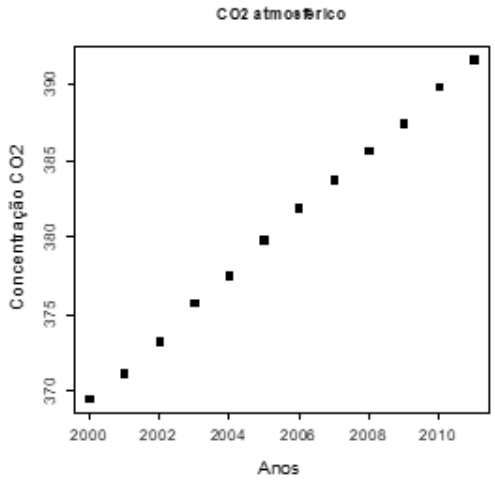
Insira os vetores no R

>Ano<- c(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011)

>CO2<- c(369.52, 371.13, 373.22, 375.77, 377.49, 379.8, 381.9, 383.77, 385.59, 387.38, 389.78, 391.57)

Faça o gráfico de dispersão:

plot(ano, CO2, xlab="anos", ylab="concentração CO2",main="CO2 atmosférico", cex=1.4,Cex.lab=1.4,Cex.axis=1.2, Pch=15)



>cor(ano, CO2)

```
[1] 0.9995346
```

Para testar a significância do mesmo, com o teste 't', utilizamos a função 'cor.test(,)'

>cor.test(ano, CO2)

```
  Pearson's product-moment correlation
data:  Ano and CO2
t = 103.6162, df = 10, p-value = 2.22e-16
alternative hypothesis: true correlation is not equal to 0
```

Para analisar mais dados, como em uma correlação múltipla, fica mais fácil trabalhar com uma tabela:

Exemplo:
Insira a tabela abaixo no R:

Biomassa	NO3	PO4	Fe2	Vitaminas	Ca	K	Mn
100	233	20	98	0.02	337	2	53
110	239	20	63	0.072	425	5	54
111	253	21	459	0.095	481	6	63

Biomassa	NO3	PO4	Fe2	Vitaminas	Ca	K	Mn
113	256	21	28	0.102	422	6	63
122	256	21	468	0.151	327	8	67
125	261	23	274	0.163	487	9	68
127	276	23	384	0.164	700	9	68
154	279	23	454	0.166	680	9	92
156	281	21	181	0.219	620	11	92
164	314	21	236	0.231	470	11	94
165	314	22	194	0.279	679	13	99

>nutrientes=read.delim("clipboard", row.names=1)

Faça a correlação múltipla:

> resultado<-cor(nutrientes)

NO3 PO4 Fe2 Vitaminas Ca K Mn

No3 1.0000000 0.4217506 0.1502284 0.9361040 0.6297038 0.9111778 0.9215926

PO4 0.4217506 1.0000000 0.5252259 0.4907243 0.7083905 0.5508838 0.4100033

Fe2 0.1502284 0.5252259 1.0000000 0.2110148 0.2632526 0.2580665 0.1936955

Vitaminas 0.9361040 0.4907243 0.2110148 1.0000000 0.6396313 0.9936334 0.9091206

Ca 0.6297038 0.7083905 0.2632526 0.6396313 1.0000000 0.6783674 0.6624910

K 0.9111778 0.5508838 0.2580665 0.9936334 0.6783674 1.0000000 0.8906264

Mn 0.9215926 0.4100033 0.1936955 0.9091206 0.6624910 0.8906264 1.0000000

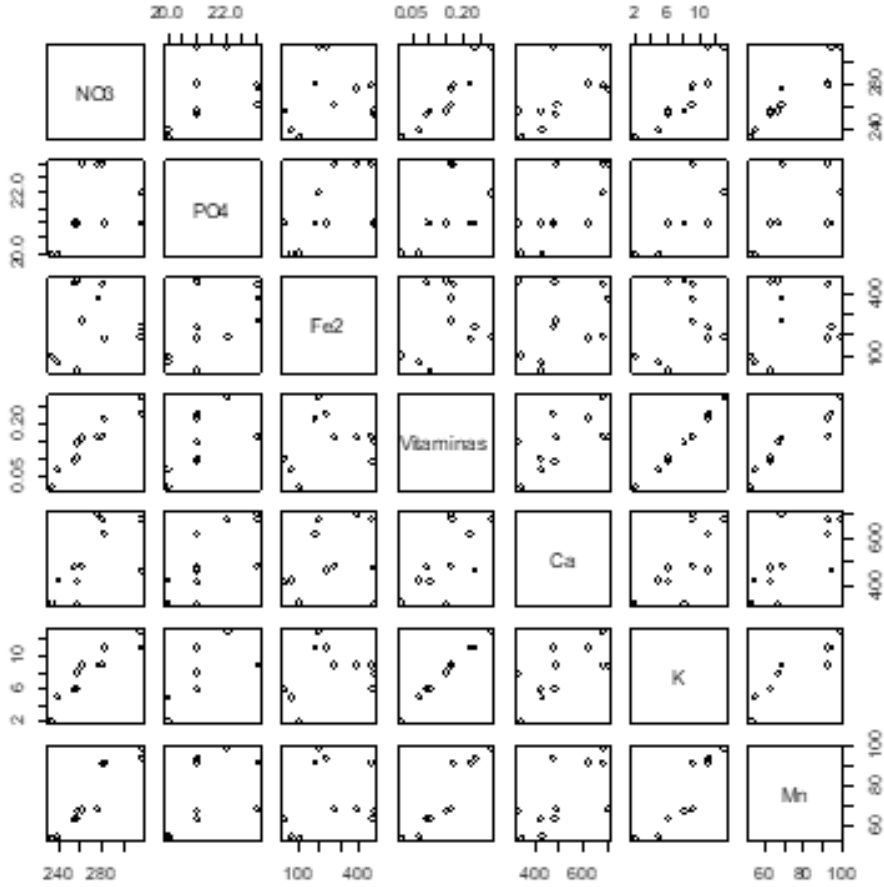
```
           biomassa        NO3       PO4        Fe2 vitaminas
Ca
biomassa  1.0000000 0.92995860 0.6790076 0.10328690 0.9872818
0.16484724
NO3       0.9299586 1.00000000 0.6027758 0.02257774 0.9246193
0.16062987
PO4       0.6790076 0.60277583 1.0000000 0.17532015 0.7302535
0.15791594
Fe2       0.1032869 0.02257774 0.1753201 1.00000000 0.0972189
0.01963637
vitaminas 0.9872818 0.92461932 0.7302535 0.09721890 1.0000000
0.15548993
Ca        0.1648472 0.16062987 0.1579159 0.01963637 0.1554899
1.00000000
K         0.9772282 0.90853287 0.7378889 0.12233796 0.9908513
0.16041821
Mn        0.9787287 0.92621377 0.6566787 0.08853480 0.9801932
0.10616714
                  K         Mn
biomassa  0.9772282 0.9787287
NO3       0.9085329 0.9262138
PO4       0.7378889 0.6566787
Fe2       0.1223380 0.0885348
vitaminas 0.9908513 0.9801932
Ca        0.1604182 0.1061671
K         1.0000000 0.9749746
Mn        0.9749746 1.0000000
```

Exporte o resultado:

```
>resultado<-data.frame(cor(nutrientes))
>attach(resultado)
>write.Csv(resultado, file = "corrmultipla.csv",row.names=TRUE)
```

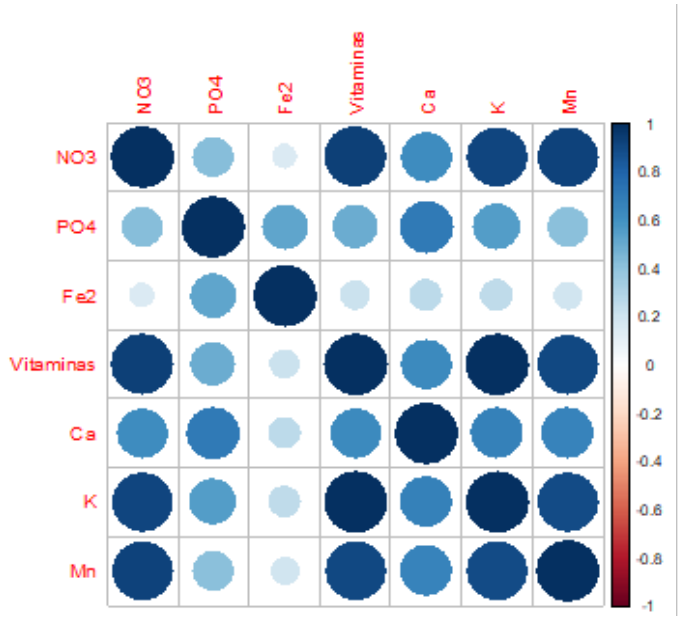
Para visualizar todos os gráficos de dispersão:

```
>pairs(nutrientes)
```



Outros pacotes do R elaboram gráficos de correlação múltipla com mais alternativas gráficas, como o 'corrplot', por exemplo.

```
>library(corrplot)
>matriz <- cor(nutrientes)
>corrplot(matriz, method = "circle")
```



#Para saber apenas a correlação entre duas variáveis da tabela:

>with(nutrientes,cor(NO3,PO4))

```
[1] 0.4217506
```

Os testes de correlação são realizados par a par.

>cor.test(nutrientes$po4, nutrientes$NO3)

```
     Pearson's product-moment correlation
data:  nutrientes$PO4 and nutrientes$NO3
t = 1.3954, df = 9, p-value = 0.1964
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2384523  0.8153452
sample estimates:
      cor
0.4217506
```

Mudando o intervalo de confiança

>with(nutrientes,cor.test(PO4,NO3, alternative=”greater”,conf. Level=.99))

```
    Pearson's product-moment correlation
data:  PO4 and NO3
t = 1.3954, df = 9, p-value = 0.09818
alternative hypothesis: true correlation is greater than 0
99 percent confidence interval:
 -0.3563238  1.0000000
sample estimates:
      cor
0.4217506
```

Correlação de Spearman

>with(nutrientes,cor.test(vitaminas, NO3, method=”spearman”, alternative= “two.sided”))

```
      Spearman's rank correlation rho
data:  Vitaminas and NO3
S = 1.0023, p-value = 1.691e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9954442
```

Correlação de Kendall

Basta mudar para method=”kendall”

>with(nutrientes,cor.test(vitaminas, NO3, method=”kendall”))

```
     Kendall's rank correlation tau
data:  Vitaminas and NO3
z = 4.1513, p-value = 3.306e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.9816498
```

REGRESSÃO

As regressões linear e múltipla são executadas pela função ‘lm()’. Esta função é a interface de uma equação. Para a regressão simples a equação tem a forma ‘DV~IV’, onde DV é a variável dependente e IV a independente.

A sintaxe básica para a análise de regressão em R é:

Lm(y ~ model)

Abaixo estão exemplos de sintaxe e respectivos modelos de regressão:

Sintaxe	Modelo	Comentários
Y ~ A	$Y= ß_0 + ß_1 A$	Regressão linear com intercepto do Y
Y ~ -1 + A	$Y= ß_1 A$	R. Linear sem intercepto Y, a reta passa pela origem (0,0)
Y ~ A + I(A^2)	$Y= ß_0 + ß_1 A+ ß_2 A^2$	Modelo polinomial. A função identidade I() permite termos no modelo incluindo símbolos matemáticos
Y ~ A + B	$Y= ß_0 + ß_1 A+ ß_2 B$	Modelo de 1ª ordem com duas variáveis independentes (A, B)
Y ~ A:B	$Y= ß_0 + ß_1 AB$	Modelo com interações entre A e B
...		

Exemplo de regressão

A concentração de gás CO_2 vem aumentando linearmente ao longo dos anos. Qual **seria o** modelo para estes dados?

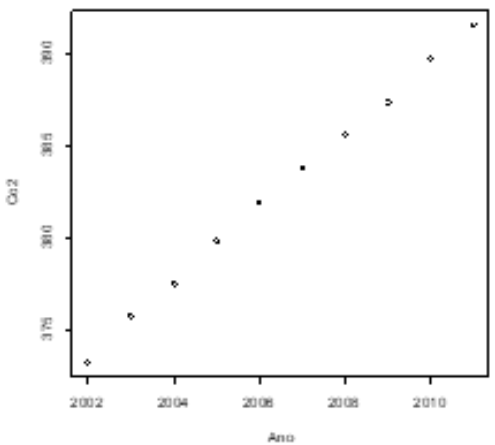
1. Inserir os dados no R.

>ano<- c(2002, 2003,2004,2005, 2006, 2007, 2008, 2009, 2010, 2011)

>CO2<- c(373.22, 375.77, 377.49, 379.8, 381.9, 383.77, 385.59, 387.38, 389.78, 391.57)

2. Plotamos os dados no gráfico de dispersão

>plot(Ano,CO2) → a variável independente vai primeiro



Os dados apresentam uma tendência linear

3. Calculando os parâmetros do modelo linear

> lm(CO2 ~ ano)

#Para calcular os parâmetros a variável dependente vai primeiro

```
Coefficients:
(Intercept)          Ano
  -3653.600        2.012
```

4. Informações úteis

Para obter mais informações úteis, e ter acesso a mais funções para manipular os dados, o melhor é criar um objeto que contém os comandos do modelo:

>lm.r = lm(CO2 ~ Ano)

Este objeto pode ser usado como argumento para outros comandos. Para obter um sumário estatístico do modelo, podemos utilizar o comando summary()

>summary(lm.r)

```
Call:
lm(formula = Co2 ~ Ano)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35491 -0.10883  0.02906  0.17194  0.27879

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.654e+03  4.987e+01  -73.26 1.34e-12 ***
Ano          2.012e+00  2.486e-02   80.93 6.06e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2258 on 8 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9986
F-statistic:  6550 on 1 and 8 DF,  p-value: 6.06e-13
```

Outros commandos úteis :

>coef(lm.r)

```
(Intercept)          Ano
-3653.599758     2.011576
```

>resid(lm.r) # dá os erros residuais de Y.

```
       1           2           3           4           5           6
-0.35490909  0.18351515 -0.10806061  0.19036364  0.27878788
0.13721212
          7           8           9          10
-0.05436364 -0.27593939  0.11248485 -0.10909091
```

>fitted(lm.r) → dá os valores preditivos de Y.

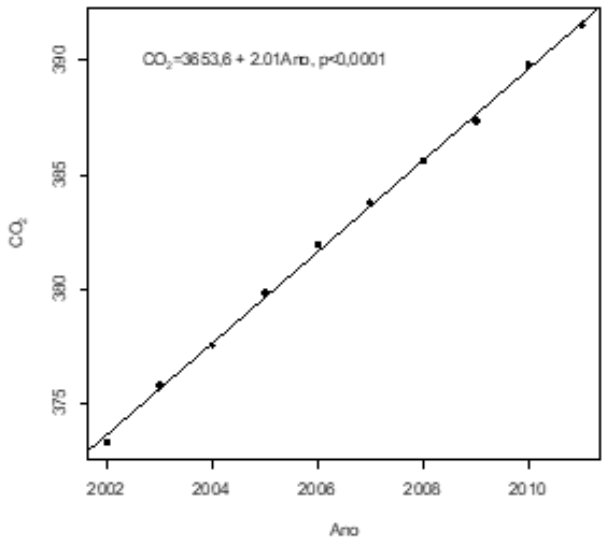
```
      1        2        3        4        5        6        7        8
373.5749 375.5865 377.5981 379.6096 381.6212 383.6328 385.6444
387.6559
       9       10
389.6675 391.6791
```

5. Fazendo o gráfico com a linha de regressão

>plot(ano,CO2, pch=16,ylab=expression("CO"[2]))

>abline(lm.r)

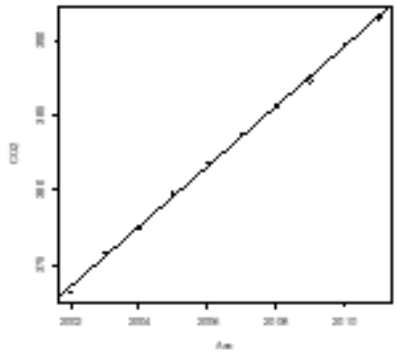
>text(2005,390,expression('CO'[2]*'=3653,6 + 2.01Ano, p<0,0001'))



O comando 'abline()' pode ser utilizado, utilizando o intercepto e inclinação: >abline(intercepto, inclinaçao)

>plot(ano,CO2)

>abline(-3653.599758, 2.011576)



O intervalo de confiança do gráfico é inserido através da função predict(lm)

Para isto, precisamos criar uma nova variável, que esteja dentro de um data.frame, e que possua o nome da variável independente:

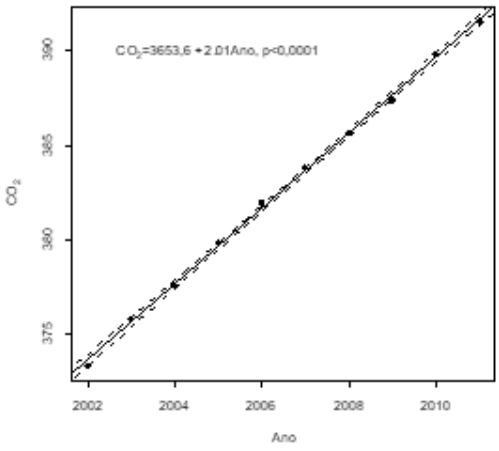
>new.x<-data.frame(Ano=seq(2001,2012, by=0.1))

Agora vamos predizer os valores da variável dependente, assim como os limites superior e inferior do intervalo de confiança (o default é de 95%).

>pred.Co<-predict(lm.r,new.X, interval="confidence")

>lines(new.X$ano,pred.Co[,2], lty=2, lwd=2)

>lines(new.X$ano,pred.Co[,3], lty=2, lwd=2)



6. Fazendo a Anova da regressão:

>anova(lm.r)

```
Analysis of Variance Table
Response: Co2
          Df Sum Sq Mean Sq F value    Pr(>F)
Ano        1 333.83  333.83  6549.6 6.06e-13 ***
Residuals  8   0.41    0.05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Avaliando os resultados da regressão linear através de gráficos diagnósticos

Podemos examinar os resíduos visualmente através de gráficos, utilizando o comando plot()

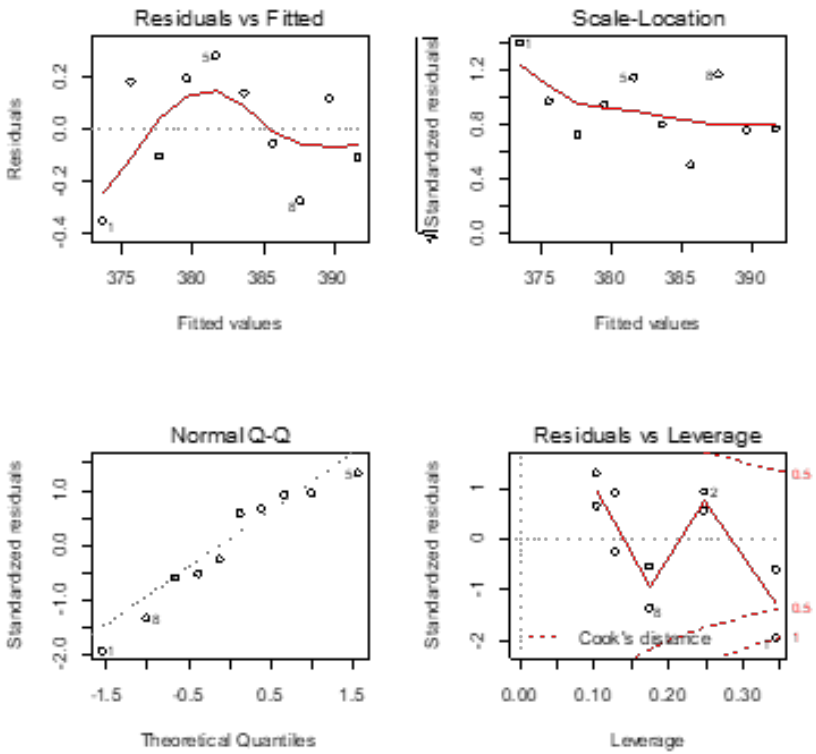
> layout(matrix(1:4,2,2))

>plot(lm.r)

Ou

>par(mfrow=c(2,2), pch=15)

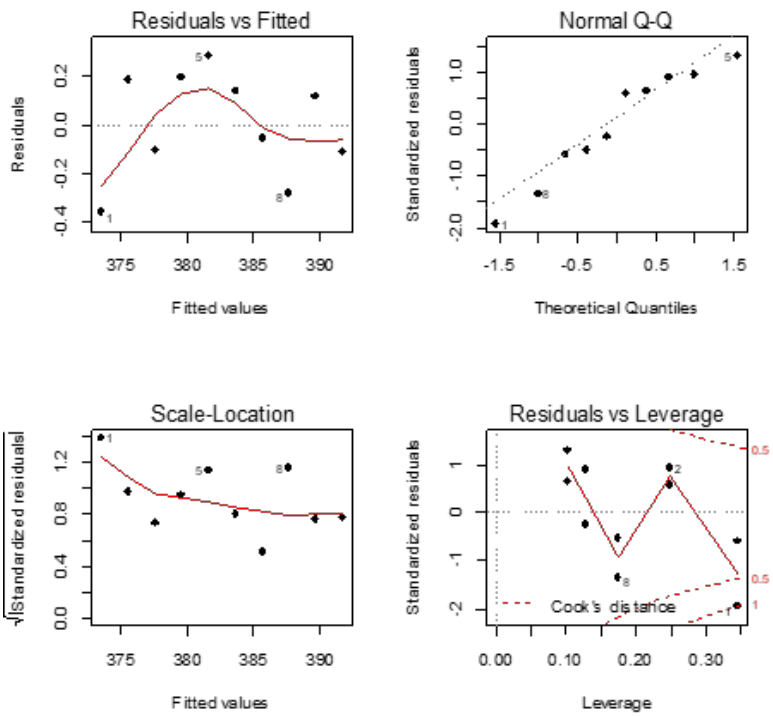
> plot(lm.r)



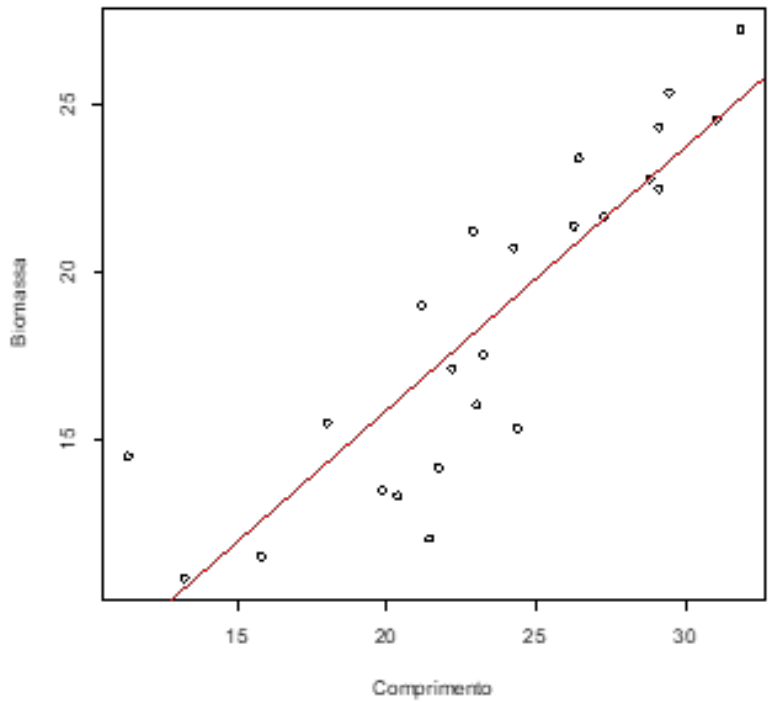
O primeiro é o gráfico clássico de resíduos, mostrando os resíduos e os valores ajustados. Os pontos que tendem a ser outliers (marginais) estão identificados. O segundo gráfico plota a raiz quadrada dos resíduos padronizados e os valores ajustados, a distribuição dos dados aqui, não deve apresentar nenhuma tendência obvia. O terceiro gráfico plota os quantis e os resíduos, a tendência linear indica normalidade dos resíduos. O último gráfico mostra os resíduos vs influencia. Os pontos com rótulo representam casos que devemos investigar para uma influencia indevida na regressão

>dado[2,]
Comprimento biomassa
2 20.32 13.28
>dado$fitted[2]

Outro comando para obter o mesmo resultado:
>par(mfrow=c(2,2),pch=16)
Plot(lm.r)



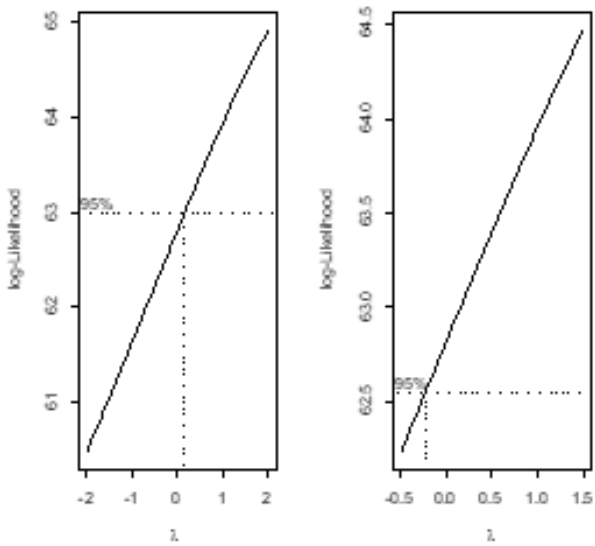
```
plot(dado)
abline(lm(dado$biomassa ~dado$Comprimento), col=”red”)
```



Transformação dos dados

Quando a distribuição dos erros não é normal, ou a variância não é homogênea, podemos transformar os dados da variável dependente ou ambos. O R tem a função boxcox(lm), que calcula o valor de λ pelo método da verossimilhança máxima. Λ define o tipo de transformação mais apropriada para a variável dependente do modelo linear. Quando λ=1, nenhuma transformação é necessária, quando λ=0, a transformação logaritmica é a mais adequada e quando λ=0,5 a raiz quadrada é a mais indicada. Esta função está no pacote [mass]

```
>require (MASS)
>par(mfrow=c(1,2))
>boxcox(lm.r)
>boxcox(lm.r, lambda=seq(-0.5,1.5, By=0.1))
```



A máxima verosimilhança nem aparece no gráfico e é maior que 1, neste caso não há necessidade de transformações.

Regressão linear múltipla

Y ~ a + b

Insira os dados no R:

>prodsoja<-c(1,1,1,1,1,2,2,2,3,3,3,3,4,5,5,5,5,5,5,6,6,6,6,8,8,8,9,10,11,11)

>pluviosidade<-c(1.5,1.57,1.69,1.72,1.78,1.8, 1.83,1.84,1.98,2.5,2.98,2.6, 3.25, 3.31, 3,3.2,2.9,2.22, 2.29,3, 2.41,2.42,2.48,2.69,2.72,2.83,3.39,3.55,3.61, 4.06)

>temperatura<-c(6.9,7.2,7.3,7.3,7.4,7.6,7.6, 7.7,7.9,8.6,8.8,9.1, 9.2, 9.3,9.3, 9.4, 9.9,10.1, 10.5,10.9,11.7, 12.1,7,7,7.2,8.7,8.8,9.1,9.8,10.1)

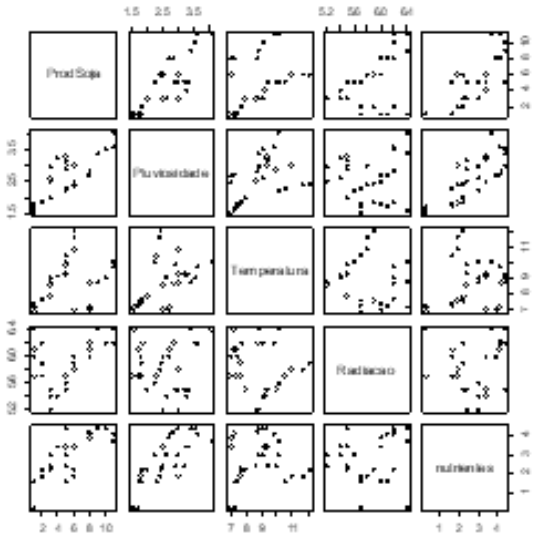
>radiacao<- c(5.7,6.4,5.7,6.1, 6,5.7,5.9,6.2,5.5,5.2,5.2, 5.4,5.4,5.5,5.5, 5.5, 5.5,5.6, 5.7, 5.8, 5.8, 5.9,6,6,6.1,6.2,6.4,6.2,6.2,6.4)

>nutrientes<-c(0.2,0.2,0.2,1.6,1.6,1.9,2.2,2.2,2.2, 2.4,3,3,3.4,3.4, 4.4,2.4,2.4,3, 1.6, 1.9, 1.9, 2.2, 3.4,3.9,4.1,4.2,4.4,4.4,4.1,3.7)

>soja<-data.frame(prodsoja, pluviosidade, temperatura, radiacao, nutrientes)

>attach(soja)

>pairs(soja)



>cor(soja)

```
             ProdSoja Pluviosidade Temperatura    Radiacao nutrientes
ProdSoja     1.0000000   0.82477865   0.4304726  0.41587372  0.7725730
Pluviosidade 0.8247786   1.00000000   0.4916020  0.05348929  0.7776800
Temperatura  0.4304726   0.49160200   1.0000000 -0.17149942  0.1860435
Radiacao     0.4158737   0.05348929  -0.1714994  1.00000000  0.1411290
nutrientes   0.7725730   0.77767996   0.1860435  0.14112896  1.0000000
```

>lm(soja$Prodsoja~soja$Pluviosidade+soja$Temperatura+soja$ Radiacao+soja$nutrientes)

```
Call:
lm(formula = soja$ProdSoja ~ soja$Pluviosidade + soja$Temperatura +
    soja$Radiacao + soja$nutrientes)

Coefficients:
      (Intercept)  soja$Pluviosidade   soja$Temperatura
soja$Radiacao
         -25.5614             1.8864             0.4692
3.2988
  soja$nutrientes
           0.8480
```

>lm.soja= lm(soja$Prodsoja~soja$Pluviosidade+soja$Temperatura+soja$Radiacao+ soja$nutrientes)

> summary(lm.soja)

```
Call:
lm(formula = soja$ProdSoja ~ soja$Pluviosidade + soja$Temperatura +
    soja$Radiacao + soja$nutrientes)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84099 -0.94995  0.08475  0.88699  2.10228

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -25.5614     4.1134  -6.214 1.69e-06 ***
soja$Pluviosidade   1.8864     0.6062   3.112  0.00461 **
soja$Temperatura    0.4692     0.1881   2.494  0.01958 *
soja$Radiacao       3.2988     0.6338   5.205 2.19e-05 ***
soja$nutrientes     0.8480     0.2984   2.842  0.00879 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.164 on 25 degrees of freedom
Multiple R-squared:  0.8733,    Adjusted R-squared:  0.8531
F-statistic:  43.1 on 4 and 25 DF,  p-value: 7.228e-11
```

Regressão múltipla stepwise

Utilizamos o mesmo exemplo acima:

>names (soja)

```
[1] "ProdSoja" "Pluviosidade" "Temperatura" "Radiacao"  "Nutrientes"
```

Criamos um modelo nulo, apenas com o intercepto:

>rm<-lm(ProdSoja~1,data=soja)

Depois iremos adionar os outros termos, com a função ‘add1()’.

>add1(rm, scope=soja)

```
Single term additions

Model:
ProdSoja ~ 1
             Df Sum of Sq    RSS    AIC
<none>                     267.47 67.634
Pluviosidade  1   181.947  85.52 35.427
Temperatura   1    49.563 217.90 63.486
Radiacao      1    46.259 221.21 63.937
Nutrientes    1   159.643 107.82 42.379
```

O menor valor de aic (akaike information criterion) será selecionado para o proximo passo. O aic é determinado neste caso, pela equação:

Aic = 2k + n[1n(RSS)], onde k é o número de termos do modelo de regressão e rss é a soma dos quadrados dos resíduos. Neste exemplo, pluviosidade é o proximo Fator A ser inlcuido no modelo, pois tem o menor AIC.

>rm= lm(prodsoja~Pluviosidade,data=soja)

> summary(rm)

```
Call:
lm(formula = ProdSoja ~ Pluviosidade, data = soja)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3615 -1.0468 -0.1071  1.4518  2.6975

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.5211     1.2575  -3.595  0.00123 **
Pluviosidade   3.6519     0.4732   7.718 2.08e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.748 on 28 degrees of freedom
Multiple R-squared:  0.6803,    Adjusted R-squared:  0.6688
F-statistic: 59.57 on 1 and 28 DF,  p-value: 2.083e-08
```

Verificamos que o modelo é significativo, e utilizamos novamente a função add1() para ver outros fatores potenciais

>add1(rm, scope=soja)

```
Single term additions

Model:
ProdSoja ~ Pluviosidade
            Df Sum of Sq    RSS    AIC
<none>                    85.520 35.427
Temperatura  1     0.221 85.299 37.349
Radiacao     1    37.071 48.449 20.379
Nutrientes   1    11.642 73.878 33.036
```

Vamos adicionar a radiação ao modelo:

>rm= lm(Prodsoja~Pluviosidade+Radiacao,data=soja)

> summary(rm)

```
Call:
lm(formula = ProdSoja ~ Pluviosidade + Radiacao, data = soja)

Residuals:
   Min     1Q Median     3Q    Max
-2.155 -1.059 -0.161  1.163  2.140

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -23.0243     4.1834  -5.504 7.90e-06 ***
Pluviosidade   3.5636     0.3632   9.812 2.13e-10 ***
Radiacao       3.2164     0.7076   4.545 0.000103 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.34 on 27 degrees of freedom
Multiple R-squared:  0.8189,    Adjusted R-squared:  0.8054
F-statistic: 61.03 on 2 and 27 DF,  p-value: 9.621e-11
```

Repetimos o processo, pois radiação também foi significativa.

>add1(rm, scope=soja)

```
Single term additions

Model:
ProdSoja ~ Pluviosidade + Radiacao
            Df Sum of Sq    RSS    AIC
<none>                    48.449 20.379
Temperatura  1    3.6273 44.822 20.045
Nutrientes   1    6.1408 42.308 18.314
```

Acrescentamos nutrientes ao modelo:

>rm= lm(prodsoja~pluviosidade+radiacao+nutrientes,data=soja)

> summary(rm)

```
Call:
lm(formula = ProdSoja ~ Pluviosidade + Radiacao + Nutrientes,
    data = soja)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8966 -1.0559  0.3155  0.9627  2.0907

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.2398     4.0883  -5.195 2.01e-05 ***
Pluviosidade   2.7288     0.5516   4.947 3.87e-05 ***
Radiacao       3.0062     0.6825   4.405 0.000162 ***
Nutrientes     0.5984     0.3080   1.943 0.062966 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.276 on 26 degrees of freedom
Multiple R-squared:  0.8418,    Adjusted R-squared:  0.8236
F-statistic: 46.12 on 3 and 26 DF,  p-value: 1.5e-10
```

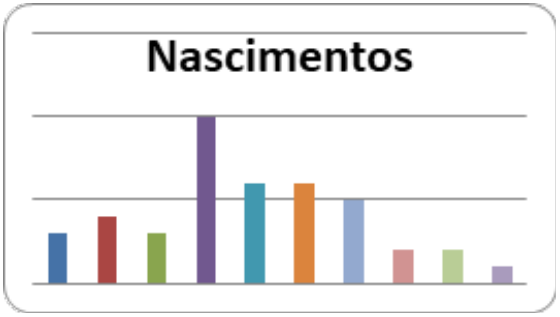
O fator nutrientes não foi significativo. Encerramos o processo.

RESPOSTAS EXERCÍCIOS

Capitulo 1: 1) Soma= 111,5, Média= 6,99, Mediana= 7,25, Moda= 9, Máximo= 9, Mínimo= 4, Variância= 3,08, Desvio Padrão=1,76, Coeficiente Variação= 25,13, Q_1= 5.75, Q_3=8,62. 2) Soma= 479 e 566, Média= 39,92 e 47,17, Mediana= 21 e 41, Moda= 7 e 38, Máximo= 99 e 96, Mínimo= 7 e 1, Variância= 1156,45 e 937,37, Desvio Padrão=34,01 e 30,63, Coeficiente Variação= 85,19 e 64,93, Q_1= 15,25 e 26,5, Q_3=64 e 66,25. 3) O comprimento apresenta o menor coeficiente de variação(49,21). Portanto, é a variável com menor variação. 4) O mais confiável é o paquímetro (c.v.=4,48). 5) $\sqrt[3]{1,53 \times 1,32 \times 1,14} = \sqrt[3]{2,302344} = 1,320454$, $13 \times 1,320454 = 17,16591$, $17,16591 \times 1,320454 = 22,6668$, $22,6668 \times 1,320454 = 29,93$(resultado final).

6)

Classes	Freqüência
1	3
2	4
3	3
4	10
5	6
6	6
7	5
8	2
9	2
10	1



Capítulo 2: 1) a) 30,567%, c) 0,03%, 2) dois filhotes= 8%, sete filhotes= 10%, 3) seis, MB,MP,MML,FB,FP,FML, 4) a) 5,26%, b) 73,68%, c) 11,76%, 5) P= 0,015625. 6) 19,2%. 7) 14%.

Capítulo 3: 1) Regra de decisão é o teste que possibilitará aceitarmos ou rejeitarmos a hipótese proposta. Isto é, serão estabelecidos os valores para se aceitar ou rejeitar H_0. Por exemplo num teste Z bilateral com α=0,05, Z_{crit}=1,96. Se o Z_{calc} for maior que o Z_{crit}, rejeitamos H_0. 2) O alfa representa a probabilidade do erro do tipo 1 (rejeitar H_0, quando ela é verdadeira). O objetivo é ter a maior confiança de que não acontecerá o erro 1. 3) A decisão depende do erro assumido, geralmente utilizamos α=0,05. Neste caso, não rejeito H_0, já que o valor calculado foi menor que o valor crítico do teste.

Capítulo 4:1) Z_{crit}= -1,65 > Z_{calc}= -2,59, rejeitamos H_0, a diferença é significativa. 2) Z_{crit}= 1,96 < Z_{calc}= |-3,65|, rejeitamos H_0. 3) t_{calc}=2,055 < t_{crit}=2,228, Não rejeito H_0, a floração pode ser considerada simultânea. 4) t_{calc}=3,795> t_{crit}=2,021, rejeito H_0. 5) t_{calc}=-1,91 < t_{crit}=2,08, não rejeito H_0. 6) teste Binomial, p=0,000656, rejeitamos H_0. 7) X^2_{calc}= 0,8 < X^2_{crit}=3,84, não rejeitamos H_0. 8) X^2_{calc}= 33,92 > X^2_{crit}=9,488, rejeitamos H_0. 9) X^2_{calc}= 10,04 > X^2_{crit}=7,815, rejeitamos H_0. 10) X^2_{calc}= 40,05 > X^2_{crit}=7,815, rejeitamos H_0. 11) X^2_{calc}= 68,42 > X^2_{crit}=5,991, rejeitamos H_0.

Capítulo 5: 1)F_{calc}= 2,29 < $F_{0,05,12,12}$=3,28 → Não rejeito Ho, as variâncias são homogêneas, t_{calc} = 3,47> t_{crit} =2,064, rejeito H_0. 2) F_{calc}= 3,4 < $F_{0,05,6,6}$=5,82 → Não rejeito Ho, as variâncias são homogêneas, t_{calc} = |-0,47|< t_{crit} =2,179, rejeito H_0. 3) t_{calc} = |-2,54| > t_{crit} =2,365, rejeito H_0. 4) t_{calc} = |-0,46| < t_{crit} =2,228, não rejeito H_0. 5) t_{calc} = |9,88| > t_{crit} =2,306, rejeito H_0.

Capítulo 6: 1)

	G.L.	S.Q.	Q.m.	F
Grupos	3	1129	376.3333	12.97701
ERRO	16	464	29	
TOTAL	19			

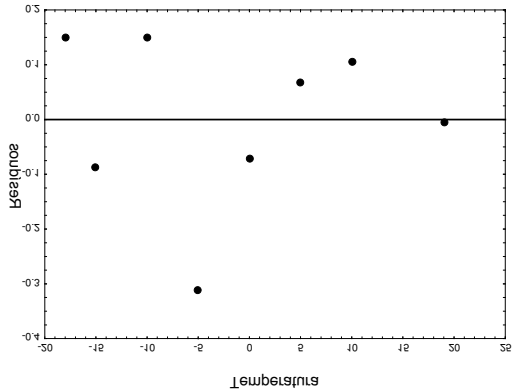
F_{calc}= 12,98> $f_{(0,05,3,16)}$=3,24, rejeito H_0. 2) F_{calc}= 6,16> $f_{(0,05,4,13)}$=3,18, rejeito H_0. 4) F_{calc}= 134,58> $f_{(0,05,4,45)}$=2,579, rejeito H_0. 5) Diferenças significativas: x1-x2, x1-x3, x2-x3, x3-x4.

Capítulo 7: 2) r= 0,576, t_{calc}= 1,726 < t_{crit}= 2,447, não rejeitamos H_0. 3) r= 0,934, t_{calc}= 7,85 > t_{crit}=2,26, rejeitamos H_0, r^2= 0,87 a abundância de árvores explica 87% do número de ninhos. 5) t_{calc}= -1,987 < t_{crit}=2,447, não rejeitamos H_0, 6) r_{calc}=|-0,603| < r_{tabela}= 0,648, não rejeito H_0. 7) r_{calc}= 0,936 > r_{tabela}= 0,648, rejeito H_0. 8) r_{calc}= 0,920 > r_{tabela}= 0,786, rejeito H_0. 9) Z_{calc}=0,824 < Z=1,96,

não rejeito H_0. 10) r= 0,657, t_{calc}= 2,753 > $t_{(0,05,\ 10)}$ = 2,228, rejeito H_0, há correlação positiva e significativa entre as duas variáveis.

Capítulo 8: 1) massa= -0,67+ 20,85 idade, t_{calc}=17,041> $t_{crít}$=2,306, rejeito H_0, r^2=0,973, 97% da variação da massa é explicada pela idade. 2) B_1= 0,163; B_0= 0,093; t_{calc}= 16,88 > $t_{(0,05;7)}$=2,36, Rejeito H_0. 3) Batimento=26,9*Esforço^{0,50}. 4) Não houve um aumento dos desvios com o aumento dos valores, não caracterizando heterocedasticidade. Não é possível dizer se o modelo é adequado.

n	1	2	3	4	5	6	7	8
ŷ	5.05	4.79	4.35	3.91	3.47	3.03	2.59	1.80



Capítulo 10: 3) a- =DIST.nORM.n(1.6,1.61,0.05,VERDADEIRO)- =DIST.nORM.n(1.55,1.61,0.05,VERDADEIRO)=0,305671. As mulheres, neste intervalo de altura, representam 30,6% da população feminina. c- = 1-DIST.nORM.n(1.78,1.61,0.05,VERDADEIRO)=0,000337. 4) a- =DIST.pOISSON(2,5,FALSO)=0,084 ou 8,4%. b- =DIST.pOISSON(7,5,FALSO)=10,4%

Capítulo 11: 1) =TESTE.QUIQUA(obs,esp)=0,001; rejeito H0, assaltam-se significativamente mais homens. 2) =TESTE.QUIQUA(obs, esp)=0,371,Não rejeito H0, as diferenças de nascimentos não são significativas . 3) =TESTE.QUIQUA(obs, esp)<0,0001. 4) =TESTE.QUIQUA(obs, esp)= 0,018,

Capítulo 12: 1) TESTE.f(matriz1,Matriz2)=0,165, as variâncias são homogêneas; =TESTE.T(matriz1,matriz2,2,2)=0,002, as duas amostras são

significativamente diferentes. 2) TESTE.T(matriz1,matriz2,2,1)=0,038, as duas amostras são significativamente diferentes

Capítulo 14: 1)correl(matriz1,matriz2)=0,934; r^2=0,87, 87% da variação do número de ninhos é explicada pela ocorrência de árvores mortas.

Capítulo 15: 8) >notas<-c(4.5,8,9, 5, 8.5, 6,5, 9,9,6, 8, 6.3, 9,4, 7, 7.5); > summary(notas); >sum(notas);>lenght(notas); >var(notas).

Capítulo 17: 1) pnorm(1.6, mean=1.61, sd=0.05)- pnorm(1.55, mean=1.61, sd=0.05)=0,30567; c- 1-pnorm(1.78, mean=1.61, sd=0.05)= 0.0003369293. 2) dpois(2,lambda=5)= 0,084; b- dpois(7,lambda=5)=0,1044. 3) dpois(3,lambda=5)= 0,1403739

Capítulo 18: 1) z.test(430, mu=450, stdev=50/SQrt(42)), z=-2,59; p= 0,01, a diferença é significativa, rejeito H_0. 2) x<-c(0, 0,1, 0, 0,2,3, 0,1, 0, 0); t.test(x, mu=0), t=2,055,p= 0,067. 3) fo<-c(35,25,30,30), prop<--c(0.2,0.15,0.125,0.525), chiSQ.test(x=fo,p=prop),chiSQ=40,05; p<0,0001. 4) var.test(v1, v2); p= 0,165; t.test(v1,v2, var.equal=TRUE,alternative="two. sided"); t=3,47; p= 0,002.

Tabela 1: Valores de z e áreas entre a média (0) e z, em valores absolutos.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706

1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Tabela 2: Valores críticos da distribuição t de student.

2α	*0,400*	*0,200*	*0,10*	*0,05*	*0,002*	*0,01*	*0,002*
g.l.\α	0,200	0,100	0,050	0,025	0,010	0,005	0,001
1	1,376	3,078	6,314	12,706	31,821	63,657	318,309
2	1,061	1,886	2,920	4,303	6,965	9,925	22,327
3	0,978	1,638	2,353	3,182	4,541	5,841	10,215
4	0,941	1,533	2,132	2,776	3,747	4,604	7,173
5	0,920	1,476	2,015	2,571	3,365	4,032	5,893
6	0,906	1,440	1,943	2,447	3,143	3,707	5,208
7	0,896	1,415	1,895	2,365	2,998	3,499	4,785
8	0,889	1,397	1,860	2,306	2,896	3,355	4,501
9	0,883	1,383	1,833	2,262	2,821	3,250	4,297
10	0,879	1,372	1,812	2,228	2,764	3,169	4,144
11	0,876	1,363	1,796	2,201	2,718	3,106	4,025
12	0,873	1,356	1,782	2,179	2,681	3,055	3,930
13	0,870	1,350	1,771	2,160	2,650	3,012	3,852
14	0,868	1,345	1,761	2,145	2,624	2,977	3,787
15	0,866	1,341	1,753	2,131	2,602	2,947	3,733
16	0,865	1,337	1,746	2,120	2,583	2,921	3,686
17	0,863	1,333	1,740	2,110	2,567	2,898	3,646
18	0,862	1,330	1,734	2,101	2,552	2,878	3,610
19	0,861	1,328	1,729	2,093	2,539	2,861	3,579
20	0,860	1,325	1,725	2,086	2,528	2,845	3,552
21	0,859	1,323	1,721	2,080	2,518	2,831	3,527

22	0,858	1,321	1,717	2,074	2,508	2,819	3,505
23	0,858	1,319	1,714	2,069	2,500	2,807	3,485
24	0,857	1,318	1,711	2,064	2,492	2,797	3,467
25	0,856	1,316	1,708	2,060	2,485	2,787	3,450
26	0,856	1,315	1,706	2,056	2,479	2,779	3,435
27	0,855	1,314	1,703	2,052	2,473	2,771	3,421
28	0,855	1,313	1,701	2,048	2,467	2,763	3,408
29	0,854	1,311	1,699	2,045	2,462	2,756	3,396
30	0,854	1,310	1,697	2,042	2,457	2,750	3,385
31	0,853	1,309	1,696	2,040	2,453	2,744	3,375
40	0,851	1,303	1,684	2,021	2,423	2,704	3,307
50	0,849	1,299	1,676	2,009	2,403	2,678	3,261
60	0,848	1,296	1,671	2,000	2,390	2,660	3,232
120	0,845	1,289	1,658	1,980	2,358	2,617	3,160
	0,842	1,282	1,646	1,962	2,330	2,581	3,098

Tabela 3: Distribuição Qui-quadrado. Valores de c tais que p(χn>-c)=p, onde 'n' é o número de graus de liberdade.

g.l.	*0,99*	*0,95*	*0,9*	*0,5*	*0,1*	*0,05*	*0,025*	*0,01*	*0,005*
1	0,0002	0,0039	0,0158	0,4549	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0201	0,1026	0,2107	1,3863	4,6052	5,9915	7,3778	9,2103	10,5966
3	0,1148	0,3518	0,5844	2,3660	6,2514	7,8147	9,3484	11,3449	12,8382
4	0,2971	0,7107	1,0636	3,3567	7,7794	9,4877	11,1433	13,2767	14,8603
5	0,5543	1,1455	1,6103	4,3515	9,2364	11,0705	12,8325	15,0863	16,7496
6	0,8721	1,6354	2,2041	5,3481	10,6446	12,5916	14,4494	16,8119	18,5476
7	1,2390	2,1673	2,8331	6,3458	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,6465	2,7326	3,4895	7,3441	13,3616	15,5073	17,5345	20,0902	21,9550
9	2,0879	3,3251	4,1682	8,3428	14,6837	16,9190	19,0228	21,6660	23,5894
10	2,5582	3,9403	4,8652	9,3418	15,9872	18,3070	20,4832	23,2093	25,1882
11	3,0535	4,5748	5,5778	10,3410	17,2750	19,6751	21,9200	24,7250	26,7568
12	3,5706	5,2260	6,3038	11,3403	18,5493	21,0261	23,3367	26,2170	28,2995
13	4,1069	5,8919	7,0415	12,3398	19,8119	22,3620	24,7356	27,6882	29,8195
14	4,6604	6,5706	7,7895	13,3393	21,0641	23,6848	26,1189	29,1412	31,3193
15	5,2293	7,2609	8,5468	14,3389	22,3071	24,9958	27,4884	30,5779	32,8013
16	5,8122	7,9616	9,3122	15,3385	23,5418	26,2962	28,8454	31,9999	34,2672
17	6,4078	8,6718	10,0852	16,3382	24,7690	27,5871	30,1910	33,4087	35,7185
18	7,0149	9,3905	10,8649	17,3379	25,9894	28,8693	31,5264	34,8053	37,1565

19	7,6327	10,1170	11,6509	18,3377	27,2036	30,1435	32,8523	36,1909	38,5823
20	8,2604	10,8508	12,4426	19,3374	28,4120	31,4104	34,1696	37,5662	39,9968
21	8,8972	11,5913	13,2396	20,3372	29,6151	32,6706	35,4789	38,9322	41,4011
22	9,5425	12,3380	14,0415	21,3370	30,8133	33,9244	36,7807	40,2894	42,7957
23	10,1957	13,0905	14,8480	22,3369	32,0069	35,1725	38,0756	41,6384	44,1813
24	10,8564	13,8484	15,6587	23,3367	33,1962	36,4150	39,3641	42,9798	45,5585
25	11,5240	14,6114	16,4734	24,3366	34,3816	37,6525	40,6465	44,3141	46,9279
26	12,1981	15,3792	17,2919	25,3365	35,5632	38,8851	41,9232	45,6417	48,2899
27	12,8785	16,1514	18,1139	26,3363	36,7412	40,1133	43,1945	46,9629	49,6449
28	13,5647	16,9279	18,9392	27,3362	37,9159	41,3371	44,4608	48,2782	50,9934
29	14,2565	17,7084	19,7677	28,3361	39,0875	42,5570	45,7223	49,5879	52,3356
30	14,9535	18,4927	20,5992	29,3360	40,2560	43,7730	46,9792	50,8922	53,6720
31	15,6555	19,2806	21,4336	30,3359	41,4217	44,9853	48,2319	52,1914	55,0027
40	22,1643	26,5093	29,0505	39,3353	51,8051	55,7585	59,3417	63,6907	66,7660
50	29,7067	34,7643	37,6886	49,3349	63,1671	67,5048	71,4202	76,1539	79,4900
60	37,4849	43,1880	46,4589	59,3347	74,3970	79,0819	83,2977	88,3794	91,9517
120	86,9233	95,7046	100,6236	119,3340	140,2326	146,5674	152,2114	158,9502	163,6482

Tabela 4: Quantis da distribuição f para probabilidade p = p(f≥ ft) = 0,05, colunas = grau de liberdade do numerador, linhas = grau de liberdade do denominador,

G.L.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87
1000	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80

12	*13*	*14*	*16*	*18*	*20*	*30*	*40*	*60*	*120*	*1000*
243,91	244,69	245,36	246,46	247,32	248,01	250,10	251,14	252,20	253,25	254,19
19,41	19,42	19,42	19,43	19,44	19,45	19,46	19,47	19,48	19,49	19,49
8,74	8,73	8,71	8,69	8,67	8,66	8,62	8,59	8,57	8,55	8,53
5,91	5,89	5,87	5,84	5,82	5,80	5,75	5,72	5,69	5,66	5,63
4,68	4,66	4,64	4,60	4,58	4,56	4,50	4,46	4,43	4,40	4,37
4,00	3,98	3,96	3,92	3,90	3,87	3,81	3,77	3,74	3,70	3,67
3,57	3,55	3,53	3,49	3,47	3,44	3,38	3,34	3,30	3,27	3,23
3,28	3,26	3,24	3,20	3,17	3,15	3,08	3,04	3,01	2,97	2,93
3,07	3,05	3,03	2,99	2,96	2,94	2,86	2,83	2,79	2,75	2,71
2,91	2,89	2,86	2,83	2,80	2,77	2,70	2,66	2,62	2,58	2,54
2,79	2,76	2,74	2,70	2,67	2,65	2,57	2,53	2,49	2,45	2,41
2,69	2,66	2,64	2,60	2,57	2,54	2,47	2,43	2,38	2,34	2,30
2,60	2,58	2,55	2,51	2,48	2,46	2,38	2,34	2,30	2,25	2,21
2,53	2,51	2,48	2,44	2,41	2,39	2,31	2,27	2,22	2,18	2,14
2,48	2,45	2,42	2,38	2,35	2,33	2,25	2,20	2,16	2,11	2,07
2,42	2,40	2,37	2,33	2,30	2,28	2,19	2,15	2,11	2,06	2,02
2,38	2,35	2,33	2,29	2,26	2,23	2,15	2,10	2,06	2,01	1,97
2,34	2,31	2,29	2,25	2,22	2,19	2,11	2,06	2,02	1,97	1,92
2,31	2,28	2,26	2,21	2,18	2,16	2,07	2,03	1,98	1,93	1,88
2,28	2,25	2,22	2,18	2,15	2,12	2,04	1,99	1,95	1,90	1,85
2,25	2,22	2,20	2,16	2,12	2,10	2,01	1,96	1,92	1,87	1,82
2,23	2,20	2,17	2,13	2,10	2,07	1,98	1,94	1,89	1,84	1,79
2,20	2,18	2,15	2,11	2,08	2,05	1,96	1,91	1,86	1,81	1,76
2,18	2,15	2,13	2,09	2,05	2,03	1,94	1,89	1,84	1,79	1,74
2,16	2,14	2,11	2,07	2,04	2,01	1,92	1,87	1,82	1,77	1,72
2,15	2,12	2,09	2,05	2,02	1,99	1,90	1,85	1,80	1,75	1,70
2,13	2,10	2,08	2,04	2,00	1,97	1,88	1,84	1,79	1,73	1,68
2,12	2,09	2,06	2,02	1,99	1,96	1,87	1,82	1,77	1,71	1,66
2,10	2,08	2,05	2,01	1,97	1,94	1,85	1,81	1,75	1,70	1,65
2,09	2,06	2,04	1,99	1,96	1,93	1,84	1,79	1,74	1,68	1,63
2,00	1,97	1,95	1,90	1,87	1,84	1,74	1,69	1,64	1,58	1,52
1,92	1,89	1,86	1,82	1,78	1,75	1,65	1,59	1,53	1,47	1,40
1,83	1,80	1,78	1,73	1,69	1,66	1,55	1,50	1,43	1,35	1,27
1,76	1,73	1,70	1,65	1,61	1,58	1,47	1,41	1,33	1,24	1,11

BIBLIOGRAFIA

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Snow G (2016). TeachingDemos: Demonstrations for Teaching and Learning. R package version 2.10. https://CRAN.R-project.org/package=TeachingDemos

Sokal RR & Rohlf FJ (1995) Biometry. 3rd ed., W. H. Freeman & Co., New York. 887 pp

Wei T & Simko V (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Wickham H, Hester J & Chang W (2018). devtools: Tools to Make Developing R Packages Easier. R package version 1.13.6. https://CRAN.R-project.org/package=devtools

Zar, JH (2010) Biostatistical Analysis. 5nd edition. Prentice Hall, New Jersey, 944 p.

Título Bioestatística: curso prático utilizando R e Excel
Autoria José Roberto Botelho de Souza

Formato *E-book*
Tipografia Minion Pro