ESTATÍSTICA APLICADA À BIOTECNOLOGIA

Pós Graduação em Toxinologia – Instituto Butantan

Pedro da Silva Peixoto
Instituto de Matemática e Estatística
Universidade de São Paulo
(pedrospeixoto@yahoo.com.br)

jun-12

Aulas

- Aulas: 6
- Dias: 28/05 até 02/07 - Segundas-feiras
- Horário: 14h00 – 17h00
- Local: Sala Barbosa Rodrigues

Avaliação

- Projeto Final

Meu contato

- Por e-mail: pedrospeixoto@yahoo.com.br

Conteúdo

Aula 1	Análise descritiva
Aula 2	Inferência estatística paramétrica
Aula 3	Inferência não paramétrica
Aula 4	Inferência multifatorial
Aula 5	Análises de regressão e de dados categóricos
Aula 6	Tópicos em estatística

Projeto Final

- Desenvolver um projeto de análise estatística que envolva as principais ferramentas estudadas
- A cada aula será proposta uma parte do projeto, para ser desenvolvido ao longo do curso
- A nota será dada pelo relatório final que será entregue contendo as análises realizadas

Parte teórica

- Serão apresentados os conceitos pertinentes.
- Demonstrações? Cálculos? Matemática?

Parte prática

- Aplicar os conceitos em dados reais
- Será necessário que o aluno tenha acesso a um computador contendo:
  - Microsoft Excel, ou similar gratuito LibreOffice Calc
  - Microsoft Word, ou similar gratuito LibreOffice Writer
  - Possibilidade de instalação de outros softwares (Bioestat)

Referências

Teórica

- *Concepts & Applications of Inferential Statistics* - Richard Lowry http://faculty.vassar.edu/lowry/webtext.html
- Princípios de Bioestatística – Pagano e Gauvreau
- Estatística Básica – Bussab e Morettin
- Biostatistical Analysis – Zar

Prática

- Funções estatísticas no Excel (com exemplos)
- Em Português: http://office.microsoft.com/pt-br/excel-help/funcoes-estatisticas-HP005203066.aspx
- Em inglês: http://office.microsoft.com/en-us/excel-help/statistical-functions-HP005203066.aspx

CONCEITOS GERAIS E ANÁLISE DESCRITIVA

Estatística Aplicada à Biotecnologia

jun-12

INTRODUÇÃO

Propósitos

Síntese

- Resumir informações para melhorar compreensão dos dados
- Média
- Mediana
- Devio Padrão
- Outras Estatísticas

Visualização

- Interpretações e análises baseadas em elementos visuais
- Gráficos
- Diagramas

Inferência

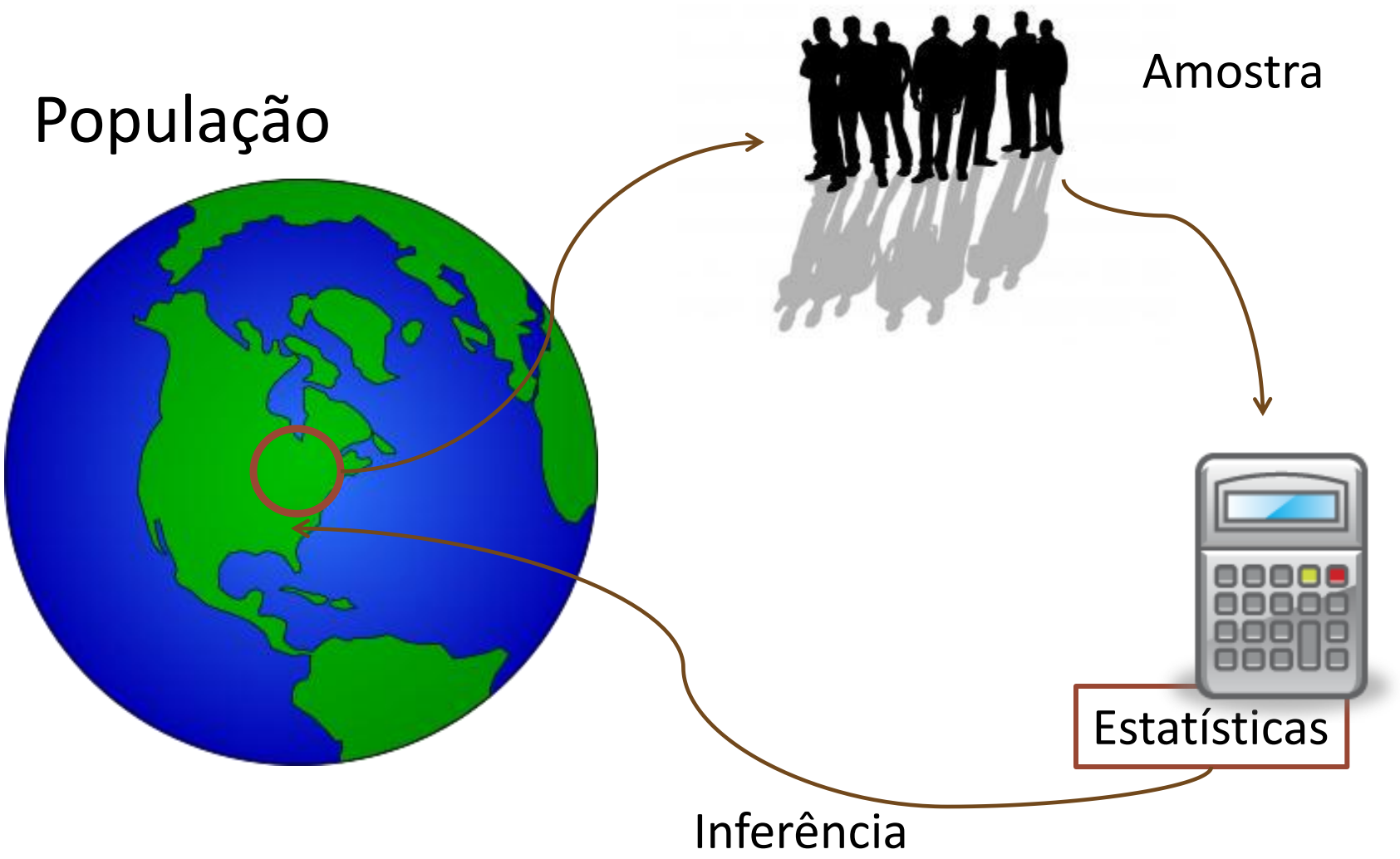
- Inferir o comportamento de uma população usando como base uma amostra
- Testes de hipóteses

Importância da Estatística no Projeto de Pesquisa

- Planejando a coleta de dados
- Organizando os dados
- Análises ao longo do processo para guiar a pesquisa
- Interpretação dos resultados
- Comparações, discussões e conclusões

A AMOSTRA

Inferência Estatística



Amostra

População

É o conjunto de todos os elementos sob investigação.

Em um experimento desejamos entender melhor as características da população.

Em geral NÃO temos acesso aos dados de toda a população**.** Temos uma amostra.

Exemplo

Desejamos saber se um novo tratamento para hipertensão é ou não eficaz.

Gostaríamos que o tratamento pudesse ser usado por TODOS aqueles que sofrem de hipertensão.

Nossa população é o conjunto de todas as pessoas que sofrem de hipertensão

Amostra e População

- Uma **amostra refere-se a uma certa população**, e as conclusões sobre esta amostra inferem sobre esta população.
- EX: Amostra de ratos suiços recém nascidos irão trazer informações sobre ratos suiços recem nascidos. Nada pode-se dizer em relação a outros tipos de ratos.

Aleatoriadade

- Uma vez determinada a população, a escolha dos elementos não pode ter viés.
- EX: Se vamos analisar coelhos de um certo tipo, não podemos, por exemplo, pegar só os coelhos mais “calminhos”

Tamanho da amostra

- Devido ao custo de obtenção de amostras vivas, o tamanho é baixo.
- Cuidado para não ser tão baixo!
- EX: Não é possivel realizar alguns testes estatísticos com menos de 4 elementos na amostra

Dicas

Anotações

- Anote o máximo de informações que puder sobre a amostra coletadas
- EX: Equipamento usado, se teve alguma dificuldade ou demorou mais tempo, quem fez a coleta,...

Tréplicas

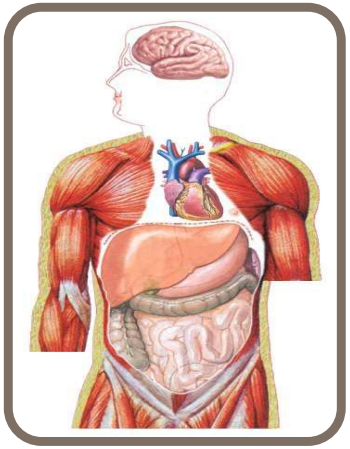
- Para reduzir o erro de medição realize a coleta em triplicata
- EX: Faça a medição do fator de interesse 3 vezes e considere a média das medições como valor a ser analisado.

Outras formas de amostragem

- Sistematizada
- Estratificada
- Tenha segurança de que você sabe o que está fazendo!

PRIMEIRAS ESTATÍSTICAS

POR ONDE COMEÇAR?



Variáveis Qualitativas

- **Nominal**
  - EX: Orgão afetado por um tratamento
  - Moda, proporções
- **Ordinal**
  - EX: Pouca, muita dor
  - Mediana, proporções

Variáveis Quantitativas

- **Ordinal**
  - EX: 1,2,3,4,5,....
  - Mediana, quartis, percentis
- **Intervalar ou de razão**
  - EX: Absorbância, pressão, comprimento, volume, %
  - Média, DP, Mediana, quartis, %, ...

Variáveis Qualitativas

	Grupo A	Grupo B	Grupo C	Total
Infectado	13	9	5	27
Não infectado	7	11	15	33
Total	20	20	20	60

Variáveis Quantitativas

Amostra	Grupo A	Grupo B	Grupo C
1	2,5	3,5	4
2	3	2,7	3,5
3	1,3	2,4	3,1
4	2,6	.	.
5	.	.	.
.	.	.	.
.			
.			

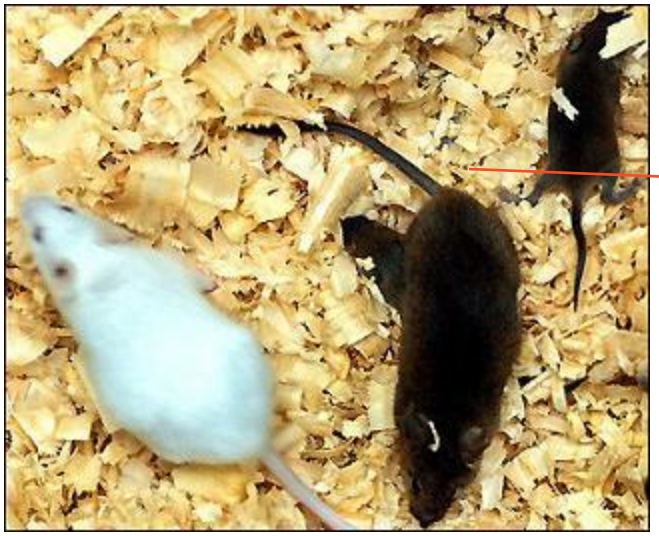
Grupos dependentes ou independentes ?

DEPENDENTES

Medidas em vários órgãos de um mesmo elemento amostral

Medidas em vários tempos de um mesmo elemento amostral

Pergunte-se: Se eu tirar essa medida deste grupo, e logo tirar o elemento amostral, ele irá também ser tirado de outros grupos? Se sim, então há dependência!

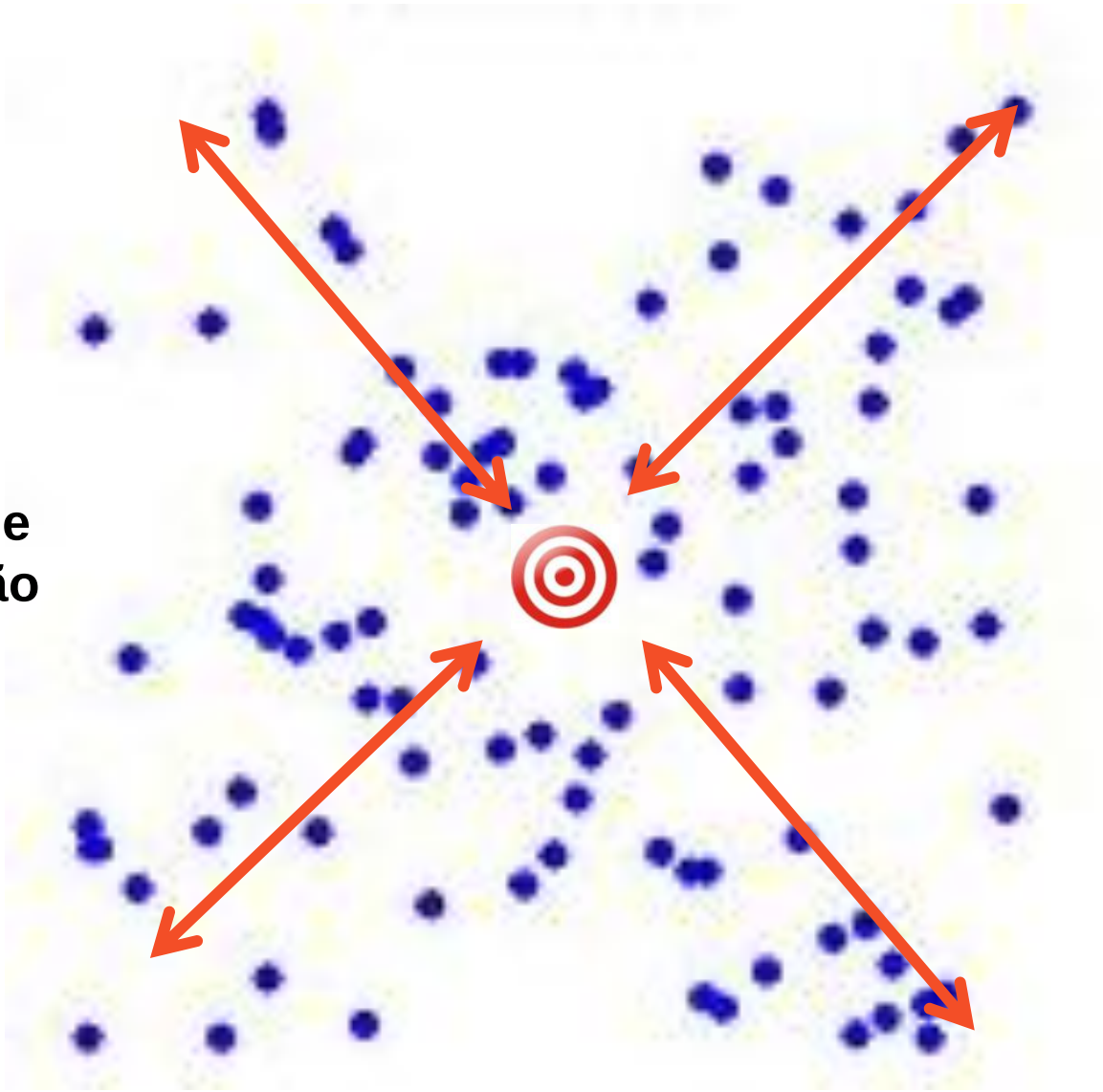


INDEPENDENTES

Cada medição provém de elementos amostrais distintos

Ao tirar uma medida, ou um elemento, os outros grupos não são afetados

Posição e Dispersão



Medidas de posição (centralizadoras)

MÉDIA

• É a medida centralizadora mais usada.

• Para calcular faça some os elementos e divida pelo número de elementos

• Ex: (3 + 5 + 9 + 4 + 8 + 2) / 6 = 5.16.

3	5	9	4	8	2

MEDIANA

• Leva em conta não a grandeza dos números mas sua disposição

•Ordena-se os elementos e toma-se o central (ou média dos centrais)

• Para calcular ordene os valores: 2, 3, 4, 5, 7, 8, 9.

• A mediana é vale 5

3	5	9	7	4	8	2

DESVIO PADRÃO AMOSTRAL

• É a medida de dispersão mais usada.

•Use sempre um software para fazer a conta.

• Soma-se as diferenças com relação às medias ao quadrado, e depois divide-se por N-1 (e não N, Por quê?). Depois toma-se a raíz.

• Exemplo:

3	5	9	4	8	2

•Média: 5.16

•Desvios com relação a média

-2.17	-0.17	3.83	-1.17	2.83	-3.17

•Eleva ao Quadrado

4.71	0.03	14.67	1.37	8.01	10.05

•Soma e divide por 5 : 7.76 (esta é a variância)

•Tomando a raíz temo o desejado: 2.78

Variação Inter-Quartis

- Leva a posição relativa dos números e não seu valor em si
- Os quartis apenas dividem a amostra.
- Os quartis mais famosos são os quartis de número 1 (25%) e 3 (75%)
- O quartil de número 2 é a mediana (50%)
- A Variação interquartil é a diferença entre o 3 e o 1 quartil

- Ex:

3	5	9	4	8	2	5

- Em ordem temos:

2	3	4	5	5	8	9

- 25% dos dados abaixo de : 3.5
- 50% dos dados abaixo de : 5
- 75% dos dados abaixo de : 6.5

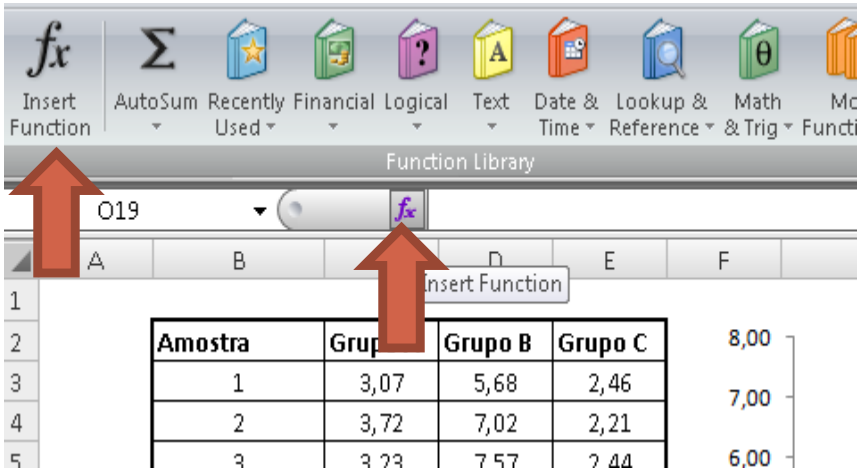
- Variação Interquartis : 3

- Deixa as contas para o computador para evitar confusão.

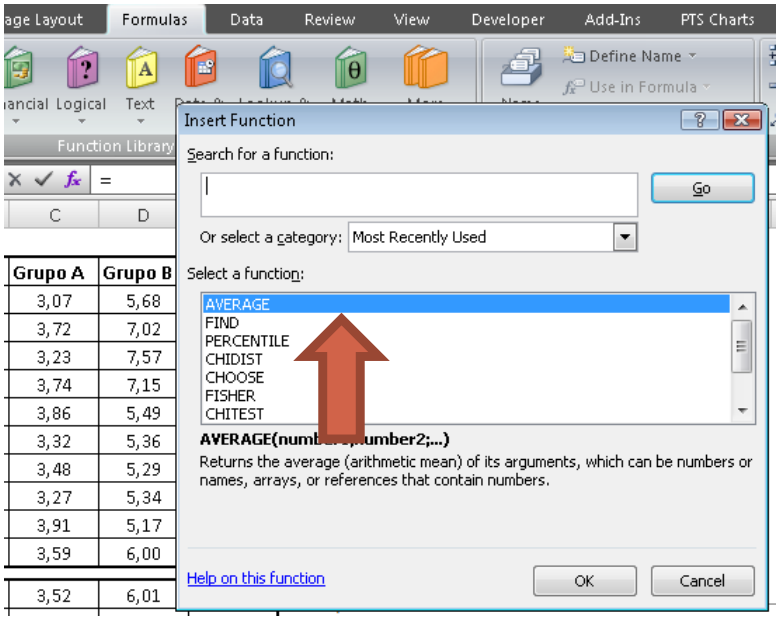
Estatísticas em Variáveis Quantitativas

Estatísticas	Tempo 1	Tempo 2	Tempo 3
Média			
Desvio Padrão			
Mediana			

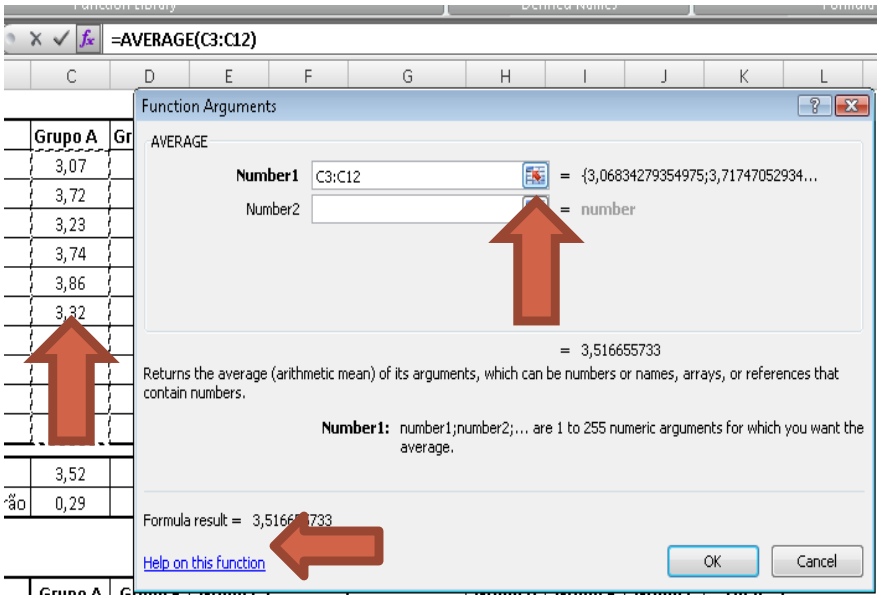
Excel - Fórmulas



Excel – Média (Average)

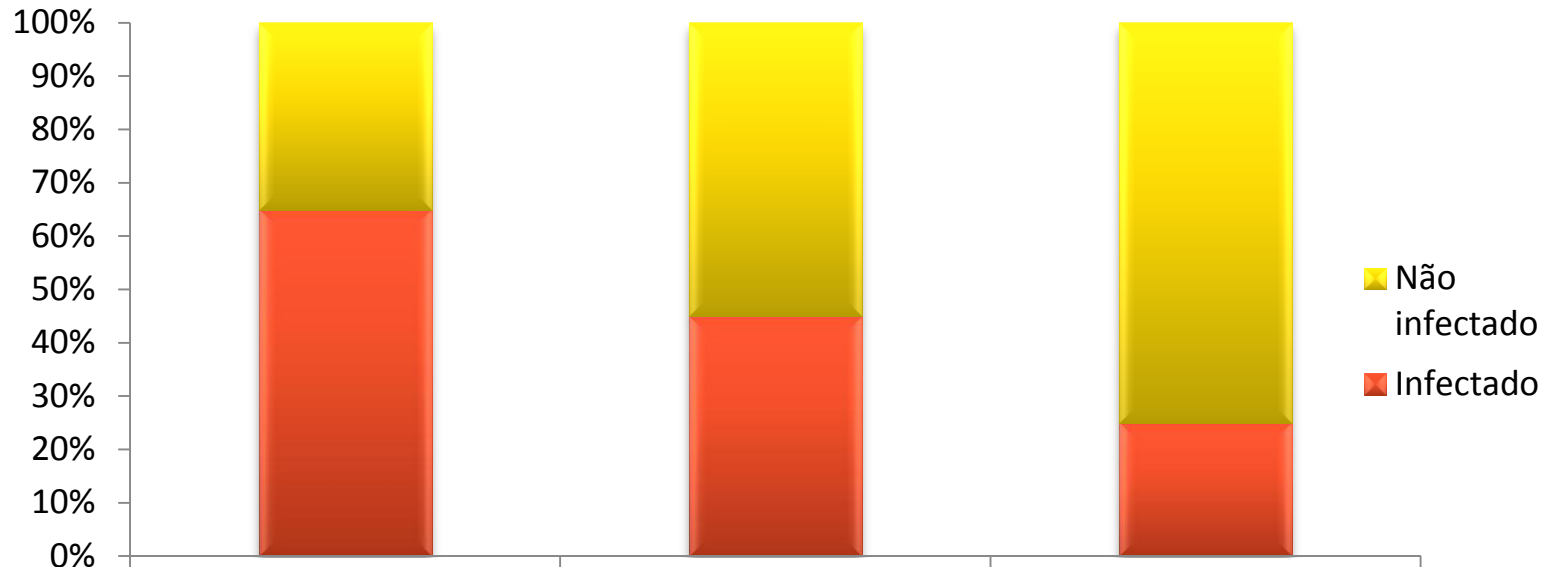


Excel – Seleção de Dados

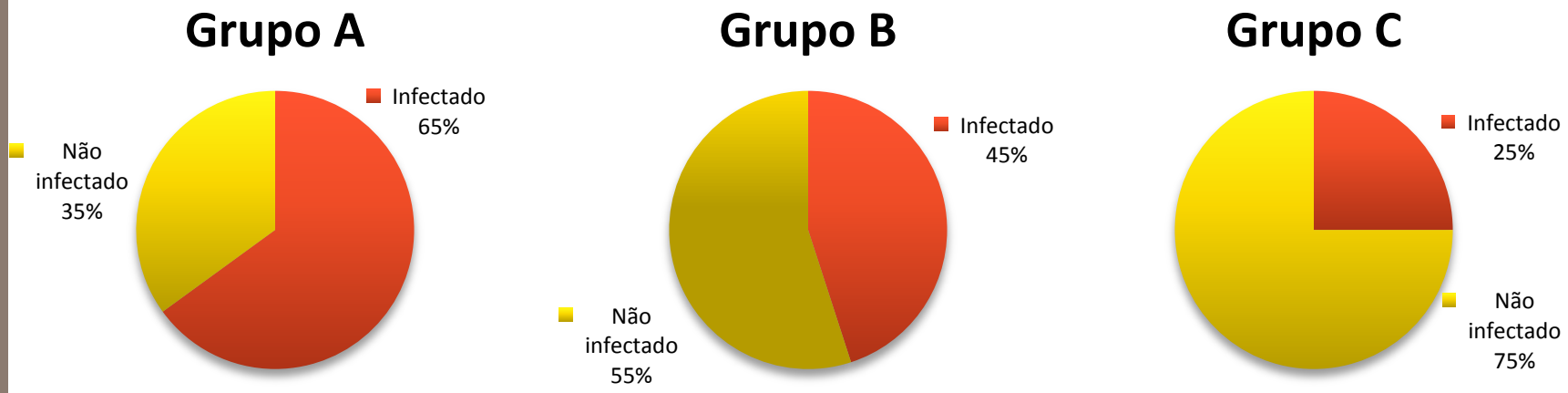


VISUALIZAÇÃO GRÁFICA

Visualização de variáveis qualitativas



	Grupo A	Grupo B	Grupo C
Não infectado	35%	55%	75%
Infectado	65%	45%	25%



Grupos INDEPENDENTES

Amostra	Grupo A	Grupo B	Grupo C
1	3,07	5,68	2,46
2	3,72	7,02	2,21
3	3,23	7,57	2,44
4	3,74	7,15	3,82
5	3,86	5,49	3,49
6	3,32	5,36	2,25
7	3,48	5,29	2,27
8	3,27	5,34	3,86
9	3,91	5,17	2,67
10	3,59	6,00	2,04
Média	3,52	6,01	2,75
Desvio Padrão	0,29	0,90	0,70

Gráfico de pontos

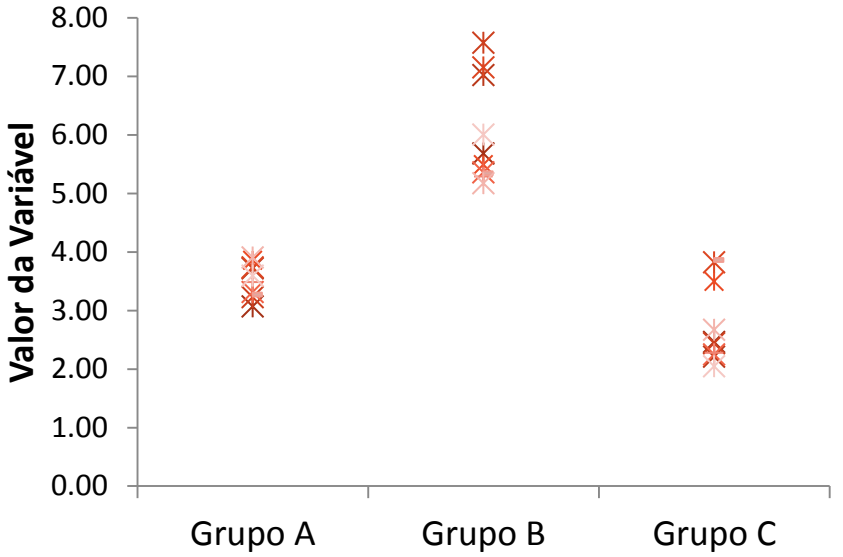


Gráfico Resumo

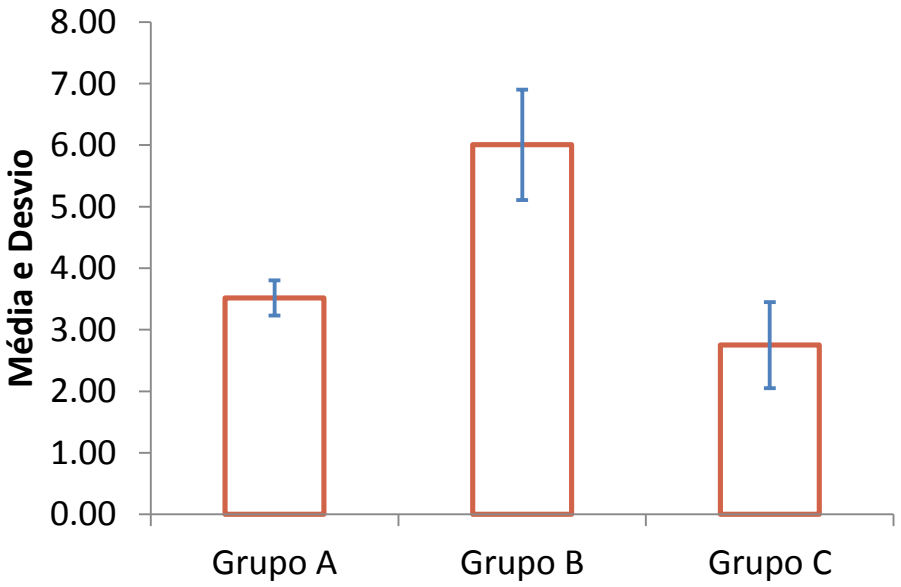
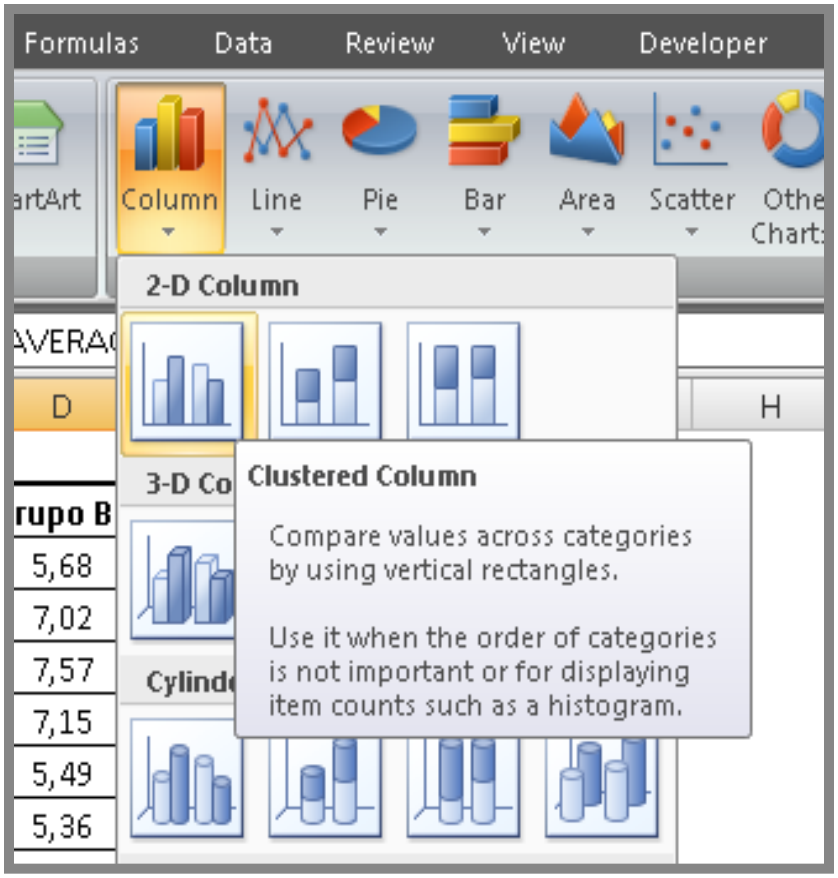
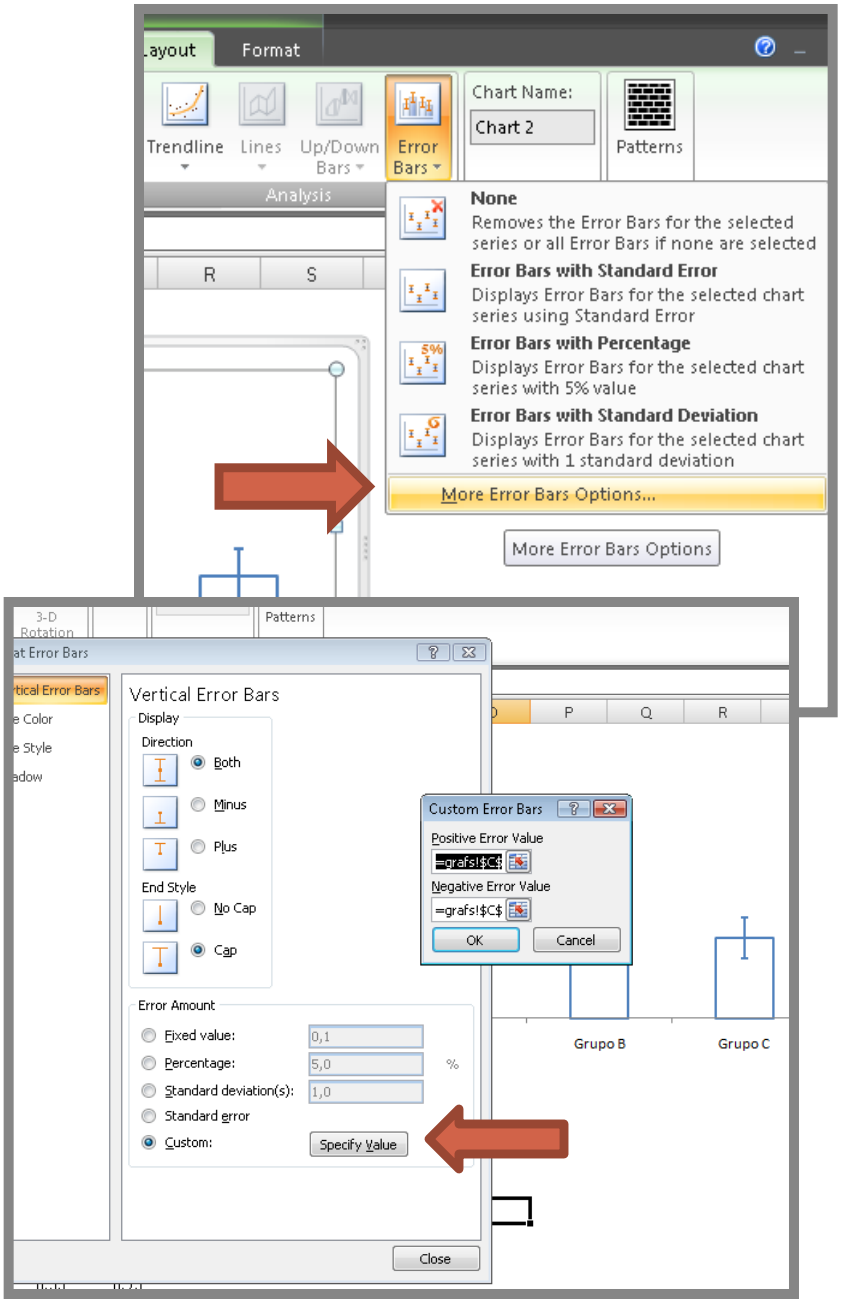


Gráfico de Barras com Erros



Barras de Erros



Variáveis CORRELACIONADAS

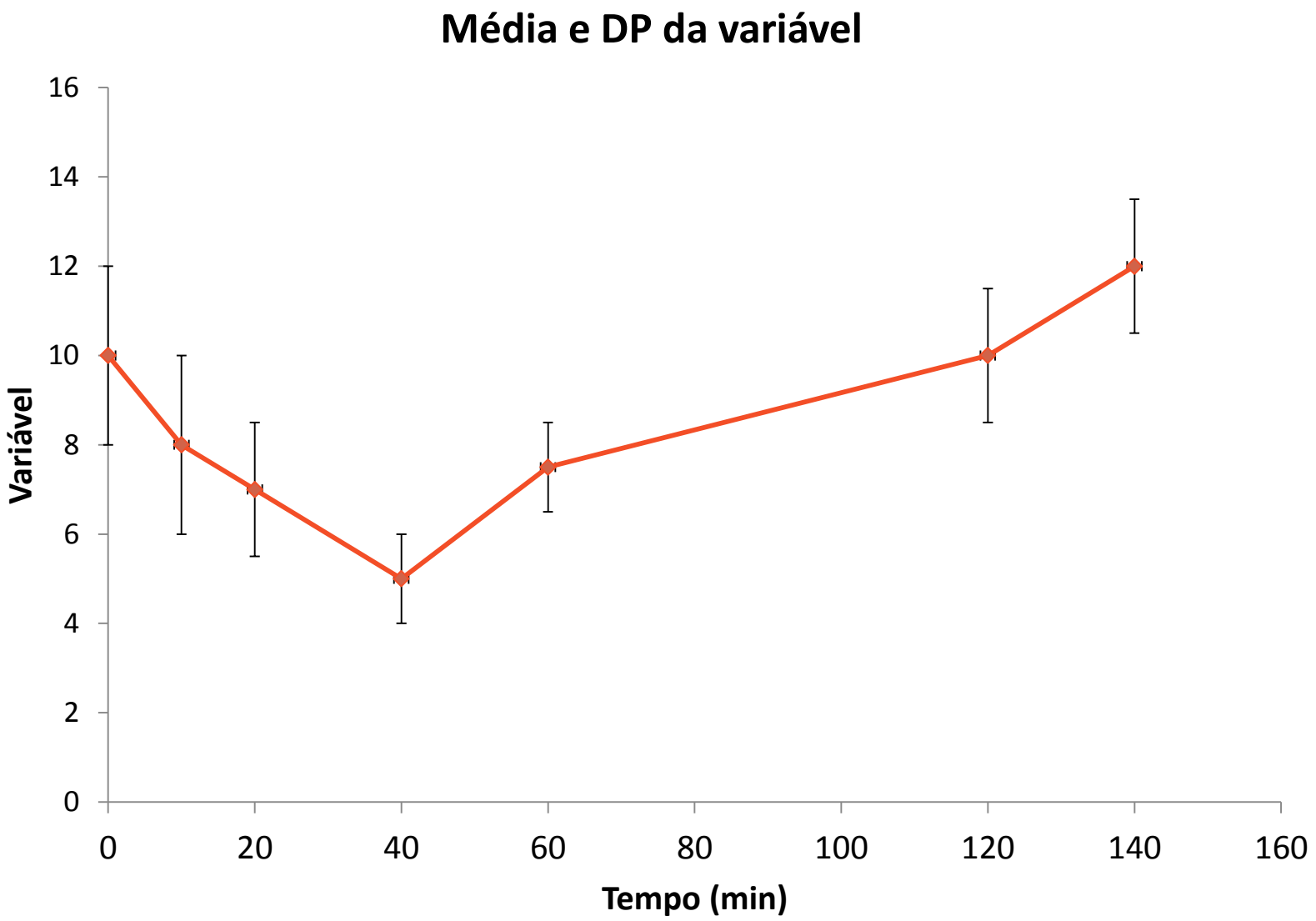
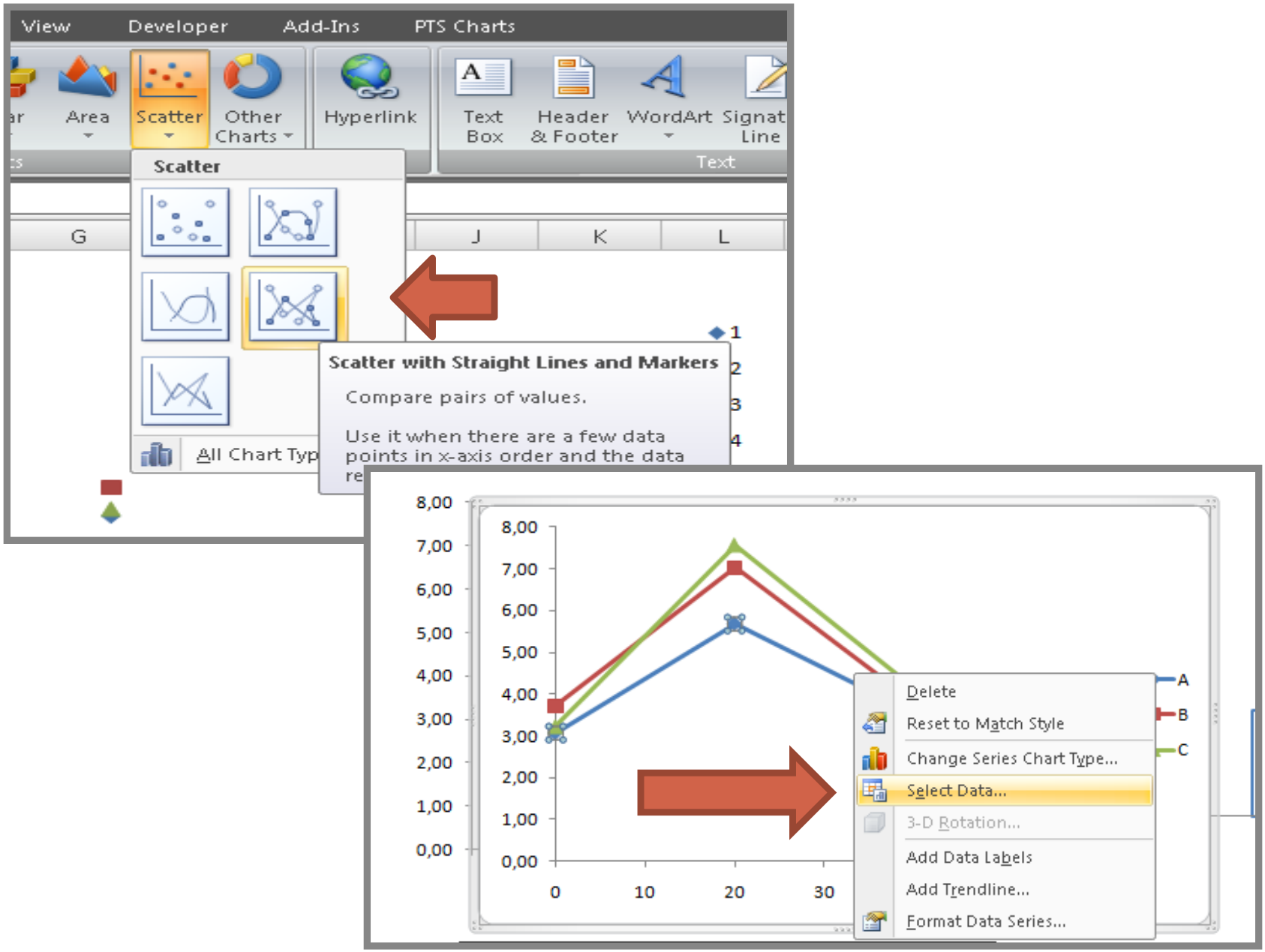


Gráfico de Linhas



VARIAÇÃO PERCENTUAL (DELTA PERCENTIL)

Motivação e Cálculos

Dados

- Aplica-se a dados quantitativos de variáveis DEPENDENTES!
- EX: Evolução do IC em 3 instantes de tempo

Movitação

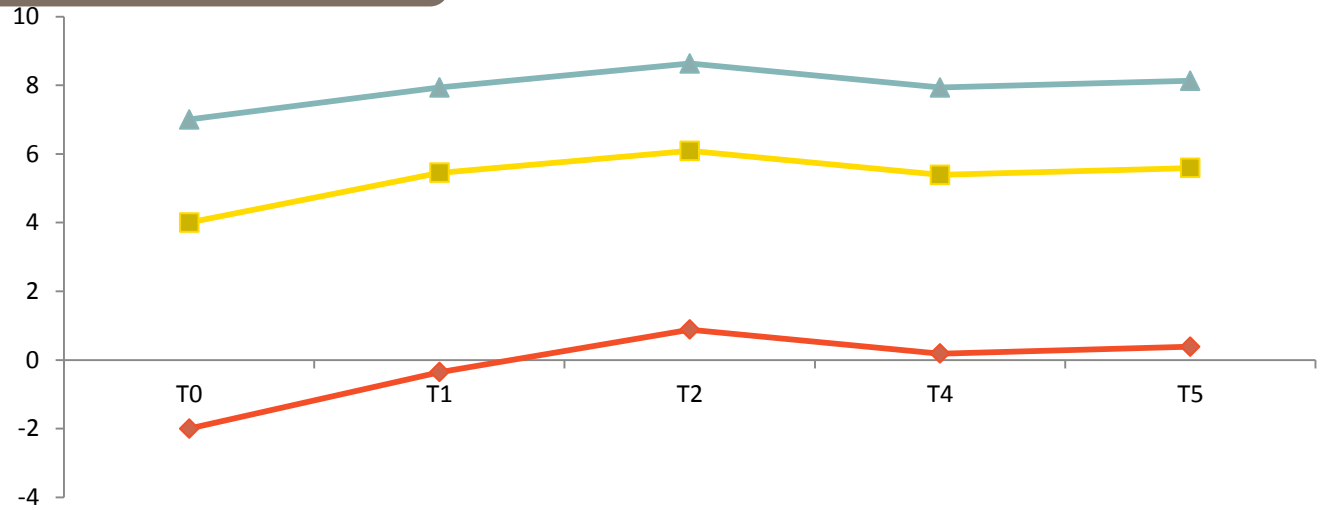
- O interesse está apenas na **variação** dos valores e não nos números em si.
- EX: Não inporta se o IC é alto ou baixo, mas sim se ele aumentou ou diminui no tempo.

Cálculo

- Variação Percentual

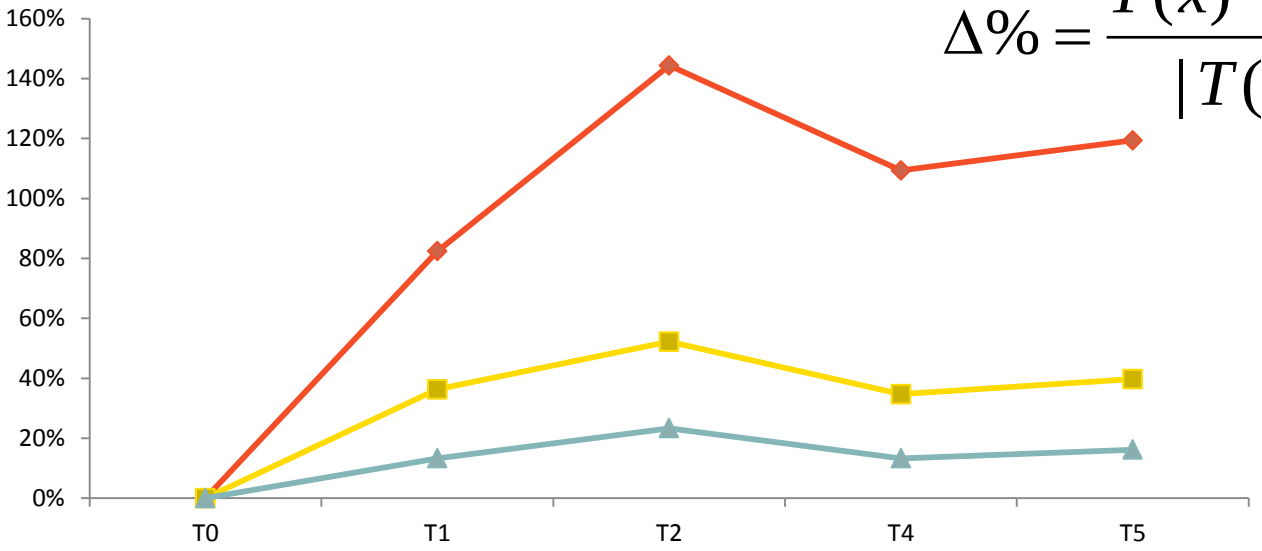
$$= \frac{(\text{Valor no instante de interesse} - \text{Valor no instante inicial})}{(\text{Valor absoluto no instante inicial})} \times 100$$

Evolução Temporal

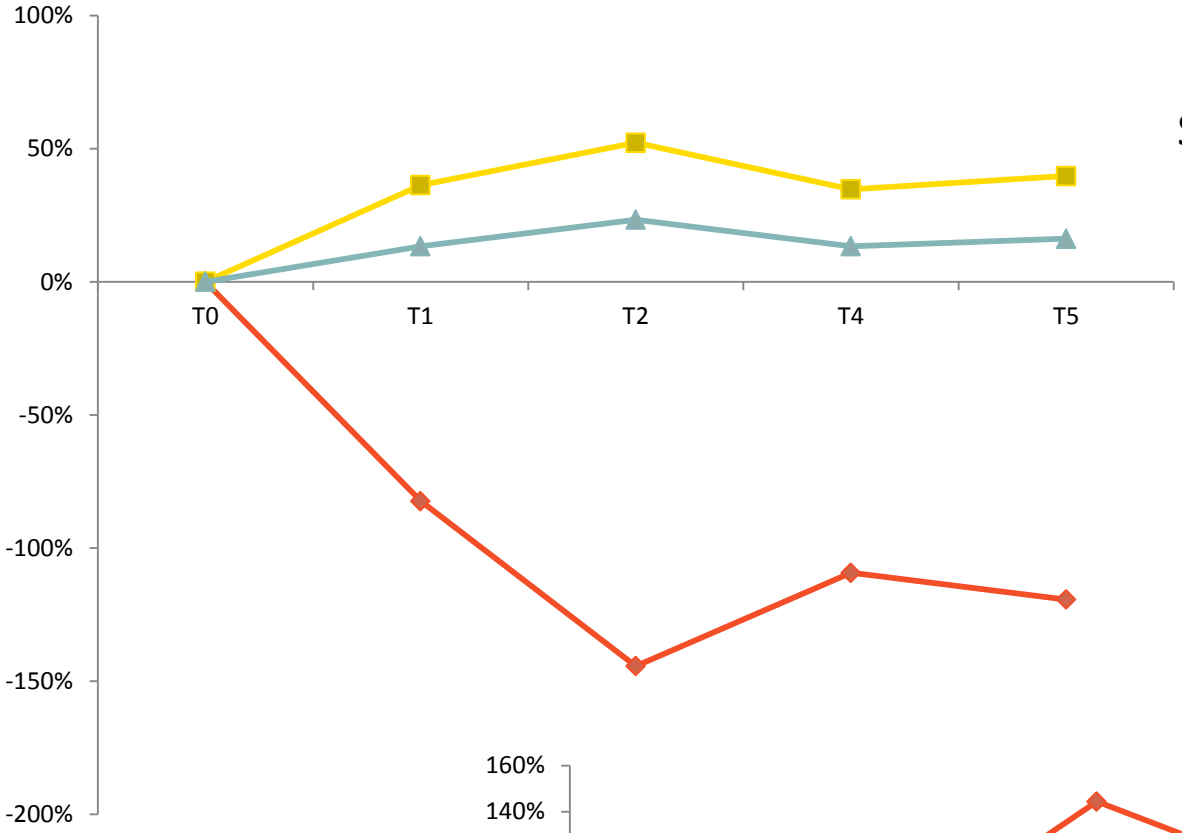


Evolução Percentual

$$\Delta\% = \frac{T(x) - T(0)}{|T(0)|}$$



CUIDADOS!



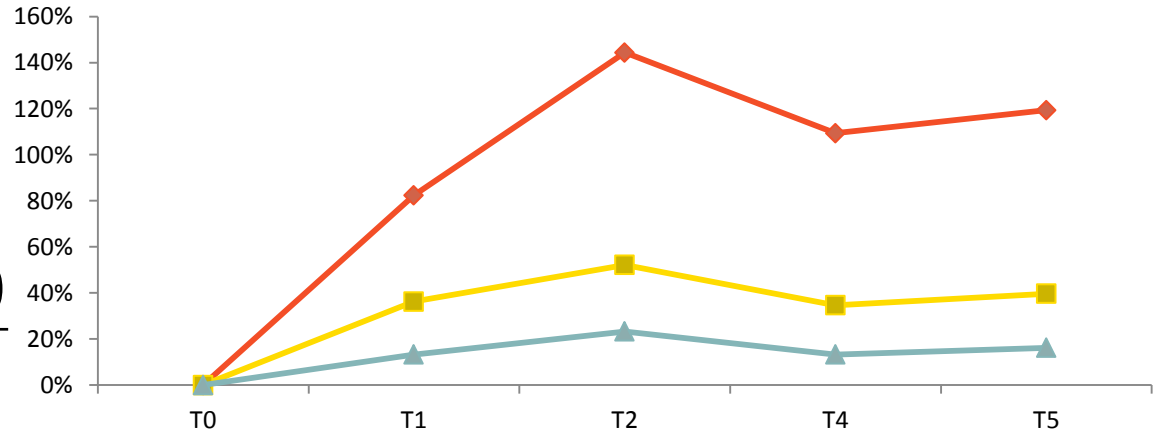
SEM O MÓDULO!!!

$$\Delta\% = \frac{T(x) - T(0)}{T(0)}$$

ERRADO!!

CORRETO:

$$\Delta\% = \frac{T(x) - T(0)}{|T(0)|}$$



BOXPLOT

Motivação e Cálculos

Dados

- Aplica-se a amostras com dados quantitativos

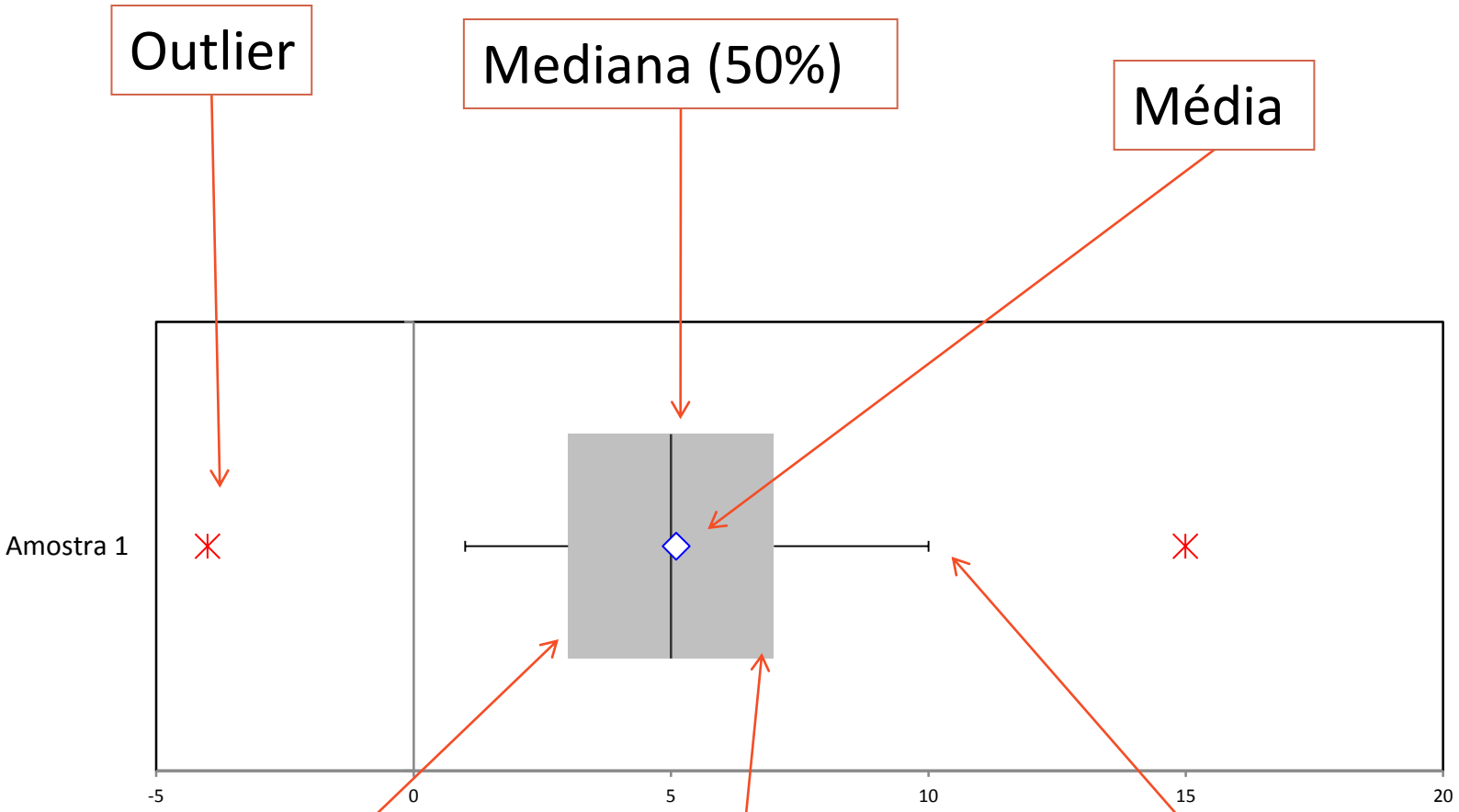
Movitação

- Resumir a distribuição dos dados de forma gráfica com a menor perda de informações possível.

Cálculos

- Mediana (Divide os dados pela metade por ordem de grandeza)
- Quartis (Divisão dos dados por ordem de grandeza na medida)
  - Q1 divide em 25%
  - Q2 divide em 50% (equivale a mediana)
  - Q3 divide em 75%
- Variação inter-quartil
  - Q3 – Q1 : Fornece uma medida de variabilidade dos dados

Descrição


Outlier
Mediana (50%)
Média
Amostra 1
-5
0
5
10
15
20
1º. Quartil (25%)
3º. Quartil (75%)
1,5*(Variação Interquartil) ou Máximo

Como fazer ?

Tabela de dados para o Boxplot:

ATUALIZAR BOXPLOT

Título: Coloque aqui seu título

Eixo y: Coloque aqui o quer escrito no eixo x

EXCEL:
www.ime.usp.br/~pedrosp

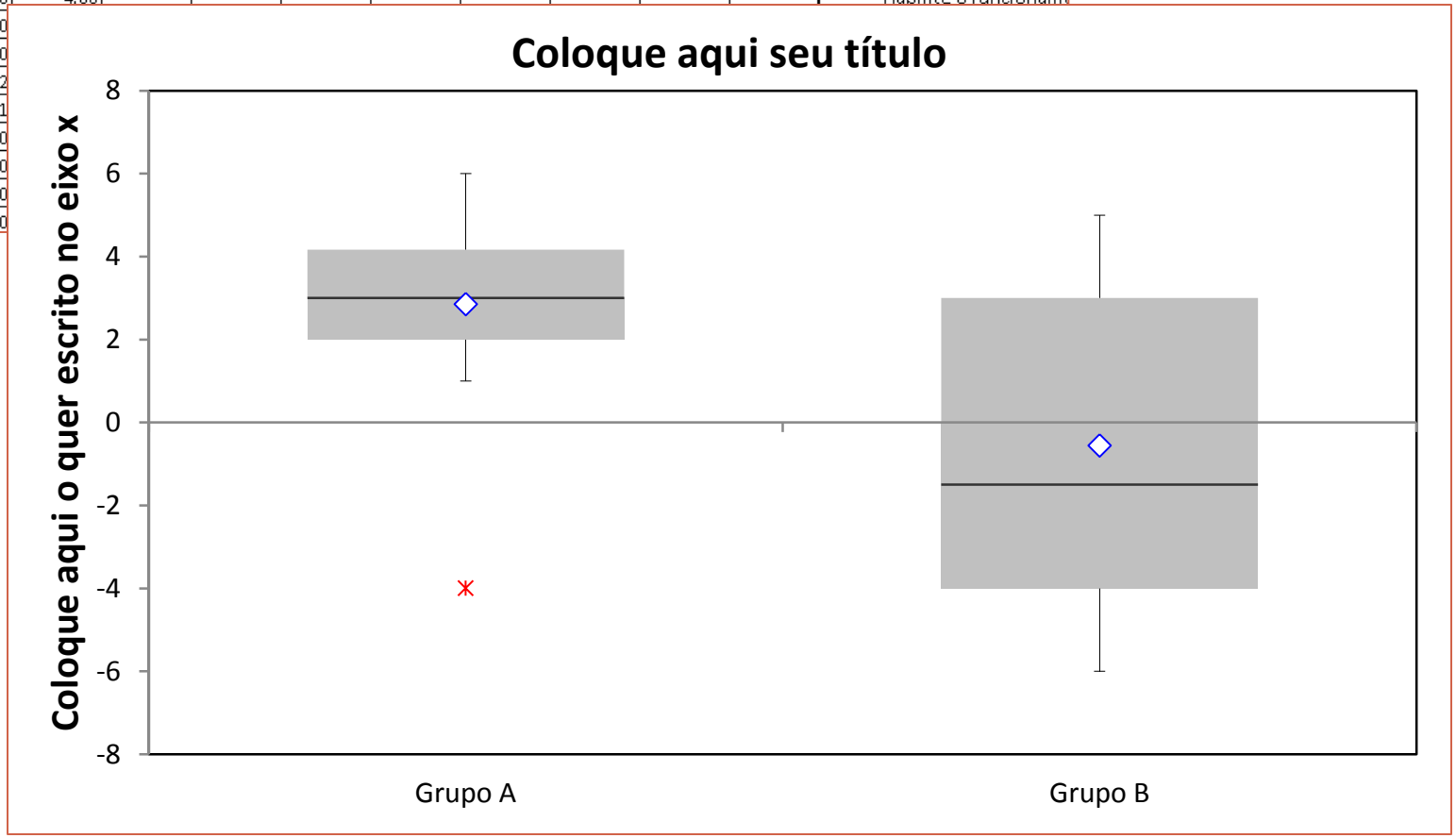
Autor

Pedro da Silva Peixot

Instruções

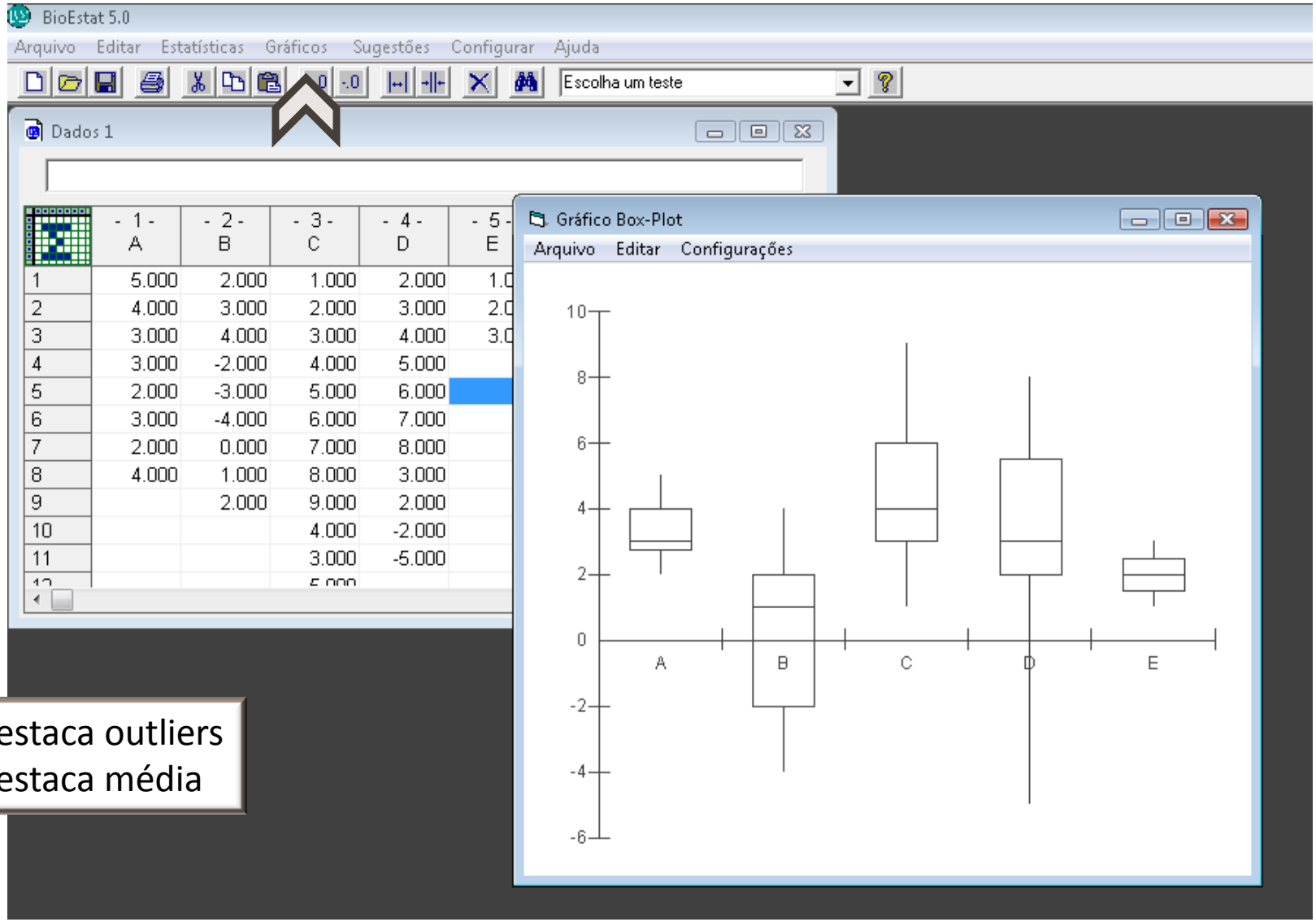
* Habilite o funcionam

	Grupo A	Grupo B						..	..	..
X1	5,00	2,00								
X2	4,17	3,00								
X3	3,00	4,00								
X4	3,00									
X5	2,00									
X6	2,62									
X7	2,01									
X8	4,00									
X9	1,00									
X10	2,00									
X11	3,00									



Outros Software

BioEstat : http://www.mamiraua.org.br/download/



Não destaca outliers
Não destaca média

SOFTWARES

Software

Cálculos das Estatísticas Básicas

- **Microsoft Excel**
  - Software comercial, mas presente em muitos computadores

Área de Biológicas

- **BioEstat (Livre) – Simples mas completo**
  - **www.mamiraua.org.br/download/**
- **OpenEPi (Livre) - Online – Só testes paramétricos**
  - **www.openepi.com**
- BioStat (comercial) - Completo
  - www.analystsoft.com/br/products/biostat/
- Sigmaplot+SigmaStat (comercial) – Completo

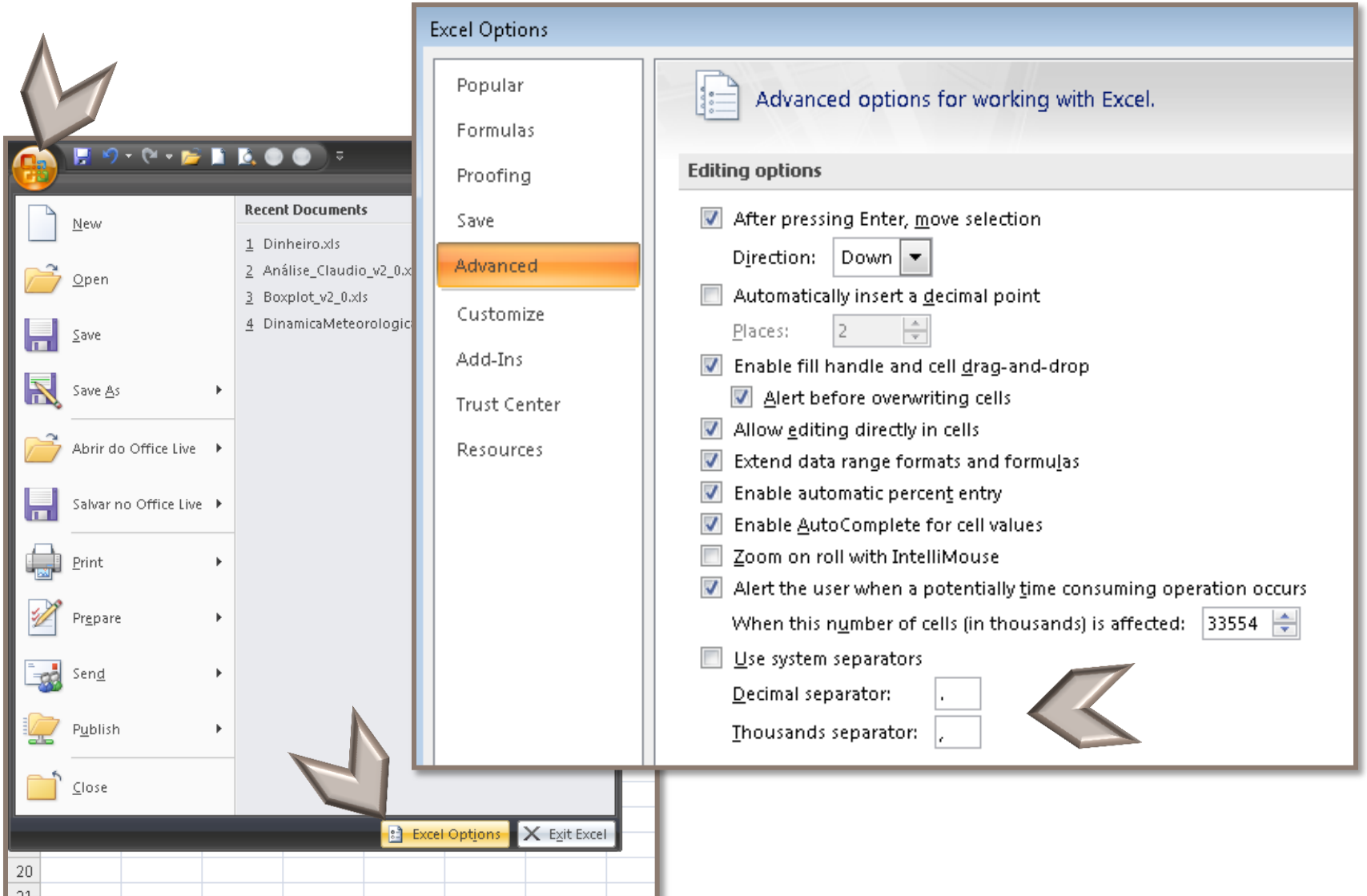
Gerais

- **Minitab (comercial)** – Completo e fácil de usar
- SPSS (comercial) – Completo e fácil de usar
- S-Plus (comercial) – Completo mas menos amigável
  - Possui versão livre – R-Statistics – exige programação
- Statistica (comercial)

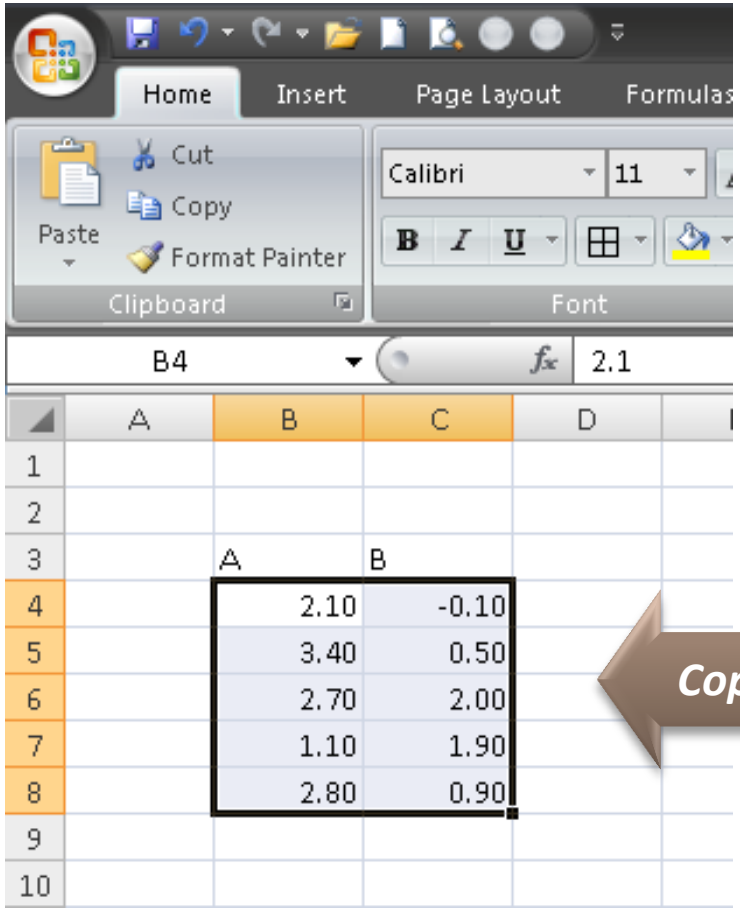
Dados no BioEstat

BioEstat utiliza padrão americano de casas decimais (usa “3.14” no lugar de “3,14”)

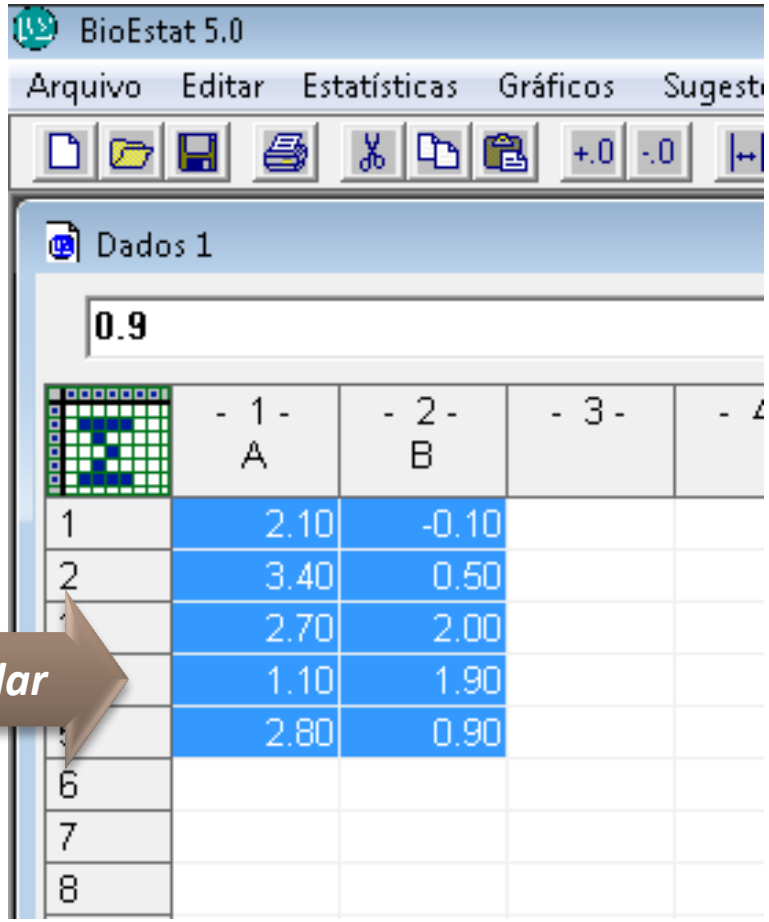
Se o Excel estiver configurado com “,” faça o seguinte:



Dados no BioEstat



Copiar e colar



PROJETO

Projeto : Parte 1 – Análise descritiva de variável independente

Encontrar 1 variável quantitativa que tenha:

- 2 ou mais grupos (1 Controle e demais de Teste)
- Tamanho entre 6 e 12 elementos em cada grupo

Organizar os dados em forma de tabela

Calcular as estatísticas: Média, Desvio Padrão, Mediana, 1º. E 3º. Quartis, Intervalo Interquatis

Fazer um gráfico de barras com médias e barras de erros com 1 desvio padrão para cima e 1 para baixo

Fazer um Boxplot com os grupos

Crair um relatório (Word - .doc), resumido, contendo informações sobre os dados, justificando independência, indicando o propósito da análise descritiva, e contendo a análise descritiva

Dicas

Organize os dados e faça os cálculos das estatísticas no Excel

Mesmo que isso demore, é um tempo investido para você economizar tempo no futuro

No Excel aperte F1 para obter ajuda e digite por exemplo: “media” que ele irá lhe ajudar!

NÃO FAÇAM CONTAS NA MÃO OU NA CALCULADORA!

Outras Dicas

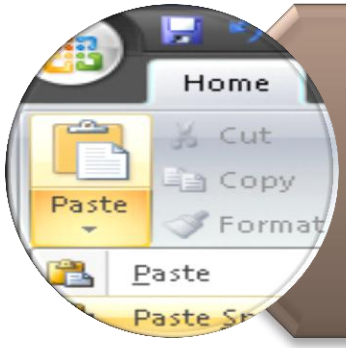


GRÁFICO NO EXCEL --> WORD

- Clicar em copiar no gráfico
- Clicar em colar ESPECIAL no WORD
- Colar como FIGURA, BITMAP ou METAARQUIVO AVANÇADO

USUÁRIOS DE MACINTOSH

- Use o Microsoft Office para Mac para os gráficos
- Utilize os programas online para os testes (ver final da aula 2)
- É possível instalar o BioEstat no Mac (mas não é tão simples)

DÚVIDAS?

OBRIGADO E ATÉ A PRÓXIMA AULA!!

pedrospeixoto@yahoo.com.br

INFERÊNCIA I – TESTES PARAMÉTRICOS

Estatística Aplicada à Biotecnologia

jun-12

INTERVALO DE CONFIANÇA E DISTRIBUIÇÕES

Erro de medida

- Uma medida obtida de uma amostra é uma ESTIMATIVA da medida real para a população
- Logo contém ERRO !
- O erro depende da distribuição da variável na população e na amostra

Intervalo de Confiança

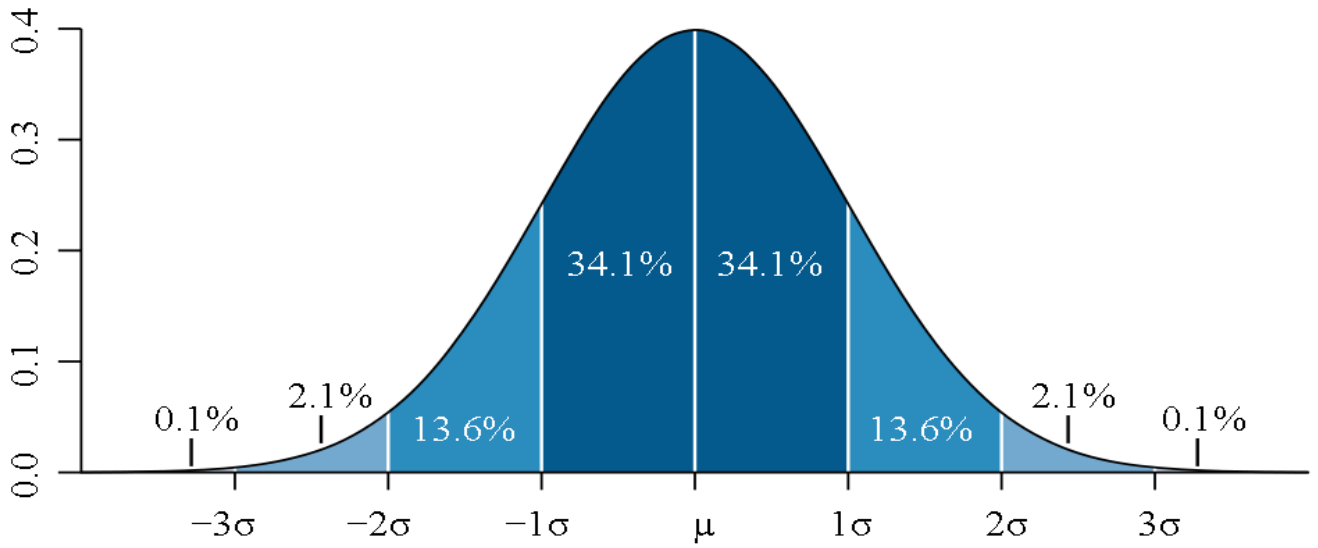
- Para cada medida é possível atribuirmos uma noção de confiabilidade usando a amostra.
- Exemplo: Calculamos uma média 5 para uma amostra, mas com base em sua distribuição, podemos estimar que a média da população está na verdade entre entre 4 e 6 com 95% de confiança

Distribuições

- Normal
- T-student
- Qui-quadrado

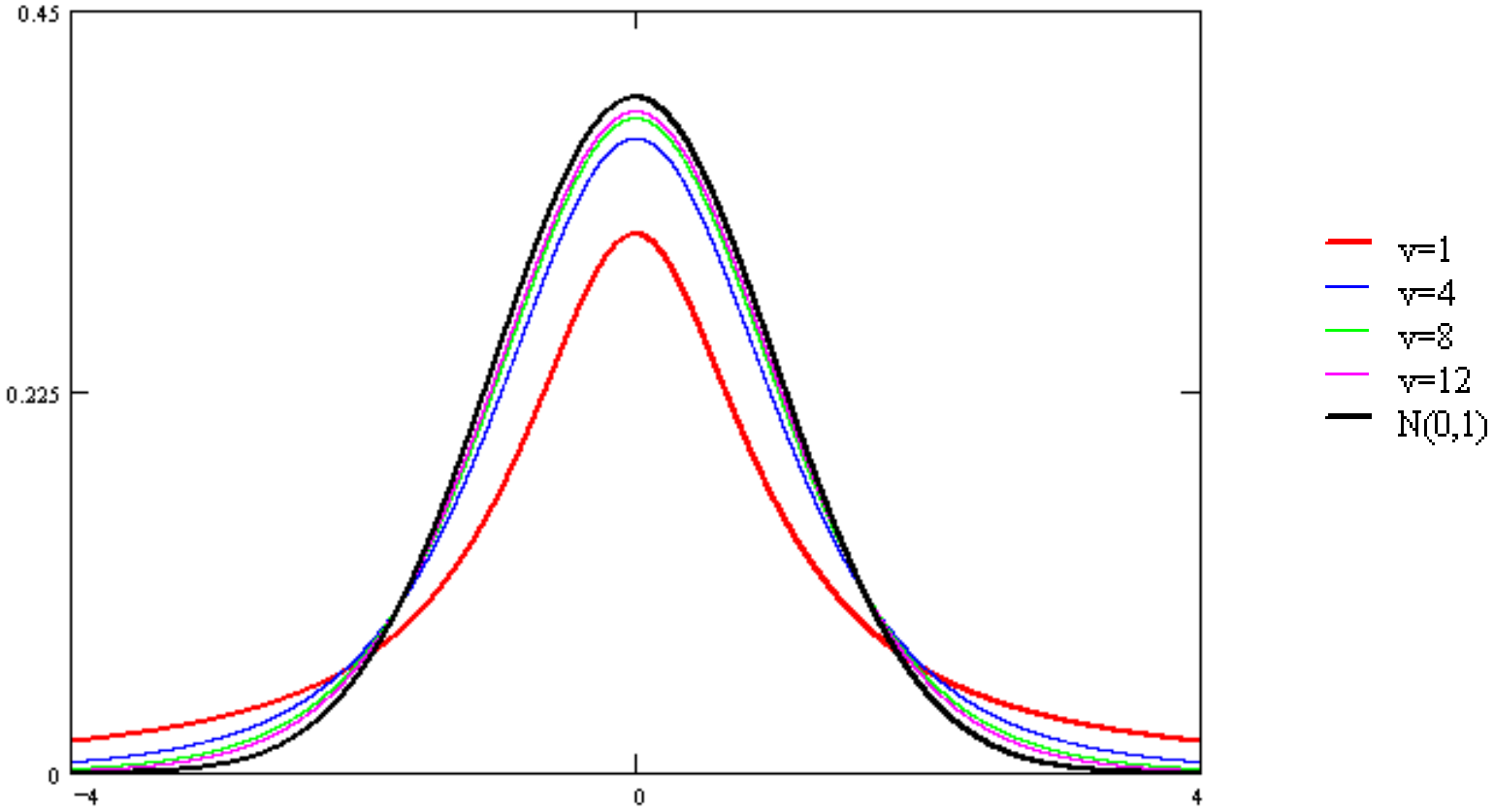
- Envolvem noções de dispersão!

Média com Distribuição Normal



T-Student

- Se a **variável tem distribuição Normal na população**, ou a amostra é suficientemente grande (>30) , mas não conhecemos o desvio da população, só da amostra, então ...
- ... A média amostral se distribui conforme uma **t-Student**
- **...** A distribuição t-Student depende dos graus de liberdade (n-1), que denotamos por **v**



Teoria de distribuições

Erro Padrão

- Usado para estimar o intervalo de confiança da média amostral
- S = Desvio Padrão Amostral
- N = Tamanho da Amostra

$$se = \frac{s}{\sqrt{n}}$$

Intervalo de Confiança

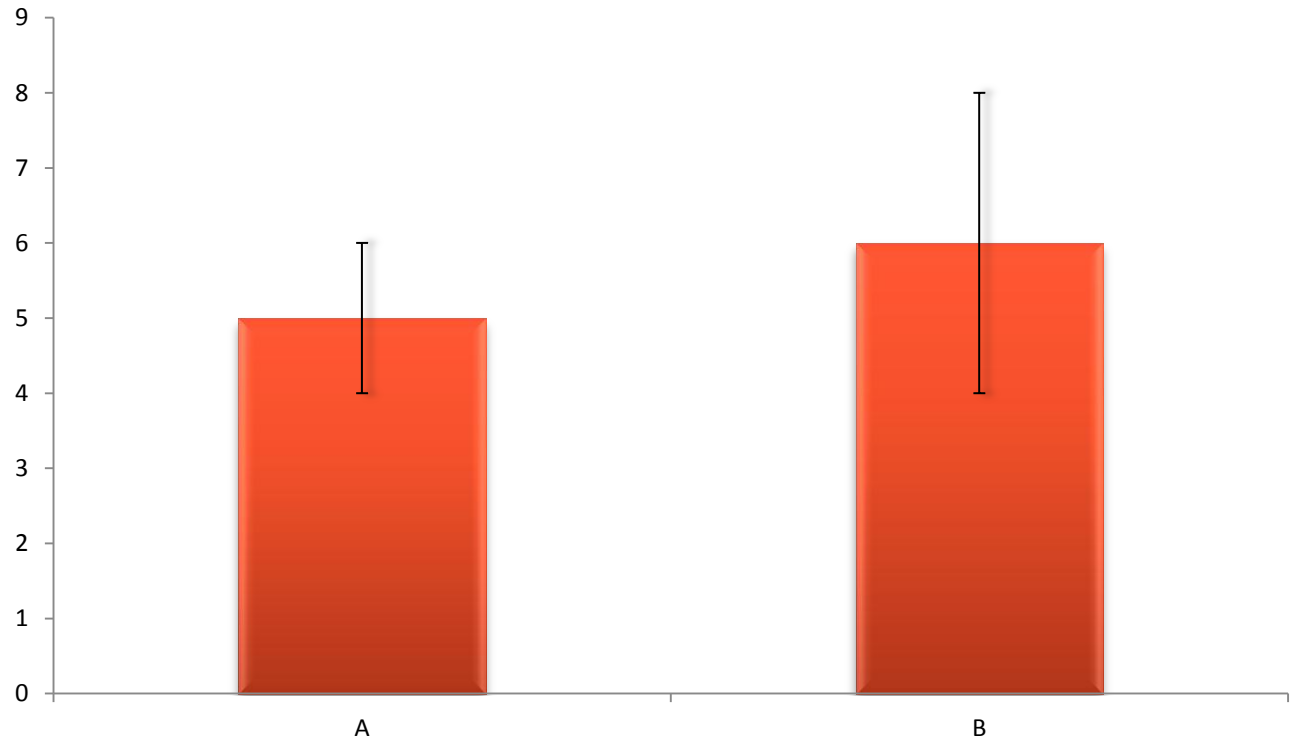
Normal (95% de confiança):
- Presupõe que conhecemos o desvio populacional
- Intervalo: $\bar{x} \pm 1{,}96\sigma / \sqrt{n}$

T-Student (95% de confiança):
- Sabemos que a variável na população tem distribuição Normal
- Mas só temos informação da amostra
- Intervalo: $\bar{x} \pm t_{\nu} se$

- t_v é um valor tabelado, em geral próximo de 2.

Na prática

- Usamos o erro padrão para termos uma noção gráfica de erro associada a uma amostra



CUIDADO: Na literatura podem aparecer gráficos com barras contendo desvio padrão ou erro padrão !!!

TESTES DE HIPÓTESES

Hipótese?

Hipótese é uma afirmação

Em uma certa população as médias de tempo de recuperação dos indíduos que tomam um certo remédio e daqueles que não tomam são iguais

O teste da hipótese é uma pergunta

Será que em uma certa população as médias de tempo de recuperação dos indíduos que tomam um certo remédio e daqueles que não tomam são iguais?

O resultado do teste é uma resposta

Com base na amostra pode-se dizer que
- Não há indícios de diferença estatisticamente significativa no tempo médio de recuperação entre os que tomam e os que não tomam o remédio, ou
- Há indícios de diferença estatisticamente significativa no tempo médio de recuperação entre os que tomam e os que não tomam o remédio

Possíveis respostas – Erros tipo I e II

ACEITAR A HIPÓTESE QUANDO ELA É VERDADEIRA

ERRO TIPO I

REJEITAR A HIPÓTESE QUANDO ELA É VERDADEIRA

ERRO TIPO II

ACEITAR A HIPÓTESE QUANDO ELA É FALSA

REJEITAR A HIPÓTESE QUANDO ELA É FALSA

Nível Descritivo (p-valor)

P-valor

- É a probabilidade de se obter o efeito observado, dado que a hipótese é verdadeira
- Ele nos fornece uma **medida de se podemos rejeitar ou não a hipótese proposta**

Na prática

- P-valor < 5%: Então com 95% de confiança estatística dizemos que podemos rejeitar a hipótese
- P-valor > 5%: Então não há evidências estatísticamente significativas que nos levem a rejeitar a hipótese

Hipóteses e parâmetros

Hipótese principal ou nula:
Média (ou mediana) dos grupos A e B são iguais

Hipótese alternativa:
Média (ou mediana) dos grupos A e B são diferentes

Testes de Hipóteses:
Que teste usar? Depende de características dos seus dados

Paramétrico:
Conheço informações de distribuição da variável na população

Não Paramétrico:
Não conheço informações de distribuição da variável na população

Unicaudal ou Bicaudal

Bicaudal

Hipótese principal ou nula:
Média (ou mediana) dos grupos A e B são iguais

Hipótese alternativa:
Média (ou mediana) dos grupos A e B são diferentes

Unicaudal

Hipótese principal ou nula:
Média (ou mediana) dos grupos A e B são iguais

Hipótese alternativa:
Média (ou mediana) do grupo A é maior que do B
ou Média (ou mediana) do grupo A é menor que do B

Paramétrico vs Não Paramétrico

- Inferência é quando tomamos decisões para a população com base em uma amostra

- Ela é **paramétrica** quando conhecemos a distribuição da população.
  - Na prática isso significa dizer qua a variável tem distribuição Normal na população

- Ela é **não paramétrica** quando não temos informações sobre a distribuição da variável na população.
  - Na prática isso não conhecemos a distribuição na população

Paramétrico vs Não Paramétrico

Paramétrico

- Distribuição da variável na população conhecida (ex: Normal)
- Estudos anteriores com amostras grandes revelam normalidade
- Amostras grandes (~> 20)
- Comparo médias e/ou variâncias
- Mais poderoso (*)

Não Paramétrico

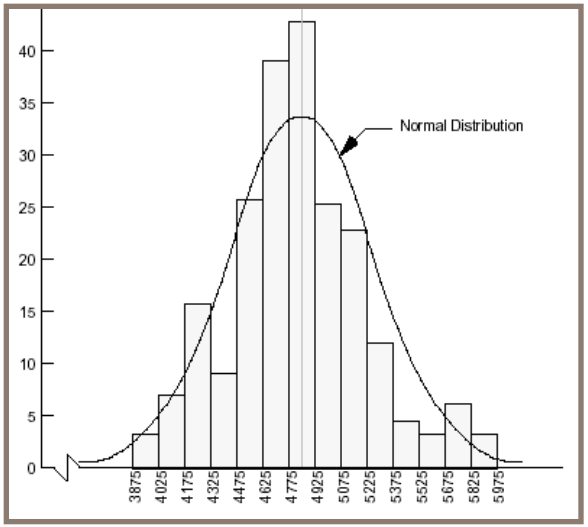
- Não conheço a distribuição da variável na população
- Amostras pequenas
- Comparo medianas e/ou distribuições
- Utiliza postos (ranks)
- Menos poderoso (*)

(*) Poder: Habilidade do teste de detectar um efeito dado que ele realmente exista

Teste de normalidade

Gráficos

- Histograma
- QQ-Normal-Plot



Testes

- Shapiro–Wilk
- Anderson–Darling
- D'Agostino
- Kolmogorov-Smirnov

p-valor < 5%	Não Normal
p-valor > 5%	Normal

Tipos de testes de hipóteses

Frequências
Variáveis Qualitativas

Teste de Homogeniedade

(1 categoria)

- Teste Qui-Quadrado
- Teste Exato de Fisher

Teste de associação ou independência

(2 ou mais categorias)

- Teste Qui-Quadrado
- Testes exatos binomiais

Relação
Variáveis Quantitativas

Correlação

- Pearson
- Spearman

Regressões

- Paramétricas
- Não paramétricas

Diferenças

Testes de diferenças

TESTES DE DIFERENÇAS PARAMÉTRICOS

T-Student - 2 grupos independentes

Teste T-Student : Comparação de médias !

Há três possíveis testes:

- ❑ Tamanhos das amostras iguais, variâncias iguais
- ❑ Tamanhos das amostras diferentes, variâncias iguais
- ❑ Tamanhos das amostras diferentes, variâncias diferentes

Variâncias Iguais?

• Exige que seja feito um teste de comparação de variâncias

• Teste-F (Excel) – Caso p-valor < 5% mostra indícios de que as variâncias são diferentes:

A	B
1	10
2	20
3	3
4	4
5	5
6	6
7	9

DP A	2.160
DP B	5.815

Test-F	2.95%

=FTEST(E4:E10;F4:F10)

T-Student - 2 grupos independentes

Caso Variâncias Iguais (homocedástico):

Usar = ttest(var1, var2, 2, 2)

Caso Variâncias Diferentes (heterocedástico):

Usar = ttest(var1, var2, 2, 3)

Caudas

Tipo

A	B
1	10
2	20
3	3
4	4
5	5
6	6
7	9

DP A	2.160
DP B	5.815

Test-F	2.95%

T-Test	=10;2;3)

TTEST

Array1 E4:E10 = {1;2;3;4;5;6;7}

Array2 F4:F10 = {10;20;3;4;5;6;9}

Tails 2 = 2

Type 3 = 3

= 0.117043657

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

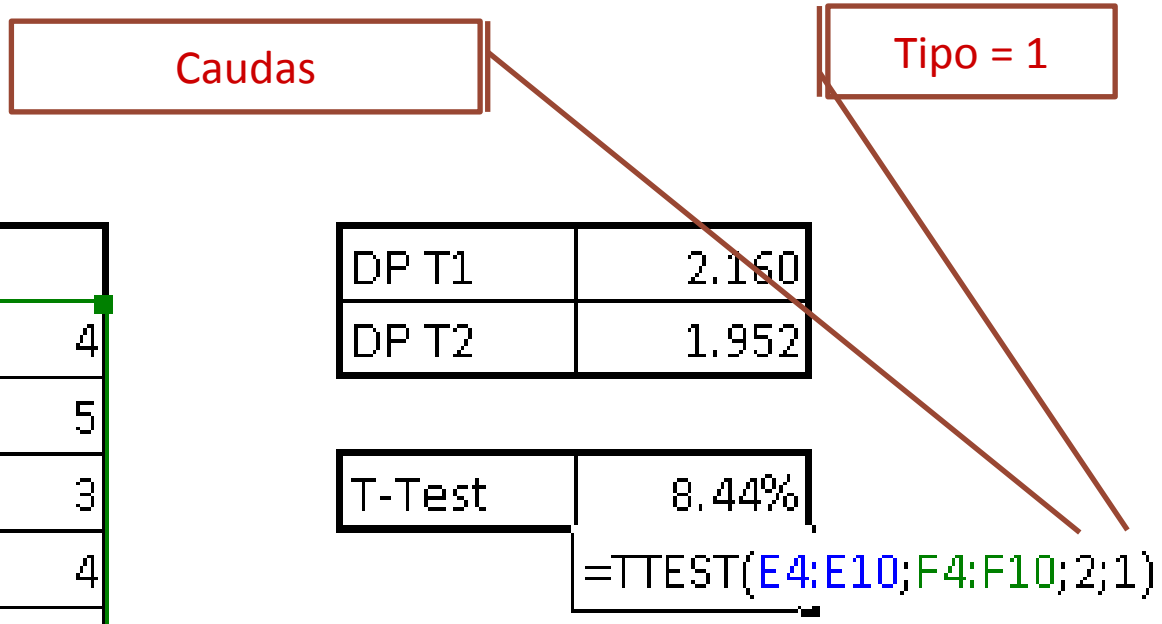
Formula result = 11.70%

T-Student - 2 grupos dependentes

Test T-Student Pareado: Comparação de médias

Amostras necessariamente com o mesmo tamanho !!!!!
É equivalente a subtrair um grupo do outro e testar se a média é zero

Excel : use tipo 1



T0	T1
1	4
2	5
3	3
4	4
5	5
6	6
7	9

ANOVA – 1 Fator – Amostras Independentes

Anova 1 fator – amostras independentes: Comparação de médias

- Comparação de diversos grupos independentes
- Assume que as variâncias são aproximadamente iguais
- Dados retirados de uma mesma população
- Distribuição normal da variável na população
- Dados na mesma escala

Hipótese:

Médias dos grupos são iguais

Alternativa:

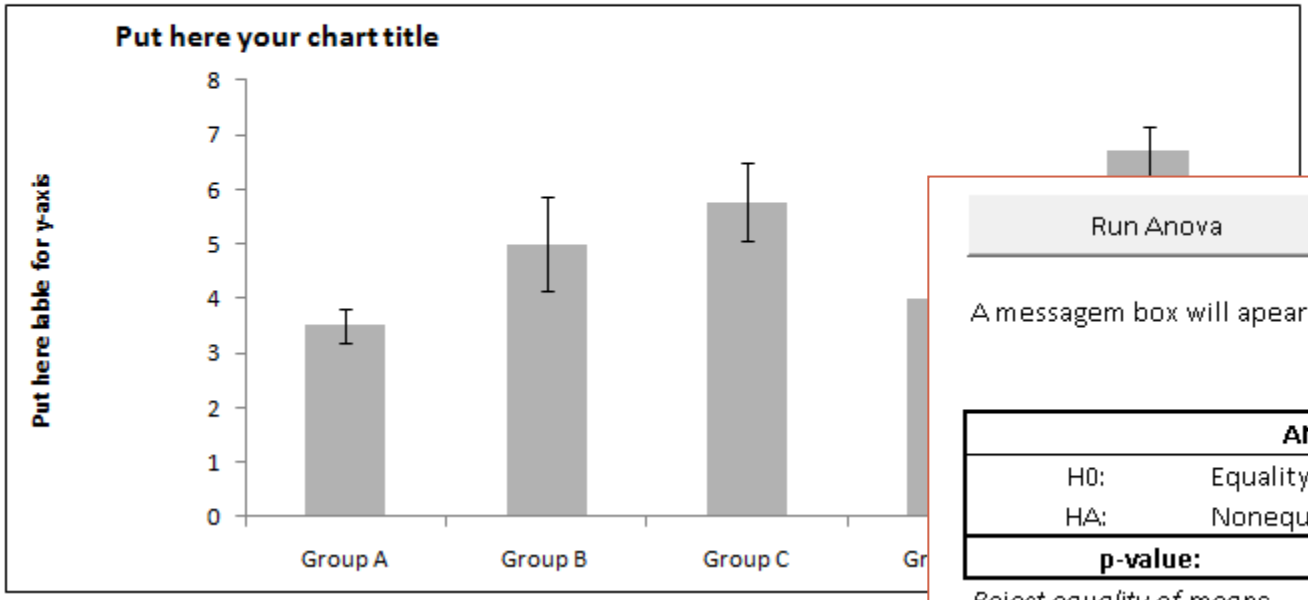
Médias dos grupos não são iguais

Group A	Group B	Group C	Group D	Group E
3.00	1.00	3.00	1.00	8.00
4.17	4.00	4.00	2.00	7.00
3.00	7.00	5.00	3.00	6.00
3.00	8.00	3.00	4.00	7.00
6.00	7.00	6.00	5.00	8.00
5.00	2.00	7.00	6.00	9.00
4.00	3.00	8.00	7.00	7.00
3.00	1.00	9.00		6.00
3.00	7.00	7.00		5.00
3.00	3.00			4.00
2.00	8.00			
3.00	9.00			

Anova - EXSTAT

Update Graphs

Averages and standard errors bar chart:



Run Anova

A messagem box will apear: please press OK

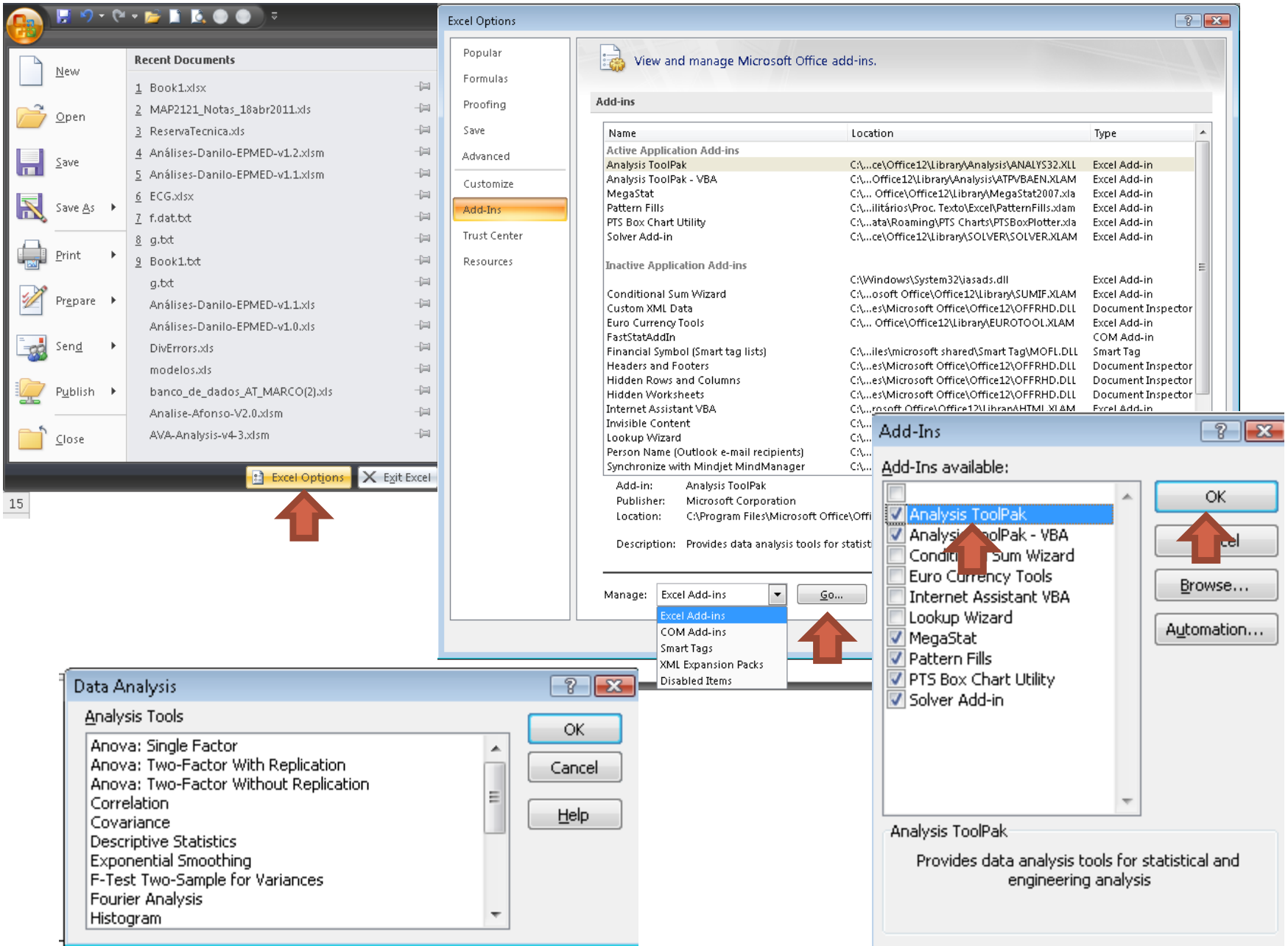
ANOVA		
H0:	Equality of Means	
HA:	Nonequality of Means	
p-value:		**0.821%**

Reject equality of means

T-Test (Heteroscedastic)		
Group A	vs	Group C
H0:	Equality of Means	
HA:	Nonequality of Means	
p-value:		**0.756%**

Reject equality of means

Ferramentas de Estatística no Excel – Analysis Toolpack



ANOVA – 1 Fator – Amostras correlacionadas

Anova 1 fator – amostras correlacionadas: Comparação de médias

- Comparação de diversos grupos dependentes !!!
- Assume que as variâncias são iguais e normalidade
- Amostras com o **mesmo tamanho** !!!
- Conhecida como Anova de 1 fator para medidas repetidas
- Conhecida também como ANOVA de 2 fatores sem replicação

Hipótese:

Médias dos tempos/medidas são iguais

Alternativa:

Médias dos temposmedidas não são iguais

T0	T1	T2	T3
3.00	2.00	3.00	1.00
4.17	3.00	4.00	2.00
3.00	4.00	5.00	3.00
3.00	5.00	3.00	4.00
6.00	4.00	6.00	5.00
5.00	3.00	7.00	6.00
4.00	4.00	8.00	7.00
2.00	5.00	4.00	5.00

Anova – Medidas Repetidas - Online

VassarStats: Website for Statistical Computation

Setup

Number of samples in analysis = 4

Independent Samples	Correlated Samples k=4
Correlated Samples	standard weighted-means analysis
Unweighted	Click this button only if you wish to perform an unweighted-means analysis. Advice: do not perform an unweighted-means analysis unless you have a clear reason for doing so.
Weighted	Click this button to return to a standard weighted-means analysis

Data Entry

Sample 1	Sample 2
3.00 4.17 3.00 3.00 6.00 5.00 4.00 2.00	2.00 3.00 4.00 5.00 4.00 3.00 4.00 5.00

Reset Calculate

standard weighted-means analysis

ANOVA Summary Correlated Samples k=4

Source	SS	df	MS	F	P
Treatment [between groups]	8.2083	3	2.7361	1.72	0.193518
Error	33.3184	21	1.5866		
Ss/Bl	38.6669	7			
Total	80.1936	31			

Graph Maker

SOFTWARES

Software

Excel

- **Não permite o cálculo de estatísticas não paramétricas diretamente**
  - Software comercial, mas presente na grande maioria das máquinas
- ADD-INS
  - EXSTAT (só alguns testes, www.ime.usp.br/~pedrosp)
  - MEGASTAT (completo, gratuíto, http://highered.mcgraw-hill.com/sites/0070983755/student_view0/megastat.html)

Área de Biológicas

- **BioEstat (Livre) – Simples mas completo**
  - **www.mamiraua.org.br/download/**
- BioStat (comercial) - Completo
  - www.analystsoft.com/br/products/biostat/
- Sigmaplot +SigmaStat (comercial) – Completo
  - www.sigmaplot.com
  - Inclui o Sigmastat a partir da versão 12
- Graphpad Prism (comercial)
  - http://www.graphpad.com/prism/Prism.htm

Testes Online!

http://faculty.vassar.edu/lowry/VassarStats.html

- *VassarStats: Website for Statistical Computation NY*
- Completo e com texto explicativo de cada teste
- Lowry, R. 2011. VassarStats: Web Site for Statistical Computation. [Online]. Available at: http://faculty.vassar.edu/lowry/VassarStats.html [May 02, 2011].

http://www.fon.hum.uva.nl/Service/Statistics.html

- IFA services (Institute of Phonetic Sciences Amesterdam)
- Menos completo, mas muito indicado para testes não paramétricos

Concepts & Applications of Inferential Statistics
Richard Lowry - http://faculty.vassar.edu/lowry/webtext.html

Concepts & Applications of Inferential Statistics

- Front Page
- Table of Contents
- Online Statistical Computation

Table of Contents

PROJETO

Projeto : Parte 2 – Inferência Paramétrica

Utilize a variável com 2 ou mais grupos independentes obtido anteriormente.

Execute o teste paramétrico apropriado para verificar diferença estatística entre as médias dos grupos

Interprete o resultado e analise-o juntamente com os gráficos obtidos anteriormentes

Complemente o relatório (Word) com as novas análises:

- Estatísticas e gráficos

- Interpretação das estatísticas

INFERÊNCIA II – TESTES NÃO PARAMÉTRICOS

Estatística Aplicada à Biotecnologia

jun-12

TESTES DE HIPÓTESES

Tipos de testes de hipóteses

Frequências
Variáveis Qualitativas

Teste de Homogeniedade

(1 categoria)

- Teste Qui-Quadrado
- Teste Exato de Fisher

Teste de associação ou independência

(2 ou mais categorias)

- Teste Qui-Quadrado
- Testes exatos binomiais

Relação
Variáveis Quantitativas

Correlação

- Pearson
- Spearman

Regressões

- Paramétricas
- Não paramétricas

Diferenças

Testes de diferenças

Paramétrico vs Não Paramétrico

Paramétrico

- Distribuição da variável na população conhecida (ex: Normal)
- Estudos anteriores com amostras grandes revelam normalidade
- Amostras grandes (~> 20)
- Comparo médias e/ou variâncias
- Mais poderoso (*)

Não Paramétrico

- Não conheço a distribuição da variável na população
- Amostras pequenas
- Comparo medianas e/ou distribuições
- Utiliza postos (ranks)
- Menos poderoso (*)

(*) Poder: Habilidade do teste de detectar um efeito dado que ele realmente exista

TESTES DE DIFERENÇAS NÃO PARAMÉTRICOS

Postos (ranks)

Valor	Posto
63,5	3
70,4	2
55,1	5
79,8	1
60,0	4
40,7	6

Usa a teoria conhecida para a distribuição de postos para o teste

Consequências

- Não leva em conta a distância entre os valores, só a ordem!
- Geralmente testa-se igualdade de medianas
- Se há um valor muito discrepante dos demais isso não afeta o teste
- Há perda de informações

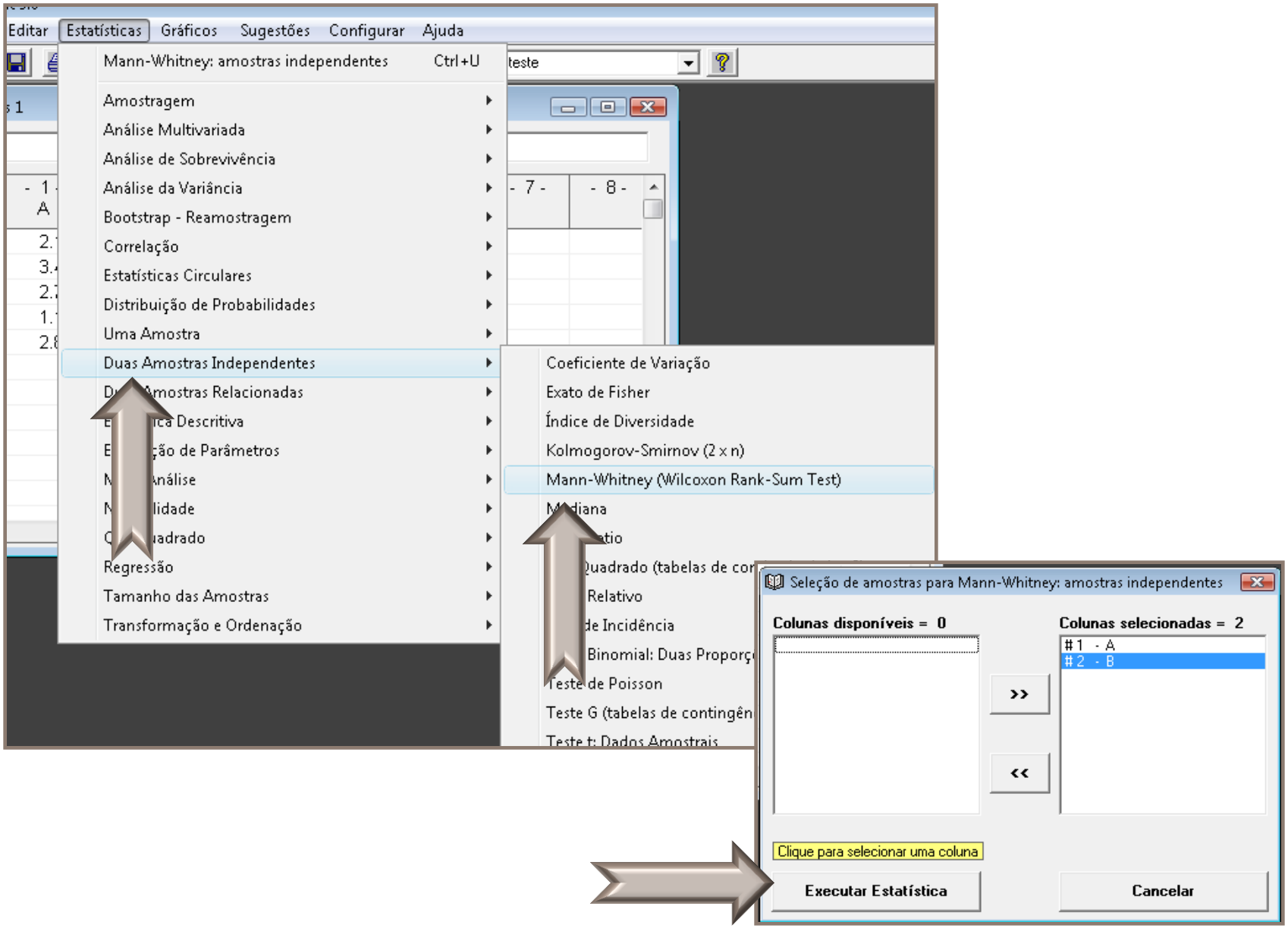
Características do Mann-Whitney

- Teste de soma de postos de Wilcoxon (W. rank-sum test)
- 2 Grupos independentes
- Hipótese: As distribuições dos grupos são iguais, ou
- Hipótese: As medianas dos grupos são iguais
- Insensível a outliers
- Os grupos não precisam ter o mesmo tamanho

Cuidados

- Se a distribuição for normal é melhor usar o teste t-Student para grupos independentes
- Precisa ter 4 ou mais elementos na amostra de cada grupo

Mann Whitney no BioEstat

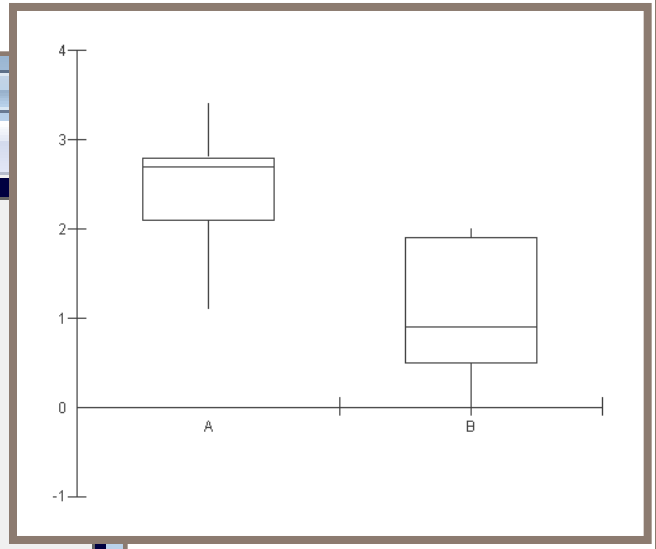


Resultados do Mann Whitney no BioEstat

Mann-Whitney: amostras independentes

Arquivo Editar Gráfico

Resultado	Amostra 1	Amostra 2
Tamanho da amostra	5	5
Soma dos Postos (Ri)	38.0	17.0
Mediana =	2.70	0.90
U =	2.00	
Z(U) =	2.1934	
p-valor (unilateral) =	0.0141	
p-valor (bilateral) =	0.0283	



Hipótese: Mediana da Amostra 1 = Mediana da Amostra 2
Alternativa: Mediana da Amostra 1 > Mediana da Amostra 2

Resultado: P-valor 1,41% -> Rejeito a hipótese com 95% de confiança

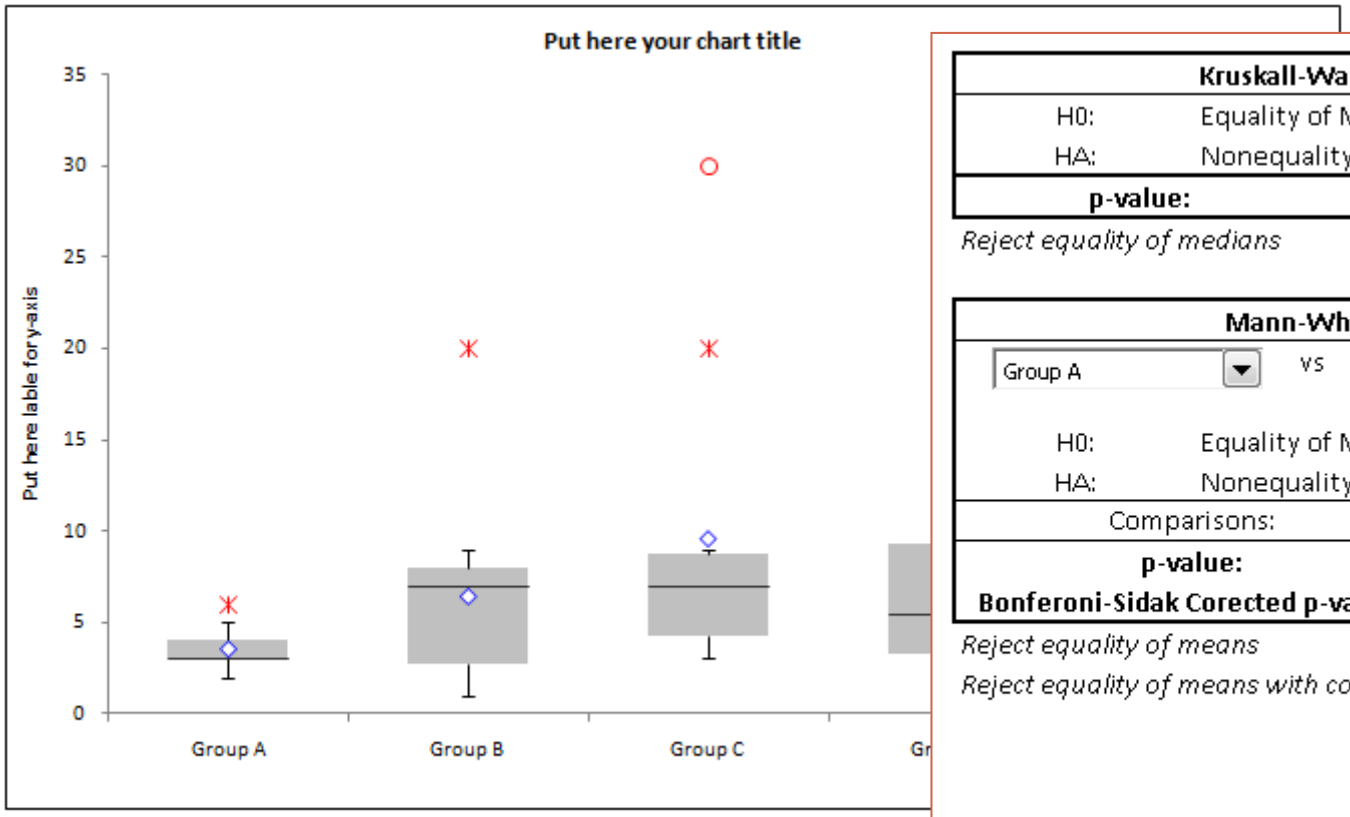
Características do Kruskal-Wallis

- Análise de variância não paramétrica
- 3 ou + grupos independentes
- Hipótese: As distribuições de todos os grupos são iguais,
- Hipótese: As medianas de todos os grupos são iguais
- Insensível a outliers
- Os grupos não precisam ter o mesmo tamanho

Cuidados

- Se a distribuição for normal é melhor usar o teste ANOVA de um critério (one-way)
- Precisa ter 4 ou mais elementos na amostra de cada grupo
- Se tiver só 2 grupos use o Mann-Whitney

Kruskal-Wallis no Template de Excel - EXSTAT



Kruskall-Wallis Test	
H0:	Equality of Medians
HA:	Nonequality of medians
p-value:	**3.379%**

Reject equality of medians

Mann-Whitney		
Group A	vs	Group C
H0:	Equality of Means	
HA:	Nonequality of Means	
Comparisons:	10	
p-value:	**0.476%**	
Bonferoni-Sidak Corected p-value:	**4.661%**	

Reject equality of means

Reject equality of means with corrected p-value

P-valor < 5% → Rejeito hipótese de igualdade entre as medianas

Mas quais são diferentes entre si?
Uma forma é usar Mann-Whitney para saber

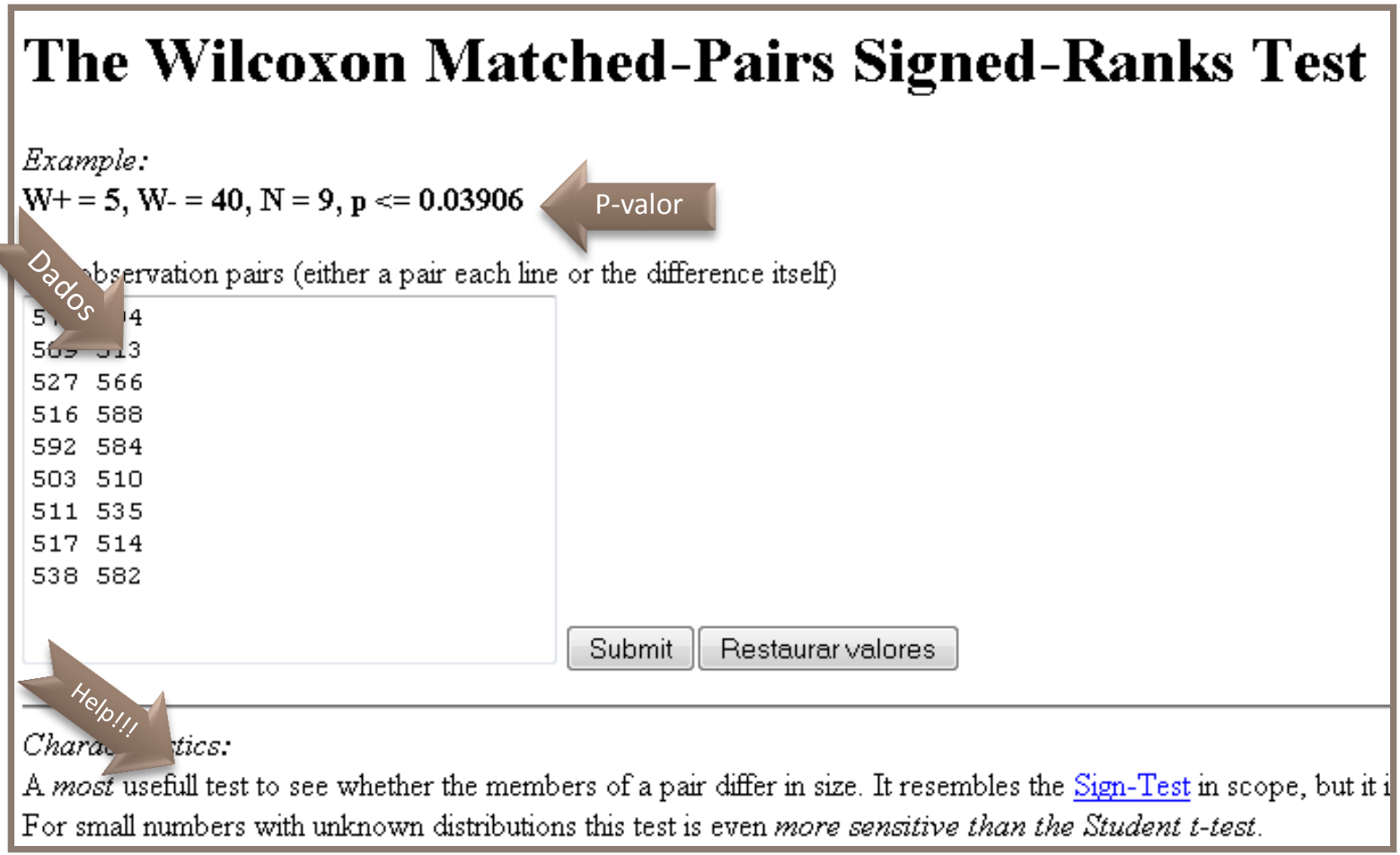
Características do Wilcoxon Pareado

- Teste de Sinais de Postos de Wilcoxon
- 2 grupos dependentes, pareados
- Hipótese: As distribuições dos grupos são iguais,
- Hipótese: As medianas dos grupos são iguais
- Insensível a outliers
- Os grupos precisam ter o mesmo tamanho

Cuidados

- Se a distribuição for normal é melhor usar o teste t-Student pareado
- Precisa ter 4 ou mais elementos na amostra de cada grupo
- Não confundir com o Wilcoxon Soma de Postos!

Wilcoxon Pareado - IFA services (Institute of Phonetic Sciences Amesterdam)



Online: http://www.fon.hum.uva.nl/Service/Statistics.html

Teste **Bicaudal** → Dividir p-valor por 2 para obter unicaudal

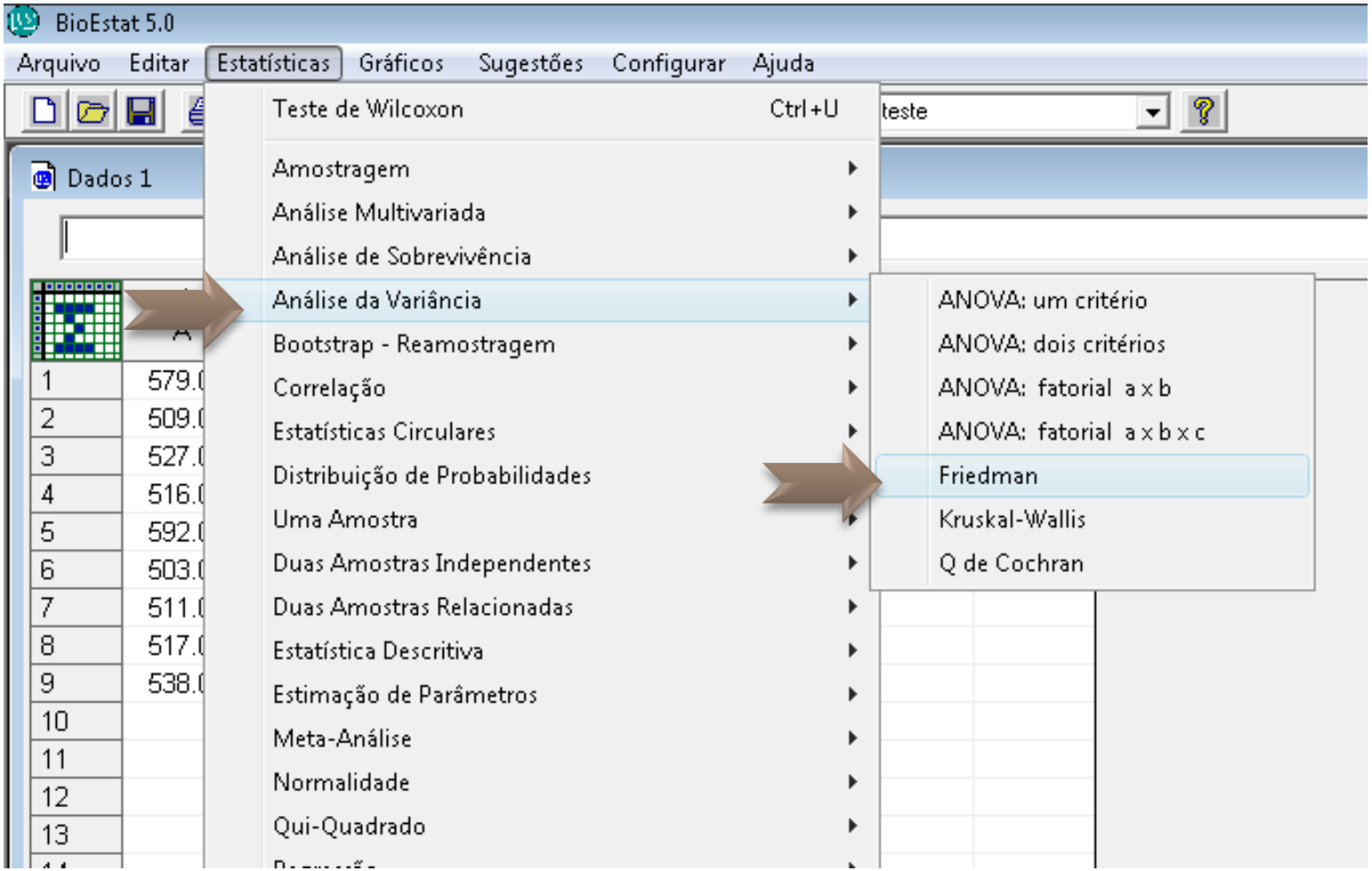
Características do Friedman

- Análise de variância não paramétrica
- 3 ou + grupos dependentes
- Hipótese: As distribuições de todos os grupos são iguais,
- Hipótese: As medianas de todos os grupos são iguais
- Insensível a outliers
- Os grupos **precisam** ter o mesmo tamanho

Cuidados

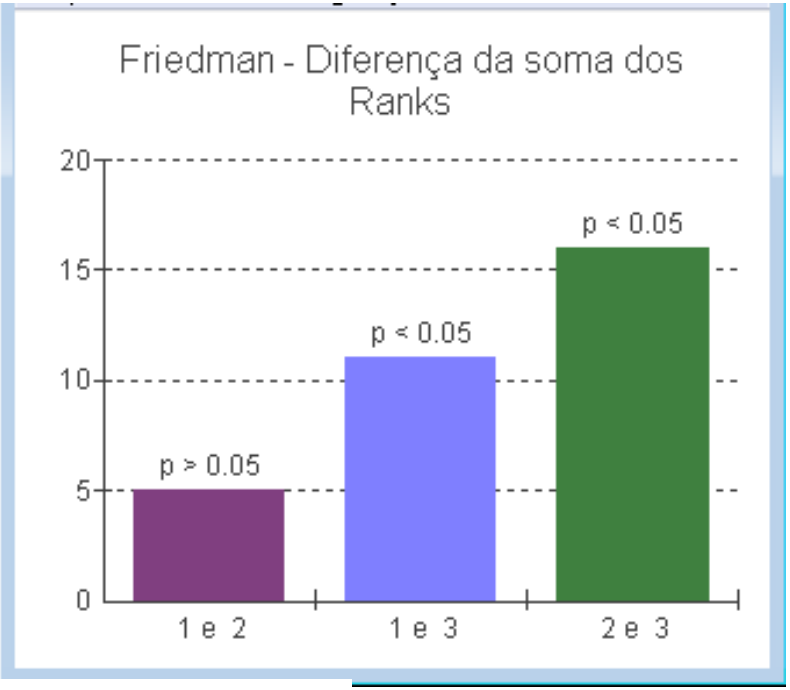
- Se a distribuição for normal é melhor usar o teste ANOVA para medidas repetidas
- Precisa ter 4 ou mais elementos na amostra de cada grupo
- Se tiver só 2 grupos use o Wilcoxon

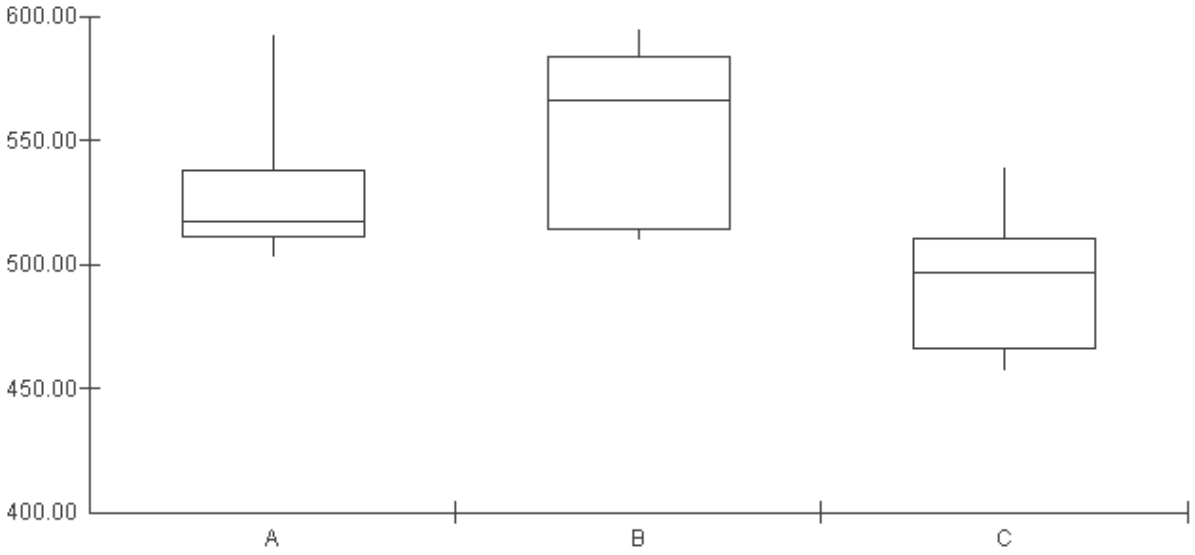
Friedman no BioEstat



Friedman – Resultados do BioEstat

	- 1 -	- 2 -	- 3 -
Soma dos Ranks =	20.0000	25.0000	9.0000
Mediana =	517.0000	566.0000	496.8227
Média dos Ranks =	2.2222	2.7778	1.0000
Média dos valores =	532.4444	554.0000	493.4979
Desvio padrão =	31.9379	35.6406	31.2195
Friedman (Fr) =	14.8889		
Graus de liberdade =	2		
(p) =	0.0006		
Comparações:	Diferença	(p)	
Ranks 1 e 2 =	5	ns	
Ranks 1 e 3 =	11	< 0.05	
Ranks 2 e 3 =	16	< 0.05	





Faz também as comparações 2 a 2 !!!

GRÁFICOS

Gráficos em testes de diferenças

CATEGORIAS NÃO CORRELACIONADAS

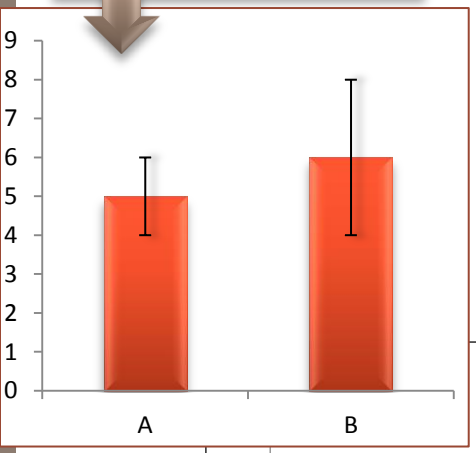
CATEGORIAS CORRELACIONADAS

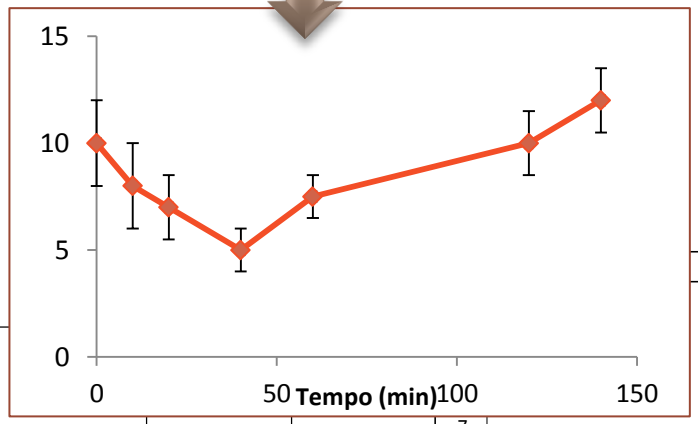
Paramétrico: Barras com erros

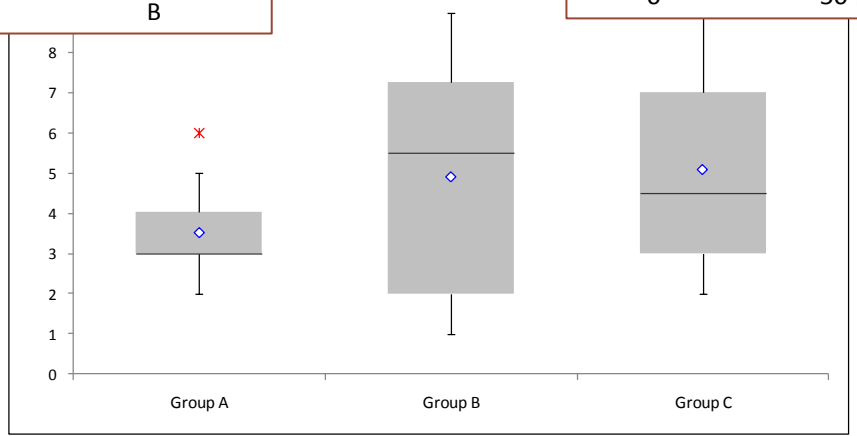
Não Paramétrico: Boxplot

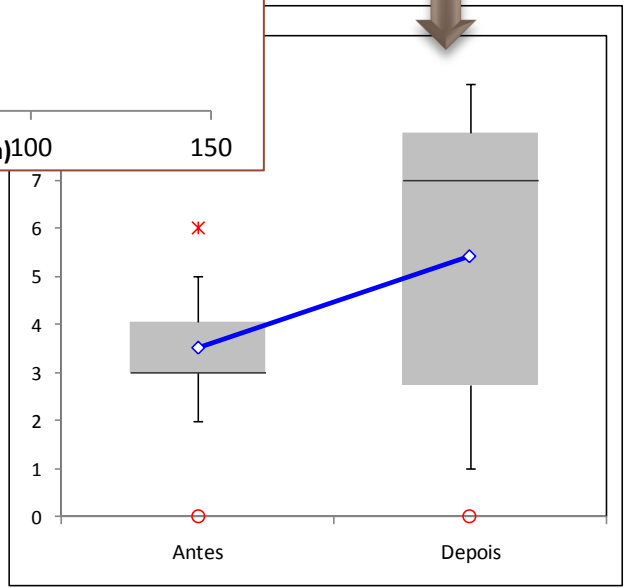
Paramétrico: Linha com erros

Não Paramétrico: Boxplot com linha









PROJETO

Projeto : Parte 3 – Inferência Não Paramétrica

Utilize a variável (apenas 1 fator) com 2 ou mais grupos independentes/dependentes obtido anteriormente.

Execute o teste apropriado para verificar diferença estatística entre as medianas dos grupos (não paramétrico)

Interprete o resultado e analise-o juntamente com os resultados obtidos anteriormente

Complemente o Relatório (Word):

- Descrição dos dados

- Tabela com os dados

- Estatísticas e gráficos

- Interpretação das estatísticas

Relatório

-Introdução :

-Discorra brevemente sobre o que se tratam os dados incluindo a justificativa de relação de dependência entre as categorias

-Metodologia:

-Quais método você vai usar para analisar os dados e por quê?

-Quais programas vai usar?

-Resultados:

-Análise descritiva (média, mediana, desvio,...)

-Gráficos de barras/linhas e boxplot

-Comparação de médias (paramétrico)

-Comparação de medianas (não paramétrico)

-Discussão/Conclusões

-Que tipo de efeito observou ?

-Era o esperado?

-Anexo/apêndice: tabela contendo os dados

Escrevendo

- Em METODOLOGIA:
  - ✓Destacar TODA a metodologia estatística usada e o motivo.
  - ✓EXEMPLOS:
    - •Como não conhecemos as distribuições de probabilidade das variáveis na população, foram usados testes não paramétricos para comparações de variáveis quantitativas, e boxplots para as representações gráficas (... Destacar quais e quando usou...)
    - •Consideramos um nível de significância de 95% para os testes de hipótese (...)

- Em RESULTADOS:
  - ✓Análise descritiva da amostra
  - ✓Interpretar os dados e as estatísticas obtidas.
  - ✓EXEMPLO:
    - • Observa-se que há uma diferença estatisticamente significativa (p-valor < 5%) nas medianas dos níveis de glicose quando consideramos o grupo controle em relação ao de teste, indicando níveis maiores no grupo de teste. Isso pode ser observado na figura 4.3, onde apresentamos o boxplot que refere-se ao teste, (...).

- DICA: Usem como referência outros artigos/teses (de qualidade)

Projeto

Entrega

- Por e-mail: pedrospeixoto@yahoo.com.br
- Mande com o assunto: Projeto de Estatística - Butantan
- Data: Até 15/7
- Entregue o relatório (.doc, .docx, .pdf) e o arquivo contendo as análises em Excel.

Avaliação

- 1/5 - Organização
- 1/5 - Análise Descrita e gráficos
- 1/5 - Testes de comparações paramétricos
- 1/5 - Testes de comparações não paramétricos
- 1/5 - Interpretações e análises

Software

- Use o que achar mais adequado para o seu perfil, sugestões:
  - BioEstat
  - Softwares Online
  - Excel + Templates Excel (MegaStat, EXSTAT)

Observações importantes

- Utilize no projeto dados de apenas 1 fator. Podem ser com grupos independentes, ou dependentes.
- O fator pode ter 2 ou mais categorias. Fique atento as escolhas dos testes pertinentes (paramétricos e não paramétricos)
- Caso a sua base de dados tenha 2 fatores escolha 1 para trabalhar, de preferência o de grupos independentes.
- Caso queira trabalhar com 2 fatores, faça a ANOVA apropriada (vamos estudar a seguir). Esta parte não será levada em conta na avaliação.
- Caso não tenha uma amostra com essas caracteristicas tente obter uma com colegas os artigos. Caso mesmo assim não consiga, pode simular os dados, assim temos o efeito didatico.

INFERÊNCIA III - ANOVAS

Estatística Aplicada à Biotecnologia

jun-12

ANOVA

Testes de diferenças

ANOVA

ANOVA – Análise de Variância (ANalysis Of VAriance)

Existem diversos tipos!!!
É usado para análises globais, com diversos grupos simultaneamente.

Já vistos:
• ANOVA de 1 fator para grupos independentes
•ANOVA de 1 fator para medidas repetidas (dados correlacionados)

A ser visto:
•ANOVA de 2 fatores para amostras independentes
•ANOVA de 2 fatores com um deles para dados correlacionados
•ANOVA de 2 fatores com com ambos fatores com dados correlacionados

Importante!!

Análises do tipo ANOVA são geralmente paramétricas,
logo pressupõe normalidade
Além disso, em geral assume se que as variâncias são iguais

ANOVA 2 fatores

Como o nome diz envolve 2 fatores.

Exemplo:

2 fatores com amostras independentes

	Serpente A	Serpente B	Serpente C
Controle	3	5	7
	2	5	8
	3	6	7
	4	5	8
	5	7	7
Trat 1	5	6	9
	4	6	7
	3	7	9
	2	7	6
	6	6	7
Trat 2	5	7	13
	4	7	11
	3	8	11
	3	7	12
	3	9	10

Médias	Serpente A	Serpente B	Serpente C
Controle	3.4	5.6	7.4
Trat 1	4	6.4	7.6
Trat 2	3.6	7.6	11.4
Erro Padrão	Serpente A	Serpente B	Serpente C
Controle	0.51	0.40	0.24
Trat 1	0.71	0.24	0.60
Trat 2	0.40	0.40	0.51

ANOVA 2 fatores

ANOVA 2 fatores com replicação (Data Analysis Excel)

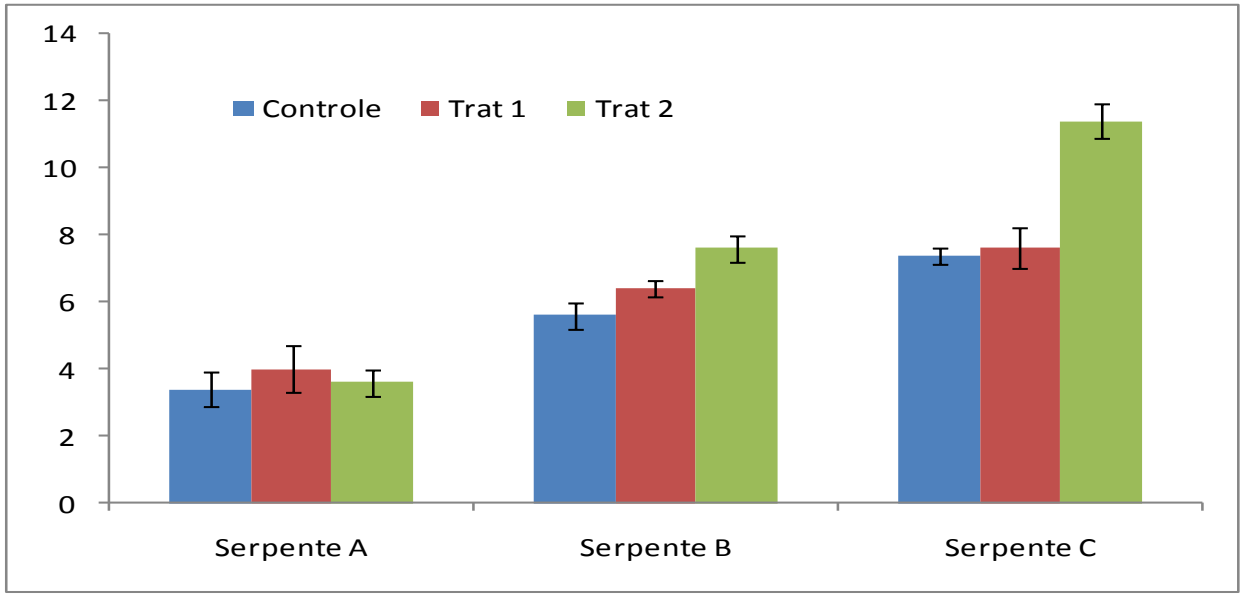
Fator das linhas (Tratamento)

Fator das colunas (Serpentes)

Interação (Tratamento x Serpentes)

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	34.53	2	17.27	15.70	0.001%	3.26
Columns	198.53	2	99.27	90.24	0.000%	3.26
Interaction	27.33	4	6.83	6.21	0.066%	2.63
Within	39.60	36	1.10			
Total	300.00	44.00				



ANOVA 2 fatores

2 fatores com um dos fatores com dados correlacionados

- Tipo de Tratamento e Estágios (Tempo)
  - Fator 1: Controle/Trat.A/Trat.B
  - Fator 2: Logo após procedimento/1dia depois/ 2dias depois
- Lago de coleta e Profundidade
  - Fator 1: LagoA/ LagoB
  - Fator 2: 0m/ 1m/ 5m
- Tratamento e diluição
  - Fator 1: Controle/Trat.A/Trat.B
  - Fator 2: 1/4000, 1/8000, 1/16000

2 fatores com um ambos fatores com dados correlacionados

- Periodo e Diluição
  - Fator 1: Antes/Depois
  - Fator 2: 1/4000, 1/8000, 1/16000
- Época e profundidade
  - Fator 1: Verão/Inverno
  - Fator 2: 0m/ 1m/ 5m

ANOVA Online

http://faculty.vassar.edu/lowry/vsanova.html

Analysis of Variance

For non-parametric alternatives to the one-way ANOVAs for independent and correlated samples, see the Kruskal-Wallis Test and the Friedman Test under 'Ordinal Data.'

One-Way ANOVA for up to five samples.
[Traducción en español]
The design can be either for independent samples or correlated samples (repeated measures or randomized blocks). This unit will also perform pair-wise comparisons of sample means via the Tukey HSD test.

Two-Way Factorial ANOVA for Independent Samples, for up to four rows by four columns. This unit will also calculate the critical values of Tukey's HSD for purposes of post-ANOVA comparisons.

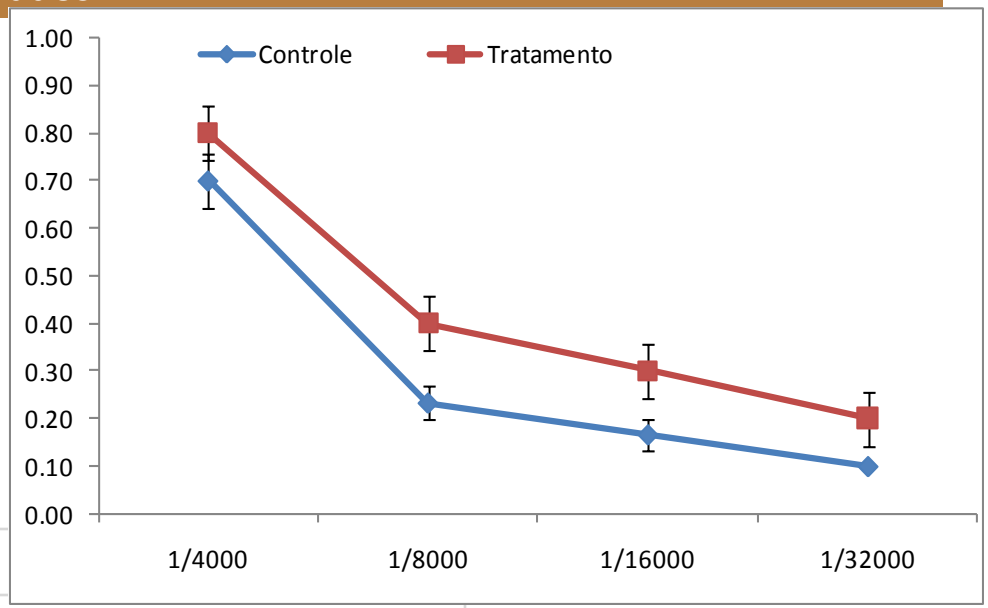
Two-Factor ANOVA with Repeated Measures on One Factor, for designs in which there are 2-4 randomized blocks of matched subjects, with 2-4 repeated measures for each subject.

Two-Factor ANOVA with Repeated Measures on Both Factors, for designs in which there are 2-4 levels of each of two variables, A and B, with each subject measured under each of the AxB combinations.

2x2x2 ANOVA for Independent Samples. For designs with three independent variables, A, B, and C, each with two levels. This situation yields 2x2x2=8 unique treatment combinations— a1b1c1, a1b1c2, and so forth— one for each of 8 independent samples of subjects.

ANOVA 2 fatores – com dados correlacionados

	Diluição			
DO	1/4000	1/8000	1/16000	1/32000
Controle	0.7	0.3	0.2	0.1
	0.8	0.2	0.2	0.1
	0.6	0.2	0.1	0.1
Tratamento	0.7	0.3	0.2	0.1
	0.9	0.5	0.4	0.3
	0.8	0.4	0.3	0.2



ANOVA Summary 2rows x 4columns

A = groups: the between-subjects variable delineated by the rows
B = the repeated-measures variable delineated by the columns

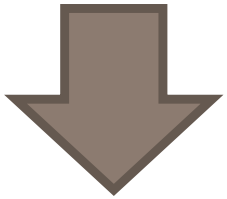
Source	SS	df	MS	F	P
Between Subjects	0.19	5			
A	0.09	1	0.09	3	0.158302
Subjects within A	0.1	4	0.03		
Within Subjects	1.31	18			
B	1.28	3	0.43	Infinity	<.0001
A x B	0	3	0	NaN	NaN
B x Subjects within A	0.03	12	0		
TOTAL	1.5	23			

Não rejeito hipótese de igualdade entre linhas (Controle/Trat)

Rejeito hipótese de igualdade entre colunas (Diluição)

ANOVA 2 fatores

ANOVA 2 Fatores com dados correlacionados:
Nem todo software faz !



Muitos estudos usam ANOVA 2 fatores para amostras independentes mesmo com dados correlacionados.
Evitem !!!!

Exemplo anterior (calculado no Excel):

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Sample	0.09	1	0.09	13.24	0.221%
Columns	1.28	3	0.43	60.45	0.000%
Interaction	0.00	3	0.00	0.22	88.407%
Within	0.11	16	0.01		

Balanceamento:

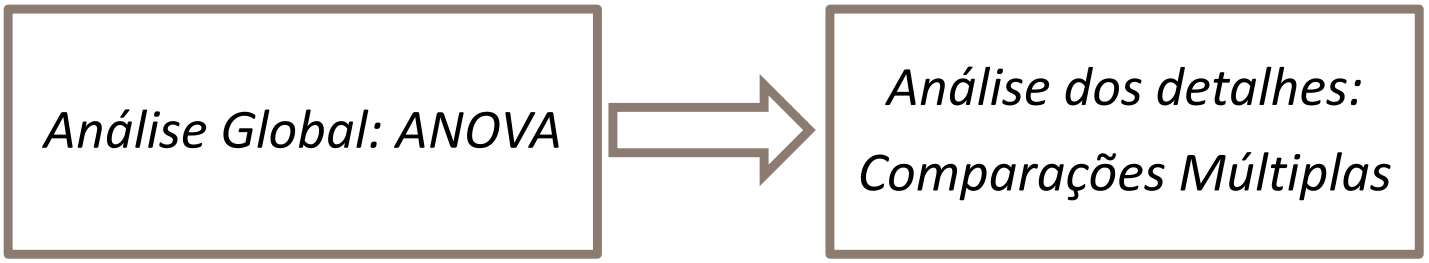
•Cada fator de linha tenha sempre o mesmo número de amostras para cada fator de coluna.

•Amostras não balanceadas exigem tratamento especial (ANOVA 2 fatores sem balanceamento) – Também conhecido como modelo linear generalizado.

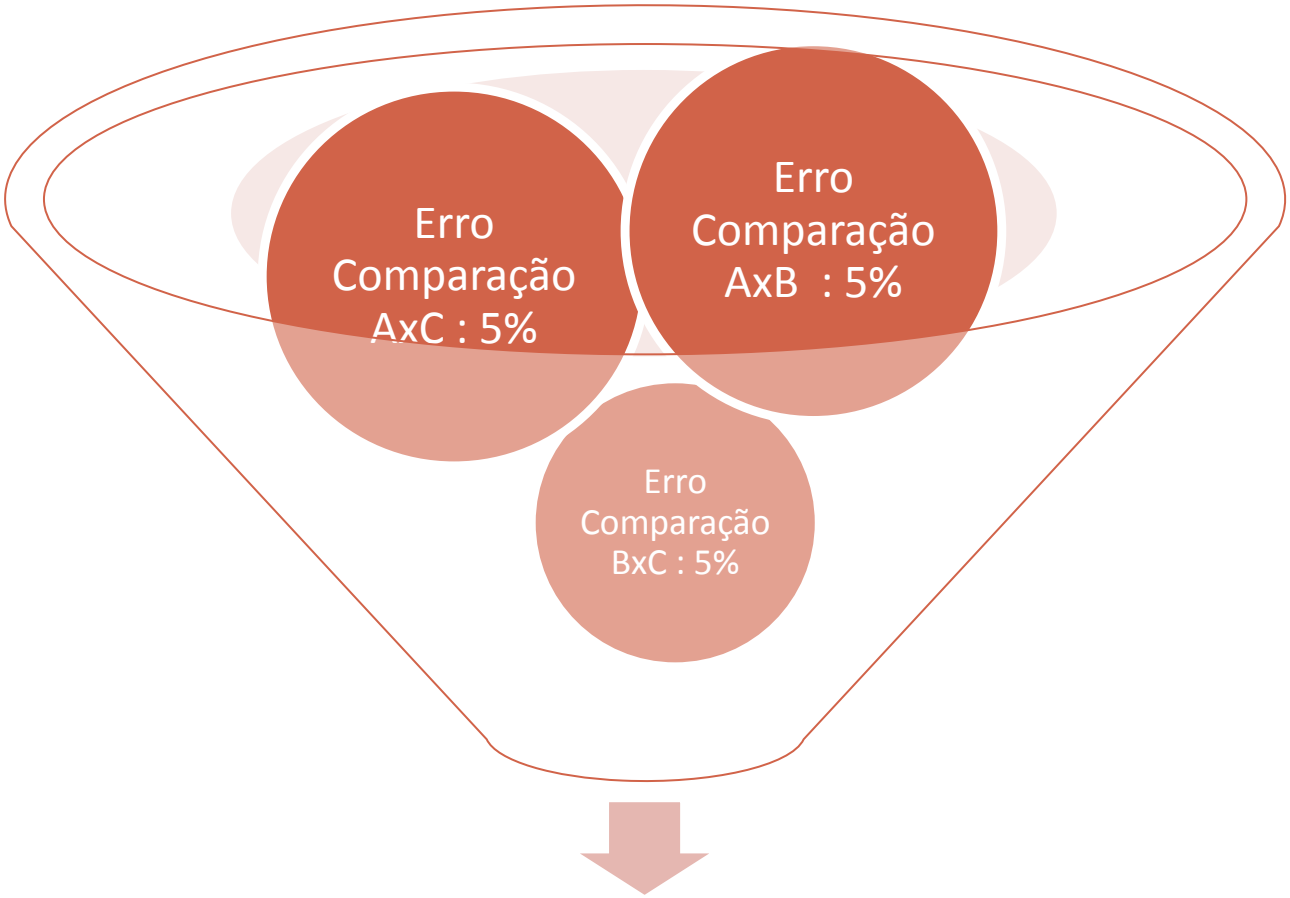
Análise dos detalhes:

•Se observamos efeito de um fator:

- •Quais das categorias/grupos se diferenciam das demais?
  - •Comparações 2 a 2**: Tukey**, Scheffé, Mann-Whitney, Teste-t, Wilcoxon, Bonferroni, ...



COMPARAÇÕES MÚLTIPLAS



Erro do Conjunto (Familywise)

100%-(95%)^3=14.2 %

Métodos de Correções

Já sei quantas comparações vou fazer e uso um nível de significância menor (<5%) ou corrijo os p-valores obtidos de testes t-student

Bonferroni	→ Muito conservador
Holm-Bonferroni	→ Mais poderoso
Dunn-Sidak	→ Assume independência
Duncan	→ Obsoleto
Fisher LSD	→ Melhor usar Holm-Sidak
Holm-Sidak	→ Em geral uma boa escolha

Testes Post Hoc

Testes realizados depois de uma ANOVA para comparações múltiplas

Teste t-Student com correções

→ Conservador, poderoso

Student-Newman-Keuls

→Detecta diferenças onde talvez não exista

Dunnet

→ Compara versus controle

Tukey (-HSD, -Kramer)

→ Uma boa escolha no caso geral

Scheffé

→ Faz comparações mais sofisticadas (não só 2 a 2)

Testes Post-Hoc para Kruskal- Wallis ?

Uso os testes anteriores para os postos !

Mann-Whitney com correções

→ Conservador, poderoso

Student-Newman-Keuls

→Detecta diferenças onde talvez não exista

Dunnet

→ Compara versus controle

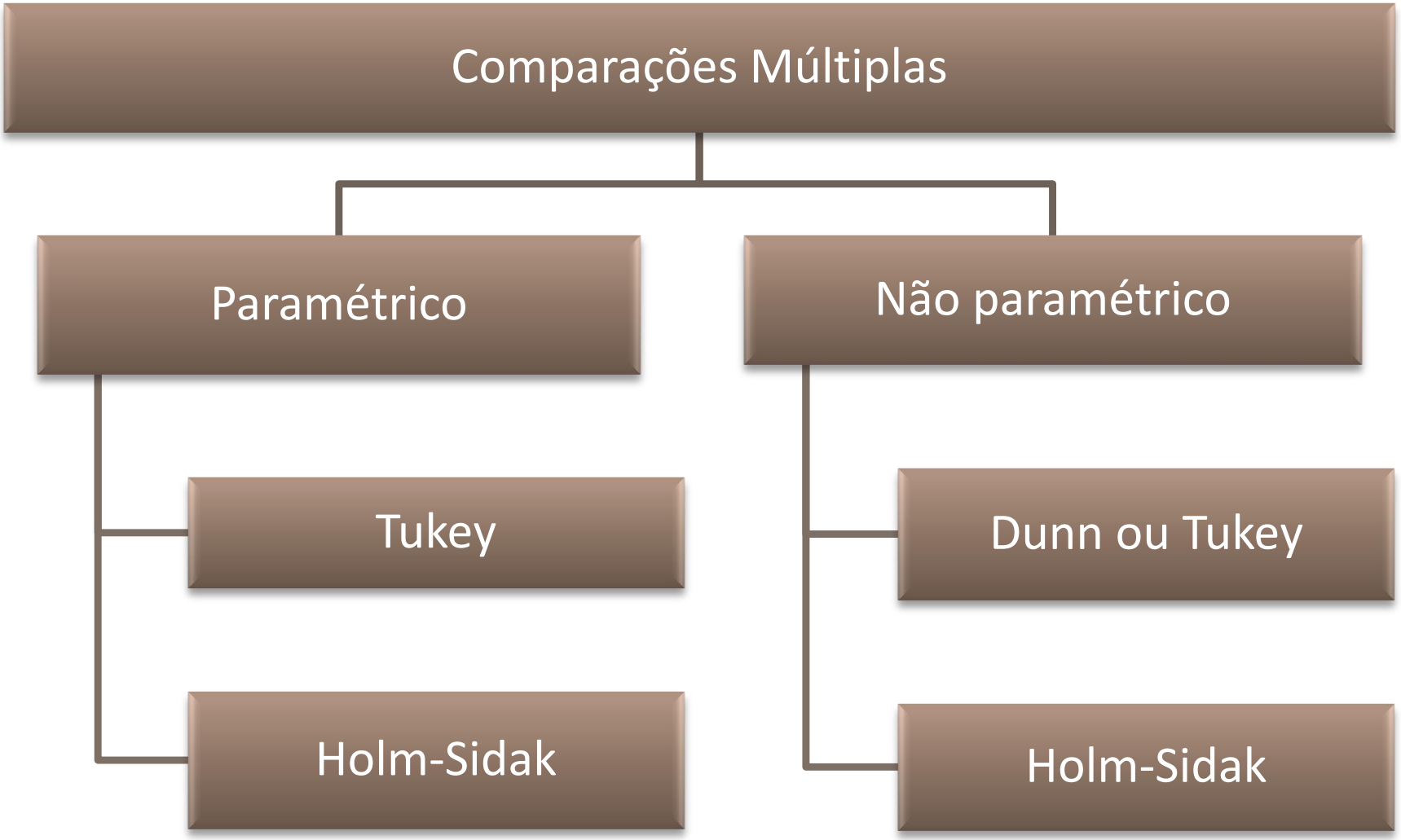
Tukey (-HSD, -Kramer)

→ Menos conservador

Dunn → Permite grupos com tamanhos diferentes

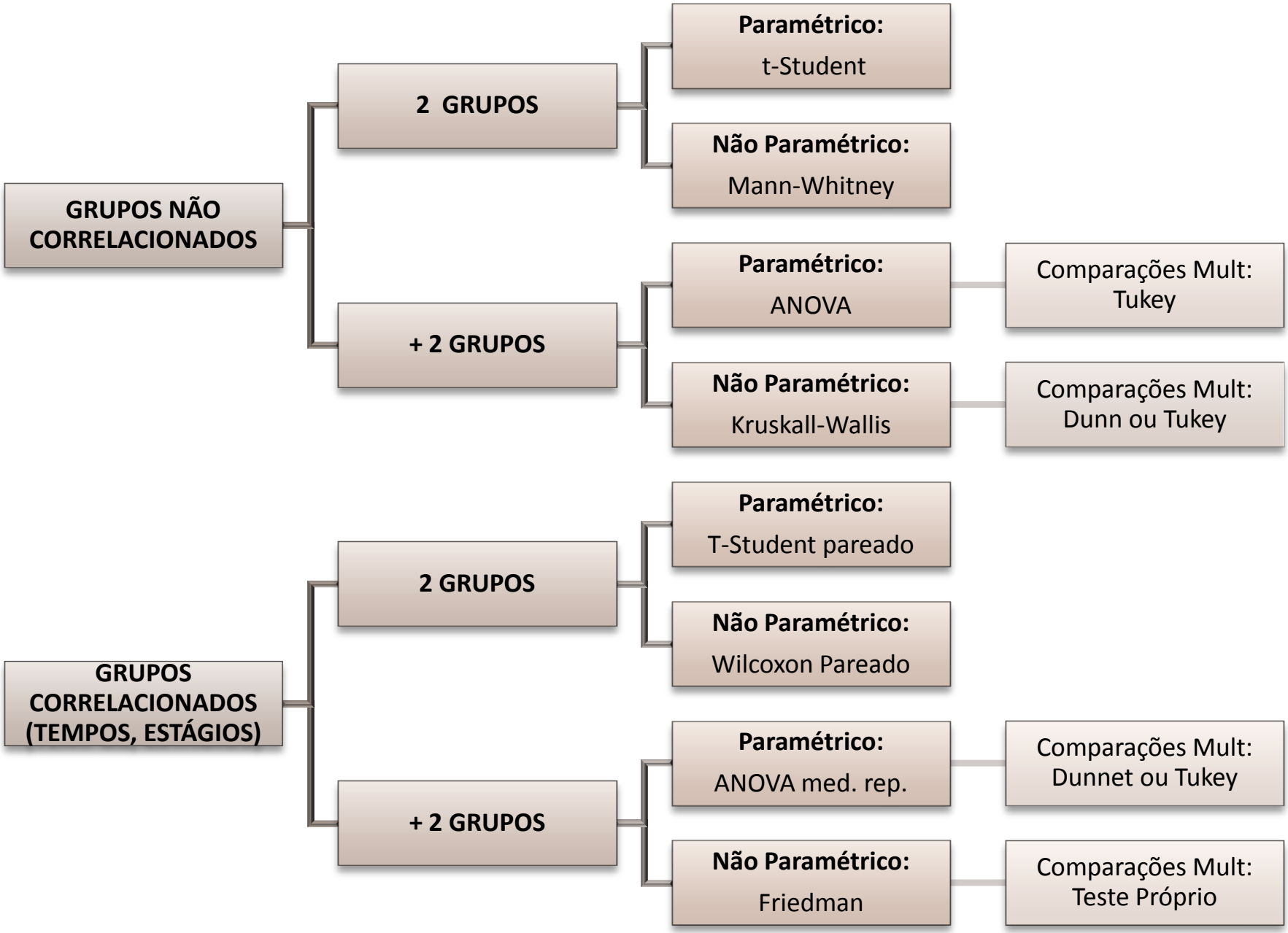
→ Indicado no caso geral

COMPARAÇÕES MÚLTIPLAS - RECOMENDAÇÕES

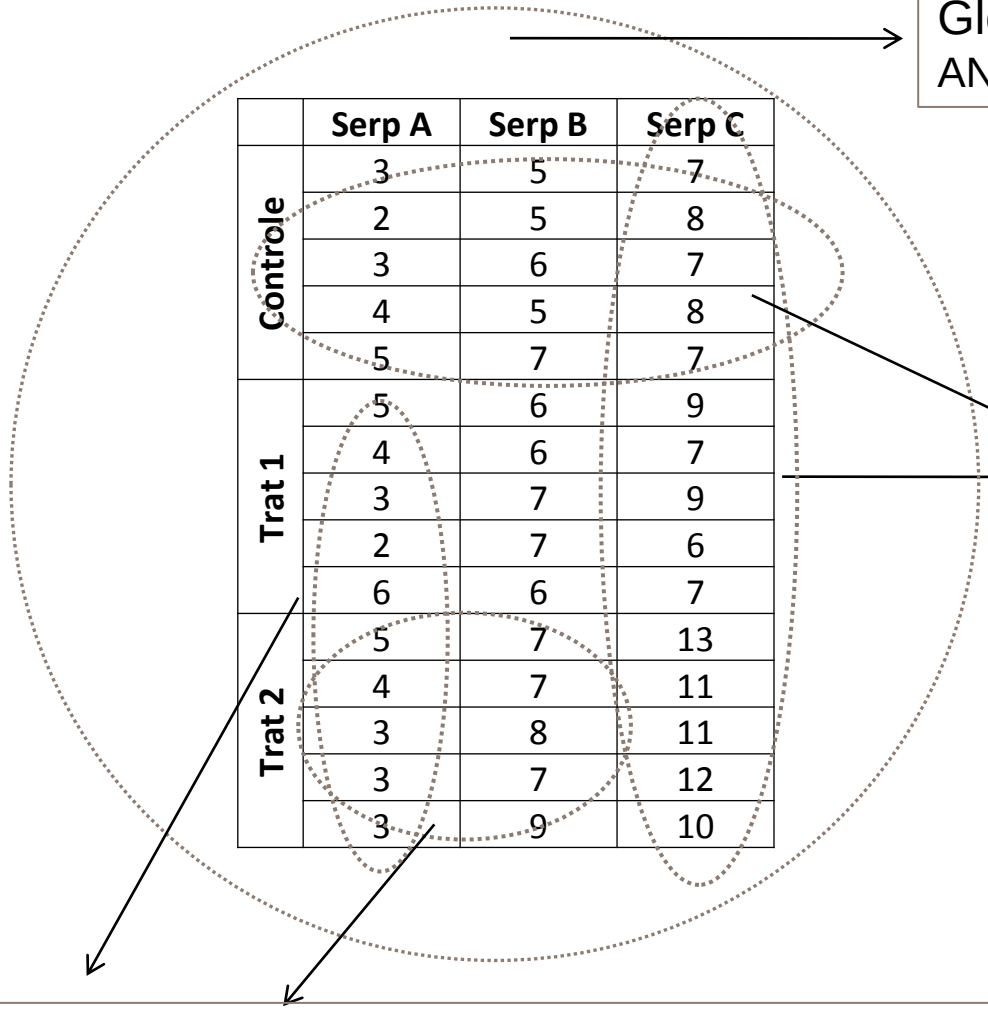


RESUMÃO

Testes de diferenças



Resumão – Grupos Independentes



	Serp A	Serp B	Serp C
Controle	3	5	7
	2	5	8
	3	6	7
	4	5	8
	5	7	7
Trat 1	5	6	9
	4	6	7
	3	7	9
	2	7	6
	6	6	7
Trat 2	5	7	13
	4	7	11
	3	8	11
	3	7	12
	3	9	10

Global:
ANOVA 2 fatores

Fixando fator de linha (ou coluna):
ANOVA 1 fator indep.
Kruskall-Wallis

Fixando Fator de Coluna e Linha (comparações 2 a 2):
Testes post hoc (Tukey, Holm-Sidak)
Teste t
Mann-Whitney

Resumão – Grupos Dependentes

Global: ANOVA 2 fatores com dados correlacionados nas colunas

	T0	T1	T2
Controle	3	5	7
	2	5	8
	3	6	7
	4	5	8
	5	7	7
Trat 1	5	6	9
	4	6	7
	3	7	9
	2	7	6
	6	6	7
Trat 2	5	7	13
	4	7	11
	3	8	11
	3	7	12
	3	9	10

Fixando Fator de Linha:
ANOVA 1 fator med. rep.
Friedman

Fixando Fator de Coluna e comparando linhas 2 a 2:
Teste t
Mann-Whitney
Testes post hoc

Fixando Fator de Linha e comparando colunas 2 a 2:
Teste t pareado
Wilcoxon
Testes post hoc

Fixando Fator de Coluna:
ANOVA 1 fator indep
Kruskall-Wallis

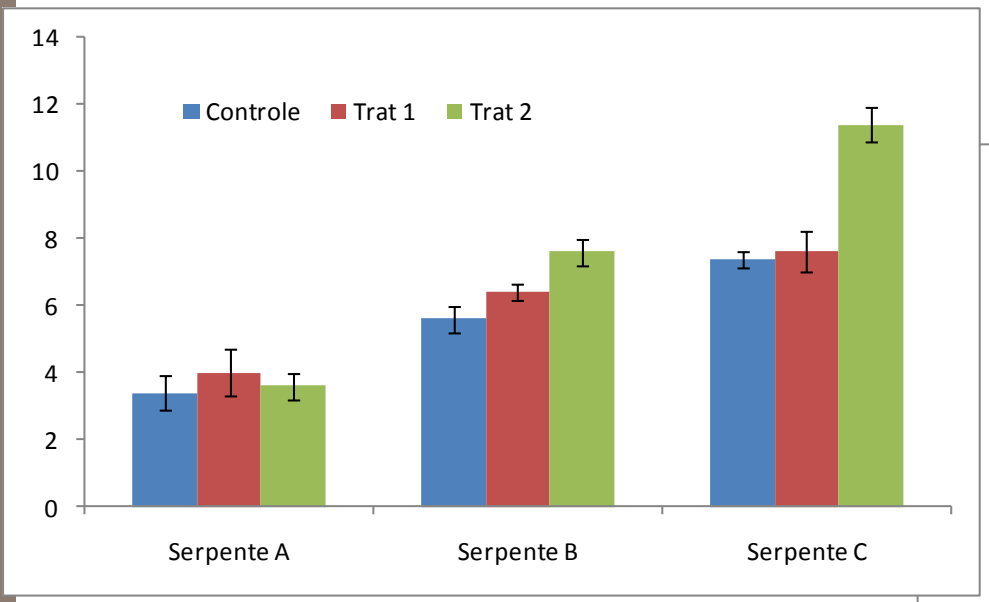
Gráficos para 2 fatores

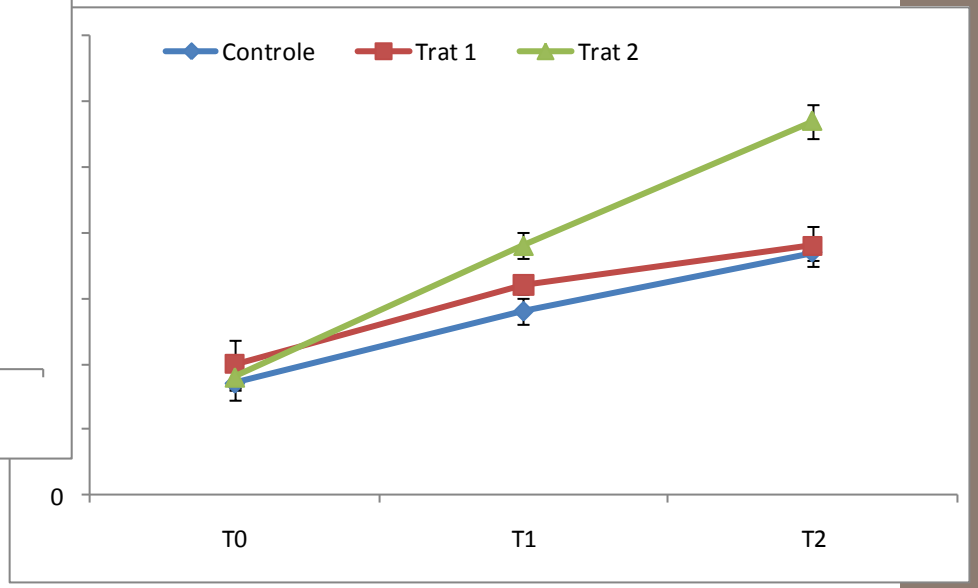
CATEGORIAS NÃO CORRELACIONADAS

Paramétrico:
Barras com erros

CATEGORIAS CORRELACIONADAS

Paramétrico:
Linhas com erros





Plus para o projeto

- Caso sua base dados tenha 2 fatores, tente usar a metodologia de ANOVA com 2 fatores
- Inclua no projeto as análises à posteriori, de comparações múltiplas, caso tenha mais de 2 grupos.
- Caso não tenha uma amostra com essas características simule um conjunto de dados para exercitar a metodologia estudada.
- Essa parte não entrará na nota do projeto

ANÁLISES DE DADOS CATEGÓRICOS E REGRESSÕES

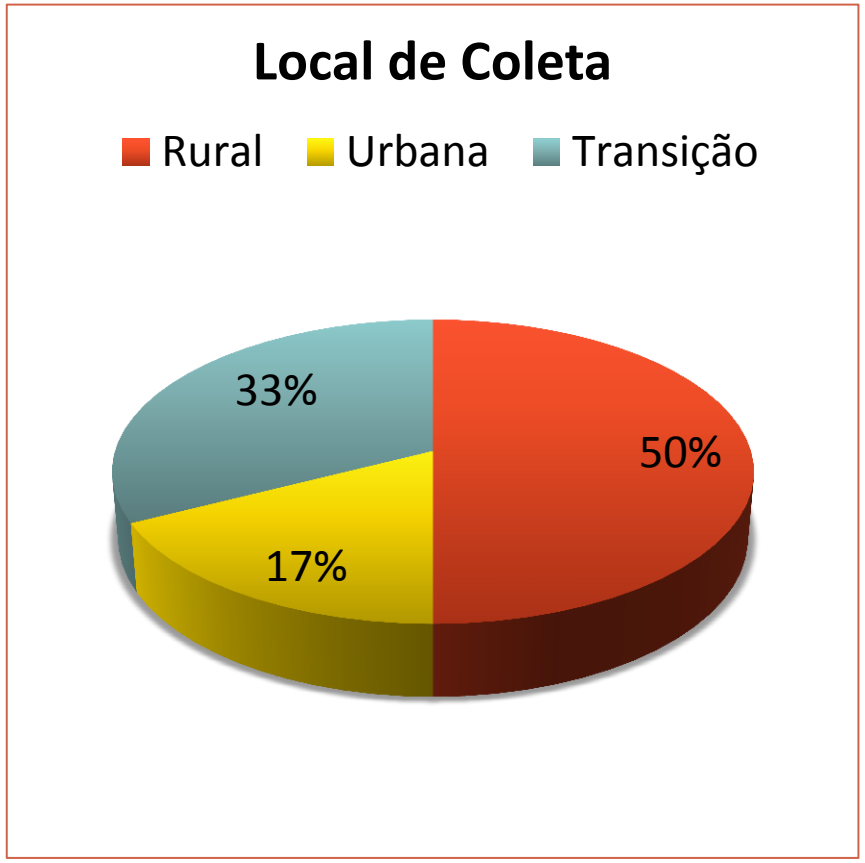
Estatística Aplicada à Biotecnologia

jun-12

ANÁLISES DE DADOS CATEGÓRICOS

Dados Nominais ou Categóricos

- Sim/Não
- Masculino/Feminino
- Pouca/Média/Muita Dor
- Escolaridade



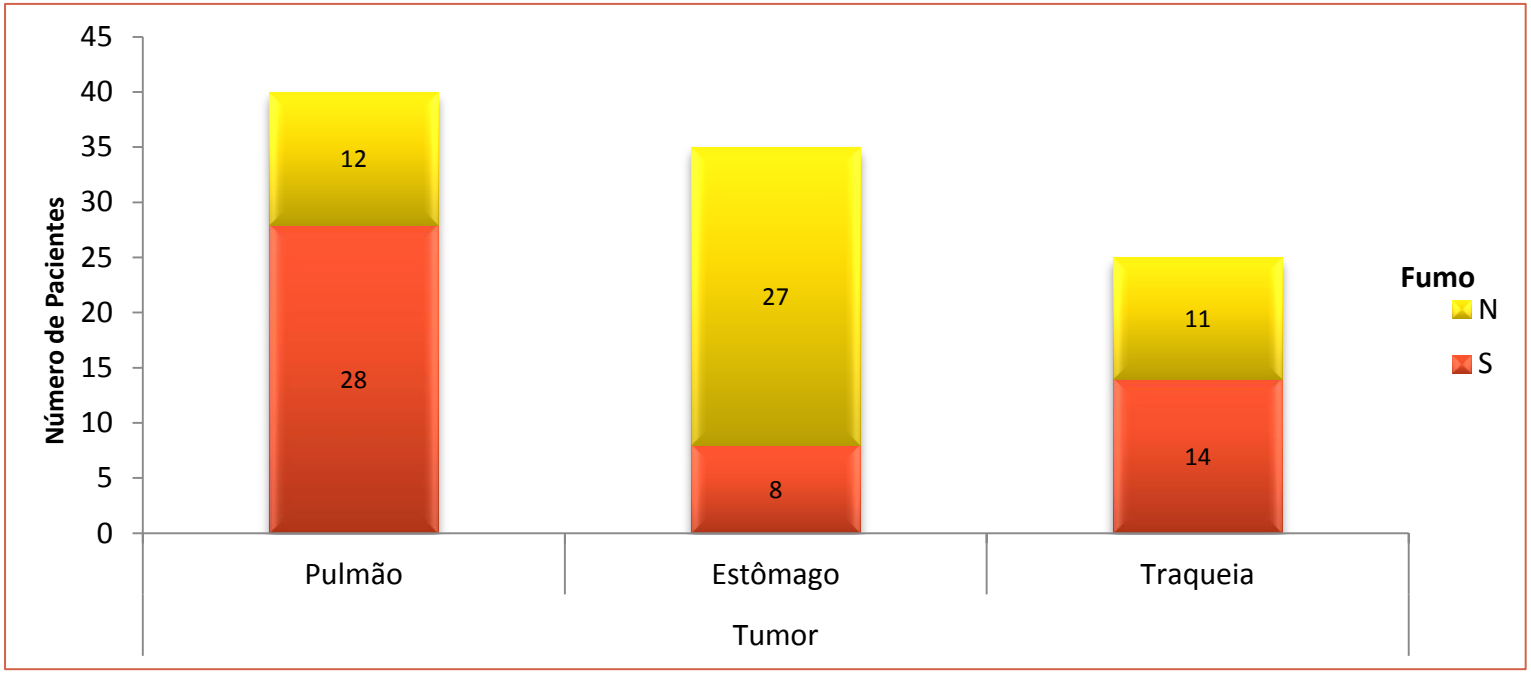
Associação ou independência

- Sexo x Fumo
- Sexo x Local de Tumor
- Fumo x Local de Tumor

Paciente	Sexo	Fumo	Local Tumor
1	M	S	Pulmão
2	F	N	Estômago
3	F	S	Traqueia
4	M	S	Pulmão
5	M	N	Estômago
6	.	.	.
.	.	.	.

Exemplo: Fumo x Tumor

Fumo	Tumor			Total
	Pulmão	Estômago	Traqueia	
S	28	8	14	**50**
N	12	27	11	**50**
Total	**40**	**35**	**25**	**100**



Existe associação?

Exemplo: Fumo x Tumor

Existe associação? Teste – Qui-quadrado

Usando Vassar Stats → Há evidência de associação com 95% de confiança

Chi-Square	df	P
17.07	2	0.0002

Valores Esperados:

Fumo	Tumor			Total
	Pulmão	Estômago	Traqueia	
S	20	17.5	12.5	**50**
N	20	17.5	12.5	**50**
Total	**40**	**35**	**25**	**100**

40*50/100=20

Para o Qui-Quadrado não pode haver valor esperado menor que 5.
Os softwares te avisam caso isso ocorra. Alternativa: Fisher

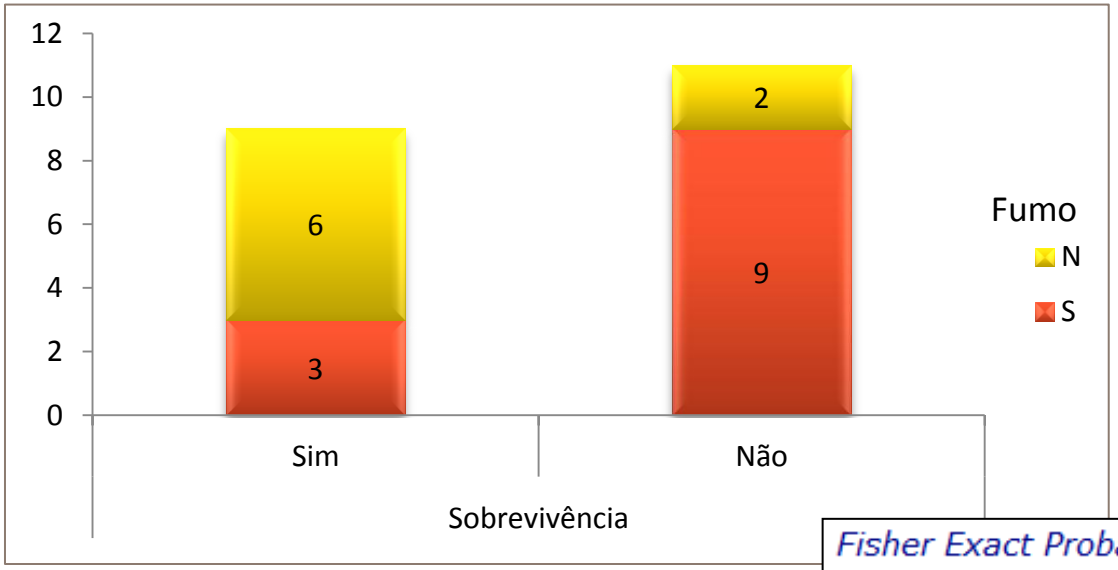
Teste alternativo para N pequeno: Teste exato de Fisher

Dos que tiveram metastese:

Fumo	Sobrevivência		Total
	Sim	Não	
S	3	9	**12**
N	6	2	**8**
Total	**9**	**11**	**20**

Esperados:

Fumo	Sobrevivência		Total
	Sim	Não	
S	5.4	6.6	**12**
N	3.6	4.4	**8**
Total	**9**	**11**	**20**



Fisher Exact Probability Test:

P	one-tailed	0.039890450107168 68
	two-tailed	0.064777327935223 14

Testes Básicos

- N grande – Todas as células com valor maior que 5: Qui-quadrado
- N pequeno: Teste exato de Fisher

Outros

- Correção de continuidade de Yates para Qui-quadrado
- Associação com dependência – McNemar
- Razão de risco, risco relativo
- Razão de chances (odds ratio)

Importante

- O teste deve ser sempre feito com as quantidades reais, e não com %

EXERCÍCIO

Exercício– Regressão

Obtenha uma tabela de contingência para fazer uma análise de regressão ou use os dados fornecidos

Calcule os valores esperados e faça os gráficos apropriados

Faça as análises de associação adequadas

Relatório (Word):

- Caso os dados estejam ligados aos dados usados anteriormente, apenas acrescente as análises no relatório do projeto.

- Caso contrário, monte um relatório simples contendo a parte de análise de regressão.

Dados para o exercício

Caso não tenha um conjunto de dados use este:

Dieta	Homem	Mulher	Total
Sim	1	9	**10**
Não	11	3	**14**
Total	**12**	**12**	**24**

REGRESSÕES

Introdução

Regressão ?

- Estabelecer um modelo com base em um conjunto de dados

Propósito

- 2 variáveis : Medir a relação entre elas
- Mais variáveis : Explicar uma variável em função das demais

Requisitos

- 2 ou mais variáveis numéricas provenientes de uma mesma amostra
- Geralmente variáveis independentes
- Distribuição Normal das variáveis na população

Alguns tipos ...

- Linear (simples e múltipla)
- Logística (Variáveis categóricas ou binárias)
- Modelos Lineares Generalizados
- Modelos temporais

Correlação de Pearson: Definição e conceitos

- Mede a relação entre 2 variáveis numéricas proveninetes de uma mesma amostra.
- A mostra deve ser pareada
- É uma medida entre -1 e 1

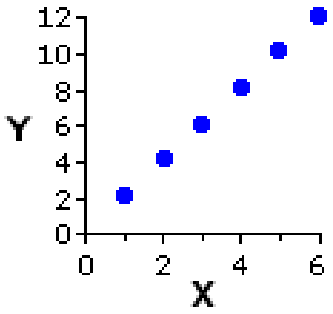
- Correlação negativa => correlação inversa
- Positiva => direta

- Quando mais perto de 1 ou -1 maior a correlação

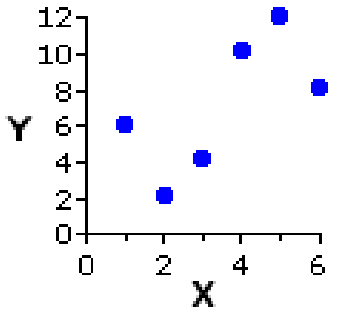
- Gráfico:
  Diagrama de Disperssão (Scatterplot)

- **Importante:** É preciso termos a hipótese de normalidade satisfeita.

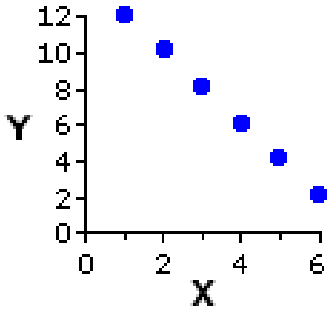
Example I.
$r = +1.0, \quad r^2 = 1.0$



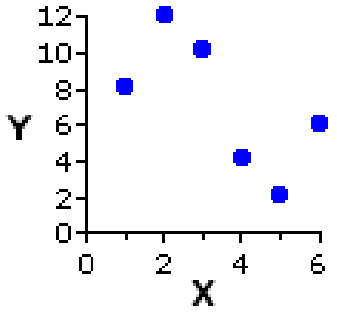
Example II.
$r = +0.66, \quad r^2 = 0.44$



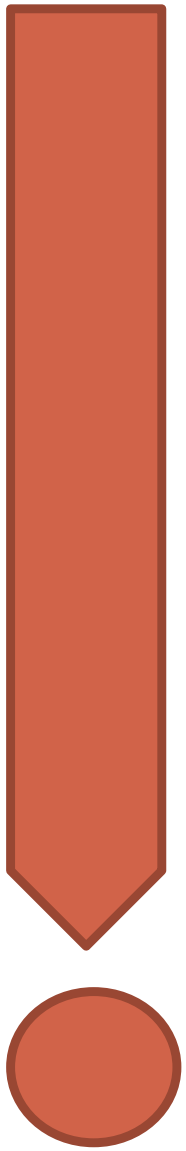
Example III.
$r = -1.0, \quad r^2 = 1.0$



Example IV.
$r = -0.66, \quad r^2 = 0.44$



CORRELAÇÃO NÃO IMPLICA CAUSA



Exemplos:

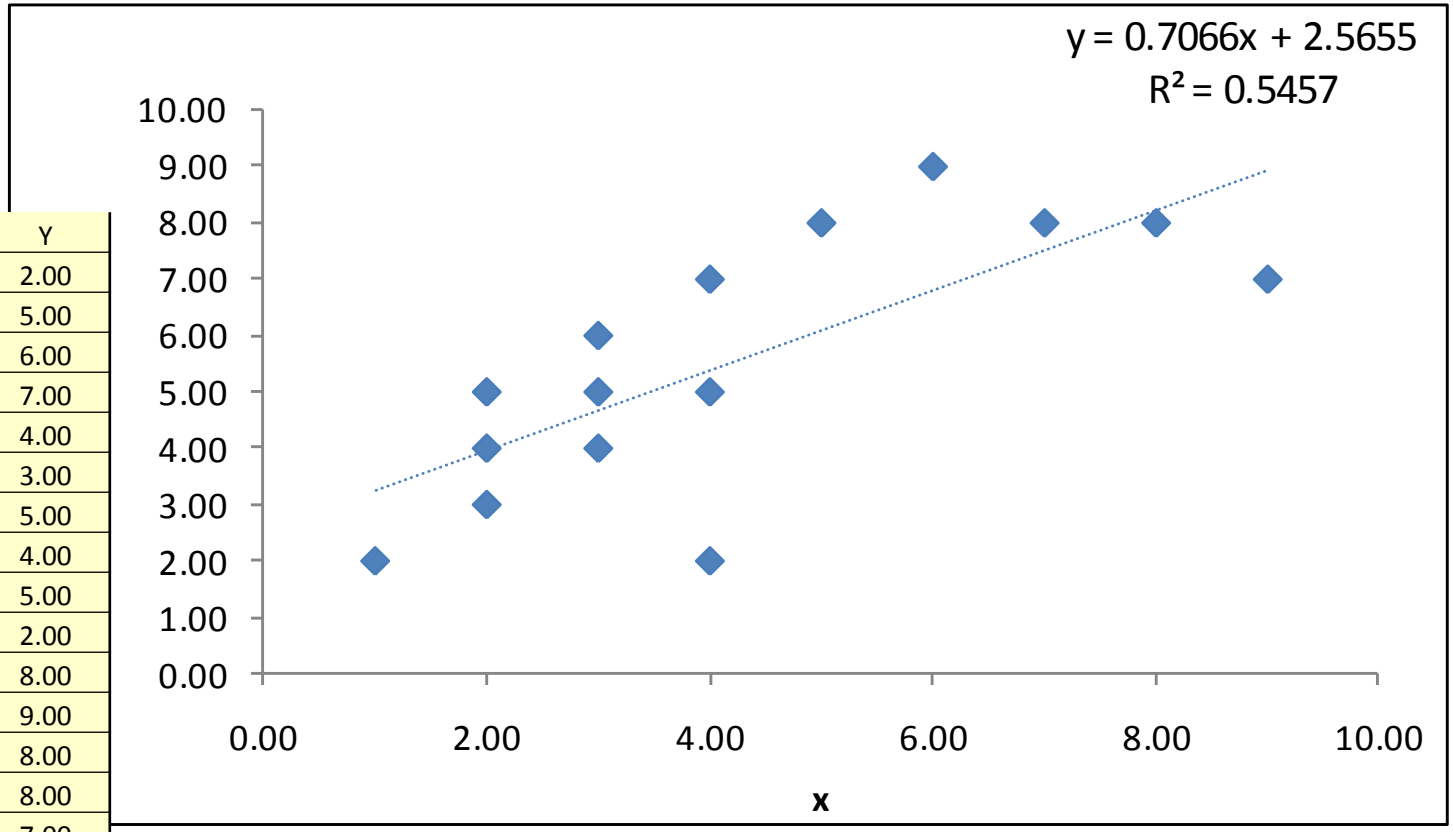
Há uma grande correlação entre a venda de sorvetes e afogamentos. Será que sorvetes causam afogamentos? Ou simplesmente no verão tomamos mais sorvetes e usamos mais a piscina/praia ?

Há um correlação alta nas últimas décadas até hoje entre o o CO_2 e a obesidade. Será que o CO_2 engorda?

Regressão Linear Simples

- Modelo simples de relação entre 2 variáveis
- Formado por um coeficiente de tendência e um de constante
- Um parâmetro de avaliação é o coeficiente de determinação R^2 entre 0 e 1

X	Y
1.00	2.00
2.00	5.00
3.00	6.00
4.00	7.00
3.00	4.00
2.00	3.00
4.00	5.00
2.00	4.00
3.00	5.00
4.00	2.00
5.00	8.00
6.00	9.00
7.00	8.00
8.00	8.00
9.00	7.00



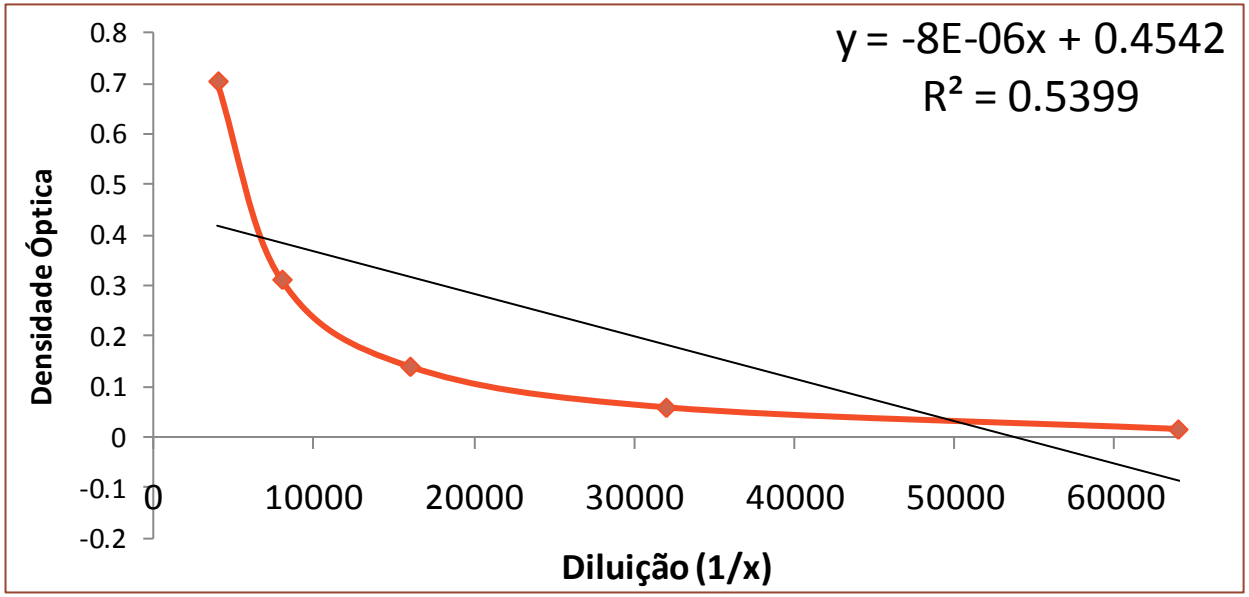
Regressão Linear Simples

Arquivo Editar Gráfico Estimar Y

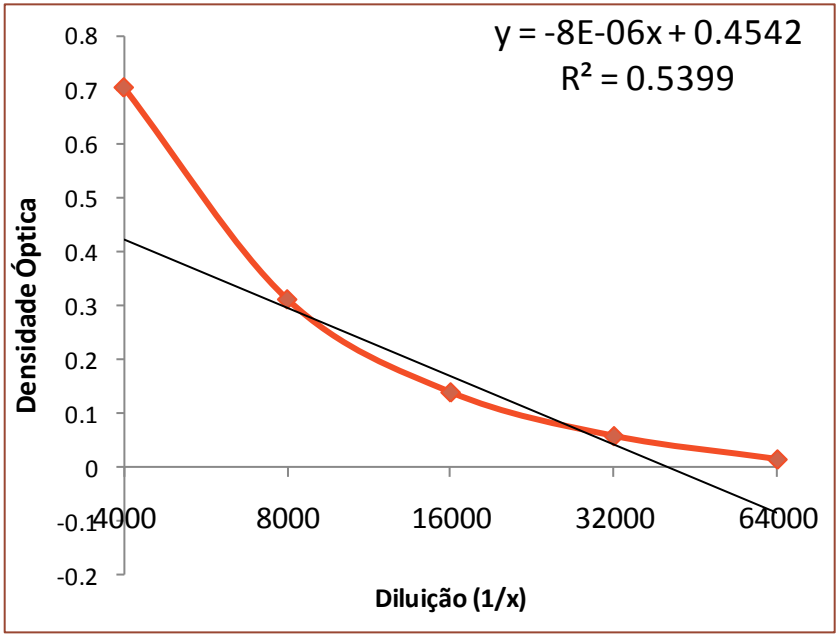
Fontes de variação	GL	SQ	QM
Regressão	1	39.1474	39.1474
Erro	13	32.5859	2.5066
Total	14	71.7333	---
F (regressão) =	15.6177	p = 0.0019	
Variável dependente =	Coluna 2		
Variável independente =	Coluna 1		
Média (X) =	4.2000		
Média (Y) =	5.5333		
Coef. de Determinação (R2) =	0.5457		
R2 (ajustado) =	0.5108		
Coeficiente de Correlação =	0.7387		
Intercepto (a) =	2.5655	t = 3.0004	p = 0.0102
Coef. de Regressão (b) =	0.7066	t = 3.9519	p = 0.0016
IC 95% (a)	0.719 a 4.412		
IC 95% (b)	0.320 a 1.093		
Equação	Y' = a + bX		

Regressões Logarítmicas

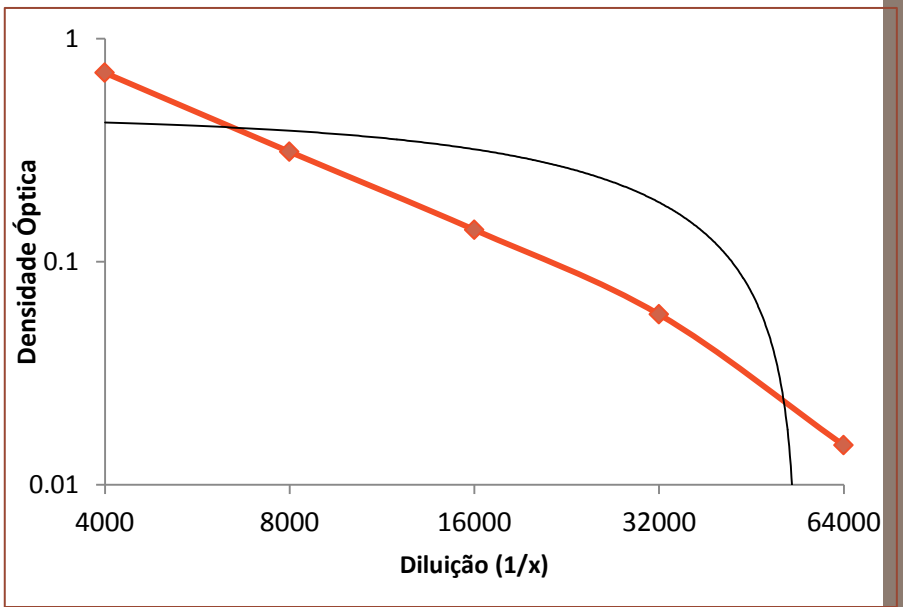
Diluição (1/x)	Densidade Óptica
4000	0.703
8000	0.311
16000	0.139
32000	0.058
64000	0.015



Escala X em Log:



Escala X e Y em Log:

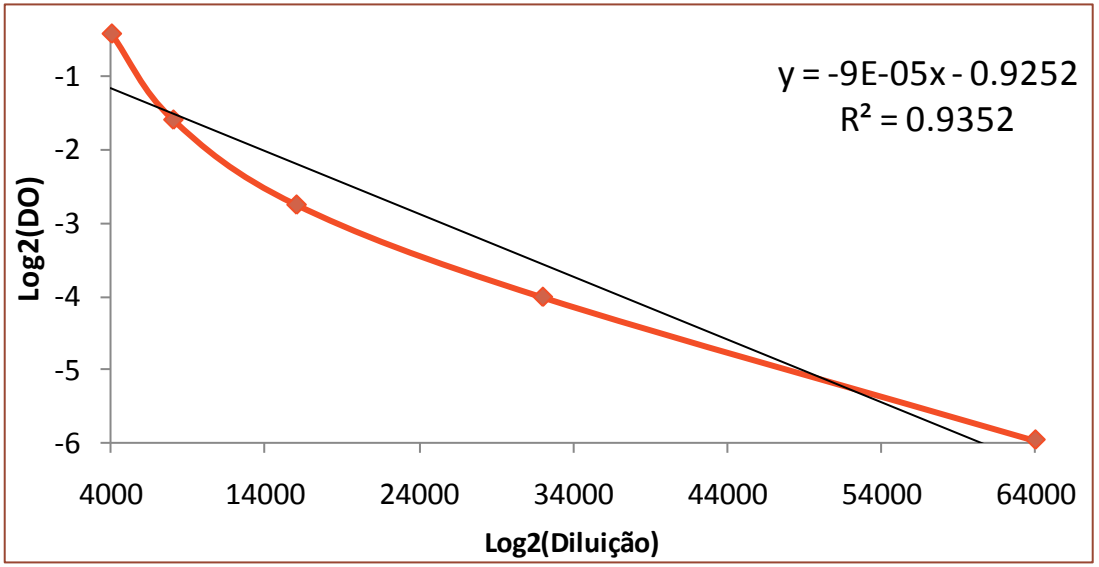


Regressões Logarítmicas

Diluição (1/x)	Log2(DO)
4000	-0.508
8000	-1.685
16000	-2.847
32000	-4.108
64000	-6.059

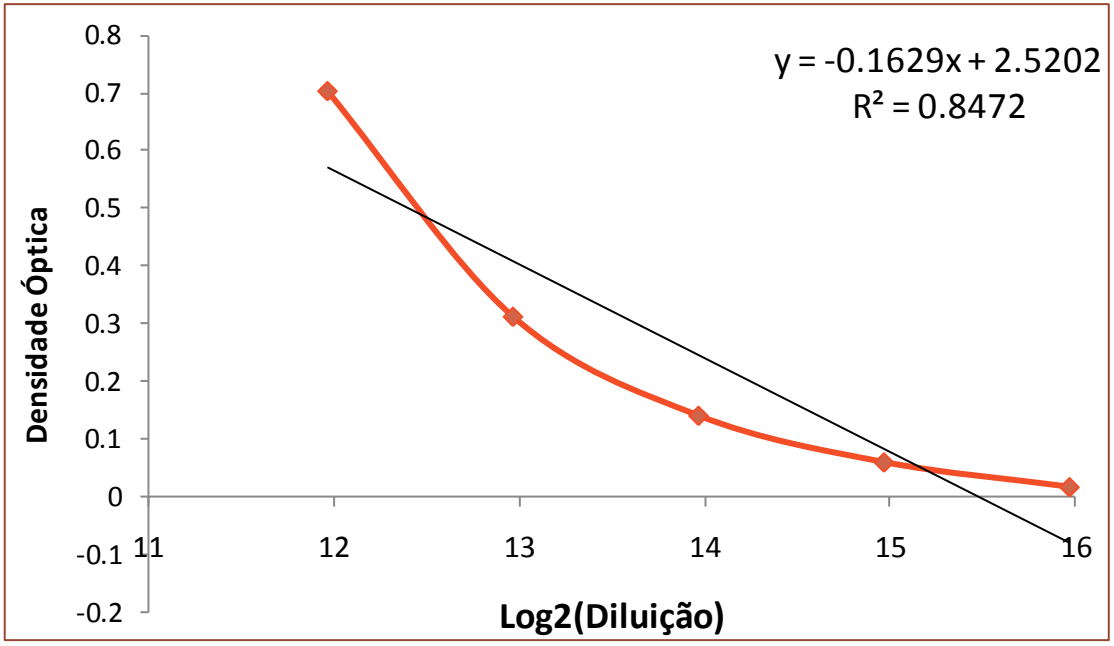
$Log_2(DO)=a*(Diluição)+b$

$DO=2^{(a*(Diluição)+b)}$



Log_2(Diluição)	Densidade Óptica
11.9658	0.703
12.9658	0.311
13.9658	0.139
14.9658	0.058
15.9658	0.015

$DO=a* log_2(Diluição)+b$



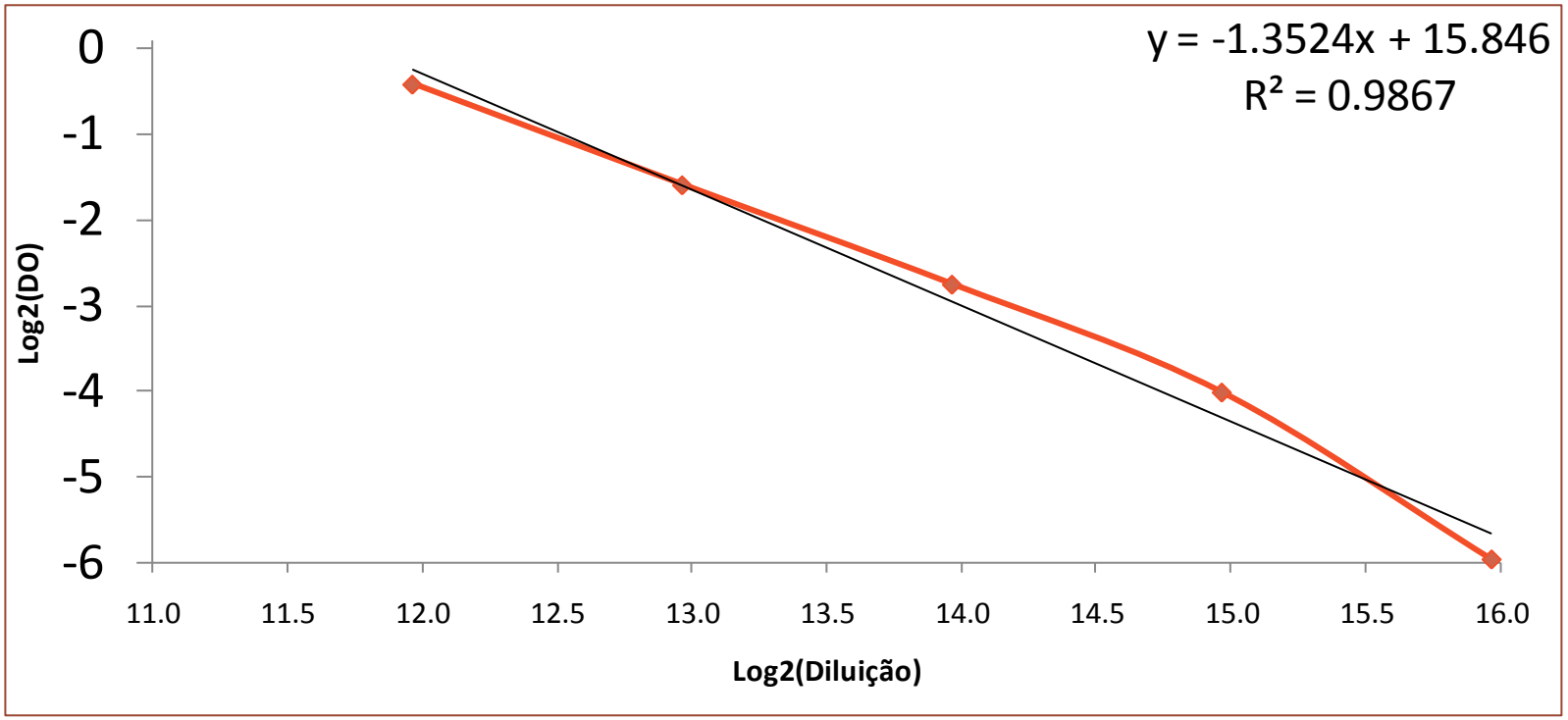
Regressões Logarítmicas

Log_2(Diluição)	Log_2(DO)
11.966	-0.508
12.966	-1.685
13.966	-2.847
14.966	-4.108
15.966	-6.059

$Log_2(DO)=a*Log_2(Diluição)+b$

$DO=2^{a*log2(Diluição)}\ 2^{b}$

$DO= 2^{b}\ (Diluição)^{a}$

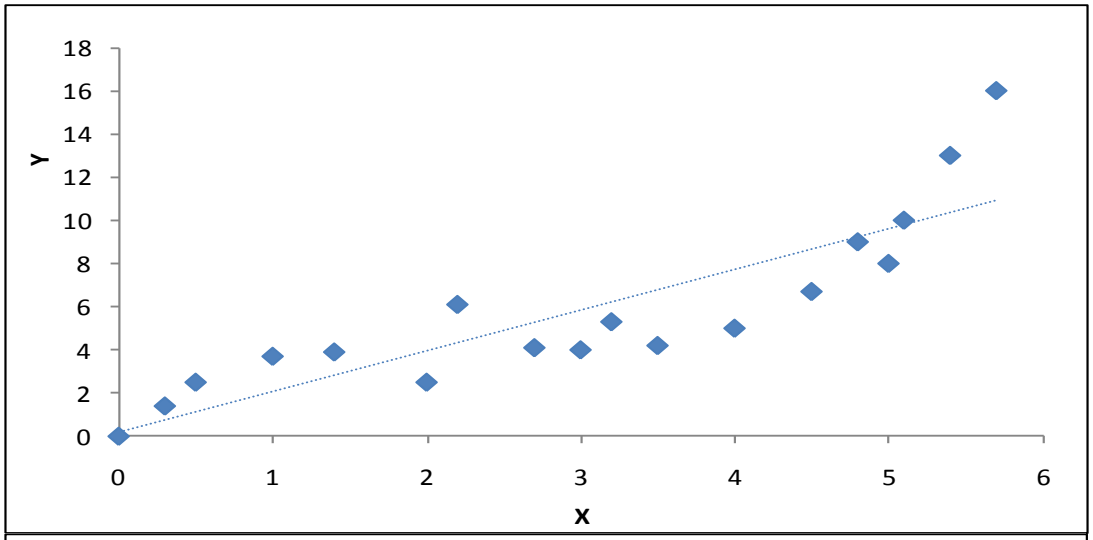


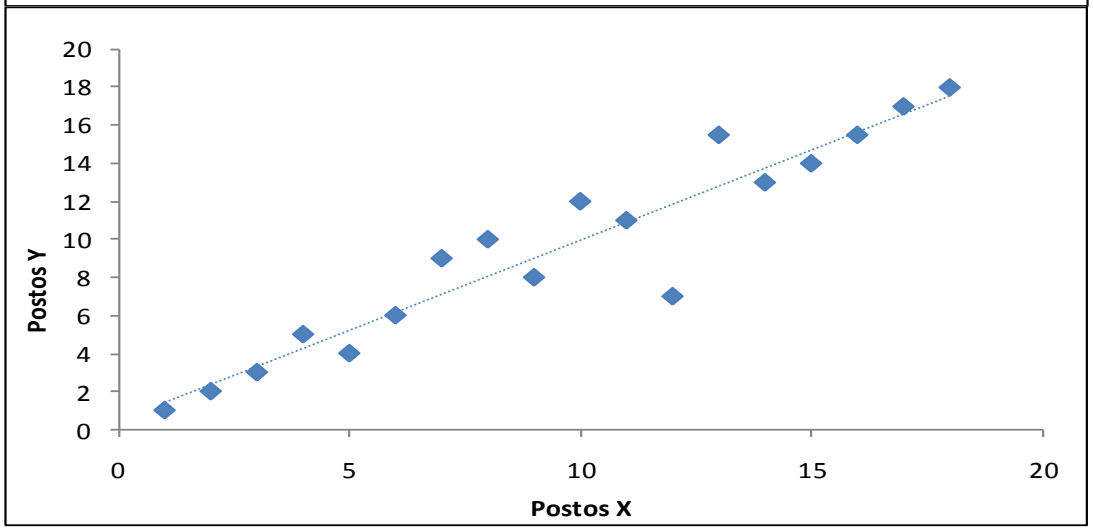
Coeficiente de Correlação de SPEARMAN:

-Não exige normalidade – É não paramétrico
-Usa postos para obter a correlação

X	Y
4.0	5.0
0.0	0.0
3.0	4.0
1.0	3.7
5.0	8.0
2.0	2.5
3.5	4.2
0.5	2.5
0.3	1.4
4.5	6.7
3.2	5.3
4.8	9.0
5.1	10.0
5.4	13.0
5.7	16.0
2.7	4.1
1.4	3.9
2.2	6.1

Pearson
0.866





Postos x	Postos y
7.0	9.0
18.0	18.0
10.0	12.0
15.0	14.0
4.0	5.0
13.0	15.5
8.0	10.0
16.0	15.5
17.0	17.0
6.0	6.0
9.0	8.0
5.0	4.0
3.0	3.0
2.0	2.0
1.0	1.0
11.0	11.0
14.0	13.0
12.0	7.0

Spearman
0.95

EXERCÍCIO

Exercício– Regressão

Obtenha 2 variáveis de uma mesma amostra para fazer uma análise de regressão ou use os dados fornecidos

Faça a parte de análise descritiva e gráfica

Calcule a correlação (Pearson e Spearman), ajuste uma regressão linear adequada e interprete os resultados

Relatório (Word):

- Caso os dados estejam ligados aos dados usados anteriormente, apenas acrescente as análises no relatório do projeto.

- Caso contrário, monte um relatório simples contendo a parte de análise de regressão.

Dados para o exercício

Caso não tenha um conjunto de dados com 2 variáveis utilize esse exemplo, onde temos os resultados de um ELISA

Grau de diluição	Proteínas/mL	Absorbância
1	2.8E+09	1.792
1/2	1.4E+09	1.522
1/4	7.0E+08	1.153
1/8	3.5E+08	0.688
1/16	1.8E+08	0.431
1/32	8.8E+07	0.237
1/64	4.4E+07	0.161

RESUMÃO DO DIA

Tipos de testes de hipóteses

Frequências
Variáveis Qualitativas

Teste de Homogeniedade

(1 categoria)

- Teste Qui-Quadrado
- Teste Exato de Fisher

Teste de associação ou independência

(2 ou mais categorias)

- Teste Qui-Quadrado
- Testes exatos de Fisher (binomiais)

Relação
Variáveis Quantitativas

Correlação

- Pearson
- Spearman

Regressões

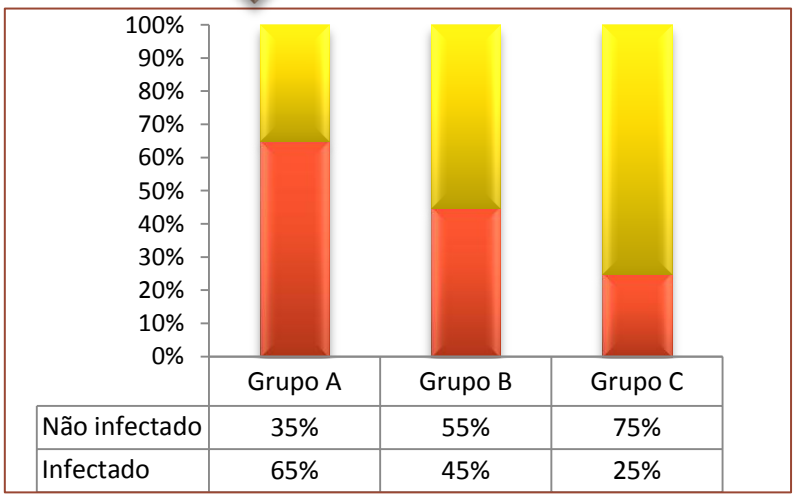
- Paramétricas
- Não paramétricas

Gráficos X Testes

Frequências
Variáveis Qualitativas

Testes de Comparação de Distribuições em Categorias

- Barras com frequências
- Pizza/torta

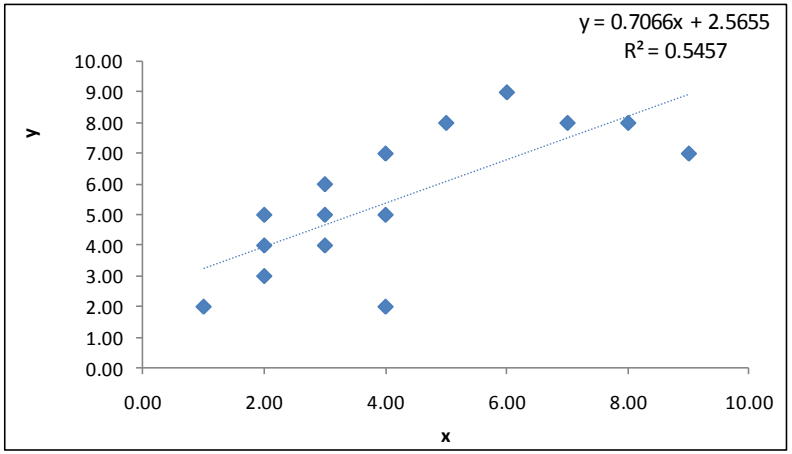


	Grupo A	Grupo B	Grupo C
Não infectado	35%	55%	75%
Infectado	65%	45%	25%

Relação
Variáveis Quantitativas

Correlação/Regressões

- Diagrama de Disperssão (Scatterplot)



OBRIGADO E BONS ESTUDOS!!!

Dúvidas?
pedrospeixoto@yahoo.com.br
www.ime.usp.br/~pedrosp