INTRODUÇÃO À ANALISE DE DADOS ESE

Prof. Paulo C. Masiero

2º. 2011

Aula 1- Capítulos 3 e 8

Agenda

- 1. Medidas e escalas
- 2. Análise Descritiva
- 3. Teste de Hipóteses

Medição

- Uma **medição** é um mapeamento entre o mundo empírico e o mundo formal.
- Uma **medida** é o número ou o símbolo atribuído a uma entidade por esse mapeamento para caracterizar um atributo.
- Exemplo:
  - LoC (Linhas de Código) é uma medida do tamanho do programa.
  - Programa-X → 1000 LoC

Medição

- Se uma medida for usada em um estudo empírico, ela deve ser válida.
- O mapeamento de um atributo para um valor de medida pode ser feito de diferentes formas e cada diferente mapeamento de um atributo é uma **escala**.
- Exemplo: o tamanho de um objeto pode ser medido em cm, m, polegada etc.

Medição (Cont.)

- Quando uma transformação de uma medida para outra preserva o relacionamento entre os objetos, então se diz que é uma transformação admissível .
- Se as conclusões ou declarações sobre objetos permanecem verdadeiras depois de uma mudança de escala, então se diz se diz que elas são significativas, senão se diz que não são significativos.
- Ex. Se a temperatura de uma sala A é 10º e de B é 20º, pode-se dizer que B é duas vezes mais quente que A. Se transformarmos para a escala Farenheit, tem-se 50°F e 68°F e não se pode afirmar a mesma coisa.

Variáveis e Valores

- As variáveis de um estudo empírico podem ser:
  - Categóricas: os valores representam tipos, formas e procedimentos
  - Numéricas: os valores representam doses ou níveis de aplicação da variável
- Os valores das variáveis são coletados em escalas:
  - Existem diversas escalas para coleta e representação dos valores: **nominal**, **ordinal**, **intervalar** e **razão**
  - As escalas determinam as operações que podem ser aplicadas sobre os valores das variáveis

Escala Nominal

- Representa diferentes tipos de um elemento, sem interpretação numérica e de ordenação entre eles. É a menos expressiva.
- Exemplos:
  - Classificações, etiquetagem e tipos de defeitos
  - Diferentes linguagens de programação: Java, C++, C#, Pascal, ...
- A escala não nos permite dizer, por exemplo, que Java é menor que C# ou que C++ é melhor que C#

Escala Ordinal

- Representam diferentes tipos de um elemento que podem ser ordenados, ainda que sem qualquer interpretação numérica. Pode-se ordenar: maior que, melhor que ...
- Exemplos em software incluem:
  - Diferentes níveis no CMMI (Nível 1, ..., Nível 5) ou MPS.BR (Nível G, ..., Nível A)
  - Diferentes graus de coesão (funcional, procedimental, temporal, sequencial, ...)
- A escala permite dizer que, no CMMI, “Nível 2” é menor do que “Nível 3”, mas não permite dizer que a diferença de qualidade entre empresas do “Nível 2” e empresas do “Nível 3” é a mesma que entre empresas do “Nível 3” e “Nível 4”

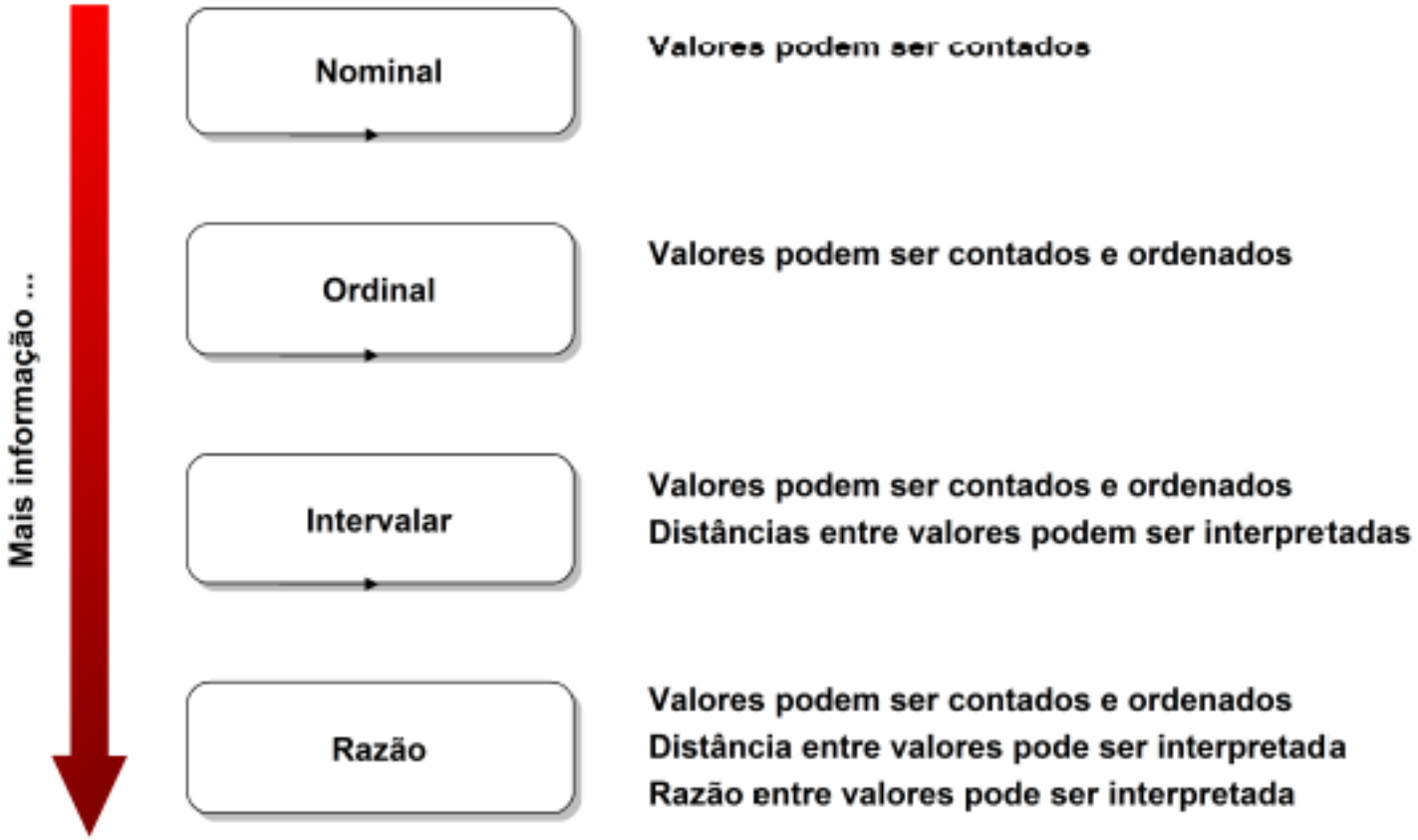
Escala Intervalar

- A diferença entre duas medidas é significativa mas não o valor nominal.
- Os valores podem ser ordenados e há uma noção de distância relativa. Porém, a razão entre estes valores não tem significado.
- Exemplo: pode-se dizer que 2006 é um ano após 2005 e um ano antes de 2007, e a distância em dias entre eles é a mesma, não faz sentido calcular a razão entre 2006 e 2007.
- Isto é possível porque toda escala intervalar possui um zero arbitrário (no caso das datas, o ano zero)

Escala Razão

- Os valores de uma escala razão podem ser ordenados, distâncias entre valores consecutivos possuem o mesmo significado e a razão entre valores pode ser interpretada
- Exemplos em software incluem o tamanho de um sistema (LoC), o esforço necessário para a sua construção e o tempo de duração do projeto que resultou no sistema
- A escala permite dizer, por exemplo, que um software com X linhas de código é duas vezes menor que um software com 2X linhas de código

Informação nas Escalas



Exemplos de medidas em Engenharia de Software

Classe	Exemplos de Objetos	Tipo de atributo	Exemplos de Medidas
Produto	Código	Interno Externo	Tamanho Confiabilidade
Processo	Teste	Interno Externo	Esforço Custo
Recurso	Pessoa	Interno Externo	Idade Produtividade

Interno: pode ser medido puramente em termos do objeto
Externo: só pode ser medido com respeito a como o objeto se relaciona com outros objetos

Dificuldades: definir medidas e usar as escalas intervalar e razão

Análise de dados (Resultados de experimentos)

- Há duas alternativas:
  - Análise descritiva dos dados
  - Testes de hipóteses

Análise Descritiva

- Estatística Descritiva
  - Medidas de Tendência Central (média, mediana, moda)
  - Medidas de dispersão (desvio padrão, variância)
  - Correlações (Pearson, Spearman)
- Análise Gráfica
  - Diagramas de dispersões
  - Histogramas e Gráficos de Pizza
  - Box Plots

Tipo de Escala	Medida de Tendência Central	Dispersão	Dependência
Nominal	Moda	Frequência	
Ordinal	Media	Intervalo de Variação	Coefs. de correlação de Spearman e de Kendall
Intervalar	Média	Desvio padrão Variância e intervalo	Coefs. de correlação de Pearson
Quociente	Média geométrica	Coeficiente de variação	

Metas da Análise Descritiva

- Identificar tendências centrais de variáveis e seus tratamentos
- Identificar o grau de dispersão
- Identificar pontos fora da curva (outliers)
- Identificar Correlações

Estatísticas relevantes por tipo de escala

Tipo escala	Med. Tend. Central	Dispersão	Dependência
Nomimal	Moda	Frequência	
Ordinal	Média, percentil	Intervalo de variação	Coefs. de correção de Spearman e de Kendall
Intervalar	Média	Desvi padrão, variância, range	Coef. De correção de Pearson
Quociente	Média geométrica	Coeficiente de variação	

Medidas de Tendência Central

- Média

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

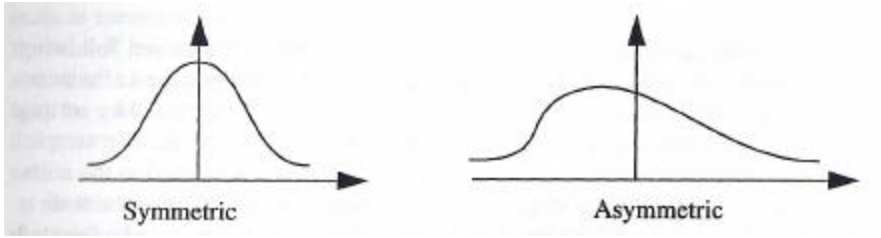
  - Ex. para o conjunto de dados (1,1,2, 4), X=2
  - A média corresponde ao percentil 50%, indicando que 50% das amostras estão abaixo da média.
- Moda
  - Representa a amostra que ocorre mais vezes.
  - Ex. A moda para o conjunto de dados (1,1,2,4) é 1

Medidas de Tendência Central

- Mediana
  - Representa o valor médio do conjunto de dados
  - Se n é impar, pega-se a amostra média do conjunto ordenado
  - Se n é par, pode-se calcular a média das duas amostras centrais.
  - Ex. a mediana de (1,1,2,4) é 1,5. A mediana de (1,1,2,4,6) é 2

Medidas de Tendência Central

- A média é a mediana são a mesma se a distribuição é simétrica



- Média Geométrica

$$\sqrt[n]{\prod_{i=1}^{n} x_i}$$

Medidas de Dispersão

- A variância é definida como: $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$
- O desvio padrão s é definido
  como a raiz quadrada da variância.
- Ele é geralmente preferido em relação à variância porque tem a mesma unidade de medida que os valores da amostra.
- O *range* de um conjunto de dados é a distância entre os valores máximos e mínimos do conjunto de dados.
  - Range = Xmax - Xmin

Medidas de Dispersão (Cont.)

- Intervalo de variação (Xmin,Xmax)
- Coeficiente de variação: é expresso como uma porcentagem da média:

$$100 \cdot \frac{s}{\bar{x}}$$

- Uma visão geral da dispersão é dada pelo frequência de cada valor.

Medidas de Dispersão (Cont.)

- A frequência relativa é calculada dividindo-se cada frequência pelo número total da amostra.
  - Ex (1,1,1,2,2,3,4,4,4,5,6,6,7), com tamanho 13. A frequência relativa dos valores é
    - 1, 23%
    - 2,15%
    - 3, 8%
    - etc.

Regressão linear

- Se o conjunto de dados contém variáveis estocásticas X e Y em pares (Xi,Yi) e suspeitamos que há uma função y= f(x) que relaciona os pares x e y.
- Se y=c1+c2.x então dizemos que a regressão é linear.

Regressão Linear

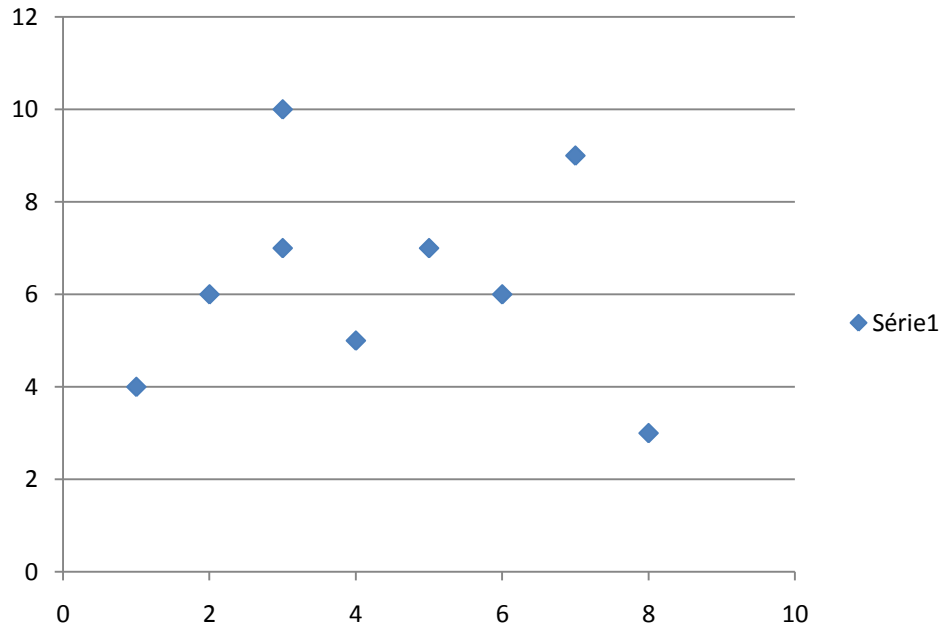
$$r = \frac{c_{xy}}{s_x \cdot s_y} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{\left(n \cdot \sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

- r= Coeficiente de correlação (de Pearson)
- Se r é igual a zero não há correlação
- O valor de r fica entre -1 e +1
- Pode haver uma correlação não linear mesmo se r=0.

Regressão Linear

- Se a escalar é ordinal ou se o conjunto de dados não está distribuído normalmente, o coeficiente de correlação de Spearman pode ser usado (Rs)
- O cálculo é feito da mesma forma mas os ranks (isto é, os números ordinais em que a amostra é ordenada) é que são usados, ao invés dos valores da amostra

Visualização gráfica (Dispersão)



Visualização gráfica (Histrograma)

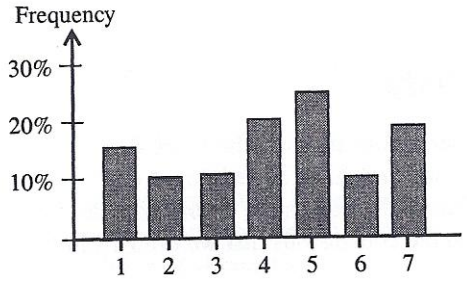
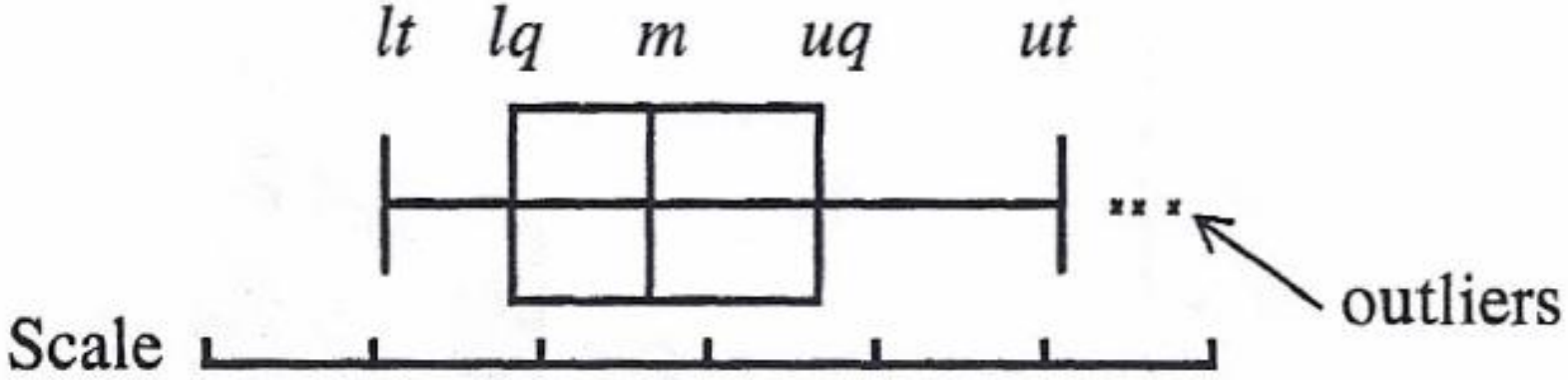
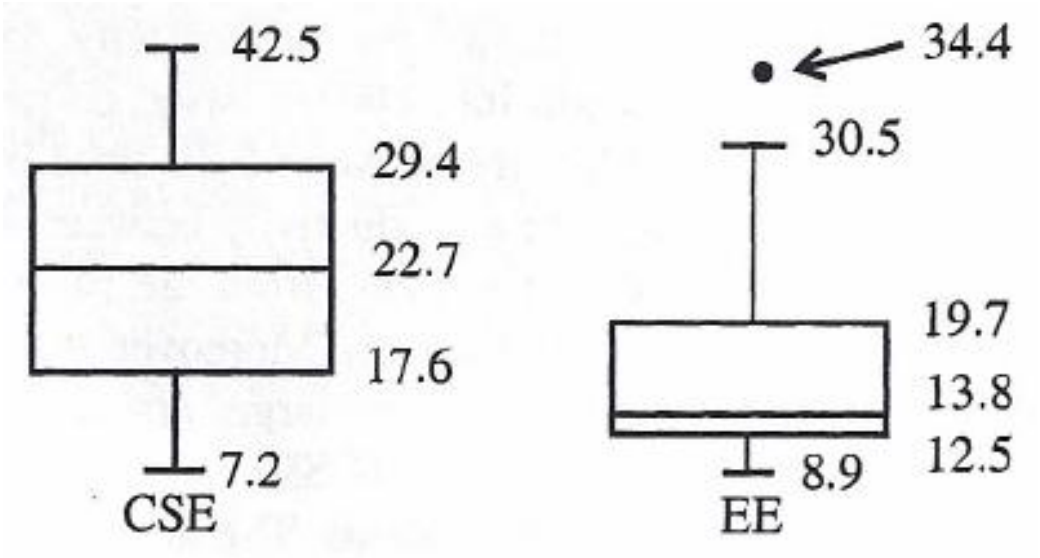


Figure 19. *A histogram.*

Box Plot



m = mediana
lq= 25%
uq = 75%
O tamanho da caixa é d = lq – uq
ut= uq+1.5d
lt= lq – 1.5d


42.5
29.4
22.7
17.6
7.2
CSE
34.4
30.5
19.7
13.8
12.5
8.9
EE

Redução do conjunto de dados

- Um critério é baseado no resultado do experimento. Ex. sujeitos que não participaram seriamente do experimento.
- Os erros podem ser sistemáticos ou ocorrer como outliers.
- Ao encontrar outliers, é necessário decidir se eles serão retirados ou não. Ex. surgiram porque pessoas inexperientes participaram do experimento.
- Olhar também dados redundantes.

Testes de Hipóteses

- Hipóteses avaliadas por testes estatísticos definidos por pesquisadores da estatística inferencial
- Normalmente são definidas duas hipóteses
  - **Hipótese nula (H0): indica que as diferenças observadas no estudo** são coincidentais, ou seja, é a hipótese que o analista deseja rejeitar com a maior significância possível
  - **Hipótese alternativa (H1): é a hipótese inversa à hipótese nula, que** será aceita caso a hipótese nula seja rejeitada
- Os testes estatísticos verificam se é possível rejeitar a hipótese nula, de acordo com um conjunto de dados observados e suas propriedades estatísticas

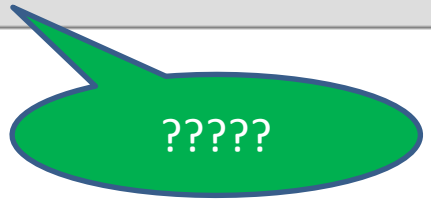
Testes de Hipótese

- Os testes comparam médias entre grupos de participantes realizando tratamentos diferentes

"Utilizando a técnica XYZ, os desenvolvedores concluem a atividade de projeto em menos tempo do que utilizando a técnica ABC"

Hipótese Nula: $\mu\ (Tempo_{XYZ}) = \mu\ (Tempo_{ABC})$

Hipótese Alternativa: $\mu\ (Tempo_{XYZ}) \neq \mu\ (Tempo_{ABC})$



Teste Estatístico

- Calculados fundamentalmente a partir de uma função de teste que considera três valores:
  - Diferença entre os valores “médios” das estatísticas para os tratamentos
  - “Dispersão” dos valores da estatística
  - Número de amostras
- A função de teste, $F(m,\sigma,N)$, depende do:
  - tipo de distribuição dos dados, e.x., normalidade e homocedasticidade.
  - Número de fatores e tratamentos

Homogeneidade da variância

Exemplo

- Quer-se testar a Hipótese “homens são mais altos que mulheres”
  - Determina-se uma amostra da população utilizando um fator e dois tratamentos
  - A certeza depende de:
    - Número de pessoas amostradas
    - Diferença entre a altura média nos tratamentos
    - Dispersão da altura nos tratamentos

Tipos de Erros

- A verificação das hipóteses sempre lida com o risco de um erro de análise acontecer
  - O erro do tipo I (a) acontece quando o teste estatístico indica um relacionamento entre causa e efeito e o relacionamento real não existe
  - O erro do tipo II (b) acontece quando o teste estatístico não indica o relacionamento entre causa e efeito, mas existe este relacionamento

α = P (erro-tipo-I) = P (H_{NULA} é rejeitada | H_{NULA} é verdadeira)

β = P (erro-tipo-II) = P (H_{NULA} não é rejeitada | H_{NULA} é falsa)

Nível de Significância

- Indica a probabilidade de se cometer um erro do tipo I:
  - Os níveis de significância (α) mais comumente utilizados são 10%, 5%, 1% e 0.1%
  - Chama-se de *p-value o menor nível de significância com que se* pode rejeitar a hipótese nula
  - Dizemos que há significância estatística quando o *p-value é menor* que o nível de significância adotado

Procedimento para o Teste de Hipótese

- Fixar o nível de significância do teste
- Obter uma estatística (estimador do parâmetro que se está testando) que tenha distribuição conhecida sob H0
- A estatística de teste e o nível de significância constroem a região crítica pela qual o teste passa
- Usando as informações amostrais, obter o valor da estatística (estimativa do parâmetro)
- Se valor da estatística pertencer à região crítica, rejeita-se a hipótese nula, aceitando-se a hipótese alternativa
- Caso contrário, não se rejeita a hipótese nula e nada se pode dizer a respeito da hipótese alternativa

Teste de Hipótese na Prática

- Na prática:
  - escolhe-se a estatística de teste
  - escolhe-se o valor P (significância)
  - Usa-se uma ferramenta estatística para aplicar o teste e verificar o valor de P
- A escolha do teste depende da determinação do tipo de distribuição dos dados e de quantos fatores e tratamentos vão ser analisados no teste
  - Testes paramétricos: assumem uma distribuição e são mais poderosos
  - Testes não paramétricos: não assumem uma distribuição . Têm uma aplicação mais abrangente

Alguns Tipos de Teste

Projeto	Teste paramétrico	Teste não-paramétrico
Um fator, um tratamento	-	Binomial Chi-2
Um fator, dois tratamentos aleatórios	Teste T Teste F	Mann-Whitney Chi-2
Um fator, dois tratamentos pareados	Teste T pareado	Wilcoxon
Um fator, mais de dois tratamentos	ANOVA	Kruskal-Wallis Chi-2

ANOVA (ANalysis Of VARiance)

- Usado para avaliar experimentos com várias quantidades de projetos.
- Baseado na análise da variabilidade total dos dados e da variabilidade de uma partição de acordo com diferentes componentes.
- Na sua forma mais simples compara a variabilidade devida ao tratamento com a variabilidade devida a erros randômicos.

Forma mais simples: compara se as amostras têm o mesmo valor médio, isto é, o projeto tem um fator e dois tratamentos.

ANOVA, one factor, more than two treatments	
Input	a samples: $x_{11}, x_{12}, \ldots x_{1n_1}$; $x_{21}, x_{22}, \ldots x_{2n_2}$;...; $x_{a1}, x_{a2}, \ldots x_{an_a}$
H_0	$\mu_{x_1} = \mu_{x_2} = \ldots = \mu_{x_a}$, i.e. all expected means are equal
Calculations	Calculate: $SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{x_{..}^2}{N}$ $SS_{Treatment} = \sum_{i=1}^{a} \frac{x_{i.}^2}{n_i} - \frac{x_{..}^2}{N}$ $SS_{Error} = SS_T - SS_{Treatment}$ $MS_{Treatment} = SS_{Treatment}/(a-1)$ $MS_{Error} = SS_{Error}/(N-a)$ $F_0 = MS_{Treatment}/MS_{Error}$ where N is the total number of measurements and a dot index denotes a summation over the dotted index, e.g. $x_{i.} = \sum_j x_{ij}$
Criterion	Reject H_0 if $F_0 > F_{\alpha,a-1,N-a}$. Here, F_{α,f_1,f_2} is the upper α percentage point of the F distribution with f_1 and f_2 degrees of freedom, which is tabulated in, for example, Table A5.1, Table A5.2 and [Montgomery97, Marascuilo88].

Teste t (Student)

- Usado para comparar duas amostras independentes (um fator com dois níveis).
- Pode ser usado com diferentes suposições.

t-test	
Input	Two independent samples: $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_m$.
H_0	$\mu_x = \mu_y$, i.e. the expected mean values are the same.
Calculations	Calculate $t_0 = \dfrac{\bar{x} - \bar{y}}{S_p \sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$, where $S_p = \sqrt{\dfrac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$, and, S_x^2 and S_y^2 are the individual sample variances.
Criterion	Two sided (H_1: $\mu_x \neq \mu_y$): reject H_0 if $\lvert t_0 \rvert > t_{\alpha/2,\, n+m-2}$. Here, $t_{\alpha,\, f}$ is the upper α percentage point of the t distribution with f degrees of freedom, which is equal to $n+m-2$. The distribution is tabulated in, for example, Table A1 and [Montgomery97, Marascuilo88]. One sided (H_1: $\mu_x > \mu_y$): reject H_0 if $t_0 > t_{\alpha,\, n+m-2}$.

Example of t-test. The defect densities in different programs have been compared in two projects. In one of the projects the result is $x = \{3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 2.68, 4.30, 2.49, 1.54\}$ and in the other project the result is $y = \{3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49\}$. The null hypothesis is that the defect density is the same in both projects, and the alternative hypothesis that it is not. Based on the data it can be seen that $n = 10$ and $m = 11$. The mean values are $\bar{x} = 2.853$ and $\bar{y} = 4.1055$.

It can be found that $S_x^2 = 0.6506$, $S_y^2 = 0.4112$, $S_p = 0.7243$ and $t_0 = -3.96$.

The number of degrees of freedom is $f = n+m-2 = 10+11-2 = 19$. In Table A1, it can be seen that $t_{0.025,\,19} = 2.093$. Since $|t_0| > t_{0.025,\,19}$ it is possible to reject the null hypothesis with a two tailed test at the 0.05 level.

Table A1. *Critical values two-tailed t-test (5%), see Section 8.3.4 and 8.3.7.*

Degrees of freedom	t-value
1	12.706
2	4.303
3	3.182
4	2.776
5	2.571
6	2.447
7	2.365
8	2.306
9	2.262
10	2.228
11	2.201
12	2.179
13	2.160
14	2.145
15	2.131
16	2.120
17	2.110
18	2.101
19	2.093
20	2.086
21	2.080
22	2.074
23	2.069
24	2.064
25	2.060

Tarefas para a próxima aula

- Instale ou use o software estatístico R. (Lab. De Estatística). Aprenda a usar as funções básicas.
- Considere os dois conjuntos de amostras do exemplo do t-test
  - Faça um box plot das duas sequências. Descubra como é calculado o "bigode" pelo R, ou se é opcional.
  - Repita o teste como no exemplo e comprove os resultados
  - Retire três amostras de cada conjunto e repita o t-test. Conclua se é possível continuar a rejeitar To com nível de significância (alfa) igual 5%.