1.0

Introdução à análise estatística com SPSS

Conceitos básicos

Pedro Sá Couto (p.sa.couto@ua.pt)

Departamento de Matemática

Universidade de Aveiro

#1

Tópicos da formação:

- **Conceitos básicos:**
  - Porquê Bioestatística?
  - Tipo de dados
  - População e amostra
  - Estatística descritiva
  - Distribuições teóricas: Normal e as outras
  - Inferência estatística
  - Potência de um teste e dimensão da amostra
  - Escolha do teste estatístico.

2

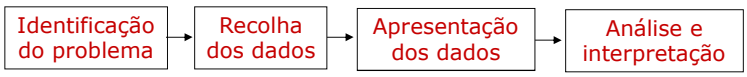
Aula 1

Estatística descritiva

#1

Basic Concepts

- **Why to study Biostatistics?**
  - A utilidade da **Bioestatística** pode ser sumariada no seguinte:
    - Permite **descrever e compreender** relações entre variáveis,
    - Permite a tomada de **melhores e mais rápidas decisões** num curto espaço de tempo,
    - Facilita a **tomada de decisões para fazer face a uma mudança**, ou seja, é fundamentada em critérios objectivos
  - As etapas que definem o **método estatístico** de resolução de problemas são:



4

Basic Concepts

- **Why to study Biostatistics?**
  - Na análise estatística, o investigador necessita sempre de **"algo" que possa medir, controlar ou manipular** durante o processo de investigação.
  - Este "algo" designa-se por **variável aleatória** e a informação que elas contém dependem de como foram medidas e da qualidade dessa medição.
  - As **variáveis estatísticas** podem ser classificadas como:
    - **Variáveis qualitativas** - variáveis cuja a escala de medida apenas indica a sua presença em categorias de classificação discreta, sendo exaustivas e mutuamente exclusivas.
    - **Variáveis quantitativas** – variáveis cuja a escala de medida permite a ordenação e quantificação de diferenças entre elas. Podem ser discretas ou contínuas.

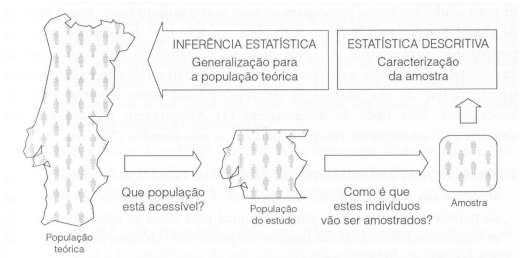
Basic Concepts

- **Measurements and data**

Níveis de medida	Exemplos	Procedimento de medida	Operações matemáticas permitidas
Nominal (qualitativa)	**Sexo, raça, religião, estado civil**	**Classificação por categorias**	**Contagens**
Ordinal (qualitativa)	**Escalas de opinião, atitude, classes social**	**Classificação por ranking de categorias**	**Maior que, igual, menor que**
Intervalo-rácio (quantitativa)	**Idade, nº de filhos, rendimento**	**Distância entre scores ou medidas em termos de unidades iguais**	**Todas as operações matemáticas**

Basic Concepts

- **Population and samples**
  - Chama-se **população** ao **conjunto de todos os valores que descrevem o fenómeno** que interessa ao investigador
  - Uma **amostra estatística** consiste de um **conjunto de indivíduos retirados de uma população** a fim de que o estudo estatístico dessa amostra possa fornecer informações cruciais sobre a população



Basic Concepts

- **Population and samples**
  - As **técnicas (ou métodos) de amostragem** dividem-se em 2 grupos:
    - Aleatórias ou probabilísticas
    - Não aleatórias
  - Uma amostra diz-se **aleatória ou probabilística** se for recolhida por um processo que assegura que todo e qualquer elemento (ou grupo de elementos) da população tem probabilidade, calculável e diferente de zero, de ser escolhido para integrar a amostra. Caso contrario diz-se **não aleatória**.
  - **Exemplo**: Há estudos em que se selecionam aleatoriamente escolas e depois dentro de cada escola selecionam-se aleatoriamente um membro (ou vários) para participar no estudo.

Basic Concepts

- **Population and samples**
  - **Vantagens** de uma amostragem ser aleatória:
    - **Não há subjetividade** ou o livre arbítrio do julgamento humano (tendência para escolher os mais disponíveis, os mais bem parecidos ou os mais simpáticos.
    - **Possibilidade de calcular a dimensão da amostra** bem como **a estimação da potência dos testes utilizados**.
  - **Dificuldades** de uma amostragem ser aleatória:
    - **Obter listagens ou registos** completos de população em estudo,
    - **Estabelecer contacto** com os potenciais elementos do estudo,
    - **Problema das não respostas** (questionários) e a taxa de participação

Basic Concepts

- **Population and samples**
  - **Vantagens de uma amostragem ser não aleatória**:
    - Fator subjetivo pode ser vantajoso na identificação de estratos ou clusters ou na definição de sub-grupos.
    - Os custos associados são mais reduzidos,
    - Permitem obter informação mais rapidamente e com menores necessidades de pessoal
  - **Desvantagens de uma amostragem ser não aleatória**:
    - Para além das razões do slide anterior, as conclusões não podem ser generalizadas para a população em estudo como acontece com as amostras aleatórias.

Basic Concepts

- **Population and samples: Descriptive vs inferential**
  - **Estatística descritiva** consiste na recolha, apresentação, análise e interpretação de dados através da criação de instrumentos adequados: quadros, gráficos e medidas de estatística descritiva (ex: médias, medianas, desvio-padrões…)
  - A estatística descritiva procura descrever ou sumariar a distribuição de uma variável ou descrever a relação entre 2 ou mais variáveis.
    - Exemplo: como descrever o rendimento familiar de 10000 famílias? E o rendimento familiar pelo nº de filhos por casal?
  - **Estatística inferencial** é quando pretende-se generalizar os resultados encontrados a partir de uma amostra aleatória para uma população (ex: através da realização de testes de hipóteses, intervalos de confiança…)
    - Exemplo: a altura média de uma população masculina é 1.75m.

Basic Concepts

- **Population and samples: Descriptive Statistics**
  - **Representação analítica**
    - **Medidas de localização** (uma indicação sobre a tendência central dos dados)
    - **Medidas de dispersão** (uma indicação sobre a variabilidade dos dados)
    - **Medidas de assimetria e achatamento** (indicação sobre onde as frequências mais altas estão localizadas)
    - **Proporções, percentagens, rácios e taxas**
  - **Representação tabelar/gráfica**
    - **Tabelas de frequência** (variáveis nominais, ordinais e quantitativas)
    - **Gráficos de barras e circulares** (variáveis nominais e ordinais)
    - **Gráficos de dispersão, gráficos de médias e desvio-padrões, histograma** (variáveis quantitativas)
    - **Caixa de bigodes** (variáveis ordinais e quantitativas)

Basic Concepts

- **Descriptive statistics: Exploring data**
  - Introduction
  - Describing data: measure of location
  - Describing data: measures of variability
  - Describing data: measures of shape
  - Displaying data graphically
  - Missing data
  - Data transformation

Basic Concepts

- **Introduction**
  - Os dados podem estar em diferentes formatos
  - Planear a introdução dos dados é fundamental
  - Dados qualitativos (nominais e ordinais)
    - Codificação das variáveis (uma possibilidade de resposta):
    - Exemplo: Género (1-Masc;2-Fem)
    - Exemplo: Satisfação (1-Mau;2-Aceitável; 3-Bom)
    - Codificação das variáveis (multiplas respostas de resposta):
    - Exemplo: Indique quantos sintomas sofre:

      Asma (1-Sim; 2- Não)

      Hypertensão (1-Sim; 2- Não)

      ...

Age group	gender	before diet	after diet
young	F	65	60
young	M	75	73
kid	M	35	35
adult	[illegible]	[illegible]	78
adult	[illegible]	[illegible]	60
kid	F	40	38
young	M	57	55
adult	M	65	62
adult	M	71	68

data file

Basic Concepts

- **Introduction**
  - Dados quantitativos:
    - Não há necessidade de codificação
    - Devem ser introduzidos com a mesma precisão que foram medidos e com a mesma unidade física associada (por ex: Kgs)
  - Identificação do paciente ou do nº do formulário ou do registo clinico
  - As datas devem ser introduzidas sempre da mesma forma na base de dados (por ex: dia/mês/ano)
  - Codificar os valores em falta
    - Exemplo para variáveis qualitiativas: 99-Não disponivel; 98- Não respondeu
    - Exemplo para variáveis quantitativas: Escolher um valor que não seja possivel ser obtido através da experiência

Basic Concepts

- **Describing data: measures of location**
  - **Fornecem** uma indicação sobre a tendência central dos dados
  - Média aritmétrica amostral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^*$
  - Mediana é definida como o valor que têm 50% da amostra à sua esquerda, após uma ordenação crescente dos dados:

$$Med(x) = \begin{cases} x_{\frac{n+1}{2}:n} & \text{se n impar} \\ \frac{x_{\frac{n}{2}:n} + x_{\frac{n}{2}+1:n}}{2} & \text{se n par} \end{cases}$$

  - Moda é defina como o valor mais frequente dessa amostra.

Basic Concepts

- **Describing data: measures of location**
  - Quantis ou percentis é definido como quantil ou percentil de ordem p (p ϵ [0;1]). O valor Qp detém a sua esquerda p.100% das observações que compõem a amostra:

$$Q_p = \begin{cases} x_{[n*p]+1:n} & \text{se } n*p \text{ não for inteiro} \\ \dfrac{x_{n*p:n} + x_{n*p+1:n}}{2} & \text{se } n*p \text{ for inteiro} \end{cases}$$

  - Quartis são os quantis de ordem p=1/4 (ou F_L), p=1/2 e p=3/4 (ou F_U).
  - Média geométrica deve-se utilizar quando os dados não são simétricos
  - Média pesada deve-se utilizar quando algumas observações são mais importantes que outras

Basic Concepts

- **Describing data: measures of variability**
  - Fornecem uma indicação sobre a dispersão (variabilidade) dos dados
  - Variância corrigida quantifica a variabilidade dos dados em torno da média:

$$s_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i^* - \bar{x})^2$$

  - Desvio-padrão corrigido (standard deviation) é dado por:

$$s_c(SD) = \sqrt{s_c^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - Estimativa do erro padrão da média (standard error of mean) é dado por:

$$SE = \frac{s_c}{\sqrt{n}}$$

Basic Concepts

- **Describing data: measures of variability**
  - Distância inter-quartil é dada pela diferença entre o quartil 75% menos o quartil 25%: IQR= $Q_{0.75}$-$Q_{0.25}$
  - Amplitude da amostra consiste na diferença entre o seu máximo e o seu mínimo: Amp=Max-Min
  - O coeficiente de dispersão (CV) é uma medida de dispersão relativa e é dado por:

$$CV = \frac{s_c}{\bar{x}}$$

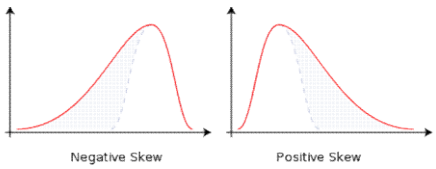
  - O coeficiente de dispersão é classificada como:
    - Dispersão fraca ≤ 15%
    - Dispersão média entre]15%;30%]
    - Dispersão elevada ≥30%

Basic Concepts

- **Describing data: measures of shape**
  - Coeficiente de assimetria (Skewness) e o seu desvio padrão (Skewness std error) são dados por:

$$B = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{s_c^3} = \frac{n\sum_{i=1}^{n}F_i(x_i-\bar{x})^3}{(n-1)(n-2)s_c^3};$$

$$Std\,Error\,B = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$



  - Se o rácio Skewness/Skewness std error for:
    - Entre [-1.96; +1.96], não se rejeita simetria.
    - Superior a +1.96, há evidências de assimetria positiva, ou seja as frequências mais altas tem tendência a estar no lado esquerdo do gráfico.
    - Inferior a -1.96, há evidências de assimetria negativa, ou seja as frequências mais altas tem tendência a estar no lado direito do gráfico.

Basic Concepts

- **Describing data:**
  - Dizem-se medidas robustas aquelas que não são afetadas por valores outliers e extremos Distância inter-quartil é dada pela diferença entre o quartil 75% menos o quartil 25%: $d_F = Q_{0.75} - Q_{0.25}$
  - Amplitude da amostra consiste na diferença entre o seu máximo e o seu mínimo: Amp=Max-Min
  - O coeficiente de dispersão (CV) é uma medida de dispersão relativa e é dado por:

$$CV = \frac{s_c}{\bar{x}}$$

  - O coeficiente de dispersão é classificada como:
    - Dispersão fraca ≤ 15%
    - Dispersão média entre]15%;30%]
    - Dispersão elevada ≥30%

#2 21

Basic Concepts

- **Describing data: Proportions and ratios**
  - Uma proporção (p) é definida como o nº de casos favoráveis (f) a dividir pelo nº de casos totais (N):

$$p = \frac{f}{N}$$

  - Uma percentagem (%) é uma proporção multiplicado por 100:

$$\% = \frac{f}{N} * 100 = p * 100$$

  - Um rácio é uma divisão entre os nºs de casos favoráveis de 2 categorias e são especialmente úteis para comparar categorias em termos de frequência relativa:

$$Ratio = \frac{f1}{f2}$$

  - Uma taxa é definida como o nº actual de ocorrências de um fenómeno a dividir pelo o nº total de ocorrências por unidade de tempo.

#2 22

Basic Concepts

- **Describing data: Outliers and extreme values**
  - Valores outliers são observações que são distintas da maioria dos restantes dados.
  - Este valores poderão ser genuínos, mas poderão também ser resultados devido a erros de equipamento ou inserção de dados.
  - Outliers são classificados como moderados ou severos (alguns autores definem como valores extremos) e são facilmente identificados através da caixa de bigodes.
  - Quando existem, deve-se fazer uma análise com e sem outliers:
    - Se os resultados forem semelhantes então os outliers tem pouca influência e devem ficar. Nesta situação os valores devem ser reportados em termos de Média ± Desvio-Padrão (estatísticas não robustas)
    - Se os resultados forem diferentes, então os outliers devem ser removidos.
    - Existem testes estatísticos que não são influenciáveis por outliers. Neste caso os valores devem ser reportados em termos de Medianas e amplitude inter-quartis (estatísticas robustas)

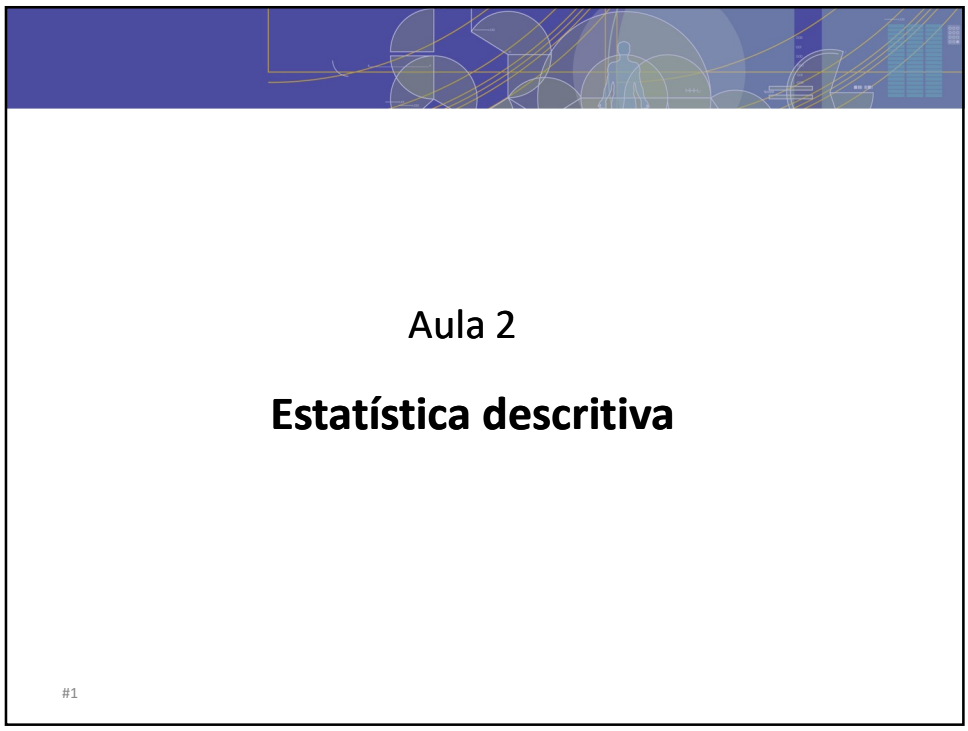
#2 23

Basic Concepts

- **Describing data: Resume**
  - Variável nominal:
    - Moda, proporções
  - Variável ordinal:
    - Moda, proporções
    - Estatísticas de ordem: Mediana , quartis, quantis,…
    - Amplitude inter-quartil
  - Variável quantitativa:
    - Moda
    - Estatísticas de ordem: Mediana , quartis, quantis,…
    - Amplitude inter-quartil
    - Amplitude total
    - Média
    - Desvio padrão e variância
    - Coeficiente de variação
    - Coeficiente de assimetria
    - …

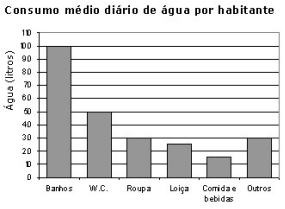
24


1.0
Introdução à análise estatística com SPSS
Conceitos básicos
Pedro Sá Couto
Departamento de Matemática
Universidade de Aveiro
#1

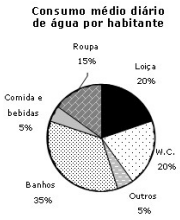

Aula 2
Estatística descritiva
#1

Basic Concepts

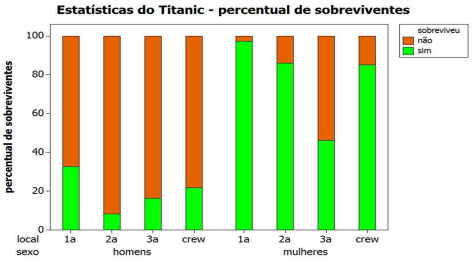
- **Displaying data graphically**
  - Gráfico de barras e gráficos circulares (variáveis nominais ou ordinais)



Jornal *Alnia*

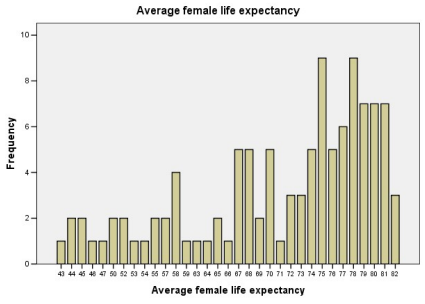


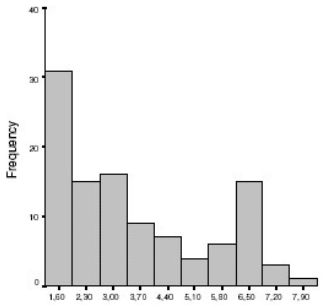
Jornal *Belnia*



Basic Concepts

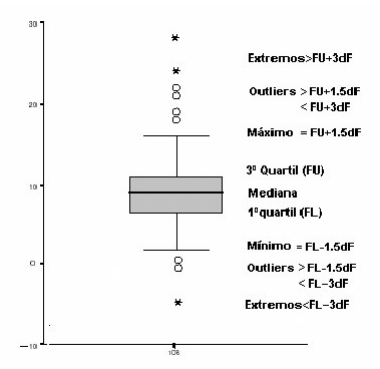
- **Displaying data graphically**
  - Histograma (variáveis quantitativas)

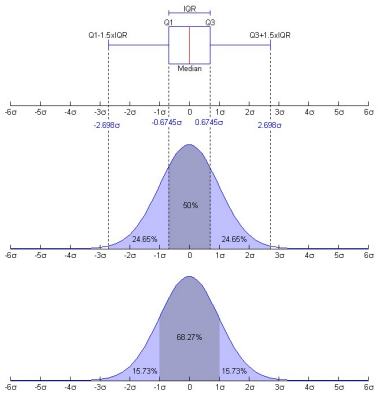




Basic Concepts

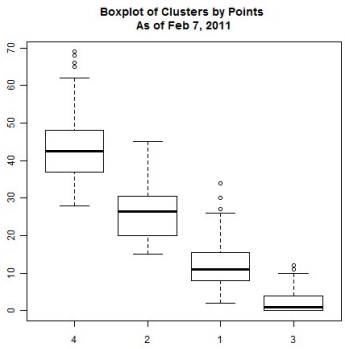
- **Displaying data graphically**
  - Caixa de bigodes (variáveis ordinais e quantitativas)

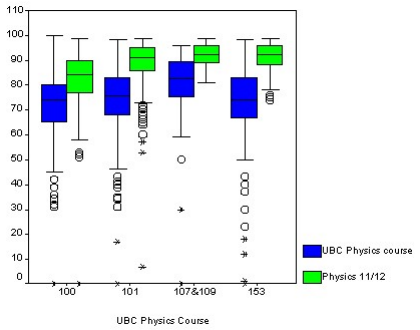




Basic Concepts

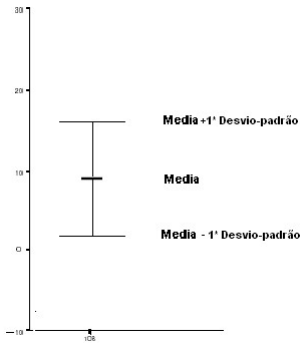
- **Displaying data graphically**
  - Caixa de bigodes (variáveis ordinais e quantitativas)

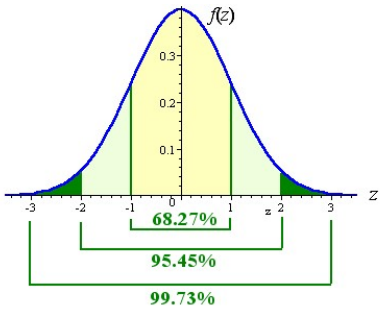




Basic Concepts

- **Displaying data graphically**
  - Médias e desvios-padrão (variáveis quantitativas)

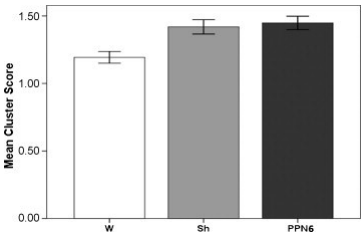


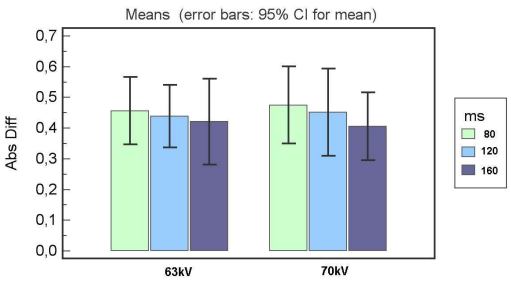


#2 31

Basic Concepts

- **Displaying data graphically**
  - Médias e desvios-padrão (variáveis quantitativas)

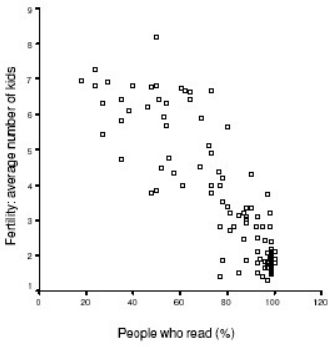


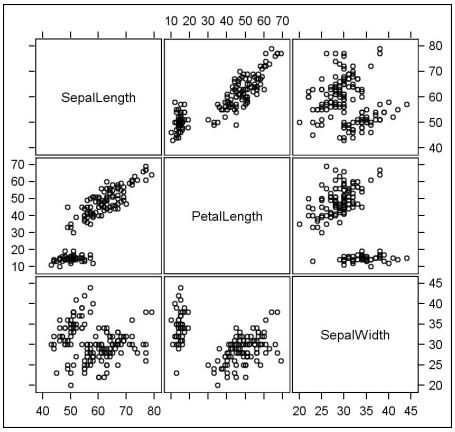


#2 32

Basic Concepts

- **Displaying data graphically**
  - Gráficos de dispersão (2 ou + variáveis quantitativas)





#2 33

1.0

Introdução à análise estatística com SPSS
Conceitos básicos

Pedro Sá Couto
Departamento de Matemática
Universidade de Aveiro

#1

Aula 3

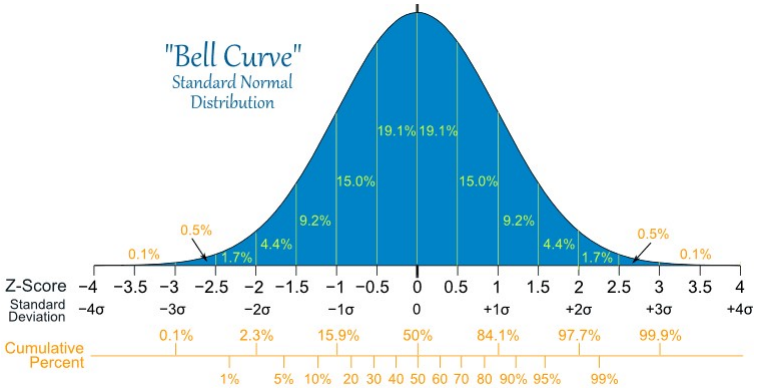
Estatística Inferêncial

Basic Concepts

- **Theoretical distributions: Normal distribution and others**
  - **Parâmetros de uma distribuição:**
    - **Valor esperado (E[X])** é um parâmetro de **localização** que nos dá uma ideia da **tendência central da distribuição** de uma variável aleatória X
    - **Variância (Var[X])** é um parâmetro de **dispersão** que nos dá uma ideia sobre a **variabilidade/dispersão da distribuição** de uma variável aleatória X.
  - **Propriedades:**
    - $E[a]=a$
    - $E[aX+b] = a.E[X]+b$
    - $E[X+Y] = E[X]+E[Y]$; Se X e Y forem independentes: $E[X.Y] = E[X].E[Y]$
    - $Var[b] = 0$,
    - $Var[aX+b] = a^2\ Var[X]$
    - $Var[X]=E[X^2]-(E[X])^2$

Basic Concepts

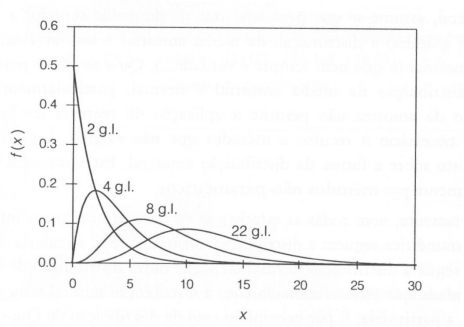
- **Theoretical distributions: Normal distribution and others**
  - **Distribuição Normal ou Gaussiana**: $X \sim N(\mu, \sigma^2)$



  - **Propriedades:**
    - $E[X]=\mu$; $V[X]=\sigma^2$

Basic Concepts

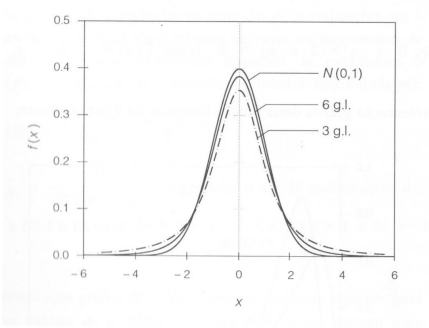
- **Theoretical distributions: Normal distribution and others**
  - **Distribuição Qui-quadrado** com n graus de liberdade: $X \sim \chi^2(n)$



  - **Propriedades:**
    - $E[X]=n$
    - $V[X]=2*n$
    - Quanto maior for os graus de liberdade, mais a distribuição do qui-quadrado se aproxima da distribuição Normal

Basic Concepts

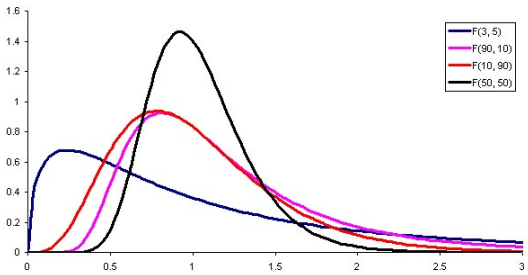
- **Theoretical distributions: Normal distribution and others**
  - **Distribuição t-Student** com n-1 graus de liberdade: X~t(n-1)



  - **Propriedades:**
    - E[X]=0
    - V[X]=n/(n-2) para n>2
    - Quanto maior for os graus de liberdade, mais a distribuição do t-Student se aproxima da distribuição Normal

Basic Concepts

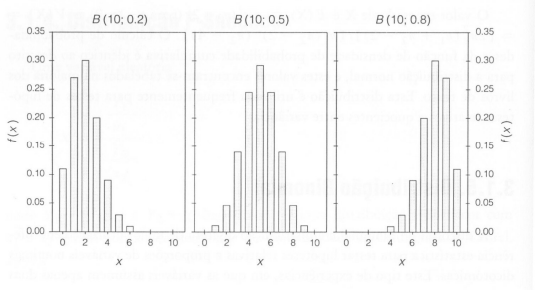
- **Theoretical distributions: Normal distribution and others**
  - **Distribuição F-Snedecor**: X~F(n_1,n_2)



  - **Propriedades:**
    - E[X]=n_2/(n_2-2) para n_2>2
    - V[X]=(2n_2^2(n_1+n_2-2))/(n_1(n_2-2)^2(n_2-4))
    - Quanto maior for os graus de liberdade, mais a distribuição do F-Snedecor se aproxima da distribuição Normal

Basic Concepts

- **Theoretical distributions: Normal distribution and others**
  - Distribuição Binomial: X~Bi(n,p), onde n é o nº total de experiência e p é a probabilidade de se obter um sucesso



  - **Propriedades:**
    - E[X]=n*p
    - V[X]=n*p*(1-p)
    - Se n>20 e n*p>7 a distribuição do Binomial pode ser aproximada por uma distribuição Normal

Basic Concepts

- **Statistical inference: Point estimation**
  - Como encontrar estimadores para os parâmetros da população?
  - Um das formas é o método dos momentos:
    - Quando **há só um parâmetro da população desconhecido** fica-se com uma só uma equação:

$$E[X] = \overline{X}$$

    - Quando **há só dois parâmetros desconhecidos** é usual utilizar o sistema equivalente:

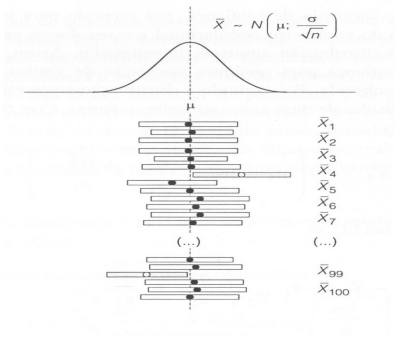
$$\begin{cases} E[X] = \overline{X} \\ Var[X] = S^2 \end{cases}$$

    - Exemplo: Se tivermos amostras aleatórias em que sabemos que E[X]=μ e a Var[X]=σ², encontre os estimadores para μ e σ²:

$$\begin{cases} E[X] = \overline{X} \\ Var[X] = S^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = S^2 \end{cases}$$

Basic Concepts

- **Statistical inference: Intervalar estimation**
  - Um **intervalo de confiança (IC)** para um parâmetro , a um nível de confiança 1 - α é um intervalo aleatório (θ_1, θ_2) tal que: $P(\theta_1 < \theta < \theta_2) = 1-\alpha$, **$\alpha$ é um valor reduzido** para termos confiança elevada e designa-se por **nível de significância.**
  - Os **ICs mais utilizados são de 90%, 95% e 99%,** que correspondem a um α de 10% (ou 0.1), 5% (ou 0.05) e 1% (ou 0.01), respetivamente.
  - **Exemplo**: Um **intervalo de confiança a 95% para μ** significa que em cada 100 intervalos obtidos de 100 amostras aleatórias, 95 destes intervalos possuirão o verdadeiro valor de μ. No entanto, **o seu verdadeiro valor nunca será conhecido**. A interpretação gráfica deste exemplo:



#1

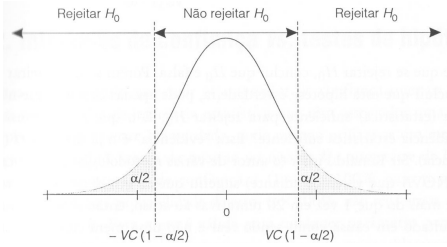
Basic Concepts

- **Statistical inference: Statistical hypothesis**
  - Os **testes de hipóteses (TH) contribuem para a tomada de decisões**.
  - Num **TH há sempre um par de hipóteses**:
    - Hipótese nula (H0) vs Hipótese alternativa (H1)
  - A **tomada de decisão (rejeição ou não rejeição da hipótese nula)** será então baseada **na análise de uma amostra aleatória dessa população.** Os **testes estatísticos** são sempre realizados **sobre H0**
  - **Exemplos** de THs**:**
    1. H0: $\mu = 0$ vs H1: $\mu \neq 0$
    2. H0: $\mu \leq 3$ vs H1: $\mu > 3$
    3. H0: $\sigma^2 \geq 1$ vs H1: $\sigma^2 < 1$
    4. H0: $\sigma^2 = 1$ vs H1: $\sigma^2 \neq 1$
  - **Tipos de testes:** Os testes podem ser bilaterais (1 e 4) ou unilaterais (2 e 3)

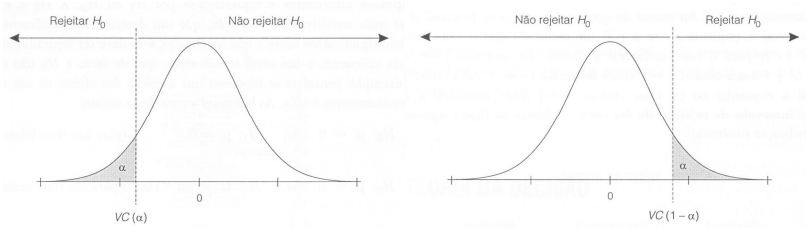
#1 44

Basic Concepts

- **Statistical hypothesis: Critical region**
  - TH bilaterais:



  - TH unilaterais á esquerda e á direita:



Basic Concepts

- **Statistical hypothesis: P-value**
  - Definição de **p-value** (ou abreviado p) do teste: Ao **menor valor de α a partir do qual se rejeita H0** chama-se **probabilidade de significância** (p-value).
  - Este valor **representa uma medida complementar do grau de certeza** a partir do qual assumimos **como real** (representativo da população) **o resultado** (ou estatística) obtido no estudo.
  - Outra def. usual para p-values é a **probabilidade dos resultados serem atribuídos por sorte ou por erro aleatório**:
  - Qualquer que seja a definição as seguintes regras são válidas:
    - Se o valor do **p-value for muito pequeno**, concluímos que **os resultado são significativos, ou seja, rejeitamos H0 (p-value < α)**.
    - Caso contrário, **os resultados não serão significativos** o que **leva a não rejeição de H0 (p-value > α)**.

Basic Concepts

- **Statistical hypothesis: Errors in hypothesis testing**
  - Num TH, as hipóteses são geralmente colocadas da seguinte forma:
    - H0: não tem doença/diferenças,… vs H1: têm doença, diferenças,…
    - ou seja, o que se pretende provar coloca-se em H1, sendo o H0 sempre o seu complementar. Para provar o que pretende, o investigador tem de ter provas/evidências (estatísticas) para rejeitar H0.
  - Exemplo: Nos tribunais, usa-se o mesmo principio que o individuo não é culpado (não se rejeita H0) até prova em contrário (rejeita-se H0).
  - Nos **testes de hipóteses** baseados neste principio, existem dois tipos de erros:
    - **Erro de tipo I**, rejeitar H0 (decisão do teste) sendo H0 verdadeira (situação real), está associado aos **falsos positivos**
    - **Erro de tipo II**, não rejeitar H0 (decisão do teste) sendo H0 falso (situação real), está associado aos **falsos negativos;**

#1 47

Basic Concepts

- **Statistical hypothesis: Errors in hypothesis testing**
  - Tabularmente:

Decisão do teste	Situação real	
	H_0 verdadeira	H_0 falsa (H1)
Rejeito H_0	Erro de tipo I	Decisão correcta
Não rejeito H_0	Decisão correcta	Erro de tipo II

$$P(erro\ tipo\ I) = P(\text{Rejeitar } H0 \mid H0\ verdadeiro) = \alpha$$

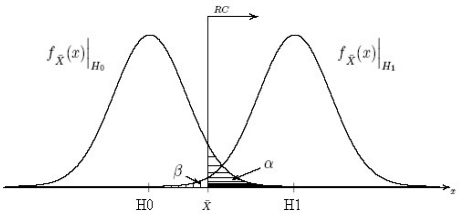
$$P(erro\ tipo\ II) = P(Não\ \text{Rejeitar } H0 \mid H0\ falso) = \beta$$

  - Os níveis de significância (α) definem-se á partida e os mais usuais são 0.1, 0.05 e 0.01, desta forma minimizando a probabilidade do erro tipo I
  - **Exemplo:** Um teste de HIV acusa positivo (rejeitou H0), mas na realidade o sujeito não tem HIV (mas não devia ter rejeitado H0), ou seja, um falso positivo.
  - **Exemplo:** Um teste de HIV acusa negativo (não rejeitou H0), mas na realidade o sujeito tem HIV (mas devia ter rejeitado H0), ou seja, um falso negativo.

48

Basic Concepts

- **Statistical hypothesis: Errors in hypothesis testing**
  - Qual dos dois erros é mais perigoso?
    - Um **erro tipo I** muito pequeno é necessário quando o tratamento ou diagnóstico é potencialmente perigoso para o paciente (mentalmente ou fisicamente).
    - Um **erro tipo II** muito baixo é necessário quando o tratamento e o diagnóstico precoce são benéficos e quando a doença é contagiosa.
  - Seria desejável que α e β fossem o mais pequenos possível. Os valores de α e β variam em relação inversa, ou seja, se diminuir o valor de α então o valor de β irá ser maior e vice-versa. A única forma de diminuir α e β simultaneamente é aumentar o tamanho da amostra.



#1 49

1.0

Introdução à análise estatística com SPSS
Conceitos básicos

Pedro Sá Couto
Departamento de Matemática
Universidade de Aveiro

#1

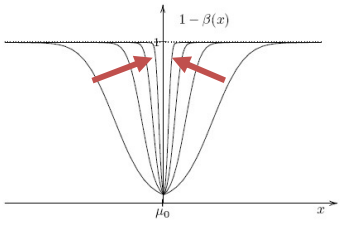
Aula 4

Estatística Inferêncial

Basic Concepts

- **Statistical power and sample size**
  - A **potência do teste** ($\pi(\theta)$) é dada por:

$$\pi(\theta) = 1 - \beta = 1 - P(\text{Não Rejeitar } H0 / H0 \ falso)$$



  - O **objectivo de um estudo** é a obtenção de uma curva da função **potência que seja um vale estreito e abrupto**. Assim, para pequenos desvios de H0, o valor da potência será muito elevado e por conseguinte representará um erro tipo II bastante reduzido

Basic Concepts

- **Statistical power and sample size**
  - **Fatores** que condicionam a **potência de um teste**:
    - A escolha do **valor do erro tipo I**. Se diminuir o valor de α então o valor de β irá ser maior (e a potência do teste menor) e vice-versa.
    - A **distância entre** os valores definidos para **H0 e H1**. Quanto maior for a diferença entre os valores do parâmetro considerados nas hipóteses H0 e H1, mais fácil é detetar qual das hipóteses é verdadeira e portanto menor será a probabilidade de errar e maior será a potência do teste.
    - O valor da **variabilidade do estudo**. Uma variabilidade elevada resulta sempre numa potência reduzida, dado o grau de incerteza que daí resulta.
    - **Dimensão da amostra**. Quando a dimensão da amostra aumenta, a variabilidade diminui e as curvas da função potência tornam-se cada vez mais próximas do ideal, ou seja, um vale estreito e abrupto.

Basic Concepts

- **Statistical power and sample size**
  - Uma parte essencial do planeamento de qualquer investigação é a **decisão de quantas amostras o vosso estudo necessita**.
  - Muitas vezes os **estudos têm uma dimensão demasiado pequena ou demasiado grande** porque a dimensão da amostra foi escolhida por **motivos de logística ou por comparação de outros trabalhos**.
  - As **vantagens** deste procedimento:
    - Menor perda de tempo,
    - Risco potencial para os participantes do estudo é menor,
    - Erro tipo I e tipo II estão controlados
    - Pouca recursos.

Basic Concepts

- **Statistical power and sample size**
  - Para o **cálculo da dimensão da amostra é necessário**:
    - **Especificar o valor do erro tipo I (α).** Quanto menor for o valor de α (menor será o erro cometido), maior será o valor da dimensão da amostra.
    - **Especificar a potência** que pretende-se atingir para se observar um "verdadeiro efeito". Quanto maior for o valor da potência (e por conseguinte, menor será o erro tipo II), maior será a dimensão da amostra.
    - **Especificar o valor da variabilidade** que se irá observar. Quanto maior for a variabilidade, maior terá de ser a dimensão da amostra. Esta variável é a mais difícil de estabelecer. Geralmente recorre-se a estudos semelhantes ou a estudos pilotos para ter uma noção do seu valor.
    - **Especificar a diferença** que se pretende observar entre os valores escolhidos **para H0 e H1**. Quanto menor for esta diferença, maior terá de ser a dimensão da amostra.

55

Basic Concepts

- **Statistical power and sample size**
  - A estimação dos dois últimos pontos pode ser muito problemática. Alternativamente, pode-se utilizar o conceito designando por **tamanho do efeito (d - effect size).** Este conceito relaciona os dois últimos itens num único cálculo:

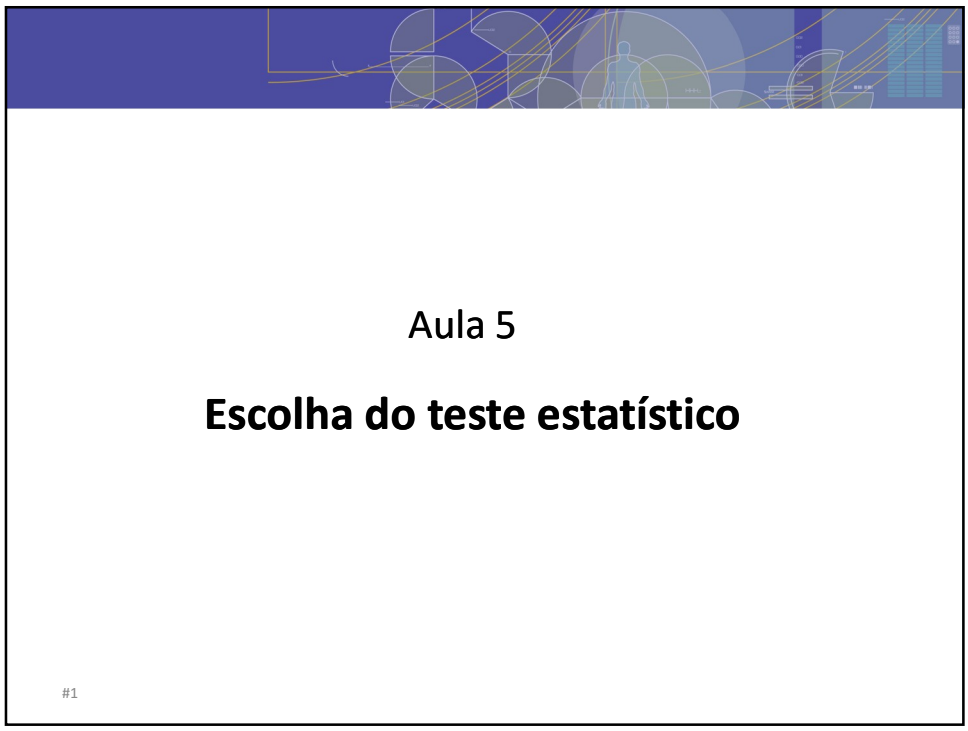
$$d = \frac{\mu_0 - \mu_1}{\sigma}$$

  - Exemplo: Para os testes baseados na distribuição t-Student existem **3 categorias** para o tamanho do efeito:
    - Pequeno ($0.2 \leq d < 0.5$)
    - Médio ($0.5 \leq d < 0.8$)
    - Grande ($d \geq 0.8$)
  - Quanto **menor for o tamanho do efeito** pretendido ou especificado, **maior será o tamanho da amostra**.

#1

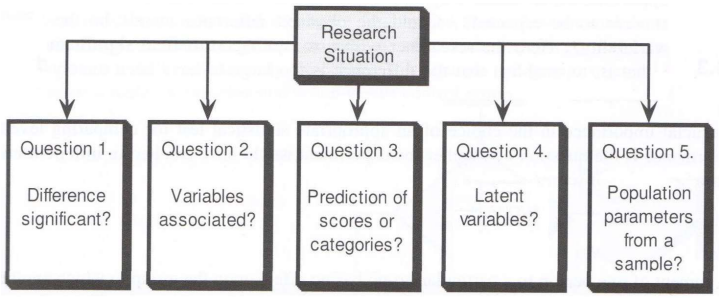
56


1.0
Introdução à análise estatística com SPSS
Conceitos básicos
Pedro Sá Couto
Departamento de Matemática
Universidade de Aveiro
#1


Aula 5
Escolha do teste estatístico
#1

Basic Concepts

- **Choosing a statistical test**
  - A escolha de um teste estatístico depende de várias considerações: A **questão de investigação, o plano** ou do **desenho da experiência e a natureza dos dados** (nível de medida) que se pretende analisar
  - Graficamente, as questões de investigação podem ser divididas nas seguintes categorias:



#1 59

Basic Concepts

- **Choosing a statistical test**
  - Em relação ás **questões de investigação**, geralmente podemos agrupar **em 5 grandes níveis**:
    - A **diferença entre médias/proporções/medianas é significativa**?
    - Exemplo: Será que o batimento cardíaco é o mesmo antes ou depois de um curso de relaxamento?
    - Como **as variáveis X e Y estão associadas**?
    - Exemplo: será que batimento cardíaco está relacionado com a temperatura?
    - Será que é possível **realizar previsões sobre uma determinada variável a partir de outras variáveis**?
    - Exemplo: será que a performance universitária pode ser predita através das pontuações obtidas em testes de aptidão?

60

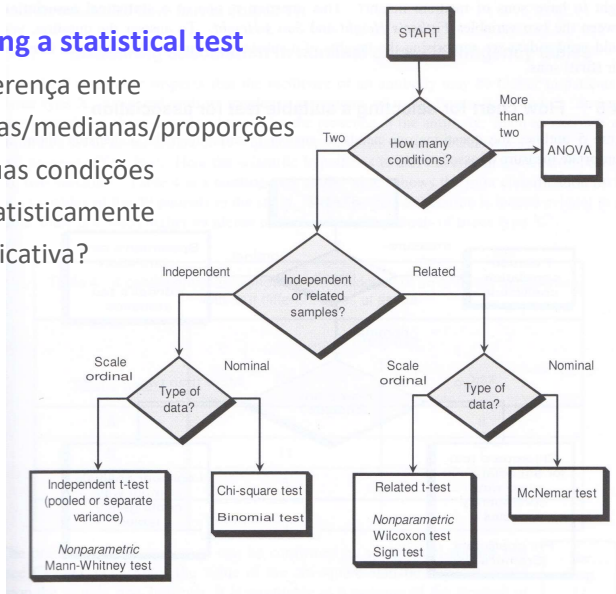
Basic Concepts

- **Choosing a statistical test**
  - Em relação ás **questões de investigação**, geralmente podemos agrupar **em 5 grandes níveis**:
    - Será possível observar **variáveis ou fatores latentes** que estão por detrás dos resultados obtidos?
    - Exemplo: Utilizou-se um conjunto de itens para medir a depressão. Será que podemos reduzir este conjunto de itens em poucas dimensões ou fatores latentes?
    - Técnicas como análise fatorial exploratória ou análise fatorial confirmatória não irão ser lecionadas neste curso
    - Os **parâmetros da população** são refletidos nos dados recolhidos da amostra?
    - Exemplo: Obteve-se a pontuação do coeficiente de inteligência de 100 crianças com uma determinada idade. Será que os valores medidos estão de acordo com a literatura para esta faixa etária?

61

Basic Concepts

- **Choosing a statistical test**
  - A diferença entre médias/medianas/proporções de duas condições é estatisticamente significativa?



#1

62

Basic Concepts

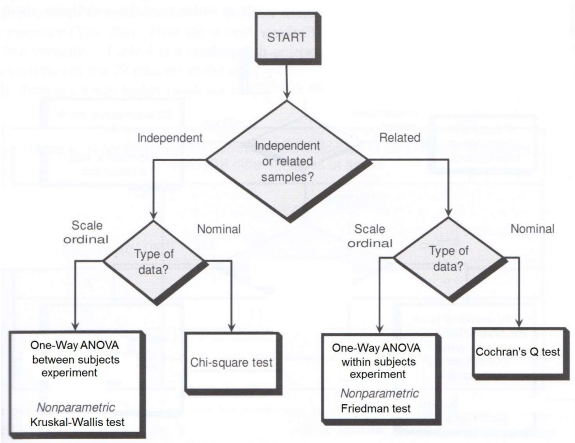
- **Choosing a statistical test**
  - Nas **amostras independentes**, os elementos que pertencem a cada condição/grupo são diferentes. Exemplo: Será que o peso dos homens é diferente das mulheres?
  - Nas **amostras emparelhadas** os mesmos indivíduos podem ser medidos em várias situações experimentais, e neste caso, as amostras dizem-se de medições repetidas ou emparelhadas. Exemplo: Será que o peso antes e depois da dieta é diferente?
  - A principal **vantagem das amostras emparelhadas** consiste no **controlo** que asseguram sobre as **diferenças individuais** existente entre sujeitos, levando a uma redução da variabilidade associada ás diferenças individuais.
  - As **desvantagens das amostras emparelhadas** são associadas ao problema da **ordem na qual se apresentam os tratamentos** (as repercussões de A sobre B não podem ser as mesmas que na sequência inversa), mesmo com periodos de wash-out entre A e B.

Basic Concepts

- **Choosing a statistical test**
  - Os testes paramétricos exigem que a forma da distribuição amostral seja conhecida (a distribuição Normal é a mais conhecida).
  - Os testes não paramétricos não exigem o conhecimento da distribuição amostral e são uma alternativa aos testes paramétricos.
  - Mostra-se que de uma forma geral a **potência** de um teste **paramétrico é superior a um teste não paramétrico**.
  - Deve-se usar um teste não paramétrico sempre que:
    - **Não é possível** demonstrar que os dados quantitativos têm uma **distribuição amostral conhecida**
    - **A dimensão da amostra é reduzida**
    - A **escala de medida** é uma variável **qualitativa (nominal ou ordinal)**

Basic Concepts

- **Choosing a statistical test**
  - A diferença entre médias/medianas/proporções com mais do que duas condições é estatisticamente significativa (ANOVAS de um factor)?



65

Basic Concepts

- **Choosing a statistical test**
  - **As ANOVAS** podem ter **mais do que um fator.** Chama-se um fator á generalização das condições existentes numa ANOVA.
  - **Exemplo:** Fator idade: jovem, adulto, sénior; Fator género: Masculino/feminino
  - **Exemplo: Anova de um fator de amostras independentes**: "Um estudo dividiu 150 sujeitos em 3 grupos mediante o seu peso (reduzido, normal, elevado) e mediu-se a frequência cardíaca".
  - O fator é o peso categorizado (reduzido, normal, elevado) enquanto a frequência cardíaca é a variável dependente ou medida.
  - **Exemplo: Anova de dois fatores de amostras independentes**: "Um estudo dividiu 150 sujeitos em 3 grupos mediante o seu peso (normal, excesso, mórbido) e género (Masculino/Feminino). De seguida mediu-se a frequência cardíaca".
  - O fatores são o peso (3 níveis) e o género (2 níveis) enquanto a frequência cardíaca é a variável dependente ou medida.

#1

66

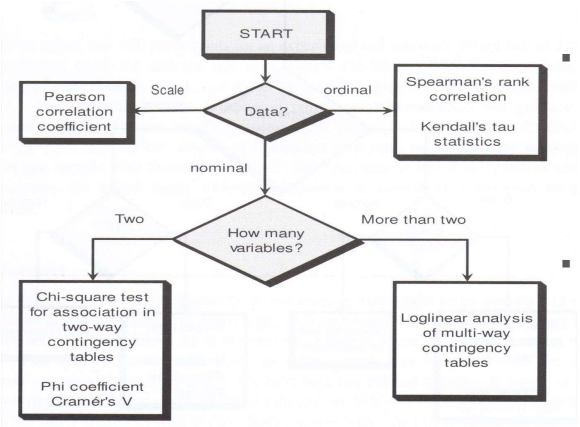
Basic Concepts

- **Choosing a statistical test**
  - **Exemplo: Anova de dois fatores de amostras mistas**: "Um estudo com 30 sujeitos controlou o peso em 3 diferentes ocasiões (Inicio, Fim, Follow up) e o género (Masculino/Feminino). De seguida mediu-se a frequência cardíaca".
  - Os fatores são o género (2 níveis, amostras independentes) e o peso (3 níveis, amostras repetidas ou emparelhadas) enquanto a frequência cardíaca é a variável dependente ou medida.
  - **Exemplo: Anova de um fator de medidas repetidas**: "Um estudo com 30 sujeitos controlou o peso em 3 diferentes ocasiões (Inicio, Fim, Follow up).
  - O fator é ocasião (3 níveis) e a frequência cardíaca é a variável dependente.
  - **Exemplo: Anova de dois fatores de amostras repetidas:** "Um estudo com dividiu 30 sujeitos em 3 regimes alimentares (A, B, C) e 2 tipos de exercícios (calmo, intenso) diferentes. Todos os sujeitos passaram por todos os regimes alimentares e tipos de exercício. Em todos os momentos mediu-se a frequência cardíaca".
  - O fatores são o regime alimentar (3 níveis) e o tipo de exercício (2 níveis) enquanto a frequência cardíaca é a variável dependente ou medida.

67

Basic Concepts

- **Choosing a statistical test**
  - Será que as variáveis estão associadas/correlacionadas?



- Para aplicar a correlação de Pearson ambas as variáveis devem ter:
  - Distribuição Normal
  - Linearidade entre si
- As estatísticas de Kendall's tau estão mais relacionadas com questões de concordância/fiabilidade do que com correlação

#1

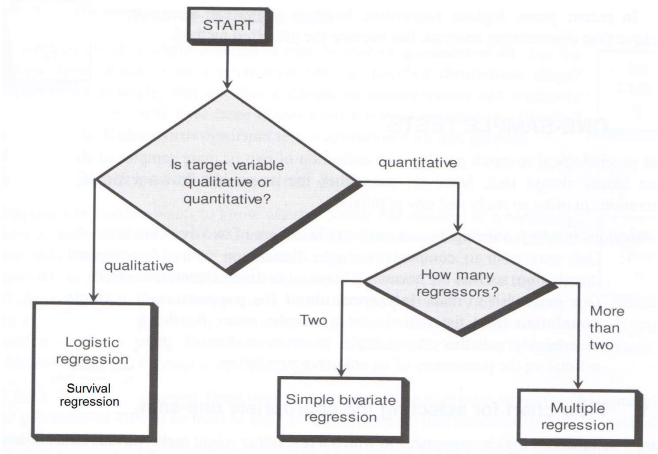
68

Basic Concepts

- **Choosing a statistical test**
  - **Exemplo: Correlação (Pearson/Spearman)**: "Um estudo procurou uma relação de associação entre o peso e álcool consumido diariamente (medido em litros). Existirá uma relação entre estas duas variáveis?
  - **Exemplo: Concordância (Kendall tau):** Pediu-se a dois juízes para classificarem 20 trabalhos realizados por alunos. Será que os juízes pontuam da mesma maneira?
  - **Exemplo: Associação de variáveis nominais (chi-quadrado):** Um anticorpo parece estar associado a um determinado tipo de tecido muscular X. O investigador recolheu amostras de vários tipos de tecido (A, B, C, X) e testou se o anticorpo estava presente ou ausente. Será que existe uma relação entre o tipo de tecido (A, B, C e D) e o anticorpo (presente/ausente)

Basic Concepts

- **Choosing a statistical test**
  - Será que é possível construir um modelo de previsão?



Basic Concepts

- **Choosing a statistical test**
  - Será que é possível construir um modelo de previsão?
  - Variáveis independentes vs variáveis dependentes
    - Uma **variável dependente** é definida como aquela que resulta ou poderá resultar de uma **combinação de variáveis independentes**.
    - Um **conjunto de variáveis independentes não têm qualquer associação estatística entre si**.
  - Se a **variável dependente for quantitativa**, utiliza-se uma **regressão linear simples** (se houver uma variável dependente e uma independente) ou **múltipla** (se houver uma variável dependente e várias independentes)
  - Se a **variável dependente for qualitativa binária (2 níveis)**, utiliza-se uma **regressão logística ou regressão de análise de sobrevivência**.
  - As **variáveis independentes** podem ser **quantitativas e qualitativas (através da utilização de dummy variables)**

#1 71

Basic Concepts

- **Choosing a statistical test**
  - **Exemplo: Regressão múltipla:** "Um estudo procura um modelo de previsão entre um determinado end-point (ex: frequência cardíaca) e um conjunto de biomarcadores fisiológicos". Será possível estabelecer uma modelo de previsão?
  - **Exemplo: Regressão logística binária:** Um estudo tenta prever se admissão de um licenciado (sim/não) numa determinada instituição depende de um conjunto de fatores como a sua média final de curso, o seu QI, atividades curriculares, atividades extra-curriculares. Quais serão as variáveis mais relevantes para a previsão?
  - **Exemplo: Análise de sobrevivência:** Num estudo mediu-se o tempo de sobrevivência de pacientes que estavam num programa experimental em dois grupos: controle e experimental. A variável dependente é se sobreviveu ao fim de um ano de seguimento (Sim/Não). Que fatores mais influenciaram nos resultados?

#1 72

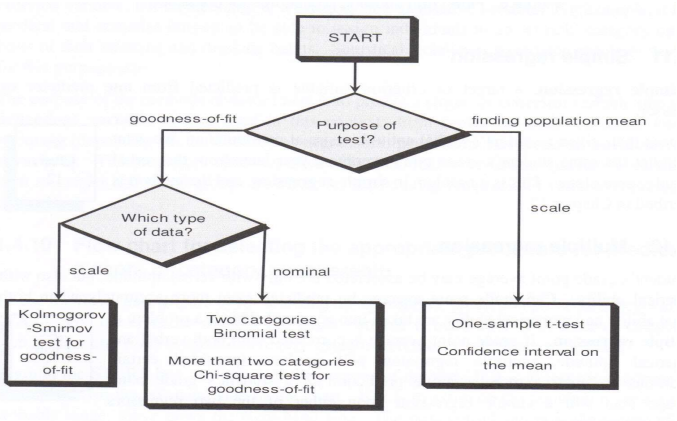
Choosing a statistical test

- **Procurando variáveis latentes ou fatores**
  - **Análise factorial** é uma técnica baseada na **matriz de correlações**, que tenta **agrupar ou classificar um conjunto grande de itens num conjunto relativamente pequeno de dimensões latentes ou fatores**:
    - Na **análise exploratória fatorial**, o objectivo é encontrar um número mínimo de fatores que contabilizem o máximo de correlação existente entre os itens.
    - Na **análise confirmatória fatorial**, modelos específicos são testados entre si, de forma a encontrar o melhor modelo para aquele conjunto de dados.
  - **Estatística multivariada** são um conjunto de métodos desenhados para a análise de dados multivariados onde existem duas ou mais variáveis dependentes:
    - Em investigação experimental, **os testes t-student e as ANOVAS são generalizadas em análise de variância multivariada** (MANOVAS) ou **análise de co-variância multivariadas** (MANCOVAS) no caso de existirem co-variantes.

73

Basic Concepts

- **Choosing a statistical test**
  - Testes sobre uma amostra.

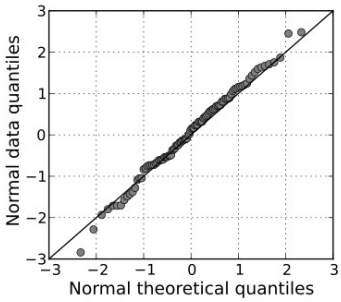


#1

74

Basic Concepts

- **Choosing a statistical test**
  - Testes sobre uma amostra.
  - **Testes de ajustamento (goodness of fit)** permitem averiguar **se uma distribuição amostral é próxima de uma distribuição teórica conhecida**.
  - Para demonstrar a **normalidade de um conjunto de dados** utiliza-se o **teste Kolmogorov-Smirnov** ou o **teste Shapiro-Wilks** (usado preferencialmente para amostras reduzidas, n<30)
  - Para **dimensões muito elevadas (n>100),** estes testes de ajustamentos devem ser substituídos por **QQ plots,** onde visualmente se avalia se a distribuição amostral se ajusta á distribuição teórica.
  - **Inferência sobre uma população** é útil quando se pretende saber **se um valor obtido pela estimação pontual** ou pela **estimação intervalar tem significado estatístico ou não**.



#1

Basic Concepts

- **Choosing a statistical test**
  - **Exemplo: Normalidade (teste K-S; S-W):** "Será que o conjunto de amostras obtidas poderão ser ajustada a uma população que tem distribuição Normal?"
  - **Exemplo: Binomial (teste chi-quadrado):** "Será que o conjunto de amostras obtidas poderão ser ajustada a uma população que tem distribuição Binomial?"
  - **Exemplo: Parâmetros populacionais:** Obteve-se a pontuação do coeficiente de inteligência de 100 crianças com uma determinada idade. Será que os valores medidos estão de acordo com os valores populacionais referidos na literatura para esta faixa etária? Qual o seu intervalo de confiança?

#1 76


1.0
Introdução à análise estatística com SPSS
Conceitos básicos
Pedro Sá Couto
Departamento de Matemática
Universidade de Aveiro
#1