



Engenharia de Dados Experimentais

Inferência Estatística

Raquel Guiné



Conteúdo

1. Objetivos da inferência estatística
2. Testes de hipóteses
3. Exemplos de seleção de testes



1. OBJETIVOS DA INFERÊNCIA ESTATÍSTICA

- A estatística inferencial permite, a partir dos dados obtidos para a amostra, generalizar no que respeita às características da população.
- A esta extrapolação está associado sempre um certo grau de erro.
- Os principais **objetivos** são:
 - a) Avaliar parâmetros
 - b) Avaliar relações
 - c) Fazer previsões
 - d) Avaliar diferenças entre amostras



a) Avaliar parâmetros

- Em que medidas os valores estimados para a amostra são ou não representativos da população?

Amostra:

\bar{X}



s





População:

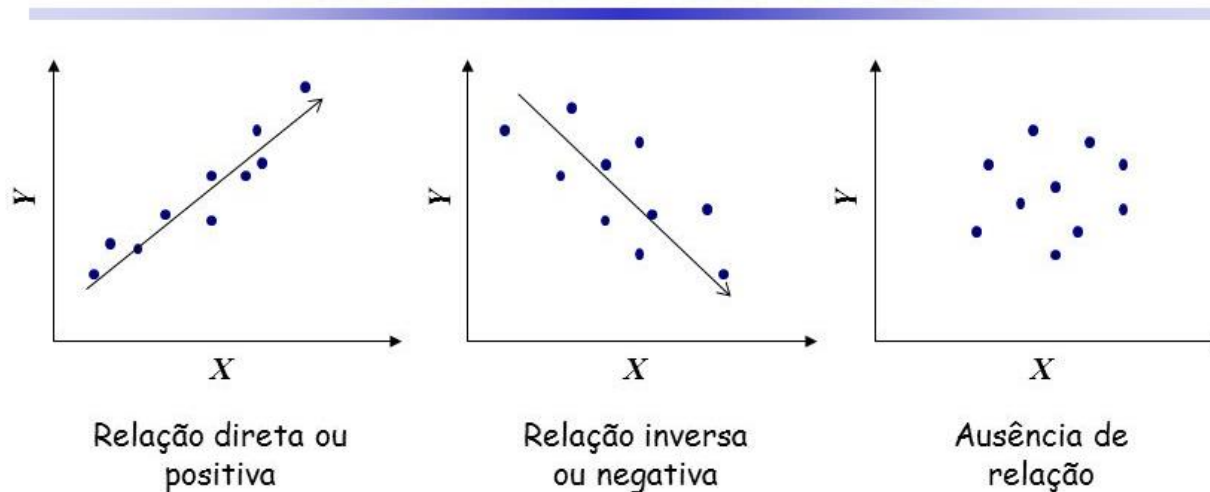
μ

σ



a) Avaliar relações

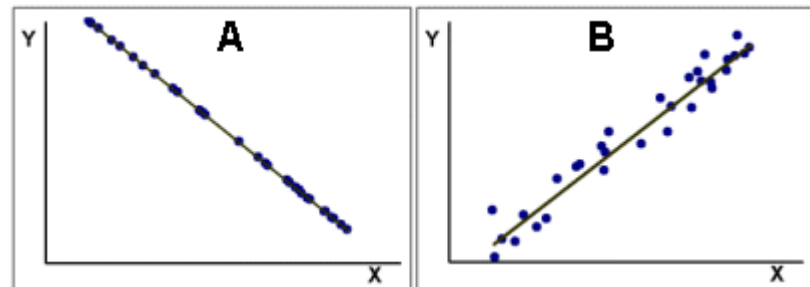
- Uma correlação caracteriza-se por um sinal e um valor numérico.
- No que respeita ao sinal, as correlações (r) podem ser positivas (+) ou negativas (-):
 - $r > 0$  ambas as variáveis variam no mesmo sentido, ie, existe uma correlação direta.
 - $r < 0$  as variáveis variam em sentido inverso, ie, existe uma correlação inversa



- No que respeita ao valor numérico, este indica a força da correlação:

$$-1 \leq r \leq 1$$

- $r = 0$ Não há correlação
- $r = -1$ ou $r = +1$ a correlação é perfeita (quer seja direta ou inversa)
- $|r| \in]0.0; 0.2[$ a correlação é muito fraca
- $|r| \in [0.2; 0.4[$ a correlação é fraca
- $|r| \in [0.4; 0.6[$ a correlação é moderada
- $|r| \in [0.6; 0.8[$ a correlação é forte
- $|r| \in [0.8; 1,0[$ a correlação é muito forte



c) Fazer previsões

- Se X está correlacionado com Y, por exemplo através de uma relação linear do tipo:

$$Y = a * X + b$$

em que a é o declive da reta e b a ordenada na origem,
então,

se eu souber X, posso determinar Y (e vice versa), através da equação.

A esta previsão está associado um erro que tem a ver com o nível de significância considerado

d) Avaliar diferenças entre amostras

- É possível determinar se uma diferença observada entre duas amostras é devida a uma causa sistemática ou se é simplesmente o reflexo da variabilidade na amostra.

2. TESTES DE HIPÓTESES

Noção de hipótese

- É uma relação entre duas (ou mais variáveis), em que uma dessas variáveis é a variável independente e a outra a variável dependente.
- É uma proposição a uma ou mais populações e mais particularmente à forma como elas se relacionam.
- Uma hipótese diz sempre respeito a uma ou várias populações mas não às amostras, cujas características são objetivamente conhecidas.

Exemplos:

“as populações de que se extraem as amostras A, B, C e D têm médias iguais?”

ou

“a população donde se extrai a amostra E está distribuída normalmente?”

Hipóteses estatísticas

Hipótese nula:

H_0 – significa que **NÃO** há diferenças significativas entre as variáveis em estudo



Se aceito H_0 = não há diferenças = as variáveis são independentes = não são explicativas = uma não explica a outra = uma não influencia a outra

Se rejeito H_0 então aceito H_1



Hipótese alternativa:

H_1 – significa que **HÁ** diferenças significativas entre as variáveis em estudo









Se aceito H_1 = há diferenças significativas = as duas variáveis são dependentes = são explicativas = uma prediz a outra

Níveis de significância

- O nível de significância é a probabilidade de erro que podemos cometer ao rejeitar H_0 .
- Representa-se por p ou α
- Os níveis de significância mais utilizados são: $p = 5\%$ ou $p = 1\%$
 - Quando $p = 5\%$  em 100 casos 95% são verdadeiros e 5% são falsos
 - Quando $p = 1\%$  em 100 casos acerto em 99% e só erro em 1%

■ Valores de referência para o nível de significância:

- quando $p > 0,05$ (5%)  **não** há diferenças significativas
 Aceito H_0 (as variáveis são independentes)
- quando $p < 0,05$ (5%)  há diferenças **significativas**
 Rejeito H_0 e aceito H_1 (as variáveis são dependentes)
- quando $p < 0,01$ (1%)  há diferenças **bastante significativas**
- quando $p < 0,001$ (0,1%)  há diferenças **altamente significativas**

NOTA:

Há situações em que $p > 0,05$ mas $p < 0,10$ (ou seja p está entre 5 e 10 %)

Embora não seja significativo a 95% é significativo a 90 % de confiança

Neste caso, diz-se que há **significância marginal**.

Erros de interpretação

- Se se adotar ou rejeitar a hipótese nula, corre-se o risco de cometer um ou outro dos seguintes erros:
 - **Erro TIPO I** consiste em rejeitar a hipótese nula, H_0 , embora, na realidade, ela seja verdadeira. Por outras palavras, o erro do tipo I consiste em admitir a existência de uma diferença sistemática ainda que, na realidade, não exista.
 - **Erro TIPO II** consiste em aceitar H_0 embora ela seja falsa.

Graus de liberdade

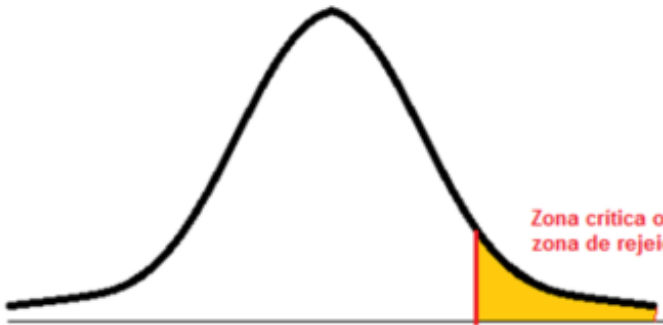
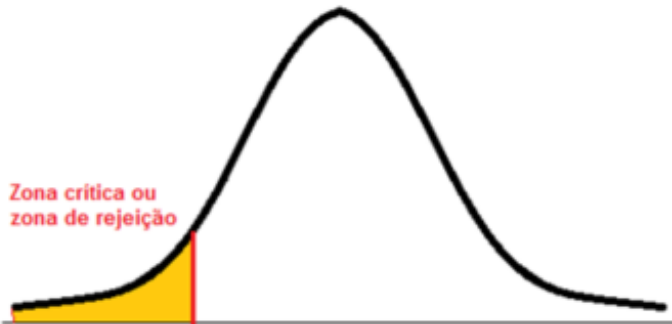
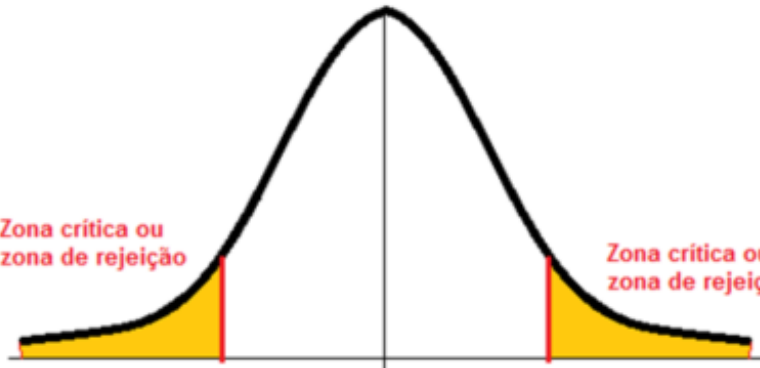
- Fórmula geral para calcular os graus de liberdade:

$$gl = N - 1$$

em que N é número de casos

Testes unilaterias *versus* bilaterias

- O teste pode ser unilateral à direita, à esquerda ou bilateral
- O mais comum é usar os testes bilaterais

Formulação de H1	Lateralidade	Localização da Zona Crítica
$>$	Unilateral à direita	
$<$	Unilateral à esquerda	
\neq	Bilateral	

Testes paramétricos *versus* não paramétricos

- Em regra geral, para cada teste paramétrico há um teste não paramétrico correspondente.
- Os testes paramétricos são mais fortes e oferecem maior flexibilidade do que os não paramétricos e por esse motivo se utilizam com maior frequência.

Testes paramétricos

- Baseiam-se na estimativa de parâmetros, pelo menos um
- A sua aplicação pressupõe que sejam verificadas 3 condições: variáveis numéricas (razão ou intervalar), distribuição normal, homogeneidade de grupos

Testes não paramétricos

- Não se baseiam em estimar parâmetros. Os métodos não paramétricos incluem suposições menos restritivas que as anteriores sendo por isso designadas de distribuição livre e só devem aplicar-se quando os dados se medem numa escala nominal ou ordinal.
- Não importa se as distribuições são ou não normais.
- Não têm em consideração o tamanho da amostra.

▪ Pressupostos para utilização dos testes paramétricos

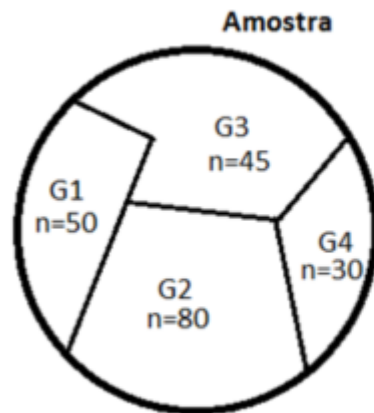
- ✓ Aplicáveis a variáveis numéricas (intervalares ou de razão)
- ✓ Requerem que as variáveis tenham uma distribuição normal

Para amostras grandes ($N > 120$) o comportamento aproxima-se da distribuição normal.

- ✓ O nº de casos em cada grupo deve ser no mínimo 30
- ✓ Os grupos devem ser homogéneos

Se $N/n < 1,5$ os grupos são homogéneos

em que N = dimensão do maior grupo e n = dimensão do menor grupo



Neste caso, $\frac{80}{30} = 2,7 > 1,5$

os grupos Não são homogéneos

Não se pode usar testes paramétricos

Usar testes Não paramétricos

QUADRO COMPARATIVO

Testes paramétricos (para var. intervalares ou de razão)

Testes NÃO paramétricos

Teste t de student para amostras independentes



Teste de U Mann Whitney (UMW)⁰

Análise de variância a um factor¹
(ANOVA)



Teste de Kruskal Wallis (KW)

Análise de variância a dois factores
(Two way)



Não tem correspondente

Teste t emparelhado²



Teste de Wilcoxon

ANOVA emparelhada



Teste de Friedman

MANOVA



Não tem correspondente

⁰usa-se em variáveis nominais ou também em ordinais

¹variável qualitativa com 3 ou mais grupos

²é necessário um mínimo de 30 casos

Testes paramétricos (para var. intervalares ou de razão)

Testes NÃO paramétricos

Correlação de Pierson (r)³



Correlação de spierman (ρ)³

Regressão linear simples



Não tem correspondente

Regressão linear múltipla



Não tem correspondente

Não tem correspondente



Análise discriminante de função

Não tem correspondente



Testes de proporção⁴

Não tem correspondente



Regressão logística binomial

Não tem correspondente



Regressão logística ordinal

Não tem correspondente



Regressão logística policórica

³usar uma ou outra é igual

⁴para variáveis nominais

3. EXEMPLOS DE SELEÇÃO DE TESTES

a)

Comparar a idade de rapazes e raparigas desta escola: há dois grupos (Masculino e Feminino)



posso usar o teste t de student para amostras independentes

b)

Verificar até que ponto a idade é diferente em função do sexo F ou M, mas neste caso a variável não apresenta uma distribuição normal (tem por exemplo enviesamentos ou achatamentos)



não se pode usar o teste t de student



usar o teste de UMW (U Mann Whitney), que é o não paramétrico correspondente

c)

Comparar a idade entre 3 grupos (pelo menos)

Idade *versus* EC { Solteiro
Casado
Viúvo

Tem-se uma variável quantitativa dependente (idade) e uma variável qualitativa independente a três factores (tricotómica), o estado civil

Se houver uma distribuição normal



teste ANOVA a 1 fator

Caso contrário (se não se verificarem as condições para aplicação do teste paramétrico)



teste de Kruskal Wallis

d)

Aplicar um teste de diagnóstico no início e no fim do semestre à mesma turma (mesmo grupo de sujeitos)



teste t emparelhado

e)

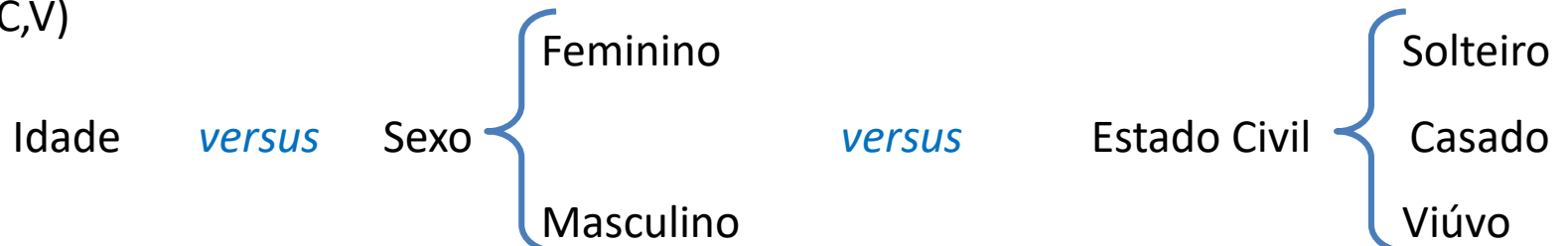
Se há emparelhamento, mas com 3 ou mais aplicações: p. ex. vamos medir a tensão arterial a um grupo de pessoas de manhã, ao meio do dia e à noite



teste ANOVA emparelhada

f)

Fazer uma análise entre as variáveis idade com sexo (M,F) e ainda com estado civil (S,C,V)



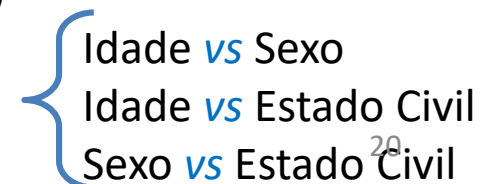
Tem-se uma variável quantitativa dependente (idade) e duas variáveis qualitativas independentes: uma dicotómica (sexo) e uma tricotómica (est. civil)



Análise de variância ANOVA a dois fatores (Two way)



dá-nos informação sobre as relações



g)

Se há diversas variáveis dependentes quantitativas e pelo menos uma variável independente qualitativa (podem ser mais que uma)

Glicémia

Colesterol

Hemoglobina

Plaquetas

Var. Dep. quantitativas

Sexo

Feminino

Masculino

Var. Indep. qualitativas



teste MANOVA

h)


Para avaliar duas variáveis quantitativas X e Y, em que X é a causa (var. independente) e Y é o efeito (var. dependente), e verificar se há relação causa-efeito entre elas, e se a variável é preditora



regressão linear simples

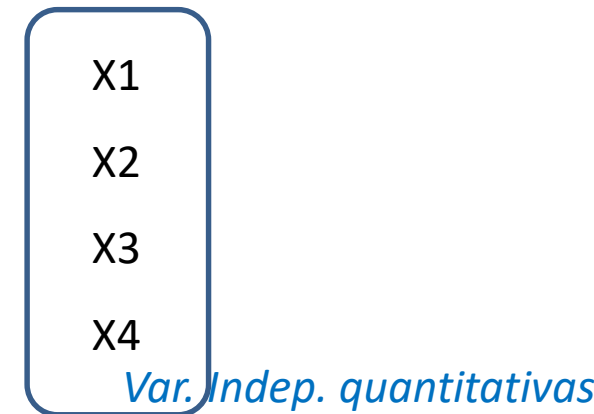
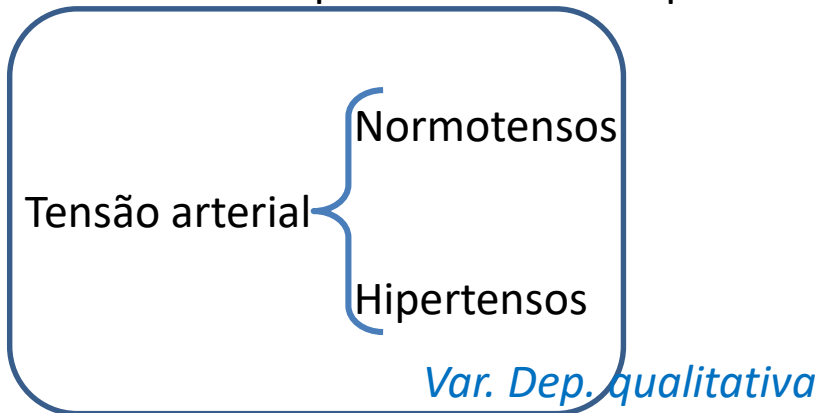
i)

Se há diversas variáveis independentes (X_1, X_2, \dots, X_n) e uma variável dependente (Y) (todas quantitativas) e quero verificar se as variáveis são preditoras

 regressão linear múltipla

j)

Tenho uma variável dependente qualitativa dicotómica (que tem dois grupos), ou também pode ser com mais grupos (ex. tricotómica = 3 grupos) que vou analisar com diversas variáveis independentes todas quantitativas (X_1, X_2, \dots)



 análise discriminante de função

k)

Se a var. dependente é qualitativa dicotômica e as var. independentes são qualitativas e/ou quantitativas (umas podem ser qualitativas e as outras quantitativas)



regressão logística binomial

l)

Se a var. dependente é qualitativa ordinal e as var. independentes são qualitativas e/ou quantitativas



regressão logística ordinal

m)

Se a var. dependente é qualitativa com vários grupos e as var. independentes são qualitativas e/ou quantitativas

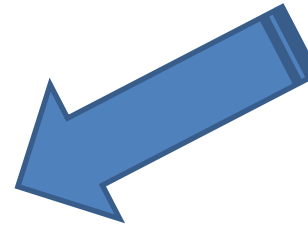


regressão logística policórica

n)

Há testes só para variáveis qualitativas:

testes de proporção
(NÃO paramétricos)



- teste de χ^2 – estuda a proporção entre duas variáveis qualitativas
- há dois testes relacionados com o χ^2 :
 - coeficiente de Fi
 - coeficiente de contingência
- quando não se pode aplicar o χ^2 (quando não é tabelas 2x2)



usa-se um outro teste: o teste de Fisher

FIM