Estatística e Delineamento

Jorge Cadima

Secção de Matemática (DCEB)
Instituto Superior de Agronomia (ULisboa)

2019-20

1. **Professores:**
   - Jorge Cadima (Responsável)
   - Elsa Gonçalves
   - Fernanda Valente
2. ***Webpage*:** `Sistema Fénix-Edu`
3. ***Software*:** 
   - *Homepage*: `www.r-project.org`
   - Repositório (para descarregar): `cran.r-project.org`

Objectivos

Admite-se que houve frequência duma disciplina introdutória de Estatística no primeiro ciclo (semelhante à existente no ISA).

Na UC Estatística e Delineamento admite-se que são conhecidos:

- principais indicadores descritivos (média, variância, covariância, coeficiente de correlação linear, etc.) e suas propriedades;
- conceitos básicos de probabilidades;
- variáveis aleatórias e sua caracterização;
- principais distribuições de probabilidades (Normal, χ^2, t-Student, F, Binomial, Poisson, etc.);
- conceitos de intervalos de confiança e testes de hipóteses.

AVISO: Vejam-se os materiais de apoio da UC Estatística dos primeiros ciclos do ISA (ou equivalentes), disponíveis na página *web* dessa UC (seguir o apontador na página *web* de ED).

Enquadramento

Nas disciplinas introdutórias de Estatística aborda-se o estudo das observações de uma variável.

A UC Estatística e Delineamento é uma disciplina de aprofundamento, que procura relacionar uma variável de interesse com outras variáveis, ou com hipóteses explicativas.

O fundamental do programa da UC diz respeito ao principal modelo estatístico: o Modelo Linear.

Existem apontamentos desta parte da matéria.

Aulas e horários de dúvidas

O semestre lectivo tem 14 semanas. As aulas práticas começaram na segunda-feira, dia 16 de Setembro.

- Aulas teóricas (2 vezes 1h por semana) - Há dois blocos diferentes, ambos com aulas 3as. e 5as.-feiras.
- Aulas práticas (2 vezes 1h30 por semana) - Há oito turmas (inscrições via Fénix).

AVISOS:

- Material de apoio às aulas na página *web* da UC (ver secção lateral de nome *Materiais de Apoio*).
- Para as aulas práticas é preciso ter conta informática de aluno (em caso de problema contactar a Divisão de Informática)
- Para as aulas práticas levar (i) enunciados dos Exercícios Introdutórios; e (ii) uma *pen* para guardar a sessão de trabalho.
- Há horários de dúvidas 3 dias por semana (ver *webpage*). Todos os horários de dúvidas são para qualquer aluno da UC.

Avaliação de conhecimentos

- Por testes (dois testes: um a meio do semestre e outro na data da primeira chamada de exame); ou
- Por exame final.

AVISOS:

- Aprovação por testes: classificação média igual ou superior a 9,5 valores no conjunto de dois testes, desde que em nenhum dos testes a classificação seja inferior a 8,0 valores.
- Aprovação em exame: classificação igual ou superior a 9,5 valores.
- Qualquer aluno inscrito na UC pode apresentar-se a avaliação de conhecimentos, desde que regularmente inscrito.
- Nas avaliações não são admitidas calculadoras gráficas, nem qualquer tipo de equipamento electrónico.
- Proposta primeiro teste: Sexta-feira, 15 Novembro, fim da tarde.

Programa

1. Testes de hipóteses para dados de contagem (baseados na estatística χ^2 de Pearson).
2. Modelo Linear.
    1. Regressão Linear Simples
    2. Regressão Linear Múltipla
    3. Análises de Variância (ANOVA) e variantes

1. Testes de Hipóteses para dados de contagem (estatística χ^2 de Pearson)

Objectivo: Testar se dados de contagem são compatíveis com uma dada hipótese explicativa.

Exemplo: Em viticultura há técnicas de enxertia chamadas "enxertos prontos". O sucesso depende de se verificar, ou não, o *pegamento* (formação do calo de enxertia). Deseja-se comparar o comportamento de três porta-enxertos específicos (1103P, SO4 e R99) para a casta Castelão. Eis os resultados dum estudo:

	1103P	SO4	R99
não pegamento	8	12	32
pegamento	954	943	939

Estes resultados são compatíveis com a hipótese de os porta-enxertos serem equivalentes, em matéria de pegamento? Ou a diferença nos resultados observados pode ser considerada estatisticamente significativa, havendo porta-enxertos melhores do que outros?

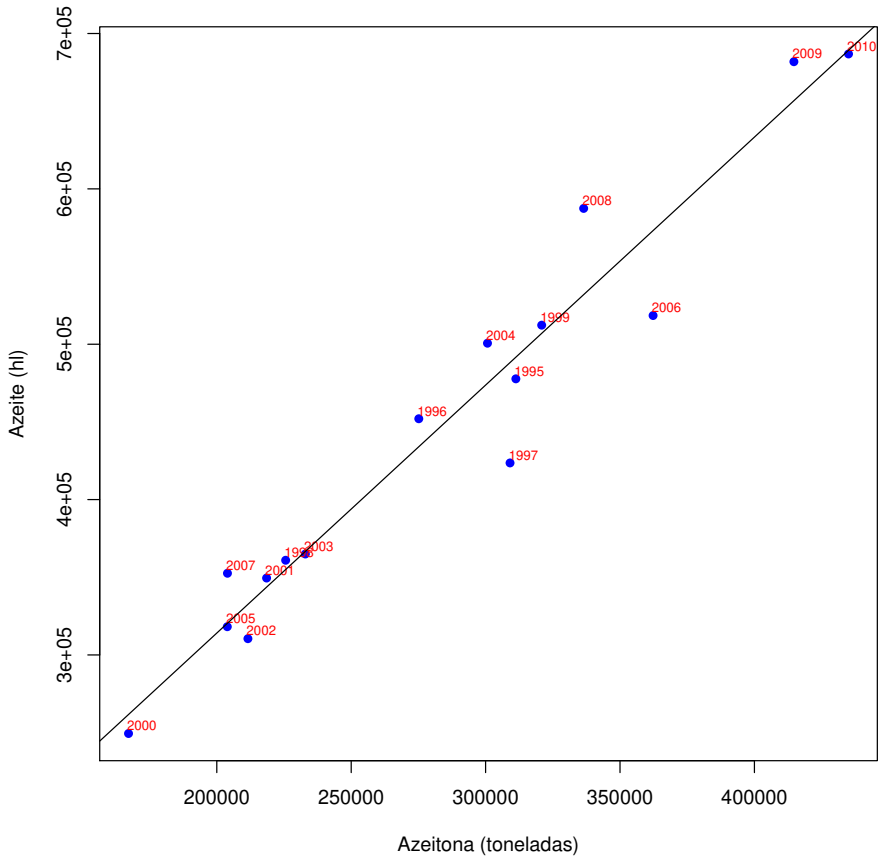
Testes χ^2 para dados de contagem (cont.)

- Breve revisão da teoria de testes de hipóteses.
- Teste de ajustamento duma distribuição unidimensional das contagens (as contagens duma experiência são compatíveis com uma distribuição Binomial? Ou uma Poisson?).
  - Probabilidades totalmente conhecidas
  - Probabilidades estimadas
- Testes para tabelas de contingência (contagens em tabela de dupla entrada, como no exemplo acima).
  - Probabilidades conhecidas (aplicações à teoria genética).
  - Teste de homogeneidade
  - Teste de independência

2.1. Modelo Linear: regressão linear simples

Objectivo: Relacionar linearmente duas variáveis numéricas.

Exemplo: Produção de azeitona e de azeite em Portugal, entre os anos 1995 e 2010 (Fonte:INE)



Modelo Linear: Regressão Linear Simples

- Contexto descritivo (o ajustamento da recta na amostra)
  - Equação da recta ajustada (método dos mínimos quadrados);
  - Propriedades da recta ajustada;
  - Relações não-lineares e transformações linearizantes.
- Inferência (sobre a recta populacional, com base numa amostra)
  - O modelo;
  - Propriedades distribucionais dos estimadores do modelo;
  - Intervalos de confiança para os parâmetros;
  - Testes de hipóteses para os parâmetros;
  - Análise dos resíduos para validação do modelo e identificação de observações especiais.

2.2. Modelo linear: Regressão Linear Múltipla

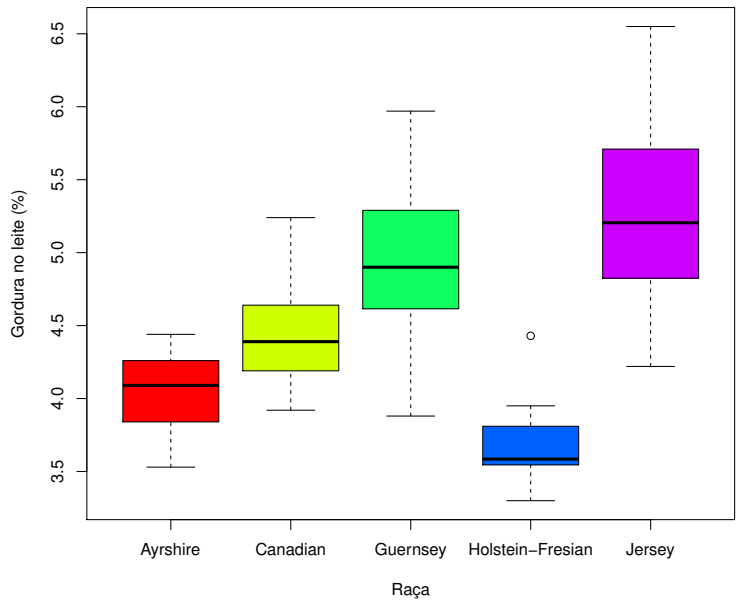
Objectivo: Relacionar uma variável resposta numérica com dois ou mais preditores numéricos, através duma relação linear.

- Contexto descritivo
  - Uma ferramenta: a notação matricial;
  - A estimação dos parâmetros do (hiper)plano ajustado;
  - Propriedades do (hiper)plano ajustado.
- A inferência
  - O modelo;
  - Propriedades distribucionais dos estimadores;
  - Intervalos de confiança e testes de hipóteses para os parâmetros;
  - Submodelos e selecção de submodelos;
  - Análise dos resíduos.

2.3. Modelo Linear: Análise de Variância (ANOVA)

Objectivo: Relacionar uma variável resposta numérica com um ou mais preditores categóricos (factores).

Exemplo: Comparar a % de gordura no leite de 5 raças de vacas (*Ayrshire, Canadian, Guernsey, Holstein-Fresian, Jersey*). Eis os diagramas de extremos e quartis para 20 vacas de cada raça:



Há raças com leite mais gordo, ou as diferenças são obra do acaso?

Modelo Linear: ANOVAs (cont.)

- Introdução ao delineamento experimental
- Delineamento a um factor totalmente casualizado e o modelo correspondente (efeitos fixos)
- Delineamento factorial a dois factores. O modelo sem interacção e o modelo com interacção (efeitos fixos).
- Delineamento a dois factores hierarquizados e respectivo modelo (efeitos fixos)
- Extensão: O modelo a um factor, com efeitos aleatórios

Bibliografia

1. **Referências Base:**
   - **Kutner, M.H.; Nachtsheim, C.J.; Neter, J. e Li, W.** (2005), *Applied Linear Statistical Models*, Irwin **[BISA: U10-727 e CD-236]**
   - **Slides e apontamentos das aulas teóricas** (disponibilizados na página *web* da UC)
2. **Outras referências:**
   - **Draper, N.R. e Smith, H. (1998)**, *Applied Regression Analysis*, 3a. edição, John Wiley & Sons **[BISA: U10-734] + [SI-78]** (**[BISA: U10-412]** a primeira edição de 1981).
   - **Montgomery, D.C. e Peck, E.A. (1982)**, *Introduction to Linear Regression Analysis*, John Wiley & Sons **[BISA: U10-329]**
   - **Seber, G.A.F. (1977)**, *Linear Regression Analysis*, John Wiley & Sons **[BISA: U10-416]**

Bibliografia (continuação)

3. **Referências de apoio à utilização do R**
   - **Docentes da disciplina de Estatística (2008/09)**, *Introdução à Aplicação R*, `ver apontador na página web da disciplina - Materiais de Apoio - Aulas Práticas`
   - **Maindonald, J. e Brown, W.J. (2003)**, *Data Analysis and Graphics using R*, Cambridge University Press **[BISA: U10-722]**
   - **Torgo, L. (2006)**, *Introdução à Programação em R*, `cran.r-project.org/doc/contrib/Torgo-ProgrammingIntro.pdf`
   - **Venables, W.N. e Ripley, B.D. (2002)**, *Modern Applied Statistics with S (fourth edition)*, Springer-Verlag **[BISA: U10-733]**

SLIDES DE APOIO

(aulas teóricas)

Estatística Descritiva

É útil dividir a metodologia estatística em duas grandes classes: descritiva e inferencial.

Estatística Descritiva: Métodos visando organizar, apresentar e extrair informação dum conjunto de dados.

- Os dados podem ser relativos a uma população inteira (censo), a uma amostra (aleatória ou não).
- As conclusões apenas dizem respeito às entidades observadas.
- Exemplos de ferramentas descritivas:
  - Para dados de uma só variável
    - ⋆ Cálculo de indicadores (média, variância, quantis, etc.).
    - ⋆ Tabelas de frequências.
    - ⋆ Histogramas, *boxplots* ou outras ferramentas gráficas.
  - Para dados relativos a duas variáveis
    - ⋆ Indicadores (Coeficientes de correlação, covariâncias, etc..)
    - ⋆ Nuvens de pontos (e, se for adequado, rectas de regressão)

Inferência Estatística

O problema conceptualmente mais complexo, de procurar conclusões relativas a um conjunto vasto de elementos (a população), a partir da observação apenas dum subconjunto dessa população (a amostra) designa-se inferência.

- Para que se possa falar em inferência estatística, é necessário que a amostra tenha sido escolhida de forma aleatória.
- A inferência estatística baseia-se na Teoria de Probabilidades, que estuda os fenómenos aleatórios.
- Exemplos de ferramentas inferenciais:
  - Estimadores e estudo das suas propriedades.
  - Intervalos de confiança para parâmetros populacionais.
  - Testes de Hipóteses.

A Inferência Estatística (cont.)