ANÁLISE DE DADOS

Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Faculdade de Economia da Universidade do Porto

RELATÓRIO

Maria Paula de Pinho de Brito Duarte Silva
1 de Março de 2018

Conteúdo

1 Introdução

O presente relatório foi elaborado para os efeitos do disposto na alínea b) do artº 5 do Decreto-Lei n. 239/2007 de 19 de Junho, referente à atribuição do título académico de agregado. Descreve o programa, conteúdos e método de ensino da unidade curricular de Análise de Dados do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão (MADSAD) da Faculdade de Economia da Universidade do Porto (FEP). Esta unidade curricular tem escolaridade semestral, o que corresponde a 14 semanas lectivas, com uma carga horária de 3 horas semanais. A candidata foi (co-)regente desta unidade curricular, neste curso e no Mestrado em Análise de Dados e Sistemas de Apoio à Decisão (ADSAD) que o precedeu, nos anos lectivos 1997/98 a 2002/03, 2004/05 a 2009/10 e 2011/12 a 2016/17 (interrupções em anos de licença sabática), tendo leccionado aulas teórico-práticas em todos aqueles anos lectivos. O programa aqui apresentado é essencialmente aquele que foi seguido durante essa leccionação. Nos últimos anos a unidade curricular tem sido lecciconada em inglês, e frequentada por estudantes estrangeiros em programas de mobilidade, razão pela qual os documentos anexos estão em língua inglesa.

Começamos por identificar os objectivos da unidade curricular, e apresentar o programa proposto. Explicamos então o enquadramento curricular da unidade, e explicitamos pré-requisitos, permitindo identificar destinatários naturais. Detalhamos em seguida o programa proposto e o desenrolar da sua leccionação e discutimos o processo de avaliação. Referimos o inquérito pedagógico efectuado aos estudantes, discutindo os seus resultados. O relatório é concluído por uma reflexão critíca e perspectivas de desenvolvimento futuro.

2 Objectivos e Programa

2.1 Objectivos

O objectivo geral da unidade curricular de Análise de Dados é o de formar os estudantes em métodos de análise univariada, bivariada e multivariada de dados. Num primeiro tempo, e de um ponto de vista uni e bi-variado, pretende-se sistematizar a análise exploratória e inferencial de uma amostra de uma ou duas variáveis. O estudante deverá depois aprender, de modo aprofundado, algumas metodologias de análise multivariada de dados, quer a nível exploratório - análise(s) factorial(ais) e análise classificatória - quer a nível estatístico - análise discriminante linear. O modelo de regressão linear múltipla não é considerado no programa proposto, tendo em conta que este tópico é estudado de forma desenvolvida no primeiro ciclo de licenciaturas em Economia e em Gestão, de onde provém a maioria dos estudantes deste curso.

Assume-se claramente que se trata de uma unidade curricular de formação universitária avançada, e não de um curso na perspectiva do utilizador. Sendo assim, pretende-se que os estudantes adquiram conhecimentos sólidos e aprofundados das metodologias estudadas e dos seus fundamentos, que lhes permitam no futuro aplicá-las de forma informada e interpretar os resultados de forma crítica.

Por outro lado, espera-se que os estudantes adquiram uma visão integrada das metodologias apresentadas, de modo a saber, perante um conjunto de dados a analisar, identificar o tipo de problema e a abordagem adequada, aplicar sequencialmente diferentes métodos e relacionar os respectivos resultados, por forma a obter uma visão geral do problema multivariado em análise.

Numa perspectiva aplicada, pretende-se ainda dotar os estudantes de competências na utilização de *packages* estatísticos, adequados ao programa leccionado. São efectuadas opções, face também aos meios disponíveis na UP e na FEP, mas pretende-se em última análise que, face a um *package* particular, os estudantes saibam identificar os métodos a aplicar, seleccionar correctamente os respectivos parâmetros, e interpretar os diferentes *outputs* de forma adequada.

2.2 Programa

Tendo em conta os objectivos acima expostos, o programa proposto, estruturado em cinco capítulos de extensão desigual, é o seguinte:

- Análise Preliminar de Dados (Revisão).
  - Tipologia das variáveis.
  - Análise Univariada.
    Variáveis qualitativas: representações gráficas, tabelas de frequências.
    Variáveis quantitativas: representações gráficas, medidas de localização, dispersão e forma, interpretação. Estandardização. *Outliers*.
    Ajustamento de distribuições: *Q-Q plot*, teste de Kolmogorov-Smirnov.
  - Análise Bivariada.
    Pares de variavéis quantitativas: diagrama de dispersão, coeficiente de correlação linear de Pearson, coeficientes de correlação ordinal de Spearman e de tau-de-Kendall, testes para a correlação.
    Uma variável quantitativa e uma variável qualitativa: comparações gráficas, testes paramétricos e não paramétricos para comparações de médias e de medianas.
    Pares de variavéis qualitativas: tabelas de contingência - introdução.
- Tabelas de Contingência.
  - Testes do Qui-quadrado de independência e de homogeneidade. Análise dos resíduos.
  - Medidas de associação.
  - Teste exacto de Fisher.
  - Teste de McNemar para amostras emparelhadas.
- Análise Factorial.
  - Análise em Componentes Principais.
    Objectivo, critério, solução. Análise da inércia explicada, decisão do número de componentes a reter. Interpretação das componentes principais com base nas variáveis e nos indivíduos, representações gráficas.
  - Análise em Factores Comuns e Específicos.
    Objectivos. Modelo geral, modelo ortogonal, propriedades. Comunalidades. Extracção do modelo factorial por componentes principais. Interpretação. Rotação.
  - Análise Factorial das Correspondências Simples.
    Objectivo. Perfis-linha e perfis-coluna. Métrica do Qui-quadrado. Interpretação dos factores. Representações gráficas.
  - Análise Factorial das Correspondências Múltiplas.
    Objectivo. Matriz disjuntiva completa. Interpretação dos factores. Representações gráficas.
- Análise Classificatória.
  - Medidas de Comparação.
  - Classificação Hierárquica:
    Noções gerais. Algoritmo de classificação hierárquica ascendente, índices de agregação. Ultramétrica induzida. Determinação do número de classes adequado.

  - Classificação Não Hierárquica:
    Algoritmo de Fisher. Algoritmo do Líder. Métodos de transferências, métodos centroïdes. Métodos de Forgy e das das k-médias de MacQueen, convergência. Método dos medoïdes. Metodologia das nuvens dinâmicas.
  - Interpretação de uma partição.
  - *Tandem analysis*.

- Análise Discriminante Linear.
  - Análise Discriminante em 2 grupos:
    Testes preliminares, Λ de Wilks. Determinação e interpretação da função discriminante linear de Fisher. Análise Discriminante passo-a-passo (*stepwise*). Classificação por minimização do custo esperado.
  - Análise Discriminante em K grupos:
    Funções discriminantes lineares, testes sequenciais. Funções de classificação.

2.2.1 Software

- SPSS - *Statistical Package for the Social Sciences*
- SPAD - *Système Portable pour l'Analyse des Données - Coheris Analytics SPAD*

3 Enquadramento Curricular

3.1 Pré-Requisitos

Para uma boa compreensão da matéria leccionada na presente unidade curricular, os estudantes devem dominar, a nível de primeiro ciclo, os seguintes tópicos:

- Álgebra Linear: cálculo matricial, formas quadráticas, valores e vectores próprios.
- Probabilidades e Estatística: estatística descritiva, noções de cálculo de probabilidades, distribuição Normal e Multinormal, testes de hipóteses.

Métodos de análise multivariada de dados tratam tabelas de dados em que n entidades ("indivíduos") são descritas simultaneamente por p variáveis. A análise conjunta destas observações multivariadas utiliza intensivamente cálculo matricial. Em particular, métodos de Análise Factorial e Análise Discriminante baseiam-se na optimização de formas quadráticas, cujas soluções são obtidas por análise espectral de matrizes apropriadas. Assim, conceitos de Álgebra Linear, usualmente leccionados nos primeiros anos de licenciaturas com uma componente quantitativa, são agora necessários e aplicados num contexto de análise de dados.

Por outro lado, esta unidade curricular assume uma abordagem estatística, e situa-se assim no prolongamento, agora numa perspectiva multivariada, das unidade curriculares de Probabilidades e Estatística leccionadas no primeiro ciclo. Conceitos e propriedades de estatística descritiva uni e bivariadas são dados como adquiridos, sem prejuízo de serem agora revistos numa perspectiva que é mais de interpretação do que de operacionalização. Noções muito gerais de cálculo de probabilidades são utilizados em particular na análise de tabelas de contingência. A distribuição Gaussiana é uma das hipóteses de base do modelo da análise discriminante linear, sendo aqui naturalmente necessária a sua extensão multivariada. Testes de hipóteses são aplicados na análise uni e bivariada, assim como na análise estatística do modelo de análise discriminante linear.

Sendo assim, esta unidade curricular dirige-se a estudantes com formação em métodos quantitativos a nível de primeiro ciclo, cobrindo métodos matemáticos e estatísticos. Estão neste caso os licenciados em Matemática, Estatística, Economia, Gestão, Engenharias, candidatos naturais ao Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão.

3.2 Enquadramento no Curso de Mestrado

O actual Plano de Estudos do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão (MADSAD) da Faculdade de Economia da Universidade do Porto (Anexo A) tem a duração de dois anos lectivos, e corresponde a um total de 120 ECTS. O 1º ano, correspondente ao "Curso de Mestrado", com 60 ECTS, é composto por um total de oito unidade curriculares semestrais (quatro em cada semestre), das quais seis são obrigatórias e duas optativas. A unidade curricular de Análise de Dados é uma unidade curricular obrigatória do 2º semestre, com 7.5 ECTS correspondentes a uma carga horária de 3 horas semanais.

Ela ocorre após uma unidade curricular semestral de Estatística Aplicada e uma unidade curricular semestral de Extracção de Conhecimento de Dados no 1º semestre do Curso. Na unidade curricular de Estatística Aplicada são revistos os conceitos ligados a testes de hipóteses estatísticos, e o processo de aplicação de um teste, e são leccionados Testes Não-Paramétricos, e Análise de Variância (ANOVA), métodos susceptíveis de serem aplicados numa análise de dados e que serão referidos no primeiro capítulo da unidade curricular de Análise de Dados. Na unidade curricular de Extracção de Conhecimento de Dados os estudantes tomam contacto com vários tipos de tarefas de extracção de conhecimento de dados (*Data Mining*) e aprendem os principais métodos/algoritmos para cada tipo de tarefa. São abordados em particular problemas de análise exploratória de dados, classificação e agrupamento (*clustering*) que serão estudados numa perspectiva mais formal e extensa na presente unidade curricular.

3.3 Enquadramento Noutros Programas de Segundo Ciclo da FEP

A presente unidade curricular é também explicitamente proposta como optativa noutros programas de mestrado da FEP, a saber, no Mestrado em Economia e no Mestrado em Finanças. Ao formar estudantes em métodos de análise multivariada de dados, a presente unidade curricular reveste-se de especial importância para estudos de áreas económicas e financeiras onde seja necessário analisar e extrair informação de um conjunto de dados multivariados. Técnicas de redução da dimensão - análise factorial, de agrupamento e de análise discriminante são aqui particularmente úteis.

Assim, estudantes que pretendam dominar estas metodolologias com vista à sua aplicação em sede de dissertação, e/ou em estudos futuros, inscrevem-se na unidade curricular de Análise de Dados.

É ainda de salientar que esta unidade curricular é obrigatória para os estudantes do Mestrado em Finanças que pretendam obter a dupla titulação com a Universidade de Kozminski (Polónia).

4 Programa Passo a Passo

Nota : 1 aula = 3 horas, conforme tem sido leccionado.

4.1 Análise Preliminar de Dados - 2 aulas

Tipologia das Variáveis

Variáveis qualitativas ou categóricas, nominais e ordinais; variáveis quantitativas ou numéricas, discretas e contínuas, de escala intervalar e de escala de razão.

Análise Univariada

Variáveis qualitativas: representações gráficas, tabelas de frequências.
Variáveis quantitativas: representações gráficas, medidas de localização, dispersão e forma, interpretação. Estandardização. *Outliers*.
Análise exploratória univariada com SPSS.
Ajustamento de distribuições: *Q-Q plot*, teste de Kolmogorov-Smirnov.

Análise Bivariada

Pares de variáveis quantitativas: diagrama de dispersão, coeficiente de correlação linear de Pearson, coeficientes de correlação ordinal de Spearman e de tau-de-Kendall, testes para a correlação.
Uma variável quantitativa e uma variável qualitativa: comparações gráficas, testes paramétricos e não paramétricos para comparações de médias e de medianas.
Testes estatísticos no SPSS.

Comentário

Este é um capítulo de alguma forma introdutório à unidade curricular de Análise de Dados. Sendo certo que os conceitos aqui referidos são já conhecidos dos estudantes, parece-nos importante relembrar definições e propriedades e fixar notação. Insiste-se ainda na importância de efectuar uma análise uni e bivariada como etapa preliminar à análise multivariada, por forma a revelar o comportamento dos dados e identificar situações atípicas ou até problemáticas: é importante conhecer os dados antes de lhes aplicar métodos sofisticados. Assume-se que os dados constituem amostras aleatórias multivariadas de populações bem definidas.

Este capítulo começa com a definição de matriz (ou tabela) $n \times p$ de dados multivariados e o estudo da tipologia das variáveis. Este último aspecto reveste-se da maior importância, pois só com uma identificação adequada dos tipos das variáveis se poderão seleccionar os métodos adequados para as analisar. São apresentados exemplos com variáveis de diferentes tipos, insistindo-se na sua distinção. Neste ponto efectua-se a introdução ao *package* SPSS, apresentando-se as janelas de dados e de descrição das variáveis.

O objectivo essencial do capítulo é então a revisão dos conceitos de análise preliminar uni e bivariada para os diferentes tipos de variáveis, insistindo-se na interpretação dos resultados mais do que na sua operacionalização.

Para as variáveis qualitativas, ou categóricas, relembram-se tabelas de frequências e representações gráficas, referindo-se que frequências acumuladas apenas têm sentido para variáveis ordinais. Introduz-se aqui o comando FREQUENCIES do SPSS. Refere-se que as variáveis ordinais poderão em certos casos ser tratadas como numéricas, desde que estejamos em presença

de "escalas de *Lickert*". Prossegue-se com o estudo de variáveis numéricas, começando pelas representações gráficas (histograma, diagrama de caule-e-folhas). Prossegue-se com a revisão de medidas de localização, com ênfase na média aritmética e mediana, e suas propriedades, referindo-se o conceito de resistência, o qual será invocado sempre que pertinente e em particular em exemplos práticos; referem-se quartis e outros quantis, moda, média aparada. Estuda-se então a representação gráfica por *Box-Plot*, e a identificação de *outliers* moderados e severos. Passa-se de seguida ao estudo da dispersão, relembrando os conceitos de amplitude amostral, amplitude inter-quartis, variância e desvio padrão, desvio médio. Segue-se o estudo da simetria de uma distribuição, referindo-se medidas de assimetria, e finalmente o estudo da curtose. São apresentados os comandos DESCRIPTIVES e EXPLORE do SPSS e analisados os respectivos *outputs*.

Aborda-se então o ajustamento de distribuições, que por método gráfico - *Q-Q plot* - quer pelo teste não paramétrico de Kolmogorov-Smirnov.

Passa-se de seguida à análise bivariada, considerando-se os casos de duas variáveis numéricas e de uma variável numérica e outra categórica; o caso de duas variáveis categóricas, a analisar por recurso a tabelas de contingência, é apenas referido e será abordado no capítulo seguinte do programa.

Para pares de variáveis numéricas, recorda-se o conceito de correlação a partir do diagrama de dispersão, referindo-se o coeficiente de correlação linear de Pearson e suas propriedades. Apresenta-se o teste paramétrico para a correlação, para amostras de variáveis Gaussianas. Introduzem-se então os coeficientes de correlação ordinal de Spearman e tau-de Kendall, permitindo medir a associação também entre variáveis categóricas ordinais, e testar a correlação para variáveis não-Gaussianas.

Para comparar o comportamento de variáveis numéricas em sub-populações definidas por categorias de variáveis qualitativas, assume-se como hipótese de trabalho que essas categorias são fixadas, e que se dispõe de amostras aleatórias em cada sub-população. Consideram-se então testes paramétricos ou não paramétricos, conforme as populações sejam Gaussianas ou não e as amostras de grande ou pequena dimensão.

Apresentam-se exemplos em SPSS para todos os casos considerados.

No final deste capítulo os estudantes deverão:

- Distinguir os diferentes tipos de variáveis e entender o que é ou não lícito aplicar a cada tipo de dados;
- Saber representar uma tabela de dados multivariados em SPSS;
- Ter relembrado noções de estatística descritiva e saber interpretar tabelas, representações gráficas e indicadores de localização, dispersão e forma;
- Saber identificar *outliers* e avaliar o seu impacto;
- Discutir o ajustamento de distribuições específicas univariadas aos dados;
- Saber analisar a associação entre duas variáveis;
- Analisar e testar adequadamente a correlação entre duas variáveis numéricas;
- Usar SPSS para efectuar a análise preliminar dos dados e interpretar os *outputs*.

4.2 Tabelas de Contingência - 2 aulas

Testes do Qui-quadrado de independência e de homogeneidade. Correcção de Yates.
Análise dos resíduos.
Medidas de associação para tabelas de contingência.
Teste exacto de Fisher.
Teste de McNemar para amostras emparelhadas.

Comentário

Este capítulo vem na sequência do anterior, e de alguma forma completa o ponto sobre análise bivariada. Optou-se no entanto por colocar o estudo de tabelas de contingência num capítulo à parte, para lhe dar algum destaque, por um lado, e porque os conteúdos são essencialmente novos para a maioria dos estudantes da unidade curricular, por outro lado.

Noções gerais sobre tabelas de contingência são apresentadas a partir de um exemplo concreto. Aproveita-se aqui a oportunidade para definir os conceitos de perfil-linha e perfil-coluna, que serão utilizados posteriormente na Análise Factorial das Correspondências.

Introduz-se o teste de independência, fazendo apelo a conceitos de teoria das probabilidades. O teste é ilustrado com o exemplo introdutório, insistindo-se na interpretaçao do resultado. Apresenta-se o comando CROSSTABS do SPSS. Considera-se em seguida o caso particular de tabelas 2×2, com a correcção de continuidade de Yates. Estuda-se então o teste de homogeneidade. Passa-se à análise dos resíduos, permitindo identificar as células onde o desvio à hipótese nula é mais relevante. É introduzido o teste exacto de Fisher, para analisar tabelas 2×2 com frequências esperadas baixas, fazendo a ligação com a distribuição hipergeométrica. De notar que o resultado deste teste é fornecido por defeito no *output* do método CROSSTABS do SPSS para tabelas 2×2. Estudam-se então medidas de associação de diferentes tipos. O capítulo é completado com a apresentação do teste de MacNemar para amostras emparelhadas. Afigura-se-nos que este teste pode ser de especial interesse em estudos de Marketing, donde a opção pela sua inclusão no programa desta unidade curricular leccionada na FEP.

No final deste capítulo os estudantes deverão ser capazes de:

- Efectuar um teste de independência ou de homogeneidade a uma tabela de contingência bidimensional e interpretar o resultado;
- Analisar os resíduos de uma tabela de contingência;
- Seleccionar medidas de associação adequadas, e cálculá-las;
- Identificar situações em que é adequado o teste de McNemar, construir a respectiva tabela, efectuar o teste e interpretar o resultado.

4.3 Análise Factorial - 5 aulas

- Análise em Componentes Principais - 2 aulas
  Objectivo, critério, solução. Análise da inércia explicada, decisão do número de componentes a reter. Interpretação das componentes principais, com base nas variáveis e nos indivíduos, representações gráficas.
- Análise em Factores Comuns e Específicos - 1 aula
  Objectivos. Modelo geral, modelo ortogonal, propriedades. Comunalidades. Extracção do modelo factorial por componentes principais. Interpretação. Rotação.

- Análise Factorial das Correspondências Simples - 1 aula
  Objectivo. Perfis-linha e perfis-coluna. Métrica do Qui-quadrado. Interpretação dos factores. Representações gráficas.

- Análise Factorial das Correspondências Múltiplas - 1 aula
  Objectivo. Matriz disjuntiva completa. Interpretação dos factores. Representações gráficas.

Comentário

Neste capítulo estudam-se métodos factoriais de redução da dimensão, nas suas diferentes variantes consoante os objectivos e tipo de dados. Considera-se no entanto o modelo da Análise em Componentes Principais (ACP) como modelo de base, razão pela qual o capítulo é iniciado com o seu estudo e lhe é dedicado mais tempo. Na Análise em Componentes Principais e na Análise Factorial das Correspondências (Simples e Múltiplas) segue-se a abordagem dita da "escola francesa" de *Analyse des Données*, abordagem de tipo exploratório e centrada na representação no "espaço das variáveis" e no "espaço dos indivíduos". Neste contexto, é introduzido o *package* SPAD, preferencialmente utilizado para estes métodos.

Em cada caso, os métodos são apresentados alternando entre formulação teórica e ilustração prática, por forma a clarificar os objectivos e conceitos, e pôr em evidência as questões que se levantam a cada passo.

A Análise em Componentes Principais (ACP) é apresentada como um método de redução de dimensão realizada por via uma projecção num espaço de dimensão inferior. Explica-se o objectivo com base num exemplo, antes de passar a considerações teóricas. O critério para definição do subespaço a identificar é apresentado sob o ponto de vista dos indivíduos - maximizar a inércia da nuvem projectada, ou seja, diminuir o menos possível as distâncias entre os indivíduos na projecção - e sob o ponto de vista das variáveis - maximizar a variância explicada. Apresenta-se a solução, sem a deduzir teoricamente, distingindo entre análise normada e não-normada. Chega-se então à formulação do problema da selecção da dimensão do subespaço, i.e., do número de componentes a reter, a partir dos valores próprios da matriz de correlações ou de variância-covariância. São apresentados e discutidos três critérios, a saber, os critérios de Pearson, de Cattell e de Kaiser (para ACP normada), ilustrando com uma aplicação prática. Insiste-se nas propriedades das componentes principais. É então introduzido o *package* SPAD, exemplificando-se a leitura do ficheiro de dados, aplicação da ACP e exploração dos diferentes tipos de *outputs*. Prossegue-se com a interpretação das componentes principais retidas, com base nas variáveis (correlações, contribuições) e nos indívíduos (contribuições, qualidade da representação), quer a nível teórico quer a nível aplicado, e análise de representações gráficas.

A Análise em Factores Comuns e Específicos é introduzida no programa por ser a que é geralmente apresentada na literatura e *packages* informáticos de origem anglo-saxónica, e designada sob o título geral de *Factor Analysis* (ou Análise Factorial). Sendo assim, e para dar uma visão alargada da área aos estudantes, e capacidade de compreender estudos com os quais se venham a confrontar, decidimos incluir o modelo factorial no programa, sem contudo o explorar extensivamente. De facto, apenas é apresentado o modelo com factores extraídos pela ACP, aproveitando o ponto anterior do programa, e do que resulta uma análise que não deixa de ser uma perspectiva alternativa à própria ACP. Ficam por explorar, sendo no entanto referidas, abordagens paramétricas ou basedas em metodologias de tipo mínimos quadrados. Referem-se os objectivos com base em exemplos. Apresenta-se o modelo, com ênfase no model ortogonal, e estudam-se as suas propriedades. Detalha-se então a extracção do modelo a partir das componentes principais, e sua interpretação, a partir de um exemplo ilustrativo. Estuda-se finalmente a técnica de rotação dos factores. Introduz-se e explora-se o comando FACTOR ANALYSIS do SPSS.

A Análise Factorial das Correspondências Simples (AFCS) é apresentada como um método que tem o objectivo de pôr em evidência a relação entre duas variáveis qualitativas, que se revelem não independentes na análise preliminar da tabela de contingência. É introduzida como um caso particular da ACP, sobre a matriz de perfis-linha (ou de perfis-coluna), com a métrica do Qui-quadrado justificada pelo facto de que linhas (ou colunas) não podem aqui ter ponderações idênticas. Sublinha-se o papel simétrico de linhas e colunas, que há-de levar à possibilidade de uma representação gráfica simultânea. Faz-se também a ligação à análise das tabelas de contingência, em particular à estatística do Qui-quadrado, estudadas no ponto anterior do programa.

A implementação é efectuada com o *package* SPAD. A partir de um exemplo prático, e efectuando o paralelo com a ACP, referem-se os valores próprios e percentagens de inércia. A interpretação dos factores seleccionados é feita observando conjuntamente as coordenadas das categorias-linha e das categorias-coluna, e analisando as respectivas contribuições absolutas e relativas. A representação gráfica simultânea nos planos factoriais selecionados permite então pôr em evidência as relações entre categorias-linha e categorias-coluna.

A Análise Factorial das Correspondências Múltiplas completa este capítulo. É apresentada como uma extensão da AFCS a uma tabela disjuntiva completa, representando as observações de s variáveis qualitativas em n indivíduos, e particularmente adaptada à análise de inquéritos. Começa-se por chamar a atenção para o facto de que a matriz de dados original não é neste caso analisável do ponto de vista numérico, o que leva à necessidade da transformação para a forma disjuntiva completa. A partir desta representação, segue-se o modelo de análise da AFCS. Referem-se as baixas percentagens de inércia tipicamente observadas, resultantes do aumento artificial da dimensão da matriz de dados. São ainda discutidos valores da inércia associada a cada categoria e a cada variável-pergunta, permitindo concluir sobre o efeito do número de categorias de resposta. O método é ilustrado com recurso a um conjunto de dados de inquérito.

No final deste capítulo os estudantes deverão ser capazes de:

- Entender o conceito associado aos métodos de redução de dimensão por análise factorial;
- Identificar o método a aplicar em função do objectivo e tipo de dados;
- Compreender a ACP e conhecer e utilizar as propriedades das componentes principais;
- Aplicar e interpretar a análise em componentes principais em SPAD;
- Entender o modelo em factores comuns e específicos, e saber obtê-lo e interpretá-lo;
- Efectuar e interpretar uma análise factorial com SPSS;
- Entender e saber aplicar e interpretar uma Análise Factorial das Correspondências Simples, utilizando SPAD;
- Entender e saber aplicar e interpretar uma Análise Factorial das Correspondências Múltiplas, utilizando SPAD.

4.4 Análise Classificatória - 2.5 aulas

- Medidas de Comparação.
- Classificação Hierárquica:
  Noções gerais. Algoritmo de classificação hierárquica ascendente, índices de agregação. Ultramétrica induzida. Determinação do número de classes adequado.

- Classificação Não Hierárquica:
  Algoritmo de Fisher. Algoritmo do Líder. Métodos de transferências, métodos centroïdes. Métodos de Forgy e das das k-médias de MacQueen, convergência. Método dos medoïdes. Metodologia das nuvens dinâmicas.
- Interpretação de uma partição.
- *Tandem analysis*: Combinação de análise factorial e classificação.

Comentário

Este capítulo é dedicado à Análise Classificatória (AC), por vezes designada por *Análise de Clusters*, Análise de Agrupamento, ou ainda Classificação Não-Supervisionada. O capítulo inicia-se com a explicitação dos objectivos da AC, sublinhando a diferença face aos métodos ditos "de Classificação",. i.e., de Análise Discriminante ou Classificação Supervisionada. Referem-se aplicações em diferentes áreas de forma a pôr em evidência a aplicabilidade geral e pertinência da AC. É utilizada um abordagem geométrica, i.e., baseada em medidas de similaridade/dissimilaridade, não sendo no âmbito desta unidade curricular contempladas abordagens de classificação por modelos de mistura finita - *model-based clustering* - nem de classificação conceptual.

Sublinha-se que o resultado de uma Análise Classificatória, a organização dos "indivíduos" em classes, não é "certo" ou "errado", mas que será o espelho de um certo número de opções, nomeadamente em termos de medidas de comparação entre indivíduos e entre grupos de indivíduos.

Introduzem-se então os conceitos de medidas de semelhança e de dissemelhança, distância e ultramétrica, e respectivas propriedades. São apresentados exemplos de medidas, as mais frequentemente usadas para os diferentes tipos de variáveis, explicando o que se está realmente a medir em cada caso, e ilustrando com dados. No caso das variáveis numéricas, alerta-se para a influência das escalas de medida, motivando para a eventual necessidade de uma estandardização. Numa perspectiva de classificação de variáveis, referem-se medidas de associação para variáveis de diferentes tipos. Introduz-se ainda o conceito de inércia, definido a partir de uma distância quadrática, assim como a decomposição da inércia total de um conjunto repartido em classes em inércia intra-classes e inércia inter-classes.

O programa prossegue então com o modelo de classificação hierárquica, formalizando as definições de hierarquia e de hierarquia indexada. Apresenta-se o algoritmo de classificação hierárquica ascendente, e referem-se os índices de agregação mais usuais, sublinhando que se trata de mais uma escolha no processo classificatório. O algoritmo é ilustrado passo a passo com um pequeno conjunto de dados.

A classificação hierárquica é então exemplificada em SPSS, com vários exemplos, explorando-se diferentes opções de parâmetros.

Abordam-se critérios para a determinação do número de classes adequado, ou seja, da selecção de uma partição a partir da hierarquia obtida: a variação do valor do índice de agregação e a variação da inércia explicada.

Introduz-se e interpreta-se o conceito de ultramétrica induzida por uma hierarquia, e apresenta-se o teorema de bijecção de Johnson-Benzécri.

Finalmente, sublinha-se a importância da interpretação da(s) partição(ões) seleccionada(s) em termos das variáveis descritoras, como etapa última do processo classificatório, propondo-se o cálculo e comparação de indicadores numéricos e representações gráficas. O processo é ilustrado em SPSS.

Passa-se então à classificação não-hierárquica cujo objectivo é a determinação de uma partição de um conjunto de n elementos em k classes, introduzida informalmente como um problema

de optimização. Distinguem-se métodos centroïdes e métodos de transferências. Apresenta-se, resumidamente, o algoritmo óptimo de Fisher. Apresenta-se e discute-se o Algoritmo do Líder. Referem-se critérios de determinação de centros inciais, necessários para a inicialização dos métodos centroïdes.

Focando nos métodos deste tipo mais utilizados, estudam-se em detalhe os algoritmos de Forgy e das k-Médias de MacQueen. Prova-se a convergência para o caso do algoritmo de Forgy. Este método é ilustrado com um pequeno exemplo, permitindo seguir o processo passo a passo. Explora-se o comando KMEANS do SPSS. Volta-se então ao problema da identificação do número de classes, e da interpretação de uma partição. Analisa-se ainda o método dos k-Médoïdes. Introduz-se, como generalização, a metodologia das nuvens dinâmicas de Diday.

No final deste capítulo é apresentada a *Tandem analysis*, que combina a redução da dimensão por via de métodos factoriais e a classificação: trata-se de começar por efectuar uma análise em componentes principais (no caso de variáveis numéricas) ou uma análise das correspondências múltiplas (no caso de variáveis categóricas), seleccionar o número de factores pertinentes a reter, e usar as coordenadas factoriais nesses factores como base para a classificação. O processo é ilustrado com o *package* SPAD.

No final deste capítulo os estudantes deverão ser capazes de:

- Entender bem o objectivo e critérios gerais da Análise Classificatória;
- Compreender os objectivos e limitações dos diferentes métodos apresentados;
- Entender o papel das diferentes opções de medidas e parâmetros;
- Conhecer o processo de cada algoritmo passo a passo;
- Seleccionar adequadamente uma medida de comparação;
- Efectuar um classificação hierárquica e interpretar o resultado;
- Efectuar um classificação não-hierárquica;
- Discutir o número de classes adequado à luz de diferentes critérios;
- Intepretar em detalhe uma partição;
- Combinar ACP ou ACM com classificação;
- Utilizar SPSS e SPAD para Análise Classificatória.

4.5 Análise Discriminante Linear - 2.5 aulas

- Análise Discriminante em 2 grupos:
  Testes preliminares, Λ de Wilks. Determinação e interpretação da função discriminante linear de Fisher. Análise Discriminante passo-a-passo (*stepwise*). Classificação por minimização do custo esperado.
- Análise Discriminante em K grupos:
  Funções discriminantes lineares, testes sequenciais. Funções de classificação.

Comentário

Este capítulo aborda a Análise Discriminante (AD) numa perspectiva estatística, focando na Análise Discriminante Linear. É iniciado com a apresentação de exemplos de motivação, permitindo identificar claramente os objectivos. Opta-se por estudar em primeiro lugar, e com algum detalhe, a análise discriminante em dois grupos, por forma a favorecer uma melhor apreensão dos conceitos e metodologia, fazendo depois num segundo tempo a generalização para um número K $>$ 2 de grupos. Por "grupos" (termo habitualmente usado em AD) entendem-se (sub-)populações, das quais se dispõe de amostras aleatórias independentes. Em qualquer caso, distinguem-se as duas abordagens da Análise Discriminante: a descritiva, que tem por objectivo analisar como se distinguem os grupos em análise, e a classificatória, que determina regras de afectação para novos casos. Assumem-se as hipóteses fundamentais da Análise Discriminante Linear: normalidade multivariada das variáveis e igualdade das matrizes de variância-covariância em cada sub-população.

A abordagem do problema começa pelos testes preliminares de significância, com o objectivo de verificar se os grupos são efectivamente distinguíveis com base nas variáveis disponíveis. Consideram-se em primeiro lugar testes univariados - que para dois grupos são testes-t de comparação de médias de duas populações Gaussianas, e ANOVA's para K$>$2 grupos. Efectua-se então a decomposição habitual da matriz de quadrados e produtos-cruzados em matriz intra-grupos (matriz *within, W*) e matriz inter-grupos (matriz *between, B*) - deduzindo-se a estatística de Wilks, que permite efectuar o teste multivariado de significância.

Numa perspectiva descritiva, deduz-se a função discriminante linear de Fisher, passando pela formulação como um problema de determinação de valores e vectores próprios de uma matriz adequada. O processo é ilustrado com um pequeno exemplo, permitindo identificar todos os elementos necessários aos cálculos a efectuar. Estuda-se então a interpretação da função discriminante linear, sublinhando-se a importância da estandardização dos coeficientes, e discutindo a relevância dos coeficientes versus as correlações entre a variável discriminante e as variáveis descritoras. Efectua-se e analisa-se uma aplicação em SPSS. No ponto seguinte, propõe-se a análise discriminante passo a passo, com o objectivo de obter uma função discriminante parcimoniosa, introduzem-se diferentes critérios de selecção. Volta-se ao exemplo em SPSS para comparação.

Passa-se então ao problema de classificação em dois grupos. Fala-se em primeiro lugar no método *cutoff*, para introduzir de seguida a abordagem baseada na teoria da decisão, com o objectivo de minimizar o custo esperado de classificação errada. Deduz-se a regra óptima de classificação, sob as hipóteses enunciadas de normalidade multivariada e igualdade de matrizes de variância-covariância. Considera-se o caso particular de custos de classificação errada e probabilidades *a priori* idênticos para os dois grupos, referindo que neste caso se obtém a mesma solução do método *cutoff*. Obtém-se ainda a distribuição da variável discriminante em cada grupo, o que permite calcular as probabilidades de classificação errada. O processo é ilustrado com um caso prático, efectuando-se todos os cálculos passo a passo. Discute-se a variação da função/variável discriminante em função da relação entre custos de classificação errada e entre probabilidades *a priori*.

Aborda-se então o problema da validação da função discriminante, referindo-se diferentes técnicas. Discute-se a importância das hipótese consideradas, e alternativas a considerar em caso de serem violadas. Refere-se aqui a alternativa da análise dicriminante quadrática, para o caso de matrizes de variância-covariância distintas, alertando para a necessidade de grandes amostras nesse caso.

Finalmente, considera-se o problema da Análise Discriminante Linear em K grupos, com K$>$2. Do ponto de vista descritivo, referem-se as funções dicriminantes lineares de Fisher, relembra-se a estatística de Wilks, e introduzem-se os testes sequenciais de Qui-quadrado para decidir do número de funções discriminantes significativas. Obtém-se as funções de classificação para os di-

ferentes grupos, admitindo agora custos de classificação errada iguais. Efectua-se e interpreta-se uma aplicação prática com SPSS.

No final deste capítulo os estudantes deverão ser capazes de:

- Entender bem o objectivo da análise discriminante, e comprender as hipóteses subjacentes ao modelo de análise discriminante linear;
- Determinar a estatística de Wilks e efectuar e interpretar os testes preliminares;
- Distinguir entre a abordagem descritiva e a abordagem classificatória;
- Obter e interpretar a função discriminante linear de Fisher para dois grupos;
- Construir e representar graficamente a regra de classificação para dois grupos;
- Interpretar e testar a relevância das funções discriminantes lineares de Fisher para K>2 grupos;
- Construir e interpretar as funções de classificação para K>2 grupos;
- Aplicar a análise discriminante linear com SPSS e interpretar os resultados.

5 Bibliografia

1. *Análise Estatística com o SPSS Statistics*
   João Marôco, ReportNumber, 6ª ed, 2014.
2. *Applied Multivariate Data Analysis*
   B.S. Everitt & G. Dunn, Wiley, 2ª ed., 2013.
3. *Applied Multivariate Techniques*
   Subhash Sharma, Wiley, 1996.
4. *Multivariate Data Analysis - A Global Perspective*
   J.F. Hair Jr., W.C. Black, B.J. Babin & R.E. Anderson, Pearson Prentice Hall., 7ª ed., 2010. *Análise Multivariada de Dados*, Bookman, 6ª ed., 2009.
5. *Estatística Multivariada Aplicada*
   Elisabeth Reis, Edições Sílabo, 2ª ed., 2001.
6. *Statistique Exploratoire Multidimensionnelle*
   L. Lebart, A. Morineau & M. Piron, Dunod, Paris, 1995.
7. *Analyse Statistique des Données. Applications et cas pour le Marketing*
   H. Fennetea & C. Biales, Ellipses, Paris, 1993.
8. *Traitement des données statistiques: méthodes et programmes*
   L. Lebart, A. Morineau & J.P. Fénelon, Dunod, Paris, 1982.
9. *The Analysis of Contingency Tables*
   B.S. Everitt, Chapman & Hall, 2ª ed., 1992.

Comentário

O curso não segue nenhuma obra em particular, pelo que não é especificamente indicada uma obra de referência. A opção por um programa que combina elementos da chamada "escola francesa" - como a análise das correspondências - a par de tópicos mais clássicos na designada "escola anglo-saxónica", leva a que não se encontrem publicações com a totalidade dos assuntos abordados e contemplando necessariamente a forma como são apresentados. Assim, a lista de referências bibliográficas proposta combina obras em português que cobrem uma parte importante do programa, obras em língua inglesa e obras em língua francesa. Os títulos referidos têm ainda em conta o asservo disponível na biblioteca da FEP, razão pela qual se indicam obras já com alguns anos.

O livro do Prof. João Marôco [1] é central (naturalmente para estudantes que leiam o português), cobrindo muitos dos capítulos do programa (mas não todos) e combinando uma apresentação rigorosa da matéria com aplicações em SPSS, incluindo cópias de ecrã e interpretação de *outputs*, tornando-se uma obra de referência e estudo para a generalidade dos estudantes. A obra [2] é talvez a que mais se aproxima do programa proposto. O livro [3] faz uma apresentação clara da metodologias multivariadas, em particular da Análise em Componentes Principais, Análise Discriminante e Análise Classificatória, e é em geral popular entre os estudantes que procuram um livro em língua inglesa. A obra [4] é também já um clássico, cobrindo capítulos centrais do programa como a Análise Discriminante e a Análise Classificatória, mas tratando a Análise Factorial apenas segundo a tradição anglo-saxónica. As obras [6] a [8] são clássicos da escola francesa, sendo que [5] é uma versão revista e melhorada de [7]. O livro [9] é a obra de referência para a análise de Tabelas de Contingência.

6 Método de Ensino

6.1 Funcionamento das aulas

As aulas seguem um modelo teórico-prático, combinando apresentação de modelos e metodologias com ilustração prática em conjuntos de dados disponibilizados. A análise dos dados é acompanhada de explicação detalhada da aplicação dos comandos do *software* utilizado, discutindo-se os possíveis valores dos diferentes parâmetros e suas implicações. São analisados e discutidos em conjunto os resultados obtidos. As aulas são leccionadas em salas de computadores, permitindo que os estudantes aprendam a utilizar os *packages* estatísticos e possam replicar as análises que são apresentadas.

No início de cada aula é efectuada uma referência à(s) aula(s) anterior(es), de modo a re-situar o estudante no contexto do programa da unidade curricular; no final de cada aula resume-se o percurso efectuado, fazendo a ponte para a aula seguinte. Sempre que apropriado, refere-se a ligação da matéria apresentada com unidades curriculares subsequentes do curso.

6.2 Materiais disponibilizados

São disponibilizados aos estudantes, através das plataformas Sigarra e Moodle, materiais de apoio, a saber:

- Cópia (em formato PDF) das apresentações das aulas - disponibilizada antes da sua apresentação, para que possa ser completada pelos estudantes durante as aulas, e evitando a necessidade de registo de tudo o que é apresentado;
- Ficheiros dos dados utlizados nas aulas para ilustração;

- Colecções de exercícios para cada capítulo, para treino;
- Enunciados de exames de anos anteriores, sugerindo-se aos estudantes que tentem resolver os exercícios propostos numa lógica de auto-avaliação;
- Colecção de tabelas estatísticas idênticas às distribuídas para exame;
- Regras/indicações para a realização do trabalho prático.

7 Avaliação

A avaliação desta unidade curricular consiste num exame final presencial e num trabalho prático, obrigatório, a efectuar por grupos de dois estudantes. O trabalho prático será ainda objecto de uma apresentação oral no final do semestre. A classificação final da unidade curricular, quer em época normal quer em época de recurso, será a média ponderada da classificação do exame final e do trabalho prático, em que o primeiro, sendo individual, tem um peso de 60%; a aprovação fica condicionada à obtenção de pelo menos 7.0 valores no exame final.

7.1 Exame Final

O exame final - épocas normal e de recurso - tem como objectivo avaliar os conhecimentos adquiridos nos diferentes capítulos que constituem o programa da unidade curricular, verificando se o estudante domina o funcionamento dos métodos leccionados, identificando os passos a seguir, conhecendo as propriedades das estruturas envolvidas e sabendo utilizá-las adequadamente, e efectuando uma correcta interpretação dos resultados. Para este efeito, são considerados pequenos conjuntos de dados, de dimensão suficientemente reduzida para que possam ser analisados sem recurso a *software* e efectuando os cálculos necessários apenas com uma máquina de calcular, mas apresentando um nível de estrutura suficiente para que uma análise multivariada faça sentido e forneça resultados passíveis de interpretação substantiva.

Cada questão de um exame será corrigida por um único docente, de modo a assegurar a uniformidade de critérios. Discutiremos aqui um exame tipo, que se incluiu no Anexo B.

Este exame é constituído por quatro questões, de extensão e cotação desigual. A cotação de cada questão é atribuída tendo em conta as matérias avaliadas pela questão, e o tempo que foi consagrado à sua leccionação, assim como a extensão da sua resolução. Assim:

- A questão 1, com cotação de 5 valores, aborda a análise classificatória, neste caso em particular a classificação hierárquica ascendente, incluindo medidas de comparação. Nesta questão pretende-se avaliar
  - se o estudante domina o processo de estandardização dos dados (alínea a));
  - se sabe calcular a distância Euclideana entre dois indivíduos (alínea a));
  - se sabe identificar como se inicia o processo de agregação a partir de uma matriz de distâncias (alínea a));
  - se sabe como se processa a formação das classes e a que corresponde a altura (valor do índice de agregação) de cada classe formada (alínea b)) ;
  - se é capaz de interpretar a hierarquia obtida, identificando partições de interesse a partir dela e interpretando-as em termos das variáveis descritoras (alínea c)) ;

- se domina o conceito de inércia e a decomposição de inércia total de um conjunto em inércia intra-classes e inércia entre-classes, sabendo calculá-las com a informação disponível e usando a proporção de inércia explicada para avaliar a partição identificada (alínea d)).

Nesta questão a cotação está distribuída pelas diferentes alíneas da seguinte forma: a) 1.25 valores, b) 0.75 valores, c) 1.50 valores, d) 1.50 valores.

- A segunda questão, com cotação de 5 valores, diz respeito à análise de tabelas de contingência, incluindo a análise de correspondências simples, e pretendendo-se avaliar
  - se o estudante sabe efectuar e interpretar o teste de independência, e se sabe obter a respectiva estatística do Qui-quadrado a partir dos valores próprios da análise de correspondências (fórmula fornecida) (alínea a));
  - se o estudante domina o conceito de perfil-coluna (alínes b) e c));
  - se o estudante relaciona os perfis coluna com a independência (ou não) das variáveis que constituem a tabela de contingência (alínea d));
  - se o estudante sabe intrepretar os resultados de uma análise de correspondências simples (alínea e)).

  A repartição da cotação pelas alíneas é a seguinte: a) 1.75 valores, b) 0.50 valores, c) 0.50 valores, d) 0.50 valores, e) 1.75 valores.

- Na terceira questão, que tem a cotação de 6 valores, aborda-se a análise em componentes principais (ACP), e o modelo em factores comuns e específicos. Nesta questão avalia-se
  - se o estudante conhece as propriedades e sabe interpretar a lista de valores próprios de uma matriz de correlações, obtida no contexto de uma ACP normada (alínea a));
  - se o estudante domina o conceito de componente principal, e é capaz de efectuar a respectiva interpretação, a partir das correlações variável-factor (alínea b));
  - se o estudante conhece o modelo em factores comuns e específicos e é capaz de o determinar a partir das componentes principais (alínea c));
  - se o estudante sabe interpretar um modelo factorial, em termos de comunalidades e capacidade de reproduzir as correlações entre variáveis (alínea d)).

  A cotação reparte-se pelas alíneas da seguinte forma: a) 1.50 valores, b) 1.75 valores, c) 1.25 valores, d) 1.50 valores.

- A questão 4 diz respeito à análise dicriminante linear, tratando o problema da discriminação de dois grupos a partir de duas variáveis para as quais se assume uma distribuição Normal bivariada. Nesta questão, que tem a cotação de 4 valores, avalia-se
  - se o estudante é capaz de determinar a estatística de Wilks a partir dos valores das somas de quadrados e produtos cruzados e de efectuar e interpretar o teste multivariado associado (alínea a));
  - se o estudante sabe calcular, num caso concreto, a função discriminante linear a partir de estimativas de valores médios e matriz de variâncias-covarâncias amostrais (alínea b1));
  - se ao estudante é capaz de avaliar e discutir o efeito da variação de custos de classificação errada na função discriminante linear (alínea b2)).

A repartição da cotação pelas alíneas é a seguinte: a) 1.50 valores, b1) 1.50 valores, b2) 1.00 valores.

A prova está concebida para ter uma duração de 3 horas.

7.2 Trabalho Prático

O trabalho prático proposto tem como objectivo confrontar os estudantes com a aplicação a dados reais das metodologias aprendidas na unidade curricular. Os estudantes deverão constituir-se em grupos de dois. A amostra a analisar poderá ser recolhida pelo grupo, ou obtida de qualquer fonte. Grupos diferentes deverão tratar amostras diferentes. São fornecidas aos estudantes algumas regras e indicações, que os guiam na constituição do conjunto de dados analisar e no desenvolvimento do trabalho (ver Anexo C)

A experiência mostra que os estudantes aderem com entusiasmo ao trabalho prático, tendo sido referidas apreciações positivas pela ligação que este lhes proporciona entre a matéria leccionada e a vida real. Exemplos de dados tratados incluem dados referentes à actividade profissional dos estudantes, indicadores financeiros de empresas, dados de estatísticas oficiais obtidos do site do INE ou da Pordata, dados recolhidos de relatórios da ONU ou UNESCO, classificações de estudantes obtidas em diferentes provas/unidades curriculares, dados relativos à indústria automóvel, etc.

O desenvolvimento do trabalho prático compreende três fases :

Na primeira fase cada grupo de estudantes identifica a problemática a estudar, selecciona o conjunto de dados para análise e efectua algum pré-processamento (tratamento de dados omissos, construção de novas variáveis, controlo do efeito dimensão, etc.) O conjunto de dados proposto é verificado, e são eventualmente dadas indicações de adaptação, de modo a que no final desta fase cada grupo tenha um conjunto de dados apropriado e isento de problemas.

A fase seguinte consiste na análise univariada das diferentes variáveis descritivas, numéricas e/ou categóricas, e na análise bivariada dos pares de variáveis considerados interessantes. Nesta fase os estudantes não aplicam ainda metodologias multivariadas novas para eles. O objetivo desta etapa é obter conhecimento do comportamento das variáveis, familiarizar os estudantes com o *software*, e ainda garantir o desenvolvimento atempado do trabalho prático.

A última e mais importante fase do trabalho prático comprende as análises multivariadas dos dados. Esta fase deve ter em atenção a natureza intrínseca das variáveis, e, desejavelmente, ser orientada pelo conhecimento dos dados obtido na fase anterior. A escolha das metodologias a aplicar e das variáveis a incluir em cada análise deverão resultar de questões substantivas sobre o problema em análise.

O relatório a entregar (ver Anexo C) deverá apresentar os dados, colocar as questões pertinentes, e discutir de forma integrada os resultados obtidos, incluindo tabelas e gráficos que sustentem as conclusões apresentadas.

A apresentação oral final, de aproximadamente 15 minutos, tem como objectivo avaliar se os estudantes são capazes de apresentar o seu trabalho de forma clara e convincente, referindo os dados analisados, as questões colocadas, as metodologias utilizadas e principais conclusões, sempre seleccionando os aspectos mais relevantes.

7.3 Resultados da avaliação

Na Tabela 1 registam-se as estatísticas de avaliação desta unidade curricular nos últimos seis anos. Pode constatar-se que o número de estudantes inscritos na unidade curricular subiu em 2012/2013, mantendo-se depois aproximadamente constante durante quatro anos, tendo voltado a subir em 2016/2017, verificando-se no entanto que o número de estudantes que completam a

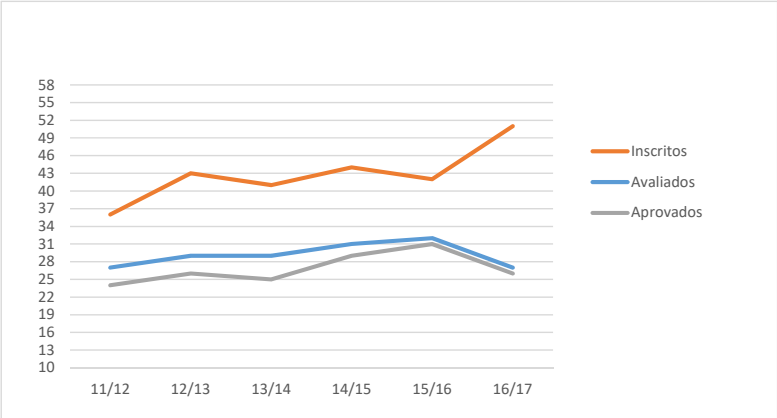


Figura 1: Número de inscritos, avaliados e aprovados na unidade curricular nos últimos seis anos.

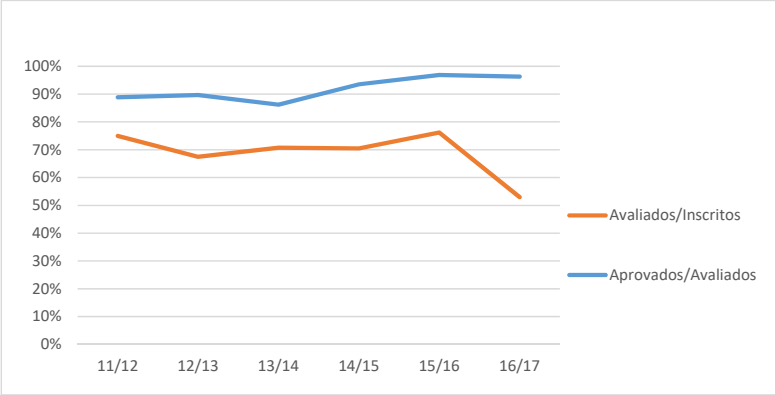


Figura 2: Rácios de avaliados sobre inscritos e aprovados sobre avaliados na unidade curricular nos últimos seis anos.

avaliação se manteve estável; o número de estudantes aprovados acompanha o número de avaliados, sendo que o rácio Aprovados/Avaliados foi sempre da ordem ou superior a 90%. Observa-se no entanto um decréscimo importante na percentagem de estudantes avaliados (em relação aos inscritos) no ano lectivo de 2016/2017.

Tabela 1: Estatísticas de avaliação.

Ano lectivo	Inscritos	Avaliados	Aprovados	Avaliados/Inscritos	Aprovados/Inscritos	Aprovados/Avaliados
11/12	36	27	24	0.75	0.67	0.89
12/13	43	29	26	0.67	0.60	0.90
13/14	41	29	25	0.71	0.61	0.86
14/15	44	31	29	0.70	0.66	0.94
15/16	42	32	31	0.76	0.74	0.97
16/17	51	27	26	0.53	0.51	0.96

8 O Inquérito Pedagógico

A Faculdade de Economia do Porto implementa um sistema de consulta aos estudantes através de um inquérito pedagógico. Este inquérito, aplicado no final da leccionação de cada unidade curricular tem como objectivos averiguar da percepção que os estudantes têm do funcionamento e dos potenciais impactos da unidade curricular e da docência (pela própria). Analisamos, com base ns resultados disponibilizados para os últimos cinco anos lectivos, as questões que visam avaliar, numa escala de 1 a 7, os seguintes aspectos :

- Sobre a disciplina:

  Q1 Pertinência dos objetivos - Apreciação e clareza

  Q2 Adequação da modalidade de avaliação aos objetivos da unidade curricular

  Q3 Grau de dificuldade dos conteúdos

  Q4 Volume de trabalho exigido em função dos objetivos e créditos da unidade curricular

  Q5 Apreciação global da unidade curricular

- Sobre o envolvimento e impacto no estudante:

  Q6 Percepção de envolvimento por parte do estudante

  Q7 Efeitos: Conhecimentos e capacidade de compreensão dos fenómenos e temas tratados

  Q8 Efeitos: A minha capacidade de resolução de problemas na investigação e/ou na prática profissional nesta área

  Q9 Efeitos: A minha capacidade de análise das implicações éticas, sociais ou políticas das diferentes matérias

  Q10 Efeitos: A minha capacidade de comunicar claramente as minhas conclusões e seus fundamentos

- Sobre a docência:

  Q11 Organização e estruturação dos conteúdos e a ctividades da unidade curricular

  Q12 Apresentação de várias perspetivas

  Q13 Uso dos contributos da investigação ou da prática profissional na docência

  Q14 Promoção da reflexão crítica dos estudantes

  Q15 Cumprimento das regras de avaliação acordadas com os estudantes

  Q16 Empenho na qualidade de ensino/aprendizagem

A Tabela 2 apresenta resultados agregados relativos aos últimos cinco anos lectivos. No Anexo D apresentam-se representações gráficas. De uma forma geral podemos afirmar que as apreciações foram razoavelmente boas, com valores médios variando entre 4.24 e 6.25. Observa-se que o número de respostas a estes inquéritos é em geral baixo, se atentarmos ao número de estudantes inscritos ou mesmo ao número de avaliados em cada ano, ficando sempre claramente abaixo dos 50%, ou até de 30%. Em quase todos os critérios observamos uma ligeira descida no último ano observado; este foi um ano atípico em que o número de inscritos foi particularmnte elevado, mas em que uma percentagem relativamente alta de estudantes não completaram o processo de avaliação, levantando-se a questão de se teria havido uma mudança no perfil dos estudantes - já que o programa e método de ensino não sofreram qualquer alteração. A seguir.

As questões sobre a unidade curricular (Q1 a Q5) mostram que os estudantes consideram os objectivos pertinentes e claramente definidos; observa-se ainda que consideram a unidade curricular com dificuldade relativamente elevada, e exigindo bastante trabalho. Lembremos que os conteúdos desta unidade curricular fazem apelo a conhecimentos de Estatística e de Álgebra Linear (ver Secção 3.1), que tipicamente os estudantes não têm presentes, e sem o domínio dos quais os conteúdos são difíceis de seguir. Há assim uma percepção de dificuldade, e necessidade de rever conceitos.

As questões Q6 a Q10 informam sobre a percepção que o estudante tem em relação ao seu envolvimento e aos efeitos da unidade curricular. Observamos que os valores mais baixos se registam nas questões sobre os possíveis efeitos das diferentes matérias da unidade curricular na capacidade de análise das implicações éticas, sociais ou políticas (Q9) assim como na capacidade de comunicar claramente conclusões (Q10). A primeira destas duas questões interpela-nos particularmente, e alerta para a necessidade de sublinhar a importância da análise e estruturação de dados e da clara representação das relações que são postas em evidência, como instrumentos de cidadania, fornecendo aos cidadãos um conhecimento da realidade social na qual se inserem, a nível político, económico, social, cultural... Por outro lado, a unidade curricular contribui certamente para desenvolver a capacidade de comunicar conclusões, nomeadamente através da apresentação do trabalho prático que é desenvolvido ao longo do semestre, este ponto deverá ser sublinhado e trabalhado.

As questões que avaliam a percepção dos estudantes sobre a docência (Q11 a Q16) mostram que esta é em geral avaliada positivamente e de forma estável ao longo dos anos. Os valores mais baixos registam-se nas questões relativas ao uso dos contributos da investigação ou da prática profissional na docência (Q13) e à promoção da reflexão crítica dos estudantes (Q14). A referência e a ligação dos tópicos leccionados a temas de investigação é um desafio interessante e motivador. A reflexão crítica, em particular no que respeita ao desenvolvimento do trabalho prático, é um aspecto que retemos para ser incentivado.

Tabela 2: Estatísticas dos inquéritos pedagógicos.

	2012-2013			2013-2014			2014-2015			2015-2016			2016-2017		
Questão	NR	TR	Média	NR	TR	Média	NR	TR	Média	NR	TR	Média	NR	TR	Média
Q1	13	30.2%	5.69	13	31.7%	5.46	8	18.2%	6.25	12	28.6%	6.00	21	41.2%	5.38
Q2	13	30.2%	5.08	13	31.7%	5.23	8	18.2%	6.13	12	28.6%	5.75	21	41.2%	4.24
Q3	13	30.2%	6.00	13	31.7%	5.38	8	18.2%	5.75	12	28.6%	5.67	21	41.2%	5.10
Q4	13	30.2%	6.31	13	31.7%	5.54	8	18.2%	6.13	12	28.6%	6.33	21	41.2%	5.62
Q5	13	30.2%	5.46	13	31.7%	5.31	8	18.2%	5.75	12	28.6%	5.58	21	41.2%	4.67
Q6	13	30.2%	5.54	13	31.7%	4.85	8	18.2%	5.75	12	28.6%	5.75	21	41.2%	5.38
Q7	13	30.2%	5.31	13	31.7%	5.23	7	15.9%	6.14	12	28.6%	5.67	21	41.2%	5.10
Q8	13	30.2%	5.23	13	31.7%	5.15	7	15.9%	6.00	12	28.6%	5.58	21	41.2%	4.76
Q9	13	30.2%	4.54	13	31.7%	4.69	7	15.9%	5.86	12	28.6%	4.50	21	41.2%	4.33
Q10	13	30.2%	4.77	13	31.7%	4.69	7	15.9%	6.14	12	28.6%	5.50	21	41.2%	4.71
Q11	13	30.2%	5.92	13	31.7%	5.15	8	18.2%	5.75	12	28.6%	6.00	16	31.4%	5.81
Q12	13	30.2%	5.62	13	31.7%	5.15	8	18.2%	5.63	12	28.6%	6.08	16	31.4%	5.75
Q13	13	30.2%	5.23	13	31.7%	5.46	8	18.2%	5.13	12	28.6%	5.67	16	31.4%	5.50
Q14	13	30.2%	5.38	13	31.7%	4.77	8	18.2%	5.13	12	28.6%	4.79	16	31.4%	5.31
Q15	13	30.2%	5.92	13	31.7%	5.85	8	18.2%	5.88	12	28.6%	6.50	16	31.4%	6.06
Q16	13	30.2%	5.69	13	31.7%	5.31	8	18.2%	5.75	12	28.6%	6.17	16	31.4%	5.94

NR : Nº de Respostas
TR : Taxa de Resposta = Nº de Respostas / Nº de Inscritos

9 A Análise de Dados Noutros Cursos de Mestrado

Nesta secção referimos unidades curriculares de cursos de mestrado de outras instituições que versam temáticas análogas às do curso aqui proposto. Não pretendendo ser exaustiva, esta análise põe em evidência o interesse generalizado pela análise multivariada de dados, justificando a sua inclusão em cursos de mestrado com focos variados, em escolas um pouco por todo o país.

- A *Information Management School* da Universidade Nova de Lisboa oferece um Mestrado em Estatística e Gestão de Informação, que inclui uma unidade curricular de "Análise de Dados" com 7.5 ECTS. O programa proposto inside sobretudo em Análise Estatística Descritiva Multivariada, com aplicações desenvolvendo análises univariadas, bivariadas e multivariadas de dados com variáveis quantitativas ou qualitativas. É utilizado o *software* SAS. `http://www.novaims.unl.pt/detalhe-disciplinas?d=200001&c=4281&r=1&o=1`

- O ICSTE - Instituto Universitário de Lisboa oferece unidades curriculares cobrindo tópicos de Análise de Dados nos Mestrados em Gestão (Escola de Gestão) e em Sistemas Integrados de Apoio à Decisão (Escola de Engenharia). O plano de estudos da unidade curricular de "Análise de Dados" do Mestrado em Gestão, com 6.0 ECTS, inclui Análise Exploratória Preliminar dos Dados, Testes de Hipóteses paramétricos e não-paramétricos e Análise de Variância, Regressão Linear, Análise em Componentes Principais e Análise Classificatória. Trata-se de um programa muito próximo do nosso, à excepção do facto que no caso da MADSAD a matéria relativa a testes de hipóteses é leccionada na unidade curricular de Estatística Aplicada (excepto a parte relacionada com Tabelas de Contingência). Já no Mestrado em Sistemas Integrados de Apoio à Decisão se encontram duas unidades curriculares apresentando estas temáticas, de 6.0 ECTS cada: (`https://www.iscte-iul.pt/curso/25/mestrado-sistemas-integrados-de-apoio-decisao/planoestudos`): "Análise de Dados para *Business Intelligence*", com objectivo de desenvolver competências para Análise de Dados, usando o SPSS (particularmente de dados recolhidos por questionário), incluindo Descrição de dados e Inferência Estatística, Regressão Linear, Análise em Componentes Principais e Análise Classificatória; e "Fundamentos de Ciência dos Dados", com forte ênfase na análise de vários casos de estudo, onde se leccionam Análise em Componentes Principais e Análise Classificatória a par de outros temas incluindo Regressão Bayesiana, Análise de Redes e Séries Temporais.

- Também no Mestrado em Matemática Aplicada à Economia e Gestão da Universidade de Lisboa encontramos uma unidade curricular de Amostragem e Análise de Dados, com 6.0 ECTS. A primeira parte da UC trata da Teoria da Amostragem, que no MADSAD é leccionada na unidade curricular de Estatística Aplicada; a segunda parte cobre métodos de análise multivariada de dados, a saber, Análise em Componentes Principais, Análise Factorial, Análise Classificatória e Tabelas de Contingência. O nosso programa inclui ainda Análise Discriminante, que não é aqui considerada.

- A Universidade de Aveiro oferece uma unidade curricular de Estatística Multivariada no Mestrado em Matemática e Aplicações, com 6.0 ECTS (`https://www.ua.pt/ensino/uc/4340`). O programa desta unidade curricular é muito próximo do nosso, cobrindo o estudo da distribuição Normal Multivariada, Análise de Componentes Principais, Análise Factorial, Análise Discriminante e Classificação e Análise de Agrupamentos (Análise Classificatória).

- O Instituto Superior de Contabilidade e Administração do Instituto Politécnico de Coimbra oferece também um Mestrado em Análise de Dados e Sistemas de Apoio à Decisão. Este curso inclui duas unidades curriculares, de 5.0 ECTS cada, com conteúdos de análise de dados multivariados, a saber: uma unidade curricular de "Análise Exploratória de Dados" (`http://www.iscac.pt/getfile.php?id=9205`) e uma unidade curricular de "Métodos Quantitativos Aplicados" (`http://www.iscac.pt/getfile.php?id=9257`). Na primeira são leccionados tópicos de Teoria das Probabilidades, Inferência Estatística (estimação, testes de hipóteses), Amostragem, Análise em Componentes Principais, Análise Factorial, Análise Classificatória e Métodos Econométricos (modelo de regressão linear simples e extensões). Os primeiros são cobertos na FEP ou a nível de primeiro ciclo, ou na unidade curricular de Estatística Aplicada do MADSAD; os modelos de análise multivariada são comuns ao nosso curso. Na disciplina de Métodos Quantitativos Aplicados encontramos a Análise Factorial e a Análise Discriminante a par de tópicos de Teoria das Probabilidades, modelos de Séries Temporais e, em geral, modelos quantitativos para Finanças (e.g. equações diferenciais estocásticas). Trata-se assim de uma unidade curricular de âmbito mais lato do que a que é aqui proposta, tendo no entanto pontos em comum.

- A Universidade de Évora oferece, no seu Mestrado em Modelação Estatística e Análise de Dados, uma unidade curricular de "Estatística de Dados Multivariados" com 9.0 ECTS, cujo programa contempla Técnicas de Dependência e de Interdependência e extensões, Análise exploratória de dados multivariados, Análise em Componentes Principais, Análise Factorial Exploratória vs Confirmatória, Análise Classificatória, Análise Discriminante e Tópicos em Modelos de Equações Estruturais. Essencialmente o nosso programa, à excepção do último ponto.

- O plano de estudos Mestrado em Estatística da Universidade do Minho inclui uma unidade curricular de "Análise Estatística Multivariada", com 6.0 ECTS. O programa parte do estudo de variáveis aleatórias multivariadas, estudando distribuições apropriadas e abordando a inferência, cobrindo depois métodos multivariadas de análise de dados, a saber, MANOVA, Análise em Componentes Principais, Análise Factorial, Análise Discriminante e Análise Classificatória. Dado o foco do Mestrado, regista-se uma preocupação com modelos estatísticos multivariados, e sua estimação, não considerados com detalhe no nosso curso; no que respeita às metodologias de análise de dados, os dois programas coincidem

Observamos assim que os programas de unidades curriculares afins são, nas suas grandes linhas, semelhantes ao aqui proposto. As diferenças registadas consistem fundamentalmente

na inclusão por vezes de tópicos de Teoria das Probabilidades e/ou de Estatística Inferencial e Amostragem, contemplados na FEP noutras unidades curriculares, ou na introdução de tópicos avançados, de interesse específico para o público correspondente, e por nós não considerados.

10 Comentários Finais e Perspectivas

O programa aqui proposto tem como objectivo dotar os estudantes do conhecimento de metodologias e de algum *know-how* em Análise de Dados, que lhes permita extrair informação essencial de um conjunto de dados multivariados. Num tempo em que são continuamente produzidos e disponibilizados conjuntos de dados de crescente dimensão e complexidade, o domínio de técnicas de representação e análise de dados é essencial para a tomada de decisão informada. As redes sociais, os sistemas de informação transacionais, instituições nacionais e supra-nacionais produzem continuamente grandes quantidades de dados, sendo mais do que nunca necessário desenvolver competências para a sua eficiente recolha, representação, análise, interpretação e organização.

Por outro lado, em todas as áreas do conhecimento, nas Ciências Sociais - Economia, Gestão, Sociologia - como nas Ciências da Engenharia, e naturalmente nas Ciências da Vida, os estudos desenvolvidos, seja numa perspectiva académica, seja numa perspectiva profissional, passam invariavelmente por uma análise de dados, um estudo quantitativo, a partir do qual serão extraídas as conclusões a apresentar. Dominar a metodologia de análise multivariada de dados - saber "o que fazer", "como fazer" e "o que se pode concluir" é fundamental.

Estudantes do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão (MADSAD), após ou mesmo durante o curso, são analistas de dados e gestores de crédito na banca e outras instituições financeiras, gestores de projectos ou responsáveis pelo desenvolvimento de aplicações de apoio à decisão em grandes grupos ou instituições nacionais, ou desenvolvem *Business Intelligence* em empresas de ponta, actividades profissionais onde o recurso a informação estatística é constante e o domínio dos métodos de análise de dados imprescindível.

Em cada um dos capítulos do programa foi necessário fazer opções de conteúdos, em função do tempo disponível. Dos tópicos não abordados, por insuficiência de tempo, destacaremos:

- classificação por modelos de misturas finitas, *model-based clustering*;
- regressão logística;
- e, naturalmente, a regressão linear múltipla.

A não inclusão do modelo de regressão linear múltipla no programa resulta do percepção de que a maioria dos estudantes do curso são licenciados em Economia ou em Gestão, tendo por isso no primeiro ciclo efectuado um estudo desenvolvido daqueles modelos em unidades curriculares de Métodos Econométricos. É o caso na Faculdade de Economia do Porto.

Finalmente, a selecção do *software* a utilizar está naturalmente sujeita a discussão e qualquer opção terá vantagens e inconvenientes. Nos anos em que a unidade curricular tem sido leccionada, optou-se pela utilização de *packages* estatísticos de aplicação generalizada, a saber SPSS e SPAD. Em particular, o *package* SPSS é ferramenta de trabalho em muitas empresas e instituições, e afigura-se-nos que dotar os estudantes de competências para a usar correcta e eficientemente na análise de dados é um objectivo relevante. O *package* SPAD permite a aplicação directa de métodos factoriais ditos "da escola francesa" - análise em componentes principais, análise das correspondências simples e múltiplas - fornecendo elementos de intrepretação muito completos.

A alternativa óbvia a esta opção é a utilização da linguagem R, para a qual estão gratuitamente disponíveis pacotes em variadíssimos temas, e cobrindo todas as temáticas da unidade

curricular. Esse será possivelmente o caminho futuro. Mas se para os estudantes do MADSAD isso será uma opção natural, tendo em conta a unidade curricular de programação no 1º semestre do curso, e a sua expectável apetência para este tipo de ferramentas, a mesma coisa já não se poderá talvez dizer dos estudantes de outras formações - Finanças, Economia... - que se inscrevem à unidade curricular, e que teriam provavelmente mais dificuldade em aderir à utilização intensiva de uma linguagem de programação.

ANEXOS

A Plano de Estudos do Mestrado

Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Plano Oficial de Bolonha

1º Ano

- 1º Semestre

Extracção de Conhecimento de Dados I	7.5 ECTS
Bases de Dados e Programação	7.5 ECTS
Optimização	7.5 ECTS
Estatística Aplicada	7.5 ECTS

- 2º Semestre

Análise de Dados	7.5 ECTS
Extração de Conhecimento de Dados II	7.5 ECTS
Unidades curriculares de opção do 2.º semestre (O estudante deverá optar por duas unidades curriculares de entre as optativas.)	15 ECTS

2º Ano

- 1º Semestre

Seminários	7.5 ECTS
Plano de Dissertação/Trabalho de Projeto/Estágio	7.5 ECTS

- Anual

Dissertação/Trabalho de Projeto/Estágio	45 ECTS

B Exame Final

FACULDADE DE ECONOMIA DO PORTO

MESTRADO EM MODELAÇÃO, ANÁLISE DE DADOS E SISTEMAS DE APOIO À DECISÃO

DATA ANALYSIS

Final Exam **Duration: 3h**

1. In a marketing study, the expenses of some clients in different products have been recorded. The table below registers the values of the expenses of 7 clients in Fresh Products, Frozen Products and Detergents (the table on the right indicates the standardized values).

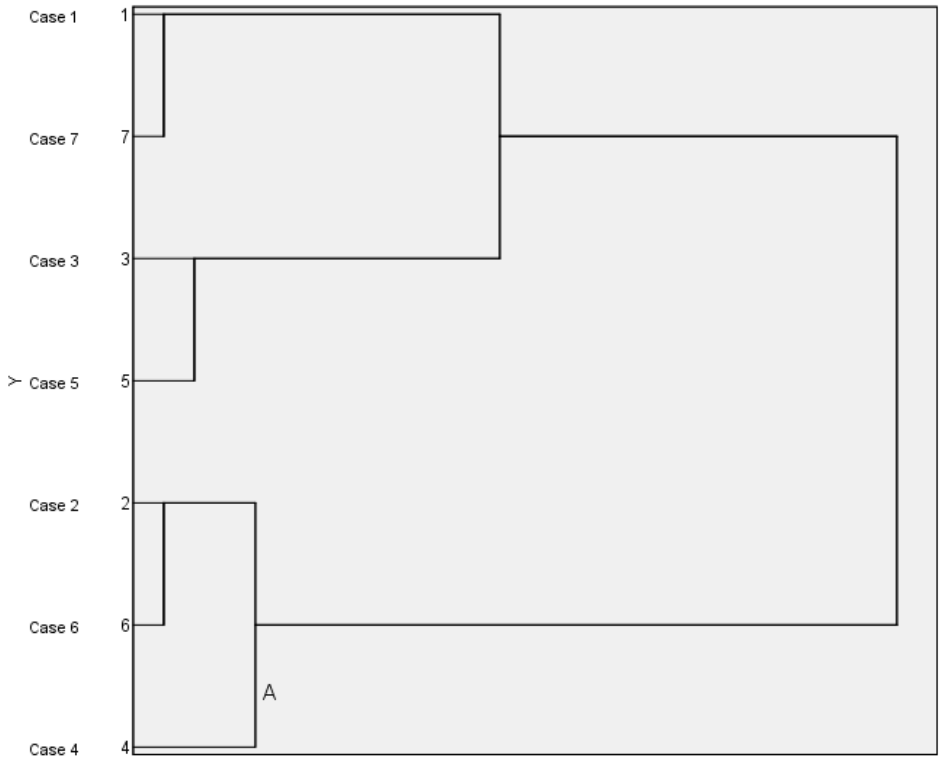
	Fresh	Frozen	Deterg.
Client 1	12669	214	2674
Client 2	56159	10002	212
Client 3	44466	7782	24171
Client 4	76237	16538	778
Client 5	35942	3254	26701
Client 6	76237	16538	778
Client 7	13146	1420	549
Mean	44979.43	7964	7980.429
St. Dev.	26515.17	6785.82	11972.73

Fresh	Frozen	Deterg.
-1.219	-1.142	V
0.422	0.300	-0.649
-0.019	-0.027	1.352
1.179	1.264	-0.602
-0.341	-0.694	1.564
1.179	1.264	-0.602
-1.201	-0.964	-0.621

The next table gathers the values of the squared Euclidean distance between the clients, based on the standardized data.

	Client 1	Client 2	Client 3	Client 4	Client 5	Client 6	Client 7
Client 1	.000	5.977	6.582	14.611	5.164	7.234	D
Client 2	5.977	.000	4.370	1.955	6.941	.313	5.188
Client 3	6.582	4.370	.000	7.771	.792	4.795	6.685
Client 4	14.611	1.955	7.771	.000	12.687	1.983	13.322
Client 5	5.164	6.941	.792	12.687	.000	7.324	5.691
Client 6	7.234	.313	4.795	1.983	7.324	.000	6.538
Client 7	D	5.188	6.685	13.322	5.691	6.538	.000

Hierarchical clustering has been applied to the data on the 7 clients, using the Squared Euclidean distance between standardized data, and the Complete aggregation index. The obtained dendrogram is represented below.



a) Determine the height of the first cluster to be formed.

b) What is the height of cluster A ?

c) Interpret the obtained clustering, indicating a partition on the clients' dataset and characterising the clusters.

d) Determine the value of the inertia explained by the partition in 2 clusters obtained from the hierarchy.

2. The data on 1404 clients of food shops have been recorded, crossing the store organization (what is the emphasis) – presentation, delicatessen, bakery, or none – with type of customer – whom he/she shops for – only him/herself, self and spouse, self and family. The results are gathered in the following table :

		Who shopping for		
		Self	Self and spouse	Self and family
Store organization	Emphasizes presentation	92	84	68
	Emphasizes delicatessen	128	128	72
	Emphasizes bakery	52	56	56
	No emphasis	212	292	164

A correspondence analysis on these data provided the eigenvalues λ_1 = 0,0075, λ_2 = 0,0047, plus the results in the two tables below.

N.B. : $\chi^2 = n \times$ total inertia

a) Check whether the type of shopping (who shopping for) is independent from the store organization.

b) Determine the table of column-profiles, and explain their meaning (using one as example).

c) What is the mean column-profile ?

d) How would table of column-profiles be in case of perfect independence ?

e) Discuss the results of the correspondence analysis.

		COORDO.		CONTRIB.		COS2	
	P.REL	1	2	1	2	1	2
Self	34.43	0.01	-0.09	0.5	65.1	0.01	0.99
S+Sp	39.91	-0.09	0.04	43.8	16.3	0.81	0.19
Fam	25.66	0.13	0.06	55.7	18.6	0.83	0.17

		COORDO.		CONTRIB.		COS2	
	P.REL	1	2	1	2	1	2
Prod	17.39	0.09	-0.06	20.5	13.9	0.70	0.30
Del	23.31	-0.04	-0.10	5.1	46.0	0.15	0.85
Bak	11.69	0.18	0.07	52.0	13.1	0.86	0.14
None	47.61	-0.06	0.05	22.4	27.0	0.57	0.43

3. A large number of US universities were analysed, and data on the following six variables were recorded v1="average salary of full professors (SFP)", v2="average salary of associate professors (SAP)", v3="average salary of assistant professors (SASP)", v4="Number of Full Professors (NFP) ", v5="Number of associate Professors (NAP)" v6="Number of Assistant Professors (NASP)"

A normed Principal Component Analysis was applied to the obtained data, leading to the following results:

eigenvalues	$\lambda 1$	$\lambda 2$	$\lambda 3$	$\lambda 4$	$\lambda 5$	$\lambda 6$
	4,3332	1,2596	A	B	0,0711	0,0672

Correlations Variable / Factor	**Axis 1**	**Axis 2**	**Axis 3**	**Axis 4**	**Axis 5**	**Axis 6**
Average salary - full professors	0,86	-0,41	0,14	0,24	0,03	-0,07
Average salary - associate professors	0,84	-0,51	-0,06	-0,03	-0,01	0,20
Average salary - assistant professors	0,86	-0,45	-0,08	-0,19	-0,02	-0,15
Number of full professors	0,85	0,43	0,28	-0,12	-0,06	0,03
Number of associate professors	0,86	0,47	-0,09	-0,01	0,21	0,01
Number of assistant professors	0,84	0,47	-0,19	0,10	-0,15	-0,01

a) Determine the eigenvalues A and B, and the percentages of dispersion recovered by all eigenvalues. Discuss the results.
(If you cannot determine all, use $\lambda_4 = 0{,}1188$).

b) Write and interpret the first two principal components. What are their mean values and variances ?

c) Write the salary and the number of full professors in the factorial model obtained by the first two principal components.

d) Estimate from the factorial model the communalities and the correlation between these two variables. Comment.

N.B. : $$r(Y_j, c_\alpha) = \frac{u_{\alpha j}\sqrt{\lambda_\alpha}}{\sigma_j}$$

4. Suppose you wish to discriminate students who qualified for a prize (Group 1) from those who have not (Group 2), based on marks on two tests (pre and post test). Assume the tests' marks follow a joint bi-Normal distribution, in each group, with common covariance matrices. In the available data, 918 students did qualify, whereas 1251 did not.

a) Given the values of sum of squares and cross-products in the table below, test if the two groups are significantly different when we consider both variables.

N.B. :
$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1 + n_2 - p - 1)}{p}\right)$$

sums of squares and cross-products			
		Pre-test	Post-test
Qualifies	Pre-test	99182,362	97443,416
	Post-test	97443,416	112676,973
Does not qualify	Pre-test	141075,276	136508,765
	Post-test	136508,765	155070,859
Total	Pre-test	392197,857	384584,607
	Post-test	384584,607	417083,720

b) Consider the sample estimates for the mean values and common covariance matrix

PRIZE		Mean
Qualifies	Pre-test	45,25
	Post-test	57,48
Does not qualify	Pre-test	62,29
	Post-test	74,38

		Pre-test	Post-test
Covariance	Pre-test	112,744	109,785
	Post-test	109,785	125,644

The classification rule that minimizes the expected value of the cost of misclassification is written:

Classify $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ in group 1 if

$$x' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \geq \frac{1}{2}(\mu^{(1)} + \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) + \ln k, \qquad k = \frac{C(1/2)}{C(2/1)} \frac{p_2}{p_1}$$

otherwise, classify x in group 2 (where $\mu(i)$ is the vector of mean-values for group i , i=1,2)

b1) Determine the classification rule if, moreover, the costs of misclassification are equal.

b2) Verify how this rule is changed if the cost of classifying an element of group 2 in group 1 is 100 times larger than that of classifying an element of group 1 in group 2. Comment.

C Trabalho Prático

PRACTICAL ASSIGNEMENT

Data Selection

The data set to be analysed should consist of a group of entities (persons, companies, countries, regions, …) described by a set of variables, at a given point in time – that is, cross-sectional data.

Do not select time-series data (data observed along time).

The number of entities should be larger than the number of variables.

The size of the data array depends in the topic, but : the number of entities should not be larger than 100-200, nor smaller than, say, 30, and the number of variables not larger than 20 nor smaller than 8-10.

Register your data in an Excel file first.

If you have counting data, for instance, nb. students in a given region, this should be transformed to relative data, dividing by the total population of the respective region – to avoid a "dimension" effect.

Introduction

Describe briefly the data, and refer its source. Explain the "questions" you have. Explain which methods you will be using to answer your "questions".

Data

Describe the data into more detail, detailing the variables, their meaning and type.

Univariate analysis

Descriptive statistics of ALL variables.

Indicators (position, dispersion, form..), graphic representations, analysis of (possible) outliers.

Please try to be compact, and not repetitive: use global tables when possible, and then comment on individual results.

Bivariate analysis

Analysis of correlations for numerical variables.

Contingency tables for some pairs (considered of particular interest) of categorical variables.

Statistical tests if pertinent.

Multivariate analysis

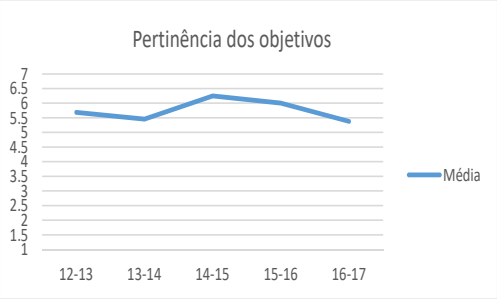
Multivariate analysis of the data set, using methods learnt in the course (not necessarily all of them !)

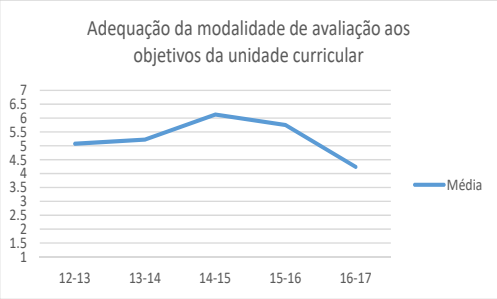
Conclusion

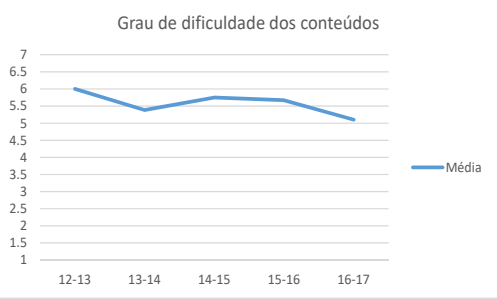
Refer the main conclusions drawn from your analysis’.

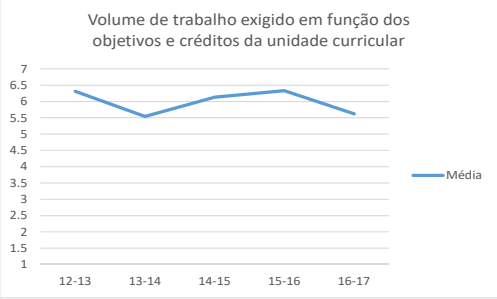
References

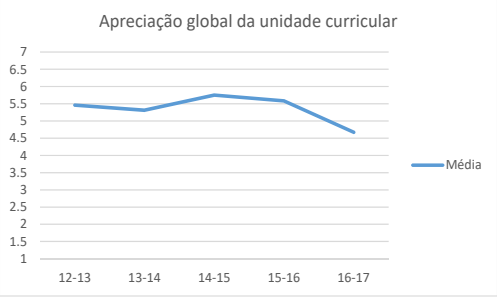
D Representações Gráficas dos Resultados dos Inquéritos Pedagógicos

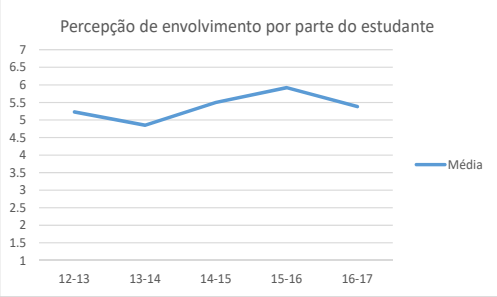


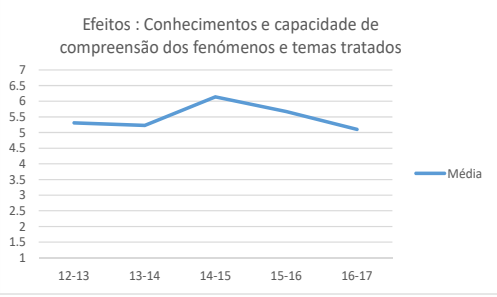


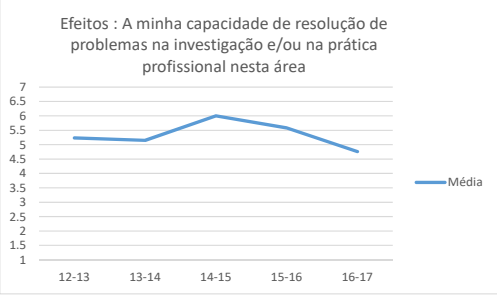


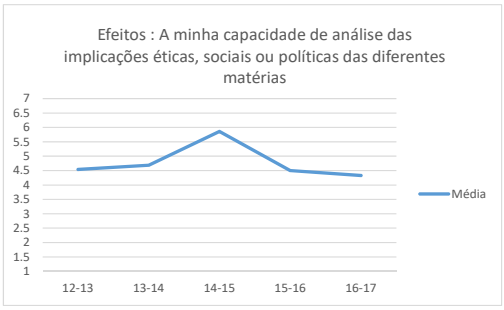


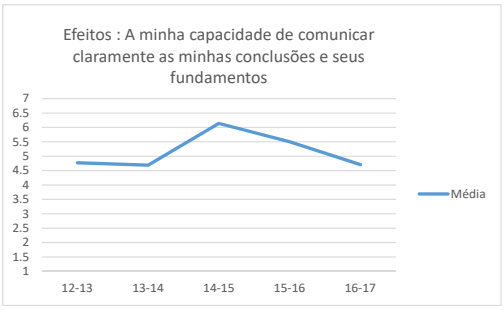


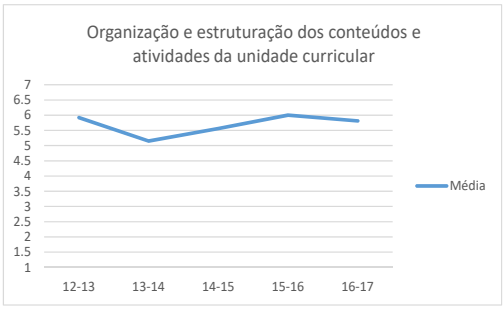


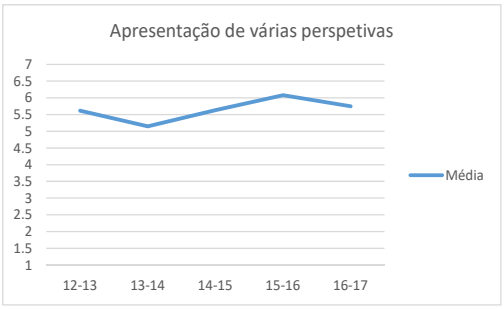


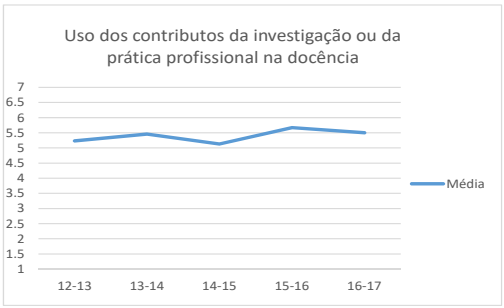


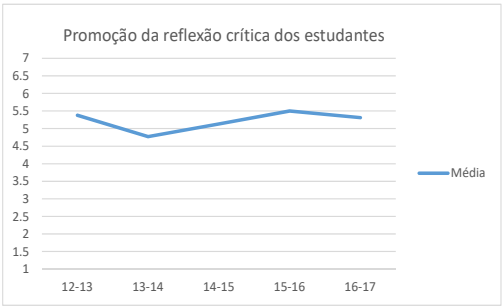

Efeitos : A minha capacidade de análise das implicações éticas, sociais ou políticas das diferentes matérias
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média

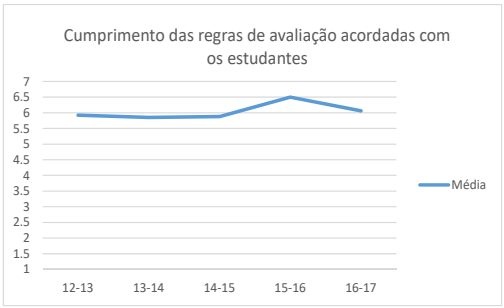

Efeitos : A minha capacidade de comunicar claramente as minhas conclusões e seus fundamentos
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média

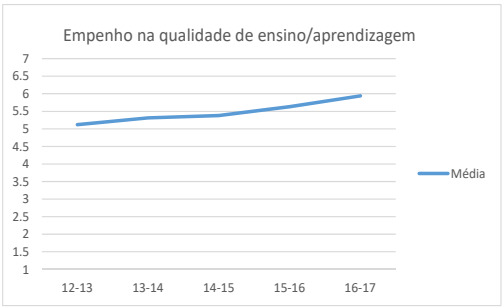

Organização e estruturação dos conteúdos e atividades da unidade curricular
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média


Apresentação de várias perspetivas
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média


Uso dos contributos da investigação ou da prática profissional na docência
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média


Promoção da reflexão crítica dos estudantes
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média


Cumprimento das regras de avaliação acordadas com os estudantes
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média


Empenho na qualidade de ensino/aprendizagem
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
12-13
13-14
14-15
15-16
16-17
Média