Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *primavera de 2013*

Estatística não - paramétrica

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe@fc.ul.pt*
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 500 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO SPE 2013

Está aberto, até **7 de junho de 2013**, o concurso para atribuição do **Prémio SPE 2013**, de acordo com o seguinte regulamento:

1. Pretendendo incentivar a participação dos setores mais jovens nas atividades da Sociedade Portuguesa de Estatística, especialmente no quadro das celebrações do Ano Internacional da Estatística,é instituído o **Prémio SPE 2013**.

2. Este prémio destina-se a estimular a atividade de estudo e investigação científica no domínio da Probabilidade e Estatística entre os jovens que trabalham nestas áreas.

3. O **Prémio SPE 2013** é constituído por uma quantia de 1000 euros.

4. Ao **Prémio SPE 2013** podem concorrer trabalhos originais sobre temas do domínio da Probabilidade e Estatística, desde que não tenham sido objeto de qualquer prémio atribuído por outra instituição.

5. Os autores dos trabalhos candidatos ao **Prémio SPE 2013** devem ser estudantes ou investigadores em alguma instituição portuguesa ou bolseiros portugueses, ser sócios da SPE e não ter atingido os 35 anos de idade até à data de submissão das candidaturas. Os autores não devem ter recebido o Prémio SPE nas quatro edições anteriores.

6. O trabalho deve ser escrito em português e não poderá exceder 25 páginas A4.

7. As candidaturas deverão vir acompanhadas do trabalho concorrente e do *curriculum vitae* dos autores e ser dirigidas ao Presidente da SPE, em carta registada, para a morada a seguir indicada. O carimbo do correio validará a data de entrega.

   Sociedade Portuguesa de Estatística
   Bloco C6, Piso 4 - Campo Grande
   1749-016 LISBOA

8. A decisão de admissibilidade e a apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição será da responsabilidade da Direção da SPE.

9. Os critérios de seleção pautar-se-ão pela exigência e precisão nos vários aspetos que o júri considerar pertinentes, nomeadamente: i) qualidade e clareza do texto; ii) inovação e rigor científico; iii) contribuição para o desenvolvimento do domínio da Probabilidade e Estatística nos planos teórico, metodológico ou aplicado.

10. O júri é soberano nas suas decisões, não havendo lugar a recurso.

11. O trabalho galardoado com o **Prémio SPE 2013** será apresentado em sessão plenária pelo seu autor ou autores no XXI Congresso da SPE e publicado nas respetivas Atas, com o devido formato editorial, de acordo com o processo de revisão dos artigos submetidos a estas.

12. A atribuição do **Prémio SPE 2013** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será feita no XXI Congresso da SPE na sessão plenária da sua apresentação.

13. O júri reserva-se o direito de não atribuir o **Prémio SPE 2013**.


XXI
CONGRESSO
SOCIEDADE PORTUGUESA DE ESTATÍSTICA
SANTO ANTONIO MILAGREIRO
http://spe2013.web.ua.pt

Editorial

… **no Ano Internacional da Estatística**…

1. O segundo ano

Os atuais órgãos administrativos da SPE eleitos no decorrer do XIX Congresso Anual tomaram posse no início de Janeiro de 2012. Decorre pois o segundo ano da atual direção. Na mensagem do Presidente neste Boletim, temos uma avaliação e um relato em tempos de Assembleia Geral ordinária. Como modesto Editor, cumpre-me felicitar a direção na pessoa do Presidente da SPE pela atividade desenvolvida e, com ênfase, relatada na Mensagem do Presidente. É relatada na modesta fórmula de notícias mas, na realidade, são palavras que cabem no conteúdo de um projeto interventivo que sempre se deseja como afirmação da SPE - muito em especial em Ano Internacional da Estatística.

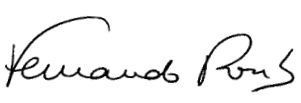
2. O Congresso Anual da SPE

A agenda nacional do Ano Internacional da Estatística pode ser consultada em http://www.spestatistica.pt/index.php?option=com_content&view=article&id=140%3Aagenda-aie-2013&catid=34%3Aeventos&lang=pt
Nesta plataforma da SPE podemos ser informados das mais diversas atividades, neste ano especial. Mantenhamos o nosso **alerta** para a concretização na divulgação do XXI Congresso Anual. Neste Boletim já divulgamos informação relevante. Apraz-nos registar e aplaudir a eficiência da equipa liderada pela Isabel Pereira que só o saber de experiência feito, em tempo muito apertado, apesar das vicissitudes, vai dar continuidade ao histórico acontecimento da SPE onde se partilha, divulga e mostra a vida da Sociedade Portuguesa de Estatística. É principalmente no Congresso que a SPE se afirma e revela o seu dinamismo anual pelo que, ao aceitar o convite, a equipa da Isabel desde já conquista o direito ao melhor sucesso.

3. O (constante) desafio editorial

O final do anterior editorial, "**… com um enorme desafio, no Ano Internacional da Estatística…**" continha:
"(…) Nos mais diversos desafios e atividades a desenvolver também o Boletim SPE pode intervir com "modesta contribuição". Como é bem sabido o Boletim SPE existe e vive da generosidade dos sócios SPE e de alguns autores que gerem as suas (às vezes muito difíceis) agendas profissionais de modo a poderem voluntariar-se para criar contribuições que formam o corpo e a mensagem desta publicação que editamos e desejamos manter bem viva e interventiva. (…) Aumentemos a "equipa de voluntários" ".
Na sequência desses desafios - com propostas, sem convite editorial - este Boletim inclui uma nova rubrica: Enigmística.
Além disso, concretiza-se uma ideia - Controvérsias - surgida logo no início do mandato da atual Direção. Para além do evidente enorme interesse do seu conteúdo, essa nova secção do Boletim exige a melhor colaboração de todos os elementos da comunidade científica que com ela se identifiquem. Fica desde já o desafio e o convite para uma participação que pode e deve incluir temáticas que estão para além do vulgar dia a dia do investigador, do professor ou do estudante. Começamos com um excelente "controverso texto" que é de uma densidade que só uma vida invulgarmente culta, intensa e dedicada permite construir toda a experiência, o suporte e o saber, das controvérsias nele inscritas. Só alguns - quero chamar-lhes *outliers*, qualquer que seja o modelo de discordância adotado - apenas esses, têm o privilégio de poder "controversar" do modo como iniciamos.
E, assim ficamos com um excelente **desafio de continuidade** para esta nova secção do Boletim SPE. A fasquia está alta mas… com a vantagem de ser controversa… decerto vai ter o enorme sucesso que os leitores quererão continuar.
Para além de uma publicação periódica, o Boletim SPE regista história científica ao mesmo tempo que divulga o estado da arte em algum tema específico. Necessita que os sócios se empenhem, pela construção e autoria, mas também pela proposta de novos Temas Centrais… Mais um desafio que o editor propõe.

O Tema Central do próximo *Boletim* será *A "Escola Bayesiana" em Portugal*



Mensagem do Presidente

Na sequência do meu texto *2013 - Ano Internacional da Estatística: como celebrá-lo?*, publicado nesta secção do Boletim de outono de 2012, procurei incentivar os membros da comunidade estatística englobada na SPE a participarem ativamente, através da promoção e organização de iniciativas, nas celebrações de 2013 - Ano Internacional da Estatística (AIE).

Têm sido vários os colegas que corresponderam a tal apelo, e a quem deixamos aqui expressa a nossa satisfação e gratidão, como transparece das informações que têm sido divulgadas pelo portal e pelo serviço de correio eletrónico da SPE, bem como das notícias que veiculo abaixo.

Não obstante, as informações que possuo não são suficientemente regozijadoras já que desconheço quaisquer ações realizadas ou programadas por vários grupos que parecem, incompreensivelmente, alhear-se desta campanha de consciencialização mundial que só pode trazer benefícios para todos nós e que, por isso mesmo, exige contrapartidas em conformidade. Quero crer que esta situação possa nos próximos tempos vir a sofrer a desejável mudança.

No prosseguimento das diligências e esforços que a Direção da SPE tem vindo a aplicar para que a comunidade estatística portuguesa celebre condignamente o AIE, cabe-me prestar as seguintes informações, posteriores à saída do anterior Boletim.

1. A Folha Informativa da RIIBES (Rede de Informação do INE em Bibliotecas do Ensino Superior) publicou no seu nº 42 (novembro 2012) uma entrevista com o Presidente da SPE onde este aborda temas relacionados com o Ano Internacional da Estatística, o relacionamento SPE-INE e papel dos congressos e outros encontros científicos promovidos pela SPE.

2. O Presidente da SPE foi convidado para estar presente na 11ª reunião plenária do Conselho Superior de Estatística, realizada em 14 de dezembro último, pela Vice-Presidente deste órgão (CSE) e simultaneamente Presidente do INE. O Presidente da SPE aproveitou essa oportunidade para, após expressar o seu agradecimento pelo convite recebido, fazer uma intervenção sobre a importância na celebração do AIE por parte das entidades representadas no CSE e a divulgação do programa corrente de atividades da SPE nesse âmbito.

3. Na sequência da decisão sobre uma emissão filatélica alusiva ao AIE por parte dos CTT, o Presidente da SPE tem vindo a assessorar a respetiva Direção de Filatelia na promoção do AIE, para o público especializado do Clube do Colecionador e do público em geral, a propósito da referida emissão.

4. Após terminada a sua elaboração, com base nos contributos dos colegas Pedro Campos (Coordenador), M. Eugénia G. Martins, Emília Oliveira, Bruno Sousa e Andreia Hall, inaugurou-se no dia 5 de fevereiro de 2013 a exposição Explorística na Escola Secundária Tomaz Pelayo (Santo Tirso), dando início à fase itinerante de propiciar a circulação da exposição pelo Pavilhão do Conhecimento e pelas escolas interessadas, cada vez em maior número, integrando-se assim nos objetivos primários do AIE.

5. Vai ser em breve distribuída aos sócios uma brochura, em edição SPE e impressa pelo INE, contendo biografias de alguns famosos estatísticos, editada pela nossa colega Emília Athayde da Universidade do Minho, com a prestativa colaboração de vários outros colegas.

6. Dando seguimento a sugestões da coordenação internacional da comemoração do AIE, a SPE e o CEAUL acordaram emitir um comunicado conjunto para os meios de informação sobre o AIE e sua participação na celebração desta iniciativa à escala global, o qual começou por ser divulgado através dos portais de comunicação do Instituto Superior Técnico e da Faculdade de Ciências da Universidade de Lisboa, bem como do serviço de comunicação e imagem do INE.

7. Em recente reunião havida com a respetiva Direção conseguiu-se obter uma primeira anuência para um patrocínio parcial da Agência Ciência Viva para a 2ª edição do projeto Radical Estatística, envolvendo de novo o grupo dos colegas Bruno Sousa, Dulce Gomes e restantes elementos. Esta é outra das estratégicas iniciativas da SPE que, alvejando o público mais jovem, apontam para um dos objetivos pivotais do AIE.

Tendo como alvo contribuir para a consecução do objetivo prioritário do AIE em ampliar o reconhecimento público do poder e impacte da Estatística na nossa vida, o Presidente da SPE decidiu impulsionar a elaboração de um texto que, em linguagem acessível, traduza o valiosíssimo papel que a Estatística desempenha nos vários domínios da sociedade. O artigo que se publica no corrente boletim, sob o título **Estatística na Sociedade – uma digressão ilustrativa por domínios de aplicação**, pretende ser uma concretização desse ambicioso texto e, nesse sentido, ajudar todos quantos estão para já empenhados na demonstração da relevância social da Estatística.

28 fevereiro 2013

O Presidente da SPE

Carlos Daniel Paulino

Notícias

• XXI Congresso SPE: Aveiro - 29 novembro a 2 dezembro 2013

O XXI Congresso da Sociedade Portuguesa de Estatística decorrerá este ano de 29 de Novembro a 2 de Dezembro no Hotel Meliá Ria em Aveiro.
É organizado pelo Departamento de Matemática da Universidade de Aveiro e pela Sociedade Portuguesa de Estatística (SPE).

Como tem sido habitual, o congresso da SPE será antecedido por um minicurso intitulado "**Análise de Valores Extremos: uma Introdução**" assegurado por M. Ivette Gomes – DEIO, FC-UL.

O programa científico do Congresso compreende sessões plenárias com a presença de conceituados conferencistas convidados; sessões temáticas organizadas e ainda comunicações livres (orais ou posters).

Neste congresso a SPE encerrará as atividades relativas ao Ano Internacional da Estatística, *Statistics2013*, com uma sessão especialmente dedicada ao tema. Pretende-se com esta iniciativa "*ampliar a compreensão pública do poder e impacto da Estatística em todos os aspetos da sociedade*" e ainda "*fortalecer a Estatística como uma carreira profissional, especialmente entre os jovens do ensino secundário e superior*".

É com enorme satisfação que convidamos todos a participarem neste **Congresso em Ano Internacional da Estatística**, deslocando-se a esta cidade maravilhosa intitulada "Veneza de Portugal" para uns dias de formação, partilha de conhecimentos e de estreitamento de laços através do convívio.

A Comissão Organizadora Local

• EVT2013 - Extremes in Vimeiro Today - Setembro 2013, 8-11

O Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) e a Sociedade Portuguesa de Estatística (SPE) estão a organizar um workshop de 8 a 11 de Setembro de 2013, no Vimeiro, em homenagem à nossa amiga e colega Ivette Gomes, por ocasião de seu 65º aniversário.
Embora o aniversário da Ivette seja no dia 21 de Julho, a data e o local escolhidos para este evento é muito especial para a comunidade estatística de Valores Extremos e, em particular, para a própria Ivette. De facto, foi há 30 anos, durante a primeira quinzena de Setembro, que teve lugar no Vimeiro, a conferência **Statistical Extremes and Applications**, que agora é designada, entre a comunidade de Valores Extremos, por ***EVA-zero-th conference***. Como é sabido por quem participou nesta conferência, a Ivette desempenhou um papel importante na sua organização e subsequente sucesso. Também por esse facto, este evento representa uma oportunidade para a homenagear. A Ivette, além de uma amiga especial de todos nós, é uma cientista líder com contribuições proeminentes na área da Teoria de Valores Extremos, sendo a principal responsável por colocar Portugal no roteiro de investigação científica nesta área. Além disso, esta celebração ocorre também no âmbito do **2013 - International Year of Statistics,** criando uma oportunidade para o encontro de antigos e novos colegas em Estatística.

Informações do EVT2013: *http://evt2013.weebly.com/*

Scientific Committee
Clive Anderson (University of Sheffield,UK)
Isabel Fraga Alves (University Lisbon, Portugal)
Jef Teugels (Katholieke Universiteit Leuven, Belgium)
Jürg Hüsler (University of Bern, Switzerland)
Kamil Feridun Türkman (University Lisbon, Portugal)
Manuela Neves (Technical University Lisbon, Portugal)
Rolf-Dieter Reiss (Universität Siegen,Germany)

Scientific Program
The workshop program will consist of Invited Speakers, Organized Topic Sessions, Oral and Poster Contributed Sessions.

Invited Speakers
Anthony Davison (École Poytechnique Fédérale de Lausanne, Switzerland)
Holger Rootzén (Chalmers University of Technology, Göteborg, Sweden)
Laurens de Haan (Erasmus University Rotterdam, The Netherlands, and University of Lisbon, Portugal)
Richard Davis (Columbia University, New York, USA)
Ross Leadbetter (University of North Carolina at Chapel Hill, USA)

Special Session
Maria Ivette Gomes (University of Lisbon, Portugal)

Invited Organized Sessions
Statistics for Univariate Extremes – Organizer: Armelle Guillou (Université de Strasbourg, France)
Spatial Extremes – Organizers: Jonathan Tawn (Lancaster University , UK) and Ben Shaby (University of California, Berkeley, US)
Multivariate Extremes – Organizer: Michael Falk (Universität Würzburg, Germany)
Extremes in Finance and Insurance – Organizer: Paul Embrechts (ETH Zurich, Switzerland)

Convidamos toda a comunidade científica em Estatística a assistir e participar neste evento. Ficaremos muito honrados com a vossa presença e para a Ivette será um prazer acrescido.
Junte-se a nós em 2013!

A **Comissão Organizadora Local**,
Antónia Amaral Turkman (Universidade de Lisboa)
Isabel Fraga Alves (Universidade de Lisboa)
Manuela Neves (Universidade Técnica de Lisboa)

• Encontros de Biometria

Caros Colegas:

Chama-se a atenção para o I Encontro Português de Biometria e o I Encontro Luso-Galaico de Biometria que decorrerão de 14 a 16 de Julho de 2013 na Escola de Ciências da Universidade do Minho em Braga, sendo organizados pela Sociedade Portuguesa de Estatística (SPE) e pela Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGAPEIO).

Estes Encontros são dirigidos a profissionais e utilizadores da Estatística, académicos, investigadores e estudantes, e integram-se nas celebrações em Portugal de 2013 como Ano Internacional da Estatística (vide agenda AIE 2013 em *http://www.spestatistica.pt/*).

Objetivos gerais:
1. Reforçar a projeção da Estatística no amplo campo da Biometria.
2. Ampliar o raio de ação das sociedades envolvidas a novos setores das Biociências.
3. Promover o intercâmbio e intensificar as relações dentro de cada comunidade e entre as duas comunidades estatísticas.

O programa científico dos Encontros, para além de comunicações (orais ou posters) selecionadas, compreende conferências plenárias proferidas por

- Alan Agresti - University of Florida
- Lucília Carvalho - Universidade de Lisboa
- Thomas Kneib - Universität Göttingen
- Guadalupe Gómez Melis - Universitat Politècnica de Catalunya

e um minicurso sobre Análise de Dados Categorizados Incompletos lecionado por

- Julio Singer - Universidade de São Paulo

A data-limite para submissão de resumos é 2 de abril de 2013 e a data-limite para pagamento da taxa de inscrição a preço reduzido é 5 de maio de 2013.

Informações mais detalhadas podem ser vistas em *http://biometria2013-pt.weebly.com*

Saudações cordiais

Pedro Oliveira (Universidade do Porto) e
Carlos Daniel Paulino (Universidade de Lisboa)

• METMA VI

O "VI International Workshop on Spatio-Temporal Modelling (METMAVI)" decorreu em Guimarães, entre 12 e 14 de Setembro de 2012, tendo sido organizado pelo Centro de Matemática da Escola de Ciências da Universidade do Minho. O METMAVI teve por objetivo promover o desenvolvimento e a aplicação de métodos estatísticos espaço-temporais em diferentes campos relacionados com as Ciências do Ambiente e da Saúde, tendo contado com a presença de mais de 100 congressistas.
Mais informação disponível em *http://www.metma6.com/.*

Raquel Menezes

• BIOSTATNET - Rede de Bioestatística em Espanha

Em Espanha há uma nova rede (BIOSTATNET) proposta por um grupo de bioestatísticos espanhóis, com alguns membros portugueses, e aprovada na forma de um projeto científico pelo Ministério (Espanhol) de Ciência e Inovação. BIOSTATNET é uma rede de bioestatísticos académicos ou a trabalhar na indústria, saúde, etc. em Espanha, incluindo investigadores de outras universidades em vários países. Os principais objetivos da BIOSTATNET são: a) coordenar a pesquisa e ensino em Bioestatística em Espanha, e reforçar a sua projeção internacional, b) promover a formação adequada em Bioestatística; (c) fomentar a transferência de conhecimentos e aplicações em Biomedicina.

BIOSTATNET reúne atualmente mais de 185 membros, organizados em 8 nós liderados por bioestatísticos universitários com projetos de investigação em Estatística, com experiência de ensino em Bioestatística e com colaborações estreitas com investigadores da área biomédica. Os líderes dos nós são Carmen Cadarso-Suárez - Universidad de Santiago de Compostela (Galiza), Guadalupe Gómez i Melis - Universidad Politécnica de Cataluña (Catalunya - BIO), Vicente Núñez-Antón - Universidad del País Vasco (País Vasco), María Jesús García Bayarri - Universidad de Valencia (Valencia - GEEITEMA), Antonio Martín Andrés - Universidad de Granada (Granada), María Durban Reguera - Universidad Carlos III (Madrid), Jesús López Fidalgo - Universidad de Castilla La Mancha (Castilla La Mancha - OED), e Pere Puig Casado - Universitat Autònoma de Barcelona (Catalunha - SEA). O nó da Galiza também inclui os seguintes estatísticos portugueses: Carlos Daniel Paulino (Instituto Superior Técnico - UTL), Luís Machado (Escola de Ciências - UM), Bruno de Sousa (Faculdade de Psicologia e Ciências da Educação - UC), Luzia Gonçalves (Instituto de Higiene e Medicina Tropical - UNL), Inês Sousa (Escola de Ciências - UM) e Giovani Silva (Instituto Superior Técnico - UTL).

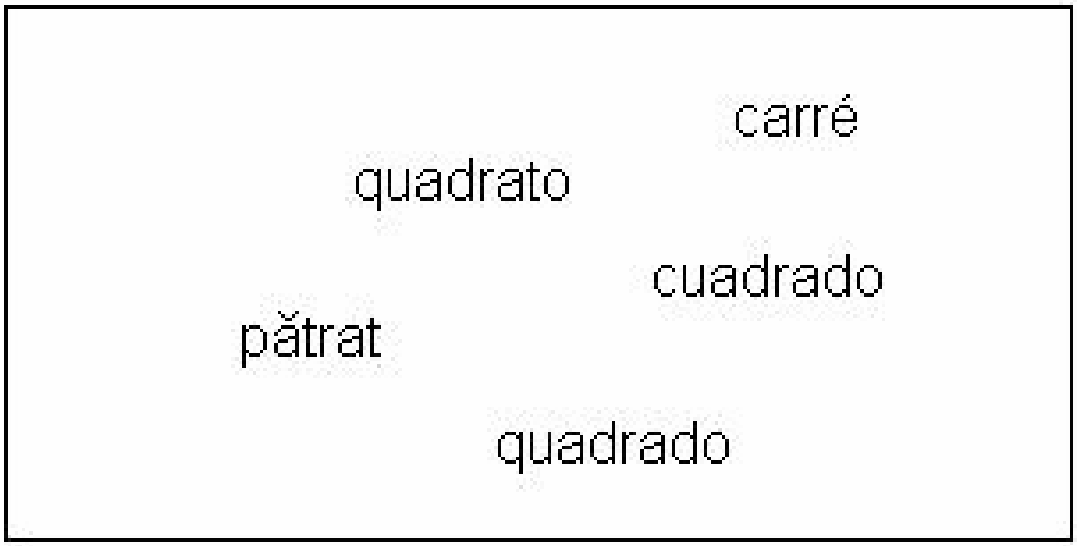
A segunda reunião da BIOSTATNET foi realizada em 25 e 26 de janeiro de 2013 na Faculdade de Medicina da Universidade de Santiago de Compostela. Esta segunda reunião serviu para discutir as metas iniciais e estabelecer novas linhas de ação, reforçar os laços existentes entre os nós da BIOSTATNET, promover a geração de novas conexões e colaborações e refletir sobre o fortalecimento e a consolidação da rede, e o seu futuro internacional. Neste sentido, o programa da segunda reunião incidiu sobre os temas: "Caminho percorrido pela rede: atividades desenvolvidas", "Bioestatísticos em Instituições Biomédicas: uma necessidade?" (mesa redonda), "Estatística, Genómica e Bioinformática" (colóquio), "Formação atual em Bioestatística" (painel), "*Past and Current Issues in Clinical Trials. A biostatistician´s perspective*" (palestra convidada, Prof. Urania Dafni - University of Athens) e "Futuro da BIOSTATNET: Relações internacionais" (colóquio).

Para mais informações, consulte *http://www.biostatnet.org*

Giovani Loiola da Silva

Enigmística de mefqa

E R A T M A O S G M



Estimação da densidade segundo o ponto de vista Bayesiano não paramétrico: o processo de Dirichlet

Vanda Inácio de Carvalho[1,2] *vanda.kinets@gmail.com*
Miguel de Carvalho[1,3] *mmbbcarvalho@gmail.com*

1 - *Departamento de Estatística, Pontificia Universidad Católica de Chile*
2 - *CEAUL, Universidade de Lisboa*
3 - *CMA, Universidade Nova de Lisboa*

Introdução

Duas abordagens comumente utilizadas na estimação da densidade são as abordagens paramétrica e não paramétrica. Enquanto que na abordagem paramétrica é assumido que os dados são provenientes de uma distribuição paramétrica conhecida que pode ser descrita utilizando um número fixo e finito de parâmetros, na abordagem não paramétrica a estrutura da distribuição que originou os dados não é especificada *a priori*, mas determinada a partir da amostra observada. Note-se que o termo não paramétrico não sugere a ausência de parâmetros, mas sim que o número e a natureza dos mesmos não é pre-determinado.
Tradicionalmente, segundo o ponto de vista não paramétrico, a estimação da densidade era feita recorrendo aos métodos clássicos, dos quais se destacam o método do núcleo (Silverman, 1986) e os splines (de Boor, 2001). Na última década, devido aos avanços computacionais, os métodos de Monte Carlo por cadeias de Markov (MCMC) tiveram um desenvolvimento ímpar e, consequentemente, a implementação dos métodos Bayesianos não paramétricos tornou-se viável.
Do ponto de vista Bayesiano paramétrico, os dados são modelados de acordo com uma família paramétrica de distribuições $\{F_\theta : \theta \in \Theta\}$ o que requer distribuições *a priori* sobre Θ. Por oposição, segundo a visão Bayesiana não paramétrica, procura-se uma classe mais geral de modelos $\{F : F \in \mathcal{F}\}$, o que requer distribuições *a priori* sobre $\mathcal{F}$, o espaço de todas as medidas de probabilidade. Uma revisão detalhada sobre métodos Bayesianos não paramétricos encontra-se em Müller e Quintana, 2004. Das diversas distribuições *a priori* existentes (árvores de Polya, *species sampling models*, etc), o processo de Dirichlet é o mais amplamente utlizado.

O Processo de Dirichlet

Definição

O processo de Dirichlet (DP) foi introduzido por Ferguson (1973) como uma distribuição *a priori*, digamos F, sobre o espaço de todas as medidas de probabilidade $\mathcal{F}$. Um processo de Dirichlet é definido por um parâmetro de concentração (ou precisão), $\alpha > 0$, e uma distribuição de centro (ou localização) F_0. Diz-se que F é distribuído de acordo com um processo de Dirichlet, DP(α, F_0), se para qualquer partição mensurável $(A_1, \ldots, A_K)$ do espaço amostral, o vector $\left(F(A_1), \ldots, F(A_K)\right)$, é distribuído segundo uma distribuição de Dirichlet com parâmetro $\left(\alpha F_0(A_1), \ldots, \alpha F_0(A_K)\right)$

$$\big(F(A_1),\dots,F(A_K)\big)\sim\text{Dirichlet}\big(\alpha F_0(A_1),\dots,\alpha F_0(A_K)\big)\,.$$

A distribuição de centro pode ser interpretada como a distribuição média do processo uma vez que

$$E\big(F(\cdot)\big)=F_0(\cdot)\,.$$

O parâmetro α é referido como o parâmetro de precisão porque controla a variância do processo

$$\text{Var}\big(F(\cdot)\big)=\frac{F_0(\cdot)(1-F_0(\cdot))}{1+\alpha}\,.$$

De facto, para valores de α elevados, a variância do processo é reduzida, havendo pois pouca variação em torno de F_0. Para uma amostra aleatória $y_1,\dots,y_n$ escrevemos então

$$y_1,\dots,y_n \mid F\sim F$$

$$F\mid\alpha,F_0\sim\text{DP}(\alpha,F_0)\,.$$

Representação quebra-vara

Distribuições obtidas a partir de um processo de Dirichlet são discretas com probabilidade um (Ferguson, 1973). Esta propriedade fica explícita na representação quebra-vara de Sethuraman (1994). Sethuraman mostrou que se F segue um processo de Dirichlet de parâmetros α e F_0, então com probabilidade um

$$F(\cdot)=\textstyle\sum_{l=1}^{\infty}\omega_l\delta_{v_l}(\cdot),$$

com $z_l\sim\text{Beta}(1,\alpha)$, $\alpha>0$, e cada ω é definido como $\omega_1=z_1$, $\omega_l=z_l\prod_{r=1}^{l-1}(1-z_r)$, $l=2,3,\dots$ e $v_l\sim F_0$, para $l=1,2,\dots$ Fica pois claro que nesta representação F é expressa como uma soma ponderada de massas pontuais. A terminologia quebra-vara surge porque começando com uma vara de tamanho equivalente à unidade, ω_1 é a proporção de vara quebrada e atribuída a v_1, ω_2 é a proporção da vara restante ($1-z_1$) atribuída a v_2, e assim sucessivamente.

A representação quebra-vara é provavelmente a definição mais versátil do processo de Dirichlet, tendo sido muitíssimo explorada para gerar algoritmos MCMC eficientes.

Mistura por processos de Dirichlet

A natureza discreta do processo de Dirichlet inviabiliza a sua utilização para modelar dados contínuos. Com o objectivo de ultrapassar esta limitação, Lo (1984) propôs uma mistura entre uma distribuição contínua e uma medida de probabilidade aleatória que segue um processo de Dirichlet. Ou seja, propôs construir um modelo de mistura via processos de Dirichlet. De forma hierárquica, temos

$$y_1,\dots,y_n\sim F$$
$$F(\cdot)\sim\int F_\theta\,dG(\theta)$$
$$G\mid\alpha,\psi\sim\text{DP}(\alpha,G_0),\quad G_0=G_0(\cdot\mid\psi)$$
$$\alpha,\psi\sim p(\alpha)p(\psi)\,.$$

Este modelo representa uma mistura de distribuição onde a medida misturadora é a distribuição G. Devido à natureza discreta do processo de Dirichlet, o modelo de mistura por processo de Dirichlet

divide as observações em grupos independentes, ficando assim claro o seu potencial para agrupamento de dados.

A representação quebra-vara permite-nos expressar F como uma mistura infinita de distribuições paramétricas

$$F(\cdot) = \sum_{l=1}^{\infty} \omega_l F_\theta(\cdot; v_l) .$$

Analogamente, a densidade pode ser expressa como

$$f(\cdot) = \sum_{l=1}^{\infty} \omega_l f_\theta(\cdot; v_l) .$$

Análise de dados reais

Nesta secção iremos ilustrar o anteriormente exposto recorrendo a um conjunto de dados de velocidades (1000 km/seg) de 82 galáxias da região de Corona Borealis (Roeder, 1990).

Como podemos constatar por observação da Figura 1, um ajustamento paramétrico baseado na distribuição normal é inadequado para este conjunto de dados.

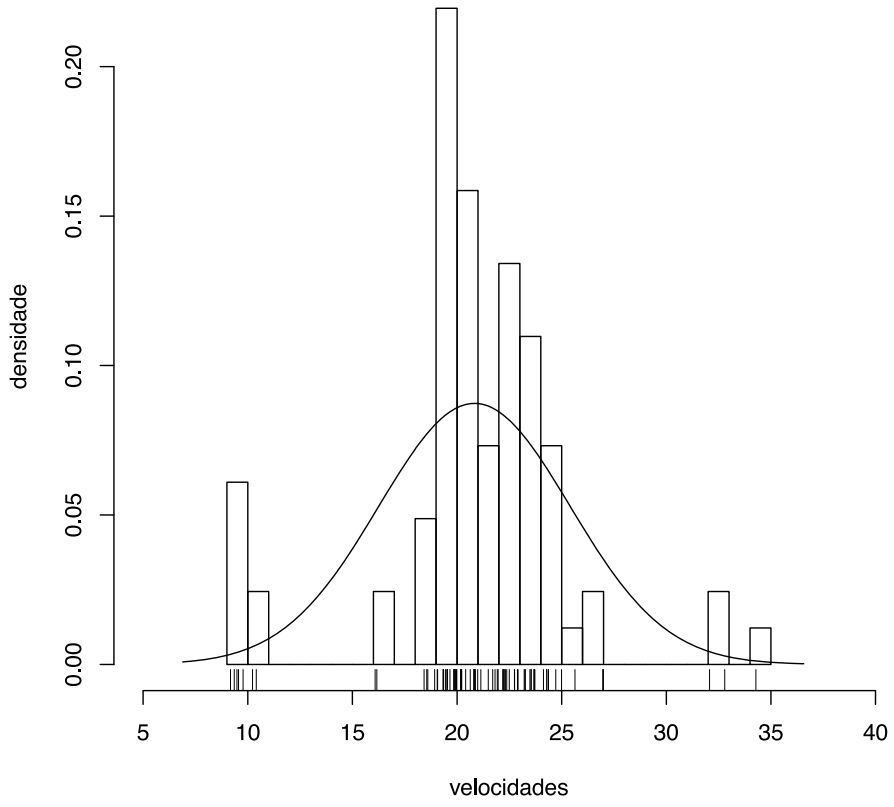


Figura 1. Histograma das velocidades das galáxias e densidade referente a um ajustamento paramétrico através de um modelo normal.

Ajustámos então o seguinte modelo de mistura por processo de Dirichlet de distribuições normais

$$y_i \sim \int \phi(y_i \mid \mu, \sigma^2) \mathrm{d}G(\mu, \sigma^2), i = 1, \ldots, 82$$

$$G \mid \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$

com $G_0 \equiv \phi(\mu \mid m, S) GI(\sigma^2 \mid a, b)$, onde ϕ representa a densidade da distribuição normal estandardizada, GI denota a distribuição gama inversa e m, S, a e b são hiperparâmetros aos quais são atribuídas distribuições *a priori* (ver Jara e outros (2011) para mais detalhes). Este modelo produziu o ajustamento que é reportado na Figura 2.

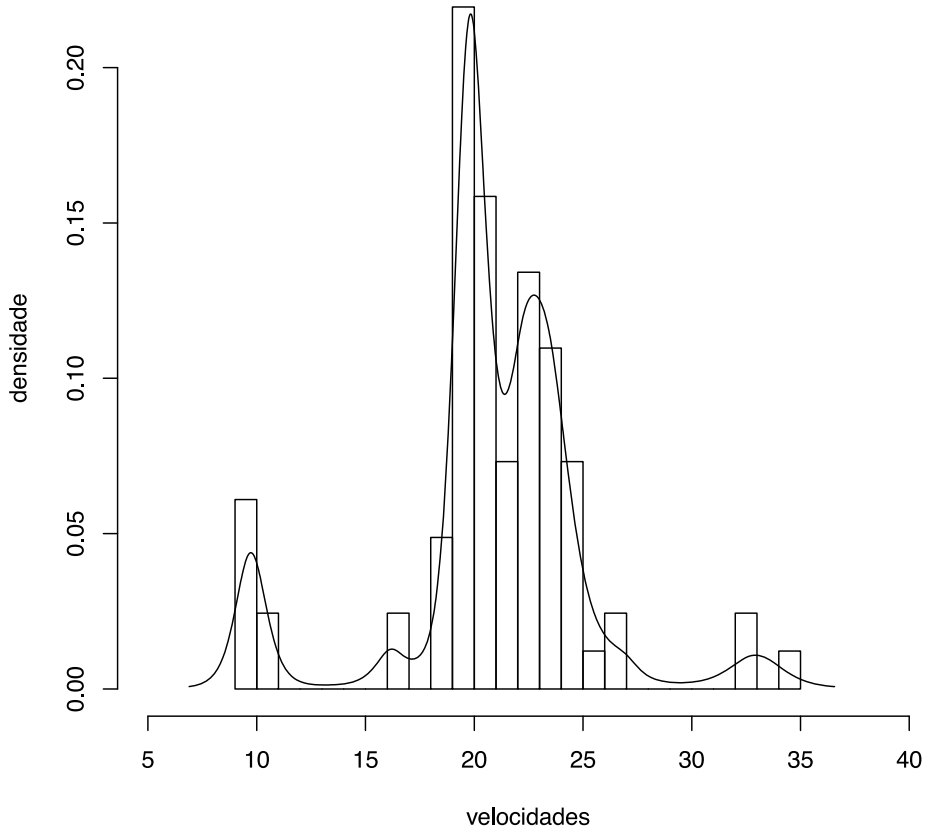


Figura 2. Histograma das velocidades das galáxias e densidade referente a um ajustamento Bayesiano não paramétrico através de uma mistura por processos de Dirichlet de distribuições normais.

Como podemos constatar este modelo é bastante mais adequado para os dados em questão. Tal facto pode também ser comprovado através da estatística LPML (do inglês, *log pseudo marginal likelihood*) que para este modelo é de -210.71, enquanto que para o ajustamento baseado na distribuição normal é de -243.1.

Referências

de Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209–230.

Jara, A., Hanson, T. E., Quintana, F. A., Muller. P. e Rosner, G. L. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40, 1–30.

Lo, A.Y. (1984). Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12, 351–357.

Müller, P. e Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science,* 19, 95–110.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, 92, 894–902.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statistica Sinica,* 19, 639–650.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Combinando testes de Mardia e BHEP na avaliação duma hipótese multivariada de normalidade

Carlos Tenreiro, *tenreiro@mat.uc.pt*

CMUC, Departamento de Matemática, Universidade de Coimbra

1. Introdução

Sendo $X_1, \ldots, X_n, \ldots$ uma sucessão de cópias independentes dum vector d-dimensional e absolutamente contínuo X, com densidade de probabilidade f, desconhecida, o problema do teste duma hipótese multivariada de normalidade (MVN) é o de, com base em $X_1, \ldots, X_n$, testar a hipótese

$$H_0 : f \in \mathscr{N}_d,$$

contra uma hipótese alternativa geral, onde $\mathscr{N}_d$ é a família das densidades de probabilidade normais sobre $\mathbb{R}^d$. Este é um problema clássico na literatura estatística sobre o qual muito trabalho tem sido desenvolvido como atestam Mecklin e Mundfrom (2000) que referem a existência de cerca de cinquenta procedimentos para testar uma hipótese MVN. Apesar disso, este assunto continua a despertar o interesse dos investigadores como confirmam os trabalhos mais recentes de Liang et al. (2005), Mecklin e Mundfrom (2005), Székely e Rizzo (2005), Sürücü (2006), Arcones (2007), Farrel et al. (2007), Chiu e Liu (2009), Liang et al. (2009), Tenreiro (2009, 2011) e Ebner (2012). O facto de muito dos métodos estatísticos multivariados como a ANOVA, regressão multivariada, análise discriminante ou correlação canónica, dependerem da aceitação duma hipótese MVN pode explicar este interesse continuado. Para mais bibliografia sobre o tema veja-se Csörgő (1986), Rayner e Best (1989, p. 98–109), Thode (2002, p. 181–224) e os artigos de revisão de Henze (2002) e Mecklin e Mundfrom (2004).

Neste texto, baseado, no essencial, nos nossos trabalhos acima citados, iremos centrar a nossa atenção em testes duma hipótese MVN que gozam da propriedade natural de serem invariantes para transformações de localização e de escala dos dados. Sendo $T_n = T_n(X_1, \ldots, X_n)$ a estatística associada a um tal teste, vale assim a igualdade

$$T_n(AX_1 + b, \ldots, AX_n + b) = T_n(X_1, \ldots, X_n),$$

para toda a matriz não-singular A e todo o vector $b \in \mathbb{R}^d$. Apesar de grande parte dos testes propostos na literatura não satisfazerem a propriedade anterior, satisfazem-na alguns dos mais utilizados testes de normalidade, como são os casos dos testes clássicos de Mardia e dos testes BHEP (Baringhaus-Henze-Epps-Pulley) a que faremos referência detalhada neste texto. Atendendo à propriedade de invariância,

os pontos críticos destes procedimentos de teste podem ser estimados através de experiências de Monte Carlo sob H_0. Sobre outros testes invariantes para transformações afins dos dados veja-se Henze (2002) e Székely e Rizzo (2005).

2. Os testes de Mardia

De entre a vasta família de procedimentos de testes para uma hipótese MVN, os testes de Mardia (1970), baseados em medidas multivariadas de assimetria e curtose, desempenham um papel importante, estando entre os testes mais recomendados para testar uma hipótese MVN (ver Romeu e Ozturk, 1993; Mecklin e Mundfrom, 2005, e as referências bibliográficas respectivas). Denotando por

$$\bar{X}_n = n^{-1}\sum_{j=1}^{n} X_j \quad \text{e} \quad S_n = n^{-1}\sum_{j=1}^{n}(X_j-\bar{X}_n)(X_j-\bar{X}_n)',$$

a média e a matriz de covariância amostrais, respectivamente, as estatísticas MS (multivariate skewness) e MK (multivariate kurtosis) de Mardia são definidas por

$$\text{MS} = nb_{1,d}$$

e

$$\text{MK} = \sqrt{n}\,|\,b_{2,d} - d(d+2)|,$$

com

$$b_{1,d} = \frac{1}{n^2}\sum_{j,k=1}^{n}(Y_j'Y_k)^3 \quad \text{e} \quad b_{2,d} = \frac{1}{n}\sum_{j=1}^{n}(Y_j'Y_j)^2,$$

onde

$$Y_j = S_n^{-1/2}(X_j-\bar{X}_n),\ j=1,\ldots,n,$$

são os resíduos standardizados e $S_n^{-1/2}$ é a raiz quadrada definida positiva da inversa da matriz de covariância amostral. Sob a hipótese H_0, valem as convergências em distribuição

$$nb_{1,d} \xrightarrow{d} 6\chi^2_{d(d+1)(d+2)/6}$$

e

$$\sqrt{n}\,(\,b_{2,d} - d(d+2)) \xrightarrow{d} N(0, 8d(d+2))$$

(Mardia, 1970), e, assim, o teste MS rejeita H_0 para valores grandes de $b_{1,d}$ enquanto que o teste MK rejeita H_0 para valores pequenos ou grandes de $b_{2,d}$.

Apesar de serem invariantes para transformações afins dos dados, os testes de Mardia, tal como a quase totalidade dos testes de normalidade propostos na literatura, não são convergentes para todas as distribuições alternativas $f \notin \mathscr{N}_d$. Mesmo para amostras de grande tamanho, potencialmente infinitas, existem distribuições alternativas que não são detectadas por tais testes. Denotando por

$$\beta_{1,d} = \text{E}((X_1-\mu)'\Sigma^{-1}(X_2-\mu))^3 \quad \text{e} \quad \beta_{2,d} = \text{E}((X_1-\mu)'\Sigma^{-1}(X_1-\mu))^2,$$

os parâmetros correspondentes às medidas amostrais anteriores de assimetria e de curtose, onde μ e Σ são a média e a matriz de covariância de X, Baringhaus e Henze (1992) mostraram que se $\text{E}(X'X)^3 < \infty$, o teste baseado em MS é convergente se e só se $\beta_{1,d} > 0$, e Henze (1994) provou que se $\text{E}(X'X)^4 < \infty$, o teste baseado em MK é convergente se e só se $\beta_{2,d} \neq d(d+2)$. Assim, apesar dos testes baseados em MS e MK poderem possuir uma potência elevada para alternativas com $\beta_{1,d} > 0$ ou $\beta_{2,d} \neq d(d+2)$, ambos os testes podem apresentar um fraco comportamento para alternativas com os mesmos coeficientes de assimetria e curtose que a distribuição normal multivariada, isto é, para distribuições $f \notin \mathscr{N}_d$ com $\beta_{1,d} = 0$ e $\beta_{2,d} = d(d+2)$. Este problema pode ser também sentido em outros testes baseados em estatísticas que combinam as medidas anteriores de assimetria e curtose de forma a produzirem um teste com boas

propriedades globais (*omnibus test*), como são os casos dos testes propostos por Mardia e Foster (1983), Horswell e Looney (1992) ou Doornik e Hansen (1994).

3. Os testes BHEP

Em alternativa aos testes de Mardia, ou a combinações destes como as que acabámos de referir, podemos optar por utilizar testes convergentes para todas as distribuições alternativas, como são os casos dos testes BHEP (Baringhaus–Henze–Epps–Pulley), introduzidos por Baringhaus e Henze (1988) e Henze e Zirkler (1990), e que estendem o teste de normalidade de Epps e Pulley (1983) ao contexto multivariado. A estatística de teste BHEP é baseada na distância L_2 ponderada entre a função característica amostral associada aos resíduos standardizados

$$\Psi_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(it'Y_j\right), \quad t \in \mathbb{R}^d,$$

e a função característica Φ da distribuição normal standard em $\mathbb{R}^d$ com densidade

$$\phi(x) = (2\pi)^{-d/2}\exp(-x'x/2), \quad x \in \mathbb{R}^d.$$

A função de peso é dada por

$$t \to |\Phi_h(t)|^2 = \exp(-h^2 t't),$$

onde Φ_h é a função característica de $\phi_h(\cdot) = \phi(\cdot/h)/h^d$ e h é um número real estritamente positivo que deve ser escolhido pelo utilizador. Assim, o teste BHEP é baseado na estatística

$$\mathrm{B}(h) = n\int |\Psi_n(t) - \Phi(t)|^2 |\Phi_h(t)|^2 dt = (2\pi)^d \frac{1}{n}\sum_{i,j=1}^{n} Q(Y_i, Y_j; h),$$

com

$$Q(u,v;h) = \phi_{(2h^2)^{1/2}}(u-v) - \phi_{(1+2h^2)^{1/2}}(u) - \phi_{(1+2h^2)^{1/2}}(v) + \phi_{(2+2h^2)^{1/2}}(0),$$

para $u, v \in \mathbb{R}^d$. A simplicidade da expressão anterior para $\mathrm{B}(h)$, justifica a escolha da função de peso considerada. Reparemos ainda que a estatística de teste depende das observações através das quantidades $||Y_i - Y_j||^2$ e $||Y_i||^2$, onde $||\cdot||$ representa a norma euclidiana em $\mathbb{R}^d$. Tais quantidades dependem apenas de S_n^{-1}, não sendo assim sequer necessário calcular $S_n^{-1/2}$ para obter $\mathrm{B}(h)$. Para uma referência recente sobre testes de ajustamento baseados na função característica ver Jiménez-Gamero et al. (2009).

O comportamento assintótico de $\mathrm{B}(h)$ sob a hipótese nula, sob uma alternativa fixa e sob uma sucessão de alternativas locais, foi estudado por diversos autores como são os casos de Baringhaus e Henze (1988), Csörgő (1989), Henze e Zirkler (1990) e Henze e Wagner (1997). Em particular, para cada $h > 0$, $\mathrm{B}(h)$ possui como distribuição nula assintótica uma soma ponderada de qui-quadrados independentes sendo o teste associado convergente para toda a alternativa fixa.

É interessante notar que no caso da densidade f ser de quadrado integrável, a estatística de teste anterior pode também ser interpretada como sendo baseada na distância L_2 entre o estimador da densidade de Parzen-Rosenblatt obtido a partir dos resíduos standardizados, com núcleo (kernel) $K = \phi$ e janela (*bandwidth*) h, e a convolução $K_h * \phi$, que pode ser vista como uma aproximação de ϕ quando h tende para zero (ver Henze e Zirkler, 1990; Bowman e Foster, 1993; Fan, 1998). Neste sentido, quando tomamos $h = h_n \to 0$, $n \to \infty$, o teste baseado em $\mathrm{B}(h)$ deve ser interpretado como um teste baseado na densidade de probabilidade de X e não na sua função característica. Deixando o parâmetro h de ser fixo e passando a desempenhar o papel de janela do estimador do núcleo da densidade, as propriedades assintóticas do teste baseado em $\mathrm{B}(h)$ são distintas das que acima descrevemos. No entanto, o teste resultante continua a ser convergente para toda a distribuição alternativa. Os testes de ajustamento baseados no estimador do núcleo da densidade de probabilidade foram primeiramente estudados por Bickel e Rosenblatt (1973). Para mais detalhes sobre estes testes vejam-se os trabalhos de Rosenblatt (1975), Fan (1994), Tenreiro (1996, 2007), Gouriéroux e Tenreiro (2001) e Henze (2002).

4. A selecção do parâmetro h nos testes BHEP

Apesar dos testes BHEP serem convergentes para todas as distribuições alternativas, independentemente do valor tomado para o parâmetro $h > 0$, a sua potência depende fortemente da escolha de h (cf. Henze e Wagner, 1997; Tenreiro, 2009). As diferentes escolhas de h consideradas na literatura foram analisadas em Tenreiro (2009) que, com base num vasto estudo de simulação para dimensões $2 \leq d \leq 15$, sugere duas escolhas empíricas para h. Assim, a janela

$$h = h_{\mathrm{L}} := 0.448 + 0.026\, d$$

mostrou ser adequada para alternativas com 'caudas leves' ou alternativas aproximadamente simétricas, e a janela

$$h = h_{\mathrm{P}} := 0.928 + 0.049\, d,$$

mostrou-se adequada para alternativas com 'caudas pesadas' ou alternativas moderadamente assimétricas. Estas escolhas estão de acordo com uma interpretação heurística das propriedades de potência do teste em termos da janela h. Para valores grandes de h, a ponderação $t \to \exp(-h^2 t't)$ coloca a maior parte da sua massa numa vizinhança da origem onde a função característica reflecte o comportamento da cauda da distribuição, sendo por isso de esperar que o teste BHEP seja sensível a alternativas com 'caudas pesadas'. Por razões análogas, será de esperar que o teste possa ser mais sensível para alternativas com 'caudas leves' para valores pequenos de h. Na ausência de informação adicional sobre o tipo de alternativa em causa, recomenda-se a utilização da janela combinada

$$h = \bar{h} := \frac{1}{2} h_{\mathrm{L}} + \frac{1}{2} h_{\mathrm{P}},$$

que conduz a um teste com boas propriedades de potência para um vasto conjunto de distribuições alternativas (cf. Tenreiro, 2009).

5. Combinando testes de Mardia e BHEP

Apesar das boas propriedades reveladas pelo teste BHEP baseado em $\mathrm{B}(\bar{h})$, para diversas distribuições alternativas este teste é claramente superado por um dos testes de Mardia. O teste MS revela-se particularmente eficiente na detecção de alternativas assimétricas ou com 'caudas pesadas', enquanto que o teste MK mostra-se especialmente eficaz na detecção de alternativas com 'caudas leves' (cf. Henze e Zirkler, 1990; Romeu e Ozturk, 1993). A ideia de combinar os testes de Mardia e BHEP de forma a obter um teste que possa usufruir das boas propriedades de cada um dos testes intervenientes na combinação surge assim de forma natural, sendo a mesma analisada em Tenreiro (2011) utilizando um método, considerado em Fromont e Laurent (2006), que pode ser interpretado como um melhoramento do método de Bonferroni clássico.

Sendo $T_{n,h}, h \in H$, um conjunto finito de estatísticas de testes invariantes para transformações afins dos dados, o teste múltiplo proposto rejeita a hipótese MVN se pelo menos uma das estatísticas $T_{n,h}$ for maior que o seu quantil de ordem $1 - u_{n,\alpha}$ sob a hipótese nula, onde $u_{n,\alpha}$ é calibrado de forma que o teste múltiplo tenha um nível de significância não superior ao nível nominal α fixado à partida. Mais precisamente, sendo $c_{n,h}(u)$ o quantil de ordem $1-u$ da estatística de teste $T_{n,h}$ sob H_0 e considerando a estatística corrigida

$$\mathbf{T}_n(u) = \max_{h \in H} \left(T_{n,h} - c_{n,h}(u) \right),$$

o teste múltiplo rejeita a hipótese H_0 sempre que

$$\mathbf{T}_n(u_{n,\alpha}) > 0$$

onde

$$u_{n,\alpha} = \sup \left\{ u \in]0,1[: \mathrm{P}_0(\mathbf{T}_n(u) > 0) \leq \alpha \right\},$$

e P_0 é a distribuição normal standard sobre $\mathbb{R}^d$. Na prática, o nível $u_{n,\alpha}$, segundo o qual cada um dos testes baseados nas estatísticas $T_{n,h}$ é aplicado, é estimado através de experiências de Monte Carlo sob H_0.

Em Tenreiro (2011) o método geral anterior é utilizado para combinar os testes baseados nas estatísticas $T_{n,1} = \mathrm{MS}$, $T_{n,2} = \mathrm{MK}$, $T_{n,3} = \mathrm{B}(h_\mathrm{L})$ e $T_{n,4} = \mathrm{B}(h_\mathrm{P})$. Outras combinações de testes invariantes duma hipótese MVN são naturalmente possíveis. Dum ponto de vista teórico, o teste resultante da combinação dos quatro testes anteriores é convergente para toda distribuição alternativa e, para n fixo, o seu nível de significância não é superior ao nível nominal $\alpha \in]0,1[$ fixado à partida. O desempenho a distância finita do procedimento múltiplo foi avaliado sob H_0, revelando o teste possuir um nível de significância efectivo muito próximo do nível nominal α, e sob um vasto conjunto de distribuições alternativas que são habitualmente consideradas na literatura em estudos deste tipo. Como seria de esperar, fixada uma distribuição alternativa, o teste múltiplo proposto nunca é o melhor dos testes incluídos na combinação. No entanto, ele herda as boas propriedades de cada um dos testes envolvidos no procedimento múltiplo, revelando um bom desempenho para todas as alternativas consideradas. Tendo em conta que numa situação real a formulação duma hipótese alternativa é em geral impossível, a propriedade anterior é muito interessante não sendo a mesma partilhada por nenhum dos testes envolvidos na combinação considerada. Além disso, o teste mostra um bom desempenho global quando comparado com os mais recomendados testes duma hipótese MVN, o que nos leva a considerá-lo uma alternativa válida aos diversos testes propostos na literatura. Uma função escrita em R para implementar este teste está disponível no endereço http://www.mat.uc.pt/∼tenreiro/publications.

Bibliografia

Arcones, M.A., 2007. Two tests for multivariate normality based on the characteristic function. *Math. Methods Statist*. 16, 177–201.

Baringhaus, L., Henze, N., 1988. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika* 35, 339–348.

Baringhaus, L., Henze, N., 1992. Limit distributions for Mardia's measure of multivariate skewness. *Ann. Statist*. 20, 1889–1902.

Bickel, P.J., Rosenblatt, M., 1973. On some global measures of the deviations of density function estimates. *Ann. Statist*. 1, 1071–1095.

Bowman, A.W., Foster, P.J., 1993. Adaptive smoothing and density-based tests of multivariate normality. *J. Amer. Statist. Assoc*. 88, 529–537.

Chiu, S.N., Liu, K.I., 2009. Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions. *Comput. Statist. Data Anal*. 53, 3817–3834.

Csörgő, S., 1986. Testing for normality in arbitrary dimension. *Ann. Statist*. 14, 708–723.

Csörgő, S., 1989. Consistency of some tests for multivariate normality. *Metrika* 36, 107–116.

Doornik, J.A., Hansen, H., 1994. An omnibus test for univariate and multivariate normality. Working Paper, Nuffield College, Oxford.

Ebner, B., 2012. Asymptotic theory for the test for multivariate normality by Cox and Small. *J. Multivariate Anal*. 111, 368–379.

Epps, T.W., Pulley, L.B., 1983. A test for normality based on the empirical characteristic function. *Biometrika* 70, 723–726.

Fan, Y., 1994. Testing the goodness of fit of a parametric density function by kernel method. *Econometric Theory* 10, 316–356.

Fan, Y., 1998. Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory* 14, 604–621.

Farrel, P.J., Salibian-Barrera, M., Naczk, K., 2007. On tests for multivariate normality and associated simulation studies. *J. Stat. Comput. Simul.* 77, 1065–1080.

Fromont, M., Laurent, B., 2006. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.* 34, 680–720.

Gouriéroux, C., Tenreiro, C., 2001. Local power properties of kernel based goodness of fit tests. *J. Multivariate Anal.* 78, 161–190.

Henze, N., Zirkler, B., 1990. A class of invariante consistent tests for multivariate normality. *Comm. Stat. Theory Methods* 19, 3595–3617.

Henze, N., 1994. On Mardia's kurtosis test for multivariate normality. *Comm. Statist. Theory Methods* 23, 1047–1061.

Henze, N., 2002. Invariant tests for multivariate normality: a critical review. *Statist. Papers* 43, 467–506.

Henze, N., Wagner, T., 1997. A new approach to the BHEP tests for multivariate normality. *J. Multivariate Anal.* 62, 1–23.

Horswell, R.L., Looney, S.W., 1992. A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *J. Stat. Comput. Simul.* 42, 21–38.

Jiménez-Gamero, M.D., Alba-Fernández, V., Muñoz-García, J., Chalco-Cano, Y., 2009. Goodness-of-fit tests based on empirical characteristic functions. *Comput. Statist. Data Anal.* 53, 3957–3971.

Liang, J., Pan, W.S.Y., Yang, Z.-H., 2005. Characterization-based Q-Q plots for testing multinormality. *Statis. Probab. Lett.* 70, 183–190.

Liang, J., Tang, M.-L., Chan, P.S., 2009. A generalized Shapiro–Wilk W statistic for testing high-dimensional normality. *Comput. Statist. Data Anal.* 53, 3883–3891.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519–530.

Mardia, K.V., Foster, K., 1983. Omnibus tests of multinormality based on skewness and kurtosis. *Comm. Statist. Theory Methods* 12, 207–221.

Mecklin, C.J., Mundfrom, D.J., 2000. Comparing of the power of classical and newer tests of multivariate normality. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 24–28, 2000.

Mecklin, C.J., Mundfrom, D.J., 2004. An appraisal and bibliography of tests for multivariate normality. *Int. Stat. Rev.* 72, 123–138.

Mecklin, C.J., Mundfrom, D.J., 2005. A Monte Carlo comparison of Type I and Type II error rates of tests of multivariate normality. *J. Stat. Comput. Simul.* 75, 93–107.

Rayner, J.C.W., Best, D.J., 1989. *Smooth tests of goodness of fit*. New York: Oxford University Press.

Romeu, J.L., Ozturk, A., 1993. A comparative study of goodness-of-fit tests for multivariate normality. *J. Multivariate Anal.* 46, 309–334.

Rosenblatt, M., 1975. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* 3, 1–14.

Sürücü, B., 2006. Goodness-of-fit tests for multivariate distributions. *Comm. Statist. Theory Methods* 35, 1319–1331.

Székely, G.J., Rizzo, M.L., 2005. A new test for multivariate normality. *J. Multivariate Anal.* 93, 58–80.

Tenreiro, C. 1996. Tests d'ajustement à une densité fondés sur un estimateur non paramétrique à noyau pour des observations dépendantes. *Ann. Économ. Statist.* 43, 129–148.

Tenreiro, C., 2007. On the asymptotic behaviour of location-scale invariant Bickel-Rosenblatt tests. *J. Statist. Plann. Inference* 137, 103–116. Erratum: 139, 2115, 2009.

Tenreiro, C., 2009. On the choice of the smoothing parameter for the BHEP goodness-of-fit test. *Comput. Statist. Data Anal.* 53, 1038–1053.

Tenreiro, C., 2011. An affine invariant multiple test procedure for assessing multivariate normality. *Comput. Statist. Data Anal.* 55, 1980–1992.

Thode, Jr., H.C., 2002. *Testing for normality*. New York: Marcel Dekker.

Estimação da distribuição de um processo espacial recorrendo a um variograma de indicatriz tipo núcleo

Raquel Menezes, *rmenezes@math.uminho.pt*

Universidade do Minho

1 Introdução

No contexto de dados geo-referenciados, existem situações práticas para as quais é bastante útil a estimação da função de distribuição do processo espacial, uma vez que define a probabilidade da variável envolvida não exceder um determinado valor de corte, permitindo a construção de mapas de risco. A aproximação da função de distribuição poderá se basear na aplicação de métodos de krigagem da indicatriz ou na estimação do *sill* (patamar), que exigem condições de estacionaridade do processo aleatório ou a estimação do variograma (ou covariograma) da indicatriz. Em Garcia-Soidán & Menezes (2012), sugere-se um estimador tipo-núcleo como alternativa não-paramétrica ao estimador proposto por Matheron para o variograma da indicatriz. Nesse trabalho, alguns estudos numéricos envolvendo dados simulados são apresentados para ilustrar o melhor desempenho desta proposta versus a clássica. Pretende-se, aqui, apresentar os principais resultados desta proposta, e descrever a sua aplicação a um conjunto de dados ambientais (concentrações de nitratos em águas subterrâneas) recolhidos na zona de Beja, permitindo a construção de mapas de risco de poluição da referida região.

2 Principais resultados

A aproximação da função de distribuição de um processo aleatório espacial $\{Z(s): s \in D \subset R^d\}$ poderá ser importante, por exemplo, para estimar depósitos de um determinado metal ou avaliar a contaminação do solo, ou em esquemas de classificação para a análise de mapas. Tipicamente, um número finito de localizações espaciais s_i são selecionadas, $1 \le i \le n$, onde medições da variável envolvida são recolhidas e usadas para calcular informação para toda a região de observação, incluindo as localizações não amostradas.

2.1 Métodos para aproximar $F(.)$

Neste contexto, poderá ser adotada a krigagem da indicatriz, um método não paramétrico eficiente para aproximar a função de distribuição (Journel, 1983).
A abordagem da indicatriz é baseada na interpretação da função de distribuição como a esperança de uma variável aleatória tipo indicatriz, nomeadamente:

$$P(Z(s) \le x) = F_s(x) = E[I_Z(s,x)]$$

onde $I_Z(s,x) = 1$ se $Z(s) \le x$ e zero caso contrário. Na prática, a distribuição é aproximada em Q valores de corte x_q, pré-definidos, e os restantes valores são obtidos por interpolação.
Pelo teorema da projeção analisado em Goovaerts (1997), o estimador de mínimos quadrados (kriging) da função indicatriz é também o estimador de mínimos quadrados da sua esperança.

Consequentemente, uma aproximação da função de distribuição na localização s e valor de corte x é dada pelo preditor de krigagem da indicatriz $I_Z(s,x)$, obedecendo à seguinte expressão:

$$\hat{I}_Z(s,x) = \sum_{i=1}^{n} \lambda_i I_Z(s_i,x)$$

onde os parâmetros λ_i são obtidos resolvendo as correspondentes equações de kriging. Tal implica a estimação de um variograma (ou função de covariância) para cada valor de corte, denominado o variograma da indicatriz (ou covariograma da indicatriz).

Journel (1983) propõe um método alternativo para a aproximação da função distribuição. Debaixo do pressuposto que o variograma da indicatriz pode ser aproximado e o respetivo *sill* (patamar) adequadamente estimado, a distribuição para o correspondente valor de corte poderá ser calculada resolvendo-se uma equação de segundo grau, em vez do sistema de equações de kriging. Para tal, deve-se considerar a expressão (1), que relaciona o *sill* $S(x)$ com a função de distribuição para o valor x.

A validade dos dois métodos referidos para a aproximação de $F_s(x)$ é garantida debaixo dos seguintes pressupostos:

a) $E\left[I_Z(s',x) - I_Z(s'',x)\right] = 0, \forall s', s'' \in D$

b) $Var\left[I_Z(s',x) - I_Z(s'',x)\right] = 2\gamma_{I_Z}(s'-s'',x), \forall s', s'' \in D$

onde γ_{I_Z} identifica o variograma da indicatriz de Z. As hipóteses a) e b) são verificadas para um processo estritamente estacionário, embora esta condição possa ser eventualmente um pouco restritiva em aplicações a dados reais. Uma alternativa poderá ser apenas considerar a hipótese a) de um forma "local", ou seja, assumir a) apenas numa vizinhança de s.
Adicionalmente, iremos considerar que

$$Z(s) = \mu(s) + Y(s)$$

onde $\{Y(s) : s \in D \subset R^d\}$ é um processo aleatório estritamente estacionário de média zero e $\mu(.)$ representa uma tendência determinística, nomeadamente $E[Z(s)] = \mu(s), \forall s \in D$. Assumindo-se que $\mu(.)$ pode ser adequadamente caracterizado, então a estimação da função F_s pode ser reduzida à estimação da distribuição de Y. Considerando-se que G e $G_{s',s''}$ identificam as funções de distribuição univariada e bivariada de Y, tem-se

$$P(Y(s') \le x) = G(x), \forall s' \in D, x \in R$$
$$P(Y(s') \le x, Y(s'') \le y) = G_{s',s''}(x,y) = G_{0,s'-s''}(x,y), \forall s', s'' \in D, x, y \in R$$

Verifica-se então que, $\forall t \in R^d$:

$$2\gamma_{I_Y}(t,x) = Var\left[I_y(s,x) - I_y(s+t,x)\right] = 2\left(G(x) - G_{0,t}(x,x)\right)$$

Por conseguinte, as condições a) e b) prevalecem para o processo Y. Os métodos "krigagem da indicatriz" ou "estimação do *sill*" podem, então, ser adotados para aproximar a função de distribuição G e, por sua vez, a distribuição de Z pode ser obtida considerando

$$F_s(x) = G(x - \mu(s)).$$

2.2 Estimadores do variograma da indicatriz

Um estimador do variograma da indicatriz é dado pelo variograma experimental, obtido pelo método dos momentos (Matheron, 1963):

$$2\hat{\gamma}_{I_Y}(t,x) = \frac{1}{|N(t)|}\sum_{(i,j)\in N(t)}(I_Y(s_i,x) - I_Y(s_j,x))^2$$

onde $|N(t)|$ denota o número de pares distintos em $N(t) = \{(i,j) : s_i, s_j \in D, s_i - s_j = t\}$.

Em Garcia-Soidán & Menezes (2012), propõe-se um variograma alternativo baseado na estimação tipo núcleo, dado por:

$$2\hat{\gamma}_{I_Y,h}(t,x) = \frac{\sum_{i\neq j} K\left(\frac{t-(s_i-s_j)}{h}\right)\left(I_Y(s_i,x) - I_Y(s_j,x)\right)^2}{\sum_{i\neq j} K\left(\frac{t-(s_i-s_j)}{h}\right)}$$

onde K representa uma função núcleo d-dimensional e h o parâmetro janela. O estimador proposto, semelhante ao apresentado em Garcia-Soidán (2007), goza de propriedades como a consistência, sendo representado por uma função mais suave do que a associada ao estimador de Matheron. Por conseguinte, é expectável que ofereça um melhor desempenho quer para o cálculo de um variograma válido (i.e. condicionalmente negativo-definido; ver por exemplo Pardo-Igúzquiza (1998)) quer para o cálculo do *sill* $S(x)$, necessários para a aproximação da função distribuição. De acordo com Journel (1983), tem-se

$$S(x) = \lim_{\|t\|\to\infty} \gamma_{I_Y,h}(t,x) = G(x) - G(x)^2 \qquad (1)$$

o que nos irá permitir aproximar $G(x)$ e, por sua vez, $F_s(x)$.

3. Aplicação a dados ambientais

Como exemplo de motivação para a aplicação da metodologia previamente apresentada, adotámos um conjunto de dados com valores de concentrações de nitratos em águas subterrâneas. Os dados foram recolhidos na zona de Beja em 1998 e 2000 (ver Figura 1). Conforme descrito em Paralta & Ribeiro (2003), os níveis da concentração de nitrato dependem da época do ano e a nossa análise irá se restringir ao mesmo mês de cada ano (Julho).

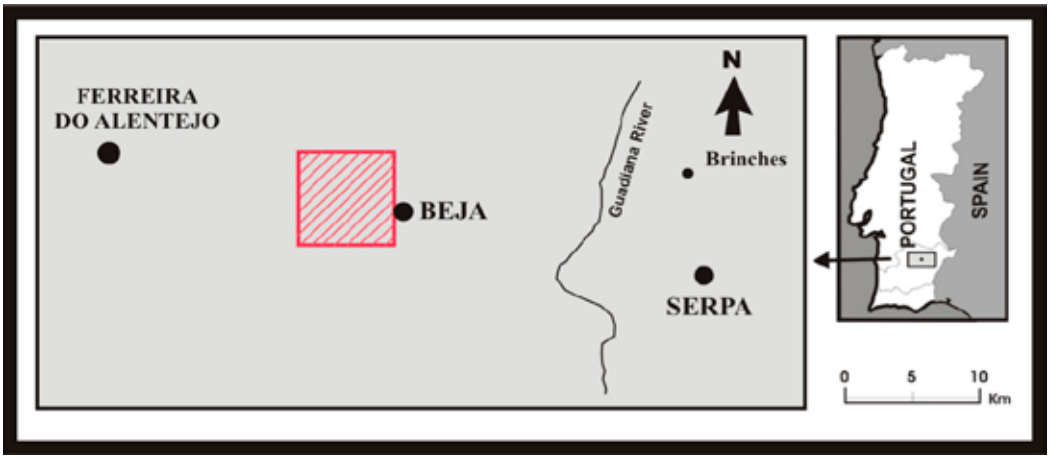


Figura 1. Localização do Aquífero Sistema de Gabros, em Beja.
Área de estudo (marcada) com cerca de 50km2.

A Tabela seguinte apresenta um sumário das principais estatísticas descritivas obtidas para os dados disponíveis, tendo sido identificados três *outliers* em ambos os anos (que foram mantidos no conjunto de dados).

	Dados originais		Dados logaritmizados	
	1998	2000	1998	2000
N.de localizações	50	69	50	69
Média	72.74	86.43	4.15	4.36
Mediana	66.50	86.00	4.20	4.45
Desvio padrão	37.00	31.86	0.56	0.53
Minímo	14	10	2.64	2.30
Máximo	190	162	5.25	5.09

Note-se que a média é mais elevada em 2000 do que em 1998, o que poderá eventualmente ser justificado por um aumento dos níveis de poluição na zona de estudo entre as duas campanhas. De acordo com a regulamentação Portuguesa e da União Europeia, sobre o controlo de poluição e classificação de água potável, o valor máximo admissível de concentração de nitrato é 50mgNO3/L. Pretende-se, então, determinar a probabilidade da concentração deste poluente não exceder este valor crítico, numa localização qualquer não amostrada, permitindo a construção de mapas de risco de poluição da referida região. A variável indicatriz $I(s,x)$ irá considerar x igual a 50mgNO3/L e $Z(s)$ será definido como o logaritmo da concentração de nitrato na localização s.

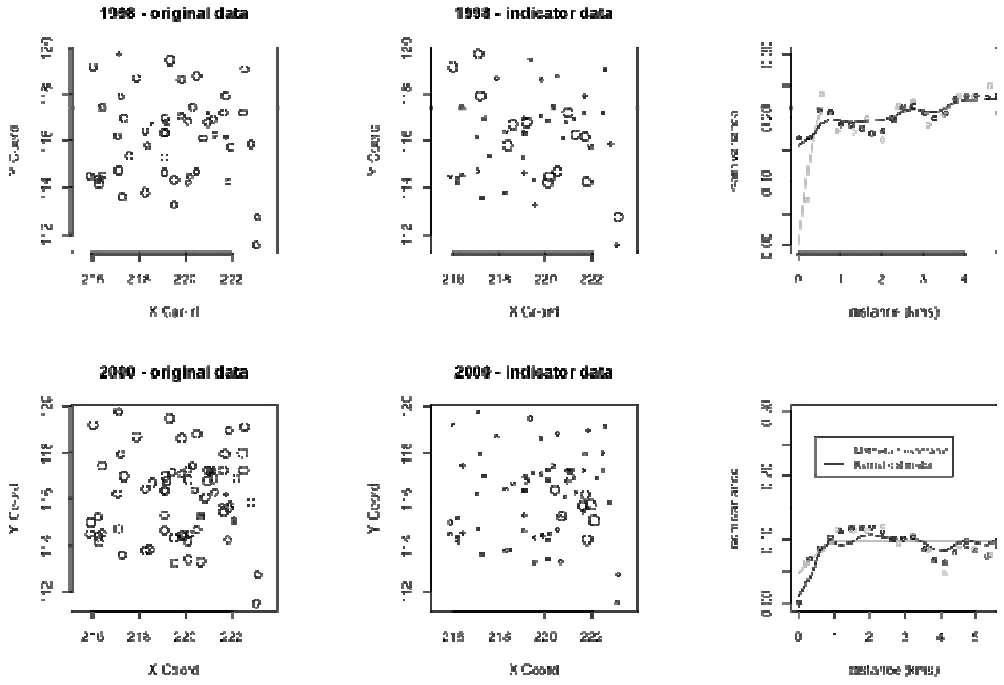


Figura 2. Representação espacial dos dados para 1998, 1.fila, e 2000, 2.fila (dados originais e dados da indicatriz na 1.coluna e 2.coluna, respetivamente). Os paneis da 3.coluna apresentam os estimadores não paramétricos dos variogramas da indicatriz e as respetivas versões válidas (linhas a cheio) para as concentrações de NO3 em Beja.

Os paneis da 1.coluna da Figura 2 mostram as localizações das estações de monitorização para cada ano, onde a distância unitária é 1km. O tamanho de cada carater é proporcional ao valor medido para a concentração de nitrato De forma semelhante, os paneis da 2.coluna apresentam os dados da variável da indicatriz, onde os carateres maiores identificam as localizações cujas concentrações são inferiores ao valor crítico (50 mg NO3/L).

As aproximações obtidas para os variogramas da indicatriz, baseadas no estimador Matheron e no estimador núcleo proposto, estão representados nos paneis da 3.coluna. Paralta & Ribeiro (2003) apresentam mapas de kriging para as mesmas variáveis da indicatriz, depois de ajustar um modelo esférico às estimativas de Matheron. No nosso estudo, iremos proceder de forma análoga para o estimador de Matheron. Adicionalmente, iremos obter uma versão válida do estimador núcleo através do método de Shapiro & Botha (1991), evitando problemas de especificação incorreta do modelo paramétrico. Com o estimador núcleo válido, obtivemos estimativas para o *sill* iguais a 0.235 e 0.119, para 1998 e 2000, respetivamente. As estimativas do raio de influência não diferem muito entre as duas campanhas, sendo também semelhantes às apresentadas em Paralta & Ribeiro (2003), sendo cerca de 1.3km para 1998 e 1km para 2000.
Pretende-se, agora, construir mapas de risco de poluição para a referida região. Começa-se por considerar o modelo $Z(s) = \mu(s) + Y(s)$, aproximando-se a superfície $\mu(s)$ por uma interpolação spline (resultados estão apresentados nos paneis da 1.coluna da Figura 3). $Y(.)$ pode-se assumir como sendo um processo aleatório estacionário de média zero, com distribuição espacial denotada por G. Consequentemente, a aproximação $F_s(50)$ pode ser obtida como $\hat{F}_s(50) = \hat{G}(50 - \mu(s))$, conforme anteriormente descrito.

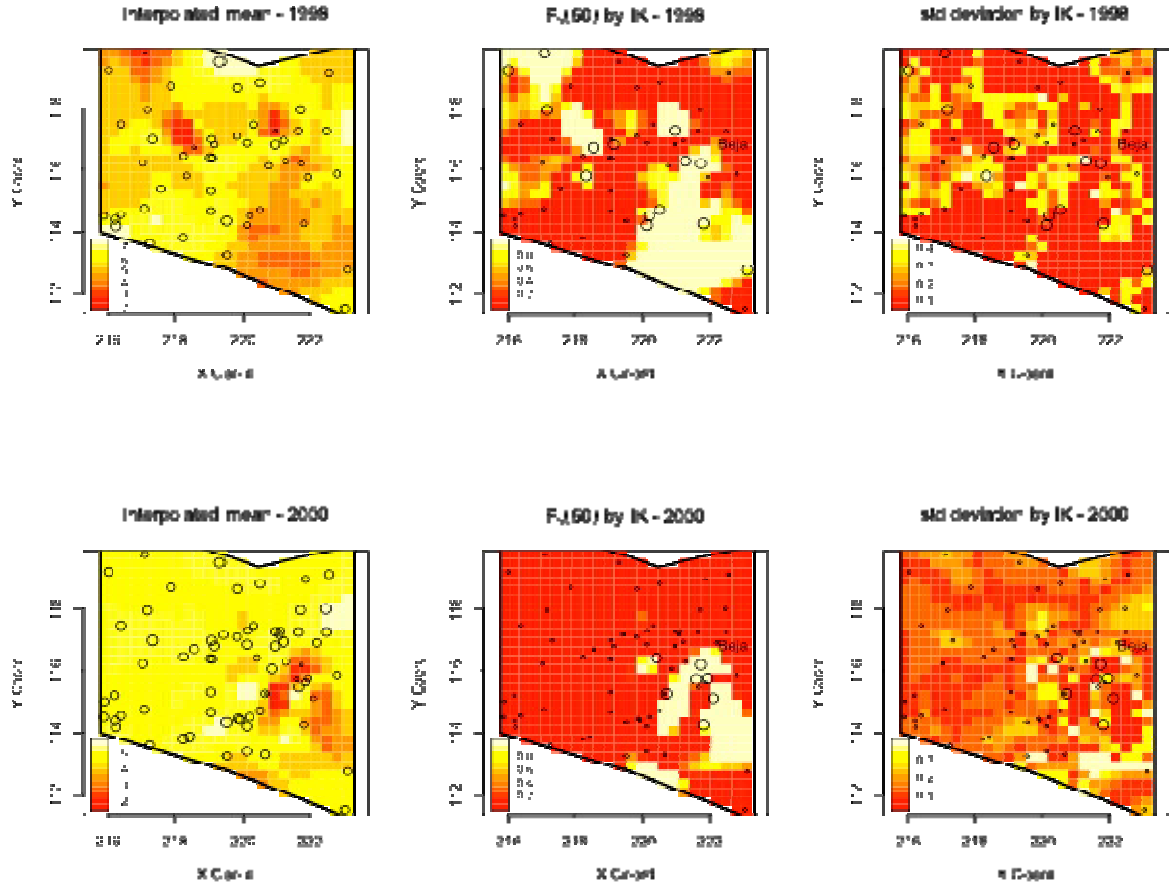


Figura 3. Paneis da 1.coluna apresentam $\hat{\mu}(s)$. Paneis da 2.coluna apresentam os mapas de risco de poluição para as concentrações de NO3 em Beja, mostrando as estimativas de $F_s(50)$, i.e. a probabilidade de não exceder o valor crítico (50mgNO3/L), baseado no método kriging da indicatriz (IK). Paneis da 3.coluna apresentam os respetivos mapas de incerteza das predições.

O nosso objetivo prende-se agora com a aplicação dos métodos apresentados, quer pela krigagem da indicatriz quer pela estimação do *sill*, para aproximar a distribuição $G(.)$. Para cada método, consideram-se as duas alternativas para estimar o variograma da indicatriz, estimador de Matheron ou tipo núcleo.

A Figura 3 apresenta os resultados obtidos para $F_s(50)$, no caso específico da krigagem da indicatriz, considerando-se que s percorre uma grelha de aproximadamente 720 pontos. Repetiu-se a construção destes mapas de risco pelo método da estimação do *sill*, tendo-se obtido resultados semelhantes, com a vantagem de apenas ser necessário resolver a equação de 2ºgrau dada em (1) em vez do sistema usual de equações de kriging. Observando-se os mapas da 2.coluna da Figura 3, conclui-se que 1998 tem associado maiores probabilidades de não exceder o valor de risco de 50mgNO3/L do que 2000, ou seja, os resultados apontam para um aumento dos níveis de poluição entre as duas campanhas.

Deste modo, confirmou-se que os métodos propostos em Garcia-Soidán & Menezes (2012), e aqui brevemente apresentados sem entrar em questões cruciais tais como a seleção adequada do parâmetro janela h, permitem identificar a localização de áreas de elevado risco de poluição, disponibilizando informação importante para a tomada de decisões relacionadas com a saúde pública.

Referências

Garcia-Soidán P. (2007). Asymptotic normality of the Nadaraya-Watson semivariogram estimator. *TEST*, 16 (3), 479-503. DOI: 10.1007/s11749-006-0016-8.

Garcia-Soidán P. & Menezes R. (2012). Estimation of the spatial distribution through the kernel indicator variogram. *Environmetrics*, vol 23 (6), 535-548.

Goovaerts P. (1997). *Geostatistics for natural resources evaluation (1st ed.)*. Oxford University Press: New York.

Journel AG. (1983). Nonparametric estimation of spatial distribution. *Mathematical Geology*, 15 (3), 445-468. DOI: 10.1007/BF01031292.

Matheron G. (1963). Principles of geostatistics. *Economic Geology* 58(8): 1246–1266, DOI: 10.2113/gsecongeo.58.8.1246.

Paralta E. & Ribeiro L. (2003). Monitorização e Modelação Estocástica da Contaminação por Nitratos do Aquífero Gabro-diorítico na Região de Beja. Resultados, Conclusões e Recomendações. *Seminário sobre Águas Subterráneas*, LNEC, Lisboa.

Pardo-Igúzquiza E. (1998). Inference of spatial indicator covariance parameters by maximum likelihood using MLREML. *Computers & Geosciences*, 24 (5): 453-464. DOI: 10.1016/S0098-3004(98)00015-6.

Shapiro A, Botha J. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis*, 11 (1): 87-96. DOI: 10.1016/0167-9473(91)90055-7.

Regularização em suportes discretos

Paulo Eduardo Oliveira, *paulo@mat.uc.pt*

CMUC, Departamento de Matemática, Universidade de Coimbra

1. Introdução, noções e dificuldades

O presente artigo apresenta alguns resultados e métodos para a estimação de distribuições de probabilidade cujo suporte é discreto. A resposta mais óbvia e simples seria a considerarmos frequências relativas, isto é, considerar para cada ponto do suporte o quociente entre o número de observações iguais ao ponto de estimação e o total de observações disponíveis. Este método para construção de aproximações é bem conhecido e, como consequência imediata das leis dos grandes números, fornece estimativas que convergem, nos vários sentidos correntes em Probabilidades e Estatística, para o valor pretendido. É também bem conhecido que este tipo de aproximações não se limitam a ser bons estimadores para o caso de suportes discretos, já que é simples a sua adaptação ao contexto de distribuições contínuas, considerando intervalos no lugar de pontos individuais e, eventualmente, uma normalização adequada. O contexto de distribuições com suporte contínuo acaba por ser mais natural para a manipulação formal e, especialmente, para a apresentação das ideias que estão subjacentes aos métodos do tipo núcleo. De forma a ser possível alguma formalização, introduz-se em seguida alguma notação. Estaremos interessados na distribuição de probabilidade de uma variável aleatória X que suporemos ter função de distribuição $F(x) = \mathbf{P}(X \leq x)$. Podemos estar interessados na estimação de $F(x)$ ou, caso esta exista, na estimação de $f(x) = F'(x)$, a densidade da distribuição. É evidente que esta última possibilidade apenas faz sentido no caso de distribuições absolutamente contínuas. A estimação de F é obtida através dos quocientes referidos acima que podem ser formalizados da seguinte forma: considera-se que dispomos de uma amostra $X_1, \ldots, X_n$ da variável X, um conjunto A de valores possíveis e constrói-se

$$\widehat{F}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(X_i), \tag{1}$$

onde $\mathbb{I}_A(X_i) = 1$ se $X_i \in A$, sendo igual a 0 no caso contrário. O somatório efetua a contagem do número de observações que caíram no conjunto A. Assumindo a habitual hipótese de independência, a lei dos grandes números fornece imediatamente a convergência, com probabilidade 1, para $\mathbf{E}(\mathbb{I}_A(X_i)) = \mathbf{P}(X \in A)$. É também simples invocar resultados do tipo Teorema do Limite Central para encontrarmos caracterizações mais precisas do comportamento de $\widehat{F}_n$ e, em particular, construir intervalos de confiança. Note-se que este resultado vale para todas as escolhas possíveis do conjunto A, pelo que permite uma

grande flexibilidade mesmo em relação às propriedades do suporte da distribuição. A aproximação da densidade f pode ser obtida da forma análoga através de

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{I}_{(x-h/2,x+h/2)}(X_i). \tag{2}$$

Note-se a introdução do parâmetro h, o que corresponde, relativamente, a uma escolha particular do conjunto A em (1) que permite uma invocação do Teorema do Valor Médio que justifica a construção. É este estimador (2) que é habitualmente designado por estimador da janela móvel. Uma observação óbvia: no caso de suportes discretos (2) reduz-se a (1) caso h seja suficientemente pequeno. O estimador (2) tem alguns inconvenientes evidentes. Há claros problemas quanto à continuidade da função $\widehat{f}_n(x)$, o que é sempre uma propriedade desejável. Por outro lado, a utilização da função $\mathbb{I}$ faz com que haja um corte abrupto na informação que é incluída na construção da aproximação $\widehat{f}_n(x)$. De facto, qualquer observação da variável aleatória que fique a uma distância maior do que $\frac{h}{2}$ do ponto x não contribui com qualquer informação. Uma extensão aparece como natural quando se observa que $\frac{1}{h}\mathbb{I}_{(x-h/2,x+h/2)}(\cdot)$ é uma função de densidade de probabilidade. Porque não substituí-la por alguma outra função de densidade de probabilidade? Note-se ainda que esta função densidade utilizada em (2) pode ser vista como uma transformação, em localização e escala, da densidade $\mathbb{I}_{(-1/2,1/2)}(x)$. Surge assim de forma natural o estimador do núcleo para a densidade de probabilidade

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{3}$$

onde K é uma densidade de probabilidade, designada por *função núcleo*, e h um parâmetro que varia com n tendendo para 0, designado por *janela*. Caso pretendamos proceder à estimação em $\mathbb{R}^d$, substitui-se h por h^d na expressão acima. Existe uma extensíssima literatura sobre este tipo de estimadores. Para um estudo das propriedades básicas destes estimadores sugere-se uma consulta a Scott [13], Silverman [14] ou Wand e Jones [17]. O parâmetro crucial para o comportamento deste tipo de estimadores é a escolha de h, já que se mostra que a função K acaba por ter um efeito secundário nas propriedades de (3).

Os métodos de estimação do tipo núcleo foram inicialmente introduzidos noutro contexto por Nadaraja [11] e Watson [18]. Para estes autores o problema era o da aproximação de uma função de esperança condicional $r(x) = \mathbf{E}(Y|X = x)$, sem necessitar de colocar à partida hipóteses sobre a forma de r. É aliás uma constatação simples de verificar, consultando a bibliografia entretanto produzida, que os dois problemas de aproximação referidos, densidade ou regressão, têm uma evolução paralela quer em metodologias quer em resultados. De facto, embora esse aspeto não seja neste texto desenvolvido, estes dois problemas podem ser vistos como casos particulares de um mesmo problema de estimação colocado num contexto mais geral, conforme foi explorado em Jacob e Oliveira [9, 10].

Voltando à expressão (3), esta traduz uma ideia bastante natural. De facto, a intuição subjacente a (3), consiste em, para obter uma aproximação no ponto x, calcular uma média (pesada) das observações numa vizinhança adequada deste ponto. É a escolha do parâmetro h mencionado acima que faz a definição desta *vizinhança adequada*. Esta argumentação é completamente natural quando se trabalha com dados com distribuição contínua. Mais ainda, apesar de se não terem aqui incluído resultados nem as suas demonstrações, uma inspeção da argumentação utilizada na prova das (boas) propriedades dos estimadores do núcleo deixa claro que são fundamentais as características da medida de Lebesgue em $\mathbb{R}$ ou na sua versão mais geral em $\mathbb{R}^d$. Para além desta medida dar um sentido natural à noção de densidade tem algo de realmente característico, que é a sua invariância relativamente a translações. É esta invariância que acaba por ser o argumento determinante em muitas das demonstrações de propriedades dos estimadores da forma (3). São estas as duas dificuldades a ultrapassar quando se pretende extender a utilização de estimadores do tipo núcleo a contextos distintos dos tradicionais espaço Euclideanos. Não é esse o objetivo deste texto, mas são estas as dificuldades técnicas que é necessário contornar na estatística com dados funcionais, que se tornou bastante popular nos últimos anos. No que respeita a

variáveis com distribuição discreta, as que irão ser abordadas no restante deste texto, as dificuldades técnicas a ultrapassar para a utilização desta metodologia dizem respeito ao tratamento das noções de vizinhança e também às transformações por efeito de escala que, naturalmente, nos podem colocar fora do suporte da distribuição.

2. Primeiras extensões

Uma das primeiras utilizações deste tipo de estimadores que incluem regularização das observações em contexto discreto aparece em Aitchison e Aitken [5] associado a um problema de discriminação. Considerando que a variável aleatória X apenas assume valores em $B = \{0,1\}^d$ pretende-se obter uma aproximação para $\mathbf{P}(X = y)$, com $y \in B$, propondo-se a seguinte estimativa:

$$\widehat{f}(y|h) = \frac{1}{n}\sum_{j=1}^{n} W(h,y,X_j), \qquad \text{onde} \qquad W(h,y,X_j) = h^{d-d^*(y,x)}(1-h)^{d^*(y,x)}, \quad h \in \left[\tfrac{1}{2},1\right], \tag{4}$$

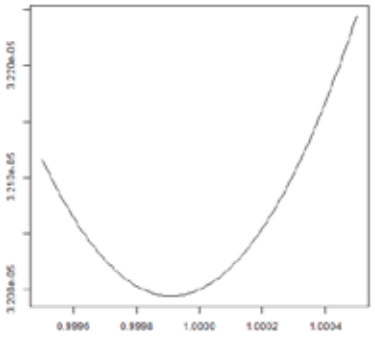
e $d^*(x,y)$ conta as discrepâncias entre as coordenadas de x e de y (recorde-se que se trata de vetores constituídos apenas por 0's e 1's), $d^*(y,x) = (x-y)^t(x-y)$. Refira-se que, apesar de h ter um papel de regularização não corresponde neste estimador à definição de uma vizinhança, tal como na definição geral (3). É simples verificar que a escolha do parâmetro $h = \frac{1}{2}$ conduz à construção de uma aproximação uniforme sobre $B = \{0,1\}^d$, enquanto que a escolha $h = 1$ conduz ao estimador das frequências relativas de cada ponto. A escolha adequada fica assim algures no interior do intervalo $\left[\frac{1}{2},1\right]$. Estes autores utilizam assim um estimador do tipo núcleo num espaço discreto resolvendo, para já, a questão da definição da função núcleo tirando partido da estrutura particular do suporte da distribuição de probabilidade. Para a escolha da janela Aitchison e Aitken [5] mostram que a maximização relativamente a h da verosimilhança $\prod_{i=1}^{n} \widehat{f}_n(x_i|h)$ conduz necessariamente a $h = 1$, isto é, ao estimador das frequências relativas. A alternativa considerada por estes autores maximiza uma pseudo-verosimilhança construída através de um argumento do tipo *jacknifing*: representemos por D a amostra, por $D - x_i$ a amostra com a observação x_i excluída e por $\widehat{f}_n(x|D,h)$ o estimador definido em (4), para tornar evidente qual a amostra que intervém na sua definição; neste caso escolhe-se $h_n = \operatorname{argmax} \prod_{i=1}^{n} \widehat{f}_n(x_i|D - x_i, h)$. Esta maximização, conforme se mostra em [5], conduz a um parâmetro de regularização h_n inferior a 1 mas que verifica $h_n \to 1$. Por um lado isto garante a consistência mas por outro dá a indicação de que para amostras grandes o estimador das frequências relativas está muito próximo de ser o mais adequado.

Uma alteração de notação permite uma abordagem um pouco distinta mas que justifica que alguma regularização dá sempre uma melhoria relativamente às frequências relativas. Para a descrição do resultado necessitamos de alterar a notação. Suponhamos agora que X é uma variável cujo suporte contém Q pontos estando associados a cada um deles as probabilidades $p_q = \mathbf{P}(X = q)$, $q = 1,\ldots,Q$. Designemos ainda por $n_1,\ldots,n_Q$ a contagem do número de observações igual a cada um dos pontos. Então, para cada $\ell = 1,\ldots,Q$ (4) pode-se reescrever na forma, conforme Titterington [16],

$$\widehat{f}_n(\ell) = \frac{1}{n}\sum_{j=1}^{Q} \lambda_{\ell,j} n_j, \qquad \text{com} \qquad \lambda_{\ell,\ell} = h, \quad \lambda_{\ell,j} = \tfrac{1-h}{Q-1}, \quad (\ell \neq j) \tag{5}$$

É ainda possível verificar que esta representação se pode escrever na forma $\widehat{f}_n(\ell) = \frac{1-\alpha}{n}(n_1,\ldots,n_Q) + \alpha\theta$, onde $\theta = (\frac{1}{Q},\ldots,\frac{1}{Q})$ e $\alpha = \frac{1-h}{Q-1}$ ($h = 1$ continua a corresponder ao estimador das frequências). Esta forma de escrever o estimador mostra que $\widehat{f}_n$ se pode interpretar como uma convolução entre a distribuição uniforme no suporte e a distribuição definida pelas frequências relativas. Brown e Rundell [6] analisam a soma dos erros quadráticos $\mathrm{SEQ}(h) = \sum_{q=1}^{Q}(\widehat{f}_n(q) - p_q)^2$ onde, h é o parâmetro

de regularização na definição de $\widehat{f_n}$. Em [6] mostra-se que existe sempre $h < 1$ tal que $\mathrm{SEQ}(h) < \mathrm{SEQ}(1)$. Isto é, no sentido do erro quadrático há sempre vantagem em proceder a alguma regularização, mesmo que a convolução a fazer guarde essencialmente as frequências relativas, isto é, sejamos levados a uma escolha de h próxima de 1, conforme se ilustra na figura ao lado que mostra uma representação gráfica típica da função $\mathrm{SEQ}(h)$. Este comportamento confirma o obtido por Aitchison e Aitken [5].



A partir de representação de SEQ como uma função quadrátrica de h Brown e Rundell [6] definem um estimador para esta quantidade que depois otimizam para propor a escolha

$$\hat{h}_v = \frac{1}{Q}\mathbb{I}_{h_v \leq \frac{1}{Q}} + h_v \mathbb{I}_{\frac{1}{Q} < h_v \leq 1} + \mathbb{I}_{h_v > 1},$$

onde

$$h_v = \frac{\mathbf{N}^t\mathbf{A}\mathbf{N}\left(n + \frac{n-1}{Q-1}\right) - n^2}{\frac{Q(n-1)}{Q-1}\mathbf{N}^t\mathbf{A}\mathbf{N}}, \qquad \mathbf{N} = (n_1, \ldots, n_Q), \quad \mathbf{A} = \begin{bmatrix} 1 & & \frac{-1}{Q-1} \\ & \ddots & \\ \frac{-1}{Q-1} & & 1 \end{bmatrix}.$$

3. Uma abordagem genérica

Antes de partir para outras variantes na estimação em suportes discretos referem-se de seguida alguns resultados que estendem caracterizações dos estimadores de tipo núcleo a situações mais gerais do que as clássicas. Suponhamos que $B \subset \mathbb{R}^d$ e que fixamos uma medida ν que admitimos ser σ-finita em B. Refira-se que o caso clássico dos estimadores do núcleo consiste em tomar $B = \mathbb{R}^d$ e ν a medida de Lebesgue. No que se segue $X_1, \ldots, X_n$ é uma amostra de uma variável aleatória X com valores em B e admitimos que a sua distribuição $\mathbf{P}_X$ é absolutamente contínua em relação a ν e pretendemos estimar a derivada de Radon-Nikodym $f = \frac{d\mathbf{P}_X}{d\nu}$. No caso clássico esta derivada não é mais do que a densidade. No caso de B ser discreto estamos a estimar a função de probabilidade. É neste contexto que é estudada a estimação pelo método do núcleo em Campos e Dorea [7] que definem um estimador da forma

$$p_n(x) = \frac{1}{n}\sum_{j=1}^{n} W(h, x, X_j), \tag{6}$$

onde W é uma função tal que $\int |W(h,x,y)|\, \nu(dy) < \infty$, para $h \in (0, h_0]$ para algum $h_0 > 0$. É imediato verificar que, no caso $B = \mathbb{R}^d$ a escolha $W(h,x,X_k) = \frac{1}{h^d}K(\frac{x - X_k}{h})$ conduz ao estimador do núcleo clássíco (3). No caso $B = \mathbb{N}$, a escolha $W(h,i,j) = \frac{h}{2u}\mathbb{I}_{|j-i|=1,\ldots,u} + (1-h)\mathbb{I}_{j=i}$, onde $u \in \mathbb{N}$ define o suporte desta função, define um núcleo uniforme em pontos vizinhos do ponto i em que se faz a estimação. O controlo desta função W é crucial para obter as boas propriedades do estimador p_n. No caso de B ser discreto e admitindo que W só assume valores não negativos, a condição de integrabilidade segue-se de

$$\int W(h,x,y)\, \nu(dy) = \sum_n W(h,x,x_n) \leq K_0(x). \tag{7}$$

Além da condição sobre a integrabilidade já referida é necessário o controlo dos valores da função W. Para isso considerem-se as seguintes condições:

$$\forall \delta > 0,\ \exists K_\delta,\ \left|W(h,x,y)\mathbb{I}_{\{|y-x|>\delta\}}\right| \leq K_\delta(x) < \infty, \qquad \text{e} \qquad \lim_{h\to 0} W(h,x,y)\mathbb{I}_{\{|y-x|>\delta\}} = 0. \tag{8}$$

A primeira destas duas condições é uma limitação uniforme dos valores de W longe da diagonal principal de B. Quanto à segunda ela é verificada trivialmente em núcleos do tipo dos do exemplo discreto

apresentado acima. Uma das condições correntes no caso clássico é a continuidade da função a estimar, que é substituída por

$$x \in C_\nu \quad \Leftrightarrow \quad \forall \varepsilon > 0, \exists \delta > 0, \nu\{y \in B : ||x-y|| \leq \delta, |f(x)-f(y)| > \varepsilon\} = 0. \tag{9}$$

Os pontos em C_ν dizem-se *pontos de ν-continuidade* de f.

Teorema 1 *Se (8) e (9) se verificam então, em todo o x ponto de ν-continuidade tem-se*

$$\lim_{h\to 0} \left| \int W(h,x,y) f(y)\, \nu(dy) - f(x) \int W(h,x,y)\, \nu(dy) \right| = 0.$$

É agora imediato obter condições que garantem que p_n definido em (6) é assintoticamente centrado.

Corolário 2 *Além de (8) e (9), admita-se que $h_n \longrightarrow 0$ e $\int W(h,x,y)\, \nu(dy) = 1$. Então $\mathbf{E}(p_n(x)) \longrightarrow f(x)$.*

No caso de B ser discreto já vimos que a condição de integrabilidade se traduz em (7) que garante também a verificação da primeira das condições em (8). Vejamos em que se traduz a ν-continuidade no caso de suportes discretos e em que ν é a medida de contagem. São imediatas as caracterizações seguintes.

Proposição 3 *Seja ν uma medida de contagem. Todos os pontos de B que não são de acumulação são pontos de ν-continuidade. Um ponto de acumulação de B é ponto de ν-continuidade se e só se for ponto de continuidade.*

Um exemplo frequente de regularização em suportes discretos consiste na escolha $W^*(h,x,y) = K(||x-y||)\mathbb{I}_{||x-y||<h}$. O estimador baseado nesta função W^* traduz-se no cálculo de uma média pesada dos pontos numa vizinhança de x. Admitindo que h é suficientemente pequeno, quanto à condição de integrabilidade, temos

$$\int K(||y-x||)\mathbb{I}_{||y-x||<h}\, \nu(dy) = \begin{cases} K(0), & \text{se 0 não é ponto de acumulação de } B, \\ K(0) + 2 \displaystyle\sum_{n:||y-x||<h} K(x_n), & \text{se 0 é ponto de acumulação de } B. \end{cases}$$

As condições (8) e (9) deduzem facilmente-se de $\sup K(x) \leq K(0) < \infty$. O Teorema 1 implica agora que $\mathbf{E}(p_n(x)) \longrightarrow K(0)f(x)$.

A convergência quase certa necessita de algum controlo adicional no parâmetro de regularização h. Recorde-se que esta é exatamente a situação no caso clássico em que se supõe $nh_n \longrightarrow +\infty$, para obter o controlo da variância do estimador. Para cada ponto de ν-continuidade de f defina-se $\gamma_n(x) = \nu\{y \in B : ||y-x|| \leq h\}$ e assuma-se que

$$\lim_{n\to+\infty} \gamma_n(x) = \gamma(x) < \infty, \qquad \lim_{n\to+\infty} n\gamma_n(x) = +\infty, \qquad |\gamma(x)W(h,x,y)| \leq K_1(x) < \infty, h \in (0,h_0). \tag{10}$$

Teorema 4 *Admita-se que x é ponto de ν-continuidade tal que (10), $h_n \longrightarrow 0$ e $\int W(h,x,y)\, \nu(dy) = 1$ se verificam. Se $\sum_n \exp(-n\gamma_n(x)) < \infty$ então $p_n(x) \longrightarrow p(x)$ quase certamente exceto eventualmente para valores de x num conjunto de medida ν nula.*

As hipóteses que figuram em (10) fazem a extensão das hipóteses clássicas sobre h. De facto, se ν for a medida de Lebesgue em $\mathbb{R}^d$ é fácil ver que $\gamma_n(x)$ é, a menos da multiplicação por uma constante, h_n^d. Num caso que nos interessa explorar aqui, em que $B \subset \mathbb{R}$ é discreto e ν a medida de contagem, temos $\gamma_n(x) = 2\lfloor h \rfloor + 1$ se x não é ponto de acumulação e $\gamma_n(x) = +\infty$ caso contrário. Assim, neste caso não há possibilidade de cumprir (10), pelo que estes resultados gerais não nos conseguem fornecer a convergência quase certa. Felizmente, é possível obter um resultado para a escolha do núcleo W^* referido acima.

Teorema 5 *O estimador (6) baseado no núcleo W^* verifica $p_n(x) \longrightarrow p(x)$ quase certamente exceto eventualmente para valores de x num conjunto de medida ν nula desde que* $\sup K(x) \leq K(0) < \infty$.

A demonstração deste resultado não aparece em Campos e Dorea [7], mas facilmente se obtém seguindo o plano de demonstração do Teorema 3 em [7]. É ainda possível deduzir, neste caso, que a ordem de convergência é $\left(\frac{\log n}{n}\right)^{1/2}$. Recuperam-se assim resultados similares aos bem conhecidos no caso Euclideano clássico.

A normalidade assintótica pode também ser demonstrada utilizando este contexto genérico para a estimação. Como habitualmente é necessário algum controlo adicional nos momentos para permitir estabilizar as variâncias das variáveis.

Teorema 6 *Seja x um ponto de ν-continuidade tal que $f(x) > 0$. Suponhamos que (10) se verifica e que, além disso,*

$$\liminf_{n\to+\infty} \int \gamma_n(x) W^2(h,x,y)\, \nu(dy) = K_1(x) > 0. \tag{11}$$

Então

$$\frac{p_n(x) - \mathbf{E}p_n(x)}{\mathrm{Var}(p_n(x))} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Este resultado enferma da mesma dificuldade que o Teorema 4 quanto à sua adaptação ao contexto discreto, devido ao significado do parâmetro $\gamma_n(x)$ que torna (10) não cumprível. Refira-se apenas que no caso discreto a condição (11) apenas levanta dificuldades se x for ponto de acumulação de B. Tal como para a convergência quase certa, também aqui é possível demonstrar um resultado de normalidade assintótica particularizando a escolha da função núcleo.

Teorema 7 *O estimador (6) baseado no núcleo W^* verifica o resultado de normalidade assintótica do Teorema 6 se* $\sup_{h,y} W^2(h,x,y) \leq K_2(x) < \infty$, $\sup K(x) \leq K(0) < \infty$ *e*

$$\frac{1}{s_n^2} \sum_{j=1}^{n} \int_{\{|Y_{n,j}| > \varepsilon s_n\}} Y_{n,j}^2\, d\mathbf{P} \longrightarrow 0,$$

onde $Y_{n,j} = W(h,x,X_j) - \mathbf{E}W(h,x,X_j)$ *e* $s_n^2 = \sum_j \mathrm{Var}\Big(W(h,x,X_j)\Big)$.

Acrescente-se apenas que, nas condições do teorema anterior $s_n^2 \sim np(x)(1-p(x))K^2(0)$. Mais uma vez este resultado não se encontra em Campos e Dorea [7], mas facilmente se demonstra por verificação da condição de Lindeberg para a soma de variáveis que define o estimador $p_n(x)$.

A caracterização de escolha ótima para o parâmetro h fica um pouco mais problemática neste contexto genérico para a estimação. Lutamos aqui com as dificuldades referidas acima por deixarmos de poder contar com as propriedades geométricas da medida de Lebesgue. Apesar disso é possível encontrar, em alguns casos particulares, critérios semelhantes aos clássicos. Desta vez será necessário colocar hipóteses adicionais sobre o suporte, a distribuição e utilizar o núcleo W^*. Admitamos que o suporte da distribuição é um subconjunto simétrico dos inteiros e que a função $f(x)$ se obtém por discretização de uma verdadeira função densidade f^*, isto é, que temos as seguinte representação $f(x) = \int_{x-1/2}^{x+1/2} f^*(u)\, du$, para $x \in \mathbb{Z}$. Se, além disso, a função K tal que $W^*(h,x,y) = K(||x-y||)\mathbb{I}_{||x-y||<h}$ for simétrica é válida a seguinte representação

$$\begin{aligned}\sum_x \mathbf{E}(p_n(x) - f(x))^2 \quad \sim \quad & (f^*)'(x)\frac{1}{n}\sum_{u=-[h]}^{[h]} uK^2(u) + p(x)\frac{1}{n}\sum_{u=-[h]}^{[h]} K^2(u) \\ & + p(x)\frac{1}{n}\left(\sum_{u=-[h]}^{[h]} K(u)\right)^2 + (f^*)''(x)\frac{1}{4n}\sum_{u=-[h]}^{[h]} u^2K(u) + o\left(\frac{h}{n} + h^2\right).\end{aligned}$$

Esta representação é similar à representação em ambiente Euclideano do erro quadrático médio com integrais no lugar dos somatórios. A escolha do parâmetro de regularização h pode agora ser feito otimizando esta expressão. Evidentemente, todas as dificuldades que é necessário resolver no contexto Euclideano, e que estiveram na base de muita da investigação produzida a propósito do método do núcleo, se mantêm embora com formulação discreta.

4. Estimação por polinómios locais

Nesta secção iremos considerar que o suporte da distribuição são os pontos $x_i = \frac{i-1/2}{Q}$, $i = 1,\ldots,Q$. Esta restrição permite considerar o suporte contido em $[0,1]$ e pensar que a distribuição se pode obter por integração de uma função conveniente definida em $[0,1]$. Recorde-se que assumimos dispor de uma amostra $X_1,\ldots,X_n$ e que representamos por n_ℓ o número de observações iguais a x_ℓ. Consideremos agora uma função de pesos $w(\cdot)$ definida nos inteiros e defina-se, para cada $\ell = 1,\ldots,Q$,

$$H_\ell = \sum_{j=1}^{Q} \left(\tfrac{n_j}{n} - \beta_\ell\right)^2 w(j-\ell). \tag{12}$$

A minimização relativamente a $\beta_1,\ldots,\beta_Q$ destas funções H_ℓ conduz a um estimador para $f(x_\ell)$. A intuição por detrás desta abordagem consiste em fazer uma média pesada das frequências relativas observadas à volta do ponto de referência x_ℓ. É fácil encontrar uma expressão explícita para o estimador:

$$\widehat{f}(x_\ell) = \tfrac{1}{n}\sum_{k=-u}^{u} \tfrac{n_{\ell-k}}{n} w(k) = \tfrac{1}{n}\sum_{k=\ell-u}^{\ell+u} \tfrac{n_k}{n} w(\ell-k),$$

onde u é o menor valor tal que $w(k) = 0$ sempre que $|k| > u$. Isto é, a minimização de cada H_ℓ conduz-nos ao estimador do núcleo para cada $f(x_\ell)$, que pode ser representado na forma (6) escolhendo $W_1(h, x_\ell, y) = \sum_{k=-h}^{h} \mathbb{I}_{x_{\ell-k}}(y) w(k)$, pelo que se lhe aplicam as propriedades referidas na secção anterior.

Uma das vantagens da construção do estimador do núcleo a partir da minimização de (12) é que é fácil obter uma extensão. A expressão (12) traduz que, em cada intervalo centrado em x_ℓ procuramos o melhor segmento de reta horizontal para aproximar as observações disponíveis. Ora, uma reta horizontal não é mais do que a representação algébrica de um polinómio de grau 0. Então porque não substituir este polinómio de grau 0 por um polinómio de outro grau. Este aumento de grau permitirá uma maior liberdade no ajuste pelo que será de esperar uma melhoria nas aproximações que daí se seguirão. É esta a ideia subjacente aos métodos de estimação por polinómios locais. Para um estudo bem completo das propriedades desta família de estimadores refere-se o leitor para a monografia de Fan e Gijbels [8]. O ponto de vista adotado em [8], tal com na maioria da literatura sobre estimadores por polinómios locais, coloca o problema no contexto Euclideano clássico. Em contexto discreto, estes estimadores foram estudados por Aerts, Augustyns e Janssen [1, 2, 3, 4], obtendo expressões matriciais explícitas e caracterizações para os resultados de consistência e de normalidade assintótica. Formalmente, o problema a resolver passa a ser o da minimização de

$$H_\ell = \sum_{j=1}^{Q} \left(\tfrac{n_j}{n} - \beta_{0,\ell} - \beta_{1,\ell}(x_j - x_\ell) - \beta_{p,\ell}(x_j - x_\ell)^p\right)^2 w(j-\ell). \tag{13}$$

O estimador que se obtém admite uma representação da forma $\widehat{f_n}(x_\ell) = \sum_{j=1}^{Q} s_{\ell,j} \frac{n_j}{n}$, sendo os coeficientes $s_{\ell,j}$ explicitamente descritos à custa de uma manipulação matricial baseada na amostra e na função de pesos que, por uma questão de brevidade no texto, se não descreve aqui. O leitor interessado poderá encontrar estas expressões em Aerts, Augustyns e Janssen [2, 3]. Há uma subtileza nos resultados demonstrados por estes autores: considera-se que o número de pontos Q no suporte da distribuição aumenta à medida que o tamanho da amostra cresce, embora este aumento seja moderado já que se supõe que

n/Q (dever-se-ia escrever agora Q_n no lugar de Q, mas opta-se para manter a notação) tem limite finito estritamente positivo. Esta hipótese pode ser interpretada como a existência de uma capacidade crescente de distinção dos valores observados à medida de que dispomos de mais informação. O contexto de estimação continua a ser discreto, mas tem em pano de fundo um contexto contínuo que permite recuperar a utilização de algumas das propriedades geométricas da medida de Lebesgue. Trata-se, portanto, de um problema e de resultados de índole distinta dos abordados nas secções anteriores.

Admitindo que a distribuição discreta se obtém por discretização de uma densidade que tem derivadas contínuas de segunda ordem e que a função de pesos é simétrica e de quadrado integrável (ou somável, já que estamos em contexto discreto), em Aerts, Augustyns e Janssen [3] mostra-se que

$$\sum_{j=1}^{Q} \mathbf{E}(\widehat{f}_n(x_j) - f(x_j))^2 \sim \frac{h^2}{Q} + \frac{1}{nQ\sqrt{h}} + o\left(\frac{h^2}{Q}\right) + o\left(\frac{1}{nQ\sqrt{h}}\right).$$

Refira-se que estes estimadores e representações podem ser reescritos admitindo que o suporte (discreto) está contido em algum espaço de dimensão superior $[0,1]^d$. Mais uma vez referem-se os leitores interessados para Aerts, Augustyns e Janssen [1, 2, 3, 4] onde se poderão encontrar as expressões completas. Esta representação para o erro quadrático médio permite concluir da consistência, neste sentido, naturalmente, e obter uma caracterização para a escolha do parâmetro h. Decorre da representação acima a escolha ótima para $h \sim n^{-1/(2p+3)}$, a que corresponde o comportamento, também ótimo, do erro quadrático médio da ordem de $n^{-(2p+2)/(2p+3)}Q^{-1}$.

Se considerarmos agora soma $\sum_{j=1}^{Q}(\widehat{f}_n(x_j) - f(x_j))^2$, cujo valor médio foi caracterizado acima, estabelece-se em Aerts, Augustyns e Janssen [4] um resultado de normalidade assintótica.

Teorema 8 *Admita-se que $f(x) = \int_{x-1/2}^{x+1/2} f^*(u)\,du$, f^* tem derivada de ordem $p+1$ contínua, que a função de pesos é simétrica e tem suporte finito, que $nh \longrightarrow +\infty$ e $hQ \longrightarrow +\infty$. Então, para p ímpar,*

$$d(n)\left(\sum_{j=1}^{Q}(\widehat{f}_n(x_j) - f(x_j))^2 - \sum_{j=1}^{Q}\mathbf{E}(\widehat{f}_n(x_j) - f(x_j))^2\right) \xrightarrow{d} Z \sim \mathcal{N}(0,\sigma^2),$$

onde

$$d(n) = \begin{cases} \frac{\sqrt{n}Q}{h^{p+1}}, & \text{se } nh^{2p+3} \to +\infty, \\ n^{\frac{4p+5}{2(2p+3)}}Q, & \text{se } nh^{2p+3} \to \lambda > 0, \\ nQ\sqrt{h}, & \text{se } nh^{2p+3} \to 0. \end{cases}$$

É possível escrever expressões explícitas para a variância limite σ^2, veja-se, mais uma vez, Aerts, Augustyns e Janssen [4].

A ideia, explorada por Aerts, Augustyns e Janssen [1, 2, 3, 4], de trabalhar com suportes discretos mas que vão sendo ajustados à dimensão da amostra pretende colocar-se na posição em que se faz estimação de uma distribuição mas se dispõe de poucas observações relativamente ao tamanho desse suporte. É neste contexto que surge uma nova medida de erros das aproximações: $\sup_\ell \left|\frac{\widehat{f}_n(x_\ell)}{f(x_\ell)} - 1\right|$. Este critério de erro foi introduzido por Simonoff [15] é denominado *distância esparsa*. É simples demonstrar que o estimador da frequências relativas não é consistente no sentido da distância esparsa se $n/Q \longrightarrow \delta \in (0,+\infty)$, conforme se ilustra pelo exemplo seguinte (veja-se Santner e Duffy [12], p. 60): considere-se a distribuição $f(\ell) = Q^{-1}$, para $\ell = 1,\ldots,Q$ com $Q = n$, o tamanho da amostra; então, para $\varepsilon \in (0,1)$,

$$\mathbf{P}\left(\sup_{1\le\ell\le Q}\left|\frac{\frac{n_\ell}{n}}{f(\ell)} - 1\right| < \varepsilon\right) = \mathbf{P}(n_\ell = 1, \ell = 1,\ldots,Q) = \frac{n!}{n^n} \longrightarrow 0,$$

relembrando a fórmula de Stirling. É evidente então que há necessidade de considerar alguma regularização das frequências relativas para poder obter consistência no sentido da distância esparsa. Recorde-se que, relativamente ao habitual erro quadrático o mesmo havia sido demonstrado por Brown e Rundell [6]. Em Aerts, Augustyns e Janssen [2] podemos encontrar o seguinte resultado.

Teorema 9 *Admita-se que* $f(x) = \int_{x-1/2}^{x+1/2} f^*(u)\,du$, f^* *tem derivada de ordem* $p+1$ *uniformemente contínua, que a função de pesos é simétrica, tem suporte finito e média nula, que* $nh \longrightarrow +\infty$. *Seja* $m_n = \min_{1\leq \ell \leq Q} f(\ell)$ *e suponhamos ainda que*

$$A_n^2 = \left(\frac{\log n + \log Q}{n}\right)\left(\frac{1}{m_n h Q} + \frac{1}{m_n^2 Q^2}\right) \longrightarrow 0,$$

$$B_n = \frac{h^{p+1}}{m_n Q} \longrightarrow 0.$$

Então

$$\sup_{\ell} \left| \frac{\widehat{f}_n(x_\ell)}{f(x_\ell)} - 1 \right| = O\left(\max(A_n, B_n)\right) \quad q.c..$$

A demonstração desta velocidade de convergência depende de resultados sobre processos empíricos e da utilização adequada de desigualdades exponenciais. Velocidades explícitas podem-se obter particularizando alguns dos parâmetros. Por exemplo, admitindo que $m_n \sim Q^{-1}$, $Q = n^\beta$ e $h = \left(\frac{\log n}{n}\right)^{1/(2p+3)}$, com $\beta > \frac{1}{2p+3}$, deduz-se uma velocidade da distância esparsa da ordem de $\left(\frac{\log n}{n}\right)^{(p+1)/(2p+3)}$.

Bibliografia

[1] Aerts, M., Augustyns, I., Janssen, P., Smoothing sparse multinomial data using loclal polynomial fitting, J. Nonparamtr. Statist. 8 (1997), 127–147.

[2] Aerts, M., Augustyns, I., Janssen, P., Sparse consistency and smoothing for multinomial data, Statist. Probab. Lett. 33 (1997), 41–48.

[3] Aerts, M., Augustyns, I., Janssen, P., Local polynomial estimation of contingency table cell probabilities, Statistics 30 (1997), 127–148.

[4] Aerts, M., Augustyns, I., Janssen, P., Central limit theorem for the total squared error of local polynomial estimators of cell probabilities, J. Statist. Plann. Inference 91 (2000), 181–193.

[5] Aitchison, J., Aitken, C.G., Multivariate Binary Discrimination by the Kernel Method, Biometrika 63 (1976), 413–420.

[6] Brown, P., Rundell, P., Kernel Estimates for Categorical, Technometrics 27 (1985), 293–299.

[7] Campos, V., Dorea, C., Kernel density estimation: the general case, Statist. Probab. Lett. 55 (2001), 173–180.

[8] Fan, J., Gijbels, I., Local polynomial modelling and its applications, Chapman & Hall, 1996.

[9] Jacob, P., Oliveira, P.E., A general approach to non-parametric histogram estimation, Statistics 27 (1995), 73–92.

[10] Jacob, P., Oliveira, P.E., Kernel estimators of general Radon-Nikodym derivatives, Statistics 30 (1997), 25–46

[11] Nadaraja, E.A., On estimating regression, Th. Probab. Appl. 9 (1964), 157–159.

[12] Santner, T.J., Duffy, D.E., The Statistical Analysis of Discrete Data, Springer, 1989.

[13] Scott, D., Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, 1992.

[14] Silverman, B., Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986.

[15] Simonoff, J., A penalty function approach to smoothing large sparse contingency tables, Ann. Statist. 11 (1983), 208–218.

[16] Titterington, D.M., A Comparative Study of Kernel-Based Density Estimates for Categorical Data, Technometrics 22 (1980), 259–268.

[17] Wand, M.P., Jones, M.N., Kernel smoothing, Chapman & Hall, 1995.

[18] Watson, G.S., Smooth regression analysis, Sankhya Ser. A 26 (1964), 359–372.

A Estatística Não-Paramétrica ao Encontro da Genética

Carina Silva-Fortes[1,3], *carina.silva@estesl.ipl.pt*
Maria Antónia Amaral Turkman[2,3], *antonia.turkman@fc.ul.pt*
Lisete Sousa[2,3], *lmsousa@fc.ul.pt*

[1]*Escola Superior de Tecnologia da Saúde de Lisboa do Instituto Politécnico de Lisboa*
[2]*Faculdade de Ciências da Universidade de Lisboa*
[3]*Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)*

1. Introdução

Os sistemas biológicos são constituídos por várias, e muitas vezes desconhecidas, interações e circuitos regulatórios e, consequentemente, a forma funcional da relação entre a mensuração e a atividade é complexa de modelar. No entanto, muitos dos métodos estatísticos aplicados nesta área são baseados em pressupostos como por exemplo de independência e aditividade. O desejo que a biologia seja linear, independente e aditiva, raramente é atendido e nem sempre a aplicação do Teorema Limite Central poderá vir em auxílio, particularmente na presença de amostras pequenas.

Desde 1995 a tecnologia denominada de *microarrays*[1] tornou possível monitorizar em simultâneo milhares de genes, possibilitando aos investigadores perceberem as suas interações (Brewster *et al.*, 2004).

A análise de dados provenientes de *microarrays* representa um grande desafio para a estatística. A complexidade advém não só da elevada quantidade de dados que é gerada, mas também pela necessidade da interação de diferentes áreas, como a biologia, a estatística e a informática. A análise estatística é efetuada num âmbito muito aplicado, pois pretende-se resolver problemas muito concretos. Contudo, há que salientar o facto de muitas vezes as conclusões não terem um carácter definitivo, ou seja, constituem um primeiro passo que leva posteriormente os bioquímicos, ou biólogos, a prosseguir com a experiência de forma mais eficaz com menor custo e tempo. Outro facto a apontar prende-se com a extrema dificuldade na modelação de dados biológicos, sendo frequentemente necessário admitir pressupostos que nem sempre se verificam.

Métodos não-paramétricos são em geral os mais adequados para descrever os sistemas biológicos, uma vez que os pressupostos necessários são mais flexíveis. No entanto, os métodos não-paramétricos são usualmente utilizados como segunda opção, apesar da perda de eficiência ser contrabalançada com a redução do risco em interpretar resultados baseados em especificações incorretas.

Um dos objetivos da tecnologia de *microarrays* é medir a expressão de milhares de genes[2] e identificar mudanças nas expressões entre diferentes estados biológicos. Na análise da expressão genética, a maior parte dos estudos tem como objetivo identificar genes com regulação positiva (os níveis de expressão são superiores na amostra experimental) e/ou com regulação negativa (os níveis de

[1] Suporte sólido onde são depositadas gotas individuais de material genético dispostas de forma matricial que permite a sua análise em simultâneo, com o objetivo de alcançar um maior rendimento e velocidade.
[2] Todas as células contêm o mesmo DNA, no entanto diferentes células sintetizam diferentes proteínas. A concentração de mRNA define o "estado biológico" de cada célula e a expressão diferencial dos genes é o reflexo dessa concentração.

expressão são superiores na amostra controlo), estes genes designam-se de genes diferencialmente expressos (DE).

Sabe-se que as amostras biológicas são heterogéneas, por exemplo, devido à presença de subtipos moleculares. Por exemplo, em estudos que envolvam a classificação de tumores é importante verificar se existem diferentes subtipos de cancro. Distribuições bimodais ou multimodais geralmente refletem a presença de misturas de subclasses. Consequentemente pode haver genes que sendo diferencialmente expressos quando se tem em conta a presença de subclasses, não são identificados pelos métodos usualmente utilizados para selecionar genes DE.

Propôs-se uma nova ferramenta, *Arrowplot* (Silva-Fortes *et al.*, 2012), de modo que a partir de uma análise gráfica seja possível identificar genes DE e os genes DE que revelam diferentes subclasses e que não são identificados pelos métodos tradicionais. Este gráfico, *Arrowplot*, baseia-se em duas medidas, nomeadamente na área abaixo da curva (AUC) *receiveroperatingcharacteristic* (ROC) e no coeficiente de sobreposição entre duas densidades (OVL). Dados provenientes de experiências que envolvem *microarrays* têm determinadas características como: um reduzido número de réplicas, valores omissos, heterogeneidade das variâncias, presença de distribuições bimodais e distribuições enviesadas; o que nos levou a considerar uma abordagem não-paramétrica, uma vez que os métodos são mais robustos por exemplo à presença de *outliers* e a transformações de escala das variáveis. Assim, para a estimação da AUC, considerou-se o método de estimação do núcleo e para o OVL desenvolveu-se um algoritmo com base em funções de densidade de probabilidade estimadas pelo método do núcleo.

2. Estimador do núcleo

Desde 1890 diferentes métodos de estimação de funções densidade de probabilidade (f.d.p.) têm sido propostos. A partir de 1956 os métodos de estimação de f.d.p. não-paramétricos têm-se consolidado como uma alternativa sofisticada no tratamento tradicional de conjuntos de dados. Esta alternativa baseia-se na possibilidade de analisar dados sem se admitir um comportamento distribucional específico.

O estimador mais simples de uma f.d.p. é o histograma, no entanto apresenta algumas limitações (Wegman, 1975;Parzen, 1962; Grenander, 1981). A partir dos trabalhos de Rosenblatt (1956) e Parzen (1962), o estimador do núcleo tem sido bastante estudado, veja-se, por exemplo, os trabalhos de Silverman (1986) e de Wand e Jones (1995). As boas propriedades contribuíram para a ampla utilização deste estimador.

Existem vários tipos de estimadores do núcleo. Fix e Hodges (1951) foram os primeiros autores na literatura que propuseram as ideias básicas do estimador do núcleo, utilizando uma função núcleo *Uniforme* (-1,1). Rosenblatt (1956) e Parzen (1962) estudaram a classe geral do estimador do núcleo univariado, conhecido na literatura como estimador do núcleo fixo ou estimador de núcleo global. A terminologia *fixo* foi adotada devido ao facto da janela h ser constante para o suporte da variável. Para uma descrição mais completa sobre o estimador do núcleo e as suas propriedades, *vide* por exemplo Silverman (1986), Scott (1992) eWand e Jones (1995).

Dada uma amostra aleatória $X_1,\ldots,X_n$ de uma distribuição univariada contínua, com f.d.p. f desconhecida, um estimador do núcleo é dado por

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \forall\, x \in S, h > 0, \qquad (1)$$

onde K é a função núcleo, ha janela e So suporte (Rosenblatt, 1956).

Várias são as funções que podem servir de núcleo, desde que sejam simétricas, unimodais e invariantes a transformações. Um aspeto-chave do estimador do núcleo está associado ao facto do seu desempenho, em termos do erro, não depender diretamente da forma funcional do núcleo. Diversos trabalhos demonstram que a qualidade do estimador do núcleo depende essencialmente da escolha da janela *h* (Silverman, 1986). Neste trabalho a função núcleo está restrita ao núcleo gaussiano. Ao longo dos anos, vários estudos foram feitos em busca de métodos que estimassem de uma forma "automática" a janela ótima. Estes métodos automáticos de seleção da janela são baseados na ideia que a quantidade ótima de suavização deve depender unicamente dos dados.

Uma vez que em experiências de *microarrays* estimam-se milhares de pares de f.d.p. (um par para cada gene), a escolha da janela ótima pretende-se automática e simples. A função `density` do `R` usa, por omissão, a função núcleo gaussiana e a escolha de *h* é feita de acordo com

$$h = \left(\frac{4}{3n}\right)^{1/5} min\left(s, \frac{R}{1{,}34}\right), \qquad (2)$$

onde *R* é a amplitude interquartil empírica e *s* o desvio padrão empírico. Existem outros métodos mais complexos para estimar *h,* por exemplo, o método de validação cruzada por mínimos quadrados (Rudemo, 1982; Bowman, 1984).

3. Coeficiente de sobreposição — OVL

O OVL é definido como a área comum entre duas funções de densidade de probabilidade (f.d.p.) (1) e é utilizado como medida de concordância entre duas distribuições (Weitzman, 1970; Inman e Bradley, 1989).

A expressão do OVL pode ser representada segundo Weitzman (1970) por

$$\mathrm{OVL}(X,Y) = \int_{-\infty}^{+\infty} min[f_X(c), f_Y(c)]dc, \qquad (3)$$

onde f_X e f_Y são as f.d.p. das variáveis aleatórias *X* e *Y*, respetivamente. No caso discreto o integral é substituído pelo somatório. OVL(*X,Y*)=1 se e só se as distribuições de *X* e *Y* forem iguais e OVL(*X,Y*)=0 se e só se não tiverem nenhum ponto interior em comum.

Por outro lado, a partir da expressão (4) é possível obter-se uma representação alternativa para o OVL (5) tomando $u=f_X(c)$ e $v=f_Y(c)$, ou seja,

$$min(u,v) = \frac{1}{2}(u+v) - \frac{1}{2}|u+v|, \qquad (4)$$

$$\mathrm{OVL}(X,Y) = 1 - \frac{1}{2}\int_{-\infty}^{+\infty} |f_X(c) - f_Y(c)|dc. \qquad (5)$$

O OVL é invariante em relação a transformações de escala, estritamente crescentes e diferenciáveis, das variáveis *X* e *Y*. Estimadores que partilhem esta propriedade de invariância não dependem das observações diretamente, mas das suas ordens. Esta propriedade é muito útil, uma vez que na análise

de dados de *microarrays* é muito comum proceder-se a transformações das variáveis X e Y para remover enviesamentos, como por exemplo o logaritmo de base 2.

Sejam $X_1,\ldots,X_n$ e $Y_1,\ldots,Y_m$ duas amostras aleatórias e independentes e sejam $\hat{f}_X$ e $\hat{f}_Y$ os estimadores do núcleo de f_X e f_Y, respetivamente, obtidos a partir de (1). A partir da expressão (5) um estimador do OVL pelo método do núcleo será dado por

$$\widehat{\mathrm{OVL}} = 1 - \frac{1}{2}\int_{-\infty}^{+\infty} \left|\hat{f}_X(t) - \hat{f}_Y(t)\right| dt. \qquad (6)$$

3.1 Algoritmo

Para a estimação do OVL a partir de duas densidades estimadas pelo método do núcleo, propôs-se um algoritmo que fosse eficiente em bases de dados provenientes de *microarrays*(Silva-Fortes *et al.,* 2012).

Ao estimar-se uma f.d.p. pelo método do núcleo, na prática o que se obtém são os pontos onde a densidade é estimada, uma vez que de acordo com (1): para cada ponto da amostra ajusta-se uma distribuição gaussiana centrada nesse ponto e com variância h, e cada valor da função de densidade de probabilidade estimada, $\hat{f}$, resulta da soma dos valores das densidades gaussianas para esse valor.

O algoritmo proposto para estimar o OVL baseia-se essencialmente na determinação dos pontos das duas densidades estimadas pelo método do núcleo que delimitam a região de interseção (pontos na Figura 1) das duas densidades. Os pontos onde as duas densidades se intersetam designam-se de *pontos de salto* (cruzes na Figura 1). Quando não existe um ponto que pertença simultaneamente às duas densidades, este é estimado por interpolação linear. Os pontos das densidades que delimitam a zona de interseção e os pontos de salto são combinados numa única lista e ordenados por ordem crescente das abcissas. Finalmente aplica-se a regra do trapézio considerando a grelha de pontos da lista final.

O pseudo-código que implementa este algoritmo pode ser consultado em Silva-Fortes *et al.* (2012). A implementação deste algoritmo em linguagem `R` numa base de dados com 10.000 genes demora aproximadamente 60 minutos num Pentium 533 MHz.

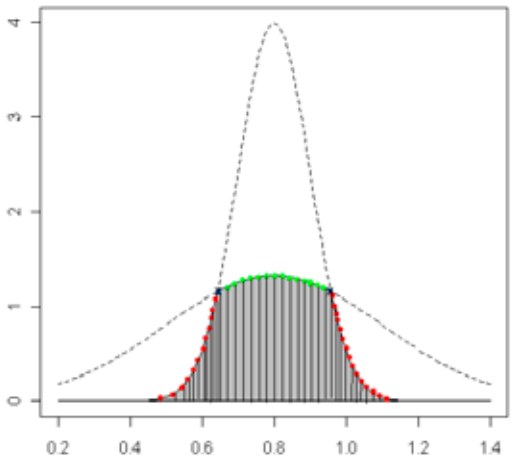


Figura 1: Representação gráfica das densidades estimadas pelo método do núcleo: os pontos delimitam a região de interseção das duas densidades e as cruzes correspondem à interseção das mesmas(correspondem a pontos de salto entre densidades) (Silva-Fortes *et al.*, 2012).

4. Área abaixo da curva ROC – AUC

Vários autores discutiram o refinamento da abordagem não-paramétrica de modo a originar curvas ROC suaves (Zou*et al.,* 1997). Um importante método não-paramétrico de estimação da AUC, no entanto pouco utilizado (talvez pela escassez de *software* que o implemente) é o estimador pelo método do núcleo da AUC. Lloyd (1997) demonstrou que, considerando um núcleo gaussiano, o estimador pelo método do núcleo da AUC é dado por

$$\widehat{\mathrm{AUC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Phi\left(\frac{Y_j - X_i}{\sqrt{h_0^2 + h_1^2}} \right), \qquad (7)$$

onden_0 e n_1 são as dimensões das amostra sem cada condição experimental, h_0 e h_1 são as janelas estimadas de acordo com (2) e $\Phi(.)$ representa a função de distribuição de uma variável aleatória com distribuição gaussiana padrão.

O método não-paramétrico mais comumente utilizado para estimar a AUC, é o método empírico, que corresponde à estatística de Mann-Whitney (McNeil e Hanley, 1984). Na prática coincide com o valor da AUC estimado pela regra do trapézio (Bamber, 1975). No entanto, as curvas ROC estimadas pelo método empírico são irregulares, fazendo com que as estimativas sejam mais otimistas.

As curvas ROC são construídas com base na sensibilidade e especificidade de um sistemas que classifique os dados em duas classes mutuamente exclusivas para todos os pontos de corte possíveis da variável de decisão. Estas probabilidades dependem da regra de classificação, sendo que na análise ROC tradicional um valor elevado da variável de decisão corresponde à presença do artefacto de interesse, *i.e.*, a variável de que representa os níveis de expressão dos genes na população controlo é estocasticamente inferior à variável que representa os níveis de expressão na população experimental. No entanto se se aplicar as curvas ROC para todos os genes de uma experiência de *microarrays* considerando a mesma regra de classificação, estas podem apresentar-se abaixo da diagonal positiva do plano unitário. Esta situação ocorre porque a regra de classificação não será a mesma para todos os genes numa experiência de *microarrays*, uma vez alguns terão regulação positiva e outros regulação negativa. A AUC varia entre 0,5 e 1 numa situação em que a relação estocástica entre as variáveis é coerente com a regra de classificação, no entanto neste trabalho por uma questão de conveniência, consideraram-se curvas ROC degeneradas, *i.e.,* curvas ROC que se apresentam abaixo ou cruzam a diagonal principal do plano unitário. Consequentemente a AUC pode variar entre 0 e 1.

5. *Arrowplot*

Neste trabalho, foram considerados apenas *arrays* de um canal (Silva-Fortes *et al*., 2012) e considera-se que se pretende comparar duas condições experimentais (*e.g.* controlo *vs.* experimental). O *Arrowplot* resulta da representação gráfica num plano unitário das estimativas da AUC e do OVL de todos os genes de uma experiência de *microarrays* (Figura 2). A partir deste gráfico é possível selecionar genes com regulação positiva, com regulação negativa e genes que revelam a presença de misturas de subclasses (genes mistos). Espera-se que estes genes apresentem valores próximos de zero para o OVL e em relação à AUC, os genes com regulação positiva terão valores próximos de um, genes com regulação negativa terão valores próximos de zero e genes mistos terão valores da AUC à roda de 0,5.

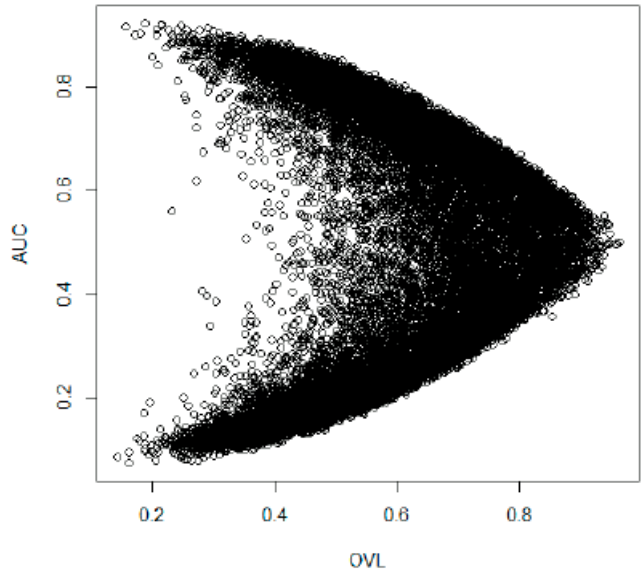


Figura 2:*Arrowplot*

A análise do gráfico é bastante intuitiva, pois permite-nos obter uma fotografia global do comportamento dos genes e com base na sua análise o utilizador pode escolher os pontos de corte para a AUC e o OVL, apesar desta escolha ser arbitrária.

4. Conclusões e trabalho futuro

Procedeu-se a um estudo de simulação para analisar o erro padrão e viés do OVL obtido a partir do algoritmo proposto, tendo sido aplicados métodos de Monte Carlo e *bootstrap*. Neste estudo apenas se simulou de distribuições gaussianas. Compararam-se os resultados obtidos com os trabalhos de Clemons e Bradley (2000) e de Schmid e Schmidt (2006), que também propõem estimadores não-paramétricos para o OVL, e a partir do algoritmo aqui proposto obtiveram-se resultados com um viés mais reduzido quando as distribuições são mais próximas.

Este algoritmo não impõe restrições no número de interseções das densidades e para determinar a curva de interseção entre as duas funções de densidade de probabilidade estimadas pelo método do núcleo, numa base ponto-por-ponto, não obriga que as estimativas das densidades pelo método do núcleo tenham as mesmas abcissas nos dois grupos em análise.

Como trabalho futuro pretende-se realizar um estudo do viés e do erro padrão mais exaustivo, considerando nos estudos de simulação não só outras distribuições unimodais (que se traduz em apenas dois pontos de salto entre as funções) mas também misturas de distribuições. Para além disto, pretende-se investigar qual a variação na eficiência do algoritmo quando se consideram outros métodos de estimação para a janela *h*.

Uma vez que dados de *microarrays* são constituídos por milhares de genes, o que se pretende é que se tenham disponíveis ferramentas que sejam eficientes e com o melhor desempenho em termos de tempo. A verificação de pressupostos é morosa e pelas características que os dados de *microarrays* apresentam, a abordagem não-paramétrica revelou-se a mais indicada para a construção do *Arrow plot*.

Agradecimentos

Este trabalho foi parcialmente financiado por fundos nacionais através da FCT- Fundação Nacional para a Ciência e Tecnologia, no âmbito dos projetos PEst-OE/MAT_UI0006/2011 e PTDC/MAT/118335/2010 e da bolsa de doutoramento SFRH/BD/45938/2008.

Referências

Bamber, D. C. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph. *Journal of Mathematical Psychology*, 12:387–415.

Bowman, A. (1984).An alternative method of cross-validation for the smoothing of densityestimates. *Biometrika*, 71:353–360.

Brewster, J. L., Beason, K. B., Eckdahl, T. T. e Evans, I. M. (2004). The microarray revolution.*The International Union of Biochesmistry and Molecular Biology*, 32(4):217–227.

Clemons, T. E. e BradleyJr., L. (2000). A nonparametric measure of the overlapping coefficient.*Computational Statistics and Data Analysis*, 31(1):51–61.

Durbin, R., Eddy, S. R., Krogh, A., e Mitchison, G. (1998).*Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press.

Ewens, W. J. e Grant, G. R. (2001).*Statistical Methods in Bioinfomatics: An Introduction.* John Wiley & Sons Ltd., England.

Grenander, U. (1981). *Abstract Inference*. John Wiley & Sons, New York.

Inman, H. F. e Bradley, E. L. (1989).The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics – Theory Methods*, 18(10):3851–3874.

Lloyd, C. J. (1997). The use of smoothed ROC curves to summarize and compare diagnostic systems. *Journal of American Statistical Association,* 93:1356–1364.

McNeil, B. J. e Hanley, J. A. (1984).Statistical approaches to the analysis of ROC curves. *Medical Decision Making*, 4(2):136–149.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. *Annals of Mathematical Statistics*, 27:832–837.

Rudemo, M. (1982).Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics*, 9: 65–78.

Schmid, F. e Schmidt, A. (2006).Nonparametric estimation of the coefficient of overlapping – theory and empirical application. *Computational Statistics and Data Analysis*, 50:1583–1596.

Silva-Fortes, C.,Amaral Turkman, M. A. e Sousa, L. (2012).Arrow Plot: a newgraphical tool for selecting up and down-regulated genes and genes differentialexpressed on subsamples. *BMC Bioinformatics*, 13:147.

Silva-Fortes, C. (2012). Aplicação da Metodologia ROC na Análise de Dados de *Microarrays.*Tese de Doutoramento, DEIO-FCUL.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*.Chapman & Hall: London.

Scott, D. (1992).*Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley:New York.

Wand, M.P. e Jones, M.C. (1995). *Kernel smoothing*. Chapman & Hall.

Weitzman, M. S. (1970). *Measure of the overlap of income distribution of white and negro families in United States.* Technical report No. 22, US Department of Commerce, Bureau of the Census, Washington, DC.

Wegman, E. J. (1975). Maximum likelihood estimation of a probability density function. *Sankhyä*, Ser. A, 37:211–224.

Zou, K. H., Hall, W. J. e Shapiro, D. E. (1997). Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16:2143–2156.

Notas breves sobre Análise de Regressão Paramétrica e Semiparamétrica

M. Manuela Neves, *manela@isa.utl.pt*

J. Amaral Santos, *josant.santos@gmail.com*

CEAUL e ISA/UTL

1. Nota Introdutória

Os modelos estatísticos são simplificações da realidade, "grosseiras" muitas vezes, mas úteis, como George Cox deixou expresso na célebre frase "*All models are wrong, some are useful*", quando falava de investigação estatística. Efetivamente a simplificação da realidade deve-se ao facto de ser necessário admitir pressupostos sobre vários aspetos dessa realidade, para modelar e inferir. A abordagem paramétrica clássica considera que a forma da função distribuição cumulativa, *F,* é conhecida pelo menos aproximadamente, exceto num parâmetro (ou vetor de parâmetros de dimensão finita), $\theta(\underline{\boldsymbol{\theta}})$. Uma grande componente da modelação estatística consiste em realizar testes de diagnóstico para assegurar que os pressupostos do modelo são verificados. O objetivo final é determinar uma estimativa desse parâmetro (vetor), $\theta^*(\underline{\boldsymbol{\theta^*}})$, ficando nestas condições feita a descrição completa da população.

O uso de um modelo paramétrico permite fazer inferência estatística e interpretar os resultados da estimação mais facilmente, mas a estimação e a inferência baseadas em pressupostos incorretos pode conduzir a resultados enganadores. Simonoff (1996) aponta que, se o modelo que se admite não é o correto, as inferências "... *can be worse than useless...*" e continua "*Unfortunately the strength of parametric modeling is also its weakness*...". Efetivamente, na inferência baseada num modelo pré-fixado ganha-se muito em eficiência, mas apenas se o modelo adotado é verdadeiro. O ajustamento de um modelo errado conduz a estimadores com viés elevado e menos eficientes.

No "extremo oposto", na modelação não paramétrica, não se admite nenhuma forma funcional para a distribuição. Essa forma é determinada a partir dos dados, quando muito considerando hipóteses muito gerais (unimodalidade, continuidade...) .

O termo <u>não paramétrico</u> não tem o significado de os modelos não terem parâmetros, mas sim que o seu número e natureza são flexíveis e não fixos à partida. Na abordagem não paramétrica, não se

procura a estimativa completa em toda a gama de valores da variável (ao contrário do caso paramétrico), mas obtêm-se estimativas em valores específicos dessa variável.

Graças aos avanços na área da computação verificados nas últimas décadas, os procedimentos não paramétricos têm vindo a ganhar cada vez mais adeptos. Citando algumas obras de referência temos Silverman (1985), Hardle (1991), Simonoff (1996), Gentle (2002), Gentle *et al*. (2004) , Hardle *et al*. (2004) e Takezawa (2006). No entanto também estes procedimentos apresentam desvantagens, sendo uma frequentemente apontada, a de que a precisão do estimador não paramétrico decresce rapidamente quando a dimensão da variável explicativa aumenta.

Quando aos procedimentos não paramétricos se alia o uso de métodos de alisamento (*smoothing methods*) fica estabelecida uma situação de "compromisso" entre:

- a não existência de qualquer hipótese sobre a estrutura formal (abordagem puramente não paramétrica);
- haver hipóteses fortes sobre a forma da distribuição (abordagem paramétrica).

Esta é a abordagem semiparamétrica, que permite uma maior flexibilidade e simplicidade na utilização de procedimentos estatísticos.

2. Análise de Regressão não paramétrica

A análise de regressão procura explicar o comportamento de uma variável (ou vetor) dependente, explicada ou resposta, *Y*, em função de outras variáveis, *X,* ditas independentes, explicativas ou regressoras.

Dados *n* pontos $\{x_i, Y_i\}_{i=1}^{n}$, o modelo de regressão simples admite que a relação entre *X* e *Y* se pode escrever como $Y_i = m(x_i) + \varepsilon_i$, sendo $m(.)$ uma função que representa a dependência sistemática de Y_i em x_i e sendo ε_i o erro aleatório.

A curva de regressão é a média condicional de *Y*, $m(x) = E[Y \mid X = x]$. Não sendo possível especificar a distribuição dos erros e/ou a forma da função $m(.)$, os métodos não paramétricos de ajustamento de curvas permitem o ajustamento de uma curva aos dados, quando pouco ou nada se sabe sobre a forma dessa curva. Aqui o termo não paramétrico significa que não são pré-especificadas nem a distribuição do erro nem a forma funcional da função $m(.)$. O pressuposto inicial de $m(.)$ ser uma função linear é aqui substituído pela hipótese mais fraca de $m(.)$ ser uma função *smooth*. Este "enfraquecimento" tem naturalmente custos: maior esforço computacional (o que hoje em dia já não é problemático) e maior dificuldade de interpretação. Ganha-se porém na obtenção de uma melhor estimativa da função de regressão.

Os alisadores (*smoothers*) permitem não só obter estimativas mais "suaves" da relação entre *X* e *Y* como ainda, nalguns casos, efetuar o diagnóstico da não linearidade e sugerir uma formulação paramétrica simples para o ajustamento (Silverman, 1985). Aplicar um alisador a um conjunto de dados $\{x_i, y_i\}$ consiste basicamente em calcular a estimativa $\hat{y}_i$ para cada x_i do seguinte modo:

- escolher um conjunto de pontos próximos de x_i;
- calcular uma média, ponderada ou não, de valores de *Y* associados a valores de *X* naquele conjunto. O modo de calcular esta média distingue os vários alisadores. Os alisadores mais conhecidos são: o alisador binário; o alisador de média móvel; os alisadores de núcleo; o alisador de regressão polinomial local e os alisadores de *splines*. Aconselhamos a consulta de algumas obras de referência como Simonoff (1996), Fan and Gijbels (1996), Gyorfi *et al*. (2002) , Takezawa (2006), Keele (2008), entre outros.

3. Análise de Regressão Semiparamétrica

Como dissemos atrás, os métodos semiparamétricos oferecem um compromisso: formular pressupostos sobre a forma funcional da curva de regressão mais fortes do que os métodos não paramétricos mas menos restritivos do que os pressupostos da abordagem paramétrica, reduzindo ou eliminando a especificação da distribuição do erro aleatório, ε_i.

O modelo de regressão semiparamétrico é uma ferramenta versátil para modelar quer relações lineares, quer relações não lineares entre variáveis.

Faremos aqui uma breve referência ao uso de um modelo semiparamétrico de regressão de dados de contagem, trabalho que foi desenvolvido por Santos (2005), Santos e Neves (2008a, b) e Santos e Neves (2010).

O modelo semiparamétrico de regressão de dados de contagem considerado baseia-se no seguinte:

- considera-se um modelo paramétrico para os dados;
- para esse modelo, é depois usado o princípio da verosimilhança local, assente no alisador de núcleo polinomial local.

3.1. O Alisador de Núcleo de Máxima Verosimilhança Local (MVL)

Os alisadores com base na verosimilhança local tiveram a sua origem no conceito de verosimilhança penalizada (O'Sullivan *et al*., 1986; Green, 1987). Tibshirani and Hastie (1987) sugeriram o conceito de verosimilhança local, que se baseia na ideia de ponderar a função

verosimilhança atribuindo maior peso às observações mais próximas do ponto de interesse. O estimador resultante é mais robusto e apresenta maior alisamento.

Considerando o modelo de regressão $E[Y \mid X = x] = m(x)$, os estimadores obtêm-se pela maximização da log-verosimilhança local, i.e. a verosimilhança calculada em cada ponto de interesse x e ponderada, ou seja, definida como

$$l[m(x)] = \sum_{i=1}^{n} K\{(x - x_i)/h\} \log f(Y_i; m(x)),$$

com $K(.)$ uma função núcleo -- $K(.)$ diz-se função núcleo se é uma função real, não negativa, contínua, limitada e simétrica tal que $\int K(u)du = 1$ --; h a largura de banda e $f(Y_i; m(x))$ a distribuição do erro aleatório na equação $Y_i = m(x_i) + \varepsilon_i$. A vizinhança local é determinada pela largura de banda h e pela função núcleo $K(.)$. Naturalmente que se coloca aqui o problema da escolha da função núcleo e da escolha da largura de banda. Essa escolha tem sido feita por divisão da amostra em duas subamostras ou por validação cruzada (Gyorfi *et al.,* 2002), portanto consiste no uso de parte da amostra, para obter informação sobre a outra parte.

No âmbito da análise de regressão em dados de contagem, em particular com excesso de zeros, Santos (2005) desenvolveu estimadores MVL para os modelos habituais de dados de contagem, no caso de uma única variável regressora: modelo de regressão de Poisson, modelo de regressão de Poisson inflacionado em zero, modelo de regressão binomial negativo, modelo de regressão binomial negativo inflacionado em zero e modelo de regressão logística.

Vamos exemplificar resumidamente a aplicação do alisador de MVL para o modelo de regressão logística.

3.2. O Alisador de MVL para o Modelo de Regressão Logística

Seja Y uma variável aleatória resposta, binária, com suporte $\{0,1\}$.Para simplificar consideremos uma única covariável contínua X.

No contexto da regressão logística, a média condicional de Y, $E[Y \mid X = x_i] = p_i$, é definida como $E[Y \mid X = x_i] = p_i = \frac{\exp\{m(x_i)\}}{1 + \exp\{(m(x_i)\}}$, onde $m(x_i)$ é uma função desconhecida a estimar com recurso a um alisador polinomial. Considerando o desenvolvimento de Taylor, de grau 1, como uma aproximação a $m(x_i)$, onde x_i está na vizinhança de x, temos $m(x_i) \approx [\beta_0 + \beta_1(x_i - x)]$. Considerando o logaritmo da função verosimilhança local, o estimador MVL obtém-se como solução de

$$\sum_{i=1}^{n}\left\{\left(y_i-\frac{\exp[\beta_0+\beta_1(x_i-x)]}{1+\exp[\beta_0+\beta_1(x_i-x)]}\right)K\{(x_i-x)/h\}\right\}\begin{bmatrix}1\\x_i-x\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}.$$

O viés, variância, intervalos de confiança assintóticos do estimador encontram-se em Santos e Neves (2010), assim como um estudo de simulação, na linha dos procedimentos indicados por Fan *et al*. (1998).

4. Referências Bibliográficas

Fan, J., Farmen, M. and Gijbels, I. (1998).Local maximum likelihood estimation and inference. J. *Roy. Statist. Soc. B*, 60, 591-608.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman & Hall, Londres.

Gentle, J. E. (2002). *Elements of Computational Statistic*s. Springer-Verlag.

Gentle, J. E., Hardle, W. and Mori, Y. (edts) (2004). *Handbook of Computational Statistics, Concepts and Methods.* Springer-Verlag.

Green, P. (1987). Penalized likelihood for general semiparametric regression models. *Int. Statist. Review*, 55, 245-259.

Gyorfi, L., Kohler, M., Kryzak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag.

Hardle, W. (1991). *Smoothing Techniques with implementation in S*. Springer-Verlag.

Hardle, W., Muller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag.

Keele, L. J. (2008). *Semiparametric Regression for the Social Sciences*. John Wiley &Sons.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986). Automatic smoothing regression. *Th. Prob. and its Appl*., 10, 186-190.

Santos, J.A. (2005*). Estimação não paramétrica em modelos de regressão de dados de contagem com excesso de zeros.* Tese de Doutoramento, Universidade Técnica de Lisboa.

Santos, J.A. and Neves, M.M. (2008a). A Local Maximum Likelihood Estimator for logistic regression. In *Proceedings of International Workshop on Statistical Modelling*, Barcelona, 536-539.

Santos, J.A. and Neves, M.M. (2008b). A Local Maximum Likelihood Estimator for Poisson Regression. *Metrika*, 68, 257-270.

Santos, J.A. and Neves, M.M. (2010). A Semiparametric Model of the Logistic Regression. *Actas do Encontro Nacional da SPM*, Leiria, 197-203.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Statist. Soc. B*, 47, 1-52.

Takezawa, K. (2006). *Introduction to Nonparametric Regression*. Wiley Series in Probability and Statistics.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Ass*., 82, 559-567.

Estatística de Extremos Univariados—Modelos Paramétricos vs Não-Paramétricos

Frederico Caeiro, *fac@fct.unl.pt*
Universidade Nova de Lisboa, FCT, DM e CMA

M. Ivette Gomes, *ivette.gomes@fc.ul.pt*
Universidade de Lisboa, DEIO e CEAUL

1 Introdução

A *Estatística de Extremos Univariados* ajuda-nos a estudar acontecimentos potencialmente desastrosos, de enorme relevo para a sociedade e de grande impacto social. Os seus domínios de aplicação são muito variados. Mencionamos as áreas de *Bioestatística*, *Engenharia Estrutural*, *Finanças*, *Hidrologia*, *Meteorologia*, *Seguros* e *Telecomunicações* (veja-se, entre outros, Reiss & Thomas, 2001, 2007; Castillo *et al.*, 2005; Markovich, 2007). Embora seja possível encontrar artigos de interesse histórico relacionados com acontecimentos extremos, o campo remonta a Gumbel, em artigos publicados a partir de 1935, e sumariados em Gumbel (1958). Gumbel desenvolve procedimentos estatísticos essencialmente baseados no teorema de Gnedenko (Gnedenko, 1943), o chamado *teorema de tipos extremais*, um dos resultados limite fundamentais em *Teoria de Valores Extremos* (EVT, do Inglês "*Extreme Value Theory*").

A diferença entre metodologias paramétricas e não-paramétricas é clara, mas dilui-se à medida que restringimos o domínio em que os métodos são válidos. Em *Estatística de Extremos* tem havido nas duas últimas décadas uma mudança do uso de metodologia paramétrica, baseada em resultados probabilísticos assintóticos em EVT, a referir na Secção 2.1, para uma abordagem semi-paramétrica, por muitos considerada não-paramétrica em sentido lato, com a estimação de parâmetros de acontecimentos extremos ou raros a ser feita em contexto muito geral. Depois de uma breve referência na Secção 2.2 ao método clássico de Gumbel e a desenvolvimentos recentes no campo paramétrico, adordamos de forma muito breve na Secção 2.3, o âmago da metodologia não-paramétrica em *Estatística de Extremos*. Na Secção 3 apresentamos alguns dos desenvolvimentos, em campo não-paramétrico, relacionados com métodos de estimação e testes direcionados para o tratamento estatístico de acontecimentos extremos. Finalmente, na Secção 4, discutimos alguns tópicos em aberto, e tecemos breves comentários sobre o tema.

2 Resultados limite na área de extremos e principais abordagens em Estatística de Extremos Univariados

Face a uma amostra aleatória, $(X_1, \ldots, X_n)$, de n variáveis aleatórias independente e identicamente distribuídas ou possivelmente estacionárias e fracamente dependentes com função de distribuição (f.d.), F, usamos a notação $(X_{1:n} \leq \cdots \leq X_{n:n})$ para a amostra de estatísticas ordinais (e.o.) ascendentes associada.

2.1 Resultados limite fundamentais em EVT

Os resultados limite fundamentais em EVT aparecem em Fréchet (1927), Fisher & Tippett (1928), von Mises (1936) e Gnedenko (1943). Qualquer resultado para máximos (e.o. superiores) pode

ser facilmente reformulados para mínimos (e.o. inferiores), face à relação simples, $\min_{1\le i\le n} X_i = -\max_{1\le i\le n}(-X_i)$. Restringir-nos-emos à cauda direita, $\overline{F}(x) := 1 - F(x)$, para x elevado. O *teorema de tipos extremais* de Gnedenko fornece o possível comportamento limite não-degenerado da sucessão de máximos parciais, linearmente normalizada, e uma caracterização incompleta, concluída por de Haan (1970), dos domínios de atração das chamadas leis *max-estáveis*, definidas como leis S para as quais é válida a equação funcional $S^n(\alpha_n x + \beta_n) = S(x)$, $n \ge 1$, para $\alpha_n > 0$, $\beta_n \in \mathbb{R}$. De forma mais específica, todas as possíveis leis limite não-degeneradas da sucessão de máximos parciais $X_{n:n}$, linearmente normalizada, são do tipo da distribuição de *valores extremos* (EV, do Inglês "*extreme value*"), i.e. se existirem sucessões de constantes normalizadoras $a_n > 0$, $b_n \in \mathbb{R}$ e uma f.d. não degenerada G tal que, para todo o x,

$$\lim_{n\to\infty} P\left\{\left(X_{n:n} - b_n\right)/a_n \le x\right\} = G(x), \tag{1}$$

podemos re-definir as constantes a_n e b_n de tal modo que,

$$G(x) \equiv EV_\gamma(x) := \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right),\ 1+\gamma x > 0 & \text{se} \quad \gamma \neq 0 \\ \exp(-\exp(-x)),\ x \in \mathbb{R} & \text{se} \quad \gamma = 0, \end{cases} \tag{2}$$

apresentada na forma de von Mises-Jenkinson (von Mises, 1936; Jenkinson, 1955). Se o limite em (1) existir, dizemos que a f.d. F pertence ao *max-domínio de atração* de EV_γ, em (2), e usamos a notação $F \in \mathcal{D}_{\mathcal{M}}\left(EV_\gamma\right)$. A f.d. $G(\cdot)$, em (1), é então *max-estável*, sendo na realidade as leis em (2) as únicas leis *max-estáveis*. O parâmetro real γ é o chamado *índice de valores extremos* (EVI, do Inglês "*extreme value index*"). O EVI, γ, regula o comportamento da cauda-direita de F. A f.d. EV_γ em (2), aparece frequentemente separada nos três tipos seguintes:

$$\begin{array}{ll} \text{Tipo I } (\textit{Gumbel}): & \Lambda(x) = \exp(-\exp(-x)),\ x \in \mathbb{R}, \\ \text{Tipo II } (\textit{Fréchet}): & \Phi_\alpha(x) = \exp(-x^{-\alpha}),\ x \ge 0 \quad (\alpha > 0), \\ \text{Tipo III } (\textit{Max-Weibull}): & \Psi_\alpha(x) = \exp(-(-x)^{\alpha}),\ x \le 0 \quad (\alpha > 0). \end{array} \tag{3}$$

O domínio de atração *Fréchet* ($\gamma = 1/\alpha > 0$) contém distribuições com cauda-direita pesada, como as distribuições *Pareto* e *t* de *Student*, i.e. cauda-direita com um comportamento de tipo polinomial negativo e um limite superior de suporte infinito. Distribuições de cauda-direita curta, com limite superior de suporte finito, como a *Beta*, pertencem ao domínio de atração *max-Weibull* ($\gamma = -1/\alpha < 0$). O caso intermédio, i.e. o domínio de atração *Gumbel* ($\gamma = 0$), é relevante em muitas ciências aplicadas, e contém uma grande diversidade de distribuições com cauda-direita de tipo exponencial, como a *Normal*, a *Exponencial* e a *Gama*, não obrigatóriamente com limite superior de suporte finito.

Para além do *teorema de tipos extremais* e da distribuição EV_γ, em (2), é também de referir o modelo *generalizado de Pareto* (GP), a possível distribuição limite não-degenerada dos excessos acima de um limiar elevado, devidamente escalados (vejam-se os artigos pioneiros de Balkema & de Haan, 1974, e de Pickands, 1975). O modelo GP tem a forma funcional

$$GP_\gamma(x) = 1 + \ln EV_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma},\ 1+\gamma x > 0,\ x > 0 & \text{se} \quad \gamma \neq 0 \\ 1 - \exp(-x),\ x > 0, & \text{se} \quad \gamma = 0, \end{cases} \tag{4}$$

com EV_γ definida em (2).

Convém ainda referir a *distribuição de extremos multivariada*, relacionada com a distribuição limite das k maiores estatísticas ordinais, $X_{n-i+1:n}, 1 \le i \le k$, muito frequentemente denotada por *processo extremal* (Dwass, 1964), com uma função densidade de probabilidade

$$h_\gamma(x_1, x_2, \cdots, x_k) = g_\gamma(x_k) \prod_{j=1}^{k-1} \left\{g_\gamma(x_j)/EV_\gamma(x_j)\right\} \quad \text{se} \quad x_1 > x_2 > \cdots > x_k, \tag{5}$$

onde $g_\gamma(x) = dEV_\gamma(x)/dx$, com $EV_\gamma(x)$ definida em(2). Outros resultados limites em EVT, importantes no desenvolvimento de abordagens paramétricas à *Estatística de Extremos Univaridos*, podem ser vistos em Gomes *et al.* (2008a) ou na versão alargada, Gomes *et al.* (2007a).

2.2 Principais abordagens paramétricas

Método de Gumbel, dos máximos anuais ou dos blocos (BM**, do Inglês "*block method*").** Com $(\lambda_n, \delta_n) \in \mathbb{R} \times \mathbb{R}^+$, um vector de parâmetros desconhecidos de localização e escala, o *teorema de tipos extremais* valida a aproximação $P\big(X_{n:n} \leq x\big) = F^n(x) \approx EV_\gamma\left((x-\lambda_n)/\delta_n\right)$. Gumbel foi pioneiro no uso de aproximações deste tipo, mas para qualquer dos modelos em (3), sugerindo o chamado modelo BM, frequentemente designado por *modelo de máximos anuais*, *modelo EV univariado* ou ainda *modelo de Gumbel*. Este modelo consiste em dividir os n dados em k sub-amostras (usualmente correspondentes a k anos) de dimensão r ($n = rk$, r razoavelmente elevado) e ajustar à amostra formada pelos k máximos de cada sub-amostra um dos modelos extremais em (3) ou o modelo EV_γ, em (2), obviamente com parâmetros adicionais de localização e escala. Hoje em dia, ao usar esta abordagem, ainda muito popular na área ambiental, ajustamos aos dados um modelo $EV_\gamma((x-\lambda_r)/\delta_r)$, com EV_γ definida em (2), $(\lambda_r, \delta_r, \gamma) \in (\mathbb{R}, \mathbb{R}^+, \mathbb{R})$ parâmetros desconhecidos de localização, escala e forma.

Método das maiores observações (LO, do Inglês "*largest observations*"). Apesar do método BM se ter revelado frutuoso em situações diversas, têm-lhe sido feitas várias críticas, pois podemos nitidamente estar a perder informação relevante para acontecimentos extremos ao usar só o máximo observado em cada bloco. Além disso, em várias áreas de aplicação, não existe uma sazonalidade natural nos dados, parecendo de certo modo artificial e subjectivo o método das sub-amostras. Parece pois mais sensato considerar um número reduzido k de observações de topo da colecção de dados original, repondo assim alguma informação acerca da amostra inicial, que o método tradicional parece desperdiçar. Esta abordagem, se colocada em campo paramétrico, vai certamente depender do comportamento distribucional conjunto referido em (5). Temos então uma segunda abordagem à *Estatística de Extremos*, em que baseamos qualquer inferência numa amostra dependente com a estrutura multivariada anteriormente referida. É o designado *modelo EV multivariado*. Mais uma vez, devem-se considerar parâmetros de localização e escala desconhecidos, λ_n e δ_n, respectivamente, a serem estimados com base na amostra das k e.o. de entre n. Neste modelo é bem mais fácil aumentar a dimensão da amostra.

Abordagem EV multi-dimensional. Note-se ainda que se pode facilmente combinar as duas abordagens anteriormente referidas, considerando que em cada uma das sub-amostras podemos recolher algumas estatísticas de topo, as quais são modeladas pelo modelo *EV multivariado*. Temos então o chamado *modelo EV multidimensional*, em que temos acesso a uma amostra multivariada $(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_k)$, onde $\underline{X}_j = (X_{1j}, \ldots, X_{i_j j}), 1 \leq j \leq k$, são *vectores extremais multivariados*.

Abordagem POT, do Inglês "*peaks over threshold*". Uma outra perspectiva equivalente ao *modelo EV multivariado* é aquela em que restringimos a nossa atenção às observações que excedem um certo *limiar* ou *threshold*, ajustando modelos estatísticos apropriados quer aos *excessos* quer aos *picos* acima desse limiar. Quando $u \to x_0^F$, o limite superior do suporte de F, tem-se $P\left[X - u \leq t | X > u\right] \approx GP_\gamma(t/\beta)$, onde $GP_\gamma(\cdot)$ é a distribuição *GP univariada*, em (4). Considera-se então um nível elevado u, trabalhando-se com os excessos (diferença entre as observações que excedem o nível e o próprio nível), os quais são modelados por uma distribuição GP. Uma comparação entre as abordagens BM e POT pode ser vista em Engeland *et al.* (2004). Este modelo é também frequentemente designado por *modelo Paretiano de excessos*.

2.3 Abordagens não-paramétricas

Mais recentemente quer o método POT, quer o método LO, das maiores observações, têm vindo a ser abordados sob um ponto de vista não-paramétrico. O tipo de ajustamento utilizado para as maiores observações não se identifica então com uma forma paramétrica dependente de parâmetros de localização λ, de dispersão δ e de escala γ. Pressupõe-se apenas que F está no domínio de atração para máximos de EV_γ, sendo $\gamma = \gamma(F)$ um dos funcionais principais a estimar, com base em algumas observações de topo, e de acordo com metodologia adequada.

3 Inferência não-paramétrica em Estatística de Extremos

Em contexto não-paramétrico, ou de forma que consideramos mais apropriada, em contexto semi-paramétrico, trabalhamos com as k e.o. de topo, associadas à amostra disponível de dimensão n, ou com os excessos acima de um nível elevado, admitindo unicamente que, para um certo $\gamma \in \mathbb{R}$, o modelo F subjacente aos dados está em $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$ ou em sub-domínios específicos de $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$, com $EV_\gamma(\cdot)$ definida em (2), sendo $\gamma = \gamma(F)$ o funcional primordial de valores extremos a ser estimado, com base nas k observações de topo, e de acordo com metodologia adequada. Usualmente, necessitamos de basear a estimação de γ num k *intermédio*, i.e. tal que $k = k_n \to \infty$ e $k = o(n)$, i.e. $k/n \to 0$, quando $n \to \infty$. Esses estimadores, em conjunto com estimadores não-paramétricos de localização e escala (veja-se, por exemplo, de Haan & Ferreira, 2006), podem então ser usados para estimar quantis extremos, períodos de retorno de níveis elevados, probabilidades de excedências de níveis elevados e outros parâmetros de acontecimentos extremos. Após uma breve introdução às condições de primeira e segunda-ordem na Secção 3.1, referimos também brevemente, na Secção 3.2, alguns estimadores clássicos do EVI. Na Secção 3.3 abordamos a estimação de viés-reduzido (RB, do Inglês "*reduced-bias*") e na Secção 3.4, referimos a estimação não-paramétrica de outros parâmetros de acontecimentos extremos. Finalmente, na Secção 3.5, mencionamos alguns testes da condição $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$.

3.1 Condições de primeira e segunda-ordem

Tal como mencionámos na Secção 2.1, a caracterização completa de $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$ foi obtida em de Haan (1970), podendo ainda ser encontrada em de Haan & Ferreira (2006), entre outros. Com a notação U para a função quantil (recíproca) associada a F, definida por $U(t) := (1/(1-F))^{\leftarrow}(t) = F^{\leftarrow}(1-1/t) = \inf\{x : F(x) \geq 1 - 1/t\}$, a propriedade de *variação regular generalizada*,

$$F \in \mathcal{D}_{\mathcal{M}}(G_\gamma) \iff \lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = h_\gamma(x) := \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{se} \quad \gamma \neq 0 \\ \ln x & \text{se} \quad \gamma = 0, \end{cases} \tag{6}$$

válida $\forall x > 0$, com $a(\cdot)$ função mensurável positiva, é uma conhecida condição necessária e suficiente para se ter $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$. Os modelos de cauda pesada, i.e. modelos $F \in \mathcal{D}_{\mathcal{M}}^+ := \mathcal{D}_{\mathcal{M}}(EV_{\gamma>0})$, são importantes numa grande diversidade de áreas. Podemos então escolher $a(t) = \gamma\, U(t)$ em (6), e podemos dizer que $F \in \mathcal{D}_{\mathcal{M}}^+$ se e só se, para qualquer $x > 0$, $\lim_{t\to\infty} U(tx)/U(t) = x^\gamma$, i.e. U é de variação regular com índice γ, o que denotaremos por $U \in RV_\gamma$. Mais geralmente,

$$F \in \mathcal{D}_{\mathcal{M}}^+ \iff \overline{F} := 1 - F \in RV_{-1/\gamma} \iff U \in RV_\gamma.$$

Num contexto não-paramétrico, para além da condição de primeira-ordem, em (6), admitimos frequentemente uma condição de segunda-ordem, especificando a velocidade de convergência na condição de primeira-ordem. É então usual admitir a existência de uma função A^*, possivelmente não mudando de sinal e tendendo para zero, quando $t \to \infty$, tal que

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - h_\gamma(x)}{A^*(t)} = H_{\gamma,\rho}(x) := \frac{1}{\rho^*}\left(\frac{x^{\gamma+\rho^*}-1}{\gamma+\rho^*} - \frac{x^\gamma - 1}{\gamma}\right) \tag{7}$$

$\forall x > 0$, onde $\rho^* \leq 0$ é um parâmetro de *segunda-ordem* que controla a velocidade de convergência de valores máximos, linearmente normalizados, para a lei limite em (2). Então, $\lim_{t\to\infty} A^*(tx)/A^*(t) = x^{\rho^*}$, $\forall x > 0$, i.e. $|A^*| \in RV_{\rho^*}$. Para caudas pesadas, admitimos usualmente que sabemos a velocidade de convergência para zero de $\ln U(tx) - \ln U(t) - \gamma \ln x$, quando $t \to \infty$. Escrevemos então a condição de segunda-ordem na forma seguinte:

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \gamma \ln x}{A(t)} = \frac{x^\rho - 1}{\rho}, \tag{8}$$

onde $\rho \leq 0$ e $A(t) \to 0$ quando $t \to \infty$. De forma mais precisa, $|A| \in RV_\rho$. Para a relação entre $(A^*(t), \rho^*)$ e $(A(t), \rho)$, veja-se de Haan & Ferreira (2006) e Fraga Alves *et al.* (2007). As condições

de terceira-ordem especificam a velocidade de convergência quer em (7) quer em (8). De forma análoga, condições de ordem superior podem ser postuladas, mas estamos assim a restringir cada vez mais as distribuições a eleger em $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$.

3.2 Estimação clássica do EVI

Os estimadores básicos do EVI, que têm motivado refinamentos diversos, são os estimadores de *Hill* (H), de *Pickands* (P), dos *momentos* (M) e de pseudo-máxima verosimilhança (ML, do Inglês "*maximum likelihood*"), denotado PORT-ML, com PORT, do Inglês "*peaks over random threshold*", terminologia cunhada em Araújo Santos *et al.* (2006). De entre os estimadores mais recentes, referiremos apenas o generalizado de Hill (GH), o dos momentos mistos (MM) e o de *média-de-ordem-p* (MOP). Detalhes adicionais sobre alguns destes estimadores podem ser vistos em de Haan & Ferreira (2006), Gomes *et al.* (2008a) e Beirlant *et al.* (2012).

Estimador H. Para caudas-direitas pesadas, i.e. em $\mathcal{D}_{\mathcal{M}}^{+}$, o estimador mais simples do EVI foi proposto por Hill (1975). O estimador H, denotado $\hat{\gamma}_{n,k}^{H}$, é a média dos *log-espaçamentos escalados* e dos *excessos das log-observações*, $U_i := i\,\{\ln(X_{n-i+1:n}/X_{n-i:n})\}$ e $V_{ik} := \ln(X_{n-i+1:n}/X_{n-k:n})$, $1 \le i \le k < n$, respectivamente, i.e.

$$\hat{\gamma}_{n,k}^{H} := \frac{1}{k}\sum_{i=1}^{k} U_i = \frac{1}{k}\sum_{i=1}^{k} V_{ik}.$$

Estimador P. Para um EVI arbitrário, $\gamma \in \mathbb{R}$, denotando $\lfloor x \rfloor$ a parte inteira de x, e considerando como base de estimação as k e.o. de topo, podemos escrever o estimador P (Pickands, 1975) como

$$\hat{\gamma}_{n,k}^{P} := \big(\ln\big(X_{n-\lfloor k/4\rfloor+1:n} - X_{n-\lfloor k/2\rfloor+1:n}\big)/\big(X_{n-\lfloor k/2\rfloor+1:n} - X_{n-k+1:n}\big)\big)/\ln 2.$$

Estimador M. Dekkers *et al.* (1989), com base em $M_{n,k}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\{\ln(X_{n-i+1:n}/X_{n-k:n})\}^j$, $j > 0$, propuseram o estimador

$$\hat{\gamma}_{n,k}^{M} := M_{n,k}^{(1)} + \tfrac{1}{2}\big(1 - \big(M_{n,k}^{(2)}/[M_{n,k}^{(1)}]^2 - 1\big)^{-1}\big),$$

válido para $\gamma \in \mathbb{R}$.

Estimador PORT-ML. Condicionalmente a $X_{n-k:n}$, com k *intermédio*, $D_{ik} := X_{n-i+1:n} - X_{n-k:n}$, $1 \le i \le k$, são aproximadamente as k e.o. de topo associadas a uma amostra de dimensão k de um modelo $GP_\gamma(\alpha x/\gamma)$, $\alpha \in \mathbb{R}$, com $GP_\gamma(\cdot)$ definido em (4). A solução das equações ML associadas a este contexto leva-nos a um estimador ML implícito, $\hat{\alpha}$, do parâmetro de escala α, e a um estimador explícito do EVI, dado por

$$\hat{\gamma}_{n,k}^{\text{PORT}-\text{ML}} := \tfrac{1}{k}\sum_{i=1}^{k}\ln(1 + \hat{\alpha}\; D_{ik}).$$

Um estudo das propriedades assintóticas deste estimador pode ser encontrado em Drees *et al.* (2004).

Estimador GH. A inclinação de um gráfico de quantis generalizado levou à introdução do estimador GH, em Beirlant *et al.* (1996). Este estimador é válido para $\gamma \in \mathbb{R}$, e tem a forma funcional,

$$\hat{\gamma}_{n,k}^{GH} = \hat{\gamma}_{n,k}^{H} + \tfrac{1}{k}\sum_{i=1}^{k}\left\{\ln\hat{\gamma}_{n,i}^{H} - \ln\hat{\gamma}_{n,k}^{H}\right\}.$$

Veja-se também Beirlant *et al.* (2005).

Estimador MM. Recentemente, Fraga Alves *et al.* (2009) introduziram e estudaram o chamado estimador MM, que envolve não só os excessos das log-observações, mas também um outro tipo de

estatísticas de momentos, dadas por $L_{n,k}^{(1)} := \frac{1}{k}\sum_{i=1}^{k}\left(1 - X_{n-k:n}/X_{n-i+1:n}\right)$. Com base em $\hat{\varphi}_{n,k} := \left(M_{n,k}^{(1)} - L_{n,k}^{(1)}\right)/\left(L_{n,k}^{(1)}\right)^2$, com $M_{n,k}^{(1)} \equiv \hat{\gamma}_{n,k}^{H}$, podemos facilmente construir o estimador MM, válido para $\gamma \in \mathbb{R}$, e dado por

$$\hat{\gamma}_{n,k}^{MM} := \left(\hat{\varphi}_{n,k} - 1\right)/\left(1 + 2\min\left(\hat{\varphi}_{n,k} - 1, 0\right)\right).$$

Estimador MOP. Trata-se de uma generalização simples do estimador de Hill, em que em vez de usar a média geométrica de $U_i := X_{n-i+1:n}/X_{n-k:n}$, $1 \le i \le k$, se usa a média-de-ordem-p destas mesmas estatísticas (veja-se Brilhante *et al.*, 2013).

Consistência e normalidade assintótica dos estimadores. A consistência fraca de qualquer um dos estimadores do EVI anteriormente mencionados é conseguida nos sub-domínios de $\mathcal{D}_{\mathcal{M}}(EV_\gamma)$ em que são válidos, sempre que temos a validade da condição em (6) e $k = k_n$ é uma sucessão *intermédia* de inteiros entre 1 e $n-1$. Sob a validade da condição de segunda-ordem, em (7), é possível garantir a normalidade assintótica desses estimadores. De forma mais precisa, se denotarmos por T qualquer um dos estimadores do EVI atrás mencionados, e denotando por $B(t)$ uma função de viés—fortemente relacionada com a função $A^*(t)$ em (7), convergente para zero quando $t \to \infty$, é possível garantir a existência de $\mathcal{C}_{_T} \subset \mathbb{R}$ e $(b_{_T}, \sigma_{_T}) \in \mathbb{R} \times \mathbb{R}^+$, tais que:

$$\hat{\gamma}_{n,k}^{T} \overset{d}{=} \gamma + \frac{\sigma_{_T}}{\sqrt{k}} P_k^T + b_{_T} B(n/k) + o_p(B(n/k)), \tag{9}$$

com P_k^T uma sucessão de variáveis aleatórias assintoticamente normais padrão. Consequentemente, para k tal que $\sqrt{k}\ B(n/k) \to \lambda$, finito, quando $n \to \infty$, $\sqrt{k}\left(\hat{\gamma}_{n,k}^{T} - \gamma\right) \longrightarrow_{n\to\infty}^{d} \text{Normal}(\lambda b_{_T}, \sigma_{_T}^2)$. A $b_{_T}$ e $\sigma_{_T}^2$ chamamos respectivamente *viés assintótico* e *variância assintótica* de $\hat{\gamma}_{n,k}^{T}$.

3.3 Estimação de viés reduzido do EVI

Os estimadores não-paramétricos clássicos de qualquer parâmetro de acontecimentos extremos exibem usualmente um viés assintótico acentuado quando k, o número de e.o. de topo envolvidas na estimação, aumenta, e mesmo para o valor óptimo de k, no sentido de *erro quadrático médio* (MSE, do Inglês "*mean square error*") mínimo, apelando para uma redução de viés, de modo a trabalharmos com $\hat{\gamma}_{n,k}^{R}$ tal que em vez de (9) se tem $\hat{\gamma}_{n,k}^{R} \overset{d}{=} \gamma + \sigma_{_R} P_k^R/\sqrt{k} + o_p(B(n/k))$.

Estimadores RB do EVI. A redução do viés destes estimadores clássicos tem sido um tema amplamente discutido na literatura mais recente. Mencionamos no entanto os artigos pioneiros de Gomes (1994), Drees (1996), Peng (1998), Beirlant *et al.* (1999), Feuerverger & Hall (1999) e Gomes *et al.* (2000), em que surge sempre o usual "trade-off"entre variância e viés, com o desenvolvimento de estimadores em que se reduz o viés, aumentado a variância, e obtendo os chamados estimadores SORB, do Inglês "*second-order reduced-bias*. Esta abordagem tem sido considerada essencialmente para caudas pesadas. As ideias chave são encontrar maneiras de nos libertarmos da componente dominante do viés $b_T B(n/k)$, em (9), ou de avançar no estudo do comportamento de segunda-ordem de estatísticas base para estimação de γ, tais como os *excessos das log-observações* ou os *log-espaçamentos escalados*, de modo a termos um termo dominante de viés assintótico de ordem inferior a $B(n/k)$. É ainda de realçar que, recentemente, Cai *et al.* (2012) deram os primeiros passos na estimação SORB do EVI, com a introdução de um estimador SORB para $\gamma \in \mathbb{R}$, na vizinhança de zero, baseado na metodologia PWM, do Inglês "*probability weighted moments*".

Estimadores RB do EVI com variância-mínima (MVRB, do Inglês "*minimum-variance reduced-bias*"). Recentemente, o "trade-off"atrás mencionado conseguiu ser ultrapassado com a estimação adequada dos funcionais de segunda-ordem, tal como foi feito, para caudas pesadas, em Caeiro *et al.* (2005) e Gomes *et al.* (2007b; 2008b), que introduziram estimadores MVRB do EVI. Esses estimadores têm uma variância assintótica coincidente com a do estimador H, e um viés de ordem inferior, ultrapassando pois os estimadores clássicos para todo o k. Algoritmos para a estimação adequada dos parâmetros de segunda-ordem podem ser encontrados em Gomes

& Pestana (2007), entre outros. O uso desses algoritmos, em que o estimador de ρ é calculado em $k_1 = [n^{1-\epsilon}]$, com ϵ pequeno, permite-nos, face a uma ligeira restrição na classe de modelos em que trabalhamos, garantir a validade de uma propriedade crucial do estimador de ρ, que permite não haver aumento da variância assintótica. Essa propriedade crucial pode ser potencialmente verificada sem qualquer restrição, caso calculemos $\hat{\rho}$ no seu nível óptimo (veja-se Caeiro *et al.*, 2009), mas a escolha adaptativa desse nível é ainda um tópico de investigação em aberto. Detalhes adicionais sobre estimação SORB e MVRB podem ser encontrados nas recensões críticas em Reiss & Thomas, Capítulo 6, Gomes *et al.* (2008a) e Beirlant *et al.* (2012).

3.4 Estimação de outros parâmetros de acontecimentos extremos

Quantis elevados de probabilidade $1-p$, com p pequeno, são talvez os parâmetros de acontecimentos extremos mais relevantes. Tratam-se de funções do EVI, bem como da localização e escala da população em estudo. Em contexto não-paramétrico, os estimadores mais usuais de $\chi_{1-p} := U(1/p)$, com p pequeno, podem ser derivados a partir de (6), através da aproximação

$$U(tx) \approx U(t) + a(t)(x^{\gamma} - 1)/\gamma.$$

O facto de se ter $X_{n-k+1:n} \overset{p}{\sim} U(n/k)$ permite-nos estimar χ_{1-p} com base nesta aproximação e na estimação adequada de γ e de $a(n/k)$. Para o caso mais simples de caudas pesadas, temos a aproximação $U(tx) \approx U(t)x^{\gamma}$, donde derivamos

$$\hat{\chi}_{1-p,k} := X_{n-k:n} \ (k/(np))^{\hat{\gamma}_k},$$

sendo $\hat{\gamma}_k$ um qualquer estimador não-paramétrico e consistente do EVI. Este tipo de estimador foi introduzido pela primeira vez em Weissman (1978). Detalhes para a estimação não-paramétria de quantis para $\gamma \in \mathbb{R}$, podem ser encontrados em Dekkers & de Haan (1989), de Haan & Rootzén (1993) e mais recentemente em Ferreira *et al.* (2003). Outras abordagens à estimação de quantis elevados podem ser vistas em Matthys & Beirlant (2003). Aconselhamos ainda a leitura de Guillou *et al.* (2010) e de You *et al.* (2010). Contudo, nenhum dos estimadores anteriores reage adequadamente a mudanças de escala nos dados. Araújo Santos *et al.* (2006) introduziram uma classe de estimadores de *quantis elevados* gozando dessa característica, a contrapartida empírica da linearidade teórica de um quantil χ_p, i.e. $\chi_p(\delta X + \lambda) = \delta\chi_p(X) + \lambda$, $\forall\lambda \in \mathbb{R}$ e escalar positivo δ. Esta classe de estimadores é baseada na metodologia PORT e fornece propriedades exactas para as medidas de risco em finanças: invariância para translações e homogeneidade positiva. A estimação da *probabilidade de excedência de níveis fixos elevados*, do *limite superior do suporte*, do *valor médio de uma distribuição com cauda-direita pesada*, e do *coeficiente de cauda Weibull* têm sido amplamente estudados em contexto não-paramétrico (vejam-se detalhes adicionais em de Haan & Ferreira, 2006, Gomes *et al.*, 2008a, e Beirlant *et al.*, 2012). Para uma estimação SORB e não paramétrica de outros parâmetros aconselhamos a consulta de bibliografia em Reiss & Thomas (2007), Capítulo 6, Gomes *et al.* (2008a) e Beirlant *et al.* (2012).

3.5 Testes em esquema não-paramétrico

Testar, em contexto não-paramétrico, a hipótese $H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_0)$ versus $H_1 : F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$, $\gamma \neq 0$, ou versus alternativas unilaterais, é obviamente sensato. Em sentido lato, já se podem encontrar testes desta natureza em artigos anteriores a 2000 (veja-se Gomes *et al.*, 2007a). Testes puramente não-paramétricos aparecem em Jurečková & Picek (2001). Mas o teste à condição de valores extremos pode ser atribuído a Dietrich *et al.* (2002), que propõem uma estatística para testar se a condição $F \in \mathcal{D}_{\mathcal{M}}(EV_\gamma)$ é ou não suportada pelos dados, juntamente com uma versão mais simples para testar $F \in \mathcal{D}_{\mathcal{M}}(EV_{\gamma\geq 0})$. Encontram-se recensões críticas do tema em Hüsler & Peng (2008) e em Neves & Fraga Alves (2008).

4 Comentários gerais e tópicos de investigação futura

Na nossa opinião a *Estatística de Extremos Univariados* continua a ser um campo interessante de investigação, onde ainda existem muitos tópicos em aberto. Recentemente, têm surgido desenvolvimentos importantes na área de *extremos espaciais*, em que *modelos paramétricos* parecem ter voltado a revelar-se de enorme importância. Neste caso, e numa altura em que temos acesso a técnicas computacionais altamente sofisticadas, existe uma grande diversidade de *modelos paramétricos*, que podem e devem ser considerados. E em contexto não-paramétrico, tópicos como a *seleção de níveis* (veja-se Gomes *et al.*, 2012), *tendências e pontos de mudança* no *comportamento da cauda*, e o "*clustering*"de observações elevadas em situações de dependência, fortemente relacionados com a existência de um *índice extremal* $\theta < 1$, são, entre outros, tópicos fortemente desafiadores. Testes não-paramétricos da condição $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$ são cruciais e já foram delineados por vários autores, tal como se referiu na Secção 3.5. E testes às condições de segunda e de terceira-ordem? A *Estatística de Extremos Univariados* para dados sujeitos a censura aleatória é também um tópico em que ainda há muito a fazer. Além disso, a estimação de parâmetros de segunda e terceira-ordem ainda merece atenção cuidada, particularmente devido à sua importância na estimação SORB de parâmetros de acontecimentos extremos.

Acknowledgements. Investigação parcialmente financiada por Fundos Nacionais através da **FCT** — Fundação para a Ciência e a Tecnologia, projectos PEst-OE/MAT/UI0006/2011 (CEAUL) e PEst-OE/MAT/UI0297/2011 (CMA/UNL).

Referências

[1] Araújo Santos, P.; Fraga Alves, M.I. and Gomes, M.I. (2006). Peaks over random threshold methodology for tail index and quantile estimation, *Revstat* **4**:3, 227–247.

[2] Balkema, A.A. and Haan, L. de (1974). Residual life time at great age, *Annals of Probability* **2**, 792–804.

[3] Beirlant, J.; Vynckier, P. and Teugels, J. (1996). Excess functions and estimation of the extreme-value index, *Bernoulli* **2**, 293–318.

[4] Beirlant, J.; Dierckx, G.; Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes* **2**, 177–200.

[5] Beirlant, J.; Dierckx, G. and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots, *Bernoulli* **11**:6, 949–970.

[6] Beirlant, J.; Caeiro, C. and Gomes, M.I. (2012). Overview and open research topics in Statistics of Univariate Extremes, *Revstat* **10**:1, 1–31.

[7] Brilhante, M.F.; Gomes, M.I. and Pestana, D.D. (2013). A simple generalization of the Hill estimator, *Comput. Statist. and Data Analysis* **57**:1, 518–535.

[8] Caeiro, C.; Gomes, M.I. and Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator, *Revstat* **3**:2, 113–136.

[9] Caeiro, F.; Gomes, M.I. and Henriques-Rodrigues, L. (2009). Reduced-bias tail index estimators under a third order framework, *Comm. Statist. – Theory and Methods* **38**:7, 1019–1040.

[10] Cai, J.; de Haan, L. and Zhou, C. (2012). *Bias correction in extreme value statistics with index around zero*, *Extremes*, DOI 10.1007/s10687-012-0158-x.

[11] Castillo, E.; Hadi, A.; Balakrishnan, N. and Sarabia, J.M. (2005). *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley, Hoboke, New Jersey.

[12] Dekkers, A.L.M. and Haan, L. de (1989). On the estimation of the extreme-value index and large quantile estimation, *Ann. Statist.* **17**, 1795–1832.

[13] Dekkers, A.; Einmahl, J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics*, **17**, 1833–1855.

[14] Dietrich, D.; de Haan, L. and Hüsler, J. (2002). Testing extreme value conditions, *Extremes* **5**, 71–85.

[15] Drees, H. (1996). Refined Pickands estimators with bias correction, *Comm. Statist. Theory and Meth.* **25**, 837-851.

[16] Drees, H.; Ferreira, A. and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index, *Annals of Applied Probability* **14**, 1179–1201.

[17] Dwass, M. (1964). Extremal processes, *Ann. Math. Statist.* **35**, 1718-1725.

[18] Engeland, K.; Hisdal, H. and Frigessi, A. (2004). Practical extreme value modelling of hydrological floods and droughts: a case study, *Extremes* **7**:1, 5–30.

[19] Ferreira, A.; de Haan, L. and Peng, L. (2003). On optimizing the estimation of high quantiles of a probability distribution, *Statistics* **37**:5, 401–434.

[20] Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution, *Ann. Statist.* **27**, 760–781.

[21] Fisher, R.A. and Tippett, L.H.C. (1928). Limiting forms of the frequency of the largest or smallest member of a sample. *Proc. Cambridge Phil. Soc.* **24**, 180–190.

[22] Fraga Alves, M.I.; Gomes, M.I.; de Haan, L. and Neves, C. (2007). A note on second order condition in extremes: linking general and heavy tails conditions, *Revstat* **5**:3, 285–305.

[23] Fraga Alves, M.I.; Gomes, M.I.; de Haan, L. and Neves, C. (2009). The mixed moment estimator and location invariant alternatives, *Extremes*, **12**, 149–185.

[24] Fréchet, M. (1927). Sur le loi de probabilité de l'écart maximum, *Ann. Société Polonaise de Mathématique* **6**, 93–116.

[25] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics* **44**:6, 423–453.

[26] Gomes, M.I. (1994). Metodologias Jackknife e Bootstrap em Estatística de Extremos. In Mendes-Lopes et al. (eds.), *Actas II Congresso S.P.E.*, 31–46.

[27] Gomes, M.I. and Pestana, D. (2007). A sturdy reduced bias extreme quantile (VaR) estimator, *J. American Statistical Association* **102**:477, 280–292.

[28] Gomes, M.I.; Martins, M.J. and Neves, M. (2000). Alternatives to a semi-parametric estimator of parameters of rare events — the Jackknife methodology, *Extremes* **3**:3, 207–229.

[29] Gomes, M.I.; Canto e Castro, L.; Fraga Alves, M.I. and Pestana, D. (2007a). *Statistics of extremes for iid data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions*, Notas e Comunicações CEAUL 16/2007.

[30] Gomes, M.I.; Martins, M.J. and Neves, M. (2007b). Improving second order reduced bias extreme value index estimation, *Revstat* **5**:2, 177–207.

[31] Gomes, M.I.; Canto e Castro, L.; Fraga Alves, M.I. and Pestana, D. (2008a). Statistics of extremes for iid data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions, *Extremes* **11**:1, 3–34.

[32] Gomes, M.I.; de Haan, L. and Henriques-Rodrigues, L. (2008b). Tail Index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses, *J. Royal Statistical Society B* **70**:1, 31–52.

[33] Gomes, M.I.; Figueiredo, F. and Neves, M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action, *Extremes* **15**, 463–489.

[34] Guillou, A.; Naveau, P. and You, A. (2010). A folding method for extreme quantiles estimation, *Revstat* **8**, 21–35.

[35] Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

[36] Haan, L. de (1970). *On Regular Variation and its Application to the Weak Convergence of Sample Extremes*, Mathematical Centre Tract 32, Amsterdam.

[37] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*, Springer Science+Business Media, LLC, New York, USA.

[38] de Haan, L. and Rootzén, H. (1993). On the estimation of high quantiles, *J. Statist. Planning and Inference* **35**, 1–13.

[39] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics* **3**:5, 1163–1174.

[40] Hüsler, J. and Peng, L. (2008). Review of testing issues in extremes: in honor of Professor Laurens de Haan, *Extremes* **11**, 99–111.

[41] Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quart. J. Royal Meteorol. Society* **81**, 158–171.

[42] Jurečková, J. and Picek, J. (2001). A class of tests on the tail index, *Extremes* **4**:2, 165–183.

[43] Markovich, N. (2007). *Nonparametric Analysis of Univariate heavy-tailed Data*, John Wiley & Sons, England.

[44] Matthys, G. and Beirlant, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models, *Statistica Sinica* **13**, 853–880.

[45] Mises, R. von (1936). La distribution de la plus grande de n valeurs, *Revue Math. Union Interbalcanique* **1**, 141-160. Reprinted in *Selected Papers of Richard von Mises*, Amer. Math. Soc. **2** (1964), 271–294.

[46] Neves, C. and Fraga Alves, M.I. (2008). Testing extreme value conditions — an overview and recent approaches, *Revstat* **6**:1, 83–100.

[47] Peng, L. (1998). Asymptotically unbiased estimator for the extreme-value index, *Statist. Probab. Letters* **38**:2, 107-115.

[48] Pickands III, J. (1975). Statistical inference using extreme order statistics, *Ann. Statist.* **3**, 119-131.

[49] Reiss, R.-D. and Thomas, M. (2001; 2007). *Statistical Analysis of Extreme Values, with Application to Insurance, Finance, Hydrology and Other Fields*, 2nd edition; 3rd edition, Birkhäuser Verlag.

[50] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *J. Amer. Statist. Assoc.* **73**, 812–815.

[51] You, A.; Schneider, U.; Guillou, A. and Naveau, P. (2010). Improving extreme quantile estimation by folding observations, *J. Statist. Plann. Infer.* **140**, 1775–1787.

Recordações e Reflexões Sobre o Ensino da Estatística (Nomeadamente a Estudante de Outras Ciências)[1] Com um *Post Scriptum* Sobre o Desafio do Presidente da SPE Para Que Haja Mais Debate de Ideias no Boletim[2]

Dinis Pestana, *ddpestana@fc.ul.pt*

CEAUL - *Centro de Estatística e Aplicações da Universidade de Lisboa*
CFCUL - *Centro de Filosofia das Ciências da Universidade de Lisboa*
Instituto de Investigação Científica Bento da Rocha Cabral

Há um notável conto do Asimov, chamado *Profession*[3], num futuro em que há um dia da escola, dia (*Reading Day*) em que as crianças vão à escola e ficam a saber qualquer coisa semelhante ao que actualmente seria chamada uma boa educação pré-universitária, e um dia chamado *Education Day*, em que após a adolescência voltam à escola, e no fim do dia são profissionais competentes, cada qual de seu ramo. Posteriormente há um *Olympics Day* (já ninguém sabe bem porquê esse nome) em que são recrutados para uma profissão — mas nem sempre a que ambicionam. Para o protagonista da história, George Platen, é um dia de inesperada decepção, pois não é escolhido para uma das profissões que ambicionava, um dos que iria usar as tecnologias mais avançadas, ou um dos enviados para áreas remotas do universo sendo os pioneiros de novas expansões da humanidade; de facto, no fim do dia não foi recrutado para nada. Parte da estória descreve a sua frustração e revolta, talvez mesmo raiva contra o "acompanhante" que foi designado para o orientar enquanto aprende coisas, já que não aprendeu uma especialidade no *Olympics Day*. E é quando ele bate no fundo da depressão, e exprime a sua frustração e raiva, que o seu guardião lhe explica que aqueles que ele inveja têm uma formação específica numa tecnologia que está sempre a mudar, e que ao fim de três ou quatro anos já não serão os pioneiros da fronteira do desenvolvimento, porque os seus conhecimentos os ligam a tecnologia que vai ficando obsoleta.
A formação que lhe foi dada, porque encontraram nele o gosto do saber pelo saber, foi a aprendizagem permanente, na esperança de que se tornasse também ele um cientista, um dos construtores do saber e

[1] Este trabalho foi financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto PEstOE/MAT/UI0006/2011.

[2] Que bonito e longo título à (i)moralista século XVIII, ou à Jorge Amado em *Os Velhos Marinheiros ou o Capitão-de-Longo-Curso* (exemplo: "*Primeiros Episódio: Da Chegada do Comandante ao Subúrbio de Periperi, na Bahia, do Relato de suas Mais Famosas Aventuras nos Cinco Oceanos, em Mares e Portos Longínquos, com Rudes Marinheiros e Mulheres Apaixonadas, e da Influência do Cronógrafo e do Telescópio Sobre a Pacata Comunidade Suburbana"*).
Experimentem a ironia luminosa de Jorge Amado dos anos 60, lendo o primeiro capítulo em *http://www.companhiadasletras.com.br/trechos/12628.pdf,* e vejam se resistem a ir comprar o livro.

[3] Encontra-se por exemplo no excelente *Nine Tomorrows*, mas aparece também com o título *Olympics*, ou *Olympics Day*, em muitas outras colectâneas de contos de Asimov. Atualmente, disponível em *http://www.abelard.org/asimov.php*.

da inovação. E o conto termina num clima de cautelosa esperança, pois honestamente o seu guia lhe diz que muitos fracassam[4].

Aprender pelo gosto de aprender é aquilo que mais ambicionamos encontrar nos nossos alunos. Em horas de sorte, conseguimos mesmo ser parte do fermento que transforma alguns alunos nessa espécie extravagante de gente, e quando olho para trás tenho pelo menos a ilusão de que em alguns casos participei nessa espécie de maiêutica de que Sócrates (o genuíno, claro) falava. Por esses alunos eu sinto gratidão, talvez mesmo mais do que a que eles sentem por mim, pois eles foram de facto um arco-íris na minha vida profissional.

Claro que também tive dissabores, mas que olhados à distância me provocam um sorriso. Num dos últimos anos em que ensinei, uma aluna de Bioestatística (cadeira em que no passado quase só tinha alunos distintos) pôs o dedinho no ar para colocar uma questão - estava eu a explorar as ideias de especificidade e sensibilidade, de valor preditivo positivo e negativo, que usei como um exemplo de aplicação do teorema de Bayes. A questão que colocou foi "Professor, isso que está para aí a dizer pode sair em exame?". Sosseguei-a, esclarecendo que nada do que era dado nas aulas podia sair em exame, e creio que a ironia da resposta foi deitar pérolas a uma porca.

Nos últimos anos, fui constatando que a preparação prévia dos alunos que entram na universidade declinou gravemente, mas felizmente nunca tive que abdicar de manter o nível das minhas cadeiras num padrão de exigência semelhante ao do ensino nas boas universidades dos países desenvolvidos. Isto é, para ser completamente realista: sempre me mantive exigente comigo, a exigência com os alunos teve que ter em linha de conta a mediania de muitos deles - e felizmente aquela ganga cuja vivência universitária parece limitar-se a praxes em geral tinha a sensatez de se auto excluir de aulas e exames.

Fui muitas vezes acusado de exagerar na quantidade de matéria dos meus programas. Discordo, e desde muito cedo comecei a disponibilizar aos alunos material de trabalho adequado, que vim a transformar num livro, e indicações de livros existentes na biblioteca, onde poderiam aprender muito mais do que o que se pode ensinar num escasso número de horas de aulas. Penso que ter passado pela Faculdade de Letras teve alguma influência no modo de encarar o ensino, pois é (ou era) uma Faculdade em que o ensino universitário tinha de facto a preocupação de rasgar horizontes. Lembro-me que o Professor Mário de Albuquerque, que todos os anos mudava o programa de História da Cultura Medieval - e a cultura de 1000 anos dá sem dúvida a possibilidade de construir inúmeros programas interessantes - na primeira aula escrevia no quadro os dez pontos do programa desse ano (e eram temas monumentais, como "a simbólica medieval", ou "heterodoxias na Idade Média"), e anunciava: "Nas aulas vou expor um destes temas, para exemplificar como os senhores alunos devem estudar os outros na biblioteca". Claro que no ensino de uma ciência tão hierarquizada como a Matemática este ensino "aberto" é impossível, mas não se deve daí concluir que devemos desistir de incutir nos alunos o desejo de saber mais do que o que ensinamos nas aulas, e talvez se tenha como recompensa de vez em quando ser o aluno a ensinar o professor.

Não consigo apreciar o ensino sem ambição, as plataformas *Moodle* que não incentivam os alunos a procurar fontes documentais mais ricas, a cultura superficial que se adquire na *net*. Felizmente, a distribuição do serviço docente que me coube nos últimos anos de exercício da profissão não me fez ter contacto direto com a miserável decadência daquilo a que hoje chamam mestrados, que nem chegam a ser uma caricatura do que antes era ensinado nos últimos anos da licenciatura. Condeno - é a única palavra que me parece adequada - o ensino baseado na exibição de *outputs* de computador, um ensino desajeitado, que não ensina nem Estatística nem sequer a fazer uma análise de dados inteligente e adequada, e que decerto é mais volátil do que éter.

Confidenciava-me o Professor Andrade e Silva, quando foi notável Presidente do Conselho Científico da FCUL, que por vezes era preciso muita pedagogia com os colegas. Não sinto que tenha tido êxito nessa componente da vida universitária, pois inutilmente tentei em muitas ocasiões melhorar o ensino

---

[4] *"Even after we've thinned out the possibilities on Education Day, nine out of ten of those who come here are not quite the material of creative genius, and there's no way we can distinguish those nine from the tenth that we want by any form of machinery. The tenth one must tell us himself.*
[…] *It won't do to say to a man, 'You can create. Do so.' It is much safer to wait for a man to say, 'I can create, and I will do so whether you wish it or not.'* […]. *We can't allow ourselves to miss one recruit to that number or waste our efforts on one member who doesn't measure up."*

da Estatística na minha instituição, e em outras em que por empréstimo trabalhei, sem o êxito que esperava.

Diz Descartes, nas *Meditações Metafísicas*, que o bom senso foi a coisa que Deus melhor distribuiu, pois cada qual está satisfeito com o que lhe coube em sorte. Este ilusório êxito individual muitas vezes redunda em fracassos coletivos. Foi um progresso assinalável a Estatística ter-se tornado cadeira obrigatória em quase todos os cursos científicos. Mas o "bom senso" de quem decidia sobre a arquitetura dos cursos, percebendo que muitas cadeiras das suas áreas só deviam ser ensinadas a estudantes já com alguma maturidade, foi recuando o ensino da Estatística, do segundo semestre do segundo ano para o primeiro semestre, e agora em alguns casos para o segundo semestre do primeiro ano, sem atender a que tão cedo quem escolheu Farmácia ou Biologia ainda não percebe porque é que lhe estão a impingir uma cadeira que lhes parece espúria. Claro que uma exemplificação cuidada, tendo o cuidado de ir chamando a atenção para questões de Amostragem e de Planeamento de Experiências, pode alterar esse primeiro repúdio por parte dos alunos, mas estes não podem colher nessa fase os benefícios que uma cadeira de Estatística pode ter quando é ensinada a alunos que já podem apreciar plenamente o seu papel na metodologia das ciências experimentais.

Numa cadeira propedêutica de Estatística, que se sabe ser a única do curso, não se pode deixar de gastar algum tempo com análise inicial dos dados (recordando estatística descritiva já esquecida, expondo algumas inovações da análise exploratória dos dados), Probabilidade, e as bases dos procedimentos inferenciais, ilustrados naturalmente com os maravilhosos resultados exatos quando se usa o modelo gaussiano (o que implica alguma referência, mesmo muito etérea, ao teorema limite central). Mas esses resultados exatos derivam da independência da média e variância empíricas, que é característica do modelo gaussiano, não se verificando em nenhum outro (a abordagem a alternativas não-paramétricas não deve ser omitida, mas tem que ser feita com uma tal brevidade que se torna impossível uma discussão comparativa de métodos paramétricos / métodos não-paramétricos, amputando o ensino da ambição que devia ter de semear devidamente ideias e discutir conceitos). Não há a possibilidade de explorar sequer os rudimentos do método de Monte Carlo e simulação, e as referências a Amostragem e Planeamento de Experiências, cujos rudimentos me parecem imprescindíveis[5] só pode ser ocasional e muito parcelar. Penso que a ausência de formação adequada em Metrologia e em recolha e organização da informação são uma das falhas graves de quase todos os cursos. A possibilidade, pelo menos opcional, de os alunos terem algum contacto com Controle de Qualidade, e com Simulação, seria decerto uma mais-valia para muitos graduados.

Outra questão em que nunca consegui "converter" devidamente os membros do Conselho Científico foi a minha reiterada esperança que todos os cursos de mestrado, ou pelo menos os estudos avançados que são atualmente o início do doutoramento, deveriam ter uma componente de metodologia da investigação científica. Recomendei tão frequentemente quanto foi possível que lessem o notável *White Paper* publicado pelo governo britânico em 1993, intitulado *Realizing Our Potential*, ou pelo menos o interessante livro *Research Methods for Postgraduates*, de T. Greenfield - que *The New Scientist* classificou como "*The most useful book any postgraduate could ever buy*" - cujo Prefácio, na primeira edição, abre com a frase

> *The* [UK] *government proposed in 1994 in their White Paper* Realising Our Potential *that all graduates who wish to study for doctorate should first take a one-year master's course in research methods. Several universities have since introduced such course and more are planned. This book is a response to that development*.

e deviam consultar o índice, para verificar o papel que a Estatística tem nessa formação (cerca de um terço do livro trata ou diretamente de Estatística ou de questões que têm diretamente que ver com Estatística — basta referir o título de alguns capítulos: 11: *Choosing and using software for statistics*; 19: *Randomized trials*; 20: *Laboratory and industrial experiments*; 21: *Agricultural experiments*; 22: *Survey research*; 23: *Principles of sampling*; 24: *Sampling in human studies*; 25: *Sources of population statistics*; 25: *Interviewing*; 29: *Elementary statistics*; 30: *Further statistical methods*; 31: *Computer support for data analysis*; 35: *Stochastic models and simulation*). Recomendações muitas

[5] Na comemoração do cinquentenário da publicação do *best-seller How to Lie with Statistics*, o fascículo especial que *Statistical Science* dedicou à efeméride tem um interessante artigo sobre *How to Lie with Bad Data*, que mereceria muita atenção.

vezes repetidas, com esperança de um dia ser escutado e não apenas ouvido, mas tão em vão quanto no belíssimo poema de David Mourão Ferreira, que começa

Em vão se agarra o Sol
Em vão se agarra o Sol às paredes do céu![6]

Consolo-me, por outro lado, a ver tantos jovens cientistas que formámos nos mestrados e doutoramentos do departamento em que tão longamente trabalhei, que tendo vocação para a docência têm levado um ensino de qualidade da Estatística para tantas universidades, tantos politécnicos, tantos institutos de investigação. Isso parcialmente me consola de esporadicamente ainda observar "carreiras" de quem, não tendo tido uma formação adequada, não teve a humildade de a procurar, e produz um ensino de qualidade lamentável, muitas vezes ainda procurando trepar para posições de destaque que não merecem. Um bom estatístico pode nascer em qualquer área - e os passos pioneiros de Galton, Pearson, Student, Fisher, e muitos outros, foram dados por biólogos, químicos, geneticistas - mas atualmente um cientista de outros campos só com sério investimento se pode transformar num estatístico profissional; não é decerto por se usar *software* estatístico - que produz sempre resultados (qualquer que seja a qualidade e fiabilidade dos dados, e a quantidade de dados omissos, e mesmo que o que se pede não seja adequado para usar com dados numa "escala de medição"[7], que não permite que as operações aritméticas envolvidas no procedimento computacional sejam válidas) - que se fica habilitado a ensinar Estatística.
A recomendação de humildade e trabalho vale também para todos aqueles que cursaram Estatística, e querem ser estatísticos profissionais: nenhum curso universitário produz um produto acabado, é um mero prefácio para o muito investimento em aprendizagem continuada, que tem que se fazer para conhecer bem uma parcela da Estatística. E isso, por sua vez, é a preparação de base que permite aprender sempre mais, quando uma consulta, a necessidade de intervir noutra área, ou as responsabilidades de dirigir, coordenar, orientar, nos obrigam a novos investimentos, aprendendo até ao fim da vida (depois disso não sei).

Post-Scriptum

Este texto foi escrito por solicitação direta do Presidente da SPE, Carlos Daniel Paulino, exprimindo o seu desejo de que houvesse uma secção regular de **Controvérsias** no *Boletim da SPE* - uma ideia que só posso aplaudir.
Pensando que o que fui capaz de escrever não se enquadrava propriamente nessa índole, enviei-lhe o texto que consegui produzir submetendo ao seu critério se queria ou não publicá-lo.
Teve a gentileza de aceitá-lo como está, mas sinto-me compelido a juntar este *Post Scriptum*, citando os comentários que me enviou - que considero uma fonte de reflexão para outros que venham a escrever para essa secção:

> ``[...] **Controvérsias**, *que queria e quero criar no Boletim para agitar as águas correntes quantas vezes demasiado estagnadas*.
>
> [...] *a título informativo, eis algumas reflexões ao correr da pena sobre tópicos que gostaria que perpassassem por aquela secção:*
>
> 1. *Independentemente de qualquer área científica, que valores estão implícitos, ou mesmo explícitos, nas atuais tendências pedagógicas dominantes?*

[6] Quem quiser ler essa extraordinária *Elegia do Outono* - será do Outono da vida? -, de *Tempestade de Verão*, pode encontrá-la em *http://montalvoeascinci.asdonossotempo.blogspot.pt/2012/01/poesia-maria-alzira-seixo-ha-la-relacao.html.*
[7] No sentido específico dado a esta expressão pela célebre e ainda actual análise de Steven, S. S. (1946), On the theory of scales of measurement, *Science* **103**, 677-680.

2. *Que competências se pretende que os quadros superiores a formar devem adquirir? O rigor e a exigência devem apenas remeter-se a uma figura de retórica?*

3. *Seja o ensino tematicamente mais abrangente ou mais confinado, como é avaliada e com que objetivos a aprendizagem?*

4. *Qual a finalidade na realidade (e não na retórica) da Universidade presentemente, além de criar e transferir conhecimento? Ensinar os seus alunos, abrindo horizontes, e avaliar consonantemente a aprendizagem conseguida, ou ao invés aplicar um receituário tolhido e aprovar em número tido a priori como suficiente de forma a ludibriar o real combate ao insucesso com falseadoras estatísticas - numa concreta manifestação de como se mente com a Estatística?*

5. *Não é precisamente a avaliação o meio mais transparente de se evidenciar quais os os objetivos que norteiam uma escola?*

6. *Que tem a este nível gerado a dita saudável competição entre escolas?*
Não estão estas mais preocupadas com a sua sobrevivência no plano mais imediatista, por força de constrangimentos externos, e menos com o futuro da sociedade em que estão inseridas? Não estará também aqui a génese do facilitismo e da colocação na prateleira ou outras formas de silenciamento daqueles que bravamente lhe resistem?

São questões de grande relevo, amplamente merecedoras de debate, e só posso aplaudir o nosso Presidente por querer vê-las discutidas, e por desafiar outros a abordá-las em vez de ter uma abordagem dogmática de expor prioritariamente as opiniões (porventura suculentamente controversas) que decerto tem sobre estes assuntos.
Não podendo discutir em detalhe qualquer delas, e ainda menos todas, decidi juntar no entanto alguns comentários:

1. Quando tive a honra de trabalhar, sob a sábia batuta do Professor Veiga Simão, na avaliação de cursos universitários que o CRUP empreendeu, usando como base documentos de auto-avaliação preparados pelos departamentos universitários que tinham a responsabilidade de organizar essas licenciaturas, imediatamente nos demos conta das limitações inultrapassáveis decorrentes de não ter sido previamente estabelecido um ``caderno de encargos", apropriado, com a antecedência necessária, explicitando devidamente as bases e critérios da avaliação. O Guião, fornecido aos responsáveis com uma antecedência exígua, era apenas uma espécie de ``borrão" do que poderia ter transformado os relatórios de auto-avaliação num instrumento de trabalho que permitisse uma análise comparativa justa entre as diversas ofertas pedagógicas sobre qualquer das áreas (decididas com critérios um pouco subjetivos, respeitando mais a tradição do que a realidade atual da Ciência, em minha opinião) em que foram classificados os cursos.

2. A Lei de Bases do Sistema Educativo é um documento admirável, e surpreende-me que muitos profissionais do ensino nunca a tenham lido.
(*http://www.dges.mctes.pt/NR/rdonlyres/AE6762DF-1DBF-40C0-B194-E3FAA 9516D79/1766/Lei46_86.pdf*)
Em particular o artigo 11º (âmbitos e objetivos do ensino superior) promove um ideário notável. A minha querida amiga Professora Maurícia Oliveira (que por azares da vida teve que deixar prematuramente a docência universitária) teve a argúcia de construir um inquérito para os alunos, levando-os a pronunciarem-se sobre a qualidade do ensino nas perspetivas dos objetivos enumerados nesse artigo. Recomendo que a construção de inquéritos seja feita atendendo à necessidade de averiguar se os objetivos do ensino superior estão ou não a nortear o ensino efetivo nas respetivas instituições.

3. A atual avaliação de desempenho, em que cada instituição inventou as suas regras, em que há uma promiscuidade de avaliadores e avaliados que só pode gerar desconfianças, em que as escolas mais inteligentes têm o cuidado de exibir um corpo docente de excelência, enquanto outras que apenas se

julgam inteligentes parecem mais apostadas em rebaixar o seu corpo docente (e se calhar os resultados junto do público farão justiça a essa argúcia e a essa mediocridade), parece-me um retrocesso ao já imperfeito sistema anteriormente usado pelo CRUP - que pelo menos tinha a menor parcialidade de os avaliadores serem externos, e um arremedo de uniformização de critérios dentro de cada área, podendo assim contribuir para estimular uma saudável competição entre as ofertas pedagógicas de diversas escolas, e permitindo a formação de uma opinião pública mais esclarecida.
Quando observo os ímpetos castigadores de alguns regulamentos, não resisto a recordar a notável conversa entre o diabo e o inquisidor em *O Físico Prodigioso* de Jorge de Sena, em que o inquisidor reage indignado à imposição do Diabo de desistirem do processo contra o Físico, perguntando "*e o castigo dos males?",* ao que o Diabo responde com ironia "*Mas a quem compete castigar, senão a mim?*" (cito de memória).
De facto, o que se deve pensar de regulamentos em que um dos critérios de base tem que ver com opiniões expressas por alunos (o que seria bom, se houvesse algum cuidado de verificar se os alunos foram às aulas que criticam, e se não se está simplesmente a incentivar o laxismo, como se fez no ensino secundário ao castigar os professores que classificavam alunos com notas "negativas", obrigando-os a aulas extras) - mas em que os dirigentes têm à cabeça pontuação máxima, sem minimamente se investigar por inquérito a opinião dos seus pares sobre a governação que efetivamente fazem, tantas vezes contrária às intenções manifestadas pré-eleitoralmente?
E que pensar de regulamentos que claramente estimulam um egoísmo que a meu ver não são as melhores "condições para a promoção da investigação científica" prometidas no artigo 15º da Lei de Bases do Sistema Educativo? Sempre partilhei as minhas ideias - e obviamente o crédito - com colegas de trabalho, e sempre considerei que esta era uma condição para promover o futuro das instituições, por isso fico perplexo com regras do tipo: se um trabalho for assinado por mais do que três autores, conta como 1/nº de autores na creditação atribuída, para avaliação, a qualquer deles (enquanto até 3 autores conta como um trabalho para cada um deles, uma descontinuidade que me parece muito injusta).
Por outro lado, parece-me um direito fundamental que os critérios de avaliação sejam amplamente divulgados com uma antecedência de dois ou três anos, por forma a permitir uma adequação do desempenho ao que as instituições decidiram considerar padrões e critérios de avaliação.

4. Ampliando um pouco os objetivos de "agitar as águas correntes", penso que esta secção de Controvérsias podia ser uma excelente base de reflexão para se submeter à Assembleia Geral da SPE, em tempo oportuno, um "código deontológico/regras de conduta" do exercício da profissão de estatístico[8].
Muitas sociedades científicas de relevo têm esse tipo de documento, recomendo a leitura de alguns, e sobretudo que a SPE tenha entre os seus objetivos propor alguma regulamentação que prestigie a Estatística na opinião pública.

Recordação de um sucesso esquecido

Como a Direção da FCUL determinou recentemente que os professores aposentados passam a dispor de espaço 0 nas instalações da instituição, 3 anos após a aposentação, ando a arrumar a desarrumação do meu gabinete (e a desarrumar um pouco mais a desarrumação de casa).
A parte boa é que vou encontrando papéis já esquecidos, que me levam a recordar algumas peripécias da minha vida de docente. Alguns dos meus exames mais originais andam pela *internet*, mas nunca lá vi o programa que entreguei aos alunos em 1996-97, e que reproduzo na página que se segue. Levei folhas para distribuir pelos alunos, e um acetato que projetei e comecei a comentar.
Só quando os sinais de estupefação eram gerais, e o silêncio um pouco consternado audível (se me posso exprimir assim) é que perguntei com ar inocente:

[8] Sim, em minha opinião a palavra acertada é "estatístico", e não "estaticista", e o argumento contrário de que o feminino seria então o nome da própria ciência, e mesmo de uma terminologia específica dentro dessa própria ciência, parece-me descabido. A menos que se passe a ter como norma que em vez de "matemático" se deve usar "matematicista", em vez de "físico" se deve usar "fisicista", etc.

— Há alguma coisa que não estejam a entender?

Na cadeira de Probabilidade e Aplicações (em que entreguei um programa semelhante), um aluno - atualmente um amigo, e durante largo período colaborador do meu projeto de investigação no CEAUL - comentou:

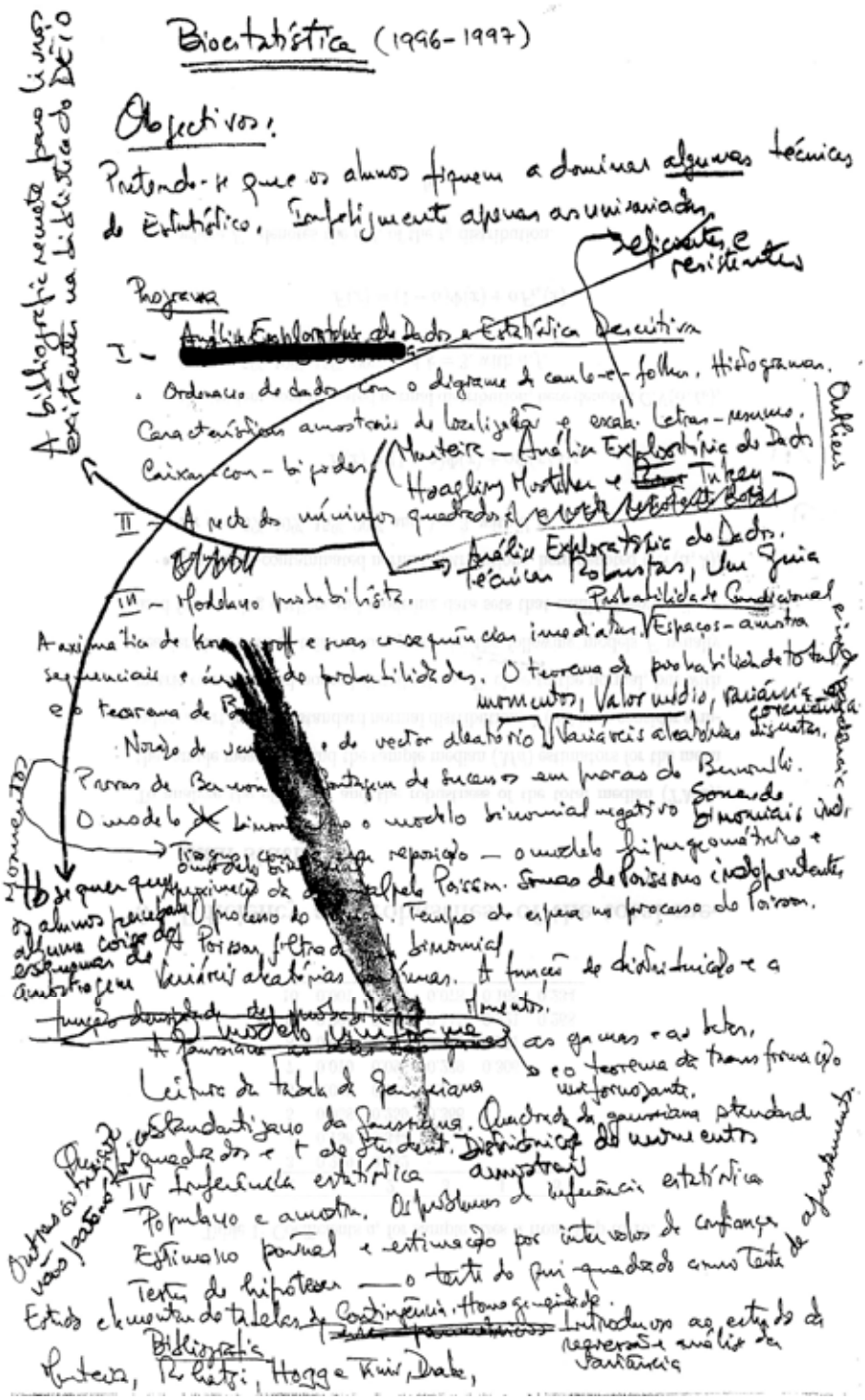
— Não, não, dificilmente poderia ser mais claro!

O que naturalmente provocou uma gargalhada geral.

Só nessa altura distribuí o programa a sério, dactilografado de forma arrumada, explicitando objetivos, bibliografia e modo de avaliação - enfim, o que se deve fazer - e disse:

— Pois, o que vos mostrei deve-se a ter estado a corrigir exames, e ter-me parecido que vos devia mostrar de uma forma visível como NÃO se deve apresentar seja o que for, incluindo um exame.

Nesse ano os exames, em geral, vinham mais cuidadosamente apresentados.



Estatística na Sociedade
-- uma digressão ilustrativa por domínios de aplicação[1]

Carlos Daniel Paulino, *dpaulino@math.ist.utl.pt*
IST, Universidade Técnica de Lisboa

Marília Antunes, *marília.antunes@fc.ul.pt*
FC, Universidade de Lisboa

Estatística e o seu papel político-cultural

A informação estatística proveniente de censos, inquéritos amostrais e outras fontes e o pensamento estatístico constituem conhecimento e meio indispensáveis à tomada de decisões sobre a estratégia e políticas de governação dos países e das suas instituições e empresas e à assunção de uma cidadania plena e responsável por parte dos seus habitantes.

A Estatística tem um papel decisivo, designadamente, no planeamento da recolha e na devida organização, tratamento e análise daquela informação. Inclusivamente, a Estatística dispõe de métodos para averiguar se na informação recolhida há suspeita de erros ou de dados forjados que, a existir, deve obrigar a uma investigação adicional. Isto é particularmente relevante em atividades de contabilidade e auditoria visando a deteção de eventuais anomalias e fraudes em documentos comerciais e financeiros.

Estatística nas Ciências da Vida

Biomedicina: campo de estudos na interface entre as ciências da *Biologia* e *Medicina* voltados para a investigação de doenças humanas no que concerne a causas, mecanismos, prevenção, diagnóstico e tratamento. Muitos desses estudos envolvem experiências que precisam de ser coerentemente delineadas e cujos dados exigem análises por metodologias estatísticas adequadas.

[1] Este trabalho foi financiado por Fundos Nacionais através da Fundação para a Ciência e a Tecnologia (FCT) no âmbito do projecto PEst-OE/MAT/UI0006/2011.

I. Estudo de potenciais fatores de risco da infeção no colo do útero pelos vírus do papiloma humano e de comportamentos de risco da infeção pelo vírus da imunodeficiência humana.

II. Estudo da diversidade do repertório dos recetores de linfócitos T (TCR), de que depende crucialmente o sistema imunitário adaptativo, e sua comparação entre os tipos CD4+ ativador e regulador da resposta imunitária (***Imunologia***).

Farmacometria: ramo da *Farmacologia*/*Biofarmacologia* Quantitativa, a que se dedica a indústria farmacêutica/biofarmacêutica envolvida em investigação, que se ocupa do desenvolvimento e aplicação de métodos estatísticos e matemáticos para caraterizar através de ensaios clínicos a *farmacocinética*, *farmacodinâmica* e resposta de pacientes de doenças biologicamente estudadas à administração de fármacos ou biofármacos (medicamentos produzidos por processos biotecnológicos).

Por outro lado, o desenvolvimento de uma nova droga requer um estudo de degradação da sua potência com o tempo decorrido após o fabrico. Para o efeito, escolhem-se de forma adequada várias amostras da droga e determina-se a degradação da sua qualidade (redução da sua potência original) ao fim de sucessivos intervalos de tempo, ao longo de períodos pré-especificados. A adoção de modelos estatísticos apropriados para tais dados permite estudar como varia a fiabilidade da droga (probabilidade de se manter suficientemente eficaz) com o tempo e estipular qual o período de validade a recomendar.

Biometria: campo de estudo de modelos e métodos estatísticos e matemáticos aplicados a problemas de análise de dados das *Biociências* (Ciências da Vida):

I. Análise da eficácia relativa de competidoras terapias de doenças através de ensaios clínicos humanos (*Biomedicina*, *Biofarmacologia*) -- como exemplo, refere-se a comparação de pacientes tratados a cancro do colon por uma de duas terapias atribuídas aleatoriamente com respeito ao tempo desde ingresso no ensaio até à ocorrência de recorrência da doença.

II. Comparação de rendimentos de diferentes variedades de um cereal, ou de produção de distintas raças de uma espécie de gado, ou de eficácia de técnicas de restauração dentária, através de ensaios agrícolas/animais/humanos previamente delineados (**Agropecuária, Medicina Veterinária, Agronomia, Zootecnia, Odontologia**);

III. Análise de efeitos de poluição atmosférica ou aquática na ocorrência de doenças nos residentes de uma dada região (*Ciência do Ambiente*, *Saúde Pública*);

IV. Estudo de efeitos da prevalência e intensidade de fogos florestais, de alterações climáticas ou de transformações humanas do solo na biodiversidade da região em análise (**Silvicultura**, *Biologia Vegetal/Animal*).

V. Estudos sobre a identificação de causas de doenças (***Genética***)

A. Os avanços científicos que permitiram a descodificação dos genomas das espécies trouxeram consigo uma mudança de paradigma na análise de dados de natureza biológica. No passado, a situação comum era a de observação de um número pequeno de variáveis sobre um número razoavelmente grande de indivíduos. As atuais tecnologias permitem a obtenção simultânea do nível de expressão de milhares de genes (as variáveis) mas, condicionados pelos elevados custos de produção, os investigadores têm, em regra, possibilidade de obter estes dados para um número reduzido de réplicas biológicas.

Dispondo de dados concretizados em matrizes de elevada dimensionalidade contendo níveis de expressão genética, é possível identificar estatisticamente genes com expressão diferenciada em casos de doença – genes que, como resultado da presença de doença, têm a sua função alterada ao ponto de se verificar um aumento ou diminuição anormal do nível de atividade.

B. Em muitos problemas os fenótipos de interesse exprimem-se de forma dicotómica, por exemplo através de suscetibilidade ou resistência a certas doenças do foro genético. A localização por etapas de regiões cromossómicas associadas com tais fenótipos é um assunto de interesse. Após uma primeira localização de cromossomas suspeitos, o recurso a um escrutínio mais refinado, denominado mapeamento genético por intervalos, usando distintos modelos estatísticos permitiu identificar mais precisamente os locais do genoma associados com a suscetibilidade à malária cerebral em cobaias.

Estatística em Epidemiologia e Saúde Pública

I. Estabelecimento de políticas de saúde pública

O registo de novos casos de doenças juntamente com a sua localização no espaço e no tempo permite um mapeamento espacial/temporal da taxa de incidência dessas doenças e a sua análise estatística pode estabelecer quais as ações a empreender no quadro de defesa da saúde pública.

Estatística em Ecologia e Ciências do Ambiente

I. Estudos de implicações na *Biogeografia* e *Biologia da conservação*

A. Dados: registo de presença/ausência ou de níveis ordinais de abundâncias de espécies em sítios amostrados num conjunto selecionado de zonas da região a analisar, bem como de valores de variáveis ambientais (e.g., temperatura, humidade, tipo de solo, etc.).

Objetivos: Análise através de modelos (espaço-temporais) apropriados para obter respostas a questões do género: Como é que populações de animais e/ou de plantas se distribuem espacialmente e/ou temporalmente? Áreas de alta biodiversidade tendem a causar menor abundância de espécies? Existem áreas ricas quer em biodiversidade quer em abundância, propiciando regiões ideais para a aplicação de esforços de conservação? Há evidência de duas ou mais espécies ocorrerem em áreas geograficamente sobrepostas (*simpatria*) ou isoladas uma da outra (*alopatria*)? Qual o efeito de transformações no terreno (para fins de agricultura, pecuária, silvicultura) ou de variações climáticas na distribuição da respetiva abundância e alterações do habitat?

B. Dados: registo de localizações com presença de dadas espécies no espaço, acompanhado de informação sobre fatores ambientais.

Objetivos: Ficar a saber como distintos fatores ambientais e tipos de solo incentivam ou desincentivam a presença de espécies ao longo de uma região; predizer a distribuição da ocorrência destas em regiões não amostradas com conhecidas condições ambientais e possibilitar o delineamento e a gestão de estratégias de conservação. A extensão em que se verifica sobreposição do habitat de várias espécies na região inspecionada constitui uma medida da alta ou baixa diversidade (grau de simpatria) ao longo dessa região, para além de poder evidenciar uma tendência para ausência ou presença de competição no consumo de recursos do solo ao ponto de suscitar a exclusão de umas por outras.

II. Estudo de monitorização ambiental com implicações em saúde pública e qualidade de vida em geral

C. Dados sobre níveis de poluição atmosférica em grandes meios urbanos/industriais

Objetivos: Análise estatística visando averiguar as suas causas e efeitos na saúde dos residentes e implementação de potenciais medidas atenuadoras ou reparadoras (e.g., restrição do tráfego automóvel, filtragem de agentes poluidores, deslocalização de fábricas, etc.).

III. Estudo de previsão de eventos hidrológicos mais ou menos raros (***Hidrologia***)

D. Dados sobre registos históricos de eventos com graves consequências sócio-económicas (cheias, desabamentos, secas) e medidas associadas (quantidade de precipitação, caudais de rios).

Objetivos: Estimar o intervalo médio de recorrência desses eventos e, com base em modelos estatísticos, avaliar parâmetros vitais para análise de risco que influencie tomada de decisões sobre a adoção de medidas preventivas e construção de infraestruturas com limites associados (diques, barragens, valas).

IV. Estudo de evolução de ecossistemas marinhos e atributos físicos oceânicos (**Oceanografia, Geografia Física – Climatologia**)

E. Dados sobre registos passados de temperatura e salinidade da água do mar, velocidade do vento e capturas de pescado.

Objetivos: Averiguar se reduções drásticas em estoques de certas espécies piscícolas têm que ver com alterações climáticas em adição a eventuais esforços pesqueiros exagerados, e se aquelas alterações se devem a fatores naturais ou humanos.

Estatística em Ciências Físicas e Químicas

I. Estudo de separação de sinais de fundo/fontes em imagens na ***Astrofísica***

Dados: Imagens astronómicas obtidas por telescópios consistindo de um fundo difuso (resultante de uma composição de emissões instrumentais e cósmicas com flutuações graduais ou abruptas da sua intensidade) no qual estão sobrepostos objetos celestiais de interesse (fontes) com morfologia e luminosidade grandemente variáveis (e.g., estrelas, galáxias em conglomerados ou não, nebulosas).

Objetivos: Deteção e caraterização de fontes indistintas de tipo pontual ou estendido com estimação do fundo através de modelos estatísticos que atendam ao ruído corruptor dos sinais presentes nas imagens e à informação *a priori* acumulada.

Este tipo de pesquisas é relevante para tentar responder a questões fundamentais da Astrofísica tais como a origem do universo e a distribuição da matéria no universo.

II. Estudos relacionados com a **Física e Química Nuclear**

As aplicações da ***Radioquímica*** (química dos material radioativos) são hoje extremamente variadas. Referem-se os setores da *Energia* com a produção de energia nuclear, das *Ciências Médicas* e

Farmacêuticas com o diagnóstico e terapia de doenças e da *Arqueologia* com a datação de restos de matéria orgânica pelo Carbono-14 (radioisótopo do Carbono-12).

Sendo a radioatividade um fenómeno tipicamente aleatório, compreende-se então a relevância do tratamento estatístico da desintegração de materiais radioativos, através nomeadamente de registo temporal de contagens de emissões de radiação por apropriados detetores.

III. Estudo de determinação indireta de composição química, do foro da **Quimiometria**

A Quimiometria é o campo da Química consistindo na aplicação de métodos estatísticos e matemáticos para o delineamento otimizado de experiências químicas e extração de informação relevante contida em dados dessas experiências.

Dados: Medição de quantidades obtidas instrumentalmente e relacionadas com a concentração de um dado composto químico na amostra em análise, bem como em várias outras amostras mas agora com composição conhecida.

Objetivos: Predizer a concentração do composto na amostra de interesse através da pesquisa de um modelo estatístico que se revele adequado no ajustamento dos dados obtidos.

Estatística em Engenharia

Em Engenharia é comum o delineamento de processos tendo em vista a avaliação de unidades experimentais (objetos físicos, formulações químicas, estruturas, materiais) quando sujeitas a algum tipo de intervenção (tratamento), especificado por uma ou mais variáveis controladas (fatores).

Os objetivos do planeamento experimental são a melhoria do desempenho e a redução da variabilidade dos processos, o desenvolvimento de novos processos e a redução do tempo e custo da operação global. O uso da Estatística no planeamento propriamente dito das experiências e na análise dos respetivos resultados permite uma maior eficiência e economia, bem como uma objetividade científica na extração das conclusões do processo experimental.

I. Investigação na resistência à compressão de betão (material composto de cimento, como aglomerante, de gravilha e areia como agregado e água) dos efeitos de distintos métodos de cura (processos de impedir a rápida evaporação da água na pasta de cimento), em vários espécimens de betão (**Eng. Civil**).

II. Estudo da evolução de danos estruturais (comprimento de rachas) provocados por fadiga do material em estudo quando sujeito a ensaios sucessivos de cargas cíclicas, tendo em vista a estimação da durabilidade residual medida pela fiabilidade, i.e., probabilidade de o tamanho da racha resistir à fratura como função do nº de ciclos de carga (**Eng. Mecânica**).

III. Averiguação do comportamento da resistência à tração de ligas metálicas (**Eng. Materiais**) ou de fibras sintéticas (**Eng. Têxtil**) em função da sua composição química ou percentagem de algodão, respetivamente.

IV. Estudo do efeito na luminosidade de válvulas integradas em televisores, do tipo de vidro e do fósforo nelas usados (**Eng. Eletrotécnica**).

V. Estudo do efeito no rendimento de um dado processo químico de fatores como o volume de solvente, a temperatura e o tempo de condensação e a quantidade de cada um dos dois materiais envolvidos no processo (**Eng. Química**).

VI. Estudo do volume de tráfego automóvel ao longo de dias úteis numa avenida de uma cidade para efeitos da correspondente regulação dos ciclos de semáforos nos cruzamentos com outras ruas (**Eng. Transportes**).

VII. Análise da distribuição espacial da qualidade de lodos de um rio, para vários níveis de profundidade, com base em amostras de sedimentos de distintas localizações, caraterizadas pelo seu grau de contaminação em diversos poluentes químicos e orgânicos (**Eng. de Georrecursos**). Tal análise por técnicas geoestatísticas reveste-se de importância para a definição de qual a estratégia de limpeza do leito do rio por dragagem em face do respetivo custo operacional.

O Controlo de Qualidade de produtos e serviços é um outro tópico crucialmente dependente da intervenção da Estatística, com enorme relevância em vários ramos da Engenharia e, em especial, da **Engenharia de Produção**.

A qualidade analisada estatisticamente é entendida no sentido de adequabilidade para uso, o que significa conformidade com as especificações requeridas. O Controlo Estatístico de Qualidade é então o conjunto de métodos estatísticos apropriado para medição, monitorização, controlo e melhoria da qualidade, tendo como objetivos fundamentais a redução da variabilidade dos resultados do processo, a eliminação de defeitos constatáveis e a otimização do seu desempenho. Tem como etapas importantes o Controlo Estatístico do processo em ação e o Planeamento Experimental exterior à produção visando a otimização ulterior do processo.

O Controlo Estatístico do Processo é assegurado por instrumentos estatísticos (diagramas) que permitam em tempo real detetar anomalias no processo, à medida que se vão analisando os resultados que vão saindo, e empreender as devidas ações corretivas para recolocar o processo sob controlo estatístico.

VIII. A título ilustrativo, suponha que uma empresa é contratada para produzir fios de cobre revestidos com um banho de prata para fins de aplicação em **Eletrónica**. O teor de prata nos fios é medido três vezes por dia durante um mês de trabalho. A análise ao longo do tempo das medições feitas através de diagramas apropriados permite constatar se o processo de aplicação do banho está ou não estatisticamente controlado.

Estatística em Ciências Sociais

As Ciências Sociais ao estudarem o comportamento humano em função do meio social com as suas relações de interdependência abrangem um largo leque de áreas como a Sociologia, Ciência Política, Psicologia, Demografia, Antropologia e Economia, entre outras, revelando ainda intercruzamentos com outros campos do conhecimento.

Não é por isso de estranhar que tal domínio possa fazer uso de metodologias múltiplas de análise que combinam técnicas qualitativas e quantitativas, em que nas últimas a Estatística tem sido chamada a intervir para dar resposta às questões solicitadas por vários estudos. Uma amostra destes é configurada pelos exemplos que se seguem:

I. Obter predições para as próximas eleições presidenciais a partir do registo histórico da percentagem de votos do candidato apoiado por um dado partido em cada um dos municípios do país, obtidos num conjunto de eleições passadas (**Ciência Política**).

II. Estudar o nível de cultura política dos cidadãos e sua relação com variáveis como o grau de instrução, tipo de emprego, nível económico, idade e género (**Sociologia**).

III. Caraterizar a privação, numa perspetiva multidimensional, dos agregados familiares de uma dada população, considerando múltiplos aspetos (capacidade económica, habitação, bens de conforto e formas de sociabilidade) determinantes do seu bem-estar e sua evolução ao longo de uma década (**Economia Social**).

IV. Analisar o desempenho de alunos em disciplinas fundamentais do ensino pré-universitário e o efeito nele do nível sócio-económico familiar e o tipo público ou privado da escola (**Ciências da Educação**, **Psicologia Social**).

V. Avaliar a capacidade de candidatos à admissão a um dado emprego através de respostas a questionários com múltiplos itens (**Psicometria**: ramo que estuda medições psicológicas – conhecimentos, competências, atitudes, crenças, inteligência - através de modelos e técnicas estatísticas e matemáticas).

Estatística em Finanças e Gestão Empresarial

I. Estudo do risco de mercado de um fundo de investimentos (**Econometria Financeira**)

Um dos objetivos em Finanças é a avaliação de risco de um ativo financeiro (ou carteira de ativos), risco este frequentemente medido em termos de variações de preços do ativo, como por exemplo os retornos.

A avaliação do risco de mercado de um ativo definido por um dado montante de milhares de euros em ações da empresa X, detido por um fundo financeiro Y, depende de dados sobre os preços diários passados dessas ações.

Supondo que se dispõe de uma série de retornos diários relacionados com os preços das ações da empresa, a sua análise estatística permite escolher uma gama de modelos com base nos quais se pode calcular e comparar os correspondentes valores em risco (o que se pode perder/ganhar) com uma dada probabilidade sobre horizontes temporais pré-especificados, e assim tomar decisões sobre a futura venda ou não ao longo desses períodos.

II. Estudo do campo da **Atuária**: análise de risco inerente à atividade seguradora e financeira (administração de seguros de vida/não vida e de fundos de pensões).

Análise estatística do montante total de indemnizações de cada tipo (danos corporais, danos materiais, danos próprios) -- função do número de sinistros e do montante envolvido em cada um destes – pago por uma dada companhia de seguros do ramo automóvel durante uma sequência de anos. O objetivo central deste estudo é o da predição para o ano seguinte do montante total de indemnizações para efeitos de planeamento da política estratégica a seguir.

III. Análise de decisão e planeamento em meio empresarial (**Gestão de informação**)

O conhecimento do comportamento dos clientes de uma empresa é fundamental para a definição de estratégias de negócio. Por isso, as empresas apostam cada vez mais na construção de bases de dados complexas, contendo não só informação relativa às transações efectuadas por cada cliente, mas também informação complementar sobre este de forma a conhecê-lo cada vez melhor.

Usando a informação constante dessas bases de dados, o recurso a adequadas técnícas estatísticas permite, nomeadamente, segmentar clientes de acordo com os seus perfis de consumo de bens ou serviços e alocar novos clientes aos grupos criados, bem como identificar potenciais necessidades/apetências/desejos dos clientes, potenciando o desenvolvimento de novas vertentes de negócio.

Estatística em Direito

I. Estudos com implicações em ***Genética Forense*** e ***Medicina Legal***.

Ainda que constituam domínios do conhecimento aparentemente afastados um do outro pelas matérias básicas que lhe são caraterísticas, a Estatística e o Direito compartilham interesses fundamentais comuns ao lidarem com interpretação de evidência, testagem de hipóteses e tomada de decisões em contextos de incerteza. Tal facto, juntamente com a experiência acumulada no recurso a periciais avaliações estatísticas em processos judiciais, explica o uso crescente em algumas sociedades da Estatística na produção de provas de identificação para julgamentos.

Dados: Informação sobre testemunhos e vestígios encontrados em casos criminais (e.g., autoria de assassínios) e perfis de ADN em casos civis (testagem de paternidade em disputa).

Objetivos: Uso da informação adequadamente quantificada e de corretos raciocínios probabilísticos para avaliação da chance de identificação do(s) criminoso(s) (do verdadeiro pai) entre o(s) suspeito(s) (entre o(s) candidato(s) em disputa), de modo a reunir evidência suplementar e evitar argumentos enganadores tais como o da falácia do acusador (que consiste em interpretar incorretamente juízos probabilísticos condicionais por inversão do respetivo condicionamento). Isto coloca um enorme desafio ao trabalho em harmonia de estatísticos e advogados pelas suas distintas formações de base.

Estatística em Literatura e Linguística Quantitativa

I. Estudos de identificação autoral

No estudo quantitativo da Literatura e da Linguística Aplicada os métodos e modelos estatísticos são de importância hoje em dia inquestionável. A ***Estilometria*** (estudo quantitativo do estilo linguístico) ao lidar com a análise de textos para fins de identificação autoral de documentos anónimos ou disputados é um bom exemplo da aplicação da Estatística à área das Humanidades. A análise de textos pode ainda ser motivada para averiguação de plágio e confirmação ou eliminação de suspeitos de crime, sendo então integrável na denominada *Estilística Forense*.

Dados: Registo de frequências de determinadas palavras ou sequências de letras em várias parcelas dos textos a analisar, obtido pelo uso de computadores. O tipo de sequências mais usado pelo seu papel mais discriminador é o das palavras gramaticais (artigos, pronomes, preposições, conjunções) nas quais a função sintática é mais relevante que a função semântica, em oposição às palavras lexicais

(substantivos, adjetivos, verbos, maioria dos advérbios). Outras variáveis possivelmente discriminativas para comparação de potenciais autorias ou de géneros literários (e.g., prosa versus poesia) são a extensão de palavras ou de frases e a proporção de pontuação (vírgulas, etc.).

Objetivos: caraterização do estilo literário de vários autores e atribuição de autoria em documentos por identificar, através de métodos estatísticos apropriados, hodiernamente aplicados através de software específico, em complemento a argumentos relacionados com as circunstâncias (pessoais, sociais, políticas, etc.).

Estatística no Desporto

A prática de virtualmente todas as provas desportivas acarreta a geração de números – de golos no futebol, de pontos no basquetebol, de tempos/distâncias em corridas/saltos de atletismo, pontos por júris em ginástica desportiva, etc.

A análise desses dados pode ser motivada por objetivos distintos que vão desde o simples entretenimento prazeroso de fãs de modalidades desportivas até à predição associada a interrogação sobre quais os limites de proezas atléticas humanas, passando pelo planeamento de melhores táticas de ação e estratégias de gestão, individuais ou coletivas, e pela necessidade de desmistificar conclusões empoladas pelos *media* (porque baseadas em pequenas amostras ou no manipular de níveis de desempenho extremos).

A Estatística tem efetivamente contribuído para

I. a definição de adequada estratégia de gestão de uma equipa com base no estudo do efeito do montante relativo despendido com salários e transferências de jogadores na classificação final conseguida na liga por cada uma de várias equipas analisadas durante um conjunto de anos;

II. a clarificação do efeito nos resultados das partidas de cada uma de várias equipas de expulsões ordenadas pelos árbitros e dos fatores de jogar como visitada versus visitante;

III. o delineamento de estratégias ótimas a adotar por cada atleta de pista, tendo em conta o seu perfil fisiológico, e a atualização de sistemas de pontuação em modalidades como o decatlo, heptatlo, triatlo, e corrida de corta-mato por equipas;

IV. evidenciar se as diferenças dos melhores desempenhos entre homens e mulheres nas provas de pista em atletismo têm vindo a esbater-se ao logo dos anos;

V. predizer a evolução de recordes por género com base em dados de recordes anteriores e de melhores marcas anuais e em modelos, com parâmetros de interpretação fisiológica e biomecânica, relacionando a distância e o tempo (ou velocidade) a percorrê-la.

• Artigos Científicos Publicados

Aleixo, S. M. and J. Leonel Rocha (2012), "Generalized Models from Beta(p,2) Densities with Strong Allee Effect: Dynamical Approach", *Journal of Computing and Information Technology*, Vol. **20**, No. 3, 201-207.

Araújo Santos, P. and Fraga Alves, M.I. (2012). A new class of independence tests for interval forecasts evaluation. *Computational Statistics and Data Analysis* Volume 56, issue 11, p. 3366-3380.

Beirlant, J., Caeiro, F. and Gomes, M.I. (2012). An overview and open research topics in the field of statistics of univariate extremes. *Revstat* 10:1, 1-31.

Brilhante, F., Gomes, M.I. and Pestana, D. (2012). Extensions of Verhulst Model in Population Dynamics and Extremes. *Chaotic Modelling and Simulation* (CMSIM) 4, 575-591.

Caiado, J., N. Crato e D. Peña (2012). Tests for comparing time series of unequal lengths. *Journal of Statistical Computation and Simulation*, **82**, 1715-1725.

Ferreira, M., Gomes, M.I. and Leiva, V. (2012). On an extreme value version of the Birnbaum-Saunders distribution. *Revstat* 10:2, 181-210.

Ferreira A., de Haan L. and Zhou C. (2012). Exceedance probability of the integral of a stochastic process. *Journal of Multivariate Analysis* 10:1, 241–257.

Figueiredo, A., Figueiredo, F., Monteiro, N.P. and Straume, O.R. (2012). Labor adjustments in privatized firms: a Statis approach. *Structural Change and Economic Dynamics* 23, 108-116.

Figueiredo, F., Gomes, M.I., Henriques-Rodrigues, L. and Miranda, C. (2012). A computational study of a quasi-PORT methodology for VaR based on second-order reduced-bias estimation. *J. Statist. Comput. and Simul.* 82:4, 587-602.

Garcia-Soidán, P., Menezes, R., (2012). "Estimation of the spatial distribution through the kernel indicator variogram". *Environmetrics*, vol **23**, issue 6, 535-548.

Gil, P. e Fernanda Figueiredo (2013). Firm size distribution under horizontal and vertical innovation. *Journal of Evolutionary Economics* **23(1),** pp. 129-161.

Gomes, M.I., Ferreira, M. and Leiva, V. (2012). The Extreme Value Birnbaum-Saunders Model, its Moments and an Application in Biometry. *Biometrical Letters* 49:2, 81-94.

Gomes, I., Fernanda Figueiredo e M. Manuela Neves (2012). Adaptive Estimator of Heavy Right Tails: Resampling-based Methods in Action. *Extremes* **15**, pp. 463-489.

Morais, M.C. e Pacheco, A. (2012). A note on the aging properties of the run length of Markov-type control charts. *Sequential Analysis* **31**, 88-98.

Moreira E. E., Mexia J.T., Pereira L.S. (2013). Assessing homogeneous regions relative to drought class transitions using an ANOVA-like inference. Application to Alentejo, Portugal. *Stochastic Environmental Research and Risk Assessmen*t, 27(1), 183-193.

Neves, M., Ivette Gomes, Fernanda Figueiredo e Dora Prata (2012). Modeling extreme events: Sample fraction adaptive choice in parameter estimation. *AIP Conference Proceedings*, **1479**, pp.1110-1113.

Papança, F. (2012). A Matemática, a Estatística e o Ensino nos Estabelecimentos de Formação de Oficiais do Exército Português no Período 1837-1926: Uma Caracterização. *Proelium VII (3) 75-86.*

Silva-Fortes, C., Amaral Turkman, M. A. e Sousa, L. (2012). Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups. *BMC Bioinformatics*, **13**:147.

Soares, M.O. and Canto e Castro, L. (2012). Continuous time simulation and discretized models for cost-effectiveness analysis. *Pharmacoeconomics* 30:12, 1101-1117.

Zuber, V., Duarte Silva, A.P. e Strimmer, K. (2012) A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. *BMC Bioinformatics*, **13**:284 .

• Teses de Mestrado

Título: *Extreme value theory: An application to sports*
Autor: Sérgio Luís Ganhão Vicente, *sergevicente1975@gmail.com*
Orientadora: Maria Isabel Fraga Alves

Título: *Estatística na Investigação Forense*
Autora: Manuela Chadreque, *manuelachadreque@gmail.com*
Orientador: Fernando Rosado

Título: *Os municípios e a qualidade de vida*
Autora: Fátima Gonçalves, *fcila2011@hotmail.com*
Orientadores: José R. Pires Manso e António F. de Matos

Título: *Estudo da Evolução do sector da Construção em Portugal recorrendo à metodologia STATIS*
Autora: Paula Cristina Marque Brás, *100485033@fep.up.pt*
Orientadoras: Fernanda Figueiredo e Adelaide Figueiredo

Título: *Metodologia STATIS Dual. Aplicação a dados sobre Infertilidade*
Autora: Ângela Marisa Nordeste Félix de Almeida, *100414039@fep.up.pt*
Orientadoras: Adelaide Figueiredo e Fernanda Figueiredo

Título: *Evolução de Alguns Indicadores de Desenvolvimento nos Países Subdesenvolvidos Africanos*
Autora: Maria da Luz da Silva Mendes, *maryluz.mendes@gmail.com*
Orientadoras: Adelaide Figueiredo e Fernanda Figueiredo

Título: *Previsões e Análise de Taxas de Juro em Economia*
Autora: Cristiana Andreia Gomes Vieira, *cristiana1710vieira@hotmail.com*
Orientadores: Raquel Menezes e Filipe Mena

Título: *Metodologias estatísticas aplicadas à relação entre Eventos Climáticos Extremos, Saúde e Desigualdades Socioeconómicas na Grande Área Metropolitana do Porto*
Autora: Alice Maria Salgado Gonçalves, *salgado.alice@gmail.com*
Orientadoras: Raquel Menezes, Susana Faria e Ana Monteiro

Título: *Identificação de Genes Associados com o Potencial Invasivo de Streptococcus pneumoniae*
Autora: Luísa Moreira Sêco, *luisa.seco@gmail.com*
Orientadores: Marília Antunes e Francisco Pinto

Título: *Modelo de Previsão de Churn em Clientes Empresariais com 1 a 5 Serviços*
Autor: Nuno Henrique Silva Monteiro, *nuno.monteiro@vodafone.com*
Orientadoras: Marília Antunes e Paula Figueiredo

• Capítulos de Livros

Morais, M.C., Okhrin, Y. e Schmid, W. (2012). Limit properties of EWMA charts for stationary processes. Em *Frontiers in Statistical Quality Control* 10, 69-84. (H.J. Lenz, P.Th. Wilrich e Schmid, W. (eds.). Physica-Verlag, Heidelberg.

Ramos, P.F., Morais, M.C., Pacheco, A. e Schmid, W. (2012). Assessing the impact of autocorrelation in misleading signals in simultaneous residual schemes for the process mean and variance: a stochastic ordering approach. *Frontiers in Statistical Quality Control* 10, 35-52. (H.J. Lenz, P.Th. Wilrich e W. Schmid (eds.). Physica-Verlag, Heidelberg.

• Livros

Título: *Probabilidades e Estatística*, Vol I
Autores: Bento Murteira e Marília Antunes
Ano: 2012. Edições Escolar Editora. ISBN: 978-9725-592-355-9

Título: *Probabilidades e Estatística*, Vol II
Autores: Bento Murteira e Marília Antunes
Ano: 2012. Edições Escolar Editora. ISBN: 978-972-592-359-7

Título: *Recent Developments in Modeling and Applications in Statistics*
Autores: Oliveira, P.E., Temido, M. G., Henriques, C. e Vichi, M.
Ano: 2013. Edições Springer. ISBN: 978-3-642-32419-2

Título: *Probabilidades - Princípios teóricos*
Autoras: Esmeralda Gonçalves e Nazaré Mendes Lopes
Ano: 2013 (2ª edição). Edições Escolar Editora. ISBN: 978-972-592-161-6

Título: *Números, Cirurgias e Nós de Gravata: 10 anos de Seminário Diagonal no IST*"
Editores: João Pedro Boavida, Rui Pedro Carpentier, Luís Cruz-Filipe, Pedro S. Gonçalves, Eloísa Grifo, David Henriques, Ana Rita Pires
Ano: 2012. Editora: IST Press. ISBN: 978-989-8481-22-1

• Teses de Doutoramento

Título: Uma aplicação da metodologia ROC na análise de dados de *microarrays*
Autora: Carina Silva Fortes, *carina.silva@estesl.ipl.pt*
Orientadoras: Maria Antónia Amaral Turkman e Lisete Sousa

Na minha tese foram desenvolvidos métodos estatísticos para dar resposta a problemas na genética. Um objetivo muito comum na análise de dados de *microarrays* é determinar que genes são diferencialmente expressos sob dois (ou mais) tipos de tecido ou sob amostras submetidas a diferentes condições experimentais. Sabe-se que as amostras biológicas são heterógeneas devido a vários fatores, como por exemplo, antecedentes genéticos e subtipos moleculares, os quais são, na maior parte das vezes, do desconhecimento do investigador. Por exemplo, em experiências que envolvam a classificação de tumores é importante que se identifiquem subtipos do cancro em investigação. Distribuições bimodais ou multimodais geralmente refletem a presença de misturas de subclasses. Consequentemente, pode haver genes que sendo diferencialmente expressos (DE) quando se tem em conta os diferentes subgrupos, não são identificados pelos métodos usualmente utilizados para selecionar genes DE. Neste trabalho propõe-se uma nova representação gráfica que não só permite identificar genes com regulação positiva e regulação negativa, mas também genes DE em subgrupos. Esta ferramenta baseia-se em duas medidas, nomeadamente na área abaixo da curva (AUC) *receiver operating characteristic* (ROC) e no coeficiente de sobreposição entre duas densidades (OVL). Para a estimação do OVL desenvolveu-se um algoritmo que permite obter uma estimativa não-paramétrica desse coeficiente e que se baseia em determinar a área de sobreposição de duas densidades estimadas pelo método do núcleo. Foi também desenvolvido um algoritmo para identificar distribuições bimodais ou multimodais estimadas pelo método do núcleo, permitindo deste modo identificar os genes com as características acima descritas. A metodologia aqui proposta foi implementada em linguagem R. Compararam-se os resultados com os resultados obtidos através de métodos usualmente aplicados na seleção de genes DE, usando-se dados simulados e duas bases de dados disponíveis publicamente. Os resultados indicam que a nova ferramenta, *Arrow plot*, apresenta um bom desempenho na seleção de genes com diferentes tipos de expressão diferencial, sendo flexível e útil na análise de perfis da expressão de genes em dados de *microarrays*.

Carina Fortes

Título: Estimating wildlife mortality at wind farms: accounting for carcass removal, imperfect detection and partial coverage

Autora: Regina Bispo, *rmcarita@fc.ul.pt*

Orientador: Dinis Pestana e Tiago A. Marques

A minha tese incidiu sobre a quantificação da mortalidade de aves e quirópteros em parques eólicos. Nos estudos de monitorização dos parques sabe-se que a mortalidade observada difere da mortalidade real fundamentalmente como consequência (1) da remoção de cadáveres, (2) da deteção imperfeita pelos observadores e (3) da prospeção parcial do parque.

Assumindo a probabilidade de encontrar um cadáver conhecida, o estimador de máxima verosimilhança do número de animais mortos presentes na região no dia da visita pode definir-se pela razão entre o número de cadáveres encontrados e a probabilidade de encontrar um cadáver. Neste contexto, a probabilidade de encontrar um cadáver pode estimar-se pelo produto entre a probabilidade de inclusão da área prospetada na amostra, a probabilidade de não ser removido, dado que se encontra na área prospetada e a probabilidade de deteção do cadáver, dado que se encontra na área prospetada e não foi removido.

A probabilidade de estar disponível para ser encontrado pode definir-se pela esperança matemática da probabilidade de permanência de um cadáver. A análise de sobrevivência é uma metodologia que se aplica sempre que interessa modelar o tempo até à ocorrência de um evento, aplicando-se pois, neste domínio. As opções de modelação incluem metodologias não-paramétricas, semi-paramétricas e paramétricas. Como ponto de partida para uma análise rigorosa, foram discutidos os métodos de discriminação entre modelos probabilísticos. A validação formal de um modelo probabilístico envolve frequentemente a realização de testes de ajustamento. Neste trabalho, foi feito um estudo por simulação da potência dos testes de ajustamento baseados nas estatísticas de Kolmogorov-Smirnov, Cramér-von Mises e Anderson-Darling, variando as distribuições sob as hipóteses nula e alternativa, a dimensão da amostra, o grau de censura e o nível de significância. Foi também efetuada a comparação das metodologias paramétricas e semi-paramétricas na modelação da função de sobrevivência e discutida a importância da seleção do modelo probabilístico no contexto da estimação da função de sobrevivência e mortalidade.

A estimação da probabilidade de deteção do cadáver é enquadrada no contexto formal da amostragem por distâncias. Na abordagem convencional assume-se que a distribuição espacial dos objetos em relação aos transetos é uniforme. Para responder à não verificação deste pressuposto no contexto em análise, consideraram-se as distribuições gama e log-normal.

A tese conclui com a apresentação de um estimador da mortalidade que integra no espaço e no tempo a mortalidade observada corrigida para a remoção de cadáveres, detetabilidade e cobertura parcial.

Regina Bispo

Título: Modelos Limite para a Fiabilidade de Grande Dimensão
Autora: Paula Cristina Martins dos Reis, $preis@est.ips.pt$
Orientadores: Luísa Canto e Castro Loura e José Caldeira Duarte

Na minha Tese de Doutoramento foram desenvolvidos modelos assintóticos (*ultimate*) e pré-assintóticos (*penultimate*) para a Função de Fiabilidade, $R_T(t) = 1 - F_T(t), t \in \mathbb{R}$, de sistemas de grande dimensão, tipo Paralelo-Série (**PS** - estrutura em paralelo com componentes distribuídas em série) e Série-Paralelo (**SP** - estrutura em série com componentes distribuídas em paralelo). A importância destes sistemas remota a um Teorema fulcral investigado na área de Fiabilidade por Barlow e Prochan (1975), no qual se estabelece que qualquer sistema de estrutura coerente, pode ser representado por um sistema paralelo-série ou série-paralelo. A abordagem que regeu esta tese e que motivou a sua realização, recorre a importantes resultados da teoria assintótica de valores extremos, nomeadamente à caracterização dos domínios de atracção para máximos e para mínimos da chamada Distribuição Generalizada de Valores Extremos (GEV), a menos de localização e escala, diferindo por isso da utilizada por outros autores em contextos análogos e por essa via, pretendeu constituir um suporte teórico alternativo no estudo do comportamento limite da função de fiabilidade para os sistemas anteriormente referidos.

Numa fase inicial, com o pressuposto dos tempos de vida das componentes serem i.i.d., foram identificadas sucessões normalizadoras, para as quais, mediante uma condição assintótica envolvendo o número de componentes em paralelo e em série, se tem a função de distribuição do tempo de vida de sistemas **PS** (ou **SP**), $F_T(t)$, a convergir para a lei limite Gumbel para máximos (ou para mínimos). Posteriormente, com base num trabalho de Gomes e De Haan (1999), admitiu-se que o número de componentes do sistema, *n*, é suficientemente grande e fixo, estipulando-se condições que permitiram garantir a existência de uma *sequência penultimate* de leis estáveis para extremos (cujo parâmetro de forma varia com *n*) e que constitui uma melhor aproximação para $F_T(t)$ relativamente à própria lei limite Gumbel (com uma velocidade de convergência uniforme válida em $\mathbb{R}$). Estes resultados teóricos foram complementados com um estudo de simulação.

A constatação de que grande parte dos sistemas reais carecem da hipótese de idêntica distribuição para os tempos de vida das componentes, levou por último, com base num artigo de Resnick (1975) para distribuições de caudas dominantes, à extensão dos resultados anteriores, quando se pretende, em particular, modelar a função fiabilidade de *m* paralelos sistemas **PS** não identicamente distribuídos**.**

Paula Reis

Edições SPE - Minicursos

Título: Modelos com Equações Estruturais
Autora: Maria de Fátima Salgueiro
Ano: 2012.

Título: Análise de Dados Longitudinais
Autoras: Maria Salomé Cabral e Maria Helena Gonçalves
Ano: 2011

Título: Uma Introdução à Estimação Não-Paramétrica da Densidade
Autor: Carlos Tenreiro
Ano: 2010

Título: Análise de Sobrevivência
Autoras: Cristina Rocha e Ana Luísa Papoila
Ano: 2009

Título: Análise de Dados Espaciais
Autoras: M. Lucília de Carvalho e Isabel C. Natário
Ano: 2008

Título: Introdução aos Métodos Estatísticos Robustos
Autores: Ana M. Pires e João A. Branco
Ano: 2007

Título: Outliers em Dados Estatísticos
Autor: Fernando Rosado
Ano: 2006

Título: Introdução às Equações Diferenciais Estocásticas e Aplicações
Autor: Carlos Braumann
Ano: 2005

Título: Uma Introdução à Análise de Clusters
Autor: João A. Branco
Ano: 2004

Título: Séries Temporais – Modelações lineares e não lineares
Autoras: Esmeralda Gonçalves e Nazaré Mendes Lopes
Ano: 2003 (2ª Edição em 2008)

Título: Modelos Heterocedásticos. Aplicações com o software Eviews
Autor: Daniel Muller
Ano: 2002

Título: Inferência sobre Localização e Escala
Autores: Fátima Brilhante, Dinis Pestana, José Rocha e Sílvio Velosa
Ano: 2001

Título: Modelos Lineares Generalizados – da teoria à prática
Autores: M. Antónia Amaral Turkman e Giovani Silva
Ano: 2000

Título: Controlo Estatístico de Qualidade
Autoras: M. Ivette Gomes e M. Isabel Barão
Ano: 1999

Título: Tópicos de Sondagens
Autor: Paulo Gomes
Ano: 1998

PRÉMIOS "ESTATÍSTICO JÚNIOR 2013"

Está aberto, até 25 de Maio de 2013, o concurso para atribuição de prémios **"Estatístico Júnior 2013"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2013"** é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística.

2. Os candidatos aos prémios **"Estatístico Júnior 2013"** devem ser alunos do 3.° Ciclo do Ensino Básico, do Ensino Secundário, dos Cursos de Educação e Formação (CEF), ou dos Cursos de Educação e Formação de Adultos (EFA), no ano lectivo 2012-2013.

3. As candidaturas podem ser individuais ou em **grupo com um máximo de 3 alunos**. Do grupo pode ainda fazer parte um professor, da categoria onde o prémio se insere, ao qual caberá o papel de orientador.

4. Os candidatos devem apresentar um trabalho cuja temática deve estar relacionada com a teoria da Probabilidade e/ou Estatística.

5. O trabalho deverá ser constituído por um texto escrito em Português com um máximo de 10 páginas A4 dactilografadas e um poster formato A2 que resuma os principais aspectos do trabalho. O trabalho (poster e texto escrito) deverá ser **enviado impresso em papel para efeitos da avaliação**.

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2013**" a sete trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.° Ciclo do Ensino Básico, aos três primeiros classificados de entre os trabalhos candidatos do Ensino Secundário e um primeiro classificado de entre os trabalhos candidatos dos Cursos CEF-EFA. Os prémios são constituídos por produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 600 euros, 300 euros e 200 euros, a atribuir, respectivamente, aos grupos cujos trabalhos sejam classificados em 1.°, 2.° e 3.° lugares, para as categorias Ensino Básico e Secundário, e 600 euros para a categoria Cursos CEF-EFA.

7. Ao professor orientador do trabalho classificado em 1º lugar, em cada categoria, é ainda atribuída uma anuidade grátis como sócio da SPE, ajudas de custo para participação na Sessão de Entrega do Prémio e produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 500 Euros.

8. Aos grupos proponentes dos trabalhos classificados em 1º lugar será também oferecida uma ampliação do correspondente poster que será exposto na Sessão de Entrega do Prémio.

9. O boletim de candidatura, acompanhado do trabalho concorrente, deverá ser dirigido ao Presidente da SPE para a morada abaixo indicada. O carimbo do correio validará a data de entrega.

Sociedade Portuguesa de Estatística – Bloco C6, Piso 4 – Campo Grande – 1749-016 Lisboa

O boletim de candidatura e este regulamento podem ser obtidos em
http://www.spestatistica.pt/ BoletimCandidaturaPEJ13.pdf
http://www.spestatistica.pt/ RegulamentoPEJ13.pdf

10. A admissibilidade e apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição e nomeação será da responsabilidade da Direcção da SPE.

11. O júri é soberano nas decisões, não havendo lugar a impugnação ou recurso.

12. A atribuição dos prémios **"Estatístico Júnior 2013"** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será realizada numa Sessão expressamente dedicada a essa entrega.

13. Os prémios **"Estatístico Júnior 2013"** poderão não ser atribuídos.

Apoio Porto Editora

PRÉMIO ESTATÍSTICO JÚNIOR 2013

INTERNATIONAL YEAR OF STATISTICS

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Candidaturas até 25 DE MAIO DE 2013

CONTACTOS

Sociedade Portuguesa de Estatística
Bloco C6, Piso 4 – Campo Grande
1749-016 Lisboa

Telef./Fax 21 750 01 20

www.spestatistica.pt

spe@fc.ul.pt

Com o apoio:

Porto Editora

INTERNATIONAL YEAR OF

STATISTICS

PARTICIPATING ORGANIZATION

Índice