

Boletim



**SOCIEDADE PORTUGUESA
DE ESTATÍSTICA**

Publicação semestral

outono de 2012



Métodos Estatísticos em Medicina

A Amostragem por Tempo - Local no Estudo de Populações de Dificil Acesso

Luzia Gonçalves 19

Modelos de Predição em Medicina: Algumas considerações

Ana Luísa Papoila 26

Sobre o ensino de Estatística em cursos de Medicina

Pedro Oliveira 36

Métodos Estatísticos para Modelar e Prever o Impacto das Epidemias de Gripe em Portugal

Maria Lucília Carvalho, Isabel Natário e Baltazar Nunes 40

Editorial	1
Mensagem do Presidente	4
Notícias	7
Pós-Doc.....	52
Ciência Estatística	
• Artigos Científicos Publicados	61
• Teses de Mestrado	61
• Teses de Doutoramento	62
• Capítulos de Livros	64
• Livros	64
Prémios “Estatístico Júnior 2012”	65
Edições SPE	68

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: 600 exemplares

Execução Gráfica e Impressão: Gráfica Sobreireense

Editor: Fernando Rosado, fernando.rosado@fc.ul.pt

Sociedade Portuguesa de Estatística desde 1980



STATISTICS2013.ORG

INTERNATIONAL YEAR OF STATISTICS

PARTICIPATING ORGANIZATION

Editorial

... à porta do Ano Internacional da Estatística...

O Ano Internacional da Estatística – “Statistics 2013” como é referenciado na página da internet <http://www.statistics2013.org/> onde é promovido – é uma celebração mundial que reconhece o valor das contribuições da Ciência Estatística. Através de uma combinação de energias das mais variadas instituições espalhadas pelo mundo, “Statistics 2013” deve promover a importância da Estatística à vasta comunidade científica, ao mundo dos negócios e dos governantes, aos utilizadores de dados, aos meios de comunicação social, aos decisores, aos empregadores, aos estudantes e ao público em geral. De facto, é uma proposta ambiciosa mas que nos habituámos a considerar ao longo da nossa ação no âmbito das atividades da SPE. Assim, podemos dizer que nada de novo nos trás este desafio “Statistics 2013” e que apenas nos solicita para que mais e melhor implementemos aqueles que são os principais objetivos da SPE e que, acima de tudo, aproveitemos esta união mundial de esforços para fazer brilhar ainda mais o sucesso da Ciência Estatística que para muitos portugueses tem sido desiderato de vida científica e profissional.

Para concretizar “Statistics 2013” é proposto que os seus apoiantes – para além da sua atividade normal – acrescentem iniciativas no sentido da maior sensibilização para a importância da Estatística. Uma grande lista de apoiantes, entre os quais a SPE, já pode ser consultada na referida página da internet. Temos de saber usar este meio para melhorar ainda mais a imagem da Ciência que nos une e que, cada vez com maior força, nos deve continuar a unir...

Na página web da SPE podemos ler: “A Sociedade Portuguesa de Estatística está a participar no Ano Internacional da Estatística 2013. A Sociedade Portuguesa de Estatística tem o orgulho de anunciar a sua participação no Ano Internacional da Estatística 2013, uma celebração mundial das contribuições da Ciência Estatística para o avanço da nossa sociedade global. Mais de 700 organizações – universidades, institutos de investigação, escolas, sociedades profissionais, agências governamentais e empresas – em cerca de 100 países estão se associando para celebrar e promover a importância da Ciência Estatística para a comunidade científica, empresas, governos, meios de comunicação, decisores, empregadores, estudantes e o público em geral. Durante este ano de celebração, a Sociedade Portuguesa de Estatística e as centenas de outras organizações participantes em todo o mundo irão:

- Aumentar a consciência pública sobre o poder e o impacto da Estatística sobre todos os aspetos da nossa sociedade;
- Fortalecer a estatística como uma profissão, especialmente entre os estudantes dos ensinos secundário e superior;
- Promover a criatividade e o desenvolvimento nas Ciências da Probabilidade e Estatística.

Não deixe de visitar o nosso site nos próximos meses para saber mais sobre como a Estatística ajuda a moldar a sua vida quotidiana”.

... em concórdia, união em torno do Ano Internacional da Estatística...

O Congresso SPE é, sem dúvida, a principal atividade da Sociedade Portuguesa de Estatística e é também aquela que mais intensamente cumpre os objetivos estatutários. De facto, logo no artigo primeiro os estatutos definem que a SPE deve “promover, cultivar e desenvolver, em Portugal, o estudo da Estatística, suas aplicações e Ciências afins” e para o conseguir procurará “realizar sessões científicas...”. Essa atividade é, decerto, aquela em que a nossa associação mais sente o pulsar e a vida

dos estatísticos portugueses. É nela que palpita o trabalho empenhado e desinteressado de cada equipa organizadora e dos seus colaboradores – também com muitos frutos e benefícios na vivência curricular dos seus elementos. É no Congresso SPE que se vive o convívio social e científico de partilha entre os sócios e destes com todos aqueles, nacionais e estrangeiros, que sempre neles participam. Acima de tudo, é em cada Congresso SPE que as mais variadas equipas de trabalho sentem a satisfação de que o seu esforço vale a pena dado o sucesso dessa realização. É, sobretudo, em torno do Congresso SPE que se sente a união (social e científica) de toda a comunidade que se deve desejar cada vez mais forte para, com maior facilidade, atingir aqueles objetivos estatutários. Se assim não fizermos mostramos fragilidade, e podemos perder o reconhecimento da grande comunidade da Ciência, que ao longo dos anos temos vindo a conquistar e se foi sedimentando em termos nacionais e também internacionais. Este é ponto fundamental que deve guiar a SPE. Todos aqueles que intensamente viveram os 20 Congressos já realizados decerto testemunham vivências intensas com muito trabalho e merecido sucesso.

Muitos colegas – sócios SPE, militantes desde há longa data – se têm manifestado ao longo do tempo sobre a importância dos Congressos e das Atas. De uma síntese excelente – no Prefácio das Atas do VI Congresso, publicadas no limiar deste século, com o muito apropriado e desafiante lema “**Afirmar a Estatística: Um desafio para o século XXI**” – permito-me repescar: “(...) Uma reflexão sobre a dimensão quantitativa e qualitativa da comunidade estatística portuguesa e o papel das Atas dos congressos SPE como expressão pública de parte da produção relevante dessa comunidade permitiu-nos reunir ideias que enformassem uma política editorial orientadora dos autores de potenciais textos e dos avaliadores do material submetido. (...) A experiência acumulada sugere-nos que este caminho deve ser prosseguido e aperfeiçoado para que não restem dúvidas sobre a seriedade profissional e científica desta publicação de produção estatística portuguesa na sua própria língua”.

Sobre as duas temáticas – Atas e Congressos – e já entrados no século XXI, é tempo para continuar a reflexão e garantir que o património até agora construído seja, cada vez mais, enriquecido, melhorado e aumentado. Os ideais (estatutários) da Sociedade Portuguesa de Estatística e dos seus fundadores devem nortear a nossa reflexão assim como a nossa ação.

As Atas são uma consequência dos Congressos e estes são excelentes causas. E, porque as Atas são constituintes do memorial SPE – e uma associação sem memória não tem futuro – devemos investir nos Congressos que as permitem criar.

Só assim defenderemos e afirmaremos a Estatística e a SPE!

No próximo ano diversas atividades SPE já estão programadas. Todas as que se realizarem – deve desejar-se – têm de cumprir no seu pleno o Ano Internacional da Estatística o mesmo é dizer: os objetivos estatutários da Sociedade Portuguesa de Estatística. Vamos ter oferta variada que permite várias opções de realização científica individual. Na escolha de cada estatístico deve ficar sempre em aberto a hipótese de poder optar pela maior participação com o maior benefício para o indivíduo e para a comunidade. Neste momento, sobre este assunto, é interessante inserir a temática de O Todo e as Partes. Sobre esta reflexão, aqui e agora, veio-me à memória um excelente seminário que, há cerca de 30 anos, foi apresentado no Departamento de Estatística da Faculdade de Ciências da Universidade de Lisboa, pelo professor Florencio Asenjo (autor de *El todo y las partes; estudios de ontologia formal*) a convite do professor Tiago de Oliveira – o fundador e principal impulsionador da nossa associação científica de estatísticos unidos em torno da Estatística. O referido livro – editado há 50 anos – e aquele seminário trazem à memória tempos de reflexão para aprofundamento e revitalização dos grandes objetivos SPE “revistos e ampliados” em ano mundial da Estatística.

E, já agora, a propósito de recordações, respigo, de um texto que escrevi no Prefácio do *Memorial da Sociedade Portuguesa de Estatística* (editado em 2005):

“Na dialética de uma recordação surge um entroncamento. Para viver é preciso esquecer; argumentam alguns grandes pensadores. Mas, dizem mal! defendem outros. A cultura da memória – a não confundir com culto do passado – é filha da informação e do discernimento e mãe da lucidez. Dela, decerto, também se vive!

Um “simbólico dado” continua como bandeira desta associação científica – no início com a sigla (5 pintas SPEIO ao alto e, desde 1991 “reequilibrado” com SPE.

Alea jacta est!

Os dados estão lançados! Mas, não apenas para registar os sucessos...

A SPE vale o que, como um todo, valerem os seus sócios. Decerto, a Sociedade Portuguesa de Estatística, onde cada sócio é parte interveniente e fundamental em algum momento, continuará cada vez com mais sucesso, na certeza de que este é uma estatística “somatório” de pequenos acasos.

Esteja cada um atento ao seu momento de ação!”

Este ano de 2013 é, talvez o momento certo, para usar todo o património ancestral da SPE no sentido de a projetar ainda mais no país e no estrangeiro. Em 2013, assim o devemos desejar, o todo (SPE) deve ser muito mais do que a soma das suas partes.

... com um enorme desafio, no Ano Internacional da Estatística...


Voltemos a “Statistics 2013”. 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. Neste sentido, extraordinariamente em 2013, siga a opinião de vários colegas que sugerem a edição de alguns números extra ao longo do ano. Desde logo, pensando em custos, esta proposta envolveria apenas edições eletrónicas – sobre temas eleitos – que seriam disponibilizadas no sítio da SPE. Esta iniciativa tem o evidente resultado de mostrar mais vida científica no momento certo, além de estimular alguns setores que possam estar menos sensibilizados para atividades deste tipo.

O Boletim SPE e o seu editor, com o maior empenho, estão disponíveis para (mais) essa tarefa. Em Ano Internacional da Estatística fica este novo desafio intrínseco à missão estatutária da SPE. O sucesso e a afirmação do Ano Internacional da Estatística como acontecimento inovador e propulsor da maior projeção nacional e internacional da Ciência que nos une, (também) depende destas iniciativas que se concretizam “apenas” com a vontade dos estatísticos.

Aumentemos a “equipa de voluntários” durante todo o próximo ano.

Aguardo todas as vossas sugestões e opiniões.



O tema central do próximo *Boletim* será *Estatística não - paramétrica*.

Mensagem do Presidente

Sessão de abertura do XX Congresso da SPE em 26/10/2012: Transcrição da Intervenção do Presidente da SPE

Boa tarde a todos

Quero começar por cumprimentar os meus colegas de Mesa e por saudar todos os presentes na sessão de abertura do XX Congresso da SPE, cuja realização foi posta em marcha pela Direção anterior da SPE e cuja concretização se deve fundamentalmente à Comissão Organizadora presidida pelo colega Pedro Duarte Silva, a quem a atual Direção nunca regateou qualquer apoio.

À Comissão Organizadora quero exprimir a nossa gratidão pelos seus esforços em levar a cabo esta iniciativa num contexto indutor de vicissitudes e dificuldades de vária ordem.

À Comissão Científica quero agradecer todo o apoio prestado às solicitações da Comissão Organizadora.

Às entidades patrocinadoras e apoiantes (e, em particular, à Universidade Católica e ao INE, representados nesta mesa pelo Professor Joaquim Azevedo e pela Dra. Helena Cordeiro, respetivamente), quero expressar os nossos agradecimentos pelo seu auxílio na concretização de uma iniciativa desta envergadura.

Aos oradores convidados quero tornar pública a nossa satisfação por terem acedido ao convite formulado pela Comissão Organizadora, prestigiando este congresso com as suas apresentações.

Por fim, aos participantes neste congresso quero dar as boas vindas e desejar vivamente que este encontro lhes seja frutuoso em termos do relacionamento científico e social que ele possibilita e promove.

A todos o nosso muito obrigado em nome da Direção da Sociedade Portuguesa de Estatística.

2013 - Ano Internacional da Estatística: como celebrá-lo?

Como a maioria dos sócios da Sociedade Portuguesa da Estatística (SPE) certamente saberá, 2013 foi instituído como Ano Internacional da Estatística (*Statistics2013*) por decisão acordada entre prestigiadas instâncias internacionais da área. Tal decisão visa projetar na sociedade global a importância da Estatística através dos esforços combinados das mais diversas instituições que em todo o mundo estão ligadas ao desenvolvimento, aplicação e disseminação da Estatística.

Os principais objetivos da celebração mundial *Statistics2013* são:

- Aumentar a tomada de consciência pública sobre o poder e o impacto da Estatística sobre todos os aspetos da nossa sociedade;
- Fortalecer a Estatística como uma profissão, especialmente entre os jovens;
- Promover a criatividade e o desenvolvimento nas ciências da Probabilidade e Estatística.

As vicissitudes ainda bem visíveis no reconhecimento social da Estatística no nosso país fazem com que assumam particular relevância os primeiros objetivos acima referidos. A SPE através da sua atual Direção aderiu desde logo sem hesitação e entusiasticamente a este movimento (<http://www.statistics2013.org/participants.cfm>), estabelecendo e preparando iniciativas próprias, propondo colaborações institucionais para a sua comemoração e incentivando muitos dos seus associados a avançarem com ações integráveis nos propósitos daquela mobilização mundial.

Como espécies de iniciativas persecutórias de tais objetivos destacamos:

- Encontros científicos de natureza diversa organizados de modo a atrair setores profissionais ainda divorciados das atividades da SPE;
- Oficinas temáticas de âmbito nacional ou internacional orientadas em sentidos cientificamente mais ou menos estritos;
- Projetos dirigidos a estudantes do ensino pré-universitário e seus professores que visem a melhoria da sua formação em ambientes mais lúdicos ou menos informais;
- Ciclos de palestras e de seminários voltados para a divulgação dos saberes e curiosidades estatísticas em distintos campos científicos ou para o público em geral;
- Minicursos em temas específicos de relevância para segmentos ainda demasiado afastados do convívio com especialistas e profissionais de Estatística;
- Cursos de formação de curta duração em assuntos de maior ou menor abrangência (*e.g.*, domínio de software e seu uso para análise estatística);
- Publicitação do Ano Internacional da Estatística através de diversos meios, incluindo a produção de materiais alusivos e a inserção do logótipo e do significado dessa celebração (ou da correspondente ligação) no sítio de cada instituição participante ou de cada ação a desenvolver.

Exemplos de iniciativas deste género com preparação já em curso incluem:

- ✓ O **1º Encontro Nacional de Biometria** em simultâneo com o **1º Encontro Luso-Galaico de Biometria** (página ainda em construção), a realizar em Braga de 14 a 16 de julho de 2013, promovidos pela SPE e SGAPEIO (Sociedade Galega para a Promoción da Estatística e Investigación de Operacións), com apoio já manifestado do nó Galiza da rede BioStatNet e do centro de investigação CEAUL. O programa científico prevê 5 palestrantes convidados e a parte portuguesa da Comissão Organizadora é constituída pelos colegas
Pedro Oliveira (Universidade do Porto) - Presidente
Giovani Silva (Universidade Técnica de Lisboa)
Luzia Gonçalves (Universidade Nova de Lisboa)
Denisa Mendonça (Universidade do Porto)
Cecília Azevedo (Universidade do Minho).

- ✓ **Workshop Extremes in Vimeiro Today** (<http://evt2013.weebly.com>), a realizar no Vimeiro de 8 a 11 de setembro de 2013, organizado pelo CEAUL e com apoio da SPE, em honra da nossa colega Ivette Gomes e em comemoração dos 30 anos do 1º encontro internacional de Valores Extremos. O programa prevê 5 palestrantes convidados e 4 organizadores de sessões convidadas e a Comissão Organizadora é constituída pelos colegas
Antónia Amaral Turkman (Universidade de Lisboa)
Isabel Fraga Alves (Universidade de Lisboa)
Manuela Neves (Universidade Técnica de Lisboa).
- ✓ **Edição 2013 do Projeto Radical Estatística** (www.radicalestatistica.net) promovido pela SPE.
- ✓ **Exposição Explorística** (<http://www.exploristica.com>): Inauguração e disponibilização promovida pela SPE e com apoio do programa Ciência Viva.

Outras atividades a cargo de instituições (*e.g.*, INE, CTT, CEAUL) e grupos de sócios, estão ainda em fase de programação, tanto quanto é do nosso conhecimento.

Para todos os que trabalham no âmbito da Estatística - docentes, investigadores, quadros técnicos e outros profissionais, integrados em escolas, universidades, unidades de investigação, institutos e serviços governamentais, empresas e sociedades científicas e profissionais -, é indubitavelmente incontroversa a importância de que se reveste a celebração em Portugal de 2013 como Ano Internacional da Estatística, através dos mais diversos eventos e atividades como forma de se contribuir decididamente para a definitiva projeção e consolidação da Estatística na sociedade portuguesa.

Nesse sentido, a Direção da SPE incentiva os seus sócios a que, através das suas instituições ou de associação em grupos regionais ou temáticos com interesses comuns, proponham e promovam iniciativas que dinamizem e intensifiquem o programa de celebrações do Ano Internacional da Estatística visando a concretização dos seus objetivos fundamentais. A SPE procurará apoiar tais iniciativas na medida das suas possibilidades, sempre que tal lhe for solicitado, e vai constituir uma comissão para efeitos de coordenação do conjunto de todas as ações por ela apoiadas.

12 outubro 2012

O Presidente da SPE

Carlos Daniel Paulino

Notícias

• XX Congresso SPE

Antiga, Mui Nobre, Sempre Leal e Invicta Cidade do Porto

O outono iniciou-se há pouco, despedindo o verão que, não raro, teima em prolongar-se por este mês de outubro. No horizonte despertavam umas nuvens carregadas que, não abdicando da sua maior arte, nos recordaram os perfumes da terra húmida mas agradecida.

Foi neste recorte de tempo, entre 26 e 29 de setembro de 2012, que decorreu o XX Congresso Anual da Sociedade Portuguesa de Estatística. Na *Cidade Invicta*, pela segunda vez, depois de 10 anos de passagem por outras paisagens nacionais.

Quem vem e atravessa o rio
Junto à serra do Pilar
vê um velho casario
que se estende até ao mar

O XX Congresso Anual da SPE teve como anfitriã a Faculdade de Economia e Gestão da Universidade Católica do Porto e por Comissão Organizadora Pedro Duarte Silva, António Andrade, Francisca Oliveira, Helena Correia, Luís Pina Rebelo e Manuela Maia, todos daquela Faculdade, e Pedro Campos da Faculdade de Economia da Universidade do Porto. Foi, pois, nas instalações da Universidade Católica Portuguesa que se reuniram muitos dos estatísticos nacionais para quatro dias de partilha de conhecimentos e experiências, coloridos por vários e agradáveis momentos de convívio, já habituais nos congressos da SPE.

O congresso contou com cerca de 200 participantes que puderam imbuir-se em variadíssimos temas e trabalhos de investigação, patentes em 86 comunicações orais e 62 posters, a acrescentar ao mini curso, que habitualmente inicia os trabalhos, e às cinco sessões plenárias.

O Curso foi o “aperitivo” que abriu o apetite para o resto do congresso, tendo decorrido na manhã e tarde de quarta-feira, dia 26 de setembro. A agradável exposição de Maria de Fátima Salgueiro sobre “Modelos com Equações Estruturais” prendeu a atenção dos participantes e poder-se-á dizer que o grau de atenção (ou interesse) suscitado na audiência é um “fator latente” tendo como indicador de medida as inúmeras “observáveis” intervenções dos participantes.

Na sempre formal e emotiva sessão de abertura do congresso, pouco depois das 17 horas do mesmo dia, foram dadas as boas vindas aos participantes, tendo contado com a intervenção de Carlos Daniel Paulino, presidente da SPE, Joaquim Azevedo, presidente da Universidade Católica do Porto, Helena Cordeiro, representante do INE e Pedro Duarte Silva, presidente da Comissão Organizadora do Congresso.

Seguiu-se a primeira conferência plenária, proferida por Herman van Dijk (University of Amsterdam) sobre o tema “Simulation-based Bayesian predictive procedures for the 21st Century: Three key research issues”. Esta primeira conferência continuou a abrir o apetite dos participantes, tanto mais que o orador conseguiu provar que quando se quer muito uma banana e apenas se dispõe de uma laranja, é possível transformar a laranja numa banana! O que um estatístico não é capaz de fazer! Ver para crer: “Cut and slice”.

E como o dia já ia longo e o apetite também – agora no sentido literal –, rumou-se para uma oportunidade de agradável convívio no Porto de Honra, o qual teve lugar no restaurante panorâmico da *Business School* da mesma Universidade, onde pudemos desfrutar de uma paisagem reconfortante protagonizada pela Foz do Douro. Acompanhando as conversas com regalos para saciar o apetite e

embarcados pela encantadora voz de Rita Santos Silva (para que se perceba que não se trata de uma “voz” qualquer, refira-se que estamos a falar da 4ª finalista da 1ª edição dos *ídolos* em 2003), vimos chegar ao fim o primeiro dia.

Quem te vê ao vir da ponte
és cascata, são-joanina
dirigida sobre um monte
no meio da neblina.

Novo dia. Às 9h voltámos ao trabalho. Durante a manhã tiveram lugar quatro sessões de comunicações orais que decorreram em paralelo, tendo como temas: Estatística Bayesiana, Processos Estocásticos, Análise Multivariada e Métodos não Paramétricos. A sessão de posters seguiu-se à pausa para café, onde estavam patentes cerca de 30 trabalhos. A manhã encerrou com a segunda conferência plenária proferida por Paulo Rodrigues (Banco de Portugal e Universidade Nova de Lisboa) que trouxe o tema “Testing for Predictability”.

Durante a tarde os congressistas puderam apreciar um passeio pedonal pela cidade do Porto, excelentemente guiado pelo conhecido e apreciado historiador e arqueólogo Joel Cleto (que teve inclusive uma solicitação externa ao congresso para dar um autógrafo). Que nos desculpe a organização do congresso e o Dr. Joel Cleto por não podermos dar nas linhas que se seguem uma verdadeira panorâmica da qualidade do passeio. Começámos na Praça da Liberdade junto à estátua de D. Pedro IV, onde Joel Cleto nos falou, de entre outras coisas, do cerco do Porto e do sofrimento e coragem dos Portuenses de então, que resistiram a um ano de cerco (1832-1833) ao lado de D. Pedro IV, contra seu irmão D. Miguel, aquando da guerra civil entre Liberais e Absolutistas. Ficámos a saber que a *Antiga, Mui Nobre e Sempre Leal Cidade do Porto*, foi então reconhecida por D. Pedro IV como *Invicta*, passando a ser a *Antiga, Mui Nobre, Sempre Leal e Invicta Cidade do Porto*. O reconhecimento de D. Pedro à cidade do Porto foi muito para além desta denominação, estando até hoje o seu coração guardado na Igreja da Lapa da cidade do Porto. D. Pedro IV deu também à cidade a honra de ter os segundos filhos do rei como Duques do Porto, associando assim o dragão (símbolo da casa de Bragança) à cidade do Porto.



Começou o passeio pedonal. Sempre dirigidos por Joel Cleto, ouvimo-lo contar inúmeras histórias da História do Porto e fazer o enquadramento histórico de praças, ruas e monumentos, com uma eloquência ímpar que só o seu talento consegue imprimir. Começámos pela Igreja dos Congregados, ou Igreja de Sto. António, e daí fomos apreciar a Estação de São Bento com os seus belos painéis pintados por Jorge Colaço. Seguimos para a Sé do Porto para “conhecer” São Vicente, São Pantaleão (antigo padroeiro da cidade do porto) e Nossa Senhora da Vandoma (padroeira atual da cidade do Porto) e conhecer também o que “se ouve” e se sabe sobre a forma como o valioso altar de prata da Sé resistiu ao saque dos Franceses em 1809. Saindo da Sé encontrámos o Paço Episcopal, o belíssimo e imponente edifício barroco do século XVIII que sobressai na paisagem a quem o “vê ao vir da ponte”. Apreciada uma bela paisagem sobre o rio Douro e sobre o Monte do Olival ou Monte da Vitória, a partir do Largo Dr. Pedro Vitorino, continuámos a descer até ao largo da Igreja e Colégio de São Lourenço de estilo renascentista (popularmente conhecido por Igreja dos Grilos ou Colégio dos Jesuítas). Seguindo pela Rua de Sant’Ana, onde se situava uma das Portas da cidade do Porto, embocámos na Rua dos Mercadores que, no Porto Medieval, era o eixo principal de ligação entre o centro mercantil da Ribeira e o burgo episcopal.



Uma curiosidade interessante, das muitas contadas por Joel Cleto: A Rua dos Mercadores, por ser uma das mais ricas da cidade, onde se concentravam as casas mais convidativas, assumia tal importância que o direito de aposentadoria detido pelos nobres (direito absoluto de ocupar uma residência particular para nela habitar temporariamente), restrito a três dias na cidade do Porto, não podia ser usufruído nesta rua. Chegámos à Praça da Ribeira e ao Muro dos Bacalhoeiros onde pudemos apreciar, mais uma vez, a paisagem que se estendia ao longo do rio, de onde se avistava, à esquerda, a Serra do Pilar e, mais à direita, as caves do Vinho do Porto. Antes de atravessar a Ponte D. Luís, parámos em frente a um painel ilustrativo da invasão Napoleónica em 1809, onde está representada a tragédia vivida então: após a entrada no Porto das tropas de Napoleão, os Portuenses tentaram fugir atravessando para o outro lado do rio, mas, sem se saber porquê, a ponte estava aberta no meio e, empurradas por outras que tentavam atravessar, as pessoas foram caindo ao rio, tendo havido muitas mortes. Já caía o sol (e doíam os pés) quando atravessámos a ponte D. Luís e fomos caminhando pela margem do rio, de novo desfrutando da paisagem, agora da margem sul do rio, sobre o Paço Episcopal e sobre o Morro da Sé. Parámos no *Mr Grill*, no cais de Gaia, para um saboroso descanso, saboroso também porque nos aguardava um lanche fantástico, vital para repor as energias.



Por ruelas e calçadas
da Ribeira até à Foz
por pedras sujas e gastas
e lampiões tristes e sós.

Sexta-feira é o dia de trabalhos mais preenchido, começando às 9h com mais uma sessão de comunicações em paralelo, subordinadas aos temas: Finanças e Econometria, Análise de Sobrevivência, Análise Multivariada, Estatística na Saúde, Amostragem e Tratamento de Inquéritos. Seguiu-se outra sessão de Posters e, pelas 12h, iniciou-se a terceira conferência plenária proferida por Pascal Ardilly (INSEE – Institut National de la Statistique et des Etudes Economiques – França) que nos falou de “Small area estimation: An appraisal about main techniques”. Depois do almoço seguiu-se mais uma sessão de comunicações orais em paralelo, desta vez encerrando os temas: Análise Multivariada, Estatística na Saúde, Modelação Paramétrica, Controlo de Qualidade e Modelação de Populações Naturais. A quarta sessão plenária teve lugar pelas 15h30m com Russel Alpizar Jara (Universidade de Évora) que proferiu sobre “Amostragem, Estimação de Parâmetros e Modelação em Populações Naturais”. Nova pausa para café para reforçar as energias, retomando-se os trabalhos pelas 16h45m com mais cinco sessões em paralelo de comunicações orais sobre Análise Multivariada, Estatística Espacial, Modelação Paramétrica, Modelação de Populações Naturais e Séries Temporais.

Pelas 17h45m teve lugar a Assembleia Geral Extraordinária da SPE que encerrou os trabalhos desse dia.

Depois do longo dia de trabalho, o merecido Jantar do Congresso compensou as mentes do esforço do dia. Deslocámo-nos de autocarro ao Hotel Infante de Sagres, onde fomos faustosamente recebidos. Antes do jantar, enquanto nos serviam os aperitivos, contámos com a descontração de mais um agradável convívio, animado pelo grupo AfrolatINE, de voluntários de danças latinas do INE, entre os quais se encontravam alguns dos nossos colegas conferencistas. O jantar foi servido numa sala de decoração clássica e majestosa (como todo o hotel!) e com uma elegância, quer no serviço quer nos sabores, que bem atestam o bom gosto de quem escolheu este sítio para nos receber! Mas ainda não tinha acabado. De volta ao pátio interior, continuaram as surpresas dedicadas pelo grupo do INE, cujo

“pé de dança” e energia nos encantaram e animaram a noite que já ia longa. Foi assim, em festa, que terminou o tão cansativo, mas bem animado, penúltimo dia do congresso.

E esse teu ar grave e sério
dum rosto e cantaria
que nos oculta o mistério
dessa luz bela e sombria

Sábado, último dia. Entre as nove e as dez horas decorreram cinco sessões paralelas de comunicações orais, desta vez dedicadas aos temas: Séries Temporais, Extremos, Variáveis Latentes, Aplicações e Bioestatística e Ecologia. Passava pouco das 10h quando começou a sessão “Radical Estatística”, uma iniciativa recente da SPE dedicada essencialmente a modelar o interesse dos alunos do secundário pela Ciência Estatística usando outros “parâmetros” que não os do ensino formal na sala de aula. A “Radical Estatística” foi abraçada pelos colegas Bruno de Sousa, Dulce Gomes, Regina Bispo, Elisa Duarte e Paulo Rodrigues, que se mostraram francamente satisfeitos com o êxito alcançado. Durante aproximadamente 30 minutos, Bruno de Sousa fez uma apresentação sumária do projeto, quer ao nível da génese quer do desenvolvimento. Usando as palavras do Bruno, “a Radical Estatística é um projeto promovido pela SPE, dirigido a todos os professores e alunos dos 10º e 11º anos, que visa divulgar a Estatística sensibilizando os alunos e professores para a sua importância, no esforço de compreender o mundo de incertezas em que vivemos. Mais de 60 alunos e professores de diferentes regiões de Portugal participaram nas três fases deste projeto: Competição Online, Campo Aventura e Ação de Formação para professores. As atividades da Radical Estatística foram concebidas de forma a serem facilmente incorporadas em sala de aula e complementares ao currículo de cada ano letivo”. Apresentado o Campo Aventura, ficámos com vontade de participar no próximo acampamento! Será que podemos? Seguiu-se a apresentação de outra iniciativa da SPE (em colaboração com a Ciência Viva) com o mesmo pendor. Pedro Campos apresentou a “Explorística - Aventuras na Estatística” com o mesmo entusiasmo do orador anterior. De acordo com as suas palavras “trata-se de uma exposição itinerante composta por vários módulos interativos com o objetivo de levar os fundamentos das Probabilidades e da Estatística às comunidades educativas (do 3º ciclo e secundário), transmitindo os conceitos de forma prática e experimental”. Se não vejamos em www.exploristica.com. Depois de uma visita guiada à exposição, chegou a última pausa para café, como todas as outras, acompanhada por uns maravilhosos bolos de manteiga! À Moda do Porto!?

Ver-te assim abandonada
Nesse timbre pardacento
Nesse teu jeito fechado
De quem mói um sentimento

Entre as 11 e as 12 horas decorreu a quinta conferência plenária proferida por Havard Rue (Norwegian University of Science and Technology), intitulada “The r-inla.org project: an overview”. Seguiu-se a já habitual entrega dos prémios *Estatístico Júnior*, pelas colegas Luísa Canto e Castro e Maria Eugénia Graça Martins. É sempre bom ver o entusiasmo dos jovens participantes, acompanhados por pais e professores. Foram todos convidados para um agradável almoço nos jardins da Universidade Católica, o qual, ao que sabemos, tendo juntado todos os colegas envolvidos neste projeto, foi mais um momento de convívio e partilha de emoções.

Por volta das 13h teve lugar a sessão de encerramento do congresso com as intervenções de Pedro Duarte Silva e Carlos Daniel Paulino. O presidente da Comissão organizadora local, agradeceu a participação de todos e dedicou umas palavras de apreço muito especial aos que com ele, trabalhando afincadamente, organizaram mais um congresso da SPE. Por fim, o presidente da SPE exprimiu os devidos agradecimentos à comissão organizadora e a todos os participantes. Carlos Daniel Paulino referiu ainda que para o próximo ano, Ano Internacional da Estatística, haverá uma enorme diversidade de eventos a que a SPE se associa ou organiza. Ficamos à espera!

E é sempre a primeira vez
Em cada regresso a casa
Rever-te nessa altivez
De milhafre ferido na asa

Três anos depois do Congresso de Sesimbra - SPE 2009 - onde tivemos o papel de “olheiras” porque estaríamos na organização do congresso do ano seguinte, em S. Pedro do Sul, termina, com este testemunho, esta nossa fatia de tempo dedicada à SPE. Já longe da “canseira” do SPE - 2010 e com a edição dos “Selected Papers” concluída, este SPE 2012 foi para nós um *Porto Sentido*, título da canção que Rui Veloso dedica ao Porto, cuja letra (de Carlos Tê) aqui introduzimos, neste nosso singelo tributo à SPE e à Cidade do Porto.

Carla Henriques e Maria da Graça Temido



• Workshops

Workshop em Estatística, Matemática e Computação e Workshop Luso-Polaco de Biometria

O encontro que reuniu o 6º Workshop em Estatística, Matemática e Computação e 3º Workshop Luso-Polaco de Biometria teve lugar na Universidade da Beira Interior, Covilhã, nos dias 3 e 4 de Julho de 2012, em homenagem ao Professor Dinis Pestana. Na cerimónia de abertura, inaugurada pelo ilustre Reitor da UBI, Professor João Queiroz, foi apresentada por Teresa Oliveira uma breve biografia do Professor Dinis Pestana, tendo sido destacada a sua obra, o seu prestígio e o seu valor humano, qualidades ainda enaltecidas nos discursos proferidos pelo Professor Pedro Oliveira, em representação da SPE e pela Professora Manuela Neves, em representação do CEAUL.

Estes encontros foram co-organizados pela Universidade Aberta, o CEAUL, a UBI, o CMUBI, o IPT e a Poznan University of Life Sciences. Os encontros foram presididos por Teresa Oliveira, Stanislaw Mejza e Dário Ferreira, tendo contado com a presença de 119 participantes, provenientes de 6 países. Estiveram representadas 11 Universidades Portuguesas, 9 Universidades estrangeiras e 16 outras instituições.

Foram proferidas 22 comunicações convidadas, num total de 85 comunicações. A primeira comunicação foi proferida pelo orientador de Doutoramento do Professor Dinis, o Professor Damodar Shanbag da Universidade de Sheffield. A última comunicação foi proferida pelo Professor Peter Cameron, da Queen Mary University of London, tendo a cerimónia de encerramento contado com a presença do ilustre Reitor da Universidade Aberta, Professor Paulo Dias.

Em nome da Comissão Organizadora, um grande Bem-Haja a todos os participantes, colegas e amigos.

Teresa Oliveira



• Joint Meeting of y-BIS and jSPE

A primeira conferência da jSPE ocorreu no Campus da FCT-UNL, Caparica, Almada, entre os dias 23 e 26 de Julho de 2012. Esta conferência foi co-organizada pela y-BIS, a secção de jovens estatísticos na International Society for Business and Industrial Statisticians e o seu nome oficial foi: **Joint Meeting of y-BIS and jSPE**. Mais informações sobre a jSPE podem ser encontradas em www.jSPE.pt.

Informação detalhada sobre o y-BIS está disponível em <https://sites.google.com/site/ybisisbis>.

O objetivo da Joint Meeting of y-BIS and jSPE foi juntar jovens estatísticos e profissionais da academia e da indústria de modo a oferecer oportunidades para que estes se conhecessem melhor e proporcionar um forum de partilha de experiências científicas e profissionais. A Joint Meeting of y-BIS and jSPE abrangeu muitas das áreas de investigação em estatística e muitas áreas de aplicação. Informação detalhada em <http://www.ybis-jspe.com/>.

Esta foi a primeira vez que uma conferência foi organizada pela jSPE e a adesão foi bastante expressiva refletindo a importância de uma iniciativa deste tipo. Estiveram presentes mais de 130 participantes, sendo que cerca de 60% trabalham em instituições Portuguesas. No entanto a exposição internacional foi enorme, tendo a conferência contado com a presença de participantes de 22 países! No total, foram apresentados mais de 105 trabalhos.

O programa científico incluiu sete keynote talks, 19 invited sessions, 10 contributed sessions e uma poster session. Gostaríamos de agradecer aos keynote speakers por partilharem a sua experiência e conhecimento connosco: Ronald Wasserstein (USA), Tapabrata Maiti (USA), Richard De Veaux (USA), João Cadete de Matos (Portugal), Agus Sudjianto (UK) and Fabrizio Ruggeri (Italy). Foi também um prazer poder contar com a presença oficial de algumas instituições e sociedades estatísticas nacionais e internacionais que culminou com a organização de invited sessions por parte das mesmas: International Association for Statistical Computing, European Network for Business and Industrial Statistics, Banco de Portugal, Instituto Nacional de Estatística, Knowledge Discovery and Bioinformatics do INESC-ID, International Association for Statistical Education, NEEUM da Minho University, e Associação Portuguesa de Classificação e Análise de Dados.

Um questionário foi delineado para analisar o grau de satisfação dos participantes em relação a diferentes aspetos da organização da conferência e para saber o que se poderá melhorar nas próximas edições. Foram obtidas um total de 21 respostas. A satisfação global com o programa científico obteve uma avaliação de 8.5 em 10. 66.7% dos participantes estabeleceram contatos profissionais durante a Joint Meeting of y-BIS and jSPE, e o nível de satisfação com as oportunidades de *networking* obteve uma média de 8.43 em 10.



O programa social incluiu três eventos que permitiram uma maior interação entre os participantes: *welcome reception*, *conference tour* e *conference dinner*. O primeiro evento social foi o cocktail de boas vindas que decorreu na praia de São João e incluiu bebidas e jantar (obteve uma média de 9.94 em 10 no questionário de satisfação). Para o passeio da conferência, na tarde do dia 24, houveram três alternativas: visita a Sintra de jeep, passeio de barco para ver golfinhos na costa da Arrábida, e snorkelling na Baía da Armação, próximo de Sesimbra. Esta atividade foi avaliada com uma média de 9.52 em 10.



O jantar da conferência foi no dia 25 de Julho, num restaurante com vista para o rio Tejo e Lisboa. Incluiu um concerto do grupo musical Almaplana (violino, guitarra acústica, voz e violoncelo). Os 21 respondentes avaliaram esta atividade com 9.75 em 10!

Foram atribuídos prémios para os dois melhores posters: *High-dimensional regression: a simulation study* (Shirin Shahriari) e *Nonparametric estimation of the tail-dependence coefficient* (Marta Ferreira). Além dos certificados, cada vencedor recebeu um prémio monetário de 250 Euro.

A avaliação do nível geral de satisfação com a Joint Meeting of y-BIS and jSPE obteve uma média de 8.95 em 10. Também gostaríamos de agradecer aos outros membros do Local Organizing Committee: Luís Ramos, Frederico Caeiro e Vanda Lourenço, e à secretária do Centro de Matemática e Aplicações da Universidade Nova de Lisboa, Vanda Martins, pelo excelente trabalho realizado que resultou numa média de 9.75 em 10 no que se refere a respostas sobre "Contact with the organizers" e de 9.57 em 10 para "Organization staff support".

Reações como esta: "Thank you very much for everything. I had a wonderful and impressive stay in Lisbon and have made a lot of friends. We had very nice chats about everything and exchanges of idea in many statistical problems", fazem-nos – o local organizing committee – esquecer todo o trabalho com a preparação da conferência e dão-nos – jSPE – o sentimento de que estamos a ir na direção correta.

Agradecemos a todos os participantes por proporcionarem um excelente programa científico e muita animação nos eventos sociais. Desejamos a todos, tudo de bom e esperamos encontrar-vos novamente em breve.

O relatório detalhado pode ser encontrado na página da conferência (<http://www.ybis-jspe.com/>) e na página da jSPE (www.jspe.pt).

Em nome do Local Organizing Committee e do International Scientific Program Committee da Joint Meeting of y-BIS and jSPE,

Paulo Canas Rodrigues
Filipe José Marques

• “Radical Estatística” para Jovens

A **Radical Estatística** é um projeto promovido pela Sociedade Portuguesa de Estatística (SPE) que, na sua 1ª edição, foi dirigido a todos os alunos dos 10º e 11º anos e que visa divulgar a Estatística através de diversas atividades onde, acima de tudo, se pretende salientar a importância do tratamento de dados reais no processo de aprendizagem, bem como sensibilizar alunos e professores para a importância da estatística no esforço de compreender o mundo de incertezas em que vivemos.

Desde o final de 2011 até princípios de Maio de 2012, diversas equipas dos 10º e 11º anos de várias Escolas Secundárias de Norte a Sul do país, acompanhadas dos respetivos professores de Matemática, participaram num jogo online - construído de raiz para o evento - constituído por oito níveis de dificuldade onde eram testados vários conhecimentos de Estatística.

As equipas vencedoras tiveram como desafio seguinte a participação num Campo Aventura, que decorreu de 11 a 13 de Maio de 2012. Durante este fim de semana, alunos e professores participaram em várias atividades radicais findas as quais os alunos tinham um desafio estatístico para solucionar.

Em complemento, o projeto incluiu ainda um Curso de Formação de 25h para os professores (acreditado pelo Conselho Científico-Pedagógico da Formação Contínua) intitulado “Radical Estatística: uma maneira diferente de ensinar estatística”. Pretendeu-se com este Curso transmitir aos formandos conteúdos e competências que lhes permitam melhorar e aprofundar os seus conhecimentos sobre Estatística e delinear estratégias de resolução de problemas de estatística visando sempre a melhoria da sua prática diária. Foi também um objetivo do Curso a promoção da utilização de novas tecnologias através da utilização do software R no apoio ao processo de ensino de aprendizagem.

Para mais informação sobre este projeto consulte www.radicalestatistica.net.

Comissão Organizadora:

Bruno de Sousa – CMDT-IHMT, Faculdade de Psicologia e Ciências de Educação, Universidade de Coimbra

Dulce Gomes - CIMA-UE, Departamento de Matemática, Universidade de Évora

Regina Bispo - CEAUL, ISPA-Instituto Universitário, Faculdade de Ciências - Universidade de Lisboa

Elisa Duarte - Doutoranda na Universidade de Santiago de Compostela

Paulo M. M. Rodrigues - Universidade Nova de Lisboa

Apoios:

Ciência Viva, Sociedade Portuguesa de Estatística, CMDT, IHMT – Universidade Nova de Lisboa, CIMA-UE, Universidade de Évora, CEAUL, CEMAT, Porto Editora, APM.



Bruno de Sousa
Dulce Gomes
Regina Duarte

• Explorística

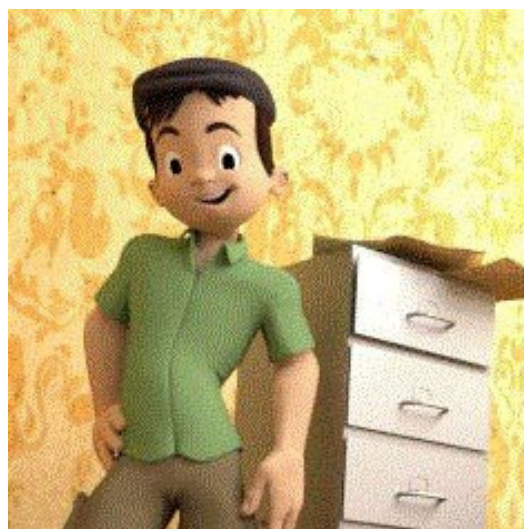


Já se imaginou no interior de um submarino a capturar animais raros e a pesá-los e medi-los, recolhendo informação estatística para posterior tratamento, como numa missão subaquática? E pensou alguma vez que poderia ser contratado por um político para entrar numa cidade e fazer uma sondagem acerca da sua notoriedade? E se um casino fizesse batota e lhe pedissem para verificar quais eram os dados viciados? Estas e outras aventuras estão disponíveis na Explorística!



A Explorística - Aventuras na Estatística - foi apresentada durante o XX Congresso da SPE. Trata-se de uma exposição itinerante concebida para ensinar de forma experimental os fundamentos da Estatística e das Probabilidades aos alunos de 3º ciclo e secundário. É uma iniciativa da SPE, apoiada pela Ciência Viva.

O anfitrião da exposição é o João Explorador, que apresenta a Explorística em cinco áreas: Seleccionar, Recolher, Descrever, Estimar e Confirmar. Cada uma das áreas é composta por módulos de materiais manipuláveis (seis, ao todo), que apresentam os seus conteúdos sob a forma de jogos e outras actividades interactivas.



A exposição

Dos seis módulos, o módulo de boas vindas define os grupos de trabalho e explica o funcionamento da exposição, recolhendo simultaneamente informação biométrica dos participantes (altura, tamanho do pé, etc.) para posterior tratamento estatístico. No Tiro com Arco, os jogadores fazem tiro com arco (virtual), usando-se as distâncias ao centro do alvo para se introduzirem os conceitos de localização e dispersão: go grupo que conseguir a menor distância média ao centro e em menos tempo. No Submarino os participantes realizam uma viagem por um ecossistema subaquático onde é preciso recolher uma amostra de uma nova espécie de réptil. Os participantes pesam, medem e identificam o sexo de cada um dos espécimens recolhidos, devolvendo-os depois de novo à água. Sabe-se que existem três subespécies deste réptil, com características diferentes, cujos elementos não se cruzam entre si. O objetivo do jogo é identificar a subespécie a que pertence a amostra recolhida. O

Quizz dos Censos é um jogo de perguntas e respostas sobre Informação dos Censos 2011, em jeito do “Quem quer ser milionário”. Com ajuda de uns pequenos computadores que servem para pesquisar dados em páginas com informação, ganha a equipa que conseguir responder mais depressa e acertadamente a um conjunto de perguntas. Nos dados ciciados os participantes descobrem qual ou quais os dados viciados com que o dono de um casino anda a enganar os seus clientes. E na Sondagem, o D. Stimacione pretende saber se tem possibilidade de ganhar as eleições na sua cidade. Com ajuda do grande gorila Conga, recolhe-se uma amostra não aleatória e outra aleatória e procura-se estimar a proporção de votantes em D. Stimacione.

Os jogos são complementados por imagens que refletem algumas aplicações práticas dos conceitos apresentados de forma lúdica: na indústria, na medicina, etc. Cada participante utiliza ainda uma folha, a fiolha do explorador, montada sobre uma tábua, onde recolhe as informações e vai escrevendo notas e conclusões, conforme vão sendo pedidas no decurso da exposição.

Onde estaremos?

A Explorística será inaugurada na Escola Tomaz Pelayo e seguirá depois para outras escolas. Qualquer escola portuguesa do ensino básico e secundária poderá albergar a Explorística, mediante a assinatura de um acordo que define o tempo e as salvaguarda dos respectivos conteúdos. O transporte dos módulos e a montagem da exposição ficarão a cargo da escola que alberga a exposição.

Para saber mais sobre como albergar a Explorística contacte-nos para a Sociedade Portuguesa de Estatística: s-estatistica@fc.ul.pt

Venha saber onde estamos para nos visitar. Ou então, requisite-nos!

Para saber mais, consulte: www.exploristica.com ou visite www.facebook.com/Exploristica.

Pedro Campos

• Prémio “Estatístico Júnior 2012”

A atribuição de prémios “Estatístico Júnior 2012” é promovida pela Sociedade Portuguesa de Estatística, com o apoio da Porto Editora, e tem como objetivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística. Neste ano foram recebidas candidaturas na categoria Ensino Básico, na categoria Ensino Secundário e na categoria Cursos de Educação e Formação de Adultos (EFA).

A cerimónia de entrega dos Prémios Estatístico Júnior 2012, conforme estipulado no Regulamento, decorreu na Sessão de Encerramento do XX Congresso Anual da Sociedade Portuguesa de Estatística, no dia 29 de setembro de 2012, nas instalações da Universidade Católica do Porto.

O Júri foi constituído pelos professores: Doutora Maria Eugénia Graça Martins (Presidente) e Doutora Luísa Canto e Castro de Loura do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa e Doutora Manuela Neves do Departamento de Matemática do Instituto Superior de Agronomia da Universidade Técnica de Lisboa.

No final deste Boletim são apresentados os premiados “Estatístico Júnior 2012”.

FR

• Prémio Savage - ISBA

O colega Carlos Daniel Paulino passou a integrar o comité do Prémio Savage 2012 da International Society for Bayesian Analysis (ISBA) por nomeação do ISBA Prize Committee. Esclarece-se que este prémio, nomeado em honra de Leonard Jimmie Savage, é outorgado em cada ano a duas notáveis dissertações doutorais em Econometria e Estatística Bayesiana, uma no campo de Teoria e Métodos e outra no de Metodologia Aplicada. Mais detalhes sobre este e outros prémios co-patrocinados pela ISBA podem ser consultados no seu sítio através do endereço <http://bayesian.org/awards/index>.

Giovani Silva

A Amostragem por Tempo - Local no Estudo de Populações de Difícil Acesso

Luzia Gonçalves, luziag@ihmt.unl.pt

*Unidade de Saúde Pública Internacional e Bioestatística,
Instituto de Higiene e Medicina Tropical,
Universidade Nova de Lisboa e CEAUL,*

Introdução

Frequentemente, na prática, os métodos clássicos de amostragem não podem ser aplicados por não existir uma base de amostragem acessível e fiável. Ultimamente, assiste-se à aplicação de novos métodos que combinam aspetos da amostragem não-probabilística com a probabilística, nomeadamente no estudo de populações ditas de difícil acesso. Numa perspetiva estatística, uma *população de difícil acesso (hard-to-reach population)* caracteriza-se pela não existência de uma base de amostragem. Numa perspetiva sociológica, corresponde a grupos de indivíduos ligados por aspetos culturais, socioeconómicos ou comportamentais, entre outros, frequentemente associados a algo considerado ilegal, podendo sofrer de estigma e de discriminação. Outros termos como *populações raras e populações ocultas* ou *escondidas* também aparecem com frequência na literatura (Kalton, 2001; Heckathorn, 1997). Nem todas as populações raras são de difícil acesso, pelo menos do ponto de vista estatístico. Nalgumas populações raras pode-se ter a lista de todos os indivíduos, por exemplo, no caso de algumas doenças raras. Heckathorn (1997) define população oculta (*hidden population*) com base em duas características: (i) a não existência de uma base de amostragem, desconhecendo-se o tamanho e os limites da população; (ii) a existência de problemas de privacidade pelos seus membros estarem associados a comportamentos ilegais e estigmatizantes, que os pode levar a não colaborarem no estudo ou a darem respostas não fiáveis de forma a preservar a sua privacidade. Analogamente, também as elites podem ser enquadradas nestes termos, redefinindo os conteúdos. No entanto, são numerosos os estudos que abordam questões de exclusão social e que consequentemente são aplicados em condições ambientais e sociais difíceis, como, no caso dos sem abrigo, dos imigrantes ilegais e dos refugiados. No âmbito da saúde pública, principalmente na vigilância do HIV/SIDA, é frequente a recolha de dados de comportamentos, de atitudes e práticas e também serológicos, em diversas populações ditas de maior risco ou vulneráveis como: profissionais do sexo, homens que fazem sexo com homens (HSH), utilizadores de drogas injetáveis, adolescentes e camionistas de longo curso (e.g., Raymond *et al*, 2007; MacKellar *et al*, 2007; Muhib *et al*, 2001).

A atualidade deste tema está patente na conferência *H2R 2012 – Survey Methods for Hard to Reach Populations* que se realiza de 31 de Outubro a 3 de Novembro do corrente ano, em Louisiana. Se nos países desenvolvidos encontramos diversas populações de difícil acesso, em ambientes mais desfavoráveis, quando efetuamos estudos em África, ainda se acrescentam mais dificuldades

operacionais e logísticas para efetuar amostragens credíveis. A intensificação da investigação em África também tem contribuído para uma maior atenção aos aspetos estatísticos, estando prevista a segunda edição da *International Conference on Applied Statistics for Development in Africa* (SADA'13) a realizar no Benin em 2013. Nos últimos anos, o foco no financiamento de projetos de investigação em HIV/SIDA, malária e tuberculose, também contribuiu para a exploração de métodos de amostragem adequados à realidade local das populações estudadas. Segundo Raymond *et al* (2007), “a escolha do melhor método de amostragem para uma determinada população alvo, assim como o conhecimento do contexto e das condições locais são fundamentais para um processo eficaz de vigilância comportamental do HIV”. A Organização Mundial de Saúde e os Centros de Controlo e Prevenção de Doenças dos Estados Unidos, entre outras instituições, têm proposto guias, cursos e conferências que abordam os desenvolvimentos da amostragem e a sua aplicação. Johnston *et al* (2010) destacam dois encontros em 2004 e 2009, respetivamente na Etiópia e Tailândia, sobre “*New Strategies for HIV/AIDS Surveillance in Resource-Constrained Countries*”.

Em África devido à rápida urbanização que tem alterado os padrões de doenças, os determinantes sociais da Saúde, incluindo as condições ambientais, assumem uma grande importância. A diversidade de morfologias, entre e nas cidades, comporta desigualdades socioeconómicas e culturais que podem afetar a saúde dos seus habitantes. Segundo Vlahov *et al* (2007), o “urbano” tem um carácter próprio e dinâmico em diferentes países que compromete o uso de estatísticas oficiais, sendo os estudos urbanos locais mais apropriados para compreender a saúde dos seus residentes. Os dados dos Inquéritos Demográficos e de Saúde, realizados para obter representatividade nacional e com intervalos regulares, geralmente de 5 anos, podem não ser suficientes para captar a realidade de zonas urbanas (WHO, 2008) e rurais de difícil acesso.

Assim, a problemática da amostragem e a qualidade da recolha de dados em estudos que envolvem comunidades rurais e urbanas de difícil acesso é fundamental pois os métodos de amostragem clássicos não são fáceis de implementar. Algumas estratégias alternativas têm sido desenvolvidas, assistindo-se a uma redução do uso de amostras por conveniência recolhidas usando amostragem por bola de neve. As estratégias de amostragem orientadas (*respondent-driven sampling* - RDS) e as que dão ênfase aos locais frequentados pelos indivíduos da população-alvo e não ao local de residência (e.g. *time – location sampling*) são frequentemente aplicadas na prática (e.g., Karon & Weijnert, 2012; Magnani *et al*, 2005). Neste trabalho, numa forma resumida, aborda-se este último tipo de amostragem. A intenção não é fornecer detalhes matemáticos, mas realçar a abrangência e a multidisciplinaridade subjacente à amostragem por tempo-local, referindo algumas desvantagens e potencialidades em termos de investigação.

Amostragem por tempo-local

A amostragem por tempo-local (*time-location sampling*) - ATL - também conhecida por outras designações: *time-space sampling* e *venue-based sampling*, tem sido difundida no âmbito de projetos de Saúde Pública, nomeadamente em estudos de HIV/SIDA em populações ditas de maior risco de infeção (Raymond *et al*, 2007). Além de esperar que um método de amostragem atinja os elementos da população-alvo, a ATL é também importante no levantamento do contexto e das condições locais, fornecendo informações úteis, por exemplo, para o planeamento, a monitorização e a avaliação de projetos de prevenção, contemplando assim simultaneamente aspetos estatísticos e de saúde pública da maior importância. A ATL exige uma equipa de trabalho experiente em ambientes difíceis e extremamente empenhada e atenta a tudo o que possa pôr em perigo a qualidade dos dados recolhidos. As questões de segurança e de ética são também extremamente delicadas.

A ATL pressupõe que a população de difícil acesso frequenta um universo de locais e/ou pontos de encontro em dias identificáveis e horários específicos. Os locais (e.g. bares, clubes, saunas, parques, ruas) poderão ser públicos ou privados, podendo haver necessidade de aprovações por parte de proprietários, gestores e/ou administradores.

A chamada *fase formativa* é fundamental para dar início à ATL, devendo a equipa de investigação do projeto obter mediadores junto das populações-alvo e efetuar uma exploração etnográfica exaustiva no terreno. Além da etnografia tradicional, a utilização de sistemas de informação geográfica e aparelhos GPS (*Global Positioning System*) é fundamental para o mapeamento no terreno, de forma a criar a lista de locais. Com mais ou menos variações, os locais podem ser considerados como as unidades primárias, sendo seleccionados aleatoriamente um conjunto de locais. As unidades secundárias resultam de associar aos locais, os dias e os horários de funcionamento, frequentemente, divididos em períodos de 4 horas (MacKellar *et al*, 2007; Barbosa Júnior *et al*, 2011). Nesta fase, para cada local seleccionado anteriormente, escolhem-se de forma aleatória os dias-horários, por exemplo, durante um calendário mensal. Note-se que de mês para mês pode haver algumas alterações devido, por exemplo, a feriados ou a encerramentos temporários. Dentro desses locais-dias-horários, os indivíduos são abordados sistematicamente para averiguar se cumprem ou não os critérios de elegibilidade e se consentem participar no estudo. Este esquema foi seguido por exemplo, por Stueve *et al* (2001) e De Boni *et al* (2012). Outras variantes consideram como unidades primárias o local-dia-horário, efectuando-se o processo em duas etapas (Johnston *et al*, 2010).

Vantagens, limitações e aspectos da implementação no terreno

A multidisciplinaridade da própria investigação e das suas equipas impõe, por si, a combinação de aspetos metodológicos quantitativos e qualitativos, sendo este tipo de amostragem integrador de várias vertentes, logo mais passível de aceitação na prática. Além das questões da inferência estatística, como já foi referido anteriormente, o processo de amostragem integra informações também relevantes para a atuação ao nível da saúde pública (ou noutros contextos). Segundo Johnston *et al* (2010) a vantagem deste método sobre a amostragem por grupos é que a população de interesse não está associada a uma localização específica, podendo contemplar movimentos entre diversas localizações, mesmo num único dia. A semelhança com a amostragem por grupos também é vantajosa para uma melhor caracterização teórica da ATL.

A questão que os estatísticos na prática enfrentam continuamente: “*A amostra será representativa?*” pode ter uma resposta indesejável pois é fácil que algo corra mal na prática. Em termos de limitações, entre outras, pode-se destacar: a própria escolha dos locais pode dar uma visibilidade não desejada, levando os potenciais participantes e proprietários dos locais a não participar ou autorizar a realização do estudo; a falta de privacidade de alguns locais pode comprometer a recolha de informação sensível; o facto de os locais seleccionados não serem frequentados por todos os indivíduos da população-alvo; a dificuldade em estimar o número de indivíduos com interesse, mas que não frequentam ou os que raramente frequentam os locais e mesmo a estimação do número dos que têm interesse. De facto, a enumeração das pessoas que poderão cumprir os critérios de elegibilidade da população-alvo e que frequentam os locais identificados pode ser um processo complicado. Raymond *et al* (2007) discutem algumas estratégias consoante o tipo de local (aberto ou fechado), a mobilidade e o tempo de permanência no local.

O uso frequente de horários com turnos de 4 horas nem sempre será adequado, devendo ser ajustado em função do número de entrevistadores e do tempo da entrevista e/ou recolha dos dados laboratoriais e mesmo das condições ambientais a que as equipas estão sujeitas. Por outro lado, deve também atender-se ao fluxo esperado de indivíduos aos locais em estudo. Porém, podem-se ter horários de menor duração para locais específicos. Por vezes escolhe-se o horário adaptado às necessidades administrativas da equipa do projeto, o que pode interferir com a representatividade.

Repare-se que os aspetos éticos, como a obtenção do consentimento informado, a confidencialidade, o anonimato podem tornar impraticável o estudo. A equipa de entrevistadores desempenha um papel muito importante neste processo e a sua seleção, atendendo às características da população a inquirir, a sua formação específica e o treino são fundamentais. Mendes (2011) descreve que num estudo de imigrantes no concelho de Oeiras, a equipa de entrevistadores teve uma participação desigual. De entre os 18 entrevistadores da equipa, 4 recolheram 60% do total de questionários. A antecipação de problemas e a proposta de soluções deve ser debatida amplamente de forma a uniformizar procedimentos entre entrevistadores. O relatório de atividades diárias, com particular enfoque nos problemas ocorridos, pode fornecer pistas importantes, não só para o estudo em curso como para a implementação de estudos futuros.

Se um estudo envolver a recolha de material biológico, as dificuldades ainda são acrescidas. Regra geral, a entrevista deve anteceder a recolha do material biológico (Raymond *et al*, 2007). Os cuidados com a recolha e armazenamento do material biológico podem ser delicados em determinados ambientes. Por exemplo, em ambiente urbano, algumas barreiras estruturais e de estacionamento podem condicionar a colocação de unidades móveis que podem ser vantajosas para a privacidade do participante e o armazenamento do material. Porém, a identificação ou algumas referências existentes nessas unidades móveis também podem ser dissuasoras da participação no estudo. Por outro lado, em zonas rurais pode haver outras barreiras em termos de acesso, impossibilitando a utilização de uma unidade móvel. Se mesmo assim, a realização de um teste rápido for viável transportando algum equipamento laboratorial, há que ter em conta que determinados materiais podem não suportar determinadas condições de temperaturas e/ou humidade.

Entre outros aspetos, é também necessário preservar o dinamismo e a sua segurança da equipa de entrevistadores e dos técnicos de saúde que podem estar expostos a condições ambientais e de segurança desfavoráveis. Em De Boni *et al* (2010) refere-se que membros da Polícia Federal ou da Brigada Militar de Porto Alegre, no Brasil, acompanharam à distância os entrevistadores, e obviamente à paisana para não inibirem os participantes, num estudo em 55 bares/turno considerados perigosos, principalmente pela alta criminalidade e tráfico de drogas.

A experiência permite afinar diversos aspetos metodológicos e os estudos que relatam as principais dificuldades e limitações de projetos de investigações anteriores podem fornecer contributos valiosos para melhorar a implementação da amostragem no terreno. No Brasil, Barbosa Júnior *et al* (2011) descrevem as dificuldades e as lições retiradas na implementação de 10 estudos-piloto que usaram amostragem por ATL e por RDS. Os custos da implementação podem ser elevados e nesse trabalho são igualmente descritos os custos por amostra nos 10 estudos analisados em 2006. Os valores descritos para os dois estudos que utilizaram ATL foram 21 400 dólares para uma amostra de 704 camionistas e 22 800 dólares para uma amostra de 718 HSH. A título comparativo, dos 8 estudos que recorreram à amostragem RDS, o valor mais elevado - 22 900 dólares - ocorreu numa amostra de 295 utilizadores de drogas. Com um sentido crítico à terminologia, Brackertz (2007) explora a ideia que ninguém na população é inacessível, o processo para aceder é que pode ser mais caro, exigindo esforços e criatividade para atingir esses grupos de indivíduos ditos de difícil acesso.

Aspetos com potencial de investigação

Relativamente ao cálculo do tamanho da amostra (n), ainda existe muito por fazer. No contexto da vigilância do HIV, aparece com alguma frequência a referência a $n=500$ (MacKellar *et al*, 2007; Raymond *et al*, 2007). Este número “fixo” merece ponderação, pois nem sempre será a dimensão mais adequada, podendo essa referência ser enganadora. A utilização frequente da fórmula para estimar uma proporção binomial como se tratasse da recolha de uma amostra aleatória simples, numa população finita ou infinita, também merece ser revista. De facto, nos estudos de vigilância predomina a estimação de prevalências, mas, pelo menos, em situações com enquadramento teórico mais definido, e aproveitando as semelhanças com a amostragem por grupos, pode-se calcular um tamanho amostral mais condizente com a realidade em estudo. Alguns autores sugerem considerar um efeito de delineamento de 2 (Raymond *et al*, 2007). Evidentemente, na prática, o tamanho da amostra final é influenciado por aspetos estatísticos, recursos humanos, financiamento, tempo e questões éticas.

Relativamente ao cálculo do tamanho da amostra e análise de dados, as informações recolhidas na fase formativa são também importantes para perceber que suposições poderão ser admitidas.

A análise estatística de dados obtidos por ATL, ainda é frequentemente efetuada como se tratasse de uma amostra aleatória simples (Karon e Wejnert, 2012). Note-se que um indivíduo pode visitar o mesmo local em turnos diferentes e pode visitar mais do que um local, com maior ou menor frequência. Em geral, os investigadores usam procedimentos para evitar selecionar um indivíduo mais do que uma vez (Karon e Wejnert, 2012). Os padrões de frequência dos visitantes dos locais são fundamentais para uma análise estatística que acomode diferentes probabilidades de amostragem e a possibilidade de agrupamentos de locais e/ou tempos amostrados, uma vez que alguns locais/tempos podem ser frequentados por indivíduos similares. Na prática, nem sempre é possível efetuar as contagens de indivíduos com rigor. Geralmente, o padrão de frequência dos locais é auto-relatado pelo participante e por vezes em escalas pouco apropriadas. Por exemplo, Karon (2005) apresenta um exemplo de um estudo com HSH realizado anteriormente, apresentando as estimativas para algumas prevalências de interesse e a exploração de associação através de modelos de regressão logística, sem e com ponderação e com e sem aglomerados. Nesse estudo, o padrão de frequência de bares e clubes noturnos foi registado na seguinte escala: nunca; menos de uma vez por mês; uma vez por mês; 2-3 vezes por mês; uma vez por semana; 2-3 vezes por semana e diariamente. Este tipo de escala não é o ideal para definir ponderadores com rigor, mas por vezes já fornece a indicação da necessidade de ajustes na análise estatística. Karon e Wejnert (2012) apresentam desenvolvimentos teóricos, mesmo partindo de algumas aproximações, dando ênfase à análise de dados obtidos por ATL em duas etapas. De Boni *et al* (2012) apresentam diferentes variantes da análise estatística para dados obtidos por ATL em 3 etapas.

No contexto da vigilância do HIV, Raymond *et al* (2007) referem que a análise estatística ajustada para a existência de conglomerados não tem sido muito usada porque se considera um número elevado de locais e o número de indivíduos por local é relativamente baixo, havendo heterogeneidade no local. No entanto, alertam para situações onde a ponderação e existência de aglomeração merecem ser contempladas. Raymond *et al* (2007) fazem a referência a programas como o Sudaan, o SAS e o STATA, interpretando o local como um grupo ou *cluster*. Assim, algumas especificidades da análise de dados por ATL estão ainda em aberto, quer na exploração teórica, quer na adequação dos comandos existentes nos pacotes estatísticos. Karon e Wejnert (2012) fornecem os códigos para o SAS e o R.

Notas finais

Apesar da ATL necessitar de aprofundamentos teóricos em diversos domínios parece ser uma temática atrativa. O elo de ligação com outras estratégias de amostragem ligadas ao estudo de populações animais e mesmo experimentação nas ciências agrárias pode contribuir para o enriquecimento da ATL e melhorar a sua aplicação no terreno. Na investigação biomédica, nota-se uma tentativa de evitar as amostras de conveniência. No entanto, as amostragens probabilísticas, como a descrita, podem também apresentar problemas que colocam em perigo a inferência estatística.

Sendo o HIV uma das infeções com mais atenção por parte de financiadores, assiste-se a um elevado número de projetos de investigação em diversas zonas do globo, contribuindo para uma expansão rápida deste método de amostragem. Porém, a exploração teórica deste método e a análise estatística dos dados recolhidos ainda necessitam de aprofundamentos diversos para uma maior credibilidade na prática.

Quer na investigação, quer no ensino, nomeadamente em ciclos de estudos (2º e 3º ciclos) ligados, entre outras áreas, à Saúde Internacional, à Saúde e ao Desenvolvimento, à Epidemiologia, à Medicina Tropical e às Doenças Sexualmente Transmitidas faz sentido nas Unidades Curricular de Bioestatística, além dos métodos probabilísticos clássicos, focar este (e outros) pois é provável que as teses e dissertações impliquem trabalhos de campo junto de comunidades de difícil acesso inseridas em contextos tropicais e/ou de pobreza.

Uma vez que no âmbito desta e de outras estratégias de amostragem para populações de difícil acesso, também existe um trabalho por realizar na uniformização e na tradução de termos do inglês para o português fica esta sugestão final que poderá contribuir para um enriquecimento do glossário da SPE/ABE.

Referências Bibliográficas

- Barbosa Júnior, A., Pascom, A.R.P., Szwarcwald, C.L., Kendall, C., McFarland, W. (2011). Transfer of sampling methods for studies on most-at-risk populations (MARPs) in Brazil. *Cad. Saúde Pública* [online], 27, Suppl.1, s36-s44.
- Brackertz, N. (2007). *Who is hard to reach and why?*. Working Paper, Melbourne, Institute for Social Research, Swinburne University of Technology.
- De Boni, R., Nascimento Silva, P.L., Bastos, F.I., Pechansky, F., Vasconcellos, M.T.L. (2012). Reaching the Hard-to-Reach: A Probability Sampling Method for Assessing Prevalence of Driving under the Influence after Drinking in Alcohol Outlets. *PLoS ONE* 7(4): e34104. doi:10.1371/journal.pone.0034104.
- De Boni, R., Vasconcellos, M.T.L., Holmer, B.P., Robin, R., Bastos, F.I., Pechansky, F. (2010). Beber e dirigir em uma amostra de condutores que frequentam bares de Porto Alegre. In Pechansky F, Duarte PCAV, De Boni R (Editores) *Uso de bebidas alcoólicas e outras drogas nas rodovias brasileiras e outros estudos*, Porto Alegre. Secretaria Nacional de Políticas sobre Drogas, p. 90-95.
- Johnston, L.G., Sabin, K., Prybylski, D. (2010). Update for sampling most-at-risk and hidden populations for HIV biological and behavioral surveillance. *JHASE*, 2,1:2.

- Heckathorn, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, Supp. 2, 174-199.
- Kalton, G. (2001). *Practical methods for sampling rare and mobile populations*. Proceeding of the Annual Meeting of the American Statistical Association, Rockville, MD.
- Karon, J. (2005). *The analysis of time location sampling study data*. ASA section on Survey Research Methods.
- Karon, J.M., Wejnert, C. (2012). Statistical Methods for the Analysis of Time-Location Sampling Data. *Journal of Urban Health*, DOI: 10.1007/s11524-012-9676-8.
- MacKellar, D., Gallagher, K., Finlayson, T., Sanchez, T., Lansky, A., Sullivan, P. (2007). Surveillance of HIV risk and prevention behaviors of men who have sex with men-a national application of venue-based, time-space sampling. *Public Health Reports*, 122, Supp. 1, 39-47.
- Magnani, R., Sabin, K., Saidel, T., Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*, 19, Suppl 2, S67-S72.
- Mendes, M. (2011) *Diagnóstico da população imigrante no concelho de Oeiras: desafios e potencialidades para o desenvolvimento local*. Alto Comissariado para a Imigração e Diálogo Intercultural. Lisboa. 298 p.
- Muhib, F.B., Lin, L.S., Stueve, A., Miller, R.L., Ford, W.L., Johnson, W.D., Smith, P.J. (2001). Community Intervention Trial for Youth Study Team. A venue-based method for sampling hard-to-reach populations. *Public Health Rep*, 116, Suppl 1, 216-22.
- Raymond, H.F., Ick, T., Grasso, M., Vaudrey, J., McFarland, W. (2007). *Resource Guide: Time Location Sampling (TLS)*, Behavioral Surveillance Unit, San Francisco Department of Public Health, 2ª Edição. [Existe versão em português: *Guia de recursos: Amostragem por Tempo-Local (TLS)*]
- Stueve, A., O'Donnell, L., Duran, R., San Doval, A., Blome, J. (2001). Time-space sampling in minority communities: Results with Latino young men who have sex with men. *American Journal of Public Health*, 91(6), 922-926.
- Vlahov, D., Freudenberg, N., Proietti, F., Ompad, D., Quinn, A., Nandi, V., Galea, S. (2007). Urban as a determinant of health. *Journal of Urban Health*, 84, 6-26
- WHO (2008) *Our cities, our health, our future. Acting on social determinants for health equity in urban settings*. Report to the WHO Commission on Social Determinants of Health from the Knowledge Network on Urban Settings.



Modelos de Predição em Medicina: Algumas considerações

Ana Luísa Papoila, *ana.papoila@fcm.unl.pt*

Faculdade de Ciências Médicas da Universidade Nova de Lisboa e CEAUL

1. Introdução

Devido à sua importância, os modelos de predição têm sido desenvolvidos em áreas como a Medicina, a Física, a Meteorologia e as Finanças. É, no entanto, em Medicina que estes modelos ganharam maior relevância e onde, nos últimos anos, a sua aplicação tem vindo a aumentar. De facto, áreas como a *saúde pública*, a *prática clínica* e a *investigação médica* têm beneficiado com a preciosa informação disponibilizada por estes modelos.

No contexto da primeira destas áreas, foram propostos vários modelos para predizer futuras ocorrências de doença em indivíduos assintomáticos; assim surgem, por exemplo, as funções de risco para a doença cardiovascular, desenvolvidas a partir de dados sobre fatores de risco para esta doença, obtidos pelo conhecido estudo de Framingham (*Framingham Heart Study*) (Wilson *et al.*, 1998). O benefício da aplicação destes modelos na Saúde Pública é evidente e advém da possibilidade de prevenir episódios graves da doença cardiovascular pelo facto da terapêutica com estatinas poder ser administrada aos indivíduos classificados como de alto risco. Muitos outros modelos de predição, úteis para esta área, foram desenvolvidos, nomeadamente na área da oncologia (Gail *et al.*, 1989; Gail 2001).

Ainda no âmbito da saúde pública, tornou-se importante prever a incidência de determinadas doenças considerando o impacto das diferenças geográficas e da evolução temporal. Assim, novas metodologias estatísticas foram propostas e estudos espaço-temporais foram realizados, tendo sido a área das doenças cancerígenas uma das que mais beneficiou com este tipo de estudos.

Na *prática clínica*, estes modelos também são úteis, particularmente na tomada de vários tipos de decisão após o diagnóstico ter sido efetuado. Baseadas nas estimativas de prognóstico, estas decisões podem incidir, por exemplo, sobre a necessidade de exames adicionais ou na forma como os tratamentos devem ser administrados (Scarvelis e Wells, 2006). Muitos modelos de prognóstico foram desenvolvidos em especialidades como a cirurgia, para estimar riscos a curto prazo de determinadas intervenções cirúrgicas com alto grau de dificuldade, efetuadas para reduzir riscos a longo prazo. Por

exemplo, em oncologia, se a curto prazo os riscos estimados de determinada cirurgia forem elevados, outras opções terapêuticas deverão ser consideradas tais como a quimioterapia, a radioterapia ou, em situações mais extremas, os cuidados paliativos. Outra grande vantagem destes modelos prende-se ainda com a informação acerca da evolução da doença, tão preciosa para o doente e familiares.

No que diz respeito à *investigação médica*, os modelos de predição são utilizados tanto em estudos experimentais como em estudos observacionais. Nos ensaios clínicos aleatorizados e controlados podem ser utilizados na seleção dos doentes (*e.g.* serem excluídos do estudo os doentes com probabilidades estimadas de morte aos 3 meses < 0.5) e, quando perdida a comparabilidade inicial dos grupos de tratamento e de controlo, para determinar a estimativa ajustada da eficácia do tratamento. De facto, quando no fim do estudo os dois grupos deixam de ser homogêneos em relação aos potenciais fatores de prognóstico utilizados no processo de aleatorização, é usual utilizar modelos de regressão para obter estimativas ajustadas. Em estudos observacionais, a utilidade dos modelos de predição surge, por exemplo, quando desejamos comparar a efetividade de dois fármacos. De facto, ao ser mais provável que indivíduos com determinadas características recebam um dos tratamentos, é impossível que os dois grupos de tratamento sejam, à partida, homogêneos em relação às variáveis que influenciam o *outcome*. Assim sendo, estas características surgem como potenciais fatores de confundimento, na medida em que se encontram simultaneamente associadas ao tratamento (exposição) recebido e ao *outcome* (efeito). Um problema adicional surge quando o efeito é representado por uma variável binária e consiste num evento raro; de facto, nesta situação, se o número de variáveis de confundimento for elevado, a modelação dos dados pode ser inviabilizada. Nesta situação é habitual recorrer ao uso dos *propensity scores* (Rosenbaum e Rubin, 1983; Braitman e Rosenbaum, 2002; Brookhart *et al.*, 2006; Cepeda *et al.*, 2003). Como é sabido, esta abordagem permite obter estimativas das probabilidades de exposição, tendo em conta a influência das variáveis de confundimento. Neste contexto é, pois, importante encontrar um modelo de predição que se ajuste aos dados de forma adequada para que faça sentido a substituição, se desejável, da variável de exposição pela estimativa da probabilidade de estar exposto. Após o seu cálculo, os *propensity scores* podem ser usados para emparelhamento, para estratificação ou ainda ser incluídos num modelo de regressão, em conjunto com o tratamento.

Como se pode constatar pelos exemplos acima referidos, a utilidade dos modelos de predição na área das Medicina é inquestionável.

2. Tipos de desenho epidemiológico

Antes de abordar os modelos que mais frequentemente se utilizam para predição em Medicina, vejamos quais os diferentes tipos de desenho epidemiológico que se encontram à disposição dos investigadores.

Assim, se o estudo é de **prognóstico**, os doentes são seguidos ao longo do tempo até à realização de determinado evento e o estudo designa-se de coorte (*cohort study*). No entanto, este tipo de estudo pode ser prospetivo ou retrospectivo. Neste último caso, é constituída uma coorte histórica em que os doentes são identificados a partir dos registos hospitalares do período em estudo. Estes doentes foram realmente acompanhados até à realização do evento de interesse mas o investigador recolhe a informação do passado. As desvantagens deste tipo de estudos (*e.g.* informação omissa ou incorretamente registada) são ultrapassadas com a realização de estudos prospetivos (*e.g.* ensaios clínicos controlados e aleatorizados). De facto, a existência de um protocolo ajuda à qualidade da informação a ser recolhida. Outro estudo prospetivo bastante eficiente, especialmente quando o

outcome é raro, é o de caso-controlo aninhado (*nested case-control study*). Este é frequentemente utilizado na avaliação de biomarcadores (Pepe *et al.*, 2008) que, naturalmente, é efetuada com base em modelos de predição.

No que diz respeito ao **diagnóstico**, embora na maioria das vezes os dados sejam resultantes de estudos de coorte, o estudo pode ser designado de transversal (*cross-sectional study*) na medida em que a relação entre as variáveis preditoras e o *outcome* é estudada em determinado instante. O diagnóstico (presença ou ausência da doença em estudo) deve ser realizado com base num teste de referência ("*gold*" *standard*) e, posteriormente, efetuados os estudos habituais do desempenho do modelo obtido, tanto no que diz respeito ao seu poder preditivo como discriminativo. Por vezes, são ainda utilizados estudos de caso-controlo (*case-control study*) uma vez que os doentes são seleccionados com base na presença ou ausência de doença, após a realização do teste de referência. No entanto, este tipo de estudo não é isento de problemas sendo o viés de seleção uma das suas principais limitações. Para mais detalhes sobre os vários tipos de desenho epidemiológico referidos consultar: Rothman e Greenland, 1998; Jewell, 2004; Clayton e Hills, 2009.

De uma forma não supervisionada, existem ainda outros estudos de predição que utilizam informação de bases de dados populacionais, criadas e atualizadas por registos nacionais como, por exemplo, no caso das doenças cancerígenas, os registos oncológicos regionais (*e.g.* em Portugal: ROR Sul, ROR Centro e RORENO). Estudos baseados neste tipo de dados são úteis, não só para modelar a incidência dos vários tipos de cancro tentando compreender diferenças regionais e temporais, mas também para efetuar projeções. De facto, o atraso nas notificações exigiu o desenvolvimento de modelos estatísticos que permitem prever a incidência de cancro em anos futuros, informação indispensável para a definição de políticas de saúde na área da oncologia.

3. Modelos de predição

Os modelos de predição podem pertencer a várias classes e, nesse sentido, serão abordados os modelos de regressão com uma referência especial aos métodos robustos, será feita uma referência às árvores e redes neuronais e, finalmente, de uma forma muito sucinta, serão tecidas algumas considerações sobre o papel dos métodos Bayesianos na predição em Medicina.

3.1. Modelos de regressão

Em relação aos modelos de regressão, como é sabido, a escolha do modelo adequado baseia-se não só no desenho do estudo mas também na escala de medida em que se encontra a variável resposta e, conseqüentemente, na respetiva distribuição.

Assim, em situações em que a referida variável é contínua, é usual utilizar o modelo de regressão linear (Kleinbaum *et al.*, 2008).

Quando a variável resposta é binária, a técnica estatística mais utilizada é o popular modelo de regressão logística (Hosmer e Lemeshow, 2000; Hilbe, 2009). É, de facto, um modelo muito importante na área da epidemiologia na medida em que permite resolver o problema da existência de variáveis de confundimento e de efeitos modificadores (interação), evitando assim procedimentos mais fastidiosos e limitados como os de Mantel-Haenszel (Jewell, 2004). Vários outros métodos estão disponíveis para resposta binária tais como o modelo de regressão *probit* e o modelo de regressão

complementar log-log. Os modelos logístico e *probit* são bastante semelhantes no que diz respeito à sua adaptabilidade dada a forma como foram definidos (Amaral-Turkman e Silva, 2000) e o modelo *complementar log-log* é habitualmente utilizado para análise de dados sobre prevalência e incidência de doenças (Martuzzi e Elliott, 1998; Penman e Johnson, 2009).

Quando a variável resposta é categórica, com mais de duas categorias, podem ainda utilizar-se os modelos de regressão logística ordinal (*e.g.* o modelo de possibilidades proporcionais) ou multinomial, consoante exista ou não uma relação de ordem entre os níveis da resposta. No que diz respeito ao primeiro destes modelos, existem alternativas quando a condição de aplicabilidade que exige a proporcionalidade das possibilidades é violada, tais como, entre outros, o modelo de regressão logística ordinal generalizado e o modelo de possibilidades proporcionais parciais. Existe *software* disponível em R para implementar estes, e outros modelos para respostas categóricas não aqui referidos (Yee, 2010). Dada a extensão do tema, aconselha-se ainda a consulta das seguintes referências: Paulino e Singer, 2006 e Agresti, 2007, 2010.

Pertencendo alguns dos modelos acima referidos à classe dos modelos lineares generalizados (McCullagh e Nelder, 1989; Amaral-Turkman e Silva, 2000; Dobson, 2002; Faraway, 2005) terão, como extensão, os modelos aditivos generalizados correspondentes (Hastie e Tibshirani, 1990; Wood, 2006). Estes são modelos mais flexíveis na medida em permitem modelar a influência das variáveis predictoras contínuas sobre o *outcome* através de funções suavizadoras, ultrapassando a obrigatoriedade de considerar a referida influência como linear. A utilização destes modelos permite, assim, contornar o problema da linearidade em *logit*, condição de aplicabilidade indispensável à utilização do modelo de regressão logística.

No caso em que pretendemos modelar dados de sobrevivência, é inevitável referir o modelo de regressão de Cox (Cox, 1972). No entanto, existem muitos outros modelos de regressão disponíveis em Análise de Sobrevivência. Sugerem-se as seguintes referências bibliográficas: Klein e Moeschberger, 1997; Therneau e Grambsch, 2000; Lawless, 2002; Collett, 2003; Rocha e Papoila, 2009; Carvalho *et al.*, 2011.

Não esquecendo que estudos de prognóstico são estudos longitudinais, é obrigatório fazer referência aos modelos que devem ser utilizados para analisar dados provenientes deste tipo de estudos. Dado o crescente número de estudos longitudinais (em Medicina e não só), o progresso observado no desenvolvimento de novas metodologias que têm em consideração a estrutura de correlação originada pelo facto de existirem várias observações do mesmo indivíduo obtidas ao longo do tempo de seguimento, tem sido substancial. Sobre este assunto, a literatura é bastante vasta (Verbeke e Molenberghs, 2000; Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2004; Twisk, 2005; Molenberghs e Verbeke, 2005; Fitzmaurice *et al.*, 2009; Cabral e Gonçalves, 2011).

Sobre métodos de regressão, podem ainda ser consultadas as seguintes referências: Vittinghoff *et al.*, 2005 e Harrell, 2010.

Finalmente, é importante recordar que o desempenho dos modelos de regressão depende da verificação de determinadas condições de aplicabilidade; de facto, a fiabilidade dos resultados obtidos pode ser questionada em situações de violação das condições requeridas. Existem várias soluções para contornar este problema tais como transformações da variável dependente e/ou de algumas (ou todas) das variáveis independentes contínuas e/ou a utilização de métodos alternativos. Entre estes, salientam-se os métodos robustos de regressão (Pires e Branco, 2007). Muito eficazes perante a existência de observações extremas (Rosado, 2006), estes modelos, entre outras vantagens, permitem manter na análise estas observações discordantes, evitando assim o procedimento radical que consiste na sua

eliminação da análise. Sendo a Medicina uma área em que abundam observações atípicas genuínas e sendo o tema sobre "predição em Medicina", é óbvio que a modelação estatística tem que ser cuidada por forma a garantir uma boa capacidade preditiva e discriminativa do modelo final.

3.2. Árvores e redes neuronais

Os métodos de partição recursiva ou árvores de classificação (*outcome* categórico) e regressão (*outcome* contínuo) (CART) (Breiman *et al.*, 1984; Hastie *et al.*, 2001), não tão populares quanto os modelos de regressão já referidos, são também utilizados para fazer predições na área da Medicina (Aguiar *et al.*, 2012; Austin *et al.*, 2012). É aconselhável a sua utilização quando se suspeita da existência de múltiplas interações de ordem superior (envolvendo mais do que duas variáveis). De facto, a estrutura recursiva dos modelos CART permite revelar interações complexas entre os preditores, contrariamente aos modelos de regressão clássicos. Existe *software* disponível em R (bibliotecas *tree* e *rpart*).

Constituindo uma nova abordagem na predição e situando-se entre os modelos aditivos generalizados e as árvores de classificação, surgem os modelos MARS (Multivariate Adaptive Regression Splines) (Friedman, 1991). Muito flexíveis, estes modelos são utilizados em diversas áreas de investigação no âmbito da predição e, em particular, em Medicina (Finnerty *et al.*, 2012; Lee *et al.*, 2012). Também já existe *software* disponível em R (biblioteca *earth*).

Quando se fala em predição é inevitável abordar o tema das redes neuronais artificiais (Bishop, 1995; Ripley, 1996). Envolvidas nalgumas controvérsias no que diz respeito ao seu desempenho, resta provar a sua superioridade em relação aos clássicos modelos de regressão, em particular aos modelos aditivos generalizados (Papoila *et al.*, 2012). Embora classificadas como "caixas negras", a sua aplicação para efeitos de predição é uma realidade (Grumett *et al.*, 2003; Pappada *et al.*, 2008). Existe *software* disponível em R (bibliotecas *nnet*, *neuralnet*, *AMORE* e *RSNNS*) (Günther e Fritsch, 2010; Bergmeir e Benitez, 2012).

3.3. Métodos Bayesianos

Finalmente, e dado o bom desempenho comprovado pelos métodos Bayesianos (Paulino *et al.*, 2003) em várias áreas de investigação, não poderíamos deixar de referir o seu papel na predição em Medicina. A aplicação dos métodos Bayesianos nesta área tem vindo a aumentar ao longo dos anos (Xian-Hong, 2006; McCormick *et al.*, 2012) com especial enfoque nos estudos espaciais e espaço-temporais (Carvalho e Natário, 2008; Lawson, 2009; Fromont *et al.*, 2010). De facto, os modelos hierárquicos Bayesianos têm-se mostrado indispensáveis na modelação e projecção da incidência de várias doenças, tendo em conta a variabilidade regional e ainda a evolução temporal (Cressie, 1993; Banerjee *et al.*, 2004; Gelfand *et al.*, 2010).

Como é sabido, têm sido as técnicas MCMC (*Markov chain Monte Carlo*) as mais utilizadas na inferência Bayesiana, no entanto, o tempo computacional dispendido é demasiado elevado, particularmente quando utilizados em modelos mais complexos como são, por exemplo, os de mapeamento espaço-temporal de doenças. Este problema motivou alguma investigação nesta área que culminou com a proposta de Rue e Chopin (2009). A metodologia desenvolvida por estes autores (INLA - Integrated Nested Laplace Approximation) constitui uma nova abordagem à inferência dos modelos GMRF (latent Gaussian Markov random field) (Rue e Held, 2005; Rue e Martino, 2007) com

custos computacionais bastante inferiores e sem perda substancial na qualidade das estimativas das distribuições *a posteriori* dos parâmetros. Não tendo a flexibilidade do *WinBUGS* e do *GeoBUGS*, o *software* disponível (e a própria metodologia), encontra-se ainda em desenvolvimento e constitui uma alternativa bastante atrativa, particularmente para os apenas utilizadores dos métodos Bayesianos. O código do INLA pode ser obtido em <http://www.math.ntnu.no/~hrue/inla>.

4. Avaliação do desempenho dos modelos de predição

Depois de aplicar um determinado método de predição, necessitamos de avaliar o seu desempenho tendo em conta a distância entre as estimativas do *outcome* e o seu verdadeiro valor. Esta avaliação pode ser efetuada de uma forma global ou, no caso de respostas binárias, ir mais além e tentar quantificar não só a capacidade preditiva através de medidas de calibração (concordância entre os valores observados do *outcome* e as estimativas) mas também o poder discriminativo (capacidade de distinguir entre os que sofreram e os que não sofreram o evento de interesse).

No caso dos modelos de regressão, é usual utilizar **medidas de variabilidade explicada** para quantificar o desempenho global dos modelos. Para respostas contínuas, a medida mais utilizada é o coeficiente de determinação R^2 , para respostas binárias é o R^2 de Nagelkerke (Nagelkerke, 1991) ou, como alternativa, o score de Brier (Brier, 1950). No caso de dados de sobrevivência, o R^2 de Nagelkerke pode ser também calculado embora tenham sido propostas muitas outras medidas que entram em consideração com o mecanismo de censura. De facto, o tema sobre medidas de variabilidade explicada é bastante complexo e mantém-se em aberto para investigação.

No que diz respeito à **calibração**, o teste de ajustamento de Hosmer-Lemeshow (Hosmer e Lemeshow, 2000), caso seja o modelo de regressão logística o escolhido para modelar os dados, é, sem dúvida, o mais utilizado. No entanto, tem alguns inconvenientes na medida em que é altamente dependente da forma como se decide agrupar as probabilidades estimadas e também não é claro que a escolha do número de grupos seja independente da dimensão da amostra. Assim sendo, é aconselhável utilizar alternativas para estudar a calibração dos modelos como, por exemplo, gráficos de calibração. Estes gráficos, de simples construção (*e.g.*, probabilidades estimadas no eixo dos x e o *outcome* no eixo dos y , a que se ajusta um suavizador, no caso de respostas binárias; no caso da regressão linear é um simples diagrama de dispersão), são caracterizados por uma ordenada na origem e por um declive (designados por *calibration-in-the-large* e *calibration slope*, no processo de validação), que constituem novas medidas de calibração bastante úteis no estudo do ajustamento do modelo.

A medida mais utilizada para avaliar a **capacidade discriminativa** de um modelo de predição é a medida de concordância c que, para respostas binárias, é idêntica à área sob a curva ROC (Receiver Operating Characteristic Curve) e para a qual foi proposta uma extensão para dados de sobrevivência (Heagerty *et al.*, 2000; Heagerty e Zheng, 2005). Também é usual calcular a diferença absoluta entre as médias das probabilidades estimadas dos que têm e dos que não têm o *outcome* (*discrimination slope*), acompanhada de um *box plot* das referidas estimativas.

Mais recentemente, surgiram novas medidas de desempenho dos modelos de predição: tabelas de reclassificação baseadas em categorias de risco (Cook, 2007) e curvas de predição (Huang *et al.*, 2007).

Cook propôs este tipo de tabelas para determinar quantos indivíduos eram reclassificados após a introdução de novas variáveis relevantes para o bom desempenho de um modelo de predição. Mais tarde, Pencina *et al.* (2008) sugeriram que esta reclassificação fosse efetuada separadamente para os

indivíduos com e sem o *outcome* (NRI - *Net Reclassification Improvement*). No entanto, estas abordagens baseiam-se na estratificação do risco em categorias sendo, por isso, possível obter diferentes valores do NRI para o mesmo modelo (ou biomarcador) ao considerar diferentes categorias. Assim sendo, Pencina *et al.* (2011) propõem um novo NRI cujo cálculo não necessita da discretização do risco designado-se, por isso, de *category-less* NRI ou *continuous* NRI. Extensível a dados de sobrevivência, esta medida permite a comparação de resultados, ao contrário das propostas anteriores.

Já existe a livreria *PredictABEL* em R que calcula o NRI e outras medidas de desempenho dos modelos de predição.

As curvas de predição (*predictiveness curves*) mostram a capacidade preditiva de um modelo (ou de um biomarcador) e permitem analisar se este identifica corretamente os indivíduos que se encontram em alto e baixo risco de ter o *outcome* de interesse. A livreria do R *pcvsuite* (funções: *roccurve*, *comproc*, *rocreg* e *predcurve*) permite calcular, entre outras medidas, a curva de predição.

Aconselha-se ainda a leitura/consulta das seguintes referências bibliográficas: Pepe, 2004; Steyerberg, 2009; Krzanowski e Hand, 2009; Harrel, 2010 e Steyerberg *et al.*, 2010.

5. Validação dos modelos de predição

Quando damos início a um estudo com a finalidade de obter um modelo de predição, isso significa que o desejamos utilizar para obter estimativas válidas do *outcome* em novos indivíduos (que não foram envolvidos no processo da modelação). Assim sendo, após a estimação do modelo, há que o validar. Existem várias técnicas de validação agrupando-se essencialmente em dois tipos: validação interna e validação externa.

Na **validação interna** o modelo é testado com a amostra que foi utilizada na modelação. Existem várias técnicas: validação aparente (o desempenho do modelo é obtido com os mesmos dados que foram utilizados na modelação); validação *split-sample* (a amostra é dividida de uma forma aleatória em dois grupos: um para a construção do modelo e o outro para a avaliação do seu desempenho); validação cruzada (*cross-validation*) (é uma extensão da validação *split-sample* considerando mais do que dois grupos) e validação *bootstrap* (um modelo de predição é desenvolvido em cada amostra *bootstrap*).

Na **validação externa** são considerados indivíduos que, de alguma forma, diferem dos que foram utilizados durante o desenvolvimento do modelo. As diferenças podem corresponder, por exemplo, a diferenças geográficas (o modelo é testado em indivíduos de outros hospitais ou regiões) ou a diferenças temporais (o modelo é testado futuramente em novos indivíduos).

Quando possível, a validação externa é mais aconselhável na medida em que aumenta a validade externa do estudo.

Seguramente que muito ficou por dizer sobre este vasto, complexo e tão apaixonante tema do ponto de vista da aplicação da Estatística. Entre outros assuntos, nada foi dito sobre a seleção de variáveis, problemas de sobreajustamento, valores omissos, discretização de variáveis, Enfim, não é em tão poucas páginas que se consegue escrever sobre todos estes tópicos. No entanto, espero que as referências que aqui foram indicadas ajudem a um aprofundar dos conhecimentos acerca desta área tão interessante que é a predição em Medicina.

Referências Bibliográficas

- Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley.
- Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed., Wiley.
- Aguiar, F.S., Almeida, L.L., Ruffino-Neto, A., Kritiski, A.L., Mello F.C.Q. e Werneck, G.L. (2012). Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. *BMC Pulmonary Medicine*, 12-40.
- Amaral-Turkman, M.A, e Silva, G.L. (2000). *Modelos Lineares Generalizados*. Livro do minicurso do VIII Congresso Anual da SPE.
- Austin, P.C., Lee, D.S., Steyerberg, E.W. e Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical Journal* **54**(5), 657–673.
- Banerjee, S., Carlin, B. e Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC.
- Bergmeir, C. e Benitez, J.M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software* Volume **46**, Issue 7. <http://www.jstatsoft.org/>.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon, Oxford.
- Braitman, L.E. e Rosenbaum, P.R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* **137**(8), 693-695.
- Breiman, L., Friedman, J.H., Olshen, R.A. e Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole.
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review* **78**, 1–3.
- Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. e Sturmer, T. (2006). Variable selection for propensity score models. *Am J Epidemiol* **163**(12), 1149-1156.
- Cabral, M.S. e Gonçalves, M.H. (2011). *Análise de dados longitudinais*. Sociedade Portuguesa de Estatística (ISBN: 978-972-8890-24-7).
- Cepeda, M.S., Boston, R., Farrar, J.T. e Strom, B.L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* **158**(3), 280-287.
- Carvalho, M.S., Andreozzi, V.L., Codeço, C.T., Campos, D.P., Barbosa, M.T.S. e Shimakura, S.E. (2011). *Análise de Sobrevida - Teorias e Aplicações em Saúde*. 2nd. ed., Editora Fiocruz, Rio de Janeiro.
- Carvalho M.L. e Natário I. (2008). *Análise de Dados Espaciais*. Sociedade Portuguesa de Estatística (ISBN: 978-972-8890-18-6).
- Clayton, D. e Hills, D. (2009). *Statistical models in epidemiology*. Oxford University Press Inc., New York.
- Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2nd edition, Chapman & Hall/CRC, Boca Raton.
- Cook, N.R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928-935.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data*. 2nd ed., John Wiley & Sons, New York.
- Diggle P.J., Heagerty P., Liang K.Y., Zeger S.L. (2002). *Analysis of Longitudinal Data*. 2nd. ed., Oxford University Press.
- Dobson, A.J. (2002). *An introduction to generalized linear models*. Chapman and Hall/CRC, Boca Raton.
- Faraway, J.J. (2005). *Linear models with R*. Chapman and Hall/CRC, Boca Raton.
- Finnerty, C.C., Ju, H., Spratt, H., Victor, S., Jeschke, M.G., Hegde, S., Bhavnani, S.K., Luxon, B.A., Brasier, A.R. e Herndon, D.N. (2012). Proteomics improves the prediction of burns mortality: results from regression spline modeling. *Clin Transl Sci*. **5**(3), 243-249.
- Fitzmaurice, G.M., Davidian, M., Verbeke, G. e Molenberghs, G. (2009). *Longitudinal Data Analysis*. Chapman and Hall/CRC, Boca Raton.
- Fitzmaurice, G.M., Laird, N.M. e Ware, J.H. (2004). *Applied Longitudinal Analysis*. Wiley, New Jersey.
- Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **1**, 1-41.

- Fromont, A., Binquet, C., Sauleau, E.A., Fournel, I., Bellisario, A., Adnet, J., Weill, A., Vukusic, S., Confavreux, C., Debouverie, M., Clerc, L., Bonithon-Kopp, C. e Moreau, T. (2010). Geographic variations of multiple sclerosis in France. *Brain* **133**, 1889–1899.
- Gail M.H. (2001). The estimation and use of absolute risk for weighing the risk and benefits of selective estrogen receptor modulators for preventing breast cancer. *Ann N Y Acad Sci.* **949**, 286-291.
- Gail M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C. *et al.* (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* **81**(24), 1879-1886.
- Gelfand, A.E., Diggle, P.J., Fuentes, M. e Guttorp, P. (2010). *Handbook of Spatial Statistics*. Taylor and Francis, Boca Raton.
- Grumett, S., Snow, P. e Kerr, D. (2003). Neural networks in the prediction of survival in patients with colorectal cancer. *Clin Colorect Cancer* **2**, 239-244.
- Günther, F. e Fritsch, S. (2010). neuralnet: Training of Neural Networks. *The R Journal* Vol. 2/1.
- Harrell, F.E. (2010). *Regression Modeling Strategies*. Springer Verlag, New York.
- Hastie, T. e Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, New-York.
- Hastie, T., Tibshirani, R. e Friedman, J. (2001). *The elements of Statistical Learning: Data mining, inference and prediction*. Springer, New York.
- Heagerty, P.J., Lumley, T., Pepe, M.S. (2000). Time-Dependent ROC curves for censored survival data and a Diagnostic Marker. *Biometrics* **56**, 337–344.
- Heagerty, P.J., Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics* **61**, 92–105.
- Hilbe, J.M. (2009). *Logistic Regression Models*. Chapman & Hall/CRC, Boca Raton.
- Hosmer, D.W. e Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd ed., Wiley, New York.
- Huang, Y., Pepe, M.S. e Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181-1188.
- Jewell, N.P. (2004). *Statistics for epidemiology*. Chapman & Hall/CRC, Boca Raton.
- Klein, J.P e Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.
- Kleinbaum, D.G., Kupper, L.L., Muller, K.E. e Nizam, A. (2008). *Applied Regression Analysis and Multivariable Methods*. 4rd ed., Duxbury Press, Pacific Grove.
- Krzanowski, W.J. e Hand, D.J. (2009). *Roc curves for continuous data*. Chapman & Hall/CRC, Boca Raton.
- Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data*. 2nd edition, Wiley, New York.
- Lawson, A.B. (2009). *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology*. Chapman & Hall/CRC, Boca Raton.
- Lee, Y.C., Lee, T.S., Lee, W.J., Lin, Y.C., Lee, C.K. e Liew, P.L. (2012). Predictors of anemia after bariatric surgery using multivariate adaptive regression splines. *Hepatogastroenterology* **59**(117), 1378-1380.
- Martuzzi, M. e Elliott, P. (1998). Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. *Ann Epidemiol.* **8**(1), 52-55.
- McCormick, T.H., Rudin, C. e Madigan, D. (2012). Bayesian hierarchical rule modeling for predicting medical conditions. *Ann. Appl. Stat.* **6**(2) , 652-668.
- McCullagh, P. e Nelder, J. (1989). *Generalized Linear Models*. 2nd ed., Chapman & Hall, London.
- Molenberghs, G. e Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.
- Nagelkerke, N.J.D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika* **78** (3), 691–692.
- Pappada, S.M., Cameron, B.D. e Rosman, P.M. (2008). Development of a neural network for prediction of glucose concentration in type 1 diabetes patients. *J Diabetes Sci Technol.* **2**(5), 792-801.
- Papoila, A.L., Rocha C., Geraldes C. e Xufre P. (2012). Generalized Linear Models, Generalized Additive Models and Neural Networks: Comparative Study in Medical Applications. *Actas do XVII Congresso da SPE (aceite para publicação)*.
- Paulino, C.D., Amaral-Turkman, M.A. e Murteira, B. (2003). *Estatística Bayesiana*. Fundação Calouste Gulbenkian. Lisboa.
- Paulino, C.D. e Singer, J.M. (2006). *Análise de Dados Categorizados*. Edgard Blucher, São Paulo.

- Pencina, M.J., D'Agostino Sr, R.B., D'Agostino Jr, R.B. e Vasan, R.S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statist. Med.* **27**, 157-172.
- Pencina, M.J., D'Agostino Sr, R.B. e Steyerberg, E.W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statist. Med.* **30**, 11-21.
- Penman, A.D. e Johnson, W.D. (2009). Complementary log-log regression for the estimation of covariate-adjusted prevalence ratios in the analysis of data from cross-sectional studies. *Biom J.* **51**(3), 433-442.
- Pepe, M.S. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.
- Pepe, M.S., Feng, Z., Janes, H., Bossuyt, P.M. Potter, J.D. (2008). Pivotal Evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Nati Cancer Inst* **100**, 1432-1438.
- Pires, A.M. e Branco, J.A. (2007). *Introdução aos métodos estatísticos robustos*. Sociedade Portuguesa de Estatística. (ISBN: 978-972-8890-10-0).
- Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge University Press.
- Rocha, C.M. e Papoila, A.L. (2009). *Análise de Sobrevivência*. Sociedade Portuguesa de Estatística. (ISBN: 978-972-8890-22-3).
- Rosado, F. (2006). *Outliers em dados estatísticos*. Sociedade Portuguesa de Estatística. (ISBN: 978-972-8890-07-9).
- Rosenbaum, P.R. e Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** (1), 41-55.
- Rothman, K.J. e Greenland, S. (1998). *Modern epidemiology*. Lippincott Williams & Wilkins, Philadelphia.
- Rue, H. e Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, London.
- Rue, H. e Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *J. Statist. Planng Inf.* **137**, 3177-3192.
- Rue, H., Martino, S. e Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society B*, **71**(2), 319-392.
- Scarvelis, D. e Wells, P.S. (2006). Diagnosis and treatment of deep-vein thrombosis. *CMAJ* **175**(9), 1087-1092.
- Schrödle, B. e Held, L. (2011). A primer on disease mapping and ecological regression using INLA. *Computational Statistics* **26**(2), 241-258.
- Steyerberg, E.W. (2009). *Clinical Prediction models*. Springer, Rotterdam.
- Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J. e Kattan, M.W. (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiol* **21**(1), 128-138.
- Therneau, T.M. e Grambsch, P.M. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer, New York.
- Twisk, J.W.R. (2005). *Applied longitudinal data analysis for epidemiology: A practical guide*. Cambridge University Press.
- Verbeke G. e Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.
- Vittinghoff, E., Glidden, D.V., Shiboski, S.C., e McCulloch, C.E. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models*. Springer-Verlag, New York.
- Xian-Hong Wang, Xiao-Hua Wu, Xiao-Nong Zhou (2006). Bayesian estimation of community prevalences of *Schistosoma japonicum* infection in China. *International Journal for Parasitology* **36**, 895-902.
- Wilson, P.W., D'Agostino R.B., Levy, D. Belanger, A.M., Silbershatz H. e Kannel, W.B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **97**(18), 1837-1847.
- Wood, S.N. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC, Boca Raton.
- Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software* **32**, 1-34.



Sobre o ensino de Estatística em cursos de Medicina

Pedro Oliveira, *pnoliveira@icbas.up.pt*

Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto

1. Introdução

O ensino de Estatística a alunos de Medicina tem vindo a constituir uma preocupação nos cursos de Medicina pelo reconhecimento do papel cada vez mais importante que a Estatística assume no exercício da profissão médica.

Os alunos de Medicina apresentam dificuldades na aprendizagem de Estatística por várias ordens de razões. Em primeiro lugar, estes alunos não têm uma formação matemática muito profunda, considerando mesmo que a matemática pré-universitária é algo que tem que ser ultrapassado mas não necessariamente com o empenho e dedicação que merecem outras disciplinas mais próximas do curso de Medicina, como sejam a Biologia ou a Química. Aliás, nem todos os cursos exigem a Matemática como disciplina específica de acesso ao curso de Medicina. Por outro lado, em geral, os cursos de Estatística são facultados no início do curso, o que se traduz nalguma frustração para quem espera sobretudo disciplinas relacionados com a essência da Medicina, como sejam a Anatomia, a Cardiologia ou, de uma forma geral, os Sistemas Orgânicos e Funcionais. Por esta última razão, os alunos tendem a encarar a disciplina de Estatística de uma forma pouco empenhada, até porque, por vezes, muito desligada daquilo que imaginam como o essencial da Medicina, pois a Estatística não é compreendida como relevante para as suas carreiras como médicos

Apesar do reconhecimento das dificuldades referidas no parágrafo anterior, a literatura sobre o assunto é muito pequena e dispersa. As dificuldades do ensino da Estatística a alunos de Medicina são reconhecidas há muito tempo. Dixon (1994) cita o relatório do *General Medical Council Education Committee Working Party Report* de 1987 sobre o ensino das ciências do comportamento, medicina comunitária e geral na educação médica, referindo que “os cursos de graduação de estatística médica devem ser ensinados antes do fim do segundo ano, com aplicações em dados usados pelos alunos, reforçado em anos subsequentes por projetos baseados em trabalho clínico”.

Tem sido reconhecido que o exercício da profissão médica requer a leitura e a interpretação dos resultados publicados de investigações de uma forma crítica, o que exige que o ensino da Estatística seja centrado nos “*principles involved and assumptions made*” [Altman, Bland (1991)] e não tanto nas competências numéricas necessárias para realizar as análises estatísticas. Assim, a aprendizagem deve ser centrada nos conceitos e não nas competências numéricas [Dawson-Saunders *et al.* (1987)].

Em geral os tópicos de estatística são ensinados quando existe um espaço no currículo e não quando o tópico é mais apropriado ou mais relevante para outras disciplinas. De facto, num estudo de Dawson-Saunders *et al.* (1987), a maioria dos tópicos de estatística eram ensinados nos primeiro e segundo ano (50 e 40%, respetivamente). Contudo, a questão não é só quando é que a Estatística é introduzida no currículo do curso de Medicina, mas também a forma como é ministrada. Switzer, Horton (2007)

referem a este propósito várias alternativas: ensino tradicional, aprendizagem direccionada, aprendizagem seleccionada pelos estudantes, ensino integrado ou ensino multidisciplinar. Parece, contudo existir hoje um consenso que a Estatística, após um curso inicial introdutório, deve ser introduzida ao longo do curso, de forma multidisciplinar, quer em disciplinas como Genética, Farmacologia, Epidemiologia, quer através dos projetos que os alunos devem realizar no fim do curso.

Switzer, Horton (2007) apresentam também uma revisão da literatura publicada, referindo que o número de artigos sem qualquer tratamento estatístico ou só com análises descritivas se tem mantido relativamente constante ao longo dos anos. Contudo, notam um aumento substancial dos artigos que usam tabelas de contingência, estatística em epidemiologia, análise de sobrevivência, regressão múltipla, comparações múltiplas e determinação da potência de testes, sendo também referidos artigos com regressão de Poisson e regressão logística. Esta revisão, conjuntamente com a constatação do uso de métodos cada vez mais sofisticados, reforça a ideia que o conhecimento estatístico dos cursos introdutórios não permitirá uma compreensão capaz do teor dos artigos publicados nas revistas científicas médicas de referência.

Por isso, Mathews, McPherson (1987), num artigo publicado no *British Medical Journal*, “*Doctors’ Ignorance of Statistics*,” afirmam que “médicos iletrados ... estão condenados a ter de aceitar sem reserva as afirmações feitas em sumários, discussões ou conclusões, e a sua prática clínica pode então ser alterada com base em evidência duvidosa ou inconclusiva”.

O Editorial do *The Lancet* de 1937 resume de alguma forma, esta longa preocupação com a importância da Estatística na Medicina:

“Statistics are curious things. They afford one of the few examples in which the use, or abuse, of mathematical methods tends to induce a strong reaction in non-mathematical minds. This is because statisticians apply, to problems in which we are interested, a technique we do not understand. It is exasperating, when we have studied a problem by methods that we have spent laborious years in mastering, to find out our conclusions questioned, and perhaps refuted, by someone who could not have made the observations himself” Editorial do *The Lancet*, “*Mathematics and Medicine*” de 1937.

2. Estatística nos cursos de Medicina

Planear um curso introdutório de Estatística em Medicina coloca a questão de quais os tópicos a abordar, sendo certo que não será possível abordar todos os tópicos e, por isso, uma escolha terá que ser realizada. Na definição dos tópicos convém ter presente algumas das ideias publicadas sobre o ensino da Estatística. Sobre este tema, dois trabalhos merecem especial relevância: Altman, Bland (1992) e Dixon (1994). O ensino da Estatística nos cursos de Medicina, como disciplina autónoma, é relativamente recente (no caso inglês desde os anos 70). Na generalidade dos cursos de Medicina são dedicadas cerca de 20 horas ao ensino da Estatística [Altman, Bland (1992)], embora este valor seja bastante variável. Em Portugal, os diversos cursos apresentam uma grande variabilidade. Assim, por exemplo, o curso da Faculdade de Medicina da Universidade do Porto apresenta logo no primeiro ano duas disciplinas (Introdução à Medicina I: Teoria da Medicina e Introdução à Medicina II: Informação em Saúde) a que correspondem 10 créditos ECTS; na Universidade do Minho, a disciplina de Introdução ao Curso de Medicina tem um total de 4 créditos ECTS; no Instituto de Ciências Biomédicas Abel Salazar, a disciplina de Métodos Quantitativos contabiliza 6 créditos ECTS; a disciplina de Biomatemática da Faculdade de Medicina da Universidade de Coimbra tem 6 créditos ECTS e a disciplina de Bioestatística e Informática do curso da Universidade Nova de Lisboa correspondem 8 créditos ECTS. Embora não seja possível uma avaliação precisa dos temas abordados e do tempo dedicado a cada um, é razoável assumir que o número de horas letivas oscila entre 2 a 4 horas semanais, ou seja, cerca de 30 a 60 horas letivas num semestre.

Tabela 1 – Teor estatístico dos artigos do New England Journal of Medicine ao longo do tempo [reproduzida de Switzer, Horton (2007)]

Procedimento	Notas	Artigos originais 1978–1979 n=332	Artigos originais 1989 n=115	Artigos originais 2004–2005 n=311	Acumulado por artigo* 2004–2005 %
Sem métodos estatísticos ou somente estatística descritiva	Estatísticas sumárias e intervalos de confiança	27%	12%	13%	13%
t-testes		44%	39%	26%	14%
Tabelas de contingência		27%	36%	53%	15%
Testes não paramétricos		11%	21%	27%	17%
Estatística epidemiológica	Risco relativo, medidas de associação, sensibilidade, etc.	9%	22%	35%	18%
Correlação de Pearson		12%	19%	3%	18%
Regressão linear simples	Regressão com uma variável preditora e uma variável resposta	8%	9%	6%	18%
Análise da variância		8%	20%	16%	20%
Transformação		7%	7%	10%	20%
Correlação não paramétrica		4%	1%	5%	21%
Análise de sobrevivência	Inclui regressão logística	11%	32%	61%	24%
Regressão múltipla	Inclui regressão passo a passo e alisamento	5%	14%	51%	39%
Comparações múltiplas		3%	9%	23%	41%
Ajuste e standardização		3%	9%	1%	41%
Tabelas multivaria	Mantel-Haenszel e modelos log-lineares	4%	10%	13%	44%
Potência	Inclui determinação da dimensão de amostras	3%	3%	39%	68%
Análise de custo benéfico		1%	0%	0%	68%
Análise de sensibilidade		0%	0%	6%	72%
Medidas repetidas*	Inclui análise longitudinal e regressão por clusters	-	-	12%	80%
Dados omissos		-	-	8%	87%
Testes de não inferioridade	Testes de equivalência	-	-	4%	91%
ROC (Receiver operating characteristic)	Área debaixo da curva/ sensibilidade e especificidade	-	-	2%	93%
Reamostragem	-	-	-	2%	94%
Componentes principais/análise factorial/cluster	-	-	-	2%	96%
Outros métodos	-	3%	9%	4%	100%

*A acumulação por artigo indica o número e a percentagem de artigos incluindo somente o correspondente procedimento e aqueles acima dele (procedimentos mais abaixo na hierarquia não são usados)

Switzer, Horton (2007) publicaram uma revisão da aplicação de métodos estatísticos em revistas científicas de Medicina, continuando o trabalho publicado por Emerson and Colditz em 1992. A tabela reproduz o resumo dos dois trabalhos, sendo as colunas de 1978-1979 e 1989 do trabalho de Emerson, Colditz (1992) e a coluna de 2004-2005 a atualização realizada pelos autores do artigo, Switzer, Horton (2007). Convém referir que as categorias foram definidas no primeiro trabalho de 1992, sendo que o último trabalho acrescentou as categorias assinaladas a negrito na primeira coluna da Tabela 1.

A coluna que reporta as percentagens acumuladas, segundo os autores, “descreve quantos artigos eram “completamente acessíveis” com o conhecimento do método referido e dos anteriores” [Switzer, Horton (2007)]. Como se pode verificar, são imensos os temas publicados que, obviamente, não podem ser todos considerados numa disciplina introdutória.

3. Em jeito de conclusão

Alguns dos métodos referidos na Tabela 1 requerem conhecimentos profundos que exigem a colaboração de estatísticos na elaboração dos artigos. Esta colaboração nem sempre é reconhecida pelo esforço que requer, por parte dos estatísticos, na compreensão do problema médico. E, tal como Altman, Bland (1992) referem, “embora a estatística seja um assunto fascinante em si mesmo, é essencialmente um conjunto de técnicas para responder a questões noutros campos”. Não sendo possível cobrir todos os temas, até porque não se pretendem formar estatísticos mas médicos, a ênfase do ensino da Estatística deve ser colocada nos princípios e não tanto no cálculo, ou seja, preparar os alunos para compreenderem, a um nível básico, a linguagem, os princípios e os métodos estatísticos mais comuns. E, para além do programa de uma disciplina, estará sempre subjacente o método de ensino e aprendizagem. A racionalidade na construção do programa não é garantia, por si só, de uma correta aprendizagem dos conceitos. Por essa razão, nos últimos anos, a investigação sobre mapas conceituais e a compreensão e aprendizagem dos conceitos tem produzido resultados muito interessantes (ver, a título de exemplo, http://nautilus.fis.uc.pt/gazeta/revistas/26_1/entrevista.pdf). Importa ter presente que os alunos de Medicina são expostos, ao longo do curso, a diferentes tópicos de Estatística, em disciplinas como Saúde Comunitária, Epidemiologia, Farmacologia, Ensaio Clínicos e Genética entre outras. A questão que se coloca é da articulação entre os diversos tópicos por forma a que haja uma filosofia de ensino e aprendizagem ao longo do curso, de maneira que os tópicos sejam introduzidos no momento necessário e em contexto relevante e com significado para os alunos. Desta forma, os futuros médicos terão os conhecimentos para formularem as questões que querem investigar com os estatísticos que os venham a acompanhar no desenvolvimento dos seus trabalhos.

Referências

- Altman, D. G., Martin Bland, J. (1991). Improving Doctors' Understanding of Statistics, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 154, 2, 223-267
- Dawson-Saunders, B., Azen, S., Greenberg, R. S. Reed, A. H. (1987). The Instruction of Biostatistics in Medical Schools, *The American Statistician*, 41, 4, 263-266.
- Dixon, R. A. (1994). Medical Statistics: content and objectives of a core course for medical students, *Medical Education*, 28, 59-67
- Emerson, J.D., Colditz, G.A. (1992). Use of Statistical Analysis in The New England Journal of Medicine. In *Medical Uses of Statistics*, 2nd Ed. (Bailar, J.C. and Mosteller, F., Eds.) Boston: NEJM Books.
- Matthews, D.R. and McPherson, K. (1987). Doctors' Ignorance of Statistics, *British Medical Journal*, 294, 856-857.
- Switzer, S. S., Horton, N. J. (2007). What Your Doctor Should Know about Statistics (but Perhaps Doesn't...), *Chance*, Vol. 20, 1, 17-21.



Métodos Estatísticos para Modelar e Prever o Impacto das Epidemias de Gripe em Portugal

M. Lucília Carvalho, *mlucilia.carvalho@gmail.com*
CEAUL, Faculdade de Ciências, Universidade de Lisboa

Isabel Natário, *icn@fct.unl.pt*
CEAUL e Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Baltazar Nunes, *baltazar.nunes@insa.min-saude.pt*
Departamento de Epidemiologia, Instituto Nacional de Saúde Dr. Ricardo Jorge.

1. Introdução

A gripe é uma infeção respiratória aguda que em Portugal, e mais geralmente no hemisfério norte, é responsável por epidemias que se desenvolvem durante o outono e o inverno com um impacto muito importante na saúde humana que se traduz muitas vezes por excessos de mortalidade e um considerável aumento da necessidade de cuidados de saúde, nomeadamente hospitalizações. Além da perda de vidas humanas e dos custos financeiros destes serviços, que podem ser, por si só, muito elevados, existem outros encargos sociais indirectos provocados pelo absentismo ao trabalho e às atividades escolares que, nos casos muito graves, podem mesmo conduzir a ruturas no fornecimento de bens e serviços.

Outro aspeto que convém a realçar, motivo de grande preocupação das autoridades de saúde mundiais, é o carácter global que as epidemias de gripe podem rapidamente alcançar devido à atual rapidez de contacto entre populações de diferentes continentes e à facilidade de propagação da doença.

Neste contexto, tornou-se indispensável implementar Sistemas de Vigilância epidemiológica da Gripe (SVG) a nível nacional e internacional com o objetivo de fornecer às autoridades de saúde informação para a avaliações de risco atualizadas que permitam uma correta aplicação de medidas de controlo e mitigação das epidemias e suas consequências.

Baseados nos dados fornecidos pelos SVG oficiais, e por alguns sistemas alternativos, têm-se desenvolvido múltiplos métodos estatísticos para dar resposta às necessidades que acabámos de referir. No presente trabalho, interessam-nos especialmente os dirigidos à medição do efeito das epidemias de gripe em termos de excessos de mortalidade e também os que se dirigem à deteção precoce do início das epidemias, pico de intensidade e termo. De facto, apesar do intenso esforço de recolha de informação por parte destes sistemas, subsistem algumas falhas, nomeadamente no que diz respeito ao registo da mortalidade atribuível diretamente à gripe e também às causadas por demoras importantes nos processos de recolha, tratamento e análise dos dados, que tornam necessária a aplicação de métodos estatísticos bastante sofisticados para estimar alguns dos principais parâmetros de avaliação dos efeitos destas epidemias.

O objetivo deste texto é apresentar resumidamente três artigos que resultaram do trabalho realizado sobre estes temas, desde 2009, pelo grupo de investigação constituído por Baltazar Nunes, Isabel Natário e M. Lucília Carvalho, sendo que os dois primeiros artigos foram elaborados no âmbito da tese de Doutoramento do primeiro dos autores citados.

Uma bibliografia mais extensa sobre os temas que se apresentam abaixo, bem como sobre os diferentes modelos e técnicas estatísticas utilizadas, pode ser encontrada em cada um dos artigos objeto deste resumo.

2. Problema da Estimação do Excesso de Mortalidade associado às Epidemias de Gripe

Durante os períodos de epidemia de gripe é comum observar-se um acréscimo da mortalidade por todas as causas, principalmente entre a população idosa (mais de 65anos). Este aumento pode ser atribuído à infeção gripal que dá origem a complicações que levam à hospitalização e/ou morte dos infetados.

O impacto da gripe em termos de hospitalizações ou mortes nunca pode ser medido através do número de mortes por gripe como causa de morte, inscrito nos registos de mortalidade nacionais, porque este valor é normalmente muito baixo, mesmo durante as epidemias mais severas. Isto deve-se ao facto de ser difícil estabelecer uma ligação entre uma complicação (pneumonia ou outras doenças respiratórias, doenças do aparelho circulatório, etc) e uma infeção prévia ou presente, por falta de confirmação laboratorial de diagnóstico.

Desta forma, os investigadores foram levados a encontrar métodos para estimar o número de mortes associadas à gripe usando como ponto de partida a série temporal de mortalidade geral, ou outra série de mortalidade que inclua, para além da gripe, outras causas de morte que com ela possam estar relacionadas e, quando disponível, informação adicional dos SVG sobre a ocorrência de epidemias sazonais e suas características.

Geralmente os métodos para estimar o excesso de mortalidade atribuível às epidemias de gripe desenvolvem-se em três passos sequenciais:

1. Obter uma linha de base do número de mortes por uma certa unidade de tempo, na ausência das epidemias de gripe.
2. Usar a linha de base para identificar os períodos onde há evidência de existir excesso de mortes atribuíveis a epidemias de gripe.
3. Subtrair a linha de base à série do número observado de mortes durante esses períodos.

Desta forma, o número de mortes observadas acima da linha de base, quando verificadas durante períodos de epidemia de gripe, pode, na ausência de outras causas que as possam explicar, ser atribuíveis à epidemia.

Embora variem consideravelmente em aspetos particulares, existem duas grandes classes de métodos que, em traços gerais, seguem o esquema apresentado. O primeiro grupo compreende os métodos estatísticos que envolvem modelos de regressão múltipla que, como variáveis explicativas incluem, além de variáveis temporais e climatéricas, indicadores de atividade gripal. Estes modelos são muito exigentes em termos da precisão destes últimos dados, cuja recolha habitualmente realizada por Sistemas Sentinela (SS) de vigilância da gripe, está muito dependente de fatores externos como, por exemplo, os períodos de férias. Os métodos do segundo grupo são caracterizados por não incluírem covariáveis e também por, no processo de estimação, utilizarem séries temporais de mortalidade das quais são excluídos os períodos onde existe evidência de ocorrência de epidemia de gripe, i.e. uma série temporal interrompida.

Uma análise mais detalhada do segundo grupo de métodos mostra, ainda assim, grandes diferenças entre eles, das quais se destacaram como essenciais as que dizem respeito aos seguintes parâmetros de escolha: tipo de modelo estatístico utilizado (regressão cíclica [1, 2, 3, 4, 5] ou modelos ARIMA [6, 7]),

método usado para construir a linha de base (não iterativo [4, 3, 5] ou iterativo [6, 1, 2, 7]) e escolha dos períodos a serem excluídos da série temporal da mortalidade (períodos epidêmicos fixos, como dezembro a abril [2, 3, 5], ou definidos através de informação produzida por sistemas de vigilância de doenças de tipo gripal [6, 1, 4, 7]).

No artigo *Time series methods for obtaining excess mortality attributable to influenza epidemics* [8], os autores, através da citada identificação das diferenças entre os métodos, conseguiram enquadrá-los numa única classe em que as características específicas variam de acordo com os três pontos referidos o que simplifica enormemente a sua descrição e comparação da respetiva aplicabilidade e resultados fornecidos. Note-se ainda que a sistematização conseguida permitiu, por combinação de todas as possíveis alternativas, a construção de três métodos que não tinham sido anteriormente considerados na literatura.

A unificação conseguida serviu também, de forma natural, como base para a construção de um conjunto de rotinas em R, de fácil utilização, o pacote *flubase*, que estima as mortes associadas à gripe por qualquer dos métodos incluídos na classe.

Em virtude do verdadeiro número de mortes atribuível às epidemias de gripe que ocorreram durante um determinado período de tempo ser desconhecido, e das desigualdades entre métodos conduzirem a diferentes estimativas desse número, a comparação entre métodos implementada no pacote *flubase* tem uma natureza empírica. Ela é realizada através da simples confrontação entre as estimativas do excesso de mortes para cada ano gripal do período, (o ano gripal inicia-se na semana 27 do correspondente ano civil), da identificação dos métodos que estimam o valor mais elevado e mais baixo do excesso de mortes, e ainda do cálculo da correlação entre as estimativas obtidas pelos diferentes métodos. Em termos estatísticos, mede-se a qualidade do ajustamento da linha de base à série dos dados, através do cálculo da média dos resíduos quadrados (SRM) e da função de autocorrelação dos mesmos, limitados, evidentemente, às semanas fora dos períodos em que se interrompeu a série.

A utilização do pacote *flubase* permite aos investigadores analisarem facilmente os resultados obtidos com os diferentes métodos nele incluídos, ter uma discussão mais frutuosa sobre as vantagens de cada um deles e adquirir, desta forma, uma melhor compreensão dos dados que estão a estudar.

Como ilustração das capacidades que acabámos de apontar, apresentam-se as principais conclusões obtidas com base no exemplo exposto no artigo em análise, em que se aplicam os métodos incluídos no pacote *flubase* à série semanal do número de mortes por pneumonia e gripe em Portugal de 1980-81 a 2003-04. Embora estas conclusões sejam suportadas por um único exemplo, é de prever que noutros casos o mesmo tipo de comparações resulte em conclusões semelhantes, nomeadamente no que diz respeito ao parâmetro de escolha que se revelou ser o mais sensível: o tipo de período epidémico, fixo ou variável, a excluir da série temporal dos dados.

Em resumo, para os três parâmetros de escolha identificados, temos os seguintes resultados:

- 1- Na ausência de uma fonte de informação externa exata, devem escolher-se métodos com períodos epidêmicos fixos. Contudo, é preferível a utilização de períodos estimados a partir de informação confiável quando se pretende um método mais conservador, aconselhável quando se realizam estudos de natureza ecológica como é o caso destes. De facto, esta última opção traduz-se normalmente em períodos epidêmicos mais curtos e adaptados às observações evidenciando uma forte variabilidade anual.
- 2- Os modelos ARIMA sazonais são mais adequados para este tipo de dados que os modelos de regressão cíclica por múltiplas razões entre as quais se destaca a não autocorrelação dos resíduos. Embora este resultado já tenha sido referido anteriormente [6] estes últimos modelos continuam a ser correntemente preferidos pelas agências governamentais (e.g. US Centre for Disease Control and Prevention), provavelmente por serem de mais fácil aplicação e por produzirem linhas de base regulares.

- 3- Os processos iterativos devem ser exclusivamente usados para séries temporais estacionárias visto produzirem resultados muito longe da realidade nos períodos em que ocorre uma modificação na tendência da série. Note-se, no entanto, que estes modelos são muito úteis quando se pretende definir uma linha de base com intuítos de identificação prospetiva e estimação do excesso de mortes num quadro de vigilância da mortalidade.

3. Problema da Previsão em Tempo Real da Evolução das Epidemias de Gripe

Durante o desenrolar de surtos de doenças infecciosas que constituem ameaças graves para a saúde pública (SARS, vírus do Nilo, pandemia de gripe, etc), é essencial conhecer, em tempo real, a tendência (crescente ou decrescente) da epidemia, a respetiva distribuição espacial e, como já referimos na secção anterior, o seu impacto em termos de consultas médicas, hospitalizações e mortes, para identificar quais as medidas que devem ser tomadas para controlar a sua difusão e mitigar os seus efeitos.

Na Europa, esse papel cabe a SVGs geralmente suportados por Redes Sentinela (RS), constituídos por Médicos de Família (MF) que fornecem semanalmente o número de pacientes que nas suas consultas apresentam sintomas de síndrome gripal (SG). Este número dividido pelo número de pacientes nas listas de utentes dos médicos participantes na RS possibilitam estimar as taxas de incidência de SG. Além disso, os MF também fornecem para alguns dos pacientes de SG, exsudados nasofaríngeos que, depois de confirmação laboratorial para infeção por gripe, suportam a vigilância virológica da atividade gripal.

A nível europeu, o sistema de vigilância da gripe é mantido pelo European Centre for Disease Prevention and Control (ECDC), através duma rede de países chamada European Influenza Surveillance Network (EISN) que enviam, cada quarta-feira, as taxas de incidência específicas por idade de SG e a informação virológica, correspondentes à semana anterior, sendo a informação completa publicada à sexta-feira no WISO, Weekly Influenza Surveillance Overview.

Pelas razões referidas inicialmente, quando um surto de gripe começa a desenvolver-se é essencial detetar precocemente o começo do período epidémico, caso este venha a ocorrer e, com esse objetivo, existe uma enorme variedade de métodos estatísticos para os quais se pode encontrar uma revisão exaustiva [9, 10, 11, 12]. A identificação da primeira semana do período epidémico, definido para este fim como o conjunto de semanas onde os níveis de incidência de SG indicam atividade gripal epidémica, é sempre feita a posteriori e, no caso desses métodos usarem os dados fornecidos pela RS, podem, na melhor das hipóteses, emitir um alerta na semana seguinte ao início da epidemia. No entanto, muitos destes sistemas de vigilância recebem diariamente informação sobre a situação corrente que, embora incompleta, pode ser suficiente para a deteção em tempo real do início, pico e fim da epidemia. O processo de predição da situação presente da epidemia, usando esta informação incompleta, foi considerado do máximo interesse e recebeu o nome de “nowcasting” [13].

No artigo que se resume de seguida, *Nowcasting influenza epidemics using non-homogeneous hidden Markov models* [14], os autores desenvolvem modelos de cadeia de Markov escondidas (Hidden Markov Models, HMM) que usam a informação coligida até à sexta-feira de cada semana, para predizer nesse mesmo dia, duas medidas: a taxa de incidência de SG e o estado de atividade gripal (epidémico ou não epidémico), que, de outra forma, só seriam reportadas ao ECDC na quarta-feira da semana seguinte, altura em que toda a informação se considera recolhida.

Geralmente, num HMM assume-se que uma dada série temporal observada y_t , $t=1, 2, \dots$ é uma realização de um processo estocástico $\{Y_t: t = 1, 2, \dots\}$, em que a distribuição de Y_t é condicionalmente determinada por uma cadeia de Markov discreta, não observada (hidden), $\{S_t: t = 1, 2, \dots\}$, que toma valores num conjunto de m estados. Na sua forma mais trivial, a cadeia de Markov é homogénea, de ordem 1, o que significa que é completamente definida pela distribuição inicial, $P(S_1 = i)$, $i =$

1, ..., m, e pela matriz $\Gamma = (\gamma_{j,i})$, $i, j = 1, \dots, m$, cujas entradas são as probabilidades de transição elementares entre estados, $\gamma_{j,i} = P(S_t = i | S_{t-1} = j)$, supostas invariantes no tempo.

Modelos deste tipo, em que a cadeia de Markov possui apenas dois estados, epidémico (1) e não epidémico (0), foram aplicados a dados de vigilância de SG [15, 16, 17] apenas com o intuito de classificar as semanas já reportadas e nunca para prever as taxas de SG ou os estados da cadeia de Markov para semanas futuras. Com o fim de realizar estas previsões de uma forma que intuitivamente parece ser mais adequada, os autores usaram um HMM não homogéneo em que as probabilidades de transição elementares são elas próprias modeladas através de uma função logística de covariáveis dependentes do tempo [18, 19],

$$\text{logit}(\gamma_{j,i}^t) = \ln \frac{\gamma_{j,i}^t}{\gamma_{j,j}^t} = \alpha_{j,i} \mathbf{Z}_t; i, j = 0, 1,$$

onde \mathbf{Z}_t é um vector de q covariáveis medidas no instante t . Essa intuição baseou-se na observação de diferenças consideráveis nas velocidades de crescimento e decrescimento das taxas de incidência de SG para diferentes epidemias, que pareciam estar associadas à circulação mais ou menos acentuada de vírus de gripe.

No exemplo de aplicação, que serviu para confirmar a mais-valia dos modelos não homogéneos, foram construídos dois modelos deste tipo em que a taxa de incidência da semana t observada na quarta-feira da semana $t+1$, $y_{t(t+1)}$, é descrita pelas seguintes equações condicionadas pelo estado de atividade gripal,

$$y_{t(t+1)} = \begin{cases} \mu + \beta_1 \cos\left(\frac{2\pi t}{52}\right) + \beta_2 \sin\left(\frac{2\pi t}{52}\right) + \theta_0 y_{t(t)0} + e_{t0} & S_t = 0, \\ \mu + \beta_1 \cos\left(\frac{2\pi t}{52}\right) + \beta_2 \sin\left(\frac{2\pi t}{52}\right) + \theta_{01} y_{t(t)1} + \theta_{02} y_{t(t)1}^2 + e_{t1} & S_t = 1, \end{cases} \quad (1)$$

em que $y_{t(t)0}$ e $y_{t(t)1}$ são modificações adequadas das observações incompletas da incidência na semana t , $y_{t(t)}$, e as variáveis aleatórias $e_{ti} \sim N(0, \tau_i)$ têm precisão, $\tau_0 > \tau_1$, $t = 1, \dots, T$. Além disso, as probabilidades de transição incluem como covariáveis, no Modelo 1, a taxa de incidência de SG observada incompletamente, $y_{t(t)}$, e no Modelo 2, além desta taxa, o número absoluto de casos de SG positivos para gripe observados na semana anterior, $v_{t-1(t)}$.

Para efeitos de comparação foi também construído um modelo homogéneo, Modelo 0, que mantém as equações (1) para a taxa de incidência mas tem matriz de probabilidades de transição invariante no tempo.

Os parâmetros e a sequência dos estados escondidos para os três modelos foram estimados usando métodos numéricos baseados em amostragem de Monte Carlo via cadeias de Markov (MCMC). Para os parâmetros usou-se o amostrador de Gibbs e o algoritmo de Metropolis-Hastings, nos casos em que o primeiro não pode ser utilizado, e para a sequência de estados o algoritmo forward filtering backward sampling [17,18,19].

Além, das estimativas anteriores foi também estimado o estado de actividade gripal da última semana observada incompletamente, \hat{s}_T , e a taxa de incidência de SG para esta última semana, que só irá ser reportada na semana $T+1$, $\hat{y}_{T(T+1)}$.

Numa primeira fase, os três modelos foram aplicados à série completa de dados disponível, referentes ao período entre a semana 40/2008 e a semana 16/2011, com o propósito de comparar as estimativas dos parâmetros, o ajustamento das séries estimadas e a classificação retrospectiva de cada semana no estado epidémico ou não epidémico, designada habitualmente por “decoding” dos estados de atividade gripal. A comparação dos ajustamentos fez-se através do cálculo do logaritmo da verosimilhança marginal e de medidas de bondade de ajustamento (RMSE e χ^2), que mostraram que os modelos não homogéneos proporcionavam um melhor ajustamento, embora as estimativas dos parâmetros comuns

não variassem muito entre os modelos. Sublinha-se que coeficientes associados a algumas das covariáveis dos modelos logísticos para as probabilidades de transição revelaram-se ser estatisticamente significativos evidenciando que a introdução de modelos não homogêneos tem justificação.

No que se refere ao “decoding” dos estados de atividade gripal encontrou-se uma consistência forte nas estimativas dos períodos epidêmicos dos anos 2008-09 e 2010-11, mas no ano de 2009-10, que corresponde à pandemia provocada pelo vírus A(H1N1)pdm09, os modelos não homogêneos detetaram dois períodos epidêmicos, ao contrário do Modelo 0 que os aglomerou num mesmo período. A divisão em dois períodos feita pelos Modelos 2 e 3 parece estar mais de acordo com as observações da generalidade dos países europeus onde parece ter existido uma primeira vaga de casos durante o verão, associada à entrada de turistas ou emigrantes, e uma segunda vaga durante a época gripal normal.

Numa segunda fase, para comparar os três modelos, no que respeita ao “nowcasting” e “decoding” em tempo real, e no fim do período epidémico, utilizou-se apenas a época gripal de 2010-11. A observação da Figura 1, em que se apresenta, para cada um dos modelos e para cada semana da época, a média da distribuição a posteriori da probabilidade do estado ser epidémico (pmp), permite ilustrar mais convenientemente os principais resultados desta comparação. Definindo como período epidémico o conjunto de semanas consecutivas em que a pmp está acima de 0.5 verifica-se (painel 3) que, no final da época, todos os modelos definem o período epidémico da semana 46/2010 à semana 10/2011. Se esse intervalo for considerado como o “verdadeiro período epidémico” a pontualidade da deteção do início da epidemia por um certo modelo pode definir-se como o tempo decorrido desde a semana 46/2010 até à semana em que é produzido o alerta de começo de epidemia por esse modelo. Assim sendo, a observação dos painéis 1 e 2 permite concluir que, enquanto o Modelo 0 tem uma pontualidade de deteção do início da epidemia de 4 semanas, o Modelo 2 melhorou essa pontualidade em 2 semanas. Definindo de forma análoga à utilizada anteriormente a pontualidade da deteção do fim da epidemia, conclui-se que os modelos não homogêneos também apresentam melhores resultados que os homogêneos.

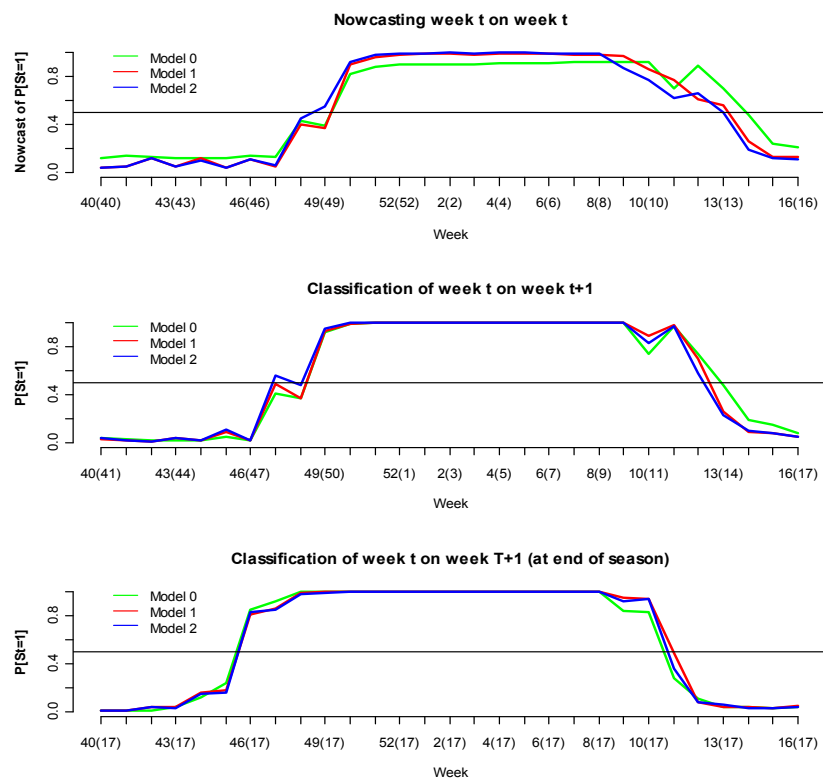


Figura 1: Média das distribuições a posteriori do estado epidémico para cada semana da época 2010-11. Painel 1: na semana corrente (nowcast); Painel 2: na semana seguinte; Painel 3: no final da época. Notação: () semana do cálculo.

Tendo em consideração que o “nowcast” da taxa de incidência de SG, não apresenta grandes diferenças nos 3 modelos, e que estas previsões se encontram em sincronia com as observações finais produzidas na quarta-feira da semana seguinte, pode dizer-se que qualquer dos modelos prevê satisfatoriamente a taxa de incidência com a informação incompleta da semana.

Não é por demais notar que o melhor desempenho dos modelos HMM não homogêneos face ao modelo homogêneo, se verifica essencialmente na capacidade de deteção rápida da atividade epidémica, resultado que não é certamente alheio à inclusão de covariáveis com informação antecipada sobre a evolução da epidemia, na modelação das probabilidades de transição da cadeia de Markov escondida.

Como conclusão realça-se que a antecipação em algumas semanas dos alertas de começo e fim duma epidemia, e a boa previsão das taxas de incidência conseguidas através dos modelos não homogêneos, os tornam num instrumento valioso na programação atempada das medidas de mitigação dos efeitos das epidemias. No entanto, não se deve deixar de referir que estes modelos apresentam algumas desvantagens face a modelos mais simples. Elas decorrem do elevado tempo de cálculo do programa de “nowcasting” (o programa que foi implementado no software R demora cerca de 12 horas a correr), e à necessidade de algum afinamento na forma de inclusão das covariáveis em qualquer das regressões utilizadas, que deve ser específica de cada aplicação real, o que requer um conhecimento especializado.

4. Problema da falta de representatividade dos sistemas de vigilância sindrómica

A gravidade do impacto na saúde pública de alguns surtos epidémicos, já aqui referida por diversas vezes, levou a que entidades públicas e privadas tenham investido na criação de sistemas de vigilância epidemiológica, frequentemente baseados em sintomas e não em diagnósticos confirmados – Sistemas de Vigilância Sindrómica (SVS).

Alguns exemplos de SVS já consagrados são: Biological Spatio-Temporal Outbreak Reasoning Module, BioStorm [21], Automated Epidemiological Geotemporal Integrate Surveillance, AEGIS [22], Eletronic Surveillance System for the Early Notification of Community-Based Epidemics, ESSENCE II [23], HealthMap [24], GRIPENET [25] e Rede de Médicos Sentinela do Instituto Dr. Ricardo Jorge, de que já falámos na secção anterior.

Embora estes sistemas de vigilância se destinem a produzir alarmes que necessitam de investigação ulterior, a sua eficácia neste campo tem sido comprovada, nomeadamente na deteção de aglomerados de casos de doença. São bem conhecidas, no entanto, as fragilidades que alguns deles apresentam na recolha dos dados que fornecem, que dificultam ou impedem a sua utilização quando se pretende extrapolar os resultados observados para universos mais amplos. Na maior parte dos casos, estas fragilidades estão ligadas à dificuldade dos sistemas distinguirem entre verdadeiros e falsos positivos, e ao facto de só serem capazes de seguir um conjunto não aleatório de indivíduos, como consequência de se basearem na participação voluntária. É este o caso do GRIPENET, sistema posto em funcionamento em Portugal em 2005, que se dedica à monitorização de doenças com sintomas semelhantes aos da gripe (Influenza-Like Illness, ILI), em que a participação se faz através do preenchimento na internet de um questionário semanal sobre sintomas de ILI.

Para tratar o problema de dados pobres e não aleatórios associado a este tipo de recolha, o artigo resumido nesta secção, *Addressing the problem of lack of representativeness on syndromic surveillance schemes* [26], utiliza uma técnica de estimação, habitualmente associada à modelação da dinâmica de populações animais, baseada em Modelos de Espaço de Estados, (State Space Models, SSM).

Tipicamente, um SSM adequa-se bem à descrição da evolução de um sistema que pode ser representado por uma série temporal não observada, o processo de estados, com o qual está associada

de algum modo, uma série temporal de observações, o processo observacional, sendo a relação entre os dois especificada pelo próprio SSM [27].

O objetivo da utilização dum SSM no contexto observacional do GRIPENET é o de estimar a verdadeira incidência de ILI na população portuguesa e, para isso, o processo de estados, aqui notado por $\{n_t, t = 0, 1, \dots, T\}$, representa o vetor não observado, constituído pelas populações com e sem sintomas de ILI em cada semana, subdivididas ainda nas que estão a participar no GRIPENET e nas que não estão,

$$n'_t = \left(n_{i,t}(\overline{GN}), n_{ni,t}(\overline{GN}), n_{i,t}(GN), n_{ni,t}(GN) \right),$$

em que, para a semana t , $n_{i,t}(\overline{GN})$ e $n_{ni,t}(\overline{GN})$ representam, respetivamente, as populações com e sem sintomas de ILI sob observação no GRIPENET, e $n_{i,t}(GN)$ e $n_{ni,t}(GN)$ o mesmo para as que não estão sob observação, tudo medido em número de pessoas em cada categoria.

Note-se que este vector pode ainda ser subdividido segundo outras características, como o género, a idade e o local de residência, conhecidos para todos os respondentes do GRIPENET, com vantagens óbvias para o ajustamento do modelo e conseqüente estimação da incidência de ILI na totalidade da população. De facto, sendo a taxa de participação marcadamente desigual entre estes grupos, um dos maiores problemas que apontámos inicialmente a este tipo de recolha, a não aleatoriedade da amostra respondente, pode ser mitigado através da assunção de diferentes distribuições e parâmetros para cada um dos grupos considerados.

De acordo com a abordagem usada na dinâmica de populações [28, 29, 30] escolheu-se uma formulação estocástica baseada no seguinte processo determinístico geral, $n_t = Pn_{t-1}$, onde P é uma matriz de projecção populacional [31]. No caso vertente, supôs-se que os nascimentos e mortes durante o período de observação eram negligenciáveis e designou-se por N a dimensão constante da população.

A complexidade da dinâmica desta população pode ser melhor capturada subdividindo o processo de estados em subprocessos que se sucedem no tempo sempre pela mesma ordem, cada um dos quais dependente exclusivamente do subprocesso anterior. Usando a mesma formulação determinística, se a cada um dos subprocessos corresponder uma matriz, a matriz de projecção P será o produto dessas matrizes. Neste caso, considerou-se $P = GN C$, em que C corresponde ao subprocesso de mudança de classe, provocada pelo aparecimento de sintomas de ILI ou pelo seu desaparecimento, e GN ao subprocesso relacionado com a entrada ou abandono do sistema GRIPENET, subprocesso esse designado de agora em diante por Gripenet.

A correspondente formulação estocástica é feita em termos dos valores esperados condicionais do processo de estados, que supomos ser uma cadeia de Markov de ordem 1,

$$E[n_t | n_{t-1}] = P n_{t-1},$$

em que P descreve o efeito médio do conjunto de subprocessos tais que :

$$E[n_t | n_{t-1}] = GN C n_{t-1}.$$

A hipótese de Markov assegura que o processo de estados fica completamente conhecido desde que, para cada instante t , se saiba a distribuição do processo condicional no valor do processo no instante anterior, $n_t \sim H_t[n_{t-1}]$, que, em virtude do já exposto, pode ser decomposto em:

$$u_t^C \sim H_t^C[n_{t-1}] \quad n_t = u_t^{GN} \sim H_t^{GN}[u_t^C],$$

em que u_t^C e u_t^{GN} representam os vetores de estados, respetivamente depois dos subprocessos mudança de classe e Gripenet.

Foram feitas as seguintes hipóteses para as distribuições destes subprocessos

Para o processo de mudança de classe:

$$u_t^c \sim H_t^c [n_{t-1}] : \begin{pmatrix} u_{i,t}^c = X[n_{ni,t-1}] \\ u_{ni,t}^c = N - u_{i,t}^c \end{pmatrix},$$

$$X[n_{ni,t-1}] \sim \text{Binomial}(n_{ni,t-1}(GN) + n_{ni,t-1}(\overline{GN}), C_i)$$

em que C_i designa a probabilidade, igual para todos, de um indivíduo saudável desenvolver sintomas de ILI.

Para o processo Gripenet:

$$n_t = u_t^{GN} \sim H_t^{GN} [u_t^c] : \begin{pmatrix} n_{i,t}(\overline{GN}) = u_{i,t}^c - n_{i,t}(GN) \\ n_{ni,t}(\overline{GN}) = u_{ni,t}^c - n_{ni,t}(GN) \\ n_{i,t}(GN) = Y[u_{i,t}^c] \\ n_{ni,t}(GN) = Z[u_{ni,t}^c] \end{pmatrix},$$

$$Y[u_{i,t}^c] \sim \text{Binomial}(u_{i,t}^c, p_i), \quad Z[u_{ni,t}^c] \sim \text{Binomial}(u_{ni,t}^c, p_{ni}),$$

em que p_i e p_{ni} designam respetivamente as probabilidades de um indivíduo com e sem sintomas de ILI estarem a ser seguidos pelo GRIPENET.

No que diz respeito ao processo observacional, designado por $\{y_t, t = 0, 1, \dots, T\}$, com componentes $(y_{i,t}, y_{ni,t})$ representando, respetivamente, medições independentes, com erro, dos estados populacionais $n_{i,t}(\overline{GN})$ e $n_{ni,t}(\overline{GN})$, assumem-se as seguintes distribuições:

$$\begin{aligned} y_{i,t} | n_t &\sim N(n_{i,t}(\overline{GN}), \varphi_1^2 n_{i,t}(\overline{GN})^2) \\ y_{ni,t} | n_t &\sim N(n_{ni,t}(\overline{GN}), \varphi_2^2 n_{ni,t}(\overline{GN})^2) \\ y_t | n_t &\sim N(n_{i,t}(\overline{GN}), \varphi_1^2 n_{i,t}(\overline{GN})^2) \times N(n_{ni,t}(\overline{GN}), \varphi_2^2 n_{ni,t}(\overline{GN})^2). \end{aligned}$$

Todas as inferências em que se possa estar interessado, resultam de integrações, raramente triviais, da especificação completa da distribuição de probabilidade dos estados (incluindo os intermédios), o que faz com que os métodos de simulação de Monte Carlo sejam uma alternativa inevitável.

Na aplicação do SSM aos dados do GRIPENET da época gripal de 2006-07, apresentada como exemplo, foram escolhidas distribuições a priori para $(C_i, p_i, p_{ni}, \varphi_1, \varphi_2)$, cujos parâmetros apresentados no artigo em análise nesta secção, se basearam em informação externa, na maioria dos casos facultada em relatórios produzidos pelo Instituto Dr. Ricardo Jorge.

Um método de *Sequential Importance Sampling with Resampling* foi utilizado para realizar as integrações necessárias.

Da análise da Figura 2, que mostra as séries de estimativas produzidas para as componentes do processo n_t , juntamente com as respectivas observações nos casos em que elas se referem às populações incluídas no GRIPENET, pode concluir-se que o ajustamento é razoável, excepto na parte final das séries, problema que pode estar relacionado com o método de estimação utilizado ou com o facto de no princípio e no fim de período de vigilância o número de participantes no GRIPENET ser consideravelmente menor.

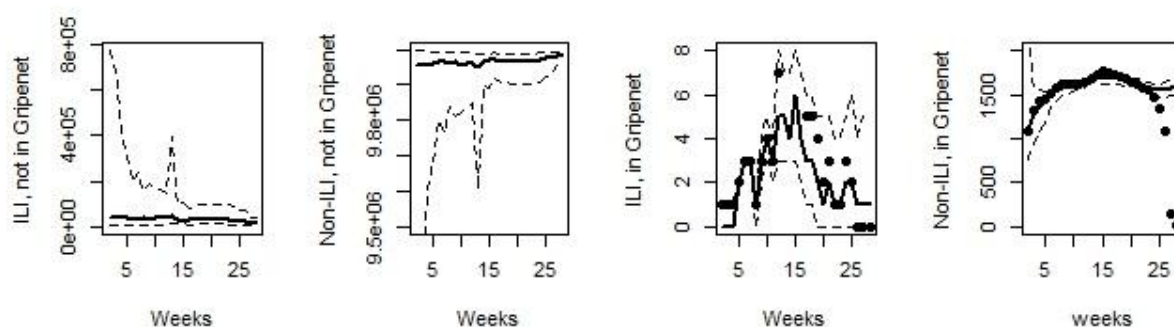


Figura 2: Estimativas dos estados, quartis (tracejado) e observações (pontos) – da esquerda para a direita, populações ILI e não ILI fora da GRIPENET e populações ILI e não ILI incluídas na GRIPENET.

Faz-se notar ainda, que uma estimativa da curva de incidência de gripe durante esta época pode ser facilmente calculada através da expressão, $(n_{i,t}(\overline{GN}) + n_{i,t}(GN))/N$, e comparada com a fornecida por outros sistemas de vigilância, nomeadamente a produzida pela rede de médicos sentinela.

Como conclusão desta secção, referimos que, embora a aplicação apresentada sugira que os SSMs podem ser um instrumento valioso na extrapolação dos resultados obtidos por sistemas de vigilância sindrómica, os modelos utilizados terão que ter uma estrutura bastante complexa para se corrigir eficazmente os efeitos da falta de representatividade seletiva das diferentes classes demográficas, típica destes sistemas de vigilância.

Dado ainda o elevado tempo de cálculo, parcialmente causado pela metodologia de estimação utilizada, recomenda-se que a aplicação destes modelos neste contexto da vigilância sindrómica se apoie num método de simulação mais expedito.

Agradecimentos

A investigação que conduziu aos resultados aqui referidos foi parcialmente suportada por fundos nacionais através da Fundação Nacional para a Ciência e Tecnologia, Portugal–FCT no âmbito do projecto PEst-OE/MAT/UI0006/2011.

Os autores gostariam também de agradecer ao Laboratório Nacional de Referência para o Vírus da Gripe do Instituto Nacional de Saúde Dr. Ricardo Jorge e à Rede Médicos Sentinela pelo acesso aos dados utilizados nesta investigação.

Bibliografia

- [1] Lui K.-J. and Kendal A.P. (1998). Impact of influenza epidemics on mortality in the United States from October 1972 to May 1985. *American Journal of Public Health*, **158**: 645-650.
- [2] Simonsen L., Clarke M.J., Williamson D., Stroup D.F., Arden N.H. and Schonberger L.B. (1997). The impact of influenza vaccination on seasonal mortality in the US elderly population. *American Journal of Public Health*, **87**(12): 1994-1950.
- [3] Simonsen L., Reichert T.A., Viboud C., Blackwelder W.C., Taylor R.J. and Miller M.A. (2005). The Impact of Influenza Epidemics on Mortality: Introducing a Severity Index. *Archives of Internal Medicine*, **165**: 265-272.
- [4] Zucs W.H.P., Buchholz U. and Uphoff H. (2005). Influenza associated excess mortality in Germany, 1985 – 2001. *Emerging Themes in Epidemiology*, **2**(6): 1-9.
- [5] Rizzo C., Viboud C., Montomoli E., Simonsen L. and Miller M.A. (2006). Influenza related mortality in the Italian elderly: No decline associated with the increasing vaccination coverage. *Vaccine*, **24**: 6468-6475.
- [6] Choi K. and Thacker S.B. (1981). An evaluation of influenza mortality surveillance 1962-1979. *American Journal of Epidemiology*, **113** 3: 215-216.

- [7] Antunes J.L., Waldman E.A., Borrell C. and Paiva T.M. (2006). Effectiveness of influenza vaccination and its impact on health inequalities. *International Journal of Epidemiology*, **36**(6): 1319–1326.
- [8] Nunes B., Natário I. and Carvalho M.L. (2011). Time series methods for obtaining excess mortality attributable to influenza epidemics. *Statistical Methods in Medical Research*, **20**(4): 331–345.
- [9] Sonesson C. and Bock D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistics Society, Series A*, **166**: 5–21.
- [10] Farrington C.P. and Andrews N. (2004). Outbreak detection: Applications to infectious disease surveillance. In: Brookmeyer R., Stroup D.F., editors. *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford: Oxford University Press, p. 203–231.
- [11] Buckridge D.L., Burkom H.S., Campell M., Hogan W.R. and Moore A. (2005). Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, **38**: 99–113.
- [12] Unkel S., Farrington C.P., Garthwaite P.H., Robertson C. and Andrews N. (2010 [cited 2011 Jun]). Statistical methods for the prospective detection of infectious disease outbreaks: a review[Internet]. Milton Keynes, UK: The Open University; 53p. Available from: <http://statswww.open.ac.uk/TechnicalReports/OutbreakReviewPaper.pdf>.
- [13] Nicoll A., Ammon A., Amato-Gauci A., Ciancio B., Zucs P., Devaux I., Plata F., Mazick A., Milbak K., Asikainen T. and Kramarz P. (2010). Experience and lessons from surveillance and studies of the 2009 pandemic in Europe. *Public Health*, **124**(1): 14–23.
- [14] Nunes B., Natário I. and Carvalho M.L. Nowcasting influenza epidemics using non-homogeneous hidden Markov Models. *Statistics in Medicine*, aceite para publicação em 2012.
- [15] Strat L and, Carrat F. (1999). Monitoring epidemiologic surveillance data using Hidden Markov Chains models. *Statistics in Medicine*, **18**: 3463–3478.
- [16] Rath T.M., Carreras M. and Sebastiani P. (2003). Automated Detection of Influenza Epidemics. University of Massachusetts.
- [17] Martinez-Beneito M.A., Conesa D., Lopez-Quilz A. and Lopez-Maside A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, **27**: 4455–4468.
- [18] Zucchini W. and MacDonald I.L. (2009). Hidden Markov Models for Time Series: An introduction using R, Boca Raton: Chapman and Hall/CRC; 275p.
- [19] Paroli R. and Spezia L. (2008). Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. *Computational Statistics and Data Analysis*, **52**: 2311–2330.
- [20] Chib S. (1996). Calculating posterior distributions and modal estimates in Markov mixtures models. *Journal of Econometrics*, **75**: 79–97.
- [21] O'Connor M.J., Buckeridge D., Choy M.K., Crubezy M., Pincus Z. and Musen M.A. (2003). BioSTORM: A System for Automated Surveillance of Diverse Data Sources. *AMIA Annual Symposium Proceedings*.
- [22] Reis B.Y., Kirby C., Hadden L.E., Olson K., McMurry A.J., Daniel J.B. and Mandl K.D. (2007). AEGIS: A robust and scalable real-time public health surveillance system. *Journal of the American Medical Informatics Association*, **14**: 581–588.
- [23] Lombardo J., Burkom H., Elbert E., Magruder S., Lewis S.H., Loschen W., Sari J., Sniegowski C., Wojcik R. and Pavlin J. (2003). A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). *Journal of Urban Health*, **80** (2Suppl 1): i32–i42.
- [24] Brownstein J.S., Freifeld C.C., Reis B.Y. and Mandl K.D. (2008). Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the Health Map Project. *PLoS Med*, **5**(7): e151.
- [25] van Noort S.P., Muehlen M., Rebelo de Andrade H., Koppeschaar C., Lima Lourenço J.M. and Gomes M.G. (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. *Euro Surveillance*, **12**(7): pii=722.

- [26] Natário I. and Carvalho M.L. (2009). Addressing the problem of lack of representativeness on syndromic schemes. *Discussion es Mathematicae - Probability and Statistics*, **29**:169-183.
- [27] Durbin J. and Koopman S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.
- [28] Buckland S.T., Newman K.B., Thomas L. and Koesters N.B. (2004). State space models for the dynamics of wild animal populations. *Ecological Modelling*, 171: 157-175.
- [29] Thomas L., Buckland S.T., Newman K.B. and Harwood J. (2005). A unified framework for modelling wild population dynamics. *Australian New Zealand Journal Statistics*, **47**:1934.
- [30] Newman K.B., Buckland S.T., Lindley S.T., Thomas L. and Fernández C. (2006). Hidden process models for animal population dynamics. *Ecological Applications*, **16**: 74-86.
- [31] Caswell H. (2001). *Matrix Population Models*. 2nd Edition. Sinauer Associates, Inc. Publishers.



Tese de Doutoramento: Séries Temporais de Memória Longa com Aplicações ao Controlo Motor – Estudo de Tarefas de Tapping Repetido

(Boletim SPE Primavera de 2009, p.107)

Ana Diniz, *adiniz@fmh.utl.pt*

*Secção de Métodos Matemáticos,
Faculdade de Motricidade Humana, Universidade Técnica de Lisboa*

Caros colegas e amigos,

Efetuei as minhas provas públicas de doutoramento na Faculdade de Motricidade Humana da Universidade Técnica de Lisboa no dia 3 de Dezembro de 2008. Foi um dia muito especial, compartilhado pela minha família e colegas, e que marcou o fim de um longo percurso. É curioso como o empreendimento de tantos dias, semanas e meses se transforma de repente em algo efémero e que se resume em cerca de duas horas. A elaboração deste estudo passou por uma série de oscilações (talvez de memória longa?!), envolvendo o tema de análise e o tratamento estatístico, numa procura constante de equilíbrio entre aplicação e teoria. Houve momentos muito difíceis, mas tive sempre ao meu lado a minha fantástica família e o meu orientador (Prof. Nuno Crato) e co-orientador (Prof. João Barreiros). Olhando para trás, sem a pressão do tempo e da obrigação, confesso que gostei muito do projeto e de todos os conhecimentos e experiências que adquiri durante a sua realização. Foi um investimento significativo, a nível profissional e também pessoal, mas que trouxe grandes mais-valias para a minha carreira e investigação científica.

A tese no ramo de Motricidade Humana e na especialidade de Métodos Matemáticos incidiu sobre a descrição e a modelação de séries temporais relativas a intervalos inter-resposta em determinadas tarefas motoras repetitivas. Mais concretamente, propôs-se um novo modelo teórico com uma interpretação biológica para explicar o comportamento dos intervalos inter-resposta em tarefas rítmicas de batimento com um dedo (*tapping*). O delineamento experimental foi o de sincronização-continuação, ou seja, cada participante devia pressionar um interruptor em simultâneo com sinais sonoros (com frequências de 1.250 Hz (800 ms) e de 0.625 Hz (1600 ms)) e, em seguida, continuar a pressionar com o mesmo ritmo sem a referência sonora. Os valores visados foram os intervalos de tempo entre instantes de respostas consecutivas da fase de continuação. A análise estatística das séries obtidas sugeriu a eventual existência de oscilações longas não periódicas. As funções de autocorrelação amostral assumiram apenas valores positivos com decaimento lento e os periodogramas normalizados evidenciaram valores maiores nas frequências baixas. Estas propriedades em conjunto com alguns estudos prévios suscitaram o interesse em explorar modelos de memória longa. A escolha recaiu sobre um modelo aditivo envolvendo um processo de mudança de regime com durações de regimes de cauda pesada.

A área das Séries Temporais aliada ao domínio da Motricidade Humana tem vindo a desenvolver-se gradualmente, dado o conjunto considerável de dados disponíveis e a necessidade crescente de encontrar respostas precisas para problemas reais. Alguns exemplos interessantes dizem respeito a séries de erros de execução em tarefas motoras, séries de velocidades em provas desportivas, séries de frequências cardíacas ou de sinais eletromiográficos em exercícios físicos, etc. Os métodos clássicos de análise de dados temporais relativos à motricidade humana são baseados em medidas descritivas, como a média e o desvio padrão, e ignoram a estrutura de correlação ao longo do tempo. Em contraste, os métodos das séries temporais, no domínio tempo e no domínio frequência, consideram o comportamento dinâmico ao longo do tempo e permitem fazer modelação e inferência. A literatura mais recente sobre este tema contém alguns estudos científicos com uma componente de modelação. No entanto, esses trabalhos são ainda bastante limitados e as metodologias estatísticas necessitam de grande refinamento.

A colaboração com colegas da Faculdade de Motricidade Humana da Universidade Técnica de Lisboa, onde trabalho há 15 anos, intensificou-se bastante durante e depois do doutoramento. De facto, a investigação realizada para a tese revelou um conjunto de questões ligadas ao controlo motor e à fisiologia extremamente interessantes do ponto de vista aplicado e teórico. Os resultados obtidos permitiram dar resposta rigorosa a alguns problemas pendentes e tirar algumas conclusões relevantes. Por um lado, as diferenças entre certos indivíduos foram notórias, o que sugere que cada sujeito deve ser tratado separadamente e os resultados médios devem ser evitados. Por outro lado, apesar da heterogeneidade individual, os parâmetros estimados exibiram uma estabilidade considerável e sugeriram existência de memória longa entre outras propriedades. Todos estes resultados foram discutidos em congressos/seminários nacionais e internacionais e foram publicados em revistas de controlo motor e de estatística. Uma certa dificuldade que encontrei na publicação foi o facto de ser considerada por vezes demasiado “teórica” pelos investigadores da motricidade e demasiado “aplicada” pelos investigadores da estatística.

A componente de interagir com colegas da Faculdade de Motricidade Humana e de outras Instituições no pós doutoramento tem sido muito enriquecedora em termos pessoais e profissionais. Parece que de certa forma o meu contributo passou a ter um valor mais significativo. Mas esta atividade implica naturalmente uma carga extra para além da lecionação das aulas de licenciatura, de mestrado e de doutoramento e pressupõe um investimento contínuo em formação e pesquisa. Na realidade, os problemas que me têm sido referidos necessitam muitas vezes de metodologias estatísticas na área das séries temporais (e outras) que estão fora do meu domínio imediato. Todavia, e face às dificuldades económicas do país e do meio universitário, é fundamental intervir em trabalhos de outros colegas e tentar produzir projetos interdisciplinares de grande qualidade. Cada vez mais é essencial a colaboração dos investigadores da estatística com investigadores de outras áreas para efetuar corretamente o delineamento experimental, o tratamento dos dados e a interpretação dos resultados e obter então respostas fiáveis.

O texto precedente mostra claramente que o doutoramento conduziu a uma nova fase da vida com mais regalias mas também com mais responsabilidades e solicitações. E no meu caso particular, o pós doutoramento coincidiu ainda com a fantástica mas absorvente experiência da maternidade. Posto isto, eu gostaria no futuro de encontrar uma forma de tornar o tempo mais elástico! Enfim, é uma fase da vida em que se torna indispensável criar prioridades e encontrar um equilíbrio entre lecionação, investigação, família e lazer.

Saudações académicas

Ana Diniz

Tese de Doutoramento: Contributos para o Estudo de Dados em Falta

(Boletim SPE Primavera de 2009, p.108)

Maria Fernanda Diamantino, *mfdiamantino@fc.ul.pt*

DEIO – Faculdade de Ciências da Universidade de Lisboa

Caros colegas e amigos:

Ao receber o convite do Professor Fernando Rosado para escrever sobre a minha atividade profissional a seguir às provas de doutoramento, respondi prontamente que o faria com todo o gosto e levei este trabalho de casa para as minhas férias (à beira-mar foi surgindo a inspiração para a escrita...).

O meu testemunho tem como guião, por um lado, o meu gosto em ensinar, enquadrando-se nesta vertente muito do que se relaciona com a docência e, por outro lado, o meu gosto pela ciência estatística enquanto ferramenta indispensável às outras ciências.

As minhas provas de doutoramento realizaram-se no dia 23 de Dezembro de 2008, o que fez delas um delicioso presente de Natal. Não vou falar do período que antecedeu as provas, não é esse o objetivo destas palavras e todos os que por lá já passaram sabem, cada um a seu modo, o “sabor” desse tempo. Naquele dia, ao sentir que uma tarefa tinha terminado, também senti que novas tarefas se iniciariam.

Uma delas diz respeito às novas responsabilidades inerentes a um professor doutorado, entre elas a regência de unidades curriculares e a orientação de alunos de mestrado. Aqui entra o meu gosto pelo ensino. Lecionar Probabilidade e Estatística a alunos que fazem a sua formação em outras áreas (muitas das disciplinas que tenho lecionado pertencem ao plano curricular das várias licenciaturas existentes na FCUL) é um desafio e uma responsabilidade. O desafio passa pela capacidade de motivar este tipo de alunos, de manter o seu interesse e aguçar o seu espírito crítico, meio caminho andado para o sucesso académico. A responsabilidade passa pela transmissão de conhecimentos que pretendemos virem a ser cuidadosamente utilizados na sua vida futura, se for esse o caso. Refiro-me, principalmente, aos alunos ligados às áreas da saúde e da medicina, áreas que tanto necessitam desta ferramenta e do uso cuidado da mesma. A noção desta responsabilidade aumentou ao lecionar uma unidade curricular de probabilidade e estatística da licenciatura em Cardiopneumologia na Escola Superior de Saúde da Cruz Vermelha Portuguesa (ao abrigo de protocolo com a FCUL) em 2009.

Além da continuação desta colaboração, no ano passado recebi a proposta de regência de uma unidade curricular no Mestrado em Cardiopneumologia a iniciar este ano letivo.

Ao propor, pela primeira vez, temas de Mestrado, além de um tema no seguimento do trabalho da minha tese, propus também temas em áreas diferentes. Na altura fui desafiada por uma colega de outro departamento da FCUL para uma coorientação que me proporcionou algo relacionado com o outro meu gosto: trabalhar em aplicações da estatística; com ela iniciei investigação na análise de dados hidrogeológicos.

Por outro lado, e ainda no âmbito das aplicações, tenho desenvolvido trabalho no tratamento de dados na área da Psicologia com o responsável pelo Gabinete de Apoio Psicopedagógico da FCUL, com quem tenho trabalhado de perto por ser a representante do meu departamento na Comissão de Acompanhamento dos Alunos com Necessidades Educativas Especiais. Este é um desafio diferente! Para mim, são alunos especiais que tenho ajudado de forma que a sua passagem pela FCUL seja a melhor possível de acordo com as limitações de cada um. Alguns deles são um bom exemplo de coragem e determinação.

Ao terminar estas breves palavras, não posso deixar de realçar o quanto tenho aprendido, nestes últimos anos, enquanto docente e investigadora. Esta é uma profissão com grandes desafios mas também com situações muito gratificantes relacionadas quer com resultados de investigação, quer com o sucesso dos nossos alunos. O reconhecimento que vou recebendo da parte deles é a compensação do que julgo ser uma das tarefas nobres e importantes de um professor: ensinar. Ainda no que diz respeito à docência, sendo a população a maior riqueza de um país, cabe-nos a nós, docentes universitários, o grande empreendimento numa boa formação e instrução das gerações futuras.

Até breve,

Fernanda Diamantino



Tese de Doutoramento: Probabilidade Cerca 1914 e a Construção de Pacheco de Amorim

(Boletim SPE Primavera de 2009, p.108)

Rui Santos, *rui.santos@ipleiria.pt*

*Escola Superior de Tecnologia e Gestão – Instituto Politécnico de Leiria
Centro de Estatística e Aplicações da Universidade de Lisboa*

A história da Teoria da Probabilidade em Portugal, tema central da minha tese de doutoramento (resumo disponível no Boletim da SPE de Primavera de 2009), nomeadamente no que se refere às dificuldades encontradas no início do século XX na procura de uma fundamentação sólida para o Cálculo das Probabilidades e para a Estatística, não permite uma investigação profícua em Portugal, consequência do reduzido número de obras originais portuguesas nesta área anteriores a 1950. Assim sendo, apesar de continuar a investir nesta temática, tentando divulgar as poucas obras lusitanas que há muito procuraram contribuir para o desenvolvimento da Probabilidade e da Estatística, nomeadamente a tese de doutoramento de Diogo Pacheco d'Amorim (1914), procurei descobrir novos temas de investigação. Perante tal situação, e uma vez que me foi solicitado escrever para a secção Pós-doc, irei tentar descrever de forma resumida os novos desafios de investigação que encontrei nestes primeiros três anos após a conclusão do doutoramento.

Na decorrência destes factores, o tema central da minha investigação durante os últimos anos tem sido, como não poderia deixar de o ser, proposto pelo meu (ainda) orientador Professor Dinis Pestana. Este é simultaneamente o tema central da tese de doutoramento que o nosso aluno Ricardo Sousa (Escola Superior de Tecnologia da Saúde de Lisboa, Instituto Politécnico de Lisboa) está a desenvolver, investigação esta em que temos ainda contado com a colaboração dos colegas João Paulo Martins e Miguel Felgueiras, ambos do Instituto Politécnico de Leiria (IPLeiria). O problema está relacionado com as análises conjuntas ou testes compostos (testes efetuados usando um sangue combinado, que é uma mistura de sangue de n indivíduos), uma teoria que na estatística foi introduzida por Dorfman (1943) com o objetivo de determinar a dimensão n ótima para cada grupo em função da taxa de prevalência p , no sentido de minimizar o número esperado de testes necessários para a identificação de todos os soldados americanos infetados com sífilis durante a segunda guerra mundial. Na metodologia proposta por Dorfman é efetuado um teste conjunto a cada grupo. Se o teste conjunto for negativo então todos os indivíduos que pertencem ao grupo são classificados como negativos (não infetados); caso contrário, significará que pelo menos um dos elementos do grupo está infetado e, consequentemente, ter-se-ão de efetuar testes individuais a todos os elementos desse grupo para identificar que elementos estão infetados. O problema inicial, que considera inexistência de erros nos resultados dos testes e unicamente a realização de testes qualitativos (identificação da presença ou da ausência de uma qualquer substância no líquido composto analisado), é extremamente simples de formalizar e resolver. Posteriormente novos algoritmos foram propostos com o objetivo de minimizar o número de testes necessário para a correta classificação de todos os indivíduos da população, como, por exemplo, as propostas de Sterret (1957), Sobel e Groll (1959), Finucan (1964) ou, mais recentemente, Kim *et al.* (2007).

A extensão deste problema a outras situações mais gerais e próximas da realidade pode gerar problemas extremamente apelativos e de resolução mais complexa. Uma das hipóteses que, desde

logo, devemos relaxar é a ausência de possibilidade de ocorrência de erro na realização dos testes (erros de classificação), recorrendo aos usuais conceitos de sensibilidade (probabilidade de um teste positivo num indivíduo infetado) e especificidade (probabilidade de um teste negativo num indivíduo saudável) para modelar cada situação. Por outro lado, esta metodologia pode ser igualmente estendida a testes quantitativos, onde a quantidade de uma qualquer substância que pretendemos analisar no sangue (e.g. o número de determinada bactéria) é descrita por uma variável aleatória X caracterizada por uma qualquer distribuição \mathbf{D} com vetor de parâmetros $\boldsymbol{\theta}$. Nestes casos estaremos a testar se X é superior (ou inferior) a um determinado limite l (que separa os testes positivos dos negativos). Por outro lado, o recurso a análises conjuntas não visa unicamente a classificação de indivíduos, podendo ser igualmente útil na estimação da taxa de prevalência, conforme Sobel e Elashoff (1975) o demonstraram, não sendo necessário, para este fim, a realização de posteriores testes individuais (uma vez que não se pretende identificar os indivíduos infetados, mas somente estimar a taxa de prevalência da doença, tendo ainda a vantagem de manter o anonimato dos indivíduos infetados uma vez que estes não são identificados). Deste modo, os estimadores obtidos pela aplicação de testes compostos têm, sob determinadas condições, melhor comportamento que os estimadores tradicionais baseados em testes individuais, cf. Sobel e Elashoff (1975), Loyer (1983) e Garner *et al.* (1989). O enviesamento, a eficiência e a robustez destes estimadores foram examinados em diversos artigos, tais como os de Chen e Swallow (1990), Lancaster e Keller-McNulty (1998) ou Hung e Swallow (1999). Assim sendo, os estimadores baseados em testes conjuntos permitem não só a obtenção de ganhos monetários (diminuindo o número de testes efetuados) mas igualmente a obtenção de estimativas mais precisas em comparação com as obtidas com base em testes individuais. Por fim, as suas aplicações não são restritas a análises sanguíneas, podendo ser aplicadas à análise de qualquer líquido (e não só) para o qual amostras provenientes de distintos indivíduos podem ser misturadas de forma homogénea, de tal forma que seja possível posteriormente retirar uma amostra conjunta para análise (consulte-se, a título de exemplo, a vasta lista de aplicações e referências bibliográficas proposta por Boswell *et al.*, 1996). Entre estas aplicações sublinhamos a potencial utilidade desta metodologia na área do controlo de qualidade, nomeadamente na amostragem para aceitação de, por exemplo, produtos alimentares.

Por conseguinte, começamos por estender os conceitos de sensibilidade e especificidade para a realização de testes conjuntos. A maioria dos trabalhos em análises conjuntas considera que as probabilidades associadas aos erros de classificação são iguais nos testes individuais e nos testes conjuntos, como por exemplo Tu *et al.* (1994), Tu *et al.* (1995), Kim *et al.* (2007) e Liu *et al.* (2011). Alguns autores consideram que estas probabilidades são distintas nos testes conjuntos, mas não dependem do número de indivíduos infetados dentro do grupo (Hwang, 1976); outros consideram modelos bastante simplificados de forma a incluir o número de indivíduos infetados em cada grupo na modelação destas medidas (Hung e Swallow, 1999) e, em algumas aplicações, a sensibilidade depende do número de indivíduos infetados no grupo, mas não é apresentada uma formalização desta condicionante, como por exemplo em Wein e Zenios (1996) e Zenios e Wein (1998). Na nossa proposta de modelação da sensibilidade e especificidade incluímos o problema da diluição e consequente rarefação, uma vez que o valor da sensibilidade será condicionado ao número de elementos infetados no grupo (se misturarmos o sangue de um indivíduo infetado com o de muitos indivíduos não infetados provavelmente a análise conjunta não irá conseguir identificar o sangue infetado devido à diluição e consequente rarefação das características que distinguem um sangue infetado de um que não o é). Sob estas hipóteses modelamos a sensibilidade e especificidade para testes conjuntos quantitativos em função da distribuição \mathbf{D} e, com base nestas, para a metodologia de classificação de Dorfman (1943), cf. Santos, R., Pestana, D. e Martins, J.P. (2012). Fórmulas para as principais distribuições discretas de probabilidades (Johnson *et al.*, 2005) foram igualmente deduzidas, nomeadamente para a família das distribuições de Katz-Panjer (distribuições cuja função de probabilidade pode ser descrita através de uma fórmula recursiva), Panjer (1981), e suas extensões, Sundt e Jewell (1981), Willmot (1987), Hess *et al.* (2002) e Pestana e Velosa (2004). Para tal o número

de bactérias na amostra composta, necessário para o cálculo da especificidade e sensibilidade em testes compostos, foi determinado recorrendo a modelos hierárquicos.

Salientemos que a realização de testes conjuntos quantitativos pretende identificar se algum dos indivíduos do grupo está infetado. Consequentemente, o objetivo da análise conjunta é identificar se o máximo (ou o mínimo) entre os indivíduos pertencentes ao grupo é superior (inferior) a determinado limiar l . Como tal, pretendemos realizar testes relacionados com o máximo de um grupo tendo como única informação a média desse grupo (uma vez que a única medida observada do grupo corresponde à média). Para analisar a qualidade de testes de hipóteses assim definidos, foram efetuadas simulações, para distintas distribuições \mathbf{D} , para medir a quantidade de informação que a média de um grupo contém acerca do máximo desse mesmo grupo (por exemplo o coeficiente de correlação entre o máximo e a média amostral para distintas distribuições \mathbf{D}), bem como a comparação dos resultados obtidos, medidos através da especificidade e sensibilidade, na aplicação de diferentes metodologias na realização do teste conjunto. Sublinhemos que unicamente em distribuições de caudas pesadas os resultados são aceitáveis, uma vez que a probabilidade de erro de classificação nas restantes distribuições é elevado, mesmo para grupos de dimensão pequena. Os referidos resultados foram já apresentados em comunicações e serão fruto de publicação brevemente.

No que se refere à estimação da taxa de prevalência, analisamos, via simulação, a qualidade de estimadores, pontuais e intervalares, que incluam sensibilidade e especificidade, procurando melhorar os existentes na literatura da área, tais como os estimadores disponíveis no package *binGroup* para o *R* (Bildler *et al.*, 2010), ou em alguns dos estimadores baseados nos propostos em Pires e Amado (2008). Seguindo uma metodologia semelhante a Finucan (1964), foram igualmente analisadas algumas sugestões para lidar com a existência de subpopulações com diferentes taxas de prevalência.

Pelo facto de cada laboratório utilizar diferentes metodologias para obter estimativas para a taxa de prevalência (aplicação de diferentes algoritmos de classificação, diferentes dimensões para cada grupo, realização ou não de testes individuais para a identificação do elementos infetados) há, por vezes, a necessidade de juntar diferentes estimativas provenientes de laboratórios distintos para se obter uma estimativa conjunta mais precisa. Chen e Swallow (1990) propõem na utilização de regressão logística ou do método dos mínimos quadrados ponderados na incorporação das diferentes estimativas. Nós investigamos o recurso a técnicas de meta-análise (Kulinskaya *et al.*, 2008) como uma metodologia alternativa, que consideramos ser mais flexível, de forma a poder incorporar na estimação os possíveis erros de testes (medidos pela sensibilidade e especificidade). Os estimadores por nós propostos estão em fase de análise à qualidade da estimativas produzidas, via simulação, nomeadamente à sua robustez perante problemas de má especificação da especificidade e sensibilidade dos testes.

Outro aspeto que se alterou após o doutoramento foi a possibilidade de poder lecionar disciplinas, da área da Estatística, em mestrados, nomeadamente da área das Finanças e Gestão. Entretanto no IPEiria ministramos igualmente um curso de Mestrado em Educação e Tecnologia em Matemática, no qual tive a oportunidade de lecionar, em conjunto com a colega Helena Ribeiro, a disciplina de Probabilidades e Análise de Dados e, posteriormente, a orientação de duas estudantes. Uma, Salomé Pedro, que orientei conjuntamente com o colega Luís Cotrim (IPEiria), desenvolveu um trabalho intitulado *Modelação e Simulação – Uma Aplicação ao Problema da Ruína do Jogador*. Este trabalho contém dois capítulos dedicados a equações às diferenças de primeira e segunda ordem onde são expostas diversas aplicações; apresenta uma resenha histórica da origem do cálculo das probabilidades, focando detalhadamente o aparecimento do problema da ruína do jogador e das suas diferentes versões; oferece algumas soluções exatas para algumas dessas variantes com recurso à modelação do problema através de equações às diferenças; propõe ainda metodologias de obtenção de soluções aproximadas do problema da ruína do jogador, recorrendo à simulação Monte Carlo, via software *R*, e aos teoremas de convergência, exemplificando a aplicabilidade da Lei dos Grandes Números, da Lei do Logaritmo Iterado, Teorema Limite Central e Teorema de Glivenko-Cantelli. O segundo trabalho que tive oportunidade de orientar foi elaborado pela Alice Martins, que orientei em

colaboração com a colega Helena Ribeiro (IPLeiria), tendo o título de *Estatística no Ensino Secundário e Básico*. Neste trabalho foi efetuada uma análise crítica a diversos manuais, elaborada uma *sebenta* que expõe todos os conceitos de Estatística focados nestes níveis de ensino, e apresentadas propostas de tarefa a serem implementadas em sala de aula recorrendo ao software *GeoGebra*, onde são sobretudo explorados os ambientes dinâmicos deste programa, em particular no que se refere à regressão linear. Estes dois trabalhos apresentaram as suas principais conclusões no Congresso Anual da SPE 2011 e prevê-se que sejam fruto de publicação, em revistas ligadas ao ensino, brevemente.

Apesar da brevidade deste texto, esperamos que estas linhas consigam retratar, pelo menos de forma aproximada, alguns dos novos desafios com que deparei nos primeiros três anos de investigação pós doutoramento.

Rui Santos

- Bilder, C.R., Zhang, B., Schaarschmidt, F., Tebbs, J. (2010). binGroup: a package for group testing, *The R Journal* **2**, 56-60.
- Boswell, M.T., Gore, S.D., Lovison, G., Patil, G.P. (1996). Annotated bibliography of composite sampling, Part A: 1936-92, *Environmental and Ecological Statistics* **3**, 1-50.
- Chen, C.L., Swallow, W.H. (1990). Using group testing to estimate a proportion, and to test the binomial model, *Biometrics* **46**, 1035-1046.
- Dorfman, R. (1943). The detection of defective members of large populations, *Ann. Math. Statist.* **14**, 436-440.
- Finucan, H.M. (1964). The blood testing problem. *Applied Statistics* **13**, 43-50.
- Garner, F.C., Stapanian, M.A., Yfantis, E.A., Williams, L.R. (1989). Probability estimation with sample compositing techniques, *Journal of Official Statistics* **5**, 365-374.
- Hwang, F.K. (1976). Group testing with a dilution effect, *Biometrika* **63**, 671-673.
- Hess, K.T.; Liewald, A. e Schmidt, K.D. (2002). An extension of Panjer's recursion, *Astin Bulletin* **32**, 283-297.
- Hung, M., Swallow, W.H. (1999). Robustness of group testing in the estimation of proportions, *Biometrics* **55**, 231-237.
- Johnson, N.L.; Kemp, A.W. and Kotz, S. (2005), *Univariate Discrete Distributions*, 3rd Edition, John Wiley & Sons.
- Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, C. (2007). Comparison of group testing algorithms for case identification in the presence of testing errors, *Biometrics* **63**, 1152-1163.
- Kulinskaya, E., Morgenthaler, S., Staudte, R.G. (2008). *Meta Analysis: a Guide to Calibrating and Combining Statistical Evidence*, Wiley, Chichester, England.
- Lancaster, V.A., Keller-McNulty, S. (1998). A review of composite sampling methods, *Journal of the American Statistical Association* **93**, 1216-1230.
- Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., Yang, T.C. (2011). Cost analysis in choosing group size when group testing for potato virus Y in the presence of classification errors, *Ann. Appl. Biology* **159**, 491-502.
- Loyer, M.W. (1983). Bad probability, good statistics, and group testing for binomial estimation, *The American Statistician* **37**, 57-59.
- Martins, M.A. (2012). *Estatística no Ensino Básico e Secundário*, mestrado em Educação e Tecnologia em Matemática, Instituto Politécnico de Leiria.
- Pacheco d'Amorim, D. (1914). *Elementos de Cálculo das Probabilidades*, Tese de doutoramento, Imprensa da Universidade de Coimbra, Coimbra. (A tradução para inglês, juntamente com original, editado por S. Mendonça, D. Pestana e R. Santos está disponível em <http://www.estg.ipleiria.pt/~rui.santos>)
- Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions, *Astin Bulletin* **12**, 22-26.
- Pedro, S. (2012). *Modelação e Simulação - Uma Aplicação ao Problema da Ruína do Jogador*, mestrado em Educação e Tecnologia em Matemática, Instituto Politécnico de Leiria.
- Pestana, D.D., Velosa, S.F. (2004). Extensions of Katz-Panjer families of discrete distributions, *REVSTAT* **2**, 145-162.
- Pires, A.M., Amado, C. (2008). Interval estimators for a binomial proportion: comparison of twenty methods, *REVSTAT* **6**, 165-197.
- Santos, R., Pestana, D., Martins, J.P. (2012). Extensions of Dorfman's theory, *Selected Papers of SPE 2010* (aceite).
- Santos R. (2008). *Probabilidade Circa 1914 e a Construção de Pacheco d'Amorim*, Tese de Doutoramento, Universidade de Lisboa.

- Sobel, K.M., Elashoff, R.M. (1975). Group testing with a new goal, estimation, *Biometrika* **62**, 181-193.
- Sobel, M., Groll, P.A. (1959) Group testing to eliminate efficiently all defectives in a binomial sample, *Bell System Technical Journal* **38**, 1179-1252.
- Sundt, B., Jewell, W.S. (1981). Further results on recursive evaluation of compound distributions, *Astin Bulletin* **12**, 27-39.
- Sterret, A. (1957). On the detection of defective members of large populations, *Ann. Math. Statistics* **28**, 1033-1036.
- Tu, X.M., Litvak, E., Pagano, M. (1994). Studies of AIDS and HIV surveillance, screening tests: can we get more by doing less?, *Statistics in Medicine* **13**, 1905-1919.
- Tu, X.M., Litvak, E., Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening, *Biometrika* **82**, 287-297.
- Wein, L.M., Zenios, S.A. (1996). Pooled testing for HIV screening: capturing the dilution effect, *Operations Research* **44**, 543-569.
- Willmot, G.E. (1987). Sundt and Jewell's family of discrete distributions, *Astin Bulletin* **18**, 17-29.
- Zenios, S., Wein, L. (1998). Pooled testing for HIV prevalence estimation exploiting the dilution effect, *Stat. Med.* **17**, 1447-1467.



• Artigos Científicos Publicados

- Bispo, R., Marques, T. A. & Pestana, D. (2012). Statistical power of goodness-of-fit tests based on the empirical distribution function for Type I right censored data. *Journal of Statistical Computation and Simulation*, 82:173-181
- Coelho, P.S. e Pereira, L.N. (2011). A Spatial Unit Level Model for Small Area Estimation. *REVSTAT Statistical Journal*, 9:2, 155-180.
- Ferreira, L.N.; Ferreira, P.L.; Pereira, L.N.; e Rowen, D. (2012). Reducing the Floor Effect in the SF-6D: A Feasibility Study. *Applied Research in Quality of Life*, 7:2, 193-208.
- Figueiredo, F., M. I. Gomes, L. Henriques-Rodrigues, M. C. Miranda (2012). A computational study of a quasi-PORT methodology for VaR based on secondorderreduced-bias estimation. *Journal of Statistical Computation and Simulation*, Vol.82 n° 4, 587-602.
- Figueiredo, A., F. Figueiredo, N. Monteiro, Straume Odd (2012). Restructuring in privatised firms: a Statis approach. *Structural Change and Economic Dynamics*, Vol.23, 108-116.
- Grilo, L. M. e Coelho, C. A. (2012). A family of near-exact distributions based on truncations of the exact distribution for the generalized Wilks Lambda statistic. *Journal of Communications in Statistics - Theory and Methods*, **41**: 1-21.
- Maroco, J.; Silva, D.; Rodrigues, A.; Guerreiro, M.; Santana, I.; Mendonça, A. De (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests. *BMC Research Notes* 4 (1): 299 – 313.
- Mendes, S. (2012). An empirical comparison of Canonical Correspondence Analysis and STATICO in the identification of spatio-temporal ecological relationships. *Journal of Applied Statistics*, Vol. 39, No. 5, 979–994
- Pereira, L.N. e Coelho, P.S. (2012). A Small Area Predictor under Area-Level Linear Mixed Models with Restrictions. *Communications in Statistics – Theory and Methods*, 41:13-14, 2524-2544.
- Roçadas, C., Oliveira, T.A., Mexia, J. (2011): Statistical Modelling: Application to Financial Sector. *Discussiones Mathematicae Probability and Statistics* 31, 103–119.
- Valente, V., Oliveira, T.A. (2011): Application of HLM to data with multilevel structure. *Discussiones Mathematicae. Probability and Statistics* 31, 87–101.

• Teses de Mestrado

Título*: *Modelação de Séries Temporais em R: Séries com suporte em números reais e números inteiros*

Autor: José Luís Santos, *jose.luis@ine.pt*

Orientadoras: Dulce Gomes e Maria Filomena Mendes

* Nota do Editor: Com pedido de desculpa pelo erro, esta informação corrige a que foi publicada na página 65 do Boletim primavera 2012.

Título: *Planeamento de Experiências: Modelos e desafios dos Planos em Blocos Incompletos*

Autora: Helga Carina Baptista Martins Correia, *helga_correia@hotmail.com*

Orientadora: Teresa A. Oliveira

Título: *Modelo de Programação Linear aplicado ao Planeamento de Recursos em Sistemas de Filas de Espera*

Autor: Paulo Jorge Ribeiro Correia da Silva, *paulo.j.silva@gmail.com*

Orientador: Manuel António Coelho do Carmo

• Teses de Doutoramento

Título: New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions

Autor: Paulo Canas Rodrigues, *paulocanas@gmail.com*

Orientadores: Stanislaw Mejza e João Tiago Mexia

Interação entre genótipo e ambiente (GEI) é frequente em ensaios multi-localização, e traduz-se por diferentes respostas dos genótipos em diferentes ambientes. Com o desenvolvimento das marcas moleculares e técnicas de mapeamento, os investigadores podem analisar todo o genoma para detetar as localizações específicas dos genes que influenciam a característica quantitativa de interesse. Estas localizações são denominadas de quantitative trait locus (QTL) e, quando estes QTLs apresentam diferentes respostas em diferentes ambientes, estamos perante interações entre QTL e ambiente (QEI), que é a base da GEI. Uma boa compreensão destas interações permite aos investigadores seleccionar melhores genótipos para diferentes condições ambientais e, conseqüentemente, melhorar colheitas em países desenvolvidos e, especialmente, em países em desenvolvimento. Nesta tese de doutoramento pretendo apresentar novas estratégias para melhorar a deteção e perceção de QTLs, especialmente QTLs associados a QEI no contexto de ensaios multi-localização, utilizando e fornecendo *open source* software.

Na primeira parte desta tese é apresentada uma comparação entre dois dos métodos mais usados na análise da GEI: a análise conjunta de regressões (JRA) e o modelo de efeitos principais aditivos e interação multiplicativa (AMMI). Esta comparação é realizada em termos de “robustez” com o aumento da proporção de valores omissos, e em termos da obtenção dos genótipos dominantes/vencedores. Nos capítulos seguintes são apresentados métodos com duas e três etapas onde os modelos AMMI são usados para aumentar a precisão dos dados fenotípicos, e os respetivos scores usados para ordenar os ambientes na procura de padrões ecológicos ou biológicos. A primeira destas abordagens (duas etapas) é apropriada quando a variância do erro é constante ao longo dos ambientes, enquanto a segunda (três etapas) é uma generalização permitindo ter em conta diferenças na variância do erro ao usar o modelo AMMI ponderado (WAMMI, proposto nesta tese). A parte final da tese ilustra uma estratégia para simular e modelar GEI e QEI em características complexas como a produção/rendimento, com base numa série de parâmetros fisiológicos dependendo apenas dos dados genotípicos. Isto é realizado usando um modelo eco-fisiológico de crescimento de colheitas com sete parâmetros dependentes de QTLs.

Paulo Canas Rodrigues

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

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: Equações diferenciais estocásticas na modelação do crescimento individual em ambiente aleatório

Autora: Patrícia A. Filipe, *pasf@uevora.pt*

Orientador: Carlos A. Braumann

Na minha tese, foi desenvolvida uma classe de modelos de equações diferenciais estocásticas (EDE) para descrever o crescimento individual de organismos (animais ou plantas) em ambiente aleatório. O crescimento de indivíduos da nascença até à maturidade deve ter em conta o efeito das flutuações aleatórias do ambiente que afetam a taxa de crescimento, e os modelos de EDE são mais apropriados do que os métodos de regressão tradicionalmente usados. A classe de modelos apresentada conduz a versões estocásticas de um modelo determinístico geral simples e flexível, que inclui como casos particulares alguns dos modelos de crescimento determinísticos mais utilizados, como é o caso, por exemplo, do modelo monomolecular, do modelo logístico, do modelo de Gompertz e do modelo de Bertalanffy-Richards. A classe de modelos de EDE desenvolvida possui um termo adicional, que incorpora o efeito das flutuações aleatórias do ambiente. Estudámos o caso em que existe uma única forma funcional para descrever a dinâmica média da curva de crescimento completa (modelo monofásico) e generalizámos para o caso em que consideramos que o coeficiente de crescimento assume valores diferentes para diferentes fases da vida do indivíduo (modelo multifásico). Desenvolvemos também o caso em que o tamanho médio assintótico varia aleatoriamente de indivíduo para indivíduo (modelo de efeitos mistos). Foram apresentadas as principais propriedades que caracterizam cada um destes três casos, em particular, as soluções dos modelos e correspondentes função densidade, função densidade assintótica e função densidade de transição. Aplicámos três métodos de estimação: o método de máxima verosimilhança, métodos bootstrap (paramétrico e não-paramétrico) e um método de estimação não-paramétrica. Para cada um destes métodos desenvolvemos o caso em que trabalhamos com os dados de uma trajetória e a extensão ao caso em que temos dados de várias trajetórias. Foi apresentado um estudo de previsão para os modelos monofásico e bifásico. Por fim, caracterizámos o tempo que um indivíduo demora a atingir um determinado tamanho. Quando se trata de um tamanho de interesse económico, podemos destacar a importância destes resultados e, neste contexto, foram abordados problemas de otimização. Para ilustração dos resultados utilizámos dados do peso de bovinos de raça Mertolenga. A aplicação dos modelos que desenvolvemos pode ser de grande utilidade em diversas áreas, e em particular, na área de criação de gado. Um maior conhecimento acerca do crescimento dos animais, permite aos criadores, por exemplo, do ponto de vista económico, otimizar o lucro médio de venda dos animais.

Patrícia Filipe

• Capítulos de Livros

Sven Trenholm, Angel A. Juan, Jorge Simosa, Amílcar Oliveira, Teresa Oliveira (2011). Long-Term Experiences in Mathematics E-Learning in Europe and the USA. Invited Chapter of Book in IGI Global Publisher - Math-E-Learning: *Teaching Mathematics Online: Emergent Technologies and Methodologies*. Chapter 12, pages 238-257. Juan, A.; Huertas, M.; Trenholm, S.; Steegmann, C. (eds.)

Available at <http://www.igi-global.com/viewtitlesample.aspx?id=57942>

Oliveira, T.A., Oliveira, A., Pérez-Bonilla, A.(2012). Data Mining and Quality in Service Industry: Review and some applications. CRC/Taylor & Francis Book: *“Decision Making under Uncertainty in the Service Industry”*. Editors: J. Faulin, A. Juan, S. Grasman & M. Fry. Chapter 4, pages 83-108.

• Livros

Título: *Modelos com Equações Estruturais*

Autora: Maria de Fátima Salgueiro

Ano: 2012. Edições SPE. ISBN: 978-972-8890-26-1



SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

PRÉMIOS “ESTATÍSTICO JÚNIOR 2012”

Trabalho classificado em 1º lugar (3º ciclo do Ensino Básico)

Título: “*Aparelhos Multimédia*”

Autores: Joana Mendes Leite Moreira e Maria Herdade Gomes Pereira de Sousa.

Professora orientadora: Marisa Alexandra Silva P. R. Barbosa.

Estabelecimento de Ensino: Externato das Escravas do Sagrado Coração de Jesus, Porto

Trabalho classificado em 2º lugar *ex-aequo* (3º ciclo do Ensino Básico)

Título: “*Ser solidário*”

Autores: Andreia Filipa Lopes Sousa, Inês da Costa Delgado e Carolina Cabeleiras Felgueira.

Professora orientadora: Maria Alice da Silva Martins.

Estabelecimento de Ensino: Agrupamento de Escolas Artur Gonçalves, Torres Novas.

Trabalho classificado em 2º lugar *ex-aequo* (3º ciclo do Ensino Básico)

Título: “*As redes Sociais*”

Autores: Sara Cristina Sousa Soares, Pedro Afonso Roque Martins e Bruno Ismael Moreira Gomes.

Professor orientador: Vasco Filipe Magalhães Ribeiro.

Estabelecimento de Ensino: EB 2,3/ S de Pinheiro, Pinheiro.

Trabalho classificado em 3º lugar (3º ciclo do Ensino Básico)

Título: “*Associação Antipoluição Atmosférica*”

Autores: André Bernardo Jorge Bandeiras, Patrícia Alexandra da Silva Fonseca e Vítor Hugo Raimundo Pontes.

Professora orientadora: Teresa Margarida Maia de Carvalho.

Estabelecimento de Ensino: Colégio Senhor dos Milagres, Leiria.

Trabalho classificado em 2º lugar (Ensino Secundário)

Título: “*Percurso dos alunos da nossa escola ao longo dos anos*”

Autores: Vítor Manuel Oliveira Ribeiro e Ricardo Ferreira Mota

Professora orientadora: Maria Irene Marques Gonçalves

Estabelecimento de Ensino: Escola Secundária Padre Benjamim Salgado, Vila Nova de Famalicão

Trabalhos classificados em 3º lugar (Ensino Secundário)

Título: “*E agora...? Aprender lá fora ou cá dentro?*”

Autor: Carlos Moura Pereira Lucas Teixeira

Professor orientador: Nuno Duarte Veríssimo Rodrigues

Estabelecimento de Ensino: Escola Secundária 3 EB Dr. Jorge Correia, Tavira

Trabalho premiado (Cursos EFA/CEF)

Título: “*Turismo em Portugal e no Mundo*”

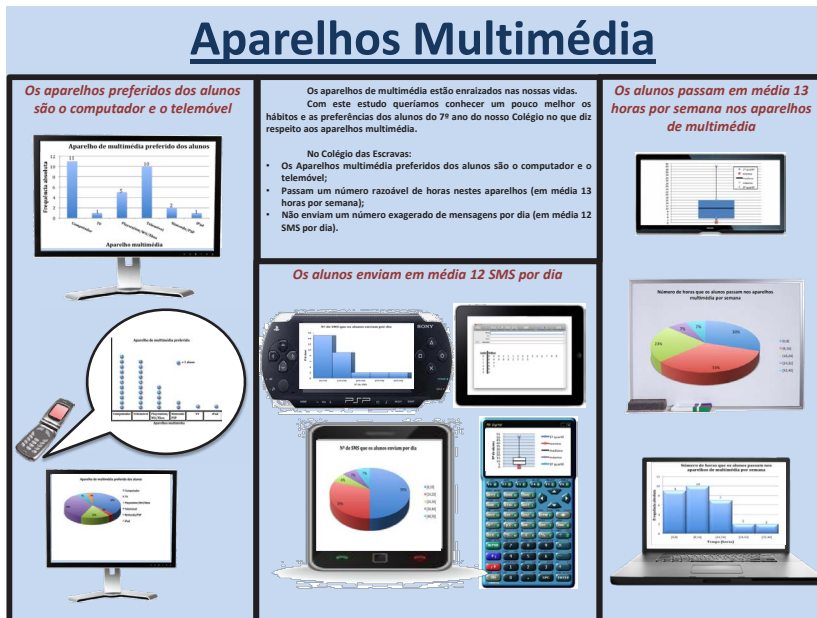
Autores: Rúben Soudo, Marco Ventura e Filipe Borges.

Professora orientadora: Helena Cristina Teixeira Fernandes

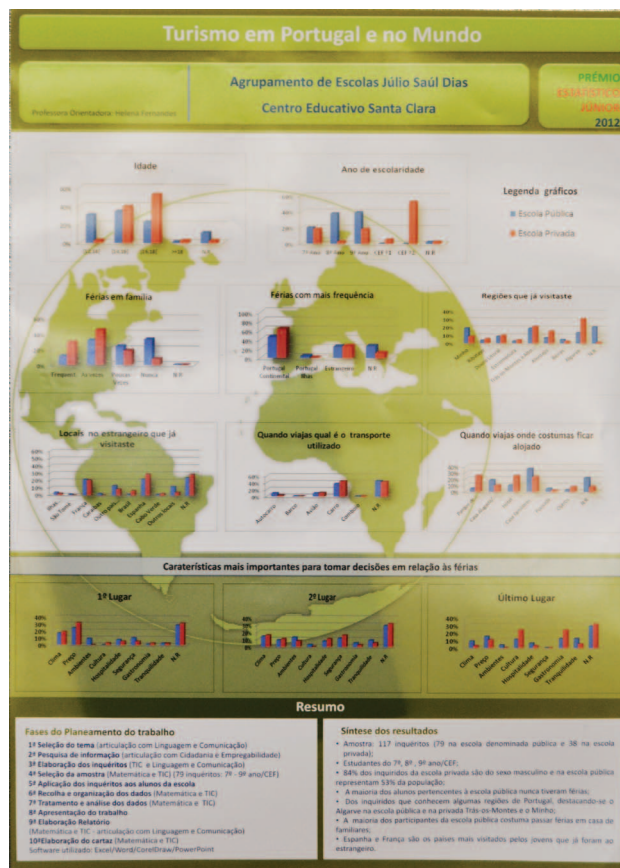
Estabelecimento de Ensino: CESC - Agrupamento de Escolas Júlio Saúl Dias, Vila do Conde

PRÉMIOS “ESTATÍSTICO JÚNIOR 2012”

Trabalho Classificado em 1.º lugar (3º Ciclo do Ensino Básico)



Trabalho Premiado (Cursos EFA / CEF)





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



SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

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.º 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.º 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.º, 2.º e 3.º 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.

Índice

Editorial	1
Mensagem do Presidente	4
Notícias	7
Métodos Estatísticos em Medicina	
A Amostragem por Tempo - Local no Estudo de Populações de Difícil Acesso <i>Luzia Gonçalves</i>	19
Modelos de Predição em Medicina: Algumas considerações <i>Ana Luísa Papoila</i>	26
Sobre o ensino de Estatística em cursos de Medicina <i>Pedro Oliveira</i>	36
Métodos Estatísticos para Modelar e Prever o Impacto das Epidemias de Gripe em Portugal <i>Maria Lucília Carvalho e outros</i>	40
Pós – Doc	
<i>Ana Diniz</i>	52
<i>Maria Fernanda Diamantino</i>	54
<i>Rui Santos</i>	56
Ciência Estatística	
<i>Artigos Científicos Publicados</i>	61
<i>Teses de Mestrado</i>	61
<i>Teses de Doutoramento</i>	62
<i>Capítulos de Livros</i>	64
<i>Livros</i>	64
Prémios “Estatístico Júnior 2012”	65
Edições SPE	68