Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral — Outono de 2007

Bioestatística

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: 1000 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

Este Boletim tem o apoio da **FCT** Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Editorial

… **Investigação e Estatística em Portugal**…

É bom! Que se crie uma história da investigação em Estatística em Portugal. Quais as potencialidades da Estatística Portuguesa? Em que áreas principais da Ciência Estatística se investiga em Portugal?
Porque a Estatística também se estabelece como uma área interdisciplinar, sabemos que a investigação se desenvolve (ou pode desenvolver) nos mais diversos domínios. Qual o ponto da situação em Portugal? Alguma pesquisa é facilitada e está disponível informação, por exemplo, nas bases de dados da FCT - Fundação para a Ciência e a Tecnologia. Mas, é importante saber o que investigam os estatísticos portugueses. Quais as principais dificuldades? Quais as áreas com maior desenvolvimento em Portugal? Quem faz o quê, onde, quando, como…
Quantos Centros com Investigação em Estatística existem em Portugal? As respostas fazem alguma luz sobre a noção de Estatístico Português!

Propomo-nos fazer um ponto da situação em próxima edição do Boletim SPE. Esta (árdua?) tarefa será muito facilitada com a colaboração dos leitores - principalmente dos seniores… aqueles que têm "mais memórias"! Mas também daqueles que bem conhecem ou que são directamente responsáveis por algum centro de investigação. Especialmente a estes se faz um apelo para manifestarem a sua presença, para já, dando notícias.
Na sequência do Memorial da Sociedade Portuguesa de Estatística, este projecto editorial específico sobre investigação científica continuará essa edição com uma actualização "mostrando" futuro.
Fundamentalmente, um ponto de partida (decerto!) pode ser uma referência para (ou de) algum centro de investigação.
Diversos são os centros estabelecidos como de investigação em Estatística; mas existem alguns onde, eventualmente, os estatísticos estão em minoria mas com actividade interdisciplinar de pesquisa estatística relevante. O somatório de todas as opiniões pode fazer uma síntese - um ponto da situação.
Uma história! É o que se deseja…

Caro(a) Colega,
Como desenvolve a sua investigação em Estatística? Está integrado em algum centro?
É responsável por um centro de investigação em Estatística? É fundamental a sua intervenção!
Vamos construir uma edição do Boletim sobre a investigação estatística.
São bem vindas todas as colaborações.
Aguardo notícias.
Obrigado!

P.S.- É devida uma palavra sobre o anterior número do Boletim Primavera. Uma lamentável gralha na edição originou que, no poema… aparecesse um "verso suplementar". Embora no contexto, o último verso do poema da página 5 do anterior Boletim é da responsabilidade do acaso e não da sua autora Manuela Magalhães Hill que assim viu perturbada a sua mensagem poética. De facto, foram acrescentadas palavras estranhas ao poema. Mas, como que por acaso o verso gralha "Os Congressistas e Seus Trabalhos", surgiu camuflado com um texto que não ficaria a despropósito na prosa embora tornando ainda mais difícil a detecção na revisão das provas; mas que incomodou a poesia…
Desculpa Manuela!

Mensagem do Presidente

Caros Colegas:

Este ano, o nosso Congresso Anual, o XV, realizou-se em Lisboa em data pouco habitual, em Agosto, de modo a permitir a participação conjugada neste evento e na 56ª Sessão do International Statistical Institute (ISI 2007), para o que, aliás, se negociou uma inscrição especial conjunta. E, de facto, a maioria dos participantes no nosso Congresso participou também no ISI 2007. Foi também uma forma de contribuirmos para uma elevada e digna participação portuguesa no grande congresso internacional da Estatística, onde participam os maiores vultos desta ciência e que contou desta vez com quase três milhares de delegados de todo o mundo.

Ambos os eventos foram muito bem sucedidos.

Está de parabéns a Comissão Organizadora do XV Congresso Anual da SPE, capitaneada pela nossa colega Manuela Magalhães Hill. Estamos todos gratos pelos incansáveis esforços e capacidade organizativa em circunstâncias mais difíceis do que as habituais. Mas estão igualmente de parabéns o ISCTE, que aceitou ser instituição anfitriã, os membros da Comissão Executiva e Científica, os presidentes das sessões de trabalho, os oradores convidados, os autores das comunicações e todos os participantes em geral. Todos deram importantes contributos para este memorável evento.

Durante o XV Congresso saíram as Actas do XIV Congresso, um registo importante dos desenvolvimentos que a Estatística teve em Portugal em 2006. Fecha assim com chave de ouro um outro Congresso, cuja Comissão Organizadora, da Universidade da Beira Interior e presidida pela nossa colega Maria Eugénia Ferrão, fez um excelente trabalho.

De parabéns também está a organização do ISI 2007, que soube vencer as enormes dificuldades de um realização tão complexa e monumental. O nosso colega Paulo Gomes, em representação do INE, foi o grande responsável por trazer o evento a Lisboa e liderou brilhantemente todas as etapas do processo. A nossa colega Ivette Gomes assumiu não menos brilhantemente a exigente responsabilidade da Comissão de Programa, fundamental para o sucesso científico do evento. Foi notável a colaboração entre as instituições organizadoras (INE, SPE e CLAD) e da comunidade estatística nacional, que se envolveu na organização e participou cientificamente no evento. Várias instituições nacionais e muitos colegas nossos assumiram a responsabilidade de organizar reuniões científicas satélite. Saiu reforçada a boa imagem internacional dos estatísticos portuguesas e a nossa capacidade organizativa.

Mas para o ano há mais: vem aí o COMPSTAT 2008, a realizar no Porto, onde igualmente esperamos uma forte participação portuguesa. Portugal está mesmo a dar cartas na Estatística.

É altura de pensarmos já no próximo Congresso anual de 2008, o XVI. A Universidade de Trás-os-Montes e Alto Douro vai ser a anfitriã e a Comissão Organizadora é presidida pela nossa colega Irene Oliveira. Estou certo que será igualmente um evento memorável.

Não queria terminar sem vos dar conta das últimas novidades em termos de actividades da SPE, não porque sejam mais importantes do que muitas outras actividades que são já rotineiras, mas porque delas poderão ainda não ter tido conhecimento.

Comecemos pelo trabalho das Comissões Especializadas. A Comissão Especializada de Educação, entre outras actividades, continua a coordenar a revisão de manuais escolares dos ensinos básico e secundário, actividade especialmente relevante que tem vindo a intensificar-se. A Comissão Especializada de Nomenclatura Estatística terminou o Glossário Estatístico Inglês-Português, para o qual se solicitou e se obteve a colaboração da Associação Brasileira de Estatística. Ele pode ser visto no nosso "site" e deverá integrar o "site" do ISI. Parabéns à Comissão, presidida pelo nosso colega

Daniel Paulino, pelo importante serviço prestado à comunidade estatística lusófona.

No XV Congresso, além do habitual Prémio SPE 2007 para trabalhos científicos, retomámos, com o apoio da Porto Editora e sob a coordenação do nosso colega Russell Alpízar-Jara (com a colaboração da Comissão Especializada de Educação), os Prémios Estatístico Júnior para trabalhos escolares dos ensinos básico e secundário. É uma das formas de suscitarmos o interesse pela Estatística nas novas gerações, mas outras iniciativas com igual objectivo estão na forja. Aproveito para agradecer aos júris de ambos os prémios.

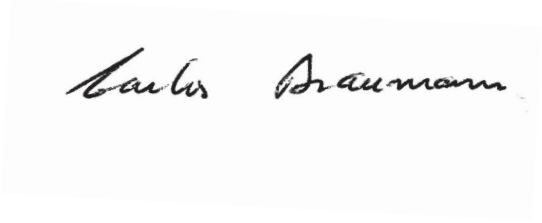
Acordámos com a CLAD - Associação Portuguesa de Classificação e Análise de Dados um Protocolo de Cooperação que facilita a participação dos sócios de uma sociedade nos eventos organizados pela outra e prevê outras formas de cooperação.

Iniciaram-se os Encontros de Estatística SPE-CIM, uma colaboração com o Centro Internacional de Matemática coordenada pela nossa colega Paula Brito, tendo o primeiro versado o tema "Estatística nas Telecomunicações" e sido organizado pelo colega António Pacheco.

Aproveitem para visitar o nosso "site" e ver as várias melhorias que foram introduzidas e, se tiverem sugestões a dar, encaminhem-nas para a coordenadora, a nossa colega Paula Brito.

E é altura de terminar esta mensagem, que já vai longa, com um "até breve".

Saudações cordiais



Notícias

• Um olhar reminiscente sobre o XV Congresso da SPE

O Palco

Lisboa, a capital, cidade fundada pelos Fenícios cerca de 1200 a.c., de onde partiram inúmeras expedições na época dos descobrimentos (como a de Vasco da Gama rumo à Índia), palco de outros importantes acontecimentos da história de Portugal (como a restauração da independência em 1640, a implantação da república em 1910 ou a revolução dos cravos em 1974). Neste ano foi também a capital da Estatística nacional e internacional durante 11 dias com o XV Congresso da SPE e logo de seguida a 56ª Sessão do ISI. Aqui se congregaram estatísticos provenientes de todos os cantos do território nacional (norte, centro e sul do país) e alguns participantes e convidados internacionais. Apesar da Sociedade Portuguesa de Estatística estar sedeada em Lisboa, foi a primeira vez que a cidade acolheu o tradicional congresso da SPE.

Os Realizadores

Coube ao grupo de Métodos Quantitativos do Instituto Superior de Ciências do Trabalho e da Empresa (ISCTE) a tarefa de organizar o XV Congresso entre 19 a 21 de Agosto. Uma tarefa que se afigurava árdua, não só pela realização quase em simultâneo com a 56ª Sessão do ISI e as várias conferências satélites, mas também por se realizar no mês de Agosto, período tradicional de férias. Também neste ano o congresso foi mais curto com o programa de trabalhos a ser completado em só 2½ dias.

Bravos organizadores! Excelente organização e óptimos espaços para a realização da conferência. O empenho, dedicação e brio profissional de Manuela Magalhães-Hill e restantes membros da Comissão Organizadora não nos passaram indiferentes. Bem hajam!

Os mais cépticos pensariam inatingível os 179 participantes!

Por outro lado, alguns preferem a realização num Hotel, onde os congressistas permanecem além do período das sessões, mas a verdade é que este congresso decorreu num ambiente muito tranquilo, diríamos familiar, possibilitando uma grande proximidade entre os participantes.

Os Actores e a metragem

Rezará a história que a primeira congressista a levantar a documentação foi Ana Paula Martins da Universidade Católica Portuguesa.

Cerca de 50 Congressistas assistiram ao minicurso "Introdução aos Métodos Estatísticos Robustos" ministrado de uma forma exemplar por Ana Pires e João Branco, ambos do Instituto Superior Técnico de Lisboa.

Robusta quanto baste foi também a Sessão de Abertura do Congresso que se realizou no bonito anfiteatro da sala B203. Carlos Braumann (Presidente do Congresso e da Direcção da SPE), Manuela Magalhães-Hill (Presidente da Comissão Organizadora Local), Rui Menezes (Presidente do Departamento de Métodos Quantitativos) e Juan Mozzicafreddo (Vice-Presidente do ISCTE) deram as boas-vindas a todos os congressistas desejando-lhes proveitosas e agradáveis sessões de trabalho. Salientamos também a presença nesta sessão de abertura da Presidente do INE, Alda Carvalho, no âmbito da cooperação entre o INE e a SPE que foi particularmente mais vincada este ano com a organização do XV Congresso da SPE e da 56ª Sessão do ISI.

Seguiu-se a primeira sessão plenária, onde Gilbert Saporta (CNAM, Paris) nos brindou com a palestra "Discriminant analysis on functional data".

O segundo dia do Congresso, já com a grande parte da documentação levantada pelos participantes, foi um dia intenso, preenchido com 4 sessões paralelas de comunicações orais de manhã (Estatística Multivariada, Estatística Aplicada à Biologia e Epidemiologia, Processos Estocásticos, Estatística e Sociedade) e outras 4 à tarde (Estatística Bayesiana, Estatística Robusta, Controlo de Qualidade I e Estatística Aplicada à Economia e Gestão I), 2 sessões de posters e 3 sessões plenárias. Neste dia, a primeira, das interessantes palestras plenárias, esteve a cargo de Ray Cambers (Universidade de Wollongton, Austrália), com o título "Measurement error in auxiliary information". A segunda intitulada "Sample design effects for social surveys: recent research and current issues" foi proferida

por Peter Lynn (ISER, Inglaterra) e a palestra que fechou o dia foi a de Pedro Portugal com o título bastante apelativo "Life after Cox".
No terceiro e último dia decorreram 4 sessões paralelas de comunicações orais (Séries Temporais, Modelos com Variáveis Latentes, Estatística Aplicada à Economia e Gestão II e Controlo de Qualidade II/Software) e uma sessão de posters.
Registe-se que as sessões orais e posters foram muito interessantes e reveladoras da grande diversidade e do grande dinamismo da investigação em estatística.
No total registámos 63 comunicações orais e 39 comunicações em poster.
Após a última sessão de Posters fomos brindados pela magnífica palestra de Ligia Rodrigues, que recebeu o Prémio SPE deste ano relativo ao trabalho "Estimação do Índice de Cauda em Modelos de Caudas Pesadas: acomodação do viés nos excessos acima de um *threshold* elevado". Foram também entregues os Prémios Estatístico Júnior 2007 (ver detalhes na notícia respectiva nesta edição do Boletim) ficando como breve testemunha dos jovens estatísticos que estão a emergir. O trabalho premiado em 1º lugar (Ensino Básico) intitulado "Ocupação dos alunos do 7º Ano da ESAG nas Férias da Páscoa", elaborado por alunos do 7º ano da Escola Secundária Artur Gonçalves, Torres Novas, foi exposto durante a última sessão de posters.
Antes da sessão de encerramento, tivemos a última plenária do congresso, onde Jeroen K. Vermunt (Univ. de Tilburg, Holanda) proferiu a palestra "Multilevel variants of discrete and continous latent variable models: The Latent GOLD framework".
A reminiscência científica poderá encontrar-se nas actas do XV Congresso Anual da SPE.

Lazer no Congresso
No final do primeiro dia fomos "mimados" com uma agradável recepção no Palácio dos Condes d´Óbidos – Sede Nacional da Cruz Vermelha Portuguesa.
O edifício, um Palácio do século XVII, "construído no alto de um rochedo, cujas escarpas desciam vertiginosamente em direcção ao Rio Tejo", com um terraço de vista magnífica sobre o Porto de Lisboa, Cacilhas, Cristo Rei, Trafaria e Baía do Seixal, deixou a todos encantados.
De realçar os seus salões ricamente decorados, com pinturas nos tectos e painéis de azulejos, e a linda Capela dedicada a Nossa Senhora da Conceição.
No final dos trabalhos do segundo dia, eis que chega a hora do lazer!
Pelas 19h, a já habitual "Foto do Congresso", tirada numa escadaria exterior do ISCTE, em que o Fotógrafo, "quase" se esquecia de tirar a tampa da máquina fotográfica e clique…clique…clique…3 fotos ao "familiar" grupo de Estatística.

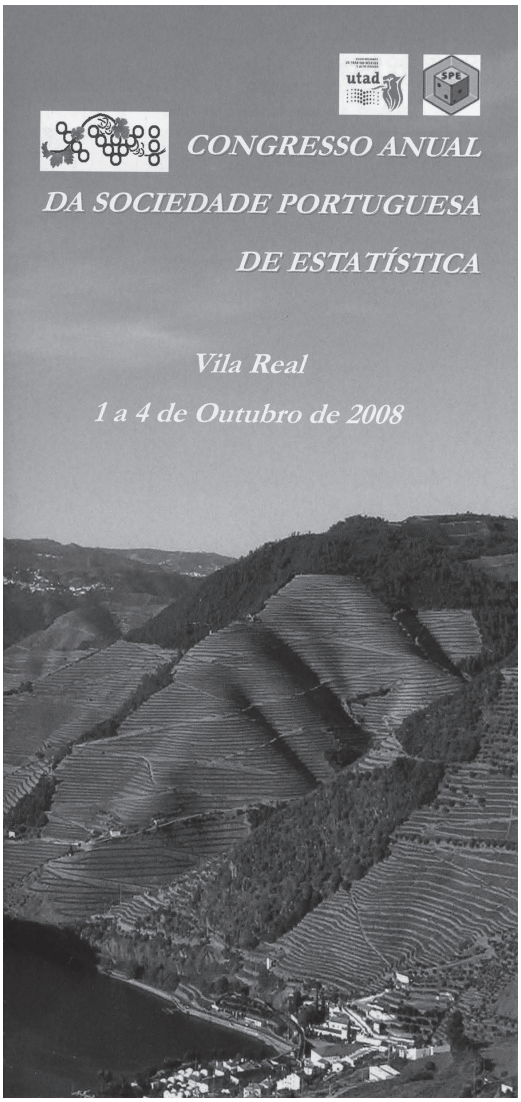
Em seguida, todos rumo ao Centro Cultural de Belém, para o jantar do Congresso. Onde o "Senhor Vento" nos levou a entrar de imediato no CCB, onde nos foi servido um delicioso aperitivo.
O jantar servido no Salão Fernando Pessoa, com uma vista magnífica sobre a Praça do Império, com a Ponte 25 de Abril ao fundo e o Mosteiro dos Jerónimos à esquerda decorreu calmamente e num ambiente de boa disposição e familiaridade (com cerca de 122 pessoas!) que já vem sendo habitual ao longo dos anos. Ficou-nos apenas um leve sabor a pouco…não, não nos referimos a quantidade (nem a qualidade, pois a comida estava deliciosa...), embora tenhamos ouvido um ou outro comentário, mas sim à falta de tempo para algumas tertúlias junto às mesas e talvez algum momento musical. Claro que o tempo para o Congresso foi mais limitado desta vez e tínhamos que deitar-nos cedo q.b. para iniciar as sessões de trabalho do dia seguinte.

O adeus e até já

E por ser mais curto, na terça-feira, dia 21 Agosto, chegou ao fim mais um Congresso da SPE, e com a promessa que para o ano haverá mais.
Lá iremos rumo a Trás-os-Montes e Alto Douro, de 1 a 4 de Outubro de 2008 para o XVI Congresso da SPE. O aperitivo servido por Irene Oliveira com um magnífico filme promocional da região, deixou-nos a todos água na boca e ansiosos pelo próximo congresso.
Até lá!...

Paulo Infante (DMAT- UÉvora)
Margarida Silva (SPE)
Russell Alpizar-Jara (DMAT- UÉvora)

• XVI Congresso SPE



CONGRESSO ANUAL DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA
Vila Real, 1 a 4 de Outubro de 2008

Presidente do Congresso

Carlos Braumann (Presidente da SPE)

Comissão Organizadora Local

Irene Oliveira (DM, UTAD)
Elisete Correia (DM, UTAD)
Fátima Ferreira (DM, UTAD)
Maria Manuel Nascimento (DM, UTAD)
Sandra Dias (DM, UTAD)

Comissão Executiva e Científica

Carlos Braumann (DM, UE)
Irene Oliveira (DM, UTAD)
Paula Milheiro (FEUP)
Isabel Pereira (DM, UA)
Pedro Oliveira (DPS, UMinho)

Conferencistas Convidados

Daniel Gianola (Univ. Wisconsin-Madison, EUA)
Carlos Daniel Paulino (IST, UTL)
Manuela Neves (ISA, UTL)
Geert Molenberghs (Univ. Hasselt, Bélgica)
Ian MacLean (Business Statistics Users Groups, Reino Unido)

Mini-Curso

Estatística Espacial
Lucília de Carvalho (FCUL)

Comunicações Orais

15 minutos para apresentação e mais 5 minutos para discussão

Sessões de Posters

• 56.ª Sessão do ISI - International Statistical Institute

Pela primeira vez, na longa história do ISI - International Statistical Institute, Portugal foi o anfitrião da comunidade estatística mundial. De 22 a 29 de Agosto de 2007 realizou-se em Lisboa a 56ª Sessão do ISI: uma organização do INE - Instituto Nacional de Estatística com a colaboração da SPE - Sociedade Portuguesa de Estatística e da CLAD - Associação Portuguesa de Classificação e Análise de Dados e que envolveu empenhadamente os estatísticos e a comunidade estatística portugueses. Num breve resumo no jornal diário do congresso Niels Keiding, Presidente do ISI, sintetizou os nomes dos "mais esforçados": " I have followed the preparations for this event for several years and I have been consistently impressed by the high level of enthusiasm and professionalism displayed at the organizational as well as the scientific level, We should all be grateful to Paulo Gomes, chair of the Organizing Committee and his staff, particularly José Pinto Martins, to Pedro Silva, chair of the Programme Coordinating Committee and its members, and to M. Ivette Gomes, chair of the Local Programme Committee, and its members."
Sobre a importância destas sessões internacionais, Ivette Gomes (no citado jornal do congresso) afirmou: "The ISI Sessions play a leading role in the development of Statistics, insofar as a large share of the presentations deal with new trends and developments; and it is a privileged forum to reinforce, with mutual benefits, the interplay between official and academic Statistics."

Foi um Congresso com cerca de 3000 participantes vindos de 125 países e que produziram mais de 1700 comunicações científicas. A comunidade portuguesa foi representada por cerca de 550 congressistas.
Com esta entusiástica adesão, demonstra-se uma vitalidade e prestígio da organização ISI como uma entidade que promove acontecimentos científicos nos diferentes campos do saber – desde o ensino das probabilidades e estatística até aos novos desafios e pesquisa de soluções; afirmou Paulo Gomes em mensagem aos congressistas.

Carlos Braumann, Presidente da SPE, numa mensagem publicada no jornal do Congresso, fez uma breve apresentação da SPE e deu as boas vindas aos participantes em nome da comunidade estatística nacional. Salientou ainda a preocupação da SPE com a melhoria da qualidade da educação estatística e a sua promoção em vários caminhos cooperando com instituições nacionais e internacionais desde logo como membro institucional do ISI.

Na cerimónia de abertura do Congresso, o Presidente da República, Cavaco Silva, no seu discurso (também como Professor de Economia) certificava que: " (…) as estatísticas são poderosos instrumentos para o conhecimento da nossa sociedade, essenciais para a tomada de decisões, para determinar e definir estratégias inclusive para o próprio debate político."

O jornal diário do congresso apresentou duas breves entrevistas sobre perspectivas e desafios que se colocam ao ISI - com Niels Keiding, o Presidente do ISI e com Denise Lievesley, a Presidente Eleita e que a seguir transcrevemos.

Com a realização deste 56ª Sessão do ISI, Portugal foi inscrito numa página cimeira no memorial do International Statistical Institute.

Entrevista a Niels Keiding

How do you qualify your presidency? What's your style: are you a conservative or an innovator?

My main motivation was that I find the broad and interdisciplinary nature of statistics interesting and challenging. I have been lucky to work in two energetic executive committees, first as president-elect under President Stephen M. Stigler from and the last two years as president. We have had a strong wish of reforming the ISI and have come far enough for this process to be completed over the coming two years.

What changed at ISI during recent years?

The ISI constantly needs to adapt to the context in which it exists, and many things are changing rapidly these years that affect how organizations can act. We have particularly had to relate to the increasing specialization, which requires the ISI to emphasize what is common for all statistics so that the ISI can redefine itself. We adopted in Sydney a new Strategic plan, with which I had been closely involved, and which more clearly outlines ISI central activities. At the more practical level, we have spent much time in digging into the financial situation and have taken several steps to consolidate it.

What were your main goals and concerns when you first started?

My main goal was the rather modest one of keeping the ship afloat, since I find the efforts of uniting all statistics important and I knew the challenges that the ISI was facing. We decide that to achieve this, we need to update the priorities, which we have done in the strategic plan. Another ambition was to strengthen the collaboration with other statistical societies, particularly those representing disciplines not so strongly involved with the ISI.

Was there anything you intended to do that turned out to be impossible?

Things take time in the ISI – but given that, I think my chain intentions will ultimately succeed. I had hoped to get further with the collaboration with other societies.

The purpose of the ISI is "to promote all aspects of statistics worldwide": What were ISI main actions and results on this matter?

The current formulation of the ISI Mission is: "to promote the understanding, development, and practice of Statistics worldwide". Most of the concrete activities go on in the biennial Sessions and in the large and varied activities of the Sections, but we hope to have modernized the framework for good continuation of these efforts.

You said once that the "desire to improve collaboration across the different sections and disciplines of statistics " is the "raison d´être of the ISI". Has this collaboration improved? Which are the main difficulties?

This issue is still a central challenge for the ISI leadership. We face a strong centrifugal force: each area is busy with its own problems, where I believe that they could often gather strength from seeing their situation in a larger perspective. One example is the conflicts of interest issue that increasingly faces biostatisticians in pharmaceutical companies – in my view they could benefit from contact with the strong professional culture of independent judgment prevalent in official statistics.

What kind of concerns are you passing on to your successor, Denise Lievesley?

The main challenge is to complete the restructuring process with the aim of a more homogeneous ISI with broad ownership and internal solidarity as its hallmarks. And the good development in the ISI financial situation needs to be continued.

Would you like to tell us about some experiences you have lived at ISI Sessions?

My first ISI Session was Warsaw 1975, where I attended the founding of the Bernoulli Society and

where the ISI and the Biometric Society (now the IBS) had jointly organized a series of meetings (topics) on statistical and mathematical approaches to epidemics. We should learn from that successful example of collaboration across organizations. Since 1983 I have only missed one ISI Session, so I have been led to many fascinating places in the world. Fortunately, there are still many nations that are willing and able to undertake the considerable effort and expense of an ISI Session. At most of these sessions I have had administrative commitments, but there has always been time to listen to meetings on topics that I do not regularly attend. I hope that more participants will take that opportunity; unfortunately one often gets the impression that most prefer to listen to their closest colleagues - that is not the primary purpose of the ISI Sessions!

What are your expectations for ISI 2007?

I have followed the preparations in some detail. I have been impressed by the dedication and professionalism displayed throughout and as we can see now that we are here, it really works! Let me thank the local organizers, headed by Paulo Gomes and with J. Pinto Martins working full time for the last years, Ivette Gomes heading the Local Programme Committee and Pedro Silva who headed the Programme Coordinating Committee. My own expectations center around the administrative obligations in my position - but I do hope to get to many talks in areas that I do not listen to so often - and to enjoy Lisboa over the weekend when my wife comes here to see me!

Entrevista a Denise Lievesley

We understand that you are the first woman President of the ISI. What are you thoughts on this?

I am very honoured to have been elected to this position but, given that women have held many high national and international offices and that there are eminent women statisticians, it has taken the ISI a long time to get there! Women constitute only 1 in 7 of the elected ISI members and I urge everyone – both men and women – to support the work of the ISI Committee on Women in Statistics, which will meet at this conference. The increasing involvement of young women in statistics in many countries leads me to be optimistic for the future.

What do you think about the programme for this ISI Session?

Pedro, Ivette, the Section representatives and others have put together such a rich and varied programme that we are all having some difficulty in choosing which talks to attend. My own fields of interest – which include the use and misuse of statistics in Public Policy – are well covered. However, one of the joys of an ISI Session is the opportunity to expand our horizons and I always try to get to talks outside my own specialism,

How can newcomers to ISI meetings find out more about what the ISI does?

The ISI General Assembly – which takes +lace next Tuesday at 15:45 – is open to all and will report on the activities of the ISI over the past two years. We shall also hear from the South Africans about their plans for the newt biennial meeting and their magnificent stand in the exhibition hall will, I'm sure, inspire delegates to include Durban in their plans for 2009. Delegates should feel free to visit the ISI Office on Level 2 or to approach me or any other member of the ISI Executive to find out more about the work of the ISI.

Do you have any further comments about this Session?

I do hope that delegates have fun and have time to enjoy the beautiful city of Lisboa. I have attended every ISI Session since Madrid 1983 and each one has been special and memorable in its own way. Mariza's wonderful performance at the Opening Ceremony has already ensured that Lisboa 2007 will continue in this great tradition.

Fernando Rosado

• Glossário Estatístico

O objectivo inicial do trabalho da Comissão Especializada da Nomenclatura Estatística (CENE) foi finalmente atingido! Após um longo mas produtivo período de estudos, debates, reflexões e contactos, deu-se por finda a elaboração da parte em língua portuguesa do glossário estatístico do ISI. Lamentamos, todavia, a escassa participação registada no processo de consulta, mormente de quem actua quotidianamente na área da Estatística, ainda que estejamos conscientes de que o calendário apertado – a que nos obrigou o compromisso de ultimar o trabalho antes da reunião do ISI em Lisboa – não tenha proporcionado as condições ideais de reacção susceptíveis de conduzir a significativas mais-valias.

Este trabalho, que se encontra disponível em versão pdf nesta página no fim do texto, é produto dominantemente do esforço conjunto de comissões criadas no seio da SPE e da ABE. A comissão portuguesa é, recordando, constituída por Carlos Daniel Paulino, João Branco, Lucília Carvalho e Dinis Pestana, e a comissão brasileira por Julio Singer, Lúcia Barroso e Wilton Bussab.

É, desde já, importante esclarecer que as diferenças terminológicas e ortográficas entre as duas comunidades estatísticas estão explicitadas nesta versão colocando-se **entre parênteses rectos ou após o símbolo // os termos ou expressões usados especificamente no meio brasileiro**, quando distintos de algum modo daqueles vigentes usualmente no meio português.

Este glossário constitui assim uma primeira fonte terminológica credível de vocabulário estatístico para o mundo lusófono, que se espera em breve estar também disponibilizada no *site* do ISI, *http://isi.cbs.nl/glossary/index.htm*. Informa-se desde já que as diferenças entre o português europeu e brasileiro poderão ficar ali registadas num formato diferente daquele indicado acima, que elimine qualquer ambiguidade.

Uma análise atenta e crítica do conteúdo deste glossário permite constatar, por um lado, a existência de muitos verbetes desnecessários ou obsoletos e, por outro, a omissão de demasiados itens, alguns dos quais suscitados por desenvolvimentos recentes na área, que não nos foi possível atenuar pelas limitações impostas ao completamento descentralizado da actual versão pelo ISI.

Neste contexto, a CENE decidiu, com a concordância dos colegas brasileiros, iniciar proximamente uma 2ª etapa deste projecto, visando elaborar um novo glossário estatístico inglês-português para a comunidade lusófona que não sofra das mencionadas deficiências do glossário do ISI.

Apelamos, mais uma vez, aos variados utentes da Estatística para que contribuam com os seus comentários e sugestões para este novo glossário, juntando-se assim aos colegas, a quem estamos imensamente gratos, que nos enviaram registos de verbetes omissos ao longo do anterior processo de consulta. Quanto mais significativa e positiva for a nossa participação, tanto melhor será o resultado deste projecto no sentido de vir a constituir um instrumento de consulta indubitavelmente útil para todos os interessados - que não se esgotam em nós, estatísticos, - e de desenvolvimento em moldes rigorosos de uma comunicação científica em língua portuguesa.

Carlos Daniel Paulino

• Região Ibérica da Sociedade Internacional de Biometria: SIM ou NÃO?

A *International Biometric Society* (IBS) é uma associação profissional sem fins lucrativos que se dedica ao desenvolvimento e aplicação da teoria e métodos estatísticos e matemáticos nas Ciências da Vida, englobando as áreas da Biomedicina e Saúde Pública, Agronomia, Ambiente, Ecologia, Psicologia e áreas afins. É neste âmbito que se insere, nomeadamente, o seu papel patrocinador de duas revistas científicas bem conhecidas, *Biometrics* e *Journal of Agricultural, Biological and Environmental Statistics*. Para mais informações consulte-se o *site* *http://www.tibs.org.*

Esta sociedade reúne no seu seio investigadores de vários perfis organizados geralmente em regiões geográficas (ou grupos nacionais). Os investigadores portugueses trabalhando naqueles campos (bioestatísticos, biomatemáticos, biólogos, epidemiologistas, etc) não estão globalmente organizados dentro da IBS, como acontece, por exemplo, com os seus colegas de Espanha que há muito constituíram a Região Espanhola da Sociedade Internacional de Biometria (através da Sociedade Espanhola de Biometria), a que por sinal estão associados alguns (poucos) estatísticos portugueses.

Nos últimos tempos tem havido diversos contactos entre membros espanhóis dessa sociedade (incluindo elementos dos seus órgãos directivos) e estatísticos portugueses, em que se discutiu o eventual interesse e viabilidade da criação da **Região Ibérica da IBS**, agrupando os membros da actual Região Espanhola e os investigadores portugueses interessados que desenvolvam a sua actividade no imenso campo da Biometria.

Julgo ser da máxima importância auscultar o sentir dos sócios da SPE que trabalham em Biometria relativamente à questão que serve de título a este trecho informativo: **Sim ou não à criação da Região Ibérica da IBS?**

Neste sentido, apela-se a esses colegas para que difundam esta notícia a sectores potencialmente interessados que não estão associados à SPE, a fim de que o máximo possível de intervenientes, sócios ou não da SPE, se possa pronunciar sobre o interesse que vêem ou não na criação da tal Região Ibérica e o desejo de se tornarem membros da IBS através daquela. Aconselho vivamente a que a expressão das opiniões seja feita por mensagem para o endereço electrónico da SPE, *spe@fc.ul.pt*. Na sequência dos resultados desta sondagem informal, espera-se tomar as iniciativas condizentes que atempadamente serão noticiadas.

Carlos Daniel Paulino

• Equipas Vencedoras na Competição sobre Literacia Estatística do ISLP/IASE

Com a colaboração de Pedro Campos do INE, Bruno de Sousa da Universidade do Minho, Maria Manuel da Silva da Universidade de Trás-os-Montes e Alto Douro e diversos recursos portugueses em Literacia Estatística, o Projecto Internacional de Literacia Estatística (*International Statistical Literacy Project, ISLP*) concluiu a sua competição no Norte de Portugal de 14 a 18 de Agosto último. As equipas vencedoras, Escolas e Professores responsáveis nesta competição foram:

(a) 1º Lugar no Secundário

Filipa Daniela Pereira Costa, Olívia Cristina Marques Gomes, Diana Cristina Oliveira Vaz

Escola Secundaria de Caldas Das Taipas

Professora: Lurdes Marques

(b) 1º Lugar no 3º Ciclo

Ana Cláudia Gomes Moreira, Diana Pereira de Carvalho

Escola EB2,3 de Valença do Minho

Professora: Deolinda Manuela Souto Agra

(c) 2º Lugar no Secundário

Mafalda Nascimento Guimarães, Ana Priscilla Ramalho Teixeira, Rafaela Sofia Oliveira Castro Bernardino

Escola Secundaria/3 Senhora de Hora

Os alunos tomaram a iniciativa de participarem e treinarem para este concurso sozinhos, com o apoio incondicional da mãe da Mafalda sempre encorajadora em todo este processo.

(d) 2º Lugar no 3º Ciclo

Soraia Correia Pacheco, David Passas Rodriguez

Escola EB2,3 de Valença do Minho

Professora: Deolinda Manuela Souto Agra

Os patrocinadores desta competição foram entre outros a Universidade do Minho, a APM, o INE, a ASA, a AREN e a UCLA. Mais pormenores sobre esta competição podem ser vistos na morada de Internet *http://www.stat.auckland.ac.nz/~iase/islp/game*

O ISLP/IASE já abriu as inscrições para a primeira Competição Internacional em Literacia Estatística (International Statistical Literacy Competition) a decorrer em 2008. Os professores já podem registar as suas turmas através do questionário disponível na morada de Internet:

http://www.stat.auckland.ac.nz/~iase/islp/competition

As inscrições estarão abertas até dia 28 de Fevereiro de 2008. Materiais de treino para a competição poderão ser encontrados no web site deste projecto. Convidamos todos os professores a registarem as suas turmas.

Bruno Sousa

Bioestatística

A Estatística na investigação científica - perspectiva de um jovem biólogo

Gil Penha-Lopes, *gil.penha-lopes@netcabo.pt*
Laboratório Marítimo da Guia, IMAR, FCUL

Introdução

Recentemente li um livro de Estatística, intitulado de *Intuitive Biostatistics*, onde o autor, Harvey Motulsky, refere que "na investigação científica os testes estatísticos não são usados para aceitar ou rejeitar hipóteses, mas apenas para perceber o peso da evidência contra ou a favor delas". Concordo plenamente com o autor mas demorei alguns anos até a compreender, não como aluno, mas como investigador científico.

Durante o curso de Biologia tive a possibilidade de trabalhar com bons investigadores e de entender, na prática, a importância dos planeamentos experimentais e estatísticos, muito referidos nas aulas teóricas. As informações do curso conjugadas com as aprendizagens feitas em laboratórios e a boa orientação, permitiram-me realizar um bom estágio na área da aquacultura. Questões fundamentais colocadas *a priori*, como "qual a pergunta concreta a que se quer responder?", "qual o número de replicados mínimo?", "existe aleatoriedade e independência dos replicados?" e uma muito importante a que me fui habituando, mas que no início raramente era respondida objectivamente "como vão ser tratados e testados os resultados?", permitiram-me realizar um bom estágio e desenvolveram em mim um gosto pela Ciência e Estatística.

Nada foi melhor, para perceber a importância destas questões, do que tentar passar pelos "terríveis" *referees* de boas revistas científicas na publicação do primeiro artigo científico como primeiro autor. Rapidamente me apercebi que é preferível perder mais duas ou três semanas a planear muito bem todas as experiências do que perder meses a justificar fraquezas (facilmente evitáveis) da experiência aos juízes científicos. Obviamente as experiências podem não ser perfeitas, mas é necessário saber que, como investigador, faço tudo para que o planeamento experimental esteja correcto e que o mesmo se adeque a responder à pergunta efectuada e saber qual o método estatístico que mais pode facilitar a compreensão dos dados. Devo também saber quais as fraquezas da experiência, dos dados obtidos e das análises estatísticas disponíveis e das que tenciono usar. Desta forma, não só respondo bem e facilmente à questão, como torno muito mais eficaz a publicação dos meus artigos científicos.

Exemplo prático

De forma a tornar mais clara a necessidade de um planeamento experimental adequado e uma estatística apropriada na investigação científica, vou dar alguns exemplos de algumas questões anteriormente colocadas e actualmente respondidas em artigos científicos já publicados pela equipa em que estive integrado.

Objectivos

Um dos estudos realizados por mim e pela equipa com quem trabalhei nos 3 últimos anos (cujo esforço levou à publicação de 2 artigos científicos, Penha-Lopes *et al*., 2005 e 2007), teve como objectivos, entre outros, comparar a sobrevivência larvar do *Mithraculus forceps* (caranguejo ornamental), tal como o desenvolvimento larvar, sob diferentes condições de cultivo (ex: densidade de cultivo, dietas, temperatura, etc…). Neste momento vou só mencionar o estudo de um desses factores, a densidade de cultivo inicial, ou seja, a densidade de larvas que colocamos inicialmente em cada tanque de cultivo.

Algumas das perguntas para as quais queríamos obter respostas eram as seguintes:

a) Qual a melhor densidade de cultivo inicial para o cultivo das larvas da espécie *M. forceps*?

b) Existe alguma diferença no desenvolvimento larvar (duração larvar, sincronismo na passagem de larva a juvenil) utilizando diferentes densidades de cultivo iniciais?

c) E tendo em conta a produtividade (número de larvas produzidas por tanque de cultivo) qual a melhor densidade de cultivo inicial a usar? (considerando que a disponibilidade larvar é infinita e o número de tanques de cultivo limitado)

Metodologia resumida

As experiências foram realizadas usando um sistema de cultivo larvar desenvolvido e descrito por Calado *et al*. (2003) e já utilizado no cultivo larvar desta espécie por Rhyne *et al*. (2004), Penha-Lopes et al. (2005) e algumas empresas de cultivo de crustáceos ornamentais. Foram mantidas em todas as experiências os seguintes factores: temperatura de 28ºC, salinidade de 35, pH de 8.0 a 8.2 e um fotoperiodo de 14h luz e 10h escuridão,.

Para descobrir a melhor densidade de cultivo inicial, a sobrevivência final foi comparada entre 4 tratamentos: 10, 20, 40 e 80 larvas L^{-1}, sendo fornecido a cada tanque *Artemia* spp. (presa) recém eclodida numa densidade de 12 naúplios.mL^{-1} (o máximo permitido pelo sistema de cultivo). Nesta experiência foram realizados 4 replicados (tanques de 10L).

A maioria dos recém-juvenis assentou no 9º dia depois das larvas serem libertas pelas fêmeas. No entanto, para o cálculo da sobrevivência até juvenil, foi contado o número de juvenis que assentou até ao dia 15. Para compreender os efeitos dos diversos tratamentos no desenvolvimento larvar, nos primeiros 4 dias do cultivo larvar, 20% do total das larvas colocadas inicialmente em cada tanque era amostrada e o estádio de desenvolvimento determinado. Para as densidades de cultivo de 40 e 80 larvas.L^{-1} amostrei apenas 10%, dado o elevado número de larvas que teriam de ser amostradas caso 20% das mesmas tivessem sido amostradas (80 e 160 larvas para os tratamentos de 40 e 80 larvas-L^{-1}, respectivamente). Nos dias 5, 7, 9, 11 e 13 todas as larvas eram observadas para a determinação do estádio de desenvolvimento. Convém referir que a técnica de amostragem e determinação do estádio larvar utilizadas não prejudicam o bom desenvolvimento das larvas.

Análise estatística

A análise estatística realizada para ajudar a responder a estas questões foi simples. A sobrevivência e o desenvolvimento das larvas (depois da transformação em arcseno, dado os valores de sobrevivência e desenvolvimento larvar serem medidas em percentagens) foram comparados entre os tratamentos (densidades de cultivo iniciais) utilizando uma ANOVA simples. Quando os requisitos de homogeneidade de variância não foram satisfeitos (testados utilizando o teste C de Cochran), testes não paramétricos equivalentes foram usados (Kruskal-Wallis). Os testes Tukey ou Dunn foram usados sempre que a ANOVA ou o Kruskal-Wallis, respectivamente, detectaram a existência de diferenças significativas entre os tratamentos. Os resultados foram considerados significativamente diferentes quando $p<0.05$.

Resultados

Os resultados obtidos da experiência e seu tratamento estatístico foram muito satisfatórios e biologicamente aceitáveis. O desenvolvimento larvar foi síncrono (2 dias no estádio de desenvolvimento Zoea I, 2 dias em Zoea II e 5 dias em Megalopa) sem diferenças significativas observadas entre as diferentes densidades de cultivo iniciais.

Tabela 1 – Sobrevivência até recém juvenil do caranguejo *M. forceps* em diferentes densidades de cultivo iniciais. Os valores são média ± desvio padrão [letras em sobrescrito indicam diferenças significativas] (Adaptado de Penha-Lopes *et al*., 2005)

		Sobrevivência a recém-juvenil (%)
Densidade de cultivo inicial (larvas.L^{-1})	10	81.5 ± 8.9^{a}
	20	$76.6\pm17.2^{a,b}$
	40	$63.9\pm22.8^{a,b}$
	80	32.9 ± 24.8^{b}

Relativamente aos dados da sobrevivência (Tabela 1), podemos observar que esta diminuiu significativamente com o aumento de densidade de "stock".

O tratamento com 40 larvas.L^{-1} foi aquele que permitiu aumentar a densidade de cultivo sem que a sobrevivência final fosse significativamente afectada, quando comparado com o tratamento com a sobrevivência final percentual mais elevada (10 larvas.L^{-1}).

Novas questões colocadas pela realidade

No entanto, os dados obtidos não me dizem que este valor é o óptimo. É bom não esquecer que as conclusões deste artigo científico têm como objectivo melhorar as condições de cultivo desta espécie pelos produtores. Esta necessidade de responder a situações práticas levou-me, juntamente com colegas de trabalho, a desenvolver modelos matemáticos, de forma a calcular valores óptimos de produção para cada um dos factores de cultivo já estudados. Dado estar a trabalhar com variáveis fixas e contínuas, os modelos matemáticos permitiram-nos encontrar valores óptimos entre os poucos valores testados experimentalmente, assim como calcular diversas variáveis dependentes que se vão alterando ao longo do tempo.

Actualmente, tais ferramentas permitem que os modelos desenvolvidos por investigadores possam ser utilizados na prática por quem mais precisa. A ciência, e especialmente a ciência aplicada, deve usar a estatística para se orientar e para ajudar a validar os seus resultados, mas também como um elo de ligação à sociedade em geral.

Desta forma, desenvolveram-se modelos matemáticos que nos vieram ajudar a obter os valores teóricos óptimos, tal como permitiram o cálculo da percentagem de larvas que fazem a metamorfose para juvenil ao longo do tempo, a duração larvar e o sincronismo de metamorfose (figura 1).

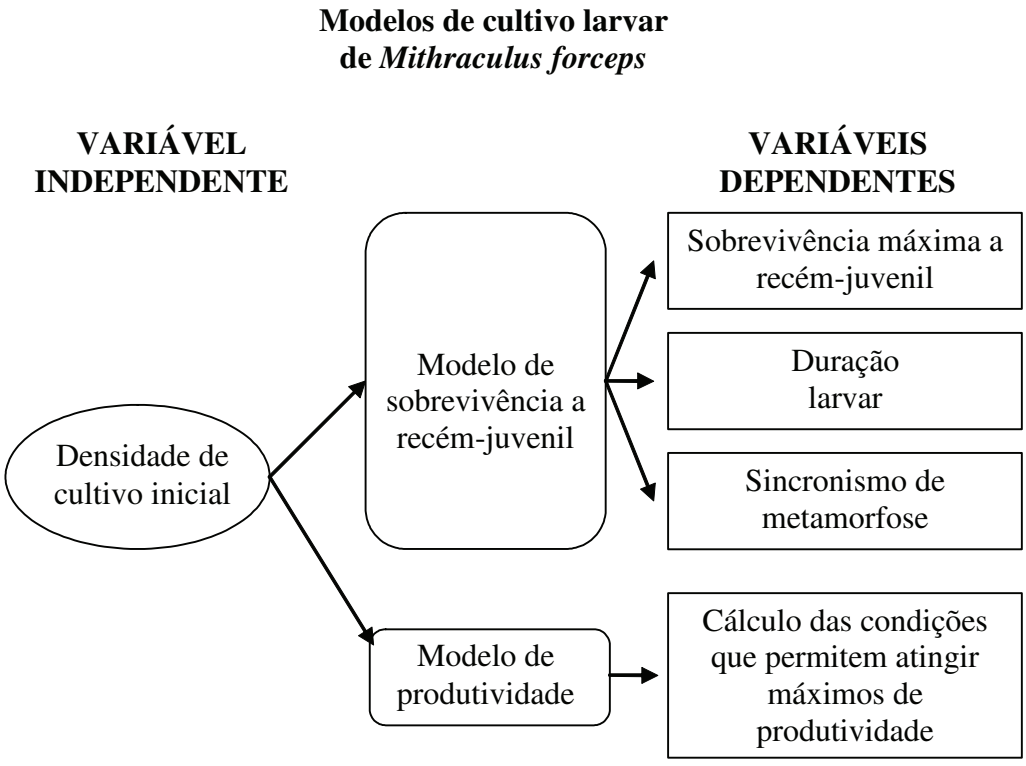


Figura 1 – Diagrama representativo da variável independente e das variáveis dependentes dos modelos de sobrevivência e de produtividade no cultivo larvar de *Mithraculus forceps*.

Nova metodologia e estatística resumidas

Para cada condição, a percentagem de larvas que sobreviveram até realizarem a metamorfose para recém-juvenil durante a experiência (sobrevivência cumulativa) foi modelada usando a biblioteca "nlme" (non-linear mixed-effects models) desenvolvida por Pinheiro e Bates (2000) no software R 2.2.1. Usámos modelos assimptóticos (equação 1) para descrever os efeitos da densidade de cultivo

inicial na sobrevivência até recém-juvenil.

$$y(x) = \phi_1 \times (1 - e^{(-e^{\phi_2} \times (x - \phi_3))}) \qquad (1)$$

Nesta formulação, os parâmetros do modelo (calculados usando a máxima verosimilhança) são:

- Φ_1 é a assímptota de $x \rightarrow + \infty$ (Tempo $\rightarrow + \infty$) e representa a sobrevivência final até recém-juvenil (%) esperada no 15º dia depois da libertação das larvas pelas fêmeas.

- Φ_2 uma avaliação da taxa de variação, obtida de $t_{0.5}$ = log2/exp(Φ_2), que nos dá uma ideia do sincronismo de assentamento; os valores mais elevados indicam um maior sincronismo;

- Φ_3 é o valor de x que corresponde a y = 0 (período correspondente a metamorfose = 0), o qual indica o tempo imediatamente anterior ao previsto da metamorfose da primeira larva a recém-juvenil (Pinheiro e Bates 2000).

Os efeitos dos factores na sobrevivência até recém-juvenil (Φ_1), o sincronismo de metamorfose (Φ_2) e o tempo imediatamente anterior ao previsto da metamorfose da primeira larva a juvenil (Φ_3) foram testados usando modelos lineares (incorporados no desenvolvimento do modelo em "nlme", descrito em detalhe por Pinheiro e Bates, 2000). Todos os resultados foram considerados significativamente diferentes quando $p<0.05$. O desenvolvimento do modelo inclui a realização de consecutivos testes de qualidade de ajustamento. Um modelo foi considerado bem ajustado quando os resíduos estandardizados fora do nível de confiança 95% (entre -1.96 e 1.96) eram minimizados e distribuídos aleatoriamente em torno de zero (Pinheiro e Bates, 2000).

Novos resultados

O modelo que melhor descreveu o efeito das densidades de cultivo iniciais na sobrevivência (%) até recém-juvenil foi:

$$Sobrevivência\ a\ recém-juvenil\ (\%) = a \times (1 - e^{-e^{(0.25)} \times (D-8)})$$

(D- Dias depois da eclosão das larvas, ou seja, da libertação das larvas pela fêmea)

onde

a=84.28 para densidade de "stock" de 10 larvas.L^{-1}
a=76.71 para densidade de "stock" de 20 larvas.L^{-1}
a=63.93 para densidade de "stock" de 40 larvas.L^{-1}
a=31.83 para densidade de "stock" de 80 larvas.L^{-1}

Felizmente, o modelo desenvolvido simulou muito bem os dados anteriormente obtidos, apresentando os seguintes resultados:

- O dia em que ocorreu a 1ª metamorfose para recém-juvenil (dia 8) não foi significativamente diferente (p=0.89) entre todas as densidades de stock (Φ_3=8);
- O sincronismo na metamorfose também não foi significativamente diferente (p=0.34) entre todas as densidade de stock (Φ_2=0.25);
- A sobrevivência final para recém-juvenil diminuiu significativamente (p<0.002) à medida que a densidade de stock aumentou (Φ_1=84.28, 76.71, 63.94 e 31.83 para 10, 20, 40 e 80 larvas.L^{-1}, respectivamente).

O modelo linear que melhor se ajustou à relação entre a produtividade e a densidade de stock (R^2=0.68) permite-nos prever que a produtividade é maximizada utilizando 60 larvas.L^{-1}:

$$Produtividade\ (juvenis.L^{-1}) = -4.129 + 9.258 \times Dens.Stock - 0.076 \times Dens.Stock^2$$

O valor "óptimo" calculado pelo modelo (60 larvas.L^{-1}) é bastante diferente da "melhor" densidade

de cultivo inicialmente sugerida aos produtores (40 larvas.L^{-1}), dado que o modelo consegue calcular a produtividade (variável dependente) obtida para todos os valores não testados da densidade de stock (variável independente).

Objectivos actuais

Dado que os modelos foram desenvolvidos para condições de cultivo bem definidas, o nosso objectivo actualmente é construir modelos holísticos (ex: Stella) que permitam que sejam colocadas no modelo qualquer combinação de variáveis independentes (determinada temperatura, pH, densidade de presa, densidade de cultivo, etc…) e solicitar ao modelo que calcule determinadas variáveis dependentes (sobrevivência, produtividade, tamanho dos juvenis, etc…). Tal ferramenta irá permitir que os modelos desenvolvidos por nós possam ser utilizados de uma forma simples e eficaz por quem mais precisa.

Questões relacionadas com a aleatoriedade e independência

Como em todos os estudos, determinadas decisões relativas ao planeamento experimental têm de ser feitas de forma a garantir a aleatoriedade e independência das amostras, não só por ser o mais correcto experimentalmente, mas porque a própria estatística o exige. Vou dar 2 exemplos de decisões que tiveram de ser feitas *a priori* permitindo a realização de uma investigação correcta.

Embora as fêmeas fossem capturadas no habitat natural em vésperas de lua cheia, período normal de libertação das larvas, existiu em laboratório um pequeno desfasamento dessa mesma libertação das larvas por parte das fêmeas. Este facto levou a que nem todos os replicados de cada tratamento começassem ao mesmo tempo. No entanto, de forma a preservar a aleatoriedade no planeamento experimental, foi decidido apenas iniciar replicados quando as larvas obtidas de diversas fêmeas (nessa mesma manhã) permitissem suprimir pelo menos um replicado de cada tratamento. Desta forma, o estudo usou larvas de várias fêmeas capturadas na natureza, tornando-o representativo, como permitiu que cada replicado (com larvas de várias fêmeas) não ficasse dependente da qualidade das larvas de uma determinada fêmea.

A outra questão colocou-se com os sistemas de cultivo. É de lembrar que nesta experiência testei 4 tratamentos diferentes com 4 replicados cada um, o que dá um total de 16 replicados (tanques). No laboratório onde realizei esta experiência existiam 4 sistemas de cultivo, cada um com 4 tanques de cultivo larvar. Até aqui as coisas são simples e claras. No entanto a questão que se colocou foi a seguinte: como distribuir os replicados pelos sistemas e respectivos tanques? Aqui não decidi ao acaso e dado que a divisão sistemática era possível, realizei um replicado de cada tratamento em cada um dos sistemas. Desta forma, em cada sistema de cultivo corria um replicado, onde a sua posição (tanque) foi determinada aleatoriamente. A outra decisão extrema seria a de colocar os 4 replicados de cada tratamento num determinado sistema, no entanto, neste caso específico, tal decisão seria injustificada e poderia colocar os resultados provenientes de cada tratamentos dependentes da qualidade e do funcionamento dos próprios sistemas.

Por exemplo, tal decisão poderia ser tomada caso estivéssemos a testar diferentes tipos de dieta, especialmente se as mesmas se diluem em água. Em geral, os sistemas de cultivo funcionam em circuito fechado, em que a água que sai de cada tanque é misturada com a dos outros tanques e, depois de filtrada, volta aos mesmos tanques desse sistema. Testar diferentes tipos de dietas num mesmo sistema não é assim aconselhável, no entanto o uso de dois ou mais sistemas por tratamento (ex: tipos de dieta) é aconselhável.

Comentário dos referees a algumas questões do estudo

Nesta secção vou colocar duas questões reais feitas por um *referee* científico quando eu tentava publicar o artigo numa das melhores revistas de aquacultura. É bom referir previamente que muitas vezes nem o *referee* nem o investigador têm razão, pois apenas têm perspectivas diferentes, quer do uso da estatística (exemplo 1) quer da função dos artigos científicos (exemplo 2).

1) Questão do *referee*: "Os autores amostraram 10% das larvas nas densidade de 40 e 80 larvas.L^{-1} e 20% das larvas nas densidades de 10 e 20 larvas.L^{-1} para o estudo do desenvolvimento larvar e asseguraram que cada estádio demorou um determinado período exacto. Porque é que os autores

usaram diferentes níveis de amostragem para a mesma experiência, podendo isto ter um impacto significativo nos resultados e análise estatística? Seria muito melhor usar um nível médio como os 15% para todos os tratamentos."

1) Resposta ao *referee*: "*A posteriori***, a análise de variância das nossas amostras demonstrou que com amostras de tamanho reduzido (n<10 em todos os casos) poderíamos obter a precisão desejada.

O protocolo foi desenhado antes da amostragem e nesta data não tínhamos qualquer informação sobre a variância dos dados, e por isso foi decidido retirar amostras de maiores dimensões, 20% do total da população quando estas eram pequenas (N < 20 larvas.L^{-1}) e 10% quando a população total era grande (N > 20 larvas.L^{-1}).

Muitos estatísticos e investigadores consideram que quanto maior a amostra melhor. No entanto, tal facto não é verdade. A "sobre-amostragem" pode causar problemas, dado que quanto maior a amostra, menor o erro padrão $s/n^{1/2}$, e como resultado pequenas diferenças triviais podem ser consideradas como significativas (Sokal, and Rohlf, 1995). Ou seja, por um lado queremos aumentar os nossos dados de forma a aumentar a precisão (20% para populações pequenas) e por outro não queremos ter "super-amostras" (10% para população grandes).

No entanto, no nosso caso,

N (larvas.tanque^{-1})	100	200	400	800
n (larvas amostradas)	20	40	40	80
$\sqrt{n}$	4,47	6,32	6,32	8,94

Parece realista acreditar que esta sobre-amostragem não afectou as nossas conclusões. No entanto, o *referee* sugere que devêssemos amostrar 15% em todos os casos, e o resultado seria:

N (larvas.tanque^{-1})	100	200	400	800
n (larvas amostradas)	15	30	60	120
$\sqrt{n}$	3,87	5,38	7,75	10,95

Desta forma é verdade que permitia uma melhor harmonização da experiência, mas do ponto de vista da análise estatística dos dados acreditamos que teria um pequeno impacto no aumento da precisão, e no caso de N=800 poderia haver o risco de considerar diferenças significativas que são, na realidade, apenas diferenças irrelevantes.

Por outro lado, para provar que amostrando "apenas" 10 e 20% da nossa população podemos garantir (com uma probabilidade > 95%) que todas as larvas estavam no estádio por nós determinado:
Para proporções $p \in (0,1)$, a pior análise possível de $\sqrt{(p(1-p)/n)}$ (a estimativa da variância) é p=0.5, $\sqrt{(p(1-p)/n)} \sim \sqrt{(0.25/n)}$. Em todos os casos obtidos, todos os elementos da amostra (p=1.0) estavam num determinado estádio, e desta forma, na perspectiva clássica, o intervalo de confiança degenera neste ponto de estimação.

Além das razões estatísticas, factos e conhecimento biológico suportam as conclusões do estudo, e por isso estamos convictos que as nossas observações, validadas pela estatística, permitem-nos afirmar que o desenvolvimento larvar para esta espécie, nas condições utilizadas, se comporta da maneira descrita no artigo."

2) Questão do *referee*: "É muito estranho observar que ambos os testes estatísticos Tukey e Dunn foram usados durante a análise estatística. É ainda mais problemático o facto de não ter sido feita referência por parte dos autores, que teste analisou determinado conjunto de dados."

(Nota: interessa mencionar que os dados apresentados neste artigo são referentes a apenas um dos factores testados e que no artigo científico foram testados mais de 4 factores)

2) Resposta ao *referee*: "Nesta experiência foi comparada a sobrevivência larvar, o estádio de desenvolvimento, a produtividade e a biometria aquando do cultivo larvar sob diferentes condições ambientais e metodológicas. Tomámos todos os cuidados recomendados ao realizar as análises estatísticas presentes neste artigo (Sokal and Rohlf, 1995). Em particular, foram observados os seguintes casos:
a) Normal qualidade de ajustamento
b) Homocedasticidade de variância
Sempre que a resposta foi afirmativa para ambas as questões, foram usados testes paramétricos padrão, dado que são mais eficientes.
A ausência de homocedasticidade pode ser remediada com o teste Smith-Welch- Satterthwaite, no entanto, heterocedasticidade está geralmente associada a um pobre ajustamento a modelos Gaussianos. Neste caso foram usados testes equivalentes não-paramétricos, dado que nenhum modelo foi assumido neste caso. A utilização de testes paramétricos seria complicada dado que $\overline{X}$ e S^2 são dependentes sempre que a distribuição parental é não-normal.
Caso o *referee* queira observar a precisão do nosso trabalho podemos enviar uma cópia dos dados e com uma descrição detalhada dos testes realizados em cada caso e as razões que levaram a essa escolha. Acreditamos que a introdução de uma descrição detalhada de cada teste utilizado vai tornar o artigo mais confuso do que esclarecedor, e neste caso não terá qualquer utilidade em ser publicado."

Aprendizagem
Apesar de não ter "oficialmente" concordado com o *referee*, aprendi a melhor planear as minhas experiências, devido ao que assimilei da resposta que tive de criar para responder às questões colocadas e ver o artigo publicado. É certo que o planeamento foi feito com cuidado, mas algumas questões foram levadas de ânimo leve. Por exemplo, hoje em dia, faço sempre que possível experiências preliminares, de forma a ter uma melhor sensibilidade da ordem de grandeza dos resultados e da variância associada. Desta forma posso calcular com um elevado grau de confiança a dimensão mínima da amostra que é necessária para ter uma boa precisão dos dados e uma elevada potência do teste estatístico escolhido. Ou seja, tento realizar boas experiências com elevada precisão e poder estatístico reduzindo ao máximo o esforço de amostragem.
Na segunda questão colocada pelo *referee*, aprendi a melhor organizar os meus dados e a realizar apenas testes estatísticos quando necessário, de forma a poder fornecer o maior número de dados e informação possível. Hoje em dia tento fornecer informação sobre o tipo de teste utilizado, a transformação realizada, o número de amostras e o valor de *p* obtido para cada comparação testada.
O relativo cuidado que tive ao preparar esta experiência e um pouco de bom senso sobre o real uso da estatística permitiu-me a publicação deste artigo. Em aquacultura, a realização de novas experiências por "ordem" do *referee* é quase sempre impossível, devido a uma variedade de factores (humanos, disponibilidade material, disponibilidade das larvas ou juvenis, etc…). Por isso aconselho um bom planeamento experimental, "temperado" com algumas mesas redondas com especialistas da área em questão, tal como estatísticos porreiros com pés bem assentes na terra.

A Estatística que actualmente um biólogo deve conhecer – vantagens, desvantagens e cuidados

Durante a minha curta carreira científica, compreendi que além deste à vontade com a estatística, a metodologia e os dados, o investigador deve possuir um bom conhecimento biológico da experiência e objectivos em questão. Além destas, deve também ter bom senso permitindo uma correcta interpretação biológica dos resultados e da sua aplicabilidade.
Como qualquer médico sabe usar o bisturi durante as suas cirurgias, o investigador deve saber usar a estatística durante os seus estudos. Não estou a dizer que o investigador deve conhecer a fundo cada teste ou compreender a sua formulação (pois o cirurgião também não percebe de metalurgia), no entanto, tem de saber quando usar a estatística e quais os testes que são mais apropriados à situação em

causa. Pelo menos o investigador deve compreender a linguagem estatística, permitindo-lhe assim solicitar ajuda a estatísticos e perceber os requisitos pedidos e os resultados obtidos. Só desta forma é possível usar a estatística para ajudar a chegar a conclusões práticas.
Em primeiro lugar, o investigador tem de ter consciência exacta do que quer saber: qual a pergunta a que quer responder. Ao formulá-la, o investigador levanta diversas questões que o ajudam a responder à questão colocada, nomeadamente a decisão da inclusão ou exclusão de determinados factores ou do tipo de dados a amostrar.
Depois, o investigador deve estudar bem a espécie em análise e a metodologia a usar e compreender as suas fraquezas. Se possível, realizar uns testes preliminares e obter uma maior sensibilidade dos dados obtidos (médias, variâncias, etc…), tal como saber se tenciona analisar dados emparelhados ou não, e decidir os níveis de resolução das amostragens (período entre amostragens, precisão do material a usar, etc…)
Relativamente à estatística, dependendo dos objectivos e do material em estudo, acredito que o investigador deve ter uma noção e saber quando usar os seguintes testes estatísticos:

a) Inferência sobre a média / comparação de médias:
   I. Um grupo :
      i. teste-*t*
      ii. teste Wilcoxon rank sum
   II. Dois grupos:
      i. teste-*t* emparelhado
      ii. teste-*t* não emparelhado
      iii. teste de Mann-Whitney
      iv. teste emparelhado de Wilcoxon
   III. 3 ou mais grupos
      i. ANOVA simples
      ii. Repeated measures ANOVA
      iii. ANOVA dupla
      iv. Teste de Kruskal-Wallis
      v. Teste de Tukey (testes paramétricos posteriores)
      vi. Teste de Dunn (testes não-paramétricos posteriores)

b) Regressões e correlações:
   I. Regressões lineares
   II. Correlações lineares (Pearson)
   III. Correlações não-paramétricas (Spearman)

c) Análise de tabelas de contingência
   I. teste do Qui-quadrado

d) Análise multivariada
   I. Escalonamento multidimensional (MDS)
   II. Análise de componentes principais (PCA)
   III. Análise BEST
   IV. Testes posteriores (ex: Anosim)

Estes são alguns dos testes que eu tenho utilizado ultimamente. É também muito importante saber olhar para os gráficos dos resultados e dos resíduos (e, naturalmente, saber executar esses gráficos), de forma a melhor compreender como se comportam os dados e melhor decidir possíveis transformações e testes a realizar.

Muitas vezes, quer por parte dos autores de artigos científicos, quer por parte dos *referee*s, é dado um excessivo valor aos testes de significância, retirando a responsabilidade das conclusões ao cientista. Dado que iniciei o meu estágio na área da aquacultura, onde o objectivo era desenvolver e melhorar protocolos de cultivo de espécies ornamentais, muitas vezes fui confrontado por parte dos produtores

com a questão: "vale a pena alterar o protocolo de cultivo para estas novas condições?". Inicialmente, utilizando apenas regressões, testes univariados paramétricos ou não-paramétricos, poderia responder se existiam diferenças significativas cultivando a 28ºC ao invés dos 26ºC, ou reduzindo a densidade de cultivo, sempre com base nos dados de sobrevivência e crescimento obtidos durante e no final da experiência. No entanto, cedo tive de perceber que para um produtor sobrevivência não é idêntico a produtividade, e desenvolvendo equações de produtividade com base nos dados obtidos as conclusões eram ligeiramente diferentes. Para além disto, também compreendi que na prática, um aumento de produtividade com determinadas condições, mesmo não sendo significativamente diferente (usando o globalmente aceite α=0.05), pode ser altamente vantajoso quando estamos perante espécies de elevado valor comercial (onde um pequeno aumento de produção equivale a muitos milhares de Euros), ou altamente desvantajoso se a tecnologia necessária para realizar o novo protocolo for cara ou dispendiosa. A estatística deve ser usada como instrumento para facilitar a tomada de decisões e não como avaliador e juiz, que considera um resultado significativo ou não comparando-o com um valor estipulado *a priori* e globalmente aceite (p=0.05).

É bom relembrar que a estatística é para ser utilizada quando os dados não conseguem falar por si, e que muita da variabilidade pode ser resolvida com um bom planeamento experimental e com um bom controlo das diversas variáveis (sempre que possível). Só quando queremos encontrar pequenas diferenças e estamos perante uma grande variabilidade é que devemos usar a estatística para nos ajudar a validar conclusões desses mesmos dados. A estatística deve ser uma ferramenta utilizada para chegar a conclusões simples e que ajude a validar a interpretação biológica dos resultados.
Sinto também por vezes, dado a estatística ser sempre bem recebida pelos *referee*s de revistas com elevado "factor de impacto", que existe um uso abusivo da mesma, permitindo por um lado uma publicação mais fácil, mas muito provavelmente tornando a publicação (com o objectivo de divulgar ao grande público) de difícil acesso à generalidade da população, científica ou não.
É também verdade que muitas vezes os investigadores são levados a "encontrar" diferenças significativas dado que as revistas científicas tendem a premiar resultados com diferenças significativas com metodologias razoáveis a dados inconclusivos com excelentes metodologias. Creio que para a ciência ser levada a sério, palavras como "dados inconclusivos", conceitos como "intervalos de confiança", e a utilização de bons planeamentos experimentais deveriam ser usados ao invés de complexas análises estatísticas.

Recomendo uma interacção mais estreita entre biólogos e estatísticos de forma a melhorar a ciência que se faz e a promover os seus resultados e aplicabilidade. Sugiro a utilização de programas estatísticos como o Graphpad Instat e o Graphpad Prism para jovens investigadores, o Statistica para quem já sabe o que quer e o Primer para quem necessita de realizar análises multivariadas. Aconselho também a leitura dos livros referenciados durante este artigo e mais alguns.
Desejo a todos a continuação de uma boa investigação biológica e estatística.

Referências bibliográficas:

- Motulsky, H. (1995). *Intuitive Biostatistics*. Oxford University Press, New York.
- Penha-Lopes, G., Rhyne, A., Lin, J., Narciso, L. (2005). "The larval rearing of the marine ornamental crab, *Mithraculus forceps* (A. Milne Edwards) (Decapoda: Brachyura: Majidae)". *Aquaculture Research* **36**, 1313-1321.
- Penha-Lopes, G., Figueiredo, J., Narciso, L. (2007). "Modelling survival and growth of *Mithraculus forceps'* larvae and juveniles (A. Milne Edwards, 1875) (Decapoda: Brachyura: Majidae) in aquaculture". *Aquaculture* **264**, 285-296.
- Pinheiro, J.C., Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS - Statistics and Computing*. Springer-Verlag New York.
- Rhyne, A.L., Penha-Lopes, G., Lin, L. (2004). Growth, development and survival of larval *Mithraculus sculptus* (Lamark) and *M. forceps* (A. Milne Edwards) (Decapoda: Brachyura: Majidae): economically important marine ornamental crabs. *Aquaculture* **245**, 183-191.
- Sokal R. & Rohlf F. (1995). *Biometry - The Principles and Practice of Statistics in Biological Research*, 3rd edn. Freeman, NewYork.

Livros recomendados:

- Boniface, D.R. (1995). *Experimental Design and Statistical Methods – for Behavioural and Social Research*. Ed. Champan & Hall, London.
- Cobb, G.W. (1998). *Introduction to Design and Analysis of Experiments*. Key College Publishing, California, US. 802 pp.
- Gilbert, N. (1989). *Biometrical Interpretation – Making Sense of Statictics in Biology*, 2nd edn., Oxford University Press, New York.

Análise e comparação de sequências biológicas através de funções vectoriais

Susana Vinga, svinga@algos.inesc-id.pt

Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID)
Departamento de Bioestatística e Informática, Faculdade de Ciências Médicas – Universidade Nova de Lisboa (FCM/UNL)

Introdução

A análise de sequências biológicas constitui uma importante área de investigação em bioinformática e biologia computacional. Estas disciplinas cresceram de forma notável nos últimos anos, sobretudo durante o desenvolvimento e após a finalização de projectos de sequenciação de grande envergadura que, um pouco por todo o mundo, têm alterado significativamente métodos e abordagens para o estudo da biologia molecular. A análise de dados experimentais quantitativos produzidos em número cada vez maior irá, sem sombra de dúvida, alterar paradigmas em biologia e fortalecer a sua componente preditiva. O objectivo desta nova área da bioinformática é a criação de novos algoritmos para avaliação de relações entre elementos de grandes conjuntos, a análise e interpretação de vários tipos de dados, incluindo DNA e proteínas, e desenvolvimento e implementação de ferramentas que permitam o acesso e gestão eficientes dos diversos tipos de informação disponíveis.

Este artigo não pretende ser uma exposição exaustiva sobre a análise de sequências biológicas, e reflecte naturalmente a experiência pessoal do autor, com todos os enviesamentos que isso acarreta. O objectivo é servir como introdução a um tema relevante num campo onde a Estatística poderá continuar a desempenhar um papel importante.

Provavelmente o DNA (ácido desoxirribonucleico), o RNA (ácido ribonucleico) e as proteínas são as moléculas mais conhecidas ao grande público, devido à grande divulgação de temáticas relacionadas com a genética e as suas possíveis aplicações. Estas sequências representam macromoléculas fundamentais a todos os seres vivos. Por meio de processos complexos que ocorrem em todas as células, aqui descritos de forma bastante simplificada, a informação armazenada nos genes permite construir uma plêiade de proteínas que desempenham variadíssimas funções, desde catalisar reacções químicas (enzimas) até alterar, modular e controlar outros processos de forma complexa através de redes de regulação.

Muito recentemente foi possível conhecer, para um conjunto variadíssimo de espécies, a respectiva constituição genómica, i.e., a descrição química do seu DNA na sua totalidade. Esta molécula é constituída por duas cadeias de nucleótidos ou bases, que podem ser de quatro tipos distintos. É, por isso, usualmente representada como uma sequência linear "escrita" num alfabeto com quatro símbolos. A questão fundamental ainda por responder é em que modo a informação contida na sequência dá origem a estruturas e processos celulares tão complexos e variados. Quais são os processos inerentes a este controlo ou, por outras palavras, como é transmitida a informação, codificada linearmente, para diversos níveis de organização molecular e celular? Visto que o material genético é idêntico em todas as células de um organismo (por exemplo, um neurónio, uma célula da epiderme, uma outra do fígado), o que provoca a sua diferenciação?

A análise desta enorme quantidade de informação, que tem crescido exponencialmente nos últimos anos, só é praticável se forem conjugados esforços de diversas áreas e disciplinas. Neste contexto a matemática aplicada, a probabilidade e estatística, as ciências da computação a informática desempenham hoje em dia um papel fundamental no estudo e análise destes dados, dificilmente

tratáveis sem recorrer às técnicas mais sofisticadas destas disciplinas. É de esperar que o impacto desta abordagem multidisciplinar seja elevado, nas áreas "bio" até à medicina e ciências da saúde, sendo a chave para a resolução de muitos problemas em aberto. Os bioestatísticos têm, neste contexto, um desafio importante pela frente, e também o perfil adequado para aplicar com sucesso a sua ciência a estas áreas (Ewens and Grant 2001). Inversamente é habitual encontrar nestes novos problemas uma fonte de inspiração para o desenvolvimento de metodologias em tópicos menos aplicados, enriquecendo o diálogo entre disciplinas (Robin, Rodolphe and Schbath 2005).

A passagem de uma análise *in silico* – computacional – para dados reais com significado biológico será o objectivo e aspiração finais desta interacção.

Métodos para análise e comparação de sequências

A análise e comparação de sequências constituem ainda uma área fulcral a todas as aplicações bioinformáticas (Durbin, Eddy, Krogh and Mitchison 1998), apesar do desenvolvimento de técnicas que abordam outros níveis de organização molecular e celular. Muitas das tarefas quotidianas dos biólogos dependem de algoritmos que processam e investigam palavras e sequências, desde a procura de DNA em bases de dados à classificação de proteínas.

As sequências biológicas são usualmente representadas como palavras cujos símbolos pertencem a um alfabeto – de bases ou nucleótidos para o DNA e aminoácidos para as proteínas.

Alinhamentos simples e múltiplos

Os métodos de análise e comparação baseados em alinhamento das sequências constituem o paradigma principal nesta disciplina. Estas técnicas consistem na sobreposição de sequências de forma a salientar as suas partes ou segmentos comuns. Desta forma é possível identificar mutações pontuais, inserções, delecções e outros acontecimentos que poderão ter tido lugar na evolução destas mesmas sequências. É, assim, possível inferir o processo pelo qual as sequências se modificaram ao longo de gerações, criando a variabilidade que observamos actualmente em todos os organismos vivos. Talvez a melhor figuração deste algoritmo seja a comparação com o deslindar da escrita hieroglífica por Champollion a partir da pedra de Rosetta, em que o mesmo texto fora escrito de forma alinhada em três línguas distintas (hieroglífica, demótica e grega), o que permitiu inferir a linguagem desconhecida a partir das duas já estudadas.

Há já algoritmos muito eficientes de alinhamento simples entre duas sequências, sendo os mais famosos o de Needleman-Wusch (Needleman and Wunsch 1970) para optimização global e o equivalente local dado pelo algoritmo de Smith-Waterman (Smith and Waterman 1981). Os dados de entrada são as penalizações dadas 1) ao alinhamento de dois símbolos e 2) à criação de uma abertura (*gap*), i.e., de alinhar um segmento com o conjunto vazio, criando uma espécie de argola. Esta informação é usualmente representada através de uma matriz quadrada (simétrica), quer para DNA quer para proteínas, e valores reais, negativos, para criar e estender uma abertura. A implementação deste algoritmo de programação dinâmica permite optimizar a pontuação global e encontrar o melhor alinhamento possível, i.e., para o qual a pontuação obtida é máxima. Existe software muito eficiente baseado nestes algoritmos que permite comparar uma determinada sequência com os largos milhões existentes em bases de dados, devolvendo o conjunto de sequências mais próximas, como por exemplo a ferramenta BLAST (Altschul, Gish, Miller, Myers and Lipman 1990).

É também de referir que houve um grande investimento no estudo da significância estatística dos alinhamentos obtidos, tendo sido desenvolvida teoria para determinação das distribuições de *scores* de alinhamento sob diversos cenários (Waterman 1995). Desta forma é possível, a partir de um dado

valor, estimar a probabilidade de estarmos perante um alinhamento espúrio, i.e., de sequências não correlacionadas. Há também algoritmos para alinhar diversas sequências criando alinhamentos múltiplos, implementadas também, e.g., no programa CLUSTAL (Thompson, Higgins and Gibson 1994).

Métodos independentes do alinhamento ou *alignment-free*

Há, no entanto, outros métodos substancialmente diferentes de analisar sequências que não estão dependentes de nenhum pré-alinhamento das mesmas. São chamados genericamente métodos sem alinhamento ou *alignment-free*. Neste tipo de técnica as sequências são previamente transformadas e subsequentemente analisadas num determinado espaço imagem, usualmente $\mathbb{R}^n$. Serão apresentadas com mais detalhe neste artigo apenas funções baseadas em composição de palavras de tamanho fixo e mapas iterativos para representação do DNA. Recentemente foi feita uma revisão destas técnicas (Vinga and Almeida 2003, Vinga 2007), onde se encontra uma descrição mais completa destas metodologias.

Sequências, palavras e motivos

Uma sequência *X* pode ser representada como uma sucessão de *N* símbolos de um dado alfabeto $\mathcal{A}$, de tamanho *r*, i.e., $X \equiv s_1 s_2 \cdots s_i \cdots s_N$, $i = 1,\ldots,N$ e $s_i \in \mathcal{A}$. Para sequências de DNA este alfabeto é composto pelos quatro símbolos que representam as 4 bases $\mathcal{A} = \{A,T,C,G\}$ enquanto que para as proteínas cada um dos símbolos representa um aminoácido. Para linguagens naturais como o Português ou Inglês, este conjunto é composto por todos os possíveis caracteres de cada idioma.

Dentro deste contexto podemos designar um segmento de *L* símbolos, com $L \leq N$, por L-tuple ou L-palavra (em algumas referências aparece também a designação L-word, L-plet, L-mer). O conjunto $\mathcal{W}_L = \{w_{L,1}, w_{L,2}, \ldots, w_{L,K}\}$ de todos os possíveis L-tuples terá $K=r^L$ elementos.

É também usual chamar ***motivo*** aos elementos deste conjunto. Um motivo será um pequeno segmento que terá significado biológico e funcional. Por exemplo, é conhecido que há motivos específicos, como o descrito pela sequência TATAAT (denominada TATA-box) que possui um papel fundamental de reconhecimento de regiões promotoras no genoma de eucariotas. Podemos interpretar estes motivos como códigos moleculares ou pequenas unidades funcionais identificados pela maquinaria celular que irá processar essa informação em processos bioquímicos determinados.

A identificação e contagem dos L-tuples existentes numa sequência *X* é geralmente efectuada utilizando uma janela deslizante de comprimento L e contabilizando todas as ocorrências de cada palavra, desde a posição 1 até à posição $n–L+1$, que corresponderá ao número máximo de palavras contidas na sequência $c_L^X = \left(c_{L,1}^X, \ldots, c_{L,K}^X\right)$.

De forma equivalente as frequências de cada palavra f_L^X estimam as probabilidades $p_L^X = \left(p_{L,1}^X, \ldots, p_{L,K}^X\right)$ de encontrar palavras específicas, $w_{L,i}^X$, $i = 1,\ldots,K$, na sequência original. Este vector de frequências é calculado de forma trivial a partir de c_L^X:

$$f_L^X = \frac{c_L^X}{\sum_{j=1}^{K} c_{L,j}^X} \Leftrightarrow f_{L,i}^X = \frac{c_{L,i}^X}{n-L+1}$$

Por conveniência este vector é muitas vezes indexado directamente pela palavra respectiva em vez da ordem relativa, i.e. $f_{w_L} \equiv f_{s_{i_1}\cdots s_{i_L}}$ representa a frequência da palavra $w_L \equiv s_{i_1}\cdots s_{i_L}$, sendo que a diferença entre as duas representações deverá ser evidente a partir do contexto onde aparecem.

Como exemplo simples, para uma sequência de DNA, cujo alfabeto é $\mathcal{A} = \{A,T,C,G\}$ e r=4, uma palavra de tamanho três, L=3, poderá ser w_3=*GTC*. Para a sequência *X*=*GTGTGA,* com n=6, o vector p_3^X é estimado pela frequência relativa de todos os tri-nucleótidos. Estas frequências, determinadas através de uma janela deslizante de tamanho três, $n-L+1$=4 vezes, seriam:

$$W_L = \{GTG, TGT, TGA, AAA, AAC, \ldots\}$$
$$c_3^X = (2,1,1,0,0,\ldots)$$
$$f_3^X = (0.5,0.25,0.25,0,0,\ldots)$$

Os vectores c_3^X e f_3^X terão comprimento $K=4^3=64$ e as coordenadas iguais a zero correspondem a palavras inexistentes, neste caso tri-nucleótidos não representados. Se utilizarmos uma representação alternativa poderíamos escrever $f_{GTG}^X = 0,5$.

Funções e mapas vectoriais

Uma função vectorial é uma transformação cuja imagem está contida em $\mathbb{R}^n$. Dada uma sequência $X \equiv s_1 s_2 \cdots s_i \cdots s_N$, $i = 1,\ldots,N$ de um dado alfabeto $\mathcal{A}$, pode-se definir uma função vectorial $F : \mathbf{S} \to \mathbb{R}^n$ que transforma *X* num vector *n*-dimensional $x = (x_1, x_2, \ldots, x_n)$, $F(X) \in \mathbb{R}^n$.

A seguinte Figura 1 ilustra alguns passos que conduzem à transformação de sequências em vectores *n*-dimensionais e alguns métodos de pós-processamento revistos neste artigo.

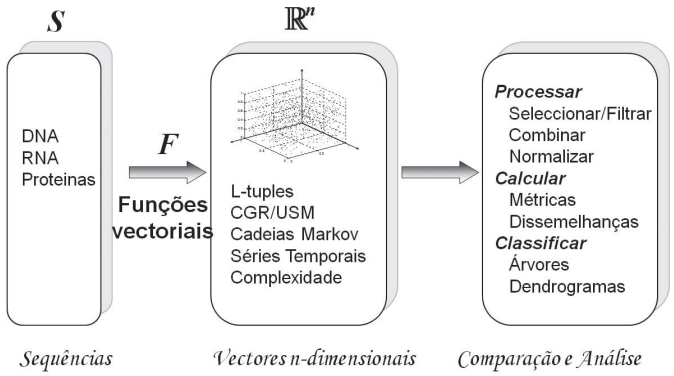


Figura 1. Funções vectoriais de sequências. Esta figura representa a transformação de uma sequência biológica *S* num vector real *n*-dimensional, o que pode ser representado por $F : \mathbf{S} \to \mathbb{R}^n$. Este artigo revê diversas metodologias que estudam e analisam as sequências neste espaço imagem.

É possível identificar diversas transformações de sequências em vectores. As secções seguintes descrevem mapas vectoriais baseados em composição de L-tuples e também em funções iterativas, nomeadamente representação por jogos do caos (CGR). Para uma revisão de outras representações baseadas em modelos de Markov, series temporais, teoria da informação e complexidade remete-se para outras referências (Vinga 2007).

Composição de L-Tuples

Uma das tarefas mais simples e intuitivas para analisar sequências foi a de calcular as frequências

relativas dos motivos que ocorrem em DNA e em proteínas, i.e., a sua composição em bases, aminoácidos ou 1-tuple. Esta abordagem foi posteriormente generalizada para palavras mais longas e são genericamente chamados métodos linguísticos.

A função vectorial considerada neste caso é simplesmente dada pela frequência de cada L-tuple.

$$F : \mathbf{S} \to \mathbb{R}^n$$
$$F\left(X\right) = f_L^S = \left(f_{L,1},\ldots,f_{L,n}\right), \text{ com } n = r^L = K$$

Estes vectores poderão ser subsequentemente processados de forma a extrair informação relevante.

Representação por jogo do caos CGR e sistemas de funções iterativas

A representação de sequências através de jogos do caos (*Chaos Game Representation – CGR* na terminologia inglesa) foi proposta em 1990 (Jeffrey 1990) como um método de transformar DNA em conjuntos de vectores bidimensionais. Pela maneira como é construída a representação, o CGR está directamente relacionado com sistemas de funções iterativas, recolhendo forte inspiração sobretudo na Teoria do Caos (Edgar 1990) e geometria fractal.

O algoritmo para construção de um mapa CGR parte de um ponto inicial no quadrado $[0,1]^2$ e calcula, iterativamente, para cada símbolo no DNA, a sua representação a partir do ponto imediatamente anterior. Formalmente, a transformação $x_i \in \mathbb{R}^2$ de uma sequência de DNA de tamanho N $X \equiv s_1 s_2 \cdots s_N$, $i = 1,\ldots,N$, $s_i \in \mathcal{A} = \{A,C,G,T\}$ é dada por:

$$\begin{cases} x_0 \sim Unif\left(0,1\right)^2 \\ x_i = x_{i-1} + \frac{1}{2}\left(y_i - x_{i-1}\right), i = 1,\ldots,N \end{cases} \text{ onde } y_i = \begin{cases} (0,0) & \text{se } s_i = 'A' \\ (0,1) & \text{se } s_i = 'C' \\ (1,0) & \text{se } s_i = 'G' \\ (1,1) & \text{se } s_i = 'T' \end{cases}$$

Este procedimento parte de um quadrado $[0,1]^2$ em que cada vértice representa cada um dos símbolos do alfabeto DNA. O algoritmo consiste em seleccionar um primeiro ponto aleatoriamente nesse conjunto e caminhar do ponto presente/actual na direcção do vértice que representa o símbolo imediatamente seguinte na sequência original, percorrendo exactamente metade dessa distância. Este processo é repetido até ao último símbolo *N*, calculando sempre a i-ésima posição x_i a partir da imediatamente anterior x_{i-1}. O gráfico final de todos os pontos x_i origina um padrão semelhante a um fractal, como exemplificado na Fig.2A. Estes mapas podem também ser interpretados como mapas contractivos, que em cada iteração transformam um determinado conjunto por homotetia de razão 0,5 e centrada em cada um dos vértices.

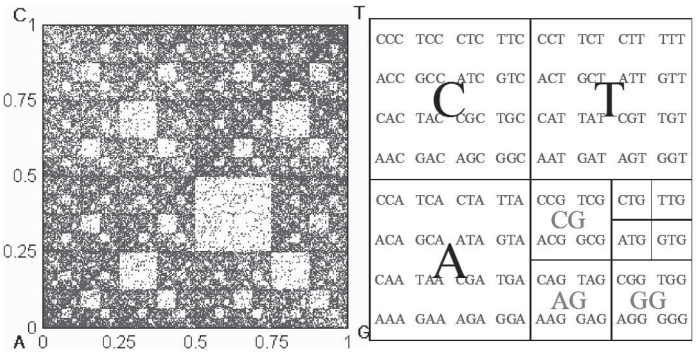


Figura 2. Exemplo de um mapa CGR e ilustração da propriedade dos sufixos. A) Exemplo da aplicação do algoritmo CGR descrito – imagem obtida para o gene da beta-globina humana, cromossoma 11 (HUMHBB), constituído por 73308 bases. Corresponde a duas das quatro proteínas constituintes da hemoglobina. B) Exemplo da propriedade dos sufixos: sequências que partilham o mesmo sufixo estarão no sub-quadrado com o respectivo rótulo da sub-palavra.

Uma das propriedades mais interessantes e notáveis desta representação é a sua capacidade em agrupar sufixos iguais. De facto, cada sufixo, independentemente do seu comprimento e da posição onde ocorre na sequência original, será mapeado no mesmo sub-quadrante, como exemplificado na Fig.2B. Este resultado é de fácil dedução para o caso de sufixos de comprimento 1 – qualquer que seja o ponto $x_i \in [0,1]^2$, se encontrarmos o símbolo A na sequência original iremos parar necessariamente ao quadrante com rótulo A na figura, devido à propriedade geométrica de contracção. O que é interessante é a repetição desta propriedade para sufixos mais longos – o motivo GA será mapeado no quadrante A e, dentro deste, no sub-quadrante G, numa repetição miniaturizada do quadrado maior, e assim sucessivamente, criando um padrão com propriedades fractais, onde cada uma das partes é uma versão pequena do todo.

Esta função pode também ser interpretada como uma representação binária de uma sequência. Por exemplo, a partir das coordenadas de um símbolo na base 2 é possível extrair todos os símbolos precedentes. Um determinado ponto com coordenadas (em base 10) x_i = $(0.65625, 0.21875)_{10}$ = $(0.10101, 0.00111)_2$ corresponde à sucessão de símbolos (1,0)→(0,0)→(1,1)→(0,1)→(1,1), i.e. ao sufixo GATC, pois as operações de divisão por 2 equivalem, em base 2, a deslocar a vírgula um algarismo para a direita (para mais propriedades consultar (Vinga 2005)).

Neste caso teremos uma colecção F_i de funções vectoriais, uma por cada símbolo num determinado contexto $i = 1,\ldots,N$, o que pode ser descrito como:

$$\begin{aligned} &F_i : S \to \mathbb{R}^n \\ &F_i(S) = (x_{i1}, x_{i2}) \end{aligned}, \text{ com } n = 2$$

Há já diversas aplicações desta representação em bioinformática, tais como a investigação de padrões em DNA, a extracção de matrizes de probabilidade de transição (Almeida, Carriço, Maretzek, Noble and Fletcher 2001), o cálculo de entropias (Vinga and Almeida 2004, Vinga 2005), e generalizações deste algoritmo para alfabetos de maior cardinalidade (Almeida and Vinga 2002) mantendo todas as propriedades originais. A literatura baseada em mapas CGR cresceu significativamente nesta última década, o que antecipa a sua importância no futuro como modelo geral para análise de sequências.

Comparação de vectores: métricas e medidas de dissemelhança

As secções anteriores descreveram vários métodos destinados a representar sequências como vectores *n*-dimensionais, através da definição de funções vectoriais. Esta secção descreve alguns procedimentos para comparar os vectores obtidos através da definição de medidas de dissemelhanças, estimando desta forma a proximidade entre as sequências originais que aqueles representam. Este objectivo é atingido definindo métricas apropriadas no espaço imagem $\mathbb{R}^n$, o que pode ser segregado da representação em si.

Formalmente, um espaço métrico (**S**, *d*) é um conjunto **S** equipado com uma função não negativa $d : \mathbf{S} \times \mathbf{S} \to \mathbb{R}_0^+$ que satisfaz, para quaisquer $X, Y, Z \in \mathbf{S}$, as propriedades de positividade, simetria e desigualdade triangular:

$$\begin{aligned} &d(X,Y) = 0 \Leftrightarrow X = Y \\ &d(X,Y) = d(Y,X) \\ &d(X,Y) \le d(X,Z) + d(Z,Y) \end{aligned}$$

Esta função *d* é uma métrica e mede a distância *d*(*X*,*Y*) entre pares de pontos *X* e *Y* em **S**, o conjunto de

todas as sequências.

Há, no entanto, funções $d(X,Y)$ que não verificam todas estas propriedades (por exemplo poderá não ser cumprida a desigualdade triangular), mas que poderão ter alguma relevância em determinadas aplicações. Neste caso, é mais adequado chamar a $d(X,Y)$ a dissemelhança entre as duas sequências.

Dissemelhança	**Equação**
Euclidiana	$d_L^E(X,Y)=\left(c_L^X-c_L^Y\right)^T\cdot\left(c_L^X-c_L^Y\right)=\sum_{i=1}^{K}\left(c_{L,i}^X-c_{L,i}^Y\right)^2$
Euclidiana ponderada	$d^2(X,Y)=\sum_{L=l}^{u}\sum_{i=1}^{K}\rho_i\left(c_{L,i}^X-c_{L,i}^Y\right)^2$
Euclidiana estandardizada	$d_L^{SE}(X,Y)=\left(c_L^X-c_L^Y\right)^T\cdot\left[\mathrm{diag}\left(s_{11},\cdots,s_{KK}\right)\right]^{-1}\cdot\left(c_L^X-c_L^Y\right)=\sum_{i=1}^{K}\frac{\left(c_{L,i}^X-c_{L,i}^Y\right)^2}{s_{ii}}$
	$d^{SE*}=\sum_{L=l}^{u}d_L^{SE}$
Mahalanobis	$d_L^M(X,Y)=\left(c_L^X-c_L^Y\right)^T\cdot\mathbf{S}^{-1}\cdot\left(c_L^X-c_L^Y\right)=\sum_{i=1}^{K}\sum_{j=1}^{K}\left(c_{L,i}^X-c_{L,i}^Y\right)\cdot s_{ij}^{inv}\cdot\left(c_{L,j}^X-c_{L,j}^Y\right)$
	$d^{M*}=\sum_{L=l}^{u}d_L^M$
Coeficiente de correlação linear	$d_L^{LCC}(X,Y)=\frac{K\sum_{i=1}^{K}f_{L,i}^X\cdot f_{L,i}^Y-\sum_{i=1}^{K}f_{L,i}^X\cdot\sum_{i=1}^{K}f_{L,i}^Y}{\left[K\sum_{i=1}^{K}\left(f_{L,i}^X\right)^2-\left(\sum_{i=1}^{K}f_{L,i}^X\right)^2\right]^{1/2}\cdot\left[K\sum_{i=1}^{K}\left(f_{L,i}^Y\right)^2-\left(\sum_{i=1}^{K}f_{L,i}^Y\right)^2\right]^{1/2}}$
Kullback-Leibler	$d_L^{KL}(X,Y)=\sum_{i=1}^{K}f_{L,i}^X\cdot\log_2\left(\frac{f_{L,i}^X}{f_{L,i}^Y}\right)$
Coseno	$d_L^{COS}(X,Y)=\theta_{XY}$, onde $\cos\left(\theta_{XY}\right)=\frac{\left(c_L^X\right)^T\cdot c_L^Y}{\left\|c_L^X\right\|\cdot\left\|c_L^Y\right\|}=\frac{\sum_{i=1}^{K}c_{L,i}^X\cdot c_{L,i}^Y}{\sqrt{\sum_{i=1}^{K}\left(c_{L,i}^X\right)^2}\cdot\sqrt{\sum_{j=1}^{K}\left(c_{L,j}^Y\right)^2}}$
Evolucionária	$d_L^{EVOL}(X,Y)=-\ln\left[\left(1+\cos\theta_{XY}\right)/2\right]$
CGR/USM	$d^{USM}(a,b)=-\log_2\left(\max_i\left\|a_i-b_i\right\|\right)$
Complexidade de Kolmogorov	$d^{KC}(X,Y)=1-\frac{K(X)-K(X\mid Y)}{K(XY)}$

Tabela 1. Definição de medidas de dissemelhança entre duas sequências X e Y, $d(X,Y)$. Estas medidas são baseadas na comparação dos vectores de contagens, frequências ou outros métodos onde a resolução L não é requerida.

A Tabela 1 apresenta algumas destas medidas, referenciadas de forma mais completa nos artigos de revisão já referidos (Vinga, et al. 2003, Vinga 2007).

De forma a ilustrar a capacidade destes métodos para classificação de sequências de qualquer tipo, a figura seguinte mostra um exemplo curioso do dendrograma obtido para dez línguas da União Europeia, a partir de um texto extraído do site da UE. Foi usada uma combinação de distâncias euclidianas entre L-tuples, com L a variar entre 1 e 4. Reconhecem-se claramente as separações linguísticas principais, nomeadamente o ramo representativo das línguas com raiz latina (sendo o Espanhol e o Português as mais próximas em termos absolutos) separadas claramente das germânicas.

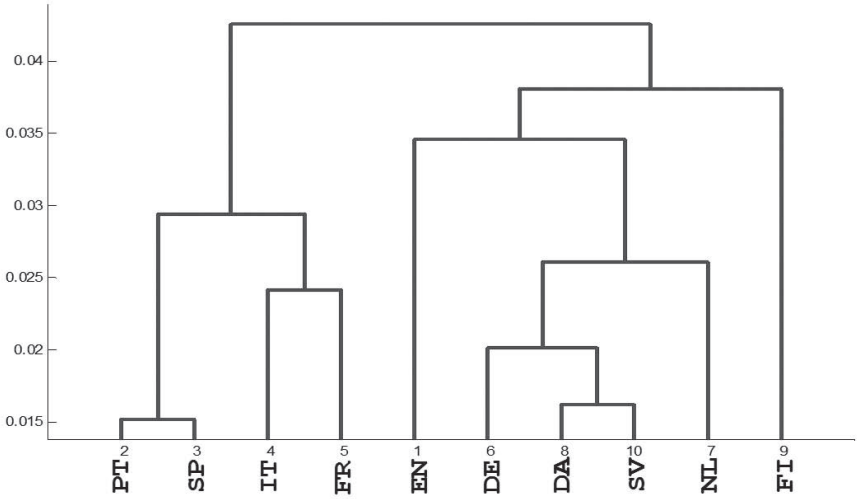


Figura 3. Dendrograma com a classificação de dez línguas da UE utilizado uma distância euclidiana combinada de L=1…4, $d^{E*}(X,Y)=\sum_{i=1}^{4} d_L^E(X,Y)$. Legenda: PT – Português; SP – Espanhol, IT – Italiano; FR – Francês; EN – Inglês; DE – Alemão; DA – Dinamarquês; SV – Sueco; NL – Holandês; FI – Finlandês.

Assinaturas genómicas: de composições de L-tuples a mapas CGR.

Foram já descritos alguns métodos relacionados com a composição de palavras em sequências. A ideia de usar frequências de palavras como características de sequências biológicas não é nova. Já nos anos 90 do século passado Samuel Karlin e os seus colegas (Karlin and Burge 1995) introduziram o conceito de assinatura genómica (*genomic signature*), no seguimento de trabalhos que estavam em curso sobre a sobre- e sub-representação de oligonucleótidos em DNA (Burge, Campbell and Karlin 1992). Os resultados sugeriam fortemente que os vectores de frequências de di-nucleótidos, obtidos para cada organismo, constituíam uma fortíssima característica da espécie. Este aspecto foi corroborado pela distinta composição encontrada intra-espécies, e pelo facto desta composição ser praticamente constante em cada organismo, mesmo em segmentos relativamente curtos do seu genoma. O cálculo envolve a determinação de uma razão das chances (*odds ratio*) $\rho_{s_i s_j}$ que representa o enviesamento de cada 2-tuple $s_i s_j$,, expresso em função da respectiva frequência $f_{s_i s_j}$ e dos valores esperados $f_{s_i} f_{s_j}$, $s_i, s_j \in \mathcal{A} = \{A,T,C,G\}$, sob a hipótese de a sequência ser bem modelada por uma cadeia de Markov de ordem 0:

$$\rho_{s_i s_j} = \frac{f_{s_i s_j}}{f_{s_i} \cdot f_{s_j}}$$

Pode ser definida uma distância entre duas sequências a partir desta medida, denominada distância-δ (Karlin and Ladunga 1994), baseada nas somas considerando todos os 4^2 di-nucleotidos existentes, usando ambas as cadeias de DNA (com notação * para a distinguir da anterior):

$$d_2^{\delta}(X,Y) = \frac{1}{16}\sum_{i=1}^{16}\left|\rho_i^{*X} - \rho_i^{*Y}\right|$$

Esta abundância relativa foi já extensivamente aplicada a sequências de DNA, verificando-se que é consistentemente mais elevada intra-espécies do que inter-espécies, o que demonstra as suas validade e congruência como possível classificador.

Outro resultado interessante é que esta assinatura se mantém, mesmo para janelas curtas do DNA. A persistência e alastramento desta característica (Jernigan and Baran 2002) demonstram mais uma vez a sua validade como característica intrínseca a cada organismo.

A ideia de assinatura genómica foi naturalmente estendida para palavras mais longas e a sua ligação com a representação através de jogos do caos CGR foi também explorada. De facto, o CGR agrupa sufixos iguais no mesmo sub-quadrante, o que faz aumentar a densidade local nesse mapa dos segmentos sobre-representados (e, inversamente, diminuir a dos sub-expressos). Nesse aspecto os mapas CGR constituem uma elegante e apelativa representação gráfica da abundância de todos os subsegmentos. Esta ideia foi explorada de forma exaustiva para caracterizar sequências desconhecidas.

A Figura 4 mostra a representação CGR de diversos genomas, calculada através do programa GENSTYLE disponibilizado na internet (Fertil, et al. 2005). Como foi referido, a representação de segmentos mais curtos de DNA conduz a imagens onde é patente o mesmo padrão, mas mais "desfocado".

A relação entre composição e assinaturas genómicas foi analisada posteriormente através da generalização deste conceito para L-nucleótidos. Num artigo recente (Wang, Hill, Singh and Kari 2005) os autores definiram o espectro de assinaturas genómicas como extensão do conceito para segmentos de ordem superior. Neste âmbito as definições de S.Karlin seriam um caso particular de perfis de abundância relativa de di-nucleótidos (*dinucleotide relative abundance profiles - DARP*), criando, desta forma, uma ponte bastante elegante entre as duas representações e contextualizando a sua validade como caracterizador de sequências e organismos.

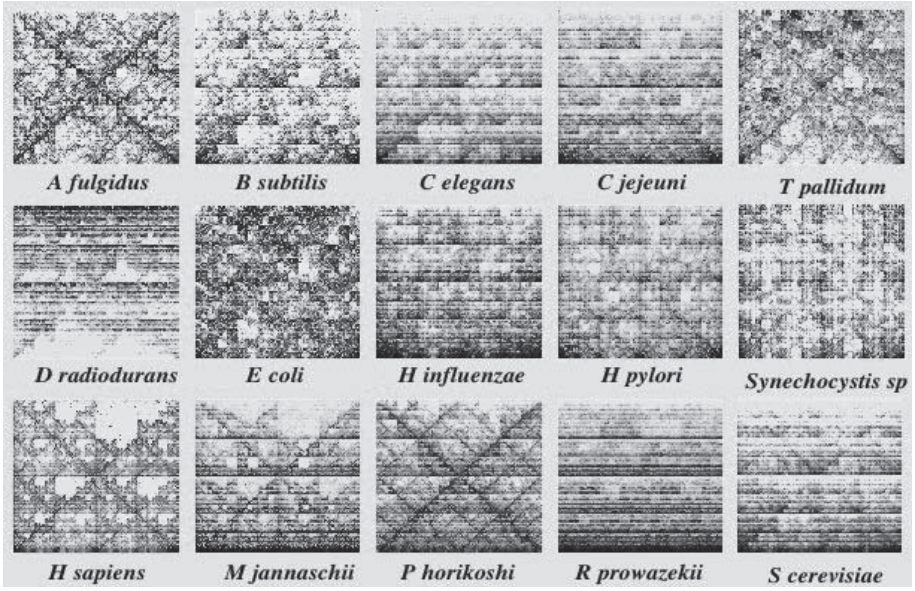


Figura 4. Mapas CGR para genomas completos de diversos organismos. O conceito de assinatura genómica é ilustrado, cada espécie apresenta um padrão de composição específico. In (Fertil, et al. 2005) © Usado com permissão dos autores.

Teoria de informação e entropia de DNA

Um outro tópico relevante na área de análise de sequências está ancorado na definição de entropia. O conceito de entropia (do grego *entropé* - em transformação) foi introduzido no estudo da termodinâmica dos gases, relacionado calor e temperatura em processos reversíveis. Mais tarde a mesma ideia foi desenvolvida no contexto de modelação de sistemas de comunicação, dando origem a uma nova área chamada Teoria da Informação (TI), introduzida por Claude Shannon num famoso artigo (Shannon 1948). A relação entre a TI e a probabilidade e estatística é clarificada com os trabalhos que conectam os conceitos de valor esperado, informação mútua, independência entre variáveis aleatórias e teoria da medida (Kullback 1968).

Intuitivamente o conceito de entropia foi inspirado por noções acerca do grau de imprevisibilidade de fenómenos aleatórios. Quando maior for a desordem de um dado sistema, maior será a sua entropia. Se a probabilidade de um estado for igual a um, então a entropia do sistema será zero, i.e., teremos um sistema determinístico.

Em termos formais a entropia de Shannon H^{Sh} associada a uma variável aleatória discreta *X* com espaço de resultados $\{x_1,\ldots,x_M\}$ e probabilidades $\{p_1,\ldots,p_M\}$ é definida como um funcional da distribuição e é medida em *bits*:

$$H^{Sh}(X) = H(p_1,\ldots,p_M) = -\sum_{i=1}^{M} p_i \log_2 p_i = E_X\left[-\log p(X)\right]$$

A entropia pode ser interpretada como o mínimo número de respostas binárias (Sim/Não) necessárias *em média* para determinar o resultado de uma observação de *X*. Por exemplo, no lançamento de um dado com seis faces teremos $H^{Sh} = \log_2 6 \approx 2{,}585$ bits.

A entropia de Shannon acima descrita pode ser obtida de forma axiomática (Ash 1990, Cover and Thomas 1991). De facto, se pretendermos uma função de um vector de probabilidades p_i com determinadas propriedades, seremos levados a definição pretendida, a menos de uma constante (que representa a base logarítmica considerada).

A entropia de Rényi de ordem $\alpha \geq 0, \alpha \neq 1$ $H_\alpha(X)$ (Rényi 1961, Rényi 1966) constitui uma generalização da anterior e é dada pelas seguintes fórmulas, apresentadas também para variáveis aleatórias contínuas e funções densidade de probabilidade *f*(*x*):

$$H_\alpha(X) = \frac{1}{1-\alpha}\ln\sum_{i=1}^{M} p_i^{\alpha}$$

$$H_\alpha(X) = \frac{1}{1-\alpha}\ln\int f(x)^{\alpha}\,dx$$

Pode-se demonstrar facilmente pela regra de l'Hôpital que $\lim_{\alpha\to 1} H_\alpha(X) = H^{Sh}(X)$.

Entropia global de Rényi de mapas CGR

Diversos trabalhos tentaram estimar a entropia de sequências biológicas, quer de DNA quer de proteínas, partindo destas definições. Tipicamente são utilizados como estimativas das probabilidades p_i dos L-tuples os respectivos valores f_i das frequências observadas, escolhendo uma resolução adequada, i.e., comprimento da palavra *L*. Estes valores para *L*=1 conduziram a resultados para o DNA de cerca de 1,9 bits, muito próximo do valor máximo de 2 bits correspondentes a uma sequência aleatória com equiprobabilidade $p_A=p_T=p_C=p_G=0{,}25$ e uma redução de apenas 1% para proteínas

(Weiss, Jimenez-Montano and Herzel 2000, Orlov, Filippov, Potapov and Kolchanov 2002). Este resultado pode parecer algo surpreendente visto a ausência aparente de estrutura de sequências que representam moléculas fundamentais ao armazenamento e transmissão de informação essencial à vida. No entanto esta aparente desordem ou aleatoriedade não é homogénea. Há zonas dos genomas extremamente repetitivas (e logo com baixa entropia) e outras praticamente indistinguíveis de sequências geradas aleatoriamente.

Um nível elevado de entropia é usualmente obtido para sequências comprimidas, sendo estas que contêm maior informação por símbolo. De facto, esta entropia reflecte provavelmente e apenas um maior nível de compactação de informação, posteriormente descodificada pela maquinaria celular e molecular. Desta forma a informação pode ser armazenada de forma mais parcimoniosa – embora seja necessário descodificá-la para poder ser utilizada.

Os mapas CGR acima descritos podem ser utilizados para determinar uma entropia generalizada desacoplada de uma determinada resolução específica, visto que o mapa pode representar todas as composições simultaneamente. Assim, num artigo recente (Vinga, et al. 2004) foi proposta uma metodologia para estimar a entropia de Rényi de uma sequência de DNA com suporte directo nesta representação. Com este objectivo seria necessário calcular a entropia de uma função densidade de probabilidade estimada a partir destes mapas, para o que foi usado um método de estimação não-paramétrico. Neste âmbito foi escolhido o método das janelas de Parzen, com aplicação de funções núcleo (*kernel*) gaussianas.

O método das janelas de Parzen consiste em estimar uma função densidade de probabilidade $f(x)$ a partir de uma amostra de N variáveis $a=(a_1,\ldots,a_N)$.independentes e identicamente distribuídas. O estimador de densidade pelo método do núcleo $\hat{f}(x)$ consiste numa combinação linear das funções do núcleo (*kernel*) $\kappa_\theta(x)$, centradas em cada um dos pontos da amostra e determinadas para uma dada largura de janela τ:

$$\hat{f}(x;a,\theta)=\hat{f}_\theta(x;a)=\frac{1}{N\tau}\sum_{i=1}^{N}\kappa_\theta\left(\frac{x-a_i}{\tau}\right)$$

A aplicação da entropia de Rényi de ordem 2 conjugada com núcleos gaussianos $\kappa(x)=g_p(x;0,\sigma^2 I_p)=\frac{1}{(2\pi)^{p/2}\sigma^p}\exp\left(-\frac{1}{2\sigma^2}x^T x\right)$ conduz a uma simplificação notável do cálculo da entropia global, cujos detalhes são descritos em (Vinga, et al. 2004). De facto, como a convolução de duas gaussianas é também ela uma gaussiana a entropia reduz-se a:

$$\begin{aligned}H_2(CGR)&=-\ln\int\hat{f}^2(x)dx\\&=-\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{16\pi^2\sigma^4}\exp\left(-\frac{1}{4\sigma^2}d_{ij}\right)\end{aligned}$$

onde d_{ij} representa o quadrado da distância euclidiana entre as coordenadas CGR a_i e a_j.

A entropia global de Rényi assim definida depende apenas das distâncias entre todos os pares de pontos do mapa, ou seja, das interacções entre coordenadas CGR, o que levou alguns autores a definir o conceito de potencial de informação (*information potential*), como analogia aos campos magnético e gravítico (Principe, Xu and Fisher 2000).

Esta simplificação pode ser aplicada directamente ao cálculo das entropias do mapa CGR onde sobressaem algumas propriedades relativas à sua dependência do parâmetro σ^2, a variância da distribuição normal utilizada, como exemplificado na Figura 5.

É também possível determinar analiticamente algumas propriedades desta função, nomeadamente no que concerne à existência de assímptotas ao gráfico (Fig.5A), iguais a:

$$H_2^+ = 2\ln\sigma^2 + \ln 16\pi^2$$
$$H_2^- = 2\ln\sigma^2 + \ln 16\pi^2 + \ln N$$

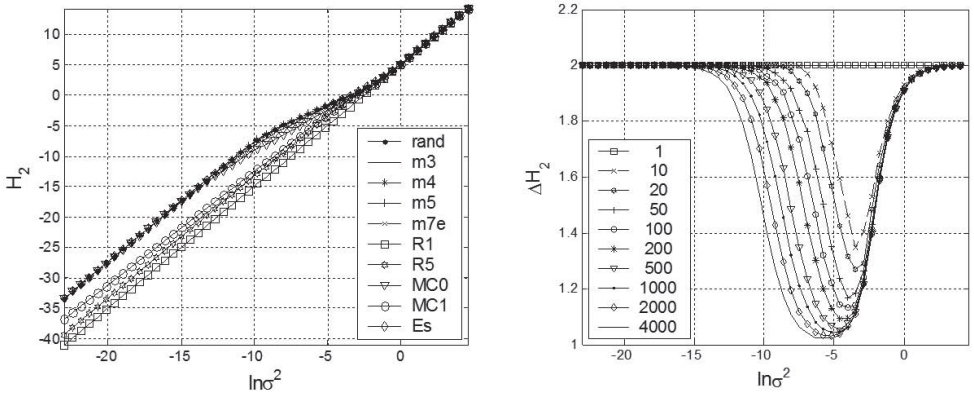


Figura 5. A) Entropia de Rényi de diversas sequências de DNA em função da variância da distribuição normal utilizada no método de Parzen. Legenda das sequências: rand –aleatória, m3, m4, m5 – artificiais com motivos inseridos de tamanhos respectivamente 3, 4, 5 e 7 com erro, R1, R5 – mesmo símbolo repetido, M0, M1 – geradas a partir de cadeias de Markov de ordem 0 e 1; Es – sequencias promotoras de *B.subtilis*. B) Entropia diferencial de Rényi de sequências aleatórias de diversos comprimentos N=1,…,4000, calculada a partir da derivada dos valores medianos de entropia obtidos por simulação.

Perfis entrópicos locais

Uma outra noção que derivou recentemente desta abordagem foi a extracção de informação local para além da estimação da entropia (global) descrita. De facto, a estimativa pontual do valor de densidade de probabilidade poderá ser usado como um valor de referência para uma posição específica.

A estimação obtida com o método de Parzen servirá não só para a estimação global, com a simplificação acima descrita, mas também cada valor $\hat{f}(x_i)$, determinado para um dado conjunto de parâmetros, revelará informação acerca do sufixo respectivo que termina no símbolo s_i. Dadas as propriedades do CGR apresentadas um valor elevado de $\hat{f}(x_i)$ significará uma grande densidade de pontos nesse quadrante, i.e., uma sobre-representação daquele sufixo específico. Inversamente, valores reduzidos correspondem a sufixos que aparecem poucas vezes na sequência original. Esta ponte entre densidades em mapas CGR e significância estatística de motivos foi explorada num trabalho recente (Vinga and Almeida 2007). Foi proposta a noção de perfil entrópico (*entropic profile*), que não é nada mais do que a estimação de densidade relativa a cada símbolo, convenientemente mapeado no CGR, em função da posição.

Houve também um estudo cuidado sobre a influência do núcleo utilizado. Funções de distribuição normais, embora vantajosas para um exame global, poderão não ser as funções mais adequadas para uma estimação mais fina, visto que o seu domínio não está definido no mesmo conjunto definido para o CGR. Deste modo, foi proposta uma outra função núcleo constituída por L blocos que preenchem

exactamente os quadrantes e sub-quadrantes (Almeida and Vinga 2006). Com este procedimento pretendeu-se eliminar potenciais fontes de "ruído" local e efeitos de fronteira entre os sub-espaços.
Esta nova função está indexada por dois parâmetros: a resolução *L*, que define o número de blocos a considerar e um parâmetro de alisamento, ϕ, que mede a razão entre volumes de dois blocos sucessivos. Obtém se uma função com propriedades fractais, exemplificada na Figura seguinte:

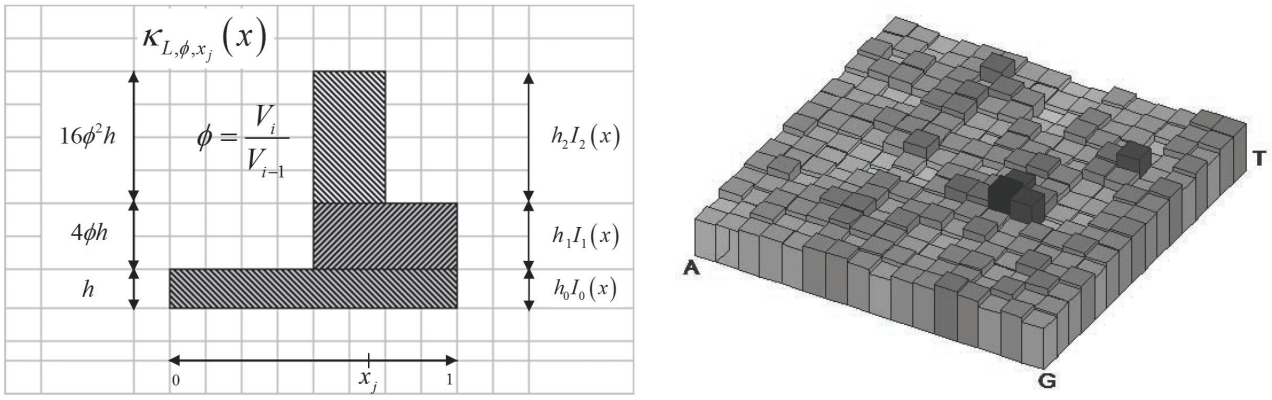


Figura 6. A) Exemplo do núcleo fractal proposto $\kappa(x)$. Construção da função, projectada para uma dimensão, e com parâmetros L=2 e ϕ arbitrário. B) Estimação de densidades com o método de Parzen. Densidades dos mapas CGR para uma sequência de DNA artificial com um motivo 'ATCG' inserido em posições controladas. Estimação obtida com o método de Parzen utilizando o kernel descrito com L=4 e ϕ=1. O sub-quadrante relativo ao motivo sobre representado sobressai nitidamente (cor mais escura) devido a uma elevada densidade de pontos nesta região.

Uma consequência em utilizar esta função núcleo com o método de Parzen é a relativa simplificação do valor estimado para cada ponto $\hat{f}(x_i)$, correspondente ao *i*-ésimo símbolo s_i da sequência.

Após algum trabalho de simplificação algébrica ficamos com o perfil entrópico para cada ponto determinado por:

$$\hat{f}_{L,\phi}(x_i) = \frac{1 + \frac{1}{N}\sum_{k=1}^{L} 4^k \phi^k \cdot c([i-k+1, i])}{\sum_{k=0}^{L} \phi^k}$$

onde $c([i-k+1, i])$ representa o número de motivos $(s_{i-k+1} \cdots s_i)$ contabilizados para toda a sequência de tamanho *N*.

Após normalização para todas as posições, podem ser obtidos gráficos com os perfis entrópicos que, para cada posição *i*, expressam a abundância relativa do motivo correspondente, em número de desvios-padrão. Esta normalização permite comparar escalas e conjuntos de parâmetros muito diferentes.

A detecção de motivos relevantes e estatisticamente significativos pode ser posteriormente efectuada de forma não supervisionada pela procura, no espaço paramétrico, de máximos e mínimos locais. De facto, se numa dada posição *i* estiver presente um motivo de comprimento igual a quatro, o perfil entrópico apresentará um máximo local para L=4, como apresentado na Figura seguinte:

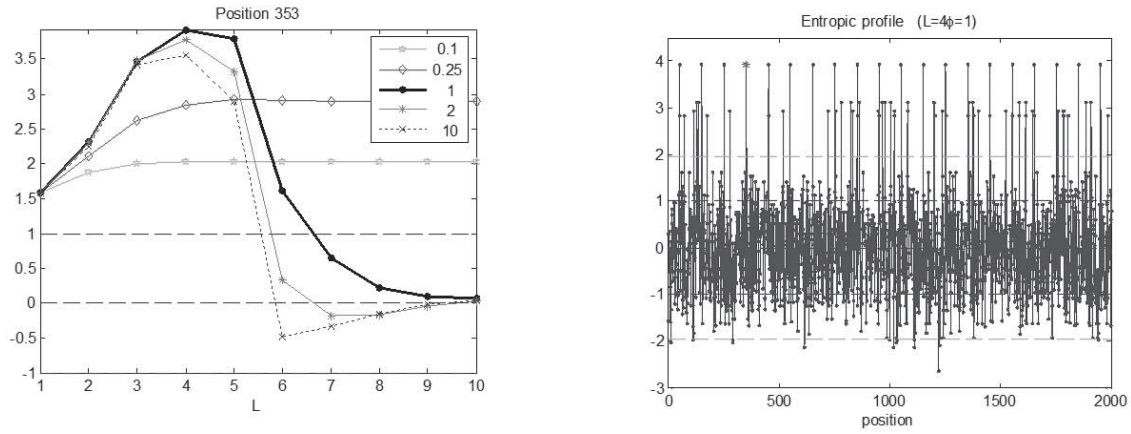


Figura 7. Perfil entrópico para uma sequência gerada aleatoriamente onde foi inserido um motivo de comprimento quatro em posições igualmente espaçadas. A) O valor de $\hat{f}(x_i) \approx 3{,}8$ para a posição onde o motivo termina tem um máximo local para L=4 e ϕ=1. B) Perfil entrópico de toda a sequência de DNA obtido para a combinação de parâmetros anteriores. Observa-se máximos locais que correspondem às posições de ocorrência do motivo, inseridos em posições equidistantes.

A aplicação desta metodologia a genomas reais reforçou a abordagem para detecção de sinais a diversas escalas (Vinga 2007). O algoritmo foi aplicado às sequências completas de DNA de duas bactérias, *Escherichia coli* K12 e *Haemophilus influenzae* Rd, com o objectivo de detectar eventuais locais onde alguma escala em particular pudesse sobressair.

Para o genoma da *E.coli* foram detectados muito facilmente os chamados **Chi sites** (*crossover hotspot instigator*), correspondentes ao motivo GCTGGTGG. Este segmento está fortemente sobre-representado no genoma e sabe-se que é biologicamente muito importante pois permite a reparação de DNA defeituoso através de enzimas específicas. De facto, este segmento ocorre 761 vezes neste genoma constituído por N=4639675 bases (símbolos); considerando um modelo de Markov de ordem zero, o valor esperado do número de ocorrências seria apenas $N/4^8$=70,8, muito abaixo do que é observado na realidade.

A Figura seguinte resume o perfil para uma janela que inclui esses motivos.

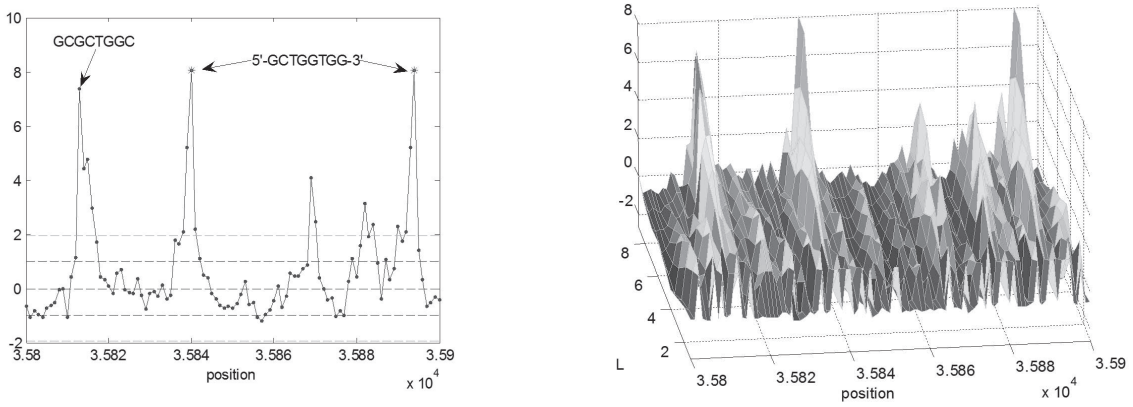


Figura 8. Perfis entrópicos para a sequência da *E.coli*. A) É apresentado uma pequena janela do genoma onde sobressaem motivos importantes detectados com a abordagem descrita. B) Perfil entrópico em função da posição e do parâmetro *L*.

Para o genoma do *H. Influenzae* foi identificado correctamente o motivo correspondente às **Uptake Signal Sequences** (USS+) , AAGTGCCGGT. Estes segmentos são também muito relevantes do ponto de vista biológico, e sabe-se que estão envolvidos num fenómeno chamado competência natural (*natural competence*), que é uma forma controlada de transferência horizontal de genes relacionada

com a capacidade de algumas bactérias extraírem DNA do ambiente circundante (Dubnau 1999). Este processo permite trocas genéticas entre estes organismos e recombinação de DNA que eventualmente esteja livre. O processo de transferência e identificação de DNA externo passa pela ligação preferencial a segmentos que contenham este motivo USS específico. Desta forma são identificados pedaços de DNA e promovida uma certa promiscuidade entre parentes próximos. Outro aspecto interessante do ponto de vista estatístico é a homogeneidade da distribuição dos USS no genoma, aparecendo em posições igualmente espaçadas (Karlin, Mrazek and Campbell 1996).

Conclusões

Este artigo pretendeu ilustrar alguns tópicos relacionados com uma área da bioinformática que apresenta fortes ligações com a estatística. Espera-se que esta breve introdução tenha despertado a curiosidade sobre um tema que se tem vindo a revelar um grande impulsionador da biologia, que muitos consideram já a ciência do séc. XXI.

Referências

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001), "Analysis of Genomic Sequences by Chaos Game Representation," *Bioinformatics*, 17, 429-437.

Almeida, J. S., and Vinga, S. (2002), "Universal Sequence Map (USM) of Arbitrary Discrete Sequences," *BMC Bioinformatics*, 3, 6.

Almeida, J. S., and Vinga, S. (2006), "Computing Distribution of Scale Independent Motifs in Biological Sequences," *Algorithms Mol Biol*, 1, 18.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990), "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, 215, 403-410.

Ash, R. B. (1990), *Information Theory*, New York: Dover Publications.

Burge, C., Campbell, A. M., and Karlin, S. (1992), "Over- and under-Representation of Short Oligonucleotides in DNA Sequences," *Proc Natl Acad Sci U S A*, 89, 1358-1362.

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory*, ed. W. Interscience, New York: Wiley.

Dubnau, D. (1999), "DNA Uptake in Bacteria," *Annu Rev Microbiol*, 53, 217-244.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998), *Biological Sequence Analysis*, Cambridge University Press.

Edgar, G. A. (1990), *Measure, Topology, and Fractal Geometry*, New York: Springer-Verlag.

Ewens, W. J., and Grant, G. R. (2001), *Statistical Methods in Bioinformatics : An Introduction*, New York: Springer.

Fertil, B., et al. (2005), "Genstyle: Exploration and Analysis of DNA Sequences with Genomic Signature," *Nucleic Acids Res*, 33, W512-515.

Jeffrey, H. J. (1990), "Chaos Game Representation of Gene Structure," *Nucleic Acids Res*, 18, 2163-2170.

Jernigan, R. W., and Baran, R. H. (2002), "Pervasive Properties of the Genomic Signature," *BMC Genomics*, 3, 23.

Karlin, S., and Burge, C. (1995), "Dinucleotide Relative Abundance Extremes: A Genomic Signature," *Trends Genet*, 11, 283-290.

Karlin, S., and Ladunga, I. (1994), "Comparisons of Eukaryotic Genomic Sequences," *Proc Natl Acad Sci U S A*, 91, 12832-12836.

Karlin, S., Mrazek, J., and Campbell, A. M. (1996), "Frequent Oligonucleotides and Peptides of the Haemophilus Influenzae Genome," *Nucleic Acids Res*, 24, 4263-4272.

Kullback, S. (1968), *Information Theory and Statistics*, New York: Dover Publications.

Needleman, S. B., and Wunsch, C. D. (1970), "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *Journal Molecular Biology*, 48, 443-453.

Orlov, Y. L., Filippov, V. P., Potapov, V. N., and Kolchanov, N. A. (2002), ""Complexity": Software

Tools for Analysis of Information Measures of Genetic Texts," in *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, North Carolina, USA.

Principe, J. C., Xu, D., and Fisher, J. W. I. (2000), "Information-Theoretic Learning," in *Unsupervised Adaptive Filtering* (Vol. 1), ed. S. Haykin, New York: John Wiley & Sons, pp. 265-319.

Rényi, A. (1961), "On Measures of Entropy and Information," in *Proc. of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, University of California Press, pp. 547-561.

Rényi, A. (1966), "Introduction a La Théorie De L'information," in *Calcul Des Probabilités*, Paris: Dunod.

Robin, S., Rodolphe, F., and Schbath, S. (2005), *DNA, Words, and Models*, New York, NY: Cambridge University Press.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *The Bell System Technical Journal*, 27, 379-423, 623-656.

Smith, T. F., and Waterman, M. S. (1981), "Identification of Common Molecular Subsequences," *Journal Molecular Biology*, 147, 195-197.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994), "Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice," *Nucleic Acids Research*, 22, 4673-4680.

Vinga, S. (2005), "*Biological Sequence Analysis by Vector Maps: Alignment-Free Comparison of DNA and Proteins*," PhD Thesis, Instituto de Tecnologia Química e Biológica - Universidade Nova de Lisboa (ITQB/UNL).

Vinga, S. (2007), "Biological Sequence Analysis by Vector-Valued Functions: Revisiting Alignment-Free Methodologies for DNA and Protein Classification," in *Advanced Computational Methods for Biocomputing and Bioimaging*, eds. T. D. Pham, H. Yan and D. I. Crane, New York: Nova Science Publishers.

Vinga, S. (2007), "Whole Genome Analysis through Rényi Entropic Profiles," in *ISMB/ECCB 2007*, Vienna, Austria.

Vinga, S., and Almeida, J. (2003), "Alignment-Free Sequence Comparison--a Review," *Bioinformatics*, 19, 513-523.

Vinga, S., and Almeida, J. S. (2004), "Rényi Continuous Entropy of DNA Sequences," *J Theor Biol*, 231, 377-388.

Vinga, S., and Almeida, J. S. (2007), "Local Rényi Entropic Profiles of DNA Sequences," *submetido*.

Wang, Y., Hill, K., Singh, S., and Kari, L. (2005), "The Spectrum of Genomic Signatures: From Dinucleotides to Chaos Game Representation," *Gene*, 346, 173-185.

Waterman, M. S. (1995), *Introduction to Computational Biology : Maps, Sequences, and Genomes : Interdisciplinary Statistics*, Boca Raton, Fla.: Chapman & Hall/CRC.

Weiss, O., Jimenez-Montano, M. A., and Herzel, H. (2000), "Information Content of Protein Sequences," *J Theor Biol*, 206, 379-386.

Bioestatística

Dinis Pestana, *dinis.pestana@fc.ul.pt*
Universidade de Lisboa
DEIO (FCUL) e CEAUL

É evidente que não posso melhorar os verbetes *Biostatistics, History of* (J. Rosser Matthews) e *Biostatistics, Overview* (W. Byron Brown, Jr.), na enciclopédia de Armitage and Colton (2005), o que me dispensa de fazer uma panorâmica da área, remetendo simplesmente para a leitura daqueles artigos magistrais.
Porém, como em minha opinião a Bioestatística é uma área disciplinar mais vasta do que a aplicação da Estatística a problemas médicos e da biologia humana, não resisto a escrever algumas linhas sobre Bioestatística e o que a distingue (?) da Estatística, porque tal tem que ver com o ensino da Bioestatística, a que estou ligado há muitos anos. A diversidade de opções no ensino da Bioestatística é grande, e saudável, indico algumas das minhas, mas sem qualquer proselitismo.
A grande profusão de artigos de Bioestatística nos congressos da SPE atesta que a área está em pleno desenvolvimento, e que não seria difícil publicar um grosso volume com contribuições de muitos cientistas de mérito. Mas havia apenas algumas escassas dezenas de páginas disponíveis, e se incluir é fácil, excluir é embaraçoso, às vezes mesmo penoso. Daí a solução fácil - e que potencialmente poderá, espero, aproximar comunidades afins que pouco se encontram: solicitei textos a cientistas jovens, cuja linha principal de investigação não é a Estatística, que não frequentam os Congressos da SPE, e de quem nunca fui professor. Enfim, escapei-me à tangente de ter que fazer escolhas difíceis, e a qualidade dos textos, em última análise, é a melhor defesa para a estratégia que usei ...

Bioestatística - existe, ou é apenas um nome?

Muitos dos livros mais conceituados de Bioestatística têm logo à cabeça afirmações como "Statistics applied to biological problems is simply called *biostatistics*" (Zar, 1996, p. 1). Parece, portanto, que na mente de muitos desses autores Bioestatística é um termo sinónimo de Estatística, que apenas explicita a área de aplicação, não havendo na essência e fundamentos qualquer distinção entre Estatística e Bioestatística.
Alguns autores são mais restritivos do que Zar, afirmando que a Bioestatística é a Estatística aplicada à Biologia humana e Medicina. No verbete *Biostatistics, Overview*, preparado por W. Byron Brown, Jr. para a Armitage and Colton (2005), é expresso esse ponto de vista: "Biostatistics focuses on the development and use of statistical methods to solve problems and questions that arise in human biology and medicine."
Lachin (2000) é uma excepção à ideia de que Bioestatística não é mais do que Estatística orientada para problemas da Biologia (humana), afirmando: "Thus I was faced with the question 'what are the foundations of biostatistics?' In my opinion, biostatistics is set apart from other statistics specialities by its focus in the assessment of risks and relative risks through clinical research."
Porventura por ter a meu cargo uma disciplina de Bioestatística inserida nas licenciaturas em Biologia, não confino a Bioestatística às fronteiras apertadas da Biologia humana. Cabem a meu ver na Bioestatística temas tão diversos quanto a avaliação de recursos faunísticos e florais (e provavelmente os desenvolvimentos notáveis da Amostragem por Distâncias e da Amostragem Adaptativa devem-se à necessidade de avaliar recursos naturais), estudos de teoria da aprendizagem e comportamento animal, que se fazem por exemplo com processos de Markov ou com lógica difusa, questões de ecologia (na minha instituição, o Prof. Henrique Cabral lecciona uma cadeira de Introdução à Ecologia Numérica, cujo programa e bibliografia, disponíveis em *http://correio.fc.ul.pt/~hcabral/index_files/EN1.pdf*, não deixam dúvida sobre a sua inserção na área da Bioestatística), e, sobretudo, Planeamento de Experiências. Cobb (1998), no seu saboroso livro sobre esta área, faz uma afirmação curiosa: cada vez se considera menos matemático e mais biólogo.
Mas voltando à linha principal: há distinção entre Estatística e Bioestatística? Ao nível do ensino

elementar, essa distinção é porventura apenas de intencionalidade de orientação temática, o que pode ser evidenciado pelos problemas propostos. No âmbito mais geral, em minha opinião a Estatística não deve ignorar as consequências éticas do seu exercício, enquanto a Bioestatística não *pode* ignorar essas consequências, nomeadamente na avaliação da experimentação que enforma. Toda a Estatística que conheço tem aplicações bioestatísticas; se tivesse que singularizar um campo que considero mais bioestatístico do que estatístico, creio que escolheria a Meta-análise: é sem dúvida em questões em que a ética de obter informação é mais controversa que as sínteses meta-analíticas, aproveitando a informação que há espalhada por diversos centros de investigação, instituições hospitalares, comunicações forenses, etc., são de mais valor.
Claro que questões desta complexidade acrescida só podem ser abordadas quando se domina razoavelmente a Bioestatísca que só no nome, intenção e objecto de estudo se distingue da Estatística.

Ensino da Bioestatística

Não sei quando começou o ensino da Bioestatística em Portugal, mas no início dos anos 60 do século passado começaram a surgir manuais para o ensino da Estatística a médicos e biólogos. Creio que Carlos Santos Reis e Alexandre Sarmento (1960) foram os pioneiros, publicando o primeiro manual português de Bioestatística. É um texto ambicioso (513 páginas), que tem a natural preocupação de, a par das questões mais matematizadas da estatística, debater com algum detalhe amostragem e planeamento experimental, discutindo a aquisição de dados e a sua análise.
No Boletim da SPE de 2002 relembrei a breve incursão do Prof. Dias Agudo no ensino da Bioestatística na FCUL quarenta anos antes, e a jóia que é a "sebenta" publicada por dois dos alunos dessa cadeira notável.
A FCUL teve o privilégio de ter no seu corpo docente o Prof. Campos Rosado, que está na origem de um grupo de investigadores cujos interesses começaram por se sediar na dinâmica e genética de populações, e evoluindo depois para incluir também o tratamento de questões epidemiológicas e bioinformáticas. Creio que se deve sobretudo aos Prof. Campos Rosado e Tiago de Oliveira a criação de uma cadeira de Bioestatística, primeiro com carácter opcional, posteriormente uma das cadeiras estruturantes das licenciaturas em Biologia.
Quando comecei a reger a cadeira introdutória de Bioestatística, esta era ainda opcional, e se então me perguntassem se a Bioestatística era um dos meus interesses, decerto diria que era apenas um serviço esporádico. A qualidade dos alunos, a diversidade dos temas, as questões que ex-alunos me colocavam quando iniciavam uma carreira de investigação, colaboração com médicos, enfermeiros, farmacêuticos, dentistas, depressa me levaram a preferir a Bioestatística a qualquer outra cadeira que ensino.
Creio que na fase inicial nem me interroguei sobre o que devia ensinar, inconscientemente partilhei a opinião daqueles para quem a Bioestatística é tão-só estatística, com o viés especial de ser tematicamente orientada para questões de Biologia. A minha opinião sobre o que é a Bioestatística mudou bastante, a forma de ensinar também, sobretudo porque a experiência e muita leitura foi enriquecendo a escolha de exemplos, mas o essencial do que ensino continua a ser Estatística.
De facto, numa cadeira inicial de Bioestatística o que se deve ensinar - se os nossos colegas que têm a responsabilidade das licenciaturas em que a cadeira se insere tiverem a percepção do tempo que ela necessita para ensinar novas formas de pensar, e nomeadamente de integrar raciocínio dedutivo e indutivo, uma característica ímpar que por si só merece o investimento - é o essencial sobre Probabilidade e Estatística:
• que a incerteza é uma fonte de conhecimento, quando a Probabilidade nos permite delimitá-la caracterizando os seus padrões, e deve ser usada em todas as ciências cujo objecto de estudo é complexo;
• que a quantificação é incontornável (algumas indicações sobre Metrologia, se houver tempo, são úteis), e consequentemente temos que saber conviver com os inevitáveis erros de medição, pondo-os ao nosso serviço;
• que mais importante do que a informação é a *transformação de* em conhecimento;
• que a informação obtida "por acaso" pode ser enganadora, enquanto a informação "ao acaso" tem uma variabilidade que é útil - por exemplo para determinar quantas observações "ao acaso" são necessárias para limitar o erro da estimativa;

• que se queremos que a informação quantitativa nos seja fornecida de forma inteligível, também devemos saber tratar a informação de forma a comunicá-la facilitando a leitura alheia;
• que a estatística descritiva e exploratória é apenas o primeiro passo (comparável às descrições dos naturalistas), mas que há que dar o salto para aspectos confirmatórios e inferenciais, sendo a alavanca que permite dar esse salto a Probabilidade — e aí deve insistir-se na importância dos modelos;
• que Linus Pauling, quando respondeu que para se ter o Nobel era preciso ter muitas ideias e a coragem de abandonar quase todas estava a expressar duas grandes verdades: a teoria dos testes de hipóteses foi desenvolvida de forma a podermos rejeitar ideias que são contrariadas pela evidência factual (contra factos não há argumentos); a Lei dos Grandes Números leva-nos a ter esperança de que algumas boas ideias escapem a essa operação de limpeza, se forem muitas;
• que o teorema da probabilidade total é o método de Descartes transposto para a modelação matemática (aproveitar para discutir "pela rama" amostragem estratificada); mas que a análise da variância é uma criação que contorna situações de comparações múltiplas sob risco, em que o método de Descartes inevitavelmente falha;
• que a Amostragem é boa, mas o Planeamento Experimental ainda é melhor (como em história religiosa, versão *grafitti* filosóficos das universidades inglesas: "*Jesus saves, but Moses invests*"), por ser um investimento na obtenção dos dados que importa analisar. A este propósito, se possível, alguma discussão sobre o papel da causalidade na ciência moderna, e referência ao notável e um pouco esquecido *The Grammar of Science*, do gigante Karl Pearson, em que se defende o abandono da causalidade em prol de formas mais fluídas de associação estatística;
• que no modelo gaussiano convivemos bem com pequenas amostras, e que o teorema limite central justifica a aproximação por esse modelo em condições muito gerais; mas que o desenvolvimento de meios de cálculo acessíveis permite actualmente a replicação diária do milagre do "pão partido em pequeninos", e as técnicas de simulação são uma forma de transformar poucas observações em muitos fingimentos de observações.
• que existem desenvolvimentos meta-analíticos que nos permitem usar dados obtidos com metodologias e "metrologias" diversas, por várias equipas, para deles extrair evidência que não é visível em cada subconjunto de dados, mas pode ser irrecusável na síntese possível;
• que se o problema do passado (e actual, no caso de doenças raras) era a escassez dos dados, actualmente o problema é frequentemente a grande profusão de dados, sendo um bom convívio com técnicas de pesquisa aleatória e *data mining* uma mais valia interessante;
• que os problemas éticos não podem ser escamoteados na investigação experimental - e esta afirmação tem dois gumes, porque não fazer investigação também tem consequências éticas!
E mais não digo, para não continuar a maçar o leitor. Não acredito num ensino que não discuta as ideias (o meu querido amigo Simões Neto teve uma vez um comentário malicioso, depois de um seminário: "Muitas fórmulas mas poucas ideias"), mas também considero que as alterações do ensino secundário justificam amplamente que na universidade se insista na importância do rigor e do pensamento dedutivo. Por isso esforço-me por seguir os excelentes conselhos de Frank Boas, que podem ser lidos no prefácio do excepcional livro de Mosteller and Rourke (1973): usar as demonstrações como o sal em cozinha, q.b., e limitá-las a situações especiais que as tornam simples e esclarecedoras de como se deve deduzir com rigor.
Tenho resistido, e creio que continuarei a resistir, a um ensino centrado na utilização de *software*. É uma habilitação que se pode ajudar os alunos a adquirir, trocar a discussão de ideias pela decifração de *outputs* (eventualmente de lixo) é em minha opinião ser tão insensato como o cão da fábula que trocou a presa pelo seu reflexo na água. O meu trabalho como consultor de médicos só tem robustecido esta opinião.
Um dos muitos livros de Bioestatística que li (creio que do Selvin) diz no prefácio que o conteúdo corresponde ao que ensina na quarta cadeira semestral de Bioestatística do curso de Berkeley. É impossível não ter inveja desta situação. Na minha instituição apenas há uma cadeira de Bioestatística da responsabilidade do Dept. Estatística, e creio que na maior parte das universidades em Portugal, nas antigas licenciaturas em 4-5 anos, a situação é análoga, tendendo a piorar porque o tempo efectivo de leccionação vai porventura ser reduzido com a nova receita de *moscas à bolonhesa*. Na FCUL há, por outro lado, muitas cadeiras cujo conteúdo é essencialmente constituído por tópicos avançados de Bioestatística, leccionadas por biólogos, uma situação privilegiada para os nossos estudantes, que lhes dá instrumentos de trabalho valiosos.

O ensino da Bioestatística a nível pós-graduado é omisso em muitos cursos, quando é a fase em que os alunos, já com alguns conhecimentos de base, experiência e ambição, mais sedentos estariam de aprender as bases de planeamento experimental e amostragem. O governo do Reino Unido publicou em 1994, se não me falha a memória, um livro branco (*Realizing our Potential*) em que manifestava a sua preocupação com a formação de cientistas que pudessem de facto ser líderes de investigação, recomendando às universidades que reservassem uma fatia dos cursos pós-graduados para a discussão aprofundada das metodologias e instrumentos da investigação científica. O memorável livro editado por Greenfield (2002) é eloquente sobre o papel da Estatística, e recomendo-o a estudantes de todas as áreas.

Continuarei a lutar pela existência de uma cadeira avançada (ou mesmo mais do que uma) de (Bio)Estatística em cursos pós-graduados, da mesma forma que defendo uma formação académica sobre como aceder a fontes documentais, usar adequadamente meios de cálculo, escrever projectos de investigação e relatórios científicos.

A escolha de autores e textos

Fazer um número temático do Boletim da SPE sobre Bioestatística, uma ideia cujo mérito cabe inteiramente ao editor, Professor Fernando Rosado, é naturalmente oportuna.

O convite que me fez para preparar esse número, quando outros - Antónia Turkman, Carlos Braumann, Daniel Paulino, Lucília Carvalho, Manuel Carmo Gomes, e muitos outros que inevitavelmente deixo de fora ou cito noutros pontos deste texto — teriam decerto mais preparação e mérito para o fazer, pela investigação que têm feito e dirigido, deve-se decerto ao grande relevo que a SPE dá ao ensino da Estatística.

Não há escolha que não comporte exclusão, e naturalmente resisti à tentação fácil de escolher os que estão mais perto de mim, colegas e alunos pós-graduados da FCUL e do CEAUL. Ao olhar para mais longe, pareceu-me que a opção mais interessante, apesar de muito redutora, era solicitar textos a investigadores em Biologia menos conhecidos da comunidade estatística, por não serem membros da SPE - esperando assim contribuir para o encontro de áreas diversas mas com uma grande partilha de interesses.

Nenhum destes dois autores que escolhi (dois casos exemplares para os estatísticos mais jovens, pela determinação e persistência na construção de uma carreira científica credível, devidamente alicerçada na publicação em revistas exigentes), a Susana Vinga e o Gil Penha-Lopes, foi meu aluno. A Susana Vinga graduou-se em Engenharia Mecânica, o Gil Penha-Lopes em Biologia (passando pela Bioestatística da Profª Luísa Canto e Castro Loura), e foi a investigação que os levou a aprofundar os conhecimentos em Estatística.

Um dos múltiplos interesses do Gil Penha-Lopes (http://biology-research.quodis.com/) é a aquacultura, uma área em que os interesses económicos prevalecem, e que por isso obriga a investigação de grande complexidade e com restrições, impostas por considerações económicas, na experimentação. O Gil Penha-Lopes, nos duelos que travou com os *referees* das revistas, depressa aprendeu que a melhor forma de poupar esforços é trabalhar arduamente na fase preparatória do trabalho experimental, planeando cuidadosamente o que deve fazer para ter bases objectivas com que dar resposta às perguntas a que quer responder. O texto que ele preparou evidencia essa faceta que todos os utilizadores da Estatística deveriam cuidar: que dados interessa recolher, que precauções se deve ter para conseguir a informação adequada, que reflexão prévia tem que se fazer sobre as perguntas a que se quer dar solução, e em que sentido essa reflexão condiciona a experimentação que se faz. Mas para quê estes comentários, que não substituem a leitura do artigo?

É um texto rico e equilibrado, que tem o mérito de mostrar que o domínio das metodologias da Estatística (e é naturalmente propositadamene que não uso o modificador Bio-) é essencial para a realização de boa investigação experimental.

Conheci a Susana Vinga quando me auxiliou num curso intensivo para alunos de doutoramento do IGC (com o Nuno Sepúlveda, que fez um trabalho excepcional, mas foi preterido por ter uma formação de base em Matemática e ser um frequentador habitual dos congressos da SPE - e ainda há quem, como Pascal, acredite na recompensa de uma vida virtuosa!). Deu-me acesso ao CV (http://algos.inesc-

id.pt/~svinga/), antes mesmo de a conhecer, e fiquei fascinado com a forma como tinha usado ideias de teoria de Informação na sua investigação. O texto que nos ofereceu, em que aborda alguns dos temas da sua preferência, mostram que os caminhos da Bioestatística, em termos de investigação, têm uma sofisticação matemática que não é inferior à da mais matematicamente hermética Inferência Estatística. Se consegui despertar a curiosidade do leitor, pode avançar para as Actas dos congressos anuais da SPE, onde vai encontrar, nos quinze anos que cobrem, centenas de artigos de Bioestatística, daqueles que fiz a injustiça, inevitável, de excluir - e a quem peço alguma compreensão pelas minhas opções.

Bibliogafia

Armitage, P., and Colton, T. (2005). *Encyclopedia of Biostatistics*, 2nd ed., Wiley, Chichester.

Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments*, Springer, New York.

Greenfield, T. (2002). *Research Methods. Guidance for Postgraduates*, 2nd ed., Arnold, London.

Lachin, J. M. (2000). *Biostatistical Methods — The Assessment of Relative Risks*, Wiley, New York.

Mosteller, F., and Rourke, R. E. K. (1973). *Sturdy Statistics. Nonparametric and Order Statistics*, Addison-Wesley, Reading, Mass. (Tradução portuguesa: *Estatísticas Firmes - Estatísticas Não Paramétricas e Estatísticas Ordinais*, Salamandra, Lisboa, 1993.)

Santos Reis, C., e Sarmento, A. (1960). *Manual de Estatística Médica*, Instituto de Medicina Tropical, Lisboa.

Zar, J. H. (1996). *Biostatistical Analysis*, 3rd ed., Prentice Hall, Upper Sadle River.

Ciência Estatística

• Artigos Científicos Publicados

Antunes, N., C. Fricker, C., Robert P., and Tibi, D. (2006) - Analysis of loss networks with routing. *Annals of Applied Probability* 16(4): 2007-2026.

Antunes, N., C. Fricker, C., Guillemin, F. and Robert, P.(2006) - Perturbation analysis of a variable M/M/1 queue: a probabilistic approach. *Advances in Applied Probability* 38(1): 263-283.

Brás Silva, H., Brito, P. e Pinto da Costa, J. (2006) - A partitional clustering algorithm validated by a clustering tendency index based on graph theory, Pattern Recognition. Vol. 39, n.5, p.766-788.

Braumann, C.A. (2007) - Harvesting in a random environment: Itô or Stratonovich calculus? *J. Theoretical Biology* 244: 424-432.

Braumann, C.A. (2007) - Itô versus Stratonovich calculus in random population growth. *Mathematical Biosciences* 206: 81-107.

Coelho, C. A., Alberto, R. P., Grilo, L. M. (2006) - A mixture of Generalized Integer Gamma distribution as the exact distribution of the product of an odd number of independent Beta random variables with first parameter evolving by 1/2. Applications. *Journal of Interdisciplinary Mathematics*, **9**, 2, 228-248.

Coelho, C. A., Alberto, R. P., Grilo, L. M. (2006) - A mixture of Generalized Integer Gamma distribution as the exact distribution of the product of an odd number of independent Beta random variables with first parameter evolving by 1/2. Applications. *Advances in Interdisciplinary Mathematics*, Edited by: Sat Gupta and B. K. Dass, Taru Publications, New Delhi, 1-20.

Curto, J. D. e Pinto, J. C. (2007) - New Multicollinearity indicators in linear regression models. *International Statistical Review*, 75 (1), 114-121;

Curto, J. D., Pinto, J. C. e Fernandes, J. E. (2006) - World equity markets: a new approach for segmentation. *Finance & Uver – Czech Journal of Economics and Finance*, 56, No. 7-8, 344-360;

Ferreira, F. and Pacheco, A. (2006) - Analysis of GI/M/s/c queues using uniformization. *Computers and Mathematics with Applications* 51(2): 291-304.

Gomes, M. I. and Pestana, D. D. (2007) - A sturdy reduced-bias extreme quantile (VaR) estimator. *J. Amer. Statist. Assoc.* 102, 280-282.

Gomes, M. I. and Pestana, D. D. (2007) - A simple second order reduced bias tail index estimator. *J. Statistical Computation and Simulation*, 77, 487-504, 2007.

Gomes, M. I., de Haan, L., and Pestana, D. D. (2006) - A note on Joint exceedances of the ARCH process. *J. Applied Probab.* 43, 1206.

Grilo, L. M. e Coelho, C. A. (2007) - Development and Comparative Study of two Near-exact Approximations to the Distribution of the Product of an Odd Number of Independent Beta Random Variables. *Journal of Statistical Planning and Inference*, **137**, 1560-1575.

Kwiatkowska, M., Norman, G. and Pacheco, A. (2006) - Model checking expected time and expected reward formulae with random time bounds. *Computers and Mathematics with Applications* 51(2): 305-316.

Morais, M. C. and Pacheco, A. (2006) - Assessing the impact of head starts in the performance of one-sided Markov-type control schemes. *Sequential Analysis* 25(4): 405-420.

Morais, M. C. and Pacheco, A. (2006) - Combined CUSUM-Shewhart schemes for binomial data. *Economic Quality Control* 21: 43-57.

Pacheco, A. and Ribeiro, H. (2006) - Algorithms for computing moments of the length of busy periods of single-server systems. *WSEAS Transactions on Computers* 11(5): 2856-2861, 2006.

Pinto, J. C. e Curto, J. D. (2007) - The organizational configuration concept as a contribution to the performance explanation. *European Management Journal*, Volume 25 (1), 60-78;

Pinto da Costa, J. e Soares, C. (2007) - Rejoinder to letter to the editor from C. Genest and J-F. Plante concerning 'Pinto da Costa, J. & Soares, C. (2005) A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*. Vol. 49, n. 2, p.205-7.

Pinto da Costa, J. e Roque, L.(2006) - Limit distribution for the weighted rank correlation coefficient, R_W. *Revstat - Statistical Journal*. Vol. 4, n. 3, p. 189-200.
Sepúlveda, N., Paulino, C. D., Carneiro, J. and Penha-Gonçalves, C. (2007) - Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity* 99:173-184.

• Teses de Mestrado

Título: Análise de Clusters, Análise de Variância e sua Aplicação à Seriação de Escolas Secundárias em Portugal
Autora: Carla Maria Esteves Miranda, *carlamem@hotmail.com*
Orientadora: Maria Teresa Alpuim

Título: Metodologias na Construção de Tabelas de Mortalidade
Autora: Eugénia Maria Sousa Carvalho, *eugeniacarvalho@sapo.pt*
Orientadora: Maria Fernanda Oliveira

Título: Análise de Dados Espaciais Referentes a Doenças Raras
Autora: Mafalda Isabel Durães Lira, *mafaldalira@mail.pt*
Orientadora: Maria Lucília Carvalho

Título: Análise de Sobrevivência – Aplicação ao Estudo do Melanoma Maligno da Pele
Autora: Joana de Lima Bastos, *joanabas@med.up.pt*
Orientadora: Cristina Simões Rocha

Título: Estudo da «*Relevância do apoio da Escola nas perspectivas profissionais dos alunos do 10º ano de escolaridade*» com aplicação dos Modelos Lineares Hierárquicos
Autor: Vitor Valente*, vmfvalente@netcabo.pt*
Orientadora: Teresa A. Oliveira

Título: Um Modelo para Estimar a Probabilidade de Morte nos Cuidados Intensivos para Vítimas de Traumatismos Múltiplos
Autora: Teresa Margarida Coimbra da Silva, *teresa_coimbra_75@hotmail.com*
Orientador: João Gomes

Título: Componentes Principais: o método e suas generalizações
Autor: Paulo Canas Rodrigues, *paulocanas@gmail.com*
Orientador: João A. Branco

Título: Famílias Exponenciais
Autora: Maria Osvaldo Dias, *mmdias@gmail.com*
Orientador: Dinis Pestana

Título: Classes de Panjer e Extensões
Autor: Iqbal Abamahoumed
Orientador: Dinis Pestana

Título: Teorema Limite Central: Extensões
Autora: Susana Maciel, *susana-maciel@sapo.pt*
Orientador: Dinis Pestana

Título: Planos Factoriais Fraccionados de Base Dois
Autora: Sandra Monteiro, *smonteiro@esce.ips.pt*
Orientadora: Teresa Oliveira

Título: Análise de clusters aplicada ao Sucesso/Insucesso em Matemática
Autora: Guida Quintal, *guida.cq@netmadeira.com*
Orientadora: Rita Vasconcelos

Título: Análise de correspondências: uma perspectiva em torno do método e das aplicações.
Autora: Patrícia Ferreira, *patriciaferreira@ist.utl.pt*
Orientador: João A. Branco

Título: Statistical models to predict electricity prices.
Autora: Tânia Silva, *tc.marquessilva@iol.pt*
Orientadores: Cláudia Nunes e António Pacheco

Título: Previsão do Rendimento da Madeira de Eucalyptus Globulus com Base em Espectroscopia Nir.
Autora: Alexandra Pereira, *ampmayerdasilva@mail.telepac.pt*
Orientadora: Maria do Rosário Oliveira.

Título: Privação em Portugal – Análise multidimensional da pobreza com base em Modelos Bayesianos de Classes Latentes
Autora: Carla Machado, *Carla.Machado@gep.mtss.gov.pt*
Orientadores: José Duarte Nunes e Carlos Daniel Mimoso Paulino

• Teses de Doutoramento

Título: Análise Multivariada de Variáveis Dependentes Não-comensuráveis
Autor: Armando Teixeira-Pinto, *tpinto.home@gmail.com*
Orientador: Sharon-Lise Normand

Na minha tese desenvolvi um modelo de variáveis latentes para situações de múltiplas variáveis de dependentes (*outcomes*) não-comensuráveis. Na investigação biomédica, é comum estarmos interessados em analisar múltiplas variáveis dependentes (VD) que, potencialmente, estão correlacionadas. Quando estas variáveis são medidos na mesma escala, ou são da mesma natureza, existem diferentes métodos que permitem a modelação das mesmas num contexto multivariado, como por exemplo, os modelos multivariados de regressão linear generalizados, modelos de efeitos aleatórios mistos e equações de estimação generalizadas (GEE). No entanto, se as VD são não-comensuráveis, i.e., de natureza distinta e medidas em diferentes escalas (e.g. contínuas e categóricas), a dificuldade de especificar uma distribuição multivariada para tais casos torna o problema da modelação complexo e a generalização dos métodos multivariados referidos não é trivial. A solução adoptada em muitas análises, é ignorar a correlação entre VD e modelar cada uma deles separadamente (equivalente a assumir independência dos variáveis de resposta no contexto multivariado). As consequências desta abordagem são a eventual perda de eficiência dos estimadores e potencial enviesamento em situações de dados omissos.
Desta forma, propus uma abordagem que, utilizando variáveis latentes para induzir a correlação entre as VD, permite modelar todas as VD num contexto multivariado. Propus ainda a extensão dos GEE para VD não comensuráveis e estudei ambas as técnicas em situações de dados omissos completamente ao acaso (MCAR) e omissos ao acaso (MAR).
Um resumo mais completo deste trabalho pode ser encontrado em http://users.med.up.pt/tpinto

Armando Teixeira-Pinto

Título: Conjuntos Difusos: Uma Abordagem Estatística
Autor: Abdul Kadir Suleman, *abdul@iscte.pt*
Orientadora: Elizabeth Azevedo Reis

A minha tese faz um levantamento de um modelo paramétrico – designado modelo GoM, *Grade of Membership*, (Woodbury & Clive, 1974) –, que se baseia na teoria dos conjunto difusos. Nela procurei atingir dois objectivos distintos: um, no domínio metodológico, propondo estimadores pseudo-bayesianos para os seus parâmetros, como alternativa ao de máxima verosimilhança; outro, no domínio empírico, aplicando-o na tipificação de acidentes domésticos e na perfilização de competências bancárias, da função comercial. O modelo foi desenvolvido especificamente para as ciências médicas, pelo que as aplicações referidas devem ser entendidas também como uma tentativa de o descolar dessa área do saber.

A proposta de estimação alternativa assenta nos trabalhos pioneiros de Good (1965, 1967) e de seus seguidores Sutherland et al. (1974), sobre a estimação de probabilidades multinomiais. Estes investigadores propõem uma correcção do estimador de máxima verosimilhança, utilizando um estimador que resulta de uma técnica bayesiana. A minha proposta difere da dos autores referidos apenas no critério utilizado para efectuar a correcção.

A experimentação, baseada em simulação estocástica, mostra que o desempenho de um dos estimadores assim obtido é consideravelmente superior ao dos outros dois – um, resultante de uma transliteração do de Sutherland et al., e o de máxima verosimilhança –, constituindo uma alternativa credível a este último.

Na avaliação de acidentes domésticos, ocorridos em Portugal, em 2003 e 2004, foi possível tipificar acidentes em cada uma das diversas divisões de uma habitação. Essa tipificação pode servir de base de trabalho na definição de políticas de prevenção de lesões, no âmbito de um programa Comunitário.

A perfilização de competências bancárias, colhida da abordagem baseada em conjuntos difusos, permitiu verificar que a função comercial se realiza por competências equilibradas a um certo nível qualitativo e não tanto por especialização em parte delas.

Paralelamente, foi estabelecido um resultado teórico (matemático) que, sob certas condições, perfaz uma ordenação em simplex unitário. Uma concretização desse resultado, designada função competência, permite obter estimativas da posição de cada trabalhador, numa classificação de competências. Essa função constitui um instrumento analítico inovador na avaliação de desempenho.

Abdul Suleman

Título: Desenvolvimento de distribuições quase-exactas para vários cenários de utilização da estatística Lambda de Wilks
Autor: Luís Grilo, *lgrilo@aim.estt.ipt.pt*
Orientador: Carlos Agra Coelho

Na minha tese, para ilustrar o conceito e procedimento utilizados no desenvolvimento de distribuições quase-exactas, obtive algumas destas aproximações para a distribuição Logbeta, quer a partir da factorização da função característica original quer a partir de truncaturas desta, quando expressa como mistura infinita de Exponenciais. Estas aproximações igualam alguns dos primeiros momentos exactos. Comparei a distribuição exacta com as quase-exactas e assimptóticas, em termos de momentos e quantis, e avaliei a proximidade destas utilizando duas medidas desenvolvidas com base nos limites de Berry-Esseen. Por aplicação directa dos resultados obtidos desenvolvi e analisei distribuições exactas, quase-exactas e assimptóticas para o produto de variáveis aleatórias independentes Beta e, ainda, para diferentes cenários de utilização da estatística Λ de Wilks: a) dois grupos, ambos com número ímpar de variáveis; b) vários grupos de variáveis, sendo: i) no máximo, três grupos com número ímpar de variáveis; e, ii) dois ou mais grupos com número ímpar de variáveis. Considerei, ainda, variações no número de variáveis por grupo e na dimensão da amostra. As distribuições quase-exactas obtidas são manejáveis e não envolvem integrais não resolvidos nem expansões em série, mostrando-se mais próximas das exactas do que as assimptóticas, nomeadamente para amostras de dimensão reduzida.

Luís Grilo

Título: Estimation and Testing for Distributions with Light, Heavy and Super-heavy Tails
Autora: Claudia Neves, *claudia@mat.ua.pt*
Orientadores: Maria Isabel Fraga Alves e Laurens de Haan

A minha tese ocupa-se essencialmente do estudo e desenvolvimento de testes de hipóteses que preconizam a selecção de domínios de atracção em Teoria de Valores Extremos. Em princípio, qualquer metodologia estatística de estimação de determinado parâmetro de interesse não se esgota na estimação pontual desse mesmo parâmetro, mas permite ainda a construção de testes de hipóteses assentes em propriedades estocásticas do estimador adoptado. Na senda desta possibilidade, é apresentado um estimador para um parâmetro α não negativo que determina o peso da cauda da distribuição subjacente às observações amostrais. As propriedades de consistência e normalidade assimptótica do estimador agora proposto são então trabalhadas também com vista à obtenção, num momento ulterior, de um teste de hipóteses que possibilita discriminar entre caudas pesadas e super-pesadas. É de notar que estas últimas, as distribuições de cauda super-pesada, não se inscrevem em nenhum domínio de atracção extremal mas formam uma classe que pode ser localizada como adjacente à classe de distribuições de cauda pesada do domínio de atracção Fréchet (e portanto com índice de cauda α positivo) ao surgirem associadas a um índice de cauda igual a zero. De referir ainda que este índice indica o número de momentos finitos da distribuição subjacente e, neste sentido, determina que uma distribuição de cauda super-pesada não possui qualquer momento finito. Finalmente, aborda-se o problema da estimação do índice de valores extremos aqui denotado por γ. Sob uma perspectiva paramétrica, o índice γ determinaria a forma da distribuição subjacente, enquanto que segundo uma abordagem semi-paramétrica, apontará o tipo da distribuição de valores extremos a alcançar.
O desempenho das metodologias e procedimentos apresentados nesta tese é também estudado para amostras de dimensão finita, recorrendo a simulação pelo método de Monte-Carlo, e ilustrado mediante aplicação a conjuntos de dados publicados tendo, alguns deles, sido amplamente divulgados na literatura.

Cláudia Neves

Título: Assessing Spatial Dependency under Non-Standard Sampling
Autora: Raquel Menezes, *rmenezes@mct.uminho.pt*
Orientadores: Manuel Febrero-Bande e Pilar Garcia-Soidán

Na minha tese, o trabalho desenvolvido centrou-se no contexto da Geoestatística, tendo-se recorrido à teoria de Processos Pontuais em parte das soluções propostas. Os objectos do nosso estudo foram a estimação da estrutura de correlação espacial, e a subsequente predição espacial, debaixo de métodos de amostragem considerados não-standard.

Em Geoestatística, supõe-se tipicamente que as localizações amostradas se encontram igualmente distribuídas sobre a região de observação, caso contrário está-se na presença de *clustering*. Adicionalmente, assume-se que o processo pontual associado às localizações (por exemplo, Poisson homogéneo) não depende do processo associado aos dados medidos nas várias localizações (por exemplo, dados Gaussianos). Quando a dependência estocástica anterior ocorre, está-se na presença de *amostragem preferencial*, um fenómeno comum em redes de monitorização ambiental onde existe a tendência natural de colocar monitores em áreas classificadas como de maior risco.

Começámos por provar que o impacto da falha dos pressupostos anteriores nos métodos tradicionais da Geoestatística não pode ser ignorado. No que diz respeito ao problema de *clustering*, propõe-se um variograma tipo núcleo que integra uma compensação para as áreas de menor densidade de pontos. Procedeu-se com o estudo teórico do estimador proposto provando-se gozar boas propriedades, tais como ser assimptoticamente centrado e consistente.

No que diz respeito ao problema de *amostragem preferencial*, propõe-se uma abordagem baseada-ao-modelo (Diggle et al, 1998), ou seja baseada em fortes assunções paramétricas. Nomeadamente, assume-se que as localizações podem ser modeladas por Processos de Cox log-Gaussianos (Moller et al, 1998). Os resultados obtidos sugerem que inferências correctas podem ser recuperadas quando os dados amostrados são tratados como realizações de Processos Pontuais Marcados.

Raquel Menezes

Título: Estimação de Parâmetros de Acontecimentos Raros
Autor: Frederico Almeida Gião Gonçalves Caeiro, *fac@fct.unl.pt*
Orientadora: Maria Ivette Gomes

Na minha tese estudou-se, num contexto i.i.d. e de caudas pesadas, a estimação semi-paramétrica de parâmetros de modelos de acontecimentos raros. Começou-se pela estimação do índice de valores extremos, o parâmetro de forma da distribuição Generalizada de Valores Extremos. Com o objectivo de melhorar o comportamento dos estimadores clássicos do índice de valores extremos, introduziram-se dois novos estimadores assintoticamente centrados. O primeiro tem um parâmetro de controlo que, se for devidamente escolhido, anula a componente dominante de viés assintótico. No entanto, a redução de viés é feita à custa do aumento significativo da variância, uma propriedade comum a muitos outros estimadores, assintoticamente centrados, existentes na literatura. No segundo estimador, e com a estimação adequada de parâmetros de segunda ordem, conseguiu-se remover o termo dominante de viés sem alterar o valor da variância assintótica. Apresentam-se algumas propriedades assintóticas e faz-se uma comparação com o estimador clássico, através de um estudo de simulação. Como ilustração apresentam-se dois conjuntos de dados financeiros e estima-se o índice de valores extremos.
Foi também abordada a estimação de um parâmetro de escala de primeira ordem. Mais uma vez procedeu-se à estimação deste parâmetro através de estimadores assintoticamente centrados. Apresentam-se algumas propriedades assintóticas e faz-se a comparação com o estimador clássico deste parâmetro, através de um estudo de simulação.
Finalmente, aborda-se a estimação de um parâmetro de "escala" de segunda ordem, um tema bastante recente na literatura. Trata-se de um parâmetro essencial para o desenvolvimento de estimadores de viés reduzido de parâmetros de primeira ordem. É estudado o comportamento assintótico, e o comportamento para amostras de dimensão finita de alguns modelos de cauda pesada.

Frederico Caeiro

Título: A Estatística Bayesiana Na Identificação Forense – Análise e avaliação de vestígios de DNA com redes bayesianas
Autora: Marina Andrade, *marina.andrade@iscte.pt*
Orientadores: Manuel Alberto M. Ferreira e João Pedro Faria

Na minha tese tomou-se como ponto de partida alguns dos problemas que se colocam aos tribunais e dizem respeito à interpretação das provas fornecidas ligadas à tomada de decisão em contexto de incerteza. Procurou-se neste trabalho relacionar a sua quantificação em termos probabilísticos, em particular no enquadramento das decisões judiciais. O uso cada vez mais comum de provas e vestígios apresentados na forma quantificada salientam a necessidade de avaliar e interpretar correctamente essas provas. Importa também eliminar muitas das dúvidas que ainda se observam. Situar estas questões no âmbito da metodologia bayesiana é o início de um caminho que tem ainda muito para percorrer.
Com base na metodologia adequada abordou-se o tratamento e análise de provas biológicas e as questões que se têm levantado, quer em disputas de paternidade quer em casos criminais.
Face à crescente complexidade dos problemas que se colocam, apresentou-se um recurso maleável para melhor responder a esses problemas – as redes bayesianas. Neste contexto, o uso de redes bayesianas "object-oriented" pela sua flexibilidade e modularidade surgem como ferramenta para representar graficamente as relações de dependência entre as variáveis de interesse.
Partindo-se de um caso simples de disputa de paternidade foram-se acrescentando elementos que pretendem aproximar os problemas a casos reais mais complicados. Da mesma forma, já num plano mais complexo que envolve misturas de perfis de DNA, partiu-se da análise de um único vestígio, com diferentes marcadores, avançando-se para problemas de dois vestígios diferentes, para diferentes marcadores considerando-se ainda o condicionamento na presença dos possíveis indivíduos desconhecidos.

Marina Andrade

Título: Estimação Não Paramétrica em Modelos de Regressão de Dados de Contagem com Excesso de Zeros
Autor: José António Santos, *JSantos@isegi.unl.pt*
Orientadora: Manuela Neves Figueiredo

Na minha tese, são desenvolvidos modelos semi-paramétricos de análise de regressão de dados de contagem, alternativos às formulações paramétricas existentes, que apresentam a vantagem da sua validade estatística, em matéria de inferência estatística, não depender de condições de especificação.
O modelo de Poisson é a base da análise de regressão de dados de contagem, que apresenta a desvantagem da restrição da igualdade entre a média e a variância condicionais, designadamente em dados de contagem com excesso de zeros.
O modelo de regressão binomial negativo admite sobredispersão, desde que não seja muito acentuada. Para ultrapassar as deficiências destes dois modelos, existem estimadores de quasi e pseudo-verosimilhança e modelos de Poisson generalizados.
Os modelos de regressão paramétricos mais comuns para dados de contagem com excesso de zeros são o modelo de Poisson inflacionado em zero (ZIP) e o modelo binomial negativo inflacionado em zero (ZINB). A validade estatística destes modelos, designadamente quanto à inferência estatística, depende da correcta especificação do modelo e de condições de regularidade. Se estas hipóteses não são verificadas, os resultados de inferência estatística deixam de ser válidos.
Neste trabalho, são apresentados cinco alisadores semi-paramétricos de máxima verosimilhança local com base nos modelos de regressão de Poisson, binomial negativo, ZIP, ZINB e de regressão logística. Deduz-se o viés, a variância e a distribuição assintótica de cada estimador. São apresentados alguns resultados de simulação e um exemplo de aplicação da literatura, na área da medicina, em homens infectados com o vírus HIV.
Os resultados são encorajadores e suscitam matéria para investigação futura, como seja a escolha dos graus dos polinómios do alisador e do estimador do seu viés, o critério de selecção da largura de banda óptima, cuja natureza deverá ser variável, e o desenvolvimento de outras formulações, para outro tipo de dados, no âmbito dos modelos lineares generalizados.

José António Santos

Título: Quase-normalidade e inferência para séries de estudos emparelhados
Autor: Luís Pedro Ramos, *lpcr@fct.unl.pt*
Orientador: João Tiago Mexia

Na minha tese desenvolvemos tecnicas de inferência para séries de estudos emparelhados. Este trabalho tem base em duas técnicas de ACT: STATIS e STATIS Dual. A metodologia STATIS utiliza distâncias euclidianas entre disposições observadas em k situações a que chamamos estudos. Estas duas metodologias permitem estudar apenas séries isoladas de estudos consistindo em projecções ortogonais sobre planos paralelos aos dois primeiros vectores próprios da matriz de Hilbert-Schmidt, sendo as conclusões empíricas. Na nossa metodologia consideramos apenas a existência de uma estrutura comum, isto é, considerando apenas o primeiro vector próprio da matriz de Hilbert Schmidt e na qual fizemos inferência para séries isoladas e séries emparelhadas de estudos.

Os estudos que fizemos assentam na técnica de quase-normalidade. Numa primeira fase apresentamos a quase-normalidade polinomial, mostrando que as distribuições de polinómios, de baixo grau, em variáveis normais independentes com pequenos coeficientes de variação, podem ser aproximadas por distribuições normais. Este resultado permitiu-nos estudar séries de estudos, admitindo a normalidade e pequenos coeficientes de variação das observações iniciais.
Utilizando a abordagem referida estudamos primeiramente séries isoladas de estudos e, em seguida, séries emparelhada. Nesta segunda fase consideramos séries associadas aos tratamentos de delineamentos ortogonais, particularizando para o caso dos factoriais de base prima.

Luís Ramos

Título: Distance-Based Methods for Classification and Clustering of Time Series
Autor: Jorge Caiado, *jcaiado@esce.ips.pt*
Orientadores: Nuno Crato e Daniel Peña

Esta tese aborda o problema da classificação e agrupamento de séries temporais. A identificação de semelhanças ou dissemelhanças entre séries temporais tem sido estudado na literatura estatística. No entanto, alguns estudos usam métodos não paramétricos para classificar e agrupar grupos de séries através da distância euclidiana entre os pontos no espaço. Esta métrica tem a importante limitação de ser invariante a transformações que modifiquem a ordem das observações ao longo do tempo e assim não toma em consideração a estrutura de autocorrelações das séries temporais.

Nesta tese, desenvolvemos novas medidas de distância entre séries temporais baseadas nas autocorrelações, nas autocorrelações parciais, nas autocorrelações inversas e nas ordenadas do periodograma. São apresentados resultados de simulação de Monte Carlo comparando estas com outras métricas paramétricas e não paramétricas. Em particular, discutimos a comparação de séries estacionárias de modelos ARMA de diferentes tipos, a comparação entre séries quase não estacionárias e séries não estacionárias, a comparação entre processos de tendência determinística e processos de tendência estocástica, e a comparação de processos de memória longa e memória curta. Introduzimos igualmente métricas no domínio tempo e frequência para classificar e agrupar séries com um número desigual de observações. São apresentados testes de hipóteses para averiguar se duas séries têm o mesmo processo de geração de dados. A potência e a dimensão dos testes são estimadas através de estudos de simulação. Como aplicações económicas, apresentamos dois exemplos ilustrativos. No primeiro, utilizamos dados de séries temporais económicas para identificar semelhanças entre sectores de produção industrial nos Estados Unidos. No segundo, fazemos a aplicação do método de classificação baseado na interpolação do periodograma de séries com amostras desiguais para agrupar países industrializados.

Jorge Caiado

Título: Distribuições Conjugadas e Aproximações
Autora: Madalena Malva, *malva@mat.estv.ipv.pt*
Orientador: Dinis Pestana

Na minha tese, como em Estatística, os resultados assintóticos como o Teorema Limite Central são a base de sustentação de muitas aproximações feitas. Naturalmente, há que estabelecer resultados fortes sobre a velocidade de convergência, para assim justificar, pelo menos parcialmente, o recurso a resultados assintóticos em situações pré-assintóticas, ou mesmo com pequenas amostras.
A teoria dos polinómios ortogonais permite uma abordagem simples a expansões em série, nomeadamente às expansões de Edgeworth. Analisámos o problema da aproximação de ângulos diferentes, mostrando que uma variante importante das expansões de Edgeworth, as expansões **diferidas** (ou *tilted Edgeworth expansions*) são uma outra forma de enquadrar o uso das distribuições conjugadas de Esscher-Cramér-Khinchine, ou as aproximações de Daniels usando pontos de sela.
Para além do teorema limite central clássico, de convergência para lei limite gaussiana, em que conseguimos demonstrar que é possível, usando os cumulantes de 3ª e 4ª ordem, melhorar a velocidade de convergência de $O(1/\sqrt{n})$ para O(1/n), investigámos também velocidades de convergência no caso de lei limite estável não gaussiana, usando um parâmetro de segunda ordem que refina a variação lenta das caudas que caracterizam os domínios de atracção das leis estáveis.
Enquadrando estas questões numa perspectiva de entropia e informação, verificamos que a melhoria da convergência de $O(1/\sqrt{n})$ para O(1/n), no teorema limite central clássico, é o que se deve esperar na família exponencial natural, sendo possível obter uma generalização da desigualdade de Cramér-Rao que perspectiva devidamente esta problemática.

Madalena Malva

Título: Análise Preditiva em Populações Finitas
Autora: Susana Rosado-Ganhão, *srosado@fa.utl.pt*
Orientadora: Maria Antónia Amaral Turkman

A minha tese incide sobre a análise preditiva em populações finitas sob a forma de intervalos de predição. Neste âmbito foram estabelecidas algumas contribuições originais, em particular na introdução da aleatoriedade da dimensão da população.
Percorreram-se diversos métodos de predição usando metodologia clássica e bayesiana e aumentou-se progressivamente o leque de possíveis aplicações deste trabalho. Partiu-se de uma população de dimensão fixa, progrediu-se para populações de dimensão aleatória cuja distribuição é totalmente conhecida e finalmente fez-se uma aplicação (na área do seguro automóvel) com recurso aos modelos bayesianos hierárquicos para considerar a situação da dimensão da população ser aleatória com distribuição desconhecida. Neste último caso aumentou-se a complexidade ao nível do cálculo, ultrapassada com recurso ao método de amostragem Gibbs obtendo-se, assim, resultados inovadores.
Neste ponto introduziu-se a reamostragem Bootstrap combinada com a amostragem Gibbs com o objectivo de melhorar a performance dos intervalos de predição verificando-se, no entanto, que o aumento de esforço computacional não era justificado.
Em termos de resultados a metodologia bayesiana foi a que, de um modo geral, teve um melhor desempenho ao nível da cobertura dos intrevalos de predição obtidos. Além disso esta metodologia aplica-se em qualquer situação sendo facilmente generalizada para o caso em que a dimensão da população é desconhecida.
A metodologia desenvolvida foi aplicada a dados reais de uma pedreira com o objectivo de predizer o total de fracturas, por piso, com determinadas características e a dados reais de uma seguradora com vista a predizer o montante total de indemnizações pagas no ramo automóvel, num determinado ano, por tipo de indemnização.

Susana Rosado-Ganhão

• Livros

Título: Estatística - Ciência Interdisciplinar. Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística
Editores: Maria Eugénia Ferrão, Célia Nunes e Carlos A. Braumann
Editora: SPE - Sociedade Portuguesa de Estatística
Ano: 2007

Título: Introdução aos Métodos Estatísticos Robustos
Autores: Ana M. Pires e João A. Branco
Editora: SPE - Sociedade Portuguesa de Estatística
Ano: 2007

Prémios Estatístico Júnior 2007

A atribuição de prémios "Estatístico Júnior 2007" é promovida pela Sociedade Portuguesa de Estatística, 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. Ao apelo para submissão de trabalhos correspondeu uma adesão elevada, tendo sido recebidos 8 trabalhos na categoria Ensino Básico, envolvendo um total de 21 alunos e 7 na categoria Ensino Secundário, envolvendo um total de 16 alunos.

A Sociedade Portuguesa de Estatística tem agora o prazer de anunciar os trabalhos premiados:

Trabalho classificado em 1º lugar (Ensino Básico)

Título: ***Ocupação dos alunos do 7º Ano da ESAG nas Férias da Páscoa***
Autoria: Sérgio Marques Ferreira
Bárbara Oliveira de Marco
Ana Sofia Mendes Oliveira
Estabelecimento de Ensino: Escola Secundária Artur Gonçalves, Torres Novas.
Ano de escolaridade: 7º Ano
Professor orientador: Maria Alice da Silva Martins

Trabalho classificado em 2º lugar (Ensino Básico)

Título: ***Abandono Escolar***
Autoria: Rita Sofia Duarte Nobre
Estabelecimento de Ensino: Escola Secundária Prof. Ruy Luís Gomes, Almada.
Ano de escolaridade: 7º Ano
Professor orientador: Andreia Alexandra de Jesus Vilas Rodrigues

Trabalhos classificados em 3º lugar – ex aequo (Ensino Básico)

Título: ***Será a Matemática apelativa entre os alunos***
Autoria: Renata Santos Reis
Catarina do Serro Pedrosa
Isa Mara da Silva Matos
Estabelecimento de Ensino: Escola E. B. 2,3 da Mexilhoeira Grande, Portimão.
Ano de escolaridade: 7º Ano
Professor orientador: Clara Maria Lourenço Marquês

Título: ***Obesidade e hábitos alimentares***
Autoria: Igor Manuel Oliveira Pacheco
Vanessa Bürgi
Carlos Rafael Alemão Furtado
Estabelecimento de Ensino: Escola E. B. 2,3 da Mexilhoeira Grande, Portimão.
Ano de escolaridade: 7º Ano
Professor orientador: Clara Maria Lourenço Marquês

Título: ***Estudar é importante***
Autoria: Catarina Mafalda Correia Costa
Verónica João Paneia
Pavlo Koryakov
Estabelecimento de Ensino: Escola E. B. 2,3 da Mexilhoeira Grande, Portimão.
Ano de escolaridade: 7º Ano
Professor orientador: Clara Maria Lourenço Marquês

A entrega dos Prémios Estatístico Júnior 2007, conforme estipulado no Regulamento, teve lugar na Sessão de Encerramento do XV Congresso Anual da Sociedade Portuguesa de Estatística, que decorreu no dia 21 de Agosto de 2007, pelas 12:30 horas, no Instituto Superior de Ciências do Trabalho e da Empresa (ISCTE), em Lisboa.

Informa-se ainda que neste ano o Júri só atribuiu prémios ao Ensino Básico por considerar que os trabalhos candidatos do Ensino Secundário não atingiram o nível desejável de modo a serem contemplados com a atribuição de um prémio.

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 Doutor Russell Alpizar-Jara do Departamento de Matemática da Universidade de Évora.

Prémio SPE 2007

Estimação do índice de cauda em modelos de caudas pesadas: acomodação do viés nos excessos acima de um "threshold" elevado[1]

Lígia Henriques Rodrigues, *ligia.henriques@aim.estt.ipt.pt*
Instituto Politécnico de Tomar e CEAUL

A metodologia POT (do inglês "Peaks Over Threshold") foi introduzida em Smith [5], onde as excedências acima de um nível elevado são modeladas por uma distribuição Generalizada de Pareto. A re-parametrização do modelo Generalizado de Pareto sugerida por Davison [1] permitiu obter uma expressão explícita para o estimador de máxima verosimilhança do índice de valores extremos γ, aqui designado de POT-ML. Smith [5] obteve o comportamento assintótico deste estimador considerando o nível elevado *u* fixo. O seu resultado foi posteriormente reescrito por Gomes [3], no contexto de um nível aleatório elevado, obtendo-se o mesmo comportamento assintótico. A generalização deste resultado pode ser encontrada em Drees *et al*. [2].

Neste trabalho, e num contexto de modelos de cauda pesada pertencentes à classe de Hall, apresentamos uma nova classe de estimadores de máxima verosimilhança do índice de cauda γ, designada de POT-MP. A metodologia usada para a derivação dos estimadores baseia-se na acomodação do viés no modelo Generalizado de Pareto dos excessos acima de um nível aleatório elevado. Os parâmetros de segunda ordem β e ρ foram estimados externamente num nível k_1 adequado de ordem superior a k, o número de estatísticas ordinais de topo utilizadas na estimação do índice de cauda. O comportamento assintótico da nova classe de estimadores é obtido face a condições fracas impostas ao parâmetro de segunda ordem ρ. Os resultados obtidos permitem-nos concluir que as duas classes de estimadores, POT-ML e POT-MP, têm a mesma variância assintótica e um viés assintótico diferente.

Analisamos ainda, num pequeno estudo de simulação, via método de Monte Carlo, o comportamento exacto dos estimadores de máxima verosimilhança em amostras de dimensão pequena. As estimativas do estimador POT-ML foram obtidas através do algoritmo proposto por Grimshaw [4] para obtenção de estimativas de máxima verosimilhança no modelo Generalizado de Pareto. Desenvolvemos, também, uma versão modificada do algoritmo de Grimshaw de forma a obtermos as estimativas do estimador POT-MP. Os resultados de simulação mostram que o estimador POT-MP apresenta, regra geral e para modelos com valores de $\gamma+\rho \neq 0$, trajectórias estáveis, sendo pois uma boa alternativa ao estimador POT-ML.

Referências:

[1] Davison, A. (1984). Modeling excesses over high threshold with an application. In J. Tiago de Oliveira ed., *Statistical Extremes and Applications*, D. Reidel, 461-482.

[2] Drees, H., Ferreira, A. and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index. *Ann. Appl. Probab.* **14**: 1179-1201.

[3] Gomes, M. I. (2002). *A note on the excesses over a high threshold*. Notas e Comunicações CEAUL 10/2002.

[4] Grimshaw, S. D. (1993). Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution. *Technometrics* **35**:2, 185-191.

[5] Smith, R.L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15**:3, 1174-1207.

Lígia Henriques Rodrigues, **galardoada com o Prémio SPE 2007**, licenciou-se em Matemática Aplicada e Computação - ramo de Probabilidades e Estatística pelo Instituto Superior Técnico. Mestre em Matemática Aplicada pela Universidade de Évora encontra-se, actualmente, a realizar o doutoramento em Probabilidades e Estatística na Faculdade de Ciências da Universidade de Lisboa, sob orientação da Professora Maria Ivette Gomes. É docente na Escola Superior de Tecnologia de Tomar, do Instituto Politécnico de Tomar.

[1] Investigação parcialmente financiada pela FCT / POCTI e POCI/ FEDER e pela bolsa de doutoramento SFRH/BD/29010/2006 da Fundação para a Ciência e Tecnologia.