Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *outono de 2011*

Análise de Sobrevivência

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

SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

PRÉMIO ESTATÍSTICO JÚNIOR 2012

Candidaturas até
25 DE MAIO DE 2012

CONTACTOS

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

Telef./Fax 21 750 01 20

www.spestatistica.pt

spe@fc.ul.pt

Com o apoio:
Porto Editora

Editorial

… a propósito de um jantar de homenagem! Algumas palavras, em 21 de Junho de 2011…

Caros Alunos e Estudantes,
Caros Colegas da SPE, do DEIO e da FCUL "a minha segunda casa durante uma vida"
Caros Amigos e Companheiros de caminhada,

Foram muito gentis, aquelas que "me desafiaram a aceitar uma homenagem" que eu não queria e que apenas a vossa gentileza torna merecida. Agradeço todas as mensagens de muitos amigos e colegas bem como a simpatia das palavras recebidas, decerto, não totalmente merecidas.
Como bem sabemos, os nossos "pequenos sucessos" estão sempre associados a mais alguém que connosco colabora ou que nos ajuda - às vezes com a parte mais significativa. É a Força dos Menores que também chamo de outliers - aqueles a quem uma obscura fama esconde. Assim, entendo que esta homenagem tem de ser repartida por muitos!
Estou aposentado! O tempo, mas também nós e as circunstâncias, de algum modo determinam as nossas opções que, como no meu caso, fizeram alterar em dois ou três anos a minha decisão em relação à aposentação. É bom sentir o desejo de mudança. Em algum momento sentimos que "outra atividade" pode prosseguir principalmente sem a pressão da agenda profissional que com o passar dos anos se vai complicando; embora eu dela não tenha grandes razões de queixa. Mas a vida é mais bela sempre que aceitamos novos desafios e de algum modo nos reformulamos ao mesmo tempo que nos vamos reformando.
A vossa presença faz-me sentir muito feliz. Esta presença, muito generosa da vossa parte, cria em mim um pouco de sentimento de dever cumprido. É um bom sentimento associado a uma carreira profissional a que dediquei uma vida. Mas, não se desempenha só. Na vida de um Professor, os estudantes e os funcionários são bases importantes do bom ambiente necessário ao nosso desempenho profissional. A todos quero neste momento recordar. Este é um grande momento com muitas emoções e muitas recordações. É um dia angular. Mas como em todos os virar de página há sempre algo de novo e de expectativa que nos prende como um novo desafio. Várias etapas já percorri e cada uma delas com os seus desafios muito próprios. O primeiro grande desafio, há mais de cinquenta anos atrás, envolveu o meu Reguengos de Monsaraz natal – um "ponto de partida". Porque "sentia que queria ir mais longe" - isso implicava a deslocação para o Liceu Nacional de Évora. Foi o primeiro grande desafio colocado aos meus pais. O segundo, dois anos depois, foi a deslocação para a Faculdade que obrigou ao êxodo para Lisboa. Foi a caminhada (o êxodo!) da família para a grande cidade muito exigente, principalmente para o meu pai. Foi um exemplo de vida e a sua memória hoje desejo invocar. Associo também a minha mãe que, no feminino de outras épocas, foi pilar da nossa grande aventura e que nos seus 92 anos continua com uma vivacidade invejável - a grande homenageada de hoje.
No livro da minha vida estou a chegar a uma página que muitos de vós conheceis. Em 1971, terminada a Licenciatura eu e alguns dos presentes, grandes amigos de longa data, fomos convidados para Assistentes Eventuais da FCUL. Dois anos depois, em tempos de guerra colonial, por longos 27 meses tive de optar pelo serviço militar obrigatório. A alternativa era desertar. Mas nós já estávamos à espera do primeiro filho, que afinal seria a primeira de 3 filhas que neste momento também quero homenagear. E aqui associo mais uma homenageada – a companheira de vida!
Segue- se o doutoramento. Orientado pelo Prof Tiago de Oliveira, tive a felicidade de ser um "pioneiro no estudo das várias formas de "acolhimento" a dados extravagantes" (citei o Prof Murteira no livro *Memorial da SPE*). Muito reconhecido, desejo também homenagear estes dois mestres de referência. Em 1991 um novo desafio: a Agregação e em 1993: Catedrático.
Permitam-me um pouco mais de recordações. A par de uma carreira profissional toda decorrida na Universidade de Lisboa fui convidado durante alguns anos a colaborar com a Universidade Católica do

Porto onde na sua Escola Superior de Biotecnologia tive a oportunidade de ajudar a criar novos cursos na área pioneira de Engenharia e Tecnologia Alimentar - há cerca de 25 anos atrás. E a esta tarefa estão ligados mais alguns dos colegas associados à minha carreira. Logo a seguir veio um convite para colaborar com a Universidade de Évora. E assim, durante alguns anos, voltei às mesmas salas que tinha frequentado 30 anos atrás como aluno do Liceu Nacional e que agora eram pertença da "recriada" Universidade de Évora e que neste momento devo saudar na pessoa do grande amigo de há muitos anos - o Reitor da Universidade e Presidente da SPE. Recordo a Universidade Federal do Rio de Janeiro onde como Professor Visitante tive uma das minhas primeiras experiências profissionais associadas à internacionalização do meu, na altura recente, trabalho de investigação e de doutoramento. Recordo todas as vivências no seio da SPE e do projeto "Direção da SPE" intensamente vivido durante 6 anos e que muito prazer tive em desempenhar. Aqui, junta-se um bom grupo de amigos que hoje também desejo juntar a esta homenagem. Recordo também as Universidades que nos ajudaram a organizar os diversos Congressos SPE e todos os colegas que, muitas vezes no limite, se esforçaram para concretizar esse desígnio da SPE - as Universidades do Algarve, de Évora, dos Açores, do Porto, de Lisboa, e da Beira Interior. Enfim uma grande viagem que foi feita com a ajuda de muitos colegas e que hoje também são homenageados. Durante esses anos assistimos a uma crescente internacionalização da SPE e, em particular, foi intensificada a colaboração com o INE para a organização da 56ª Sessão do ISI. Foi um marco histórico da Estatística em Portugal e muitos dos responsáveis por essa organização hoje também são homenageados. Recordo há 30 anos o início do DEIO e todos os colegas com quem partilhei a maior parte da minha vida profissional – somos uma família e também hoje os quero juntar a esta homenagem. Com esta tão grande lista de homenageados penso que sobra pouco espaço para me colocar a mim próprio. De todos muito bem digo. De mim, tentei que se diga que procurei servir na atividade profissional que decorre há quase 40 anos.

Já longa, esta é a minha história. Como veem, ela é uma história comum. As diferenças estão nos momentos especiais como os de hoje - aqueles que permitem construir o "nosso memorial" - com muitas variáveis. A vida ensina quais são as mais importantes e que de algum modo ajudam a organizar os sonhos de juventude. Essa universidade que todos frequentamos no dia-a-dia e que nas nossas vivências nos faz descobrir a relativização daqueles que às vezes consideramos valores fundamentais. Cria-se um processo de convergência para "o essencial". Esta é uma virtude do envelhecimento. E as variáveis que permitem construir uma carreira e a realização individual são variáveis que descobrimos algumas vezes quando menos se espera. São essas que permitem definir o nosso modelo de discordância perante o qual nos devemos afirmar como outlier. É ser (um) exemplar! Os meus outliers também isto me ensinaram. Missão cumprida!

Outros desafios se perfilam e que, para além de me reformar, me fazem "reformular". Esta mudança de página do livro da vida faz-nos sentir mais livres. Continuo, mais uns tempos, com a Estatística, com os meus outliers, com a dedicação à causa da SPE e com outros desafios que já iniciei… Enfim, novas caminhadas!

Mais do que uma homenagem - que sei não merecer - este convívio tem a grande virtude de aproximar todos aqueles que se dedicam a uma causa. Este é mais um momento em que sentimos a importância e o valor da amizade que também é um pilar na construção das relações humanas. E as relações humanas organizam e estruturam a sociedade. São um valor!

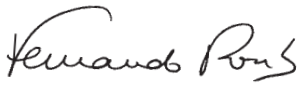
Em breve vou completar os 65 anos de idade. Serei oficialmente idoso. Como diz um economista e teólogo alemão, Anselmo Grün, que muito aprecio (e é da minha idade), por agora, sou um "jovem velho" que tenta a sublime arte de envelhecer – aceitar, renunciar e reconciliar. E as virtudes da velhice – a serenidade, a paciência, a mansidão, a liberdade, a gratidão e o amor - são bons lemas de vida.

Hoje é o dia do Solstício de Verão – data simbólica. O maior dia do ano, decerto por isso, é o Dia do Relógio de Sol. Acima de tudo, é um dia grande e com muito valor no tempo da minha vida. É, também, o Dia Europeu da Música. Temos assim uma bela conjugação com três criadores de um momento especial – o tempo, o sol e a música. Acrescento um pouco de poesia, apoiado em Ruy Belo, com o:

Remate para qualquer poema

Passeou pelo espelho dos dias / suas clandestinas alegrias /que, mal se refletiram, desertaram.

PS. Após o XIX Congresso Anual da SPE, esta mensagem teria de ser continuada. A emoção, a gratidão e todo o enorme ramo de testemunhos e palavras generosas então recebidas, criaram a necessidade de um complemento no registo do Memorial que neste Boletim se inclui a propósito da aposentação de três Presidentes da SPE.
As primeiras palavras, obviamente, são de regozijo por esta iniciativa, muito feliz e original, da Comissão Organizadora do XIX Congresso Anual da SPE e da Direcção da SPE. Desejo personalizar a gratidão nos rostos dessa liderança – os Professores António Pacheco, Alexandra Seco e Carlos Braumann. Mas, como sabemos, há sempre outros - *quos fama obscura recondit* - a quem o nosso reconhecimento se deve estender. Não elaboro a lista para, de modo mais seguro, não cometer o erro da omissão. No entanto, no meu memorial - decerto também no daqueles acima personalizados - esses nomes serão uma referência para este "grande acontecimento SPE".
As restantes palavras, ainda mais comovidas, são dirigidas à Professora Manuela Neves – grande amiga e companheira em muitos dos caminhos e desafios da minha vida profissional. Um enorme obrigado pelo tributo em sessão de homenagem. Foram palavras que muito me emocionaram pois, acima de tudo, num breve momento oratório fez agrupar (talvez) todas as principais paixões do "jovem velho" Professor Aposentado.
Terminada uma vida docente de quarenta anos, este foi também um momento desafiante para que - em continuidade e em novo patamar científico onde a experiência é mestra e o futuro uma proposta - possa consolidar o conhecimento acumulado… pois, a Ciência (Estatística) é paixão que não exige nem aceita aposentação!
Um grande obrigado à comunidade dos estatísticos que, sempre, comigo pode contar.



A partir desta edição vamos "tentar" que o texto siga o novo acordo ortográfico.

O tema central do próximo *Boletim* será *Estatística no Ensino Superior Politécnico.*

Mensagem do Presidente

Caros Colegas:

Realizou-se na Nazaré, de 28 de Setembro a 1 de Outubro de 2011, o nosso XIX Congresso Anual. A Comissão Organizadora, presidida pelo nosso Colega António Pacheco Pires, do Instituto Superior Técnico, e vice - presidida pela nossa Colega Alexandra Seco, do Instituto Politécnico de Leiria, fez um trabalho notável, reconhecido pelos cerca de 200 participantes. Foram apresentadas 88 comunicações orais e 55 comunicações em poster. Tivemos um excelente minicurso das Colegas Maria Salomé Cabral e Maria Helena Gonçalves e o respectivo manual "Análise de Dados Longitudinais", uma apresentação de *software*, a entrega dos Prémios Estatístico Júnior 2011 (com a coordenação da Comissão Especializada de Educação e o apoio da Porto Editora) e a apresentação em poster dos trabalhos vencedores. As quatro conferências convidadas foram de alto nível, três delas proferidas pelos três anteriores presidentes da SPE que se aposentaram este ano (será esta coincidência mero acaso?) em sessões em que os sócios prestaram um tributo de reconhecimento e homenagem a estas personalidades que tanto contribuíram para o desenvolvimento da SPE e da Estatística. O programa social foi muito apreciado. Ele incluiu também um *mixer* organizado pela jSPE (Secção de Jovens Estatísticos), que aqui elegeu a sua Direcção, à qual desejamos bom trabalho. A todos os que participaram, colaboraram ou patrocinaram o Congresso e as actividades que nele decorreram, a nossa sentida gratidão.

Os "Selected Papers" dos dois congressos anteriores, a publicar pelo grupo editorial Springer, estão avançados, um deles quase pronto.

A colaboração internacional entre sociedades estatísticas europeias teve um avanço notável. No decurso do 58th World Statistics Congress of the International Statistical Institute, realizado em Dublin em Agosto, um grupo de uma dezena de sociedades estatísticas de vários países europeus, entre as quais a SPE, que representei, reuniu e decidiu criar a FENStatS - Federation of European National Statistical Societies, tendo elaborado um projecto de Estatutos.

A iniciativa *Estatística Radical,* dirigida a jovens do ensino secundário, e a iniciativa Explorística (exposição interativa de Estatística) receberam financiamento do *Ciência Viva* e, sob a coordenação dos respectivos grupos de trabalho, serão brevemente uma realidade ao serviço da divulgação da Estatística.

A Universidade Católica Portuguesa no Porto será em 2012 o anfitrião do XX Congresso, sendo a Comissão Organizadora presidida pelo Colega Pedro Duarte Silva. Esperamos a vossa presença.

Terminará em Janeiro próximo o segundo mandato dos actuais órgãos sociais. Interpretando o sentimento dos seus membros, registamos o prazer e a honra que foi termos servido a SPE e a Estatística. Expressamos o agradecimento sentido a todos os sócios, particularmente àqueles que mais directamente se envolveram em Comissões, júris de Prémios e outras tarefas, pelo grande apoio que nos deram neste período, não esquecendo, naturalmente, o Editor e revitalizador deste Boletim, o Colega Fernando Rosado. Esse apoio em muito contribuiu para os assinaláveis progressos de que nos podemos todos justamente orgulhar, registados pela SPE durante estes mandatos.

Durante o Congresso foram eleitos os novos órgãos sociais, com uma boa participação e expressivo apoio dos sócios. Aos colegas e amigos que, a partir de Janeiro de 2012, vão pegar na tocha, endereçamos as nossas felicitações e expressamos os votos de que, com o apoio e colaboração de todos, a possam levar mais além.

Saudações cordiais

Mensagem do Presidente Eleito

Terminado o processo eleitoral dos órgãos administrativos da SPE para o triénio 2012-2014 que redundou numa expressiva votação na lista candidata, desejo na qualidade de Presidente da Direção eleita aproveitar esta tribuna para exprimir publicamente os meus agradecimentos

- aos companheiros que me incentivaram a apresentar uma candidatura que considerasse confiável para cumprir os desígnios da SPE, se eleita;
- aos colegas que prontamente se disponibilizaram, quando convidados, a integrar a lista em formação;
- aos sócios que transmitiram um indesmentível apoio à constituição da equipa que se apresentou ao sufrágio direto e às linhas gerais do respetivo programa de candidatura para a futura Direção.

Manifesto igualmente a minha convicção de que esta equipa saberá, com o seu empenho coletivo e as suas múltiplas capacidades, e igualmente com a prestimosa participação de atuais e futuros colaboradores ad hoc, fazer avançar a SPE na prossecução dos seus objectivos centrais, não obstante – e é prudente tomá-lo em consideração – os tempos magros e negros em que se prevê que vai desenrolar-se a sua atividade.

Lisboa, 04 de outubro de 2011

Carlos Daniel Paulino

Memorial

João Branco, perspetivas na Estatística…

Nos momentos de homenagem é sempre patente o clima de emoção. A emoção dos que homenageiam; a emoção dos familiares e amigos, mas sobretudo a emoção do homenageado. Foi neste clima de encontros e reencontros que decorreu o evento *Perspectives in Statistics, A Special Workshop in Honour of Professor João A. Branco*, a 30 de maio de 2011, no Instituto Superior Técnico de Lisboa.

Este encontro científico visou comemorar a jubilação do Professor João Branco, tendo sido uma oportunidade para honrar as suas importantes contribuições na divulgação da Estatística e homenageá-lo como Homem e Universitário.

Um breve sumário do percurso do Professor João Branco, enquanto docente e investigador, relembra-nos o seu início de carreira como investigador na Fundação Gulbenkian. Seguiu-se uma longa carreira académica no Instituto Superior Técnico de Lisboa (IST). Desde então, tem ensinado Probabilidades e Estatística para milhares de alunos graduados e para centenas de estudantes de pós-graduação. Contribuiu para a fundação do Departamento de Matemática do IST e foi também membro fundador do Centro de Matemática e Aplicações (CEMAT), estabelecendo e liderando um grupo de investigadores em probabilidade e estatística. Para além de ter sido presidente do Departamento de Matemática, pertenceu a várias comissões executivas do departamento, assim como ao senado do IST. Foi também presidente vários anos da Sociedade Portuguesa de Estatística.

Como docente e investigador, o Professor João Branco, é autor ou coautor de três livros e de várias dezenas de trabalhos de investigação em estatística. As aplicações da estatística a outras áreas do saber mereceram-lhe uma atenção especial, sendo de realçar propostas científicas que continuam num crescente de utilizadores. Organizou vários encontros científicos, foi membro de comissões científicas de diversos congressos nacionais e internacionais e proferiu dezenas de seminários de divulgação.

Professores passam, mas as marcas que deixam nas nossas vidas permanecem. Antes de alguém ser professor, é aluno. E o Professor João Branco foi aluno da Faculdade de Ciências de Lisboa, onde se licenciou, foi aluno da Universidade de Oxford, onde tirou o seu mestrado, e aluno da Universidade de Newcastle upon Tyne, onde realizou o seu doutoramento.

E, quer sejamos professores ou não, nunca esqueceremos os mestres que passam/passaram nas nossas vidas. Manuela Souto Miranda, Ana M. Pires e M. Rosário de Oliveira fizeram jus à palavra de alunas do mestre e com uma comunicação intitulada *João Branco: the Man and the Professor* salientaram a

sua importância como *Cultivador da Estatística*. Enaltecendo-o como formador dos seus alunos, no entusiasmo pelo saber; no desafio; na reflexão; no rigor e detalhe; e na escrita cuidadosa! E ainda como personalidade marcante, no que diz respeito ao relacionamento humano e à valorização da simplicidade e da comunicação.

E, quem não gosta de ouvir o Professor João Branco contar estórias sobre os professores e colegas/amigos que lhe passaram pela vida! De entre esses colegas amigos, a Professora Maria Antónia Amaral Turkman e os Professores Paulo Gomes, Trevor F. Cox e John Hinde aceitaram prontamente o desafio de participarem ativamente no encontro *Perspectives in Statistics*…, proferindo as comunicações: *Multiple-Choice: a Bayesian perspective*, *Census 2011: new information requirements and major challenges for the National Reform Programme*, *Principal Components, Simple Components and Correlated Components*, e *Overdispersed Generalized Linear Models and Random Effects*, respetivamente.

Por último, não é possível esquecer dois grandes contributos do Professor João Branco para o desenvolvimento da Estatística na comunidade nacional: a compreensão da Estatística para além da Estatística Matemática e a introdução do estudo da Robustez Estatística em Portugal, pela qual é o principal responsável.

Ao Professor Doutor João Branco O NOSSO AGRADECIMENTO!

Em nome das alunas e alunos do Professor João Branco,

Conceição Amado

Dinis, uma referência… na Estatística e seu Ensino e na Universidade

O percurso académico do Prof. Dinis Duarte Ferreira Pestana, para além de inúmeras contribuições pontuais com as mais variadas universidades foi centrado em três polos mais relevantes - a Universidade de Lisboa, a Universidade da Madeira que "fez criar" e a Universidade dos Açores.
Na Faculdade de Ciências da Universidade de Lisboa esteve na criação do departamento de Estatística, Investigação Operacional e Computação (DEIOC), que deu origem em meados da década de 80 ao atual departamento de Estatística e Investigação Operacional (DEIO) e ao qual o Prof. Dinis Pestana presidiu em épocas particularmente difíceis. De 1981 a 1987 dirigiu o CEAUL com competência e empenho e organizou o I Congresso Anual da Sociedade Portuguesa de Estatística.
Orientou um grande número de alunos de Doutoramento, contribuiu para a formação de muitos docentes de outras universidades nomeadamente da Madeira e dos Açores, e para a divulgação da Estatística em Portugal.
O Professor Dinis Pestana para além de possuir profundos conhecimentos de Probabilidade e uma vasta cultura, é também dotado de uma grande sensibilidade.
Tendo sido sua aluna de licenciatura e de doutoramento, e também sua assistente, tive a oportunidade de apreciar a sua capacidade de avaliar, criticando os aspetos negativos mas sempre com uma palavra de apreço pelo empenho e pelo esforço despendido, a sua disponibilidade para ouvir os alunos e para os ajudar, preocupando-se quando sabia de alguém que estava a passar por dificuldades de qualquer ordem e a enorme capacidade de trabalho.
A sua elevada competência pedagógica e o vasto leque de conhecimentos estão bem patentes no seu livro "Probabilidades e Estatística", que já vai na 4ª edição, e é um texto de referência tanto para alunos de licenciatura como de mestrado.
Por tudo o que referi e pelo muito que ficou por dizer, não tenho dúvidas de que o Professor Dinis Pestana foi um professor excecional e que marcou várias gerações de alunos.

Helena Iglésias Pereira
(DEIO, Faculdade de Ciências da Universidade de Lisboa)

O Professor Dinis Pestana foi um grande impulsionador do ensino superior na Região Autónoma da Madeira, em particular através do trabalho que realizou no Centro de Apoio da Faculdade de Ciências da Universidade de Lisboa na Região Autónoma da Madeira e na própria Universidade da Madeira.

No Centro de Apoio, participou na formação de muitos professores do ensino secundário que preencheram o grande número de vagas existentes na altura nas escolas da Região. Foi também no Centro de Apoio que, pela sua mão, foram encaminhados para a Faculdade de Ciências os alunos mais interessados na vertente científica dos cursos. Interessou-se pela orientação da carreira académica dos docentes do Centro de Apoio. A sua forma de trabalhar foi e ainda é estimulante para todos, incluindo para os funcionários não docentes que contaram com a sua orientação e se tornaram colaboradores entusiastas nas funções que, muitos deles, ainda hoje desempenham.

Colaborou com a Universidade da Madeira desde a sua formação. Ajudou esta Universidade a se implantar: orientou teses de doutoramento, orientou dissertações de mestrado, encorajou bons alunos a seguirem a carreira académica e apoiou a lecionação.

Eu fui uma das suas orientadas e sempre me deslumbrava quando me dizia, sem nunca falhar, em que artigos (revista, autores e ano) podia encontrar o que procurava. A sua imensa cultura, científica e não científica, é largamente conhecida. Orientou o doutoramento de outros dois docentes que se encontram atualmente em funções nesta Universidade, que ainda hoje se sentem felizes por poderem contar sempre com o seu apoio e amizade.

As suas qualidades humanas, marcaram a vida de docentes, de alunos e de funcionários e de muitas outras pessoas que, de uma forma ou de outra se cruzaram no seu caminho. Sou muitas vezes interpelada nas ruas do Funchal, por algumas dessas pessoas, interessadas em saber notícias do Professor Dinis Pestana e, naturalmente, deparo-me sempre com um sorriso nas suas caras. É o reflexo da forma como se relaciona com as pessoas com quem trabalha: vê sempre o que de melhor há em nós, quer para nos mostrar que somos capazes de seguir o percurso que nos compete, quer para valorizar o nosso trabalho.

Rita Vasconcelos
(CCCEE, Universidade da Madeira)

Foi com o Professor Dinis Pestana que a Estatística começou a ter alguma expressão como área de investigação na Universidade dos Açores. A sua colaboração com esta instituição, mais concretamente com o Departamento de Matemática, começou no final da década de oitenta do século passado, quando aceitou orientar o Professor José Carlos Rocha nas suas provas de aptidão pedagógica e capacidade científica. Mais tarde viria a orientá-lo nas suas provas de doutoramento e a parceria científica entre os dois continuaria até ao falecimento deste último. Também orientou as minhas provas de mestrado e de doutoramento, e com ele tenho mantido uma estreita colaboração científica até ao presente. Porém, a colaboração do Professor Dinis Pestana com a instituição açoriana no enriquecimento da Estatística não se tem ficado apenas no Departamento de Matemática, extrapolou para o Departamento de Economia e Gestão na altura em que decidiu orientar as provas de mestrado, e mais tarde a co-orientar as provas doutoramento, de Maria Luísa Rocha, filha do seu primeiro "discípulo" açoriano. Mais do que um mentor com qualidades científico-pedagógicas excecionais, tem sido um amigo dedicado a todos nós que temos o privilégio de o conhecer. Os (poucos) "estatísticos" açorianos terão para sempre uma dívida de gratidão para com ele.

Fátima Brilhante
(DM, Universidade dos Açores)

Por ocasião da sua aposentação em Novembro do ano passado, O DEIO, a SPE e o CEAUL, organizaram um jantar com cerca de 115 participantes de todo o país.
Em agradável convívio foram apresentados vários testemunhos de apreço ao homenageado.

Ivette, os Extremos agradecem …

e nós também!

Por Maria Isabel Fraga Alves, *uma sua humilde e eterna aluna*

Ao longo de 2011, de entre os estatísticos portugueses que se aposentaram, surge um nome de referência que foi e é um marco incontornável para muitos dos investigadores, docentes, alunos e profissionais, na área da Estatística em Portugal e além-fronteiras. Refiro-me evidentemente à Professora Doutora Maria Ivette Gomes – a Ivette como carinhosamente é tratada pelos seus pares.

Falar aqui da Ivette representa um enorme desafio só comparável ao de redigir um sintético *abstract* relativo a um *paper* de elevado gabarito. De todas as palavras que lhe possamos dedicar, "MUITO OBRIGADO" é sem dúvida a frase que mais genuinamente traduz o que sentimos por ocasião deste registo de homenagem.

Não podemos deixar de frisar que a alteração do seu estatuto profissional, não alterou a sua rotina diária de trabalhadora incondicional ao serviço da ciência estatística, com a sua sede habitual no nosso DEIO&CEAUL e inúmeros convites além fronteira para desempenhar quer funções de conferencista convidada, quer para funções editoriais associadas a diversas revista científicas de renome internacional, tendo já um palmarés internacional para a *Revstat*, publicada pelo INE. De facto, se não fossem alguns dos eventos organizados com esse pretexto, nunca seria percetível que a nossa Ivette já é afinal "aposentada".

Disso mesmo foi exemplo o elevado número de pessoas vindas de todo o país que acorreram ao jantar de 12 de Janeiro último, como homenagem à Ivette em simultâneo com aquele que é igualmente um nome sonante para toda a comunidade estatística portuguesa – o Professor Dinis Pestana, por ocasião da sua recente aposentação. Realmente, não é possível falar da Ivette sem falar do Dinis. O lugar comum de que *ATRÁS DE UMA GRANDE MULHER ESTÁ UM GRANDE HOMEM* assume aqui um particular significado. As qualidades de ambos quer de ordem científica, quer de índole docente, foram igualmente relembrados numa organização espontânea de antigos alunos, que acorreram em massa a um jantar convívio a 18 de Fevereiro. Em 11 de Julho foi a vez da Ivette ser surpreendida em Faro, por uma retrospetiva da sua vida, organizada no seio da 5th WSMC. E apostamos que muitos mais serão os eventos vindouros com propósitos análogos.

Partindo de uma formação sólida obtida em 1970 no Bacharelato em Matemática Pura da Faculdade de Ciências, o percurso da Ivette rapidamente enveredou pela aventura na área da Matemática Aplicada,

tendo rumado para terras de sua Majestade e obtido o PhD em Statistics na Universidade de Sheffield em 1978. Após concluir Agregação em Matemática Aplicada em 1982, veio a ocupar o lugar de Professora Catedrática do DEIO em 1985.

Este abreviadíssimo Curriculum ilustra bem a forma como a Ivette se posicionou na sua carreira académica. Nunca descurando a precisão teórica de todo um *background* probabilista inerente à sua tão cara Teoria de Valores Extremos, a Ivette foi mais além do que uma simples "purista": o seu *modus operandi* envolve todo um trabalho árduo no campo mais aplicado da Estatística de Extremos, através de técnicas computacionais, atuantes de forma anímica para o sucesso dos bons resultados de inferência estatística, inicialmente em meio paramétrico, e posteriormente numa abordagem semiparamétrica. Realmente, para todos nós que, de uma forma ou de outra, têm tido o privilégio de com ela privar, é manifesta a elegância com que combina a mestria de controlar os modelos extremais com a destreza computacional: pode haver outros Estatísticos que aliem esta combinação na carteira das suas qualidades, mas é sem dúvida uma qualidade invulgar. A este propósito relembramos aqui Albert Einstein, que se tivesse tido o privilégio de conhecer a Ivette lhe dedicaria com certeza este pensamento:

Computers are incredibly fast, accurate, and stupid.
Human beings are incredibly slow, inaccurate, and brilliant.
Together they are powerful beyond imagination.

Albert Einstein *(1879 - 1955)*

Notícias

• XIX Congresso SPE

Estendendo as redes da Estatística no mar da Nazaré

Foi a característica vila da Nazaré que este ano acolheu o XIX Congresso da Sociedade Portuguesa de Estatística, mais precisamente a sobranceira terra de pescadores da Pederneira, no recentíssimo Hotel Miramar Sul, inteiramente reservado ao congresso!

A escolha foi feita pela comissão organizadora, partilhada entre a Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e o Instituto Superior Técnico da Universidade Técnica de Lisboa, constituída pelos colegas Alexandra Seco, António Pacheco, Helena Ribeiro, Mª Rosário de Oliveira, Miguel Felgueiras e Rui Santos, a quem endereçamos os nossos calorosos cumprimentos pelo bom trabalho realizado.

O caminho de Lisboa não era longo e, ignorando o GPS e fazendo algumas manobras apertadas, foi possível chegar à hora do mini curso que habitualmente abre os trabalhos dos congressos. Este ano o tema versado foi a *Análise de dados Longitudinais*, tendo sido apresentado pela colega Mª Salomé Cabral, que contou com a colaboração da colega Mª Helena Gonçalves na sua preparação.

A exposição foi excelente e muito útil, não sobraram efeitos aleatórios para estimar, tudo ficou perfeitamente ajustado! Foram muitos os interessados neste assunto que certamente ficaram muito satisfeitos.

Ao fim da tarde aconteceu a sessão de abertura do congresso. Contou com a intervenção do presidente da SPE e do Congresso, colega Carlos Braumann, que nostalgicamente lembrou que estava de partida destas já longas e trabalhosas funções, do presidente da câmara da Nazaré, de representantes da Escola Superior de Tecnologia e Gestão, do Instituto Politécnico de Leiria e do Instituto Superior Técnico da Universidade Técnica de Lisboa, e de membros representantes da comissão organizadora.

Seguiu-se a 1ª sessão plenária do congresso. Ao introduzi-la, referir que neste congresso houve oportunidade de homenagear 3 ex-presidentes da SPE que se aposentaram no decorrer no passado ano letivo - os colegas Mª Ivette Gomes (1990-1994), João Branco (1994-2000) e Fernando Rosado (2000-2006).

A 1ª Plenária começou com uma breve apresentação em jeito de homenagem feita pela colega Manuela Souto de Miranda sobre a vida e obra do ex-presidente João Branco, na qualidade de sua primeira orientanda de doutoramento, e pela oferta em nome da SPE de um presente e um lindo ramo de flores.

Seguiu-se a comunicação apresentada pelo homenageado em que este se empenhou em resolver uma "disputa" antiga entre Mendel (1822-1884) e Fisher (1890-1962), sempre com o seu estilo vivo e bem disposto.

A habitual receção de boas vindas foi depois, na biblioteca municipal da Nazaré, um edifício recente em que se destacam as modernas e amplas áreas, diferenciadas por utilizações específicas a públicos alvo de diferentes faixas etárias. O guia da visita foi o próprio Sr. Presidente da Câmara tendo esta terminado com o habitual Porto de Honra (também espumante rosé, sumos ou cocktail de fruta).

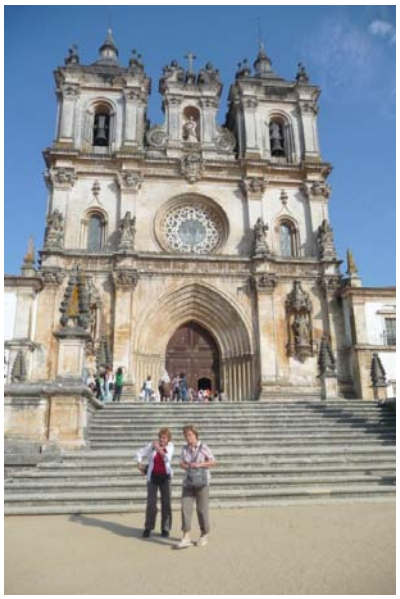
Novo dia, os trabalhos começaram às 9h, com 4 sessões em paralelo. Os congressistas eram cerca de 200, as comunicações orais chegaram quase às 90 e os posters cerca de 50. Os temas mais versados foram nas áreas de Análise Multivariada, Bioestatística, Processos Estocásticos, Extremos, Inferência Estatística e Aplicações. Houve o problema habitual da escolha de quais as sessões ir e que transições fazer (o que foi possível devido a uma longa experiência dos presidentes que permitiu uma boa sincronização dos horários das sessões). No entanto uma organização destas bem planeada permitiu que na mesma sala se encontrassem geralmente as sessões pretendidas.

As sessões de posters decorreram no átrio do hotel, durante as pausas para café, acabando por não ficarem muito bem localizadas relativamente ao local do próprio café (no pátio exterior do hotel, em frente ao mar). Inicialmente era para ser tudo no átrio, mas como estava um belíssimo tempo houve esta alteração de planos que prejudicaram um pouco a afluência aos posters.

No final da manhã a 2ª plenária, que era para ser proferida por Wolfgang Schmid, da Europa-Universität Viadrina (Alemanha), foi substituída por uma sessão de informação sobre o SAS para o ensino superior, já que o orador referido ficou retido no seu país por motivo de acidente.

O aguardado passeio do programa social foi na tarde deste dia. Existiam duas opções, um Peddy-Paper pela lindíssima vila de Óbidos ou uma visita guiada aos Mosteiros Cistercienses de Alcobaça e Cós para purificação do pecado da gula que mais tarde se haveria de cometer. O Peddy-Paper contou com um menor número de participantes e foi talvez um pouco intenso e exigente. Quem já conhecia a vila usufruiu e desfrutou mais desta atividade, mas ainda assim chegou ao final sem fôlego, agradecendo a ginginha oferecida para o recuperar. Terá sido por causa da dureza da prova que uma das equipas atrasou consideravelmente a partida de volta para a Nazaré ou será que a beleza e os recantos de Óbidos os enfeitiçou de tal forma que os impediu de acompanhar o ritmo exigido pela competição?

O passeio alternativo iniciou-se pela visita ao Mosteiro de Cós, situado no Lugar de Cós, concelho de Alcobaça. Este Mosteiro é um dos mais importantes Mosteiros femininos da Ordem de Cister mas apenas a sua Igreja e sacristia se encontram recuperadas. Foi, no entanto, um privilégio poder visitá-lo já que fomos acompanhados pela Licenciada e Mestre em História Ana Margarida Martinho, com uma tese de Mestrado exatamente sobre o Mosteiro de Cós. Sentados nos bancos da sua muito bonita Igreja, em frente a um magnífico altar de talha tivemos uma verdadeira lição de História, sobre os Mosteiros de Cister desta região e a sua relação com a História de Portugal, praticamente desde a sua Fundação. Pudemos ainda apreciar o belo cadeiral, na parte da Igreja reservada às Monjas, bem como os painéis de azulejos da sacristia.



Prosseguimos depois para Alcobaça, onde visitámos o muito famoso e muito bem conservado Mosteiro. Tivemos ocasião de apreciar a arquitetura dos diversos espaços, de que salientamos a Igreja, de estilo gótico e de grandiosas dimensões, os Claustros, em especial o de D. Dinis e a cozinha que ilustra de modo interessantíssimo diversos aspectos da vida neste Mosteiro.

No fim da visita esperava-nos, no Claustro D. Afonso VI do Mosteiro de Alcobaça, uma excelente prova de doces conventuais, acompanhada de uma Ginja de Alcobaça. Assim, todos terminámos a tarde, deliciados de corpo e alma.

De regresso à Nazaré foi ainda tempo para um mixer da jSPE em que foram discutidos alguns aspectos relacionados com o funcionamento da mesma, nomeadamente a eleição dos seus órgãos diretivos, seguido de uma muito apreciada e reconfortante refeição num terraço do hotel onde a sopa de peixe brilhou e a noite amena convidou à bem disposta cavaqueira.

O terceiro dia é sempre aquele em que, estando os congressistas já familiarizados com o cenário do congresso, se pode apreciar uma maior quantidade de trabalho em termos de número de sessões e posters. A segunda sessão plenária efetiva foi proferida neste dia antes do almoço pela colega Graciela Boente, da Universidade de Buenos Aires (Argentina), que com grande simplicidade e simpatia nos guiou através de *Some recent results for functional data analysis*, quase nos fazendo acreditar que também nós conseguimos trabalhar facilmente com dados funcionais.

A terceira sessão plenária efetiva surgiu a meio da tarde deste 3º dia começando por uma intervenção da colega Manuela Neves sobre a vida e obra do Prof. Fernando Rosado, prestando-lhe a devida homenagem, e presenteando-o em nome da SPE com um ramo de flores e um decantador de cristal. Seguidamente o ex-presidente da SPE Fernando Rosado presenteou-nos a nós com *A Força dos Menores*, em que de uma forma muito divertida ele nos fez refletir sobre a Estatística e o seu papel no desenvolvimento e investigação científicos.

Após o fim de todos os trabalhos do dia todos nos fomos aperaltar para a tradicional foto de grupo, feita com o mar de fundo e tendo um pequeno fotógrafo como incentivador de sorrisos, e também para o ponto alto do congresso, o jantar. Viajámos até à Batalha, para a encantadora Quinta do Fidalgo, propriedade do restaurante Tromba Rija (que afinal é uma tromba de porco e não de elefante, para os menos informados). Pudemos tomar um aperitivo ao som de um duo de sopro e teclas.

Depois sentámo-nos numas acolhedoras mesas de 8 pessoas para nos regalarmos com uma enorme variedade de entradas frias em buffet, regadas de vinho branco ou tinto, água ou sumo. Pratos quentes também havia, mas já era preciso ter coragem para lá chegar. As sobremesas de fruta variada, doce de bolacha, leite creme, mousse de chocolate, entre outros, coroaram o jantar de glória. Quando pensávamos que já não conseguíamos degustar nem mais uma migalha trouxeram-nos ainda avelãs, nozes, a famosa amêndoa amarga, aguardente velha e vinho do Porto.

O colega João Branco pediu a palavra durante o jantar para informar todos os sócios da SPE que, como resultado do processo eleitoral para os órgãos diretivos da SPE que tinha estado a correr durante esse dia, tinha sido eleita a lista concorrente. Apresentou os seus cumprimentos aos membros da referida lista, parabenizou-os pela apresentação da mesma lista e desejou-lhe os maiores sucessos para a tarefa que os aguarda. O Prof. Braumann secundou-o nas suas palavras.

O quarto e último dia é sempre o que desafia mais as capacidades dos congressistas, por começar após uma noite curta de sono. No entanto às 9h começaram novamente 4 sessões paralelas de comunicações seguidas pela última sessão de posters.

A última sessão plenária foi aberta pela intervenção da colega Isabel Fraga Alves que, como sua antiga orientanda de doutoramento, homenageou a ex-presidente da SPE Mª Ivette Gomes, detalhando vários aspectos da sua vida e legado científico de grande mérito. Em nome da SPE foi ainda oferecido o presente e as flores que se impunham. A homenageada apresentou de seguida a *importância de métodos de re-amostragem em Estatística de Extremos*, visivelmente emocionada, mas sempre com a energia e entusiasmo que todos lhe conhecemos.

Seguiu-se a já habitual entrega dos prémios estatístico júnior pelos colegas Carlos Braumann, Luísa Canto e Castro, Maria Eugénia Graça Martins e Russel Alpizar-Jara.

Estava-se mesmo na reta final, só faltava a sessão de encerramento. Esta iniciou com a entrega do prémio e apresentação do "Logótipo da jSPE" pelo colega Paulo Canas Rodrigues. Das 161 propostas a concurso, o júri selecionou o logótipo proposto pelo concorrente Ricardo Alves. Seguiram-se os agradecimentos à comissão organizadora do XIX congresso da SPE. Por fim, passou-se o testemunho ao colega A. Pedro Duarte Silva, presidente da comissão organizadora do XX congresso da SPE, da Universidade Católica do Porto. Assim, para o ano que vem, voltamos à invicta e fazemos desde já votos dos maiores sucessos para essa organização!

Resta voltar a agradecer à organização do XIX Congresso da SPE o seu trabalho, que nunca é de pequena monta, e desejar que possa continuar florescente a produção estatística nacional com a força e entusiasmo que tivemos oportunidade de assistir neste evento.

Isabel Natário (FCT/UNL)
Mª Teresa Arede (FEUP)
Frederico Caeiro (FCT/UNL)

P. S. - Agradecimento aos repórteres exteriores Júlia Teles e Vanda Lourenço.

XIX Congresso Anual SPE

• Eleição dos Órgãos Administrativos da SPE para o triénio 2012-14

A realização das eleições dos Órgãos Administrativos da SPE para o triénio 2012-2014 teve lugar na Nazaré, no dia 30 de Setembro de 2011, durante o XIX Congresso anual da SPE.
Apresentou a sua candidatura a seguinte lista única:

Mesa da Assembleia Geral

Presidente: Jorge Cadima (Univ. Téc. Lisboa - Instituto Superior de Agronomia)

Primeiro Vogal: Patrícia Bermudez (Univ. Lisboa - Faculdade de Ciências)

Segundo Vogal: Isabel Pereira (Univ. Aveiro)

Direcção

Presidente: Carlos Daniel Paulino (Univ. Téc. Lisboa - Instituto Superior Técnico)

Vice-Presidente: Pedro Oliveira (Univ. Porto - ICB Abel Salazar)

Secretário: Manuela Neves (Univ. Téc. Lisboa - Instituto Superior de Agronomia)

Secretário-Adjunto: Paulo Soares (Univ. Téc. Lisboa - Instituto Superior Técnico)

Tesoureiro: Marília Antunes (Univ. Lisboa - Faculdade de Ciências)

Conselho Fiscal

Presidente: Carlos Marcelo (Instituto Nacional de Estatística)

Secretário: Giovani Silva (Univ. Téc. Lisboa - Instituto Superior Técnico)

Relator: Cecília Azevedo (Univ. Minho - Escola de Ciências)

A Comissão eleitoral presidida pela Professora Conceição Amado apurou os seguintes resultados:

	Direcção	Mesa Assembleia Geral	Conselho Fiscal
Sim	100	108	110
Não	5	1	1
Brancos	9	6	4
Nulos	3	2	2
Total	117	117	117

Felicitamos todos os sócios componentes da lista candidata pelo excelente resultado alcançado. Sabendo que o trabalho dos novos órgãos administrativos é vital para a continuidade da SPE na prossecução dos seus objectivos, pedimos a todos os seus elementos o maior empenho para que a SPE venha a ser contemplada com novos e profícuos desenvolvimentos. A tarefa é grande para os órgãos administrativos, mas pode ser muito compensadora para todos. E a nossa Sociedade merece o esforço.

João Branco
Presidente da Mesa da Assembleia Geral

• jSPE - Secção de Jovens Estatísticos da SPE

A jSPE, Secção de Jovens Estatísticos da Sociedade Portuguesa de Estatística, tem por missão promover, cultivar e incentivar um intercâmbio de informação e conhecimento entre os Jovens Estatísticos. Neste sentido, propomos preparar, organizar e coordenar: (i) encontros científicos para jovens estatísticos; (ii) sessões e eventos sociais nos congressos anuais da SPE; e (iii) sessões de convívio, workshops científicos e de desenvolvimento de carreira. A jSPE conhece a importância da mobilização e participação da sociedade portuguesa no exterior e, como tal, também estabelecerá contacto com outras Secções de Jovens Estatísticos internacionais.

Podem ser membros da jSPE os Sócios da SPE que verifiquem pelo menos uma das seguintes condições:

1. Tenham obtido nos últimos 5 anos um diploma de licenciatura, mestrado ou doutoramento na área de Probabilidades e Estatística ou áreas afins;
2. Estejam inscritos num destes ciclos de estudos com vista à sua obtenção;
3. Exerçam a profissão de estatístico, há menos de 5 anos, independentemente da sua formação base.

Um dos aspetos fundamentais destacados para pertencer à jSPE é a obrigatoriedade de ser "jovem". No entanto, ressalva-se que "jovem" está associado, neste caso, ao interesse na estatística, isto é, apenas é necessário que o envolvimento científico, profissional ou de interesse na área seja recente.

Salienta-se que ser membro da jSPE não acarreta custos adicionais, pois não é cobrado qualquer valor além da cota da SPE.

Durante o XIX Congresso Anual da SPE, mais propriamente no dia 29 de Setembro de 2011, teve lugar a primeira Assembleia Geral da jSPE seguida de um *Mixer*. A Assembleia Geral consistiu na eleição da primeira direção da jSPE: Paulo Canas Rodrigues (Presidente), Rui Cruz (Vice-Presidente) e Rosa Oliveira (Vice-Presidente); e na apresentação do relatório de atividades da jSPE para 2012. Entre as atividades planeadas, destacou-se o primeiro congresso da jSPE que será realizado em conjunto com o y-BIS, a secção de young Business and Industrial Statisticians da International Society for Business and Industrial Statisticians. Este evento irá decorrer nas instalações da FCT-UNL no Campus da Caparica, entre os dias 23 e 26 de Julho de 2012. A Comissão Organizadora Local irá ser composta por Filipe Marques, Paulo Canas Rodrigues, Frederico Caeiro, Vanda Lourenço e Luís Ramos. A conferência irá contar com a participação de *Keynote Speakers* de reconhecido mérito internacional, que partilharão com os congressistas as suas experiências e sabedoria. Terão lugar, ainda, *Invited Sessions*, *Contributed Talks*, e *Contributed Posters*. Até ao momento obtivemos duas confirmações de Keynote Speakers, a apresentar:

- Narayanaswamy Balakrishnan, reconhecido Professor de Estatística na McMaster University, Hamilton, Ontario, Canada. Bala é também Executive Editor do Journal of Statistical Planning and Inference, Editor–in–Chief dos jornais Communications in Statistics–Theory and Methods, e Communications in Statistics–Simulation and Computation; e Editor–in–Chief da Encyclopedia of Statistical Sciences.
- Ronald L. Wasserstein é o Executive Director da American Statistical Association. Ron é também Presidente da Kappa Mu Epsilon National Mathematics Honor Society, 2009-2013 (President–elect, 2005–2009) e faz parte do Board of Directors da MATHCOUNT. Ron fez e continua a fazer parte de inúmeros comités com o intuito de promover a estatística.

Serão dedicados, a este congresso, dois *Special Issues* de Jornais Científicos Internacionais, sendo que um está confirmado e o outro está em negociações.

Mais informações sobre esta e outras atividades seguirão em breve para todos os membros da SPE pela sua mailing list. Como tal, solicitamos que se juntem a nós o mais breve possível, divulgando a nossa secção a fim de fortalecermos o grupo e as atividades realizadas, dando-lhe visibilidade nacional e internacional.

A direção da jSPE incita os seus membros a não faltar a este encontro e reservar desde já os dias 23 a 26 de Julho de 2012 nas vossas agendas!

Após a Assembleia Geral da jSPE, aberta a todos os membros da SPE e que contou com cerca de 45 participantes, teve lugar o *Mixer* da jSPE que juntou cerca de 80 participantes. Este *Mixer* teve como intuito promover, durante um ligeiro jantar volante, a confraternização entre membros da jSPE. Os restantes sócios da SPE não foram esquecidos e, como tal, foram convidados a juntar-se aos mais "jovens", uma vez que a recém-criada secção considera que os conhecimentos e experiência dos menos "jovens" aliados ao dinamismo, conhecimentos e experiência dos mais "jovens" trarão, com toda a certeza, mais-valias no enriquecimento pessoal e científico da comunidade estatística portuguesa e estrangeira.

Por fim, gostávamos de deixar um agradecimento especial aos patrocinadores do *Mixer* da jSPE: Comissão Organizadora do XIX Congresso Anual da SPE, SAS, Quinta do Fidalgo–Restaurante Tromba Rija.

Para os que tiveram a possibilidade de estar presente, esperamos que tenham gostado desta atividade e desejamos que nos encontremos novamente em breve. Aos que não tiveram a oportunidade de participar no primeiro *Mixer* da jSPE, estão desde já convidados a estar presente, em Setembro de 2012, nas atividades organizadas pela jSPE no XX Congresso Anual da SPE, e em todas as outras atividades organizadas pela jSPE.

As fotos do *Mixer2011* estão disponíveis no nosso grupo no Facebook!

Para finalizar apelamos aos Jovens Estatísticos que se juntem à jSPE, sendo que para o efeito, bastará que nos enviem um email a atestar a vossa vontade. Aos Estatísticos "Seniores", bem como a "Seniores" de áreas afins, pedimos que divulguem a informação e que incentivem os vossos alunos a integrarem e contribuírem para este projeto. Da nossa parte, tudo faremos para proporcionar a inspiração para a renovação do interesse estatístico que todos ambicionamos nos Jovens da Comunidade Estatística Portuguesa.

Informação mais completa sobre a jSPE pode ser encontrada na nossa página oficial: https://sites.google.com/site/jspespe/, ou no nosso grupo no Facebook, sendo que, para o efeito, basta pesquisar por jSPE.

Pela Direção da jSPE,
Paulo Canas Rodrigues
Rui Cruz
Rosa Oliveira

• Glossário Inglês-Português de Estatística - Nova edição

Divulgou-se recentemente, por via eletró(ô)nica no seio da Sociedade Portuguesa de Estatística (SPE) e Associação Brasileira de Estatística (ABE), a notícia da saída da 2ª edição do **Glossário Inglês-Português de Estatística** (GLIPE) e sua disponibilização no endereço http://glossario.spestatistica.pt/. Atendendo à relevância cultural e científica desta iniciativa de desenvolvimento e promoção da língua portuguesa no nosso campo de intervenção, tem-se feito difusão desta informação por pertinentes instituições e setores da comunidade lusófona. *A fortiori* entendemos que se justifica divulgar neste boletim alguns detalhes sobre o historial e traços distintivos desta nova edição do GLIPE, que apresenta múltiplos aperfeiçoamentos no seu conteúdo e formato, através da reprodução essencial do Prefácio que a acompanha e que se apresenta já de seguida em itálico.

No fim de 2006 constituiu-se no seio da SPE a Comissão Especializada de Nomenclatura Estatística (CENE), com o objetivo de organizar um glossário inglês-português de Estatística que pudesse servir de guia normativo na expressão oral e escrita em língua portuguesa do vocabulário estatístico.

Esta comissão nunca hesitou em compreender a relevância deste trabalho como contributo exemplar para assegurar ao nosso idioma um estatuto de língua de cultura científica, não obstante as conce(p)ções e manifestações adversas provindas de quadrantes diversos que jamais deixaram de ocorrer. Tomando em consideração a projeção mundial da língua portuguesa – considerada atualmente como a terceira língua europeia de comunicação universal, segundo a União Europeia --, e a importância daquele objetivo para toda a comunidade lusófona, a CENE conseguiu assegurar a colaboração de uma comissão brasileira, com o apoio da SPE e ABE, para numa primeira etapa se realizar a tradução organizada, consistente e o mais extensa possível do glossário de termos estatísticos do International Statistical Institute (ISI).

O trabalho enérgico e entusiástico desenvolvido ao longo de alguns (poucos) meses conduziu à produção de uma primeira edição do glossário inglês-português de Estatística, que haveria de granjear um célere reconhecimento internacional com a sua inserção no sítio oficial do ISI em julho de 2007. As limitações, decorrentes da política do ISI, à introdução a este glossário de alterações que se justificavam levou a comissão conjunta luso-brasileira a iniciar autonomamente uma segunda etapa do projeto de criação de uma fonte terminológica credível de vocabulário estatístico para o mundo lusófono, visando produzir uma nova versão que não sofresse das deficiências que apontara ao glossário do ISI.

Após um longo, vicissitudinário mas produtivo período de estudos, consulta de múltiplas fontes [1], debates, reflexões e contactos, deu-se por finda a fase conducente à segunda edição do glossário pretendido que ora se apresenta. Esta edição difere da anterior no sentido da ampliação por atualização decorrente do desenvolvimento científico de certas áreas, introdução de notas explicativas em entradas objetivando, designadamente, clarificar as opções pelas traduções incluídas e mencionar as preferências de uso por comunidade e remoção de termos de tradução óbvia e repetitiva. Além

[1] *Destaca-se, a título de exemplo, o Vocabulário Estatístico Inglês-Português editado em brochura pelo INE (1969) e o Vocabulário Brasileiro de Estatística de Milton Rodrigues publicado em boletim pela Universidade de São Paulo (1956) e pelo IBGE.*

disso, atendendo à atual fase de aplicação do Acordo Ortográfico de 1990 em Portugal e no Brasil, optou-se por (tentar) seguir as normas nele estabelecidas como este próprio prefácio exemplifica.

Também a forma de divulgação do resultado deste sério trabalho, questão nunca considerada como de somenos importância, sofreu uma substancial mudança. O uso da atrativa e eficiente plataforma informática em que agora se apresenta permite, em especial:

- *Simplificar as consultas de interesse;*
- *Agilizar atualizações sempre inevitáveis em produtos deste género;*
- *Facilitar a obtenção de cópias impressas do glossário;*
- *Ampliar a informação transmitida com material suplementar como a lista de acró(ô)nimos ingleses de potencial utilidade para tradutores que, em particular, não se enquadram dentro das fronteiras da comunidade estatística;*
- *Facultar a interação dos utilizadores com os representantes do órgão autoral para fins de apresentação de sugestões de inclusão de novos verbetes e de outras alterações.*

Queremos agradecer vivamente a várias Direções da SPE e ABE pelo interesse e apoio manifestados ao trabalho desenvolvido pela Comissão que tive o prazer e o privilégio de coordenar e, em nome desta, expressar a nossa incontida satisfação pela proficiência demonstrada pelo nosso colega Paulo Soares na construção e implementação da plataforma informática que divulga o atual Glossário Inglês-Português de Estatística.

Desejamos ainda transmitir à comunidade estatística do mundo lusófono a nossa convicção de que quanto mais significativa e positiva for a sua participação na atualização e ampliação do glossário corrente, tanto mais eficaz na sua missão será o produto deste projeto coletivo. Missão esta que é a de constituir um instrumento de consulta, indubitavelmente útil para todos os interessados -- que não se esgotam em nós, estatísticos e utentes/usuários de Estatística --, e de desenvolvimento em moldes rigorosos e harmonizados de uma comunicação científica em língua portuguesa, resistindo às pressões e tentações para a secundarização desta em tempos de asfixiante globalização.

Carlos Daniel Paulino

• Encontro Europeu de Estatísticos - *17th European Young Statisticians Meeting*

A 17.ª edição do European Young Statisticians Meeting decorreu nos dias 5–9 de Setembro na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa. Esta conferência bianual é organizada sob a alçada do comité regional da Bernoulli Society e destina-se à promoção científica dos jovens estatísticos europeus. O número de participantes é limitado a dois por cada país Europeu - sendo os participantes eleitos com base no mérito científico. Os participantes nacionais foram: Filipe Marques (Universidade Nova de Lisboa) e Luis Pereira (Universidade do Algarve). A conferência teve como oradores convidados o Professor Wolfgang Polasek, o Prof. Carlos Agra Coelho, a Prof. Ivette Gomes e o Prof. Kamil Feridun Turkman. Os chairs do comité organizador, Paulo Canas Rodrigues e Miguel de Carvalho, garantiram ainda um *special issue* associado à conferência, do qual serão editores, e que será publicado na *Communications in Statistics—Theory and Methods*. A próxima edição desta conferência irá decorrer em Agosto de 2013 na Croácia. O Relatório Final do 17th European Young Statisticians Meeting pode ser encontrado em http://isi-web.org/images/news/2011-BS-Sept-Report17thEYSM.pdf.

Paulo C. Rodrigues

• Escola Portuguesa de Extremos "Coloniza" Brasil

A Escola Portuguesa de Extremos esteve representada recentemente no Brasil na 56.ª conferência da RBRA - Secção Brasileira da *International Biometrical Society* - que decorreu de 25 a 29 de Julho em Maringá, no estado do Paraná. A representação deu-se sob a forma do curso *Modelling Statistics of Extremes* que foi lecionado por Miguel de Carvalho. O curso foi baseado em materiais recentemente escritos conjuntamente com Stuart Coles e Anthony Davison e que farão parte de uma reedição do livro *An Introduction to Statistical Modelling of Extreme Values*, a ser editado pela Springer.

Miguel de Carvalho

• Paula Brito - Presidente IASC

É com grande satisfação que vos venho informar que a Profª Paula Brito, Vice-Presidente da SPE, foi eleita Presidente do IASC (International Association for Statistical Computing), uma das associações internacionais integrantes do International Statistical Institute (ISI). Para já servirá no "Council" como Presidente eleita, passando a Presidente em 2013.

Está a comunidade estatística nacional de parabéns por esta relevante distinção a um dos seus ilustres membros que tão bem a tem servido. Em nome da SPE, venho felicitar a Profª Paula Brito e desejar-lhe as maiores felicidades para o exercício das suas funções, manifestando a disponibilidade da SPE (que é membro do ISI) para reforçar a cooperação com o IASC.

Carlos Braumann

• Prémio da *American Statistical Association*

Miguel de Carvalho foi condecorado com um prémio da *American Statistical Association* (ASA) para jovens investigadores. O prémio foi atribuído pela secção de risco da ASA e entregue no *Joint Statistical Meeting* 2011.
O trabalho vencedor com o título "Spectral Density Ratio Models for Multivariate Extremes" está disponível como pré-print do Centro de Matemática e Aplicações da Universidade Nova de Lisboa.

Miguel de Carvalho é doutorado em Matemática com especialização em Estatística pela Universidade Nova de Lisboa e é atualmente Pós-doc no Swiss Federal Institute of Technology, École Polytechnique Fédérale de Lausanne.

FR

• Prémio SPE 2011

O júri do Prémio SPE 2011, foi constituído pelos Professores João Branco (Presidente), Luísa Canto e Castro Loura e Daniel Paulino.
Foram recebidos 3 trabalhos candidatos ao Prémio SPE 2011. Verificando-se que um dos autores de um dos trabalhos excedia o limite de idade fixado no Regulamento do Prémio, o júri procedeu apenas à apreciação dos dois trabalhos admitidos a concurso.
Embora reconhecendo o valor dos resultados apresentados nesses trabalhos mas considerando que qualquer um deles necessitaria de uma revisão aprofundada, o júri decidiu não atribuir o Prémio SPE 2011.

João Branco
Presidente do Júri

• A REVSTAT está de parabéns!

A REVSTAT está na realidade de parabéns!
Foi inicialmente aceite para ser coberta pelos produtos e serviços da Thomson Reuters desde o Volume 5:1, de 2007, o primeiro dos 3 volumes que enviei para apreciação em 2007.
E em 2010 conseguiu mesmo um factor de impacto de 0.733 no ISI Web of Knowledge.

Ivette Gomes

• Prémios "Estatístico Júnior 2011"

A atribuição de prémios "Estatístico Júnior 2011" é promovida pela Sociedade Portuguesa de Estatística, com o apoio da Porto Editora, e tem como objetivo estimular e desenvolver o interesse dos alunos dos ensinos básico e secundário e cursos EFA/CEF, pelas áreas da Probabilidade e Estatística. Neste ano foram recebidas candidaturas nas 3 categorias: Ensino Básico, Ensino Secundário e Cursos CEF.

A cerimónia de entrega dos Prémios Estatístico Júnior 2011, conforme estipulado no Regulamento, decorreu na Sessão de Encerramento do XIX Congresso Anual da Sociedade Portuguesa de Estatística, no dia 1 de Outubro de 2011, nas instalações do Hotel Miramar Sul, na Nazaré.

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.

No final deste Boletim são apresentados os premiados.

A Direção

• 30.º Aniversário do DEIO - FCUL

No passado dia 25 de Setembro celebrou-se o 30º aniversário do Departamento de Estatística e Investigação Operacional (DEIO). Foi com grande alegria que toda a comunidade DEIO viu o seu Departamento atingir esta respeitável idade. Ao longo destes 30 anos, o DEIO foi pioneiro na formação de muitos dos atuais especialistas em Estatística e em Investigação Operacional e sente ter dado um forte contributo à sociedade, no contexto das suas competências.

Por ocasião desta data, o DEIO organizou uma festa no dia 26 de Setembro que contou com um bolo de aniversário, e na qual estiveram presentes docentes e alunos do Departamento.

João Telhada

Modelos com fragilidade: aplicação à modelação da heterogeneidade não observada

Cristina S. Rocha, *cmrocha@fc.ul.pt*

DEIO – Faculdade de Ciências da Universidade de Lisboa, CEAUL

1. Introdução

Na análise de dados de sobrevivência, quando é utilizado um modelo de regressão, é usual admitir a existência de homogeneidade entre os indivíduos que apresentam valores comuns das covariáveis observadas. No entanto, esta hipótese é, com frequência, pouco realista, dada a impossibilidade prática de registar todos os fatores de risco/prognóstico relevantes. Com efeito, no decorrer de um estudo clínico, é frequente constatar que os indivíduos diferem entre si na evolução natural de determinada doença, na sua reação a um mesmo tratamento ou no modo como são influenciados por vários fatores, apesar de constituírem um grupo homogéneo relativamente às covariáveis consideradas. Esta situação reflete a existência de uma heterogeneidade individual, não observada, mas que é extremamente importante e deve ser tomada em consideração na interpretação dos resultados obtidos.
Admitimos, portanto, a existência de covariáveis que não foram incluídas no modelo porque não dispomos de informação acerca dos seus valores individuais ou mesmo porque desconhecemos a sua existência. De facto, podemos argumentar que há sempre um grande número de variáveis que, caso pudessem ser medidas, dariam informação suficiente para explicar as diferenças individuais. Por outro lado, a heterogeneidade não observada pode também ser devida à existência na população de indivíduos que, na realidade, não são suscetíveis ao acontecimento de interesse.

A necessidade de desenvolver modelos de sobrevivência adequados às situações acima descritas levou ao aparecimento de modelos de heterogeneidade não observada, designados por modelos com fragilidade. O termo fragilidade foi introduzido por Vaupel *et al.* (1979) para designar uma variável não observada que descreve fatores de risco, desconhecidos ou que não se podem medir, não incluídos no modelo. Esta designação pode ser justificada pelo facto de que valores elevados desta variável correspondem a uma diminuição do tempo de vida do indivíduo, ou seja, a um aumento da função de risco em todo o intervalo de tempo considerado.
Os modelos univariados que iremos descrever, embora sejam representações bastante simplificadas da forma como a heterogeneidade pode atuar, contribuem de forma significativa para uma compreensão sistemática deste problema. Com efeito, a heterogeneidade pode explicar alguns resultados inesperados ou fornecer uma explicação alternativa em algumas situações como, por exemplo, quando se observam funções de risco não proporcionais ou decrescentes.

2. O modelo multiplicativo

Com o objetivo de desenvolver métodos de tabelas de mortalidade para populações cujos membros diferiam na sua suscetibilidade geral às causas de morte, Vaupel *et al.* (1979) propuseram um modelo multiplicativo em que a função de risco no instante t para um indivíduo com fragilidade $W = w$ é

$$\mu(t \mid w) = w\lambda(t) \qquad (1)$$

onde W é uma variável aleatória (v.a.) não negativa e $\lambda(t)$ é uma função do tempo comum a todos os indivíduos e independente de W. Vaupel *et al.* (1979) designam $\lambda(t)$ por força de mortalidade padrão, i.e., correspondente a um indivíduo com fragilidade $w=1$. Para modelar a heterogeneidade existente na população em estudo admite-se que a variável fragilidade segue uma determinada distribuição. É habitual supor que $E(W)=1$. Aalen (1988) refere que, deste modo, $\lambda(t)$ pode ser encarada como uma função de risco individual média, visto que, por (1), $\lambda(t)$ é nesse caso o valor esperado da função de risco individual $\mu(t \mid w)$.

Consideremos o modelo em que a função de risco individual é dada por (1). Então, a função de sobrevivência (f.s.) condicional de T dado $W=w$ é $S(t|w)=\exp[-w\Lambda(t)]$ onde $\Lambda(t)=\int_0^t \lambda(u)\,du$ é a função de risco cumulativa para um indivíduo com fragilidade $w=1$, que designaremos por função de risco cumulativa subjacente.

As funções $\mu(t \mid w)$ e $S(t \mid w)$ correspondem a um modelo individual que não é observável, como refere Aalen (1988). A f.s. e a função de risco populacionais, correspondentes aos dados de que efetivamente dispomos, são respetivamente, (Hougaard, 1984)

$$S(t)=\int_0^{\infty} \exp\left[-w\Lambda(t)\right] dF(w) = L(\Lambda(t)) \qquad \text{e} \qquad h(t) = -\frac{L'(\Lambda(t))}{L(\Lambda(t))}\lambda(t),$$

onde $L(s)=E[\exp(-sW)]$ é a transformada de Laplace da distribuição da fragilidade W, $F(.)$ é a correspondente função de distribuição (f.d.) e $L'(s)=dL(s)/ds$.

Suponhamos que a fragilidade W é uma v.a. absolutamente contínua com f.d.p. $f(w)$. Consideremos agora a distribuição condicional de W dado $T>t$, i.e., a distribuição da fragilidade entre os sobreviventes no instante t ou à idade t. A f.d.p. correspondente é dada por (Hougaard, 1984)

$$g(w \mid T>t) = \frac{\exp\left[-w\Lambda(t)\right] f(w)}{L(\Lambda(t))}.$$

Notemos que a definição de fragilidade pressupõe que cada indivíduo nasce com uma certa fragilidade, a qual não se altera durante toda a sua vida. Portanto, a distribuição da fragilidade coincide com a distribuição da fragilidade entre os sobreviventes no instante $t=0$, ou seja, com a distribuição da fragilidade à nascença.

Quanto ao valor médio de W dado $T>t$, ou seja, a fragilidade média entre os sobreviventes no instante t, é dada por $E(W \mid T>t) = -\frac{L'(\Lambda(t))}{L(\Lambda(t))}$.

Então, a função de risco populacional pode ser escrita como $h(t)=E(W \mid T>t)\lambda(t)$, o que evidencia que a função de risco populacional (observada) $h(t)$ é o valor esperado, entre os sobreviventes no instante t, da verdadeira função de risco $\mu(t \mid w)$.

A fragilidade média na população sobrevivente decresce com o tempo, visto que os indivíduos que possuem maior fragilidade irão morrer mais cedo do que os outros.

De facto, prova-se que $\frac{dE(W \mid T>t)}{dt} = -\lambda(t)\,\mathrm{var}(W \mid T>t) < 0$.

Comparando então $\mu(t \mid w) = w\lambda(t)$ e $h(t) = E(W \mid T > t)\lambda(t)$, podemos concluir que

- se $\lambda(t)$ for crescente, a função de risco individual cresce mais rapidamente do que a função de risco populacional.
- se $\lambda(t)$ for decrescente, a função de risco populacional decresce mais rapidamente do que a função de risco individual.
- se $\lambda(t)$ for constante, a função de risco populacional é sempre uma função decrescente.

Portanto, a existência de heterogeneidade não observada é um fator que pode enviesar as conclusões acerca da evolução da mortalidade na população. De facto, seja qual for o comportamento de $\lambda(t)$, a observação da população leva a uma avaliação mais otimista da evolução da mortalidade do que a que se verifica na realidade, a nível individual.

Nos modelos com fragilidade, o coeficiente de variação é utilizado como medida do grau de heterogeneidade da população. De facto, neste tipo de modelos, interessa-nos estudar a dispersão da distribuição da fragilidade entre os sobreviventes, visto que nos permite tirar conclusões sobre a heterogeneidade existente nessa população num dado momento. Com efeito, uma menor dispersão corresponde a uma distribuição da fragilidade mais concentrada e portanto a uma população em que os indivíduos apresentam valores da fragilidade muito semelhantes. Portanto, uma menor dispersão é indicadora de uma população mais homogénea.
Assim, o estudo do coeficiente de variação da distribuição da fragilidade entre os sobreviventes como função do tempo, quanto à sua monotonia, permite avaliar a evolução ao longo do tempo da heterogeneidade na população sobrevivente.

O modelo desenvolvido por Vaupel *et al.* (1979) destina-se a uma situação típica de tabelas de mortalidade. Quando dispomos de informação sobre fatores de risco/prognóstico e queremos modelar o efeito de uma covariável não observável podemos considerar uma extensão de um modelo de regressão como, por exemplo, a seguinte extensão do modelo de Cox para inclusão de heterogeneidade não observada e não explicada pelas covariáveis observadas **z**:

$$h(t \mid \mathbf{z}, w) = w\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}) \qquad (2)$$

A f.s. e a função de risco marginais de T dado **z** são então

$$S(t \mid \mathbf{z}) = L[\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t)] \quad \text{e} \quad h(t \mid \mathbf{z}) = -\frac{L'(\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t))}{L(\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t))}\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}),$$

onde $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$ é a função de risco cumulativa para um indivíduo com $\mathbf{z} = \mathbf{0}$.

Uma questão que naturalmente se coloca, ao reconhecermos a necessidade de utilização de um modelo que inclua heterogeneidade não observada, é a da escolha da distribuição da fragilidade, de entre as que possuem suporte em [0,∞). As distribuições que têm sido mais utilizadas como modelos para a fragilidade são:

- distribuição gama (Vaupel *et al*., 1979; Hougaard, 1984, 1991; Aalen, 1987; Klein, 1992),
- distribuição Gaussiana inversa (Hougaard, 1984, 1991; Aalen, 1988),
- família das distribuições estáveis positivas (Hougaard,1986, 1991; Aalen, 1988)
- distribuição de Poisson composta (Aalen, 1988, 1992).

A distribuição gama tem sido a mais utilizada como distribuição da fragilidade, em parte devido às vantagens que apresenta do ponto de vista do tratamento matemático. Se W seguir uma distribuição gama de parâmetros δ e θ, então a distribuição da fragilidade entre os sobreviventes no instante t é

ainda gama, de parâmetros δ e $\theta + \Lambda(t)$, o que corresponde apenas a uma mudança de escala. Neste caso, o coeficiente de variação entre os sobreviventes é constante e igual a $\delta^{-1/2}$. Portanto, o grau de heterogeneidade existente na população sobrevivente não se altera ao longo do tempo.

Ao propor a utilização da distribuição Gaussiana inversa como alternativa à distribuição gama, Hougaard (1984) referiu que neste modelo a população sobrevivente torna-se mais homogénea com o decorrer do tempo, o que é consistente com uma população sujeita a um fenómeno de seleção em que os indivíduos mais frágeis vão sendo eliminados. De facto, se a distribuição de W for Gaussiana inversa de parâmetros δ e κ, o coeficiente de variação entre os sobreviventes é dado por $2^{-1/2}[\kappa(\delta + \Lambda(t))]^{-1/4}$, portanto decresce com o tempo. Quanto à distribuição da fragilidade entre os sobreviventes no instante t, é novamente Gaussiana inversa de parâmetros $\delta + \Lambda(t)$ e κ.

Concluímos assim que as distribuições gama e Gaussiana inversa são fechadas para a seleção induzida pela mortalidade, i.e., a distribuição condicional da variável fragilidade, dada a sobrevivência até determinado instante, pertence ainda à família de distribuições da fragilidade. Portanto, a existência de truncatura à esquerda não coloca qualquer problema à análise dos dados, visto que ainda é aplicável a mesma família de distribuições.
Economou e Caroni (2005) propuseram métodos gráficos de diagnóstico para avaliação da adequabilidade da distribuição da fragilidade (gama ou Gaussiana inversa), admitindo válido o modelo multiplicativo.

Hougaard (1986) considerou como família de distribuições da fragilidade as distribuições estáveis positivas com expoente característico $\alpha \in (0,1]$, cuja transformada de Laplace, a menos de um fator de escala, é dada por $L(s) = \exp(-s^{\alpha})$.
Refira-se agora uma propriedade interessante desta família. Consideremos a extensão (2) do modelo de Cox com fragilidade estável positiva. Então, a função de risco marginal de T dado $\mathbf{z}$ é

$$h(t \mid \mathbf{z}) = \alpha\left[\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t)\right]^{\alpha-1}\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}) = \alpha\lambda_0(t)\left[\Lambda_0(t)\right]^{\alpha-1}(\exp(\boldsymbol{\beta}'\mathbf{z}))^{\alpha},$$

portanto é da forma $h(t \mid \mathbf{z}) = h_0(t)\exp(\alpha\boldsymbol{\beta}'\mathbf{z})$ onde $h_0(t) = \alpha\lambda_0(t)\left[\Lambda_0(t)\right]^{\alpha-1}$. Logo, se a fragilidade seguir uma distribuição estável positiva, a hipótese de riscos proporcionais é preservada a nível populacional e a dependência nas covariáveis retém a sua forma original, embora os coeficientes de regressão estejam agora mais próximos de zero. Assim sendo, neste caso o modelo de Cox é consistente com a possibilidade de que uma parte da variação provenha de covariáveis não observadas.
No entanto, dado que a família das distribuições estáveis positivas não é fechada para a seleção induzida pela mortalidade, Hougaard (1986) derivou, a partir dela, uma nova família de distribuições que verifica esta propriedade.

Em medicina, constata-se por vezes que alguns indivíduos na população não são suscetíveis a determinada doença, enquanto que os restantes apresentam um grau variável de suscetibilidade, possivelmente de natureza genética. Nesta situação, é adequado considerar a variável fragilidade como uma v.a. mista com massa de probabilidade não nula em zero e que tenha distribuição contínua na semirreta positiva. Aalen (1988, 1992) propôs a utilização da distribuição de Poisson composta gerada por variáveis aleatórias gama. Esta distribuição pode ser expressa como soma de um número aleatório de variáveis aleatórias gama independentes, em que o número de parcelas tem distribuição de Poisson. A probabilidade $P(W = 0) > 0$ corresponde à hipótese de não suscetibilidade a determinado acontecimento para um certo grupo de indivíduos. Se estes representarem a maior parte da população, a heterogeneidade pode ter um impacto considerável, ainda que o acontecimento em causa seja raro.
Uma distribuição com propriedades interessantes, adequada também a esta situação, é a distribuição Qui-quadrado não central com zero graus de liberdade (Rocha, 1995, 1996).

Consideremos agora a questão da escolha da distribuição condicional do tempo de vida T dado $W = w$, o que é equivalente à escolha de uma forma paramétrica para a função de risco subjacente $\lambda(t)$. Os modelos mais utilizados são: exponencial, Weibull e Gompertz.
Notemos que, no caso da distribuição exponencial, a função de risco subjacente é constante, portanto cada indivíduo tem uma probabilidade instantânea de morte que não se altera ao longo do tempo. Conforme referimos anteriormente, se existir heterogeneidade de risco entre os indivíduos, a população apresenta função de risco decrescente, resultante de uma forte seleção dos indivíduos de alto risco.

Vamos agora referir alguns exemplos de situações práticas para as quais os modelos univariados de heterogeneidade não observada, anteriormente referidos, são adequados.

• Expulsão do dispositivo intrauterino (DIU)

No início, o risco de expulsão é elevado e depois decresce rapidamente com o tempo decorrido desde a inserção do DIU. Uma explicação possível é que cada mulher apresenta um risco de expulsão constante, mas o nível do risco varia muito de mulher para mulher. A explicação biológica defende que o organismo adapta-se ao uso do DIU. Notemos que se verifica um antagonismo entre as explicações biológica e estatística. No entanto, dado que é difícil negar a existência de alguma heterogeneidade entre os indivíduos, um modelo com fragilidade será certamente um bom contributo para a compreensão do problema.
Aalen (1987) considerou um modelo com função de risco individual constante $\mu(t \mid w) = w$ e em que a variável fragilidade segue uma distribuição gama. A f.s. populacional é então da forma $S(t) = (1 + t/\theta)^{-\delta}$, o que corresponde à distribuição de Pareto de tipo II (ou distribuição de Lomax). A função de risco populacional é $h(t) = \delta/(\theta + t)$, logo decrescente.

• Transplante renal

Uma situação em que as funções de risco não são proporcionais é aquela em que se verifica o chamado declínio no efeito do tratamento: inicialmente o novo tratamento é superior ao tratamento habitual mas, após algum tempo, a razão das funções de risco correspondentes aos dois grupos de tratamento tende para um, portanto as funções de risco são convergentes. Este efeito foi encontrado por Dabrowska *et al.* (1992) na análise de dados obtidos num estudo para comparação de dois medicamentos imunossupressores usados no tratamento de doentes submetidos a transplantes renais. Embora não considerado pelos autores, o modelo (2) com fragilidade gama seria adequado a esta situação.
Com efeito, dado o vetor de covariáveis observadas $\mathbf{z}$, se admitirmos que W tem uma distribuição gama de valor médio 1 e variância γ, a f.s. populacional é dada por

$$S(t \mid \mathbf{z}) = L[\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t)] = (1 + \gamma \exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t))^{-1/\gamma}.$$

A correspondente função de risco é então da forma

$$h(t \mid \mathbf{z}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{z})\lambda_0(t)}{1 + \gamma \exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t)}.$$

É de salientar que este já não é um modelo de riscos proporcionais. De facto, seja $\mathbf{z}_j, j = 1,2$ o vetor de covariáveis associado a um indivíduo pertencente ao grupo de tratamento j. Então,

$$\lim_{t\to 0} \frac{h(t \mid \mathbf{z}_1)}{h(t \mid \mathbf{z}_2)} = \exp[\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)] \quad \text{enquanto que} \quad \lim_{t\to\infty} \frac{h(t \mid \mathbf{z}_1)}{h(t \mid \mathbf{z}_2)} = 1.$$

• Nefropatia diabética

Trata-se de uma grave doença renal que provoca elevada mortalidade nos diabéticos. Dado que um grande número de diabéticos nunca desenvolve esta doença, podemos considerar que estes indivíduos apresentam suscetibilidade nula. A função de risco do tempo até à ocorrência de nefropatia cresce até um valor máximo e depois decresce para zero. Assim sendo, uma escolha natural para a fragilidade é a distribuição de Poisson composta, podendo a função de risco subjacente ser de tipo Weibull ou Gompertz.

3. O modelo aditivo

Rocha (1995) propôs um modelo em que a fragilidade atua de modo aditivo na função de risco individual. Suponhamos então que o tempo de vida T é uma v.a. absolutamente contínua e que a função de risco no instante t para um indivíduo com fragilidade $W = w$ é dada por

$$\mu(t \mid w) = w + \lambda(t) \qquad (3)$$

onde W é uma v.a. não negativa e a função $\lambda(t)$, que designamos por função de risco subjacente, é não negativa e representa a função de risco para um indivíduo padrão com fragilidade zero. A f.s. condicional de T dado $W = w$ é então dada por

$$S(t|w) = \exp[-wt - \Lambda(t)] \ = \exp(-wt) S_0(t),$$

onde $\Lambda(t) = \int_0^t \lambda(u)\, du$ é a função de risco cumulativa subjacente e $S_0(t) = \exp[-\Lambda(t)]$ é a correspondente f.s. subjacente.

Vamos agora obter a f.s. e a função de risco populacionais (observadas). Seja $F(.)$ a f.d. da fragilidade W. Então

$$S(t) = \exp\big[-\Lambda(t)\big] \int_0^\infty \exp(-wt)\, dF(w) = \exp\big[-\Lambda(t)\big] L(t) \quad \text{e} \quad h(t) = \lambda(t) - L'(t)/L(t),$$

onde $L(t)$ é a transformada de Laplace da distribuição de W. Assim, a relação entre funções a nível individual e populacional pode ser obtida a partir da transformada de Laplace da distribuição da fragilidade, tal como Hougaard (1984) mostrou para o modelo multiplicativo.
Notemos que a f.s. individual coincide com a f.s. populacional no caso em que $L(t) = \exp(-wt)$, ou seja, quando a distribuição da fragilidade for degenerada em w. Esta situação corresponde a ausência de heterogeneidade na população, já que todos os indivíduos apresentam um mesmo valor da fragilidade. Sem perda de generalidade, podemos então considerar que a fragilidade é nula.

Se $\lambda(t) = \lambda$, constatamos que a função de risco populacional $h(t) = \lambda - L'(t)/L(t)$ é sempre decrescente, visto que $\dfrac{d\,h(t)}{dt} = -\dfrac{d^2 \log L(t)}{dt^2} < 0$.

Além disso, se a fragilidade W seguir uma distribuição gama de parâmetros δ e θ, então a sua transformada de Laplace é dada por $L(s) = (1 + s/\theta)^{-\delta}$. Então, admitindo $\lambda(t) = \lambda$, vem que $S(t) = \exp(-\lambda t)(1 + t/\theta)^{-\delta}$, que é a f.s. de uma distribuição de Pareto generalizada proposta por Davis e Feldstein (1979) como modelo para o tempo de vida.

Admitamos que a fragilidade W é uma v.a. com f.d.p. $f(w)$. Vamos agora obter a distribuição da fragilidade entre os sobreviventes no instante t. A f.d.p. condicional de W dado $T > t$ é dada por

$$g(w \mid T > t) = \frac{S(t \mid w) f(w)}{S(t)} = \frac{\exp(-wt) f(w)}{L(t)},$$

donde concluímos que a distribuição da fragilidade entre os sobreviventes num dado instante não depende da função de risco subjacente $\lambda(t)$. Portanto, no caso do modelo aditivo, a forma de que se reveste a dependência do tempo na função de risco individual não vai ter qualquer influência na distribuição da fragilidade entre os sobreviventes.

O valor médio de W dado $T > t$ é dado por $E(W \mid T > t) = -L'(t)/L(t)$. Logo, a função de risco populacional pode ser expressa como $h(t) = E(W \mid T > t) + \lambda(t)$.

Concluímos assim que, numa população heterogénea, a função de risco que observamos num dado instante é afinal o valor esperado da verdadeira função de risco, entre os sobreviventes nesse instante. Portanto, esta propriedade é característica dos modelos com fragilidade, sejam de tipo multiplicativo ou aditivo.

Como consequência do processo de seleção a que os indivíduos estão sujeitos, a fragilidade média entre os sobreviventes irá decrescer com o tempo. Com efeito, após alguns cálculos, vem que

$$\frac{dE(W \mid T > t)}{dt} = -\operatorname{var}(W \mid T > t) < 0.$$

Por outro lado, a f.d.p. de W dado $T = t$, i.e., da distribuição da fragilidade entre os indivíduos que morrem no instante t é

$$g(w \mid t) = \frac{[w + \lambda(t)]\exp(-wt) f(w)}{\lambda(t)L(t) - L'(t)}.$$

É de salientar a seguinte relação que é possível estabelecer entre a distribuição da fragilidade entre os sobreviventes no instante t e a distribuição da fragilidade entre os indivíduos que morrem nesse mesmo instante:

$$g(w \mid t) = \frac{\mu(t \mid w)}{h(t)} g(w \mid T > t).$$

• O modelo aditivo como modelo de riscos competitivos

Interpretemos então o modelo aditivo com fragilidade no contexto dos riscos competitivos. Recordemos que a f.s. para um indivíduo com fragilidade $W = w$ é dada por $S(t|w) = S_0(t)\exp(-wt)$, onde $S_0(t) = \exp[-\Lambda(t)]$ é a f.s. subjacente correspondente a uma v.a. T_0, que representa o tempo de vida para um indivíduo com fragilidade nula. Podemos então interpretar a distribuição condicional de T dado $W = w$ como sendo a distribuição de $X = \min(T_0, C)$, em que T_0 e C são variáveis aleatórias independentes e C é uma v.a. não observável com distribuição $Exp(w)$, para $w > 0$ fixo. É interessante notar que T_0 e C podem ser considerados como tempos de vida potenciais independentes correspondentes a duas causas de morte.
Temos então a nível individual um modelo de riscos competitivos independentes, em que as duas causas de morte seriam o tempo e a fragilidade específica de cada indivíduo.

• Um modelo aditivo com covariáveis

Consideremos agora o modelo de regressão linear proposto por Aalen (1989). Comecemos por admitir que a heterogeneidade populacional pode ser totalmente explicada em termos de p covariáveis fixas, cujo valor observamos para cada indivíduo. Então, a função de risco no instante t para um indivíduo a que está associado o vetor $\mathbf{z}$ é da forma $\lambda(t;\mathbf{z}) = \alpha_0(t) + \alpha_1(t)z_1 + \ldots + \alpha_p(t)z_p$, em que $\alpha_0(t)$ é interpretada como uma função parâmetro subjacente e $\alpha_1(t),\ldots,\alpha_p(t)$ são os coeficientes de regressão (designados por funções de regressão) e representam o efeito das covariáveis no instante t.

Consideremos agora uma fragilidade W não observada que descreve os fatores de risco que não foram incluídos no modelo anterior e suponhamos que a verdadeira função de risco no instante t para um indivíduo com fragilidade $W = w$ e a que está associado o vetor de covariáveis **z** é dada por

$$\mu(t \mid \mathbf{z}, w) = w + \alpha_0(t) + \alpha_1(t) z_1 + \ldots + \alpha_p(t) z_p .$$

Trata-se, portanto, de uma extensão do modelo de Aalen (1989) para inclusão de heterogeneidade não observada. Logo, a f.s. correspondente é

$$S(t \mid \mathbf{z}, w) = \exp(-wt) \exp[-A_0(t) + A_1(t) z_1 + \ldots + A_p(t) z_p],$$

onde $A_j(t) = \int_0^t \alpha_j(u)\, du$. Portanto, a f.s. marginal de T dado **z** é

$$S(t; \mathbf{z}) = \exp[-A_0(t) + A_1(t) z_1 + \ldots + A_p(t) z_p] L(t),$$

sendo a correspondente função de risco da forma seguinte:

$$h(t; \mathbf{z}) = [\alpha_0(t) - L'(t)/L(t)] + \alpha_1(t) z_1 + \ldots + \alpha_p(t) z_p .$$

Podemos então concluir que o modelo obtido é ainda um modelo linear com as mesmas funções de regressão $\alpha_1(t), \ldots, \alpha_p(t)$, portanto, o efeito das covariáveis observadas não foi alterado pela inclusão da fragilidade no modelo inicial. No entanto, a função parâmetro subjacente é agora dada por $\alpha_0^*(t) = \alpha_0(t) - L'(t)/L(t)$, ou seja, $\alpha_0^*(t) = \alpha_0(t) + E(W \mid T > t)$.

4. Modelo de heterogeneidade baseado na distribuição Qui-quadrado não central com zero graus de liberdade

Conforme já referimos, a heterogeneidade não observada pode ser devida à existência de indivíduos não suscetíveis a determinado acontecimento. Assim sendo, uma distribuição possível para a fragilidade, proposta por Rocha (1995, 1996), é a distribuição Qui-quadrado não central com zero graus de liberdade e parâmetro de não centralidade $\gamma > 0$ (Siegel, 1979), que representaremos por $\chi_0'^2(\gamma)$.

Notemos que esta distribuição pode ser expressa como uma soma de um número aleatório de variáveis aleatórias independentes com distribuição χ_2^2 , em que o número de parcelas tem distribuição de Poisson de parâmetro γ/2. Encarada desta forma, a distribuição $\chi_0'^2(\gamma)$ apresenta-se como um caso particular da distribuição de Poisson composta sugerida por Aalen (1988). No entanto, o seu estudo é plenamente justificado, visto que a sua utilização como distribuição da fragilidade no modelo aditivo leva à obtenção de um modelo de sobrevivência que pode ser interpretado como modelo de riscos competitivos. É um modelo adequado a uma população sobre a qual atuam duas causas de morte, sendo a população constituída por indivíduos suscetíveis a ambas as causas e indivíduos que, de facto, são suscetíveis apenas a uma única causa de morte.

Consideremos então o modelo (3). Dado que a transformada de Laplace da distribuição $\chi_0'^2(\gamma)$ é $L(s) = \exp[-\gamma(1-(1+2t)^{-1})/2]$, a f.s. e a função de risco populacionais são dadas, respetivamente, por

$$S(t) = \exp\left[-\frac{\gamma}{2}(1-(1+2t)^{-1}) - \Lambda(t)\right] \quad \text{e} \quad h(t) = \lambda(t) + \frac{\gamma}{(1+2t)^2}.$$

Podemos então interpretar este modelo como um modelo de riscos competitivos independentes em que existem duas causas de morte. A partir das expressões anteriores, as f.s. e funções de risco correspondentes a cada uma das causas são dadas, respetivamente, por

$$S_1(t) = \exp(-\Lambda(t)) \qquad S_2(t) = \exp\left[-\frac{\gamma}{2}(1-(1+2t)^{-1})\right]$$

$$h_1(t) = \lambda(t) \qquad h_2(t) = \frac{\gamma}{(1+2t)^2}.$$

Notemos que $\lambda(t)$, que é a função de risco associada aos indivíduos não suscetíveis (i.e. com fragilidade nula), coincide com a função de risco correspondente à causa 1. Quanto aos indivíduos suscetíveis, a respetiva função de risco não coincide com $h_2(t)$ e é dada por

$$h_S(t) = \lambda(t) + \frac{\gamma}{(1+2t)^2}\left\{1-\exp\left[-\frac{\gamma}{2}(1+2t)^{-1}\right]\right\}^{-1}.$$

Podemos então considerar que os indivíduos de suscetibilidade nula são, na realidade, suscetíveis apenas a uma única causa de morte (causa 1), enquanto que os restantes indivíduos são suscetíveis a ambas. A causa de morte 1 apresenta uma função de risco que pode ser crescente, decrescente ou constante. Por outro lado, como $h_2(t)$ é uma função decrescente, podemos interpretar a causa 2 como correspondendo à existência de defeitos congénitos.

Notemos que a fragilidade entre os sobreviventes no instante t tem distribuição $(1+2t)^{-1}\chi_0'^2(\gamma_1)$, onde o parâmetro de não centralidade é $\gamma_1 = \gamma(1+2t)^{-1}$. Conclui-se portanto que, no modelo aditivo, a família das distribuições Qui-quadrado não centrais com zero graus de liberdade é fechada para a seleção induzida pela mortalidade.

Bibliografia

Aalen, O.O. (1987). Two examples of modelling heterogeneity in survival analysis. *Scandinavian Journal of Statistics*, 14, 19-25.

Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine*, 7, 1121-1137.

Aalen, O.O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8, 907-925.

Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2, 951-972.

Dabrowska, D.M., Doksum, K.A., Feduska, N.J., Husing, R. e Neville, P. (1992). Methods for comparing cumulative hazard functions in a semi-proportional hazard model. *Statistics in Medicine*, 11, 1465-1476.

Davis, H.T. e Feldstein, M.L. (1979). The generalized Pareto law as a model for progressively censored survival data. *Biometrika*, 66, 299-306.

Economou, P. e Caroni, C. (2005). Graphical tests for the assumption of Gamma and Inverse Gaussian frailty distributions. *Lifetime Data Analysis*, 11, 565-582.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71, 75-83.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73, 387-396.

Hougaard, P. (1991). Modelling heterogeneity in survival data. *J. Appl. Prob*., 28, 695-701.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Rocha, C. S. (1995). Modelos com fragilidade em Análise de Sobrevivência. Tese de Doutoramento, Universidade de Lisboa.

Rocha, C. S. (1996). Survival Models for Heterogeneity using the Non-Central Chi-Squared Distribution with Zero Degrees of Freedom. Em *Lifetime Data: Models in Reliability and Survival Analysis*, N.P. Jewell *et al.*(Eds.), Kluwer Academic Publishers: Dordrecht, 275-279.

Siegel, A.F. (1979). The noncentral chi-squared distribution with zero degrees of freedom. *Biometrika*, 66, 381-386.

Vaupel, J.W., Manton, K.G. e Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439-454.

Censura intervalar: modelação de dados do estado atual

Ana Luísa Papoila**,** *apapoila@hotmail.com*

Faculdade de Ciências Médicas da Universidade Nova de Lisboa e CEAUL

1. Introdução

Análise de Sobrevivência é, sem dúvida, uma das áreas mais interessantes da Estatística. O seu objetivo é estudar o tempo que decorre entre um instante inicial e a ocorrência de um determinado acontecimento de interesse, que se pode revestir de diversas formas, daí a sua vasta aplicação. Num contexto diferente do da Medicina, onde indubitavelmente a utilização da Análise de Sobrevivência é extensa, encontram-se também imensas aplicações. Desde a Demografia, em que o interesse pode recair na análise da duração dos casamentos ou, de uma forma mais geral, na análise da história dos acontecimentos (*event history analysis*), passando pela Indústria, onde é importante estudar o tempo até à falha de determinadas máquinas ou componentes eletrónicas (teoria da fiabilidade) e não esquecendo, como é óbvio, a Economia e as Ciências Atuariais, todas estas são áreas de investigação em que a Análise de Sobrevivência é frequentemente utilizada aquando do estudo estatístico dos dados. Na realidade, o que torna a análise de sobrevivência a única abordagem possível nalgumas situações, tem a ver com o tipo de dados envolvidos e as consequentes limitações na sua recolha. Assim, enquanto que habitualmente qualquer variável pode ser medida instantaneamente, quando há que registar tempos de vida, é necessário "esperar" até que o acontecimento de interesse ocorra. No entanto, existem situações em que o estudo termina, por motivos financeiros ou éticos, antes da ocorrência da morte de alguns indivíduos ou, em que o único conhecimento que temos acerca da morte é o de que ela ocorreu entre dois instantes e, em ambos os casos, dizemos estar perante observações censuradas.

Indubitavelmente, as Ciências Biomédicas exerceram uma grande influência na terminologia utilizada. Assim, quando se realiza o acontecimento de interesse, diz-se que ocorreu uma morte e o tempo que decorre entre o instante inicial e a morte designa-se, habitualmente, por tempo de vida ou de sobrevivência. Este tempo é uma variável aleatória não negativa e ao longo desta exposição iremos considerar que segue uma distribuição absolutamente contínua.

Vejamos então alguns conceitos básicos de análise de sobrevivência, abordando ainda os vários tipos de censura existentes (para um estudo mais detalhado deste tema, podem ser consultadas algumas referências tais como Klein e Moeschberger (1997) e Collett (2003)).

2. Conceitos Básicos

Seja T uma variável aleatória não negativa, absolutamente contínua, que representa o tempo de vida de um indivíduo pertencente a uma população homogénea. Vejamos que funções caracterizam a distribuição desta variável aleatória.

Define-se função de sobrevivência como sendo a probabilidade de um indivíduo sobreviver para além do instante t, isto é, $S(t) = P(T > t),\ t \geq 0$. Esta função goza das seguintes propriedades: $S(0) = 1, S(\infty) = 0$, é uma função monótona, estritamente decrescente e contínua. Paralelamente à função de sobrevivência, surge a função de distribuição que se denota por F e que se define como $F(t) = P(T \leq t)$, para $t \geq 0$. Evidentemente, $F(t) = 1 - S(t)$. A função densidade de probabilidade denota-se por f e define-se como

$$f(t) = \lim_{dt \to 0^+} \frac{P[t \leq T < t + dt]}{dt}.$$

Intuitivamente, $f(t)dt$ pode ser interpretado como a probabilidade do acontecimento de interesse ocorrer exatamente no instante t.

No entanto, dada a natureza dinâmica dos dados de sobrevivência, torna-se conveniente também caracterizar a distribuição de T através de uma função que descreva a evolução, ao longo do tempo, da probabilidade instantânea de morte de um indivíduo. Surge assim a função de risco (*hazard function*), que se define através da seguinte expressão

$$\lambda(t) = \lim_{dt \to 0^+} \frac{P[t \leq T < t + dt | T \geq t]}{dt}.$$

Esta função é tal que, $\lambda(t) \geq 0$ e $\int_0^\infty \lambda(t)dt = \infty$.

Ao analisarmos as expressões anteriores, podemos estabelecer as seguintes relações entre a função de sobrevivência, a função densidade de probabilidade e a função de risco:

- $\lambda(t) = \frac{f(t)}{S(t)}$
- $S(t) = \exp\left(-\int_0^t \lambda(u)\, du\right)$
- $f(t) = \lambda(t) \exp\left(-\int_0^t \lambda(u)\, du\right)$

3. Tipos de censura

Como já foi referido, uma dificuldade inerente aos estudos de sobrevivência tem a ver com a possibilidade de existirem indivíduos para os quais não foi possível observar o seu tempo de vida com exatidão, havendo apenas uma informação incompleta sobre esse tempo.

Consideremos então uma amostra de n indivíduos e sejam $T_1, T_2, \ldots, T_n$, variáveis aleatórias independentes e identicamente distribuídas, com função de sobrevivência S e que representam os tempos de vida dos n indivíduos.

Censura à direita: se, para um determinado indivíduo, o acontecimento de interesse não ocorreu durante o período de *follow-up* mas sim após o seu final, dizemos que a observação é censurada à direita. As circunstâncias nas quais pode ocorrer este tipo de censura são várias. Assim, por exemplo, ao considerarmos um ensaio clínico ou qualquer outro estudo experimental cuja finalidade seja a análise de tempos de vida, é natural ainda existirem indivíduos, no fim do estudo, para os quais não foi

observada a morte. Por outro lado, a perda de indivíduos para o *follow-up*, a interrupção do tratamento para alguns doentes devido a fortes efeitos secundários ou a ocorrência da morte por outra causa diferente da de interesse, são também situações de censura à direita.

No entanto, para que os métodos habituais de análise de sobrevivência sejam válidos, é importante que o mecanismo de censura seja independente. Isto acontece quando o conhecimento do tempo de censura de um indivíduo não fornece mais informação sobre a sobrevivência futura do indivíduo do que a que teríamos, se ele tivesse continuado em estudo. Na realidade, trata-se de uma hipótese de independência entre os mecanismos de morte e de censura e garante que os indivíduos, cujo tempo de vida foi censurado num determinado instante, têm o mesmo risco de uma morte futura do que aqueles que ainda se encontram vivos e em observação nesse instante. Um exemplo claro de censura não independente é o enunciado anteriormente, sobre a interrupção do tratamento para alguns doentes, devido a fortes efeitos secundários.

Consideremos dois tipos de censura à direita, conhecidos por censura de tipo I e censura de tipo II.

No que diz respeito à censura de tipo I, esta ocorre quando, para cada indivíduo, o acontecimento de interesse só é observado se ocorrer antes de um determinado instante C_d, fixado previamente pelo investigador. Deste modo, para uma amostra de n indivíduos, as observações consistirão nos pares de variáveis aleatórias $(Y_i, \delta_i),\ i = 1, \dots, n$, em que $Y_i = \min(T_i, C_d)$ e $\delta_i = 1$ se $T_i \leq C_d$ e $\delta_i = 0$ se $T_i > C_d$.

A censura de tipo II surge quando o estudo termina no instante em que é observada a r-ésima morte, sendo r um número pré-determinado, $(1 \leq r \leq n)$.

Um tipo de censura mais geral é a chamada censura aleatória (*random censoring*). Neste caso, a cada indivíduo está associado um tempo de vida T_i e um tempo de censura C_i, em que T_i e C_i são variáveis aleatórias independentes. Assim sendo, as observações consistirão nos pares de variáveis aleatórias $(Y_i, \delta_i),\ i = 1, \dots, n$, em que $Y_i = \min(T_i, C_i)$ e $\delta_i = 1$ se $T_i \leq C_i$ e $\delta_i = 0$ se $T_i > C_i$.

Censura à esquerda: um tempo de vida associado a um indivíduo é considerado censurado à esquerda se é menor do que um tempo C_e, que foi registado. Neste caso, pode acontecer que o acontecimento de interesse tenha ocorrido antes da pessoa entrar em estudo. Um exemplo bastante elucidativo surge quando se pretende estudar a idade em que uma criança consegue realizar determinada tarefa. Assim, como antes do início do estudo, é possível que algumas crianças já saibam realizar essa tarefa, as observações correspondentes são censuradas à esquerda.

Nesta situação, para uma amostra de n indivíduos, as observações consistirão nos pares de variáveis aleatórias $(U_i, \varepsilon_i),\ i = 1, \dots, n$, em que $U_i = \max(T_i, C_e)$ e $\varepsilon_i = 1$ se $U_i \geq C_e$ e $\varepsilon_i = 0$ se $U_i < C_e$.

Este tipo de censura não é muito comum.

Censura intervalar: quando não é possível observar o instante exacto em que ocorre o acontecimento de interesse mas, pelo contrário, apenas sabemos ter ocorrido num certo intervalo aleatório de tempo, dizemos que estamos perante uma observação censurada num intervalo.

Existem dois tipos de censura intervalar. Assim, se a única informação de que dispomos é saber se, em determinado instante de monitorização, o acontecimento de interesse já ocorreu ou ainda não ocorreu, dizemos que se trata de censura intervalar-caso I e os dados designam-se por dados do estado atual (*current status data*).

Quando se conhece o intervalo no qual se realizou o acontecimento de interesse, dizemos que se trata de censura intervalar-caso II. Neste caso, o acontecimento de interesse ocorreu entre dois instantes observados, ou seja, $T \in [T_e, T_d]$. Dado que estes intervalos são aleatórios e há frequentemente sobreposições, não é possível usar a metodologia usual para dados agrupados.

Este tipo de censura surge com frequência em estudos longitudinais em que há *follow-up* periódico, como por exemplo, se desejarmos estudar a distribuição do tempo que decorre entre o fim do tratamento a um determinado carcinoma e o instante em que ocorre uma recidiva da doença. As observações resultantes desde estudo serão censuradas num intervalo, uma vez que a recidiva só será detetada numa visita programada de *follow-up* ou numa visita antecipada, dado a existência de queixas. Em ambos os casos, o acontecimento de interesse ocorreu entre duas visitas consecutivas.

Para uma melhor compreensão dos modelos de regressão para dados do estado atual, vejamos como se pode estabelecer uma correspondência entre modelos em análise de regressão binária e em análise de sobrevivência.

4. Correspondência entre modelos em análise de regressão binária e em análise de sobrevivência

Em qualquer problema de regressão pretende-se modelar a relação existente entre uma variável dependente Y e um vector de variáveis explicativas ou covariáveis $\boldsymbol{x}$; em análise de regressão binária, o interesse recai sobre a probabilidade $\theta(\boldsymbol{x})$ de que um indivíduo com determinados atributos, representados pelo vector de covariáveis $\boldsymbol{x} = (x_1, \dots, x_p)^T$, apresente uma determinada característica A, ou seja, $\theta(\boldsymbol{x}) = P[I(A) = 1|\boldsymbol{X} = \boldsymbol{x}]$ em que $I(.)$ representa a função indicatriz. Em análise de sobrevivência, um dos principais objetivos é estimar a função de sobrevivência $S(t|\boldsymbol{x}) = P[T > t|\boldsymbol{X} = \boldsymbol{x}]$ ou, equivalentemente, estimar a função de distribuição $F(t|\boldsymbol{x}) = P[T \leq t|\boldsymbol{X} = \boldsymbol{x}]$, em que T é uma variável contínua que representa o tempo de vida.

Se fixarmos $t \geq 0$ e considerarmos o acontecimento $A = A_t$ ="morte antes do instante t", podemos reescrever $F(t|\boldsymbol{x}) = P[I(A_t) = 1|\boldsymbol{X} = \boldsymbol{x}]$ e, então, concluímos que $F(t|\boldsymbol{x}) = \theta(\boldsymbol{x})$, ou seja, o problema de estimar $F(t|\boldsymbol{x})$ em análise de sobrevivência é equivalente ao problema de estimar $\theta(\boldsymbol{x})$ em análise de regressão binária.

Com base em Doksum e Gasko(1990), serão apresentadas duas das correspondências existentes entre modelos de regressão para análise de dados binários e modelos de sobrevivência para tempo contínuo.

Modelo de possibilidades proporcionais: em 1944, Berkson introduziu o modelo de regressão logística, um dos modelos mais utilizados na modelação de dados com resposta binária e que se define através da expressão

$$\log\left\{\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right\} = \boldsymbol{x}^T\boldsymbol{\beta} \quad \Leftrightarrow \quad \theta(\boldsymbol{x}) = L(\boldsymbol{x}^T\boldsymbol{\beta}),$$

onde L representa a função de distribuição logística $L(t) = [1 + e^{-t}]^{-1}$e $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ representa o vetor de parâmetros de regressão.

O modelo correspondente em análise de sobrevivência é

$$\log\left\{\frac{F(t|\boldsymbol{x})}{1-F(t|\boldsymbol{x})}\right\} = \log\left\{\frac{F_0(t)}{1-F_0(t)}\right\} + \boldsymbol{x}^T\boldsymbol{\beta}, \qquad (1)$$

em que $F_0(t) = F(t|\boldsymbol{X} = \boldsymbol{0})$ representa a função de distribuição subjacente (Bennett, 1983).

Se agora definirmos $\Gamma(t|\boldsymbol{x}) = \frac{F(t|\boldsymbol{x})}{1-F(t|\boldsymbol{x})}$ como sendo a função "possibilidade de morte" (*odds on death*), podemos reescrever (1) como $\Gamma(t|\boldsymbol{x}) = \Gamma_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta})$.

Notemos que a função $\Gamma_0(t)$ é crescente, contínua em $[0,\infty)$, com $\Gamma_0(0) = 0$ e $\Gamma_0(\infty) = \infty$. No modelo binário anterior, uma vez que t é fixo, $\Gamma_0(t)$ é uma constante e, portanto, $\log\Gamma_0(t)$ é absorvido pelo termo β_0.

Então, a razão entre as possibilidades de morte até ao instante t, de dois indivíduos a que estão associados os vetores $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$ é dada por

$$\frac{\Gamma(t|\boldsymbol{x_1})}{\Gamma(t|\boldsymbol{x_2})} = \frac{\exp(\boldsymbol{x}_1^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_2^T\boldsymbol{\beta})} = \exp[(\boldsymbol{x}_1^T - \boldsymbol{x}_2^T)\boldsymbol{\beta}],$$

que não depende do tempo.

Um modelo de regressão da forma (1), foi introduzido por Bennett (1983) que estudou o caso em que $\log\Gamma_0(t) = \tau\log t$.

Pensando agora numa abordagem do problema através de transformações lineares, podemos afirmar que o modelo (1) é uma transformação linear que nos conduz a um modelo linear generalizado (GLM) com função de ligação logit (McCullagh e Nelder, 1989), com a particularidade da ordenada na origem ser uma função que depende da função de distribuição subjacente do tempo de vida. Com efeito, $F(t|\boldsymbol{x}) = L[\log\Gamma_0(t) + \boldsymbol{x}^T\boldsymbol{\beta}]$, $t \geq 0$, em que $L(t) = [1+e^{-t}]^{-1}$ representa, como já vimos anteriormente, a função de distribuição logística.

Modelo Gama-logit: Aranda-Ordaz(1981) definiu uma nova transformação linearizante que, no contexto da análise de sobrevivência, é dada por:

γ-logit$(u) = \log\left\{\frac{(1-u)^{-\gamma}-1}{\gamma}\right\}$ se $\gamma > 0$ e γ-logit$(u) = \log[-\log(1-u)]$ se $\gamma = 0$.

O modelo binário vem dado por γ-logit$[\theta(\boldsymbol{x})] = \boldsymbol{x}^T\boldsymbol{\beta}$, ou, de forma equivalente por, $\theta(\boldsymbol{x}) = B(\boldsymbol{x}^T\boldsymbol{\beta})$ em que $B(.)$ é definida por:

$B(t) = 1 - [1+\gamma\exp(t)]^{-\frac{1}{\gamma}}$ se $\gamma > 0$ e $B(t) = 1 - \exp[-\exp(t)]$ se $\gamma = 0$.

É de salientar que, se considerarmos $\gamma = 1$, este modelo reduz-se ao modelo logístico, enquanto que, para $\gamma = 0$, obtém-se o modelo Gumbel. O modelo correspondente em análise de sobrevivência é dado por

$$\gamma\text{-logit}[F(t|\boldsymbol{x})] = \gamma\text{-logit}[F_0(t)] + \boldsymbol{x}^T\boldsymbol{\beta}, \qquad (2)$$

ou seja, $F(t|\boldsymbol{x}) = B\{\gamma\text{-logit}[F_0(t)] + \boldsymbol{x}^T\boldsymbol{\beta}\}$ em que, $F_0(t)$ representa a função de distribuição subjacente.

Paralelamente à definição de possibilidade de morte (*odds on death*) considerada anteriormente, também é possível definir γ-possibilidade de morte (γ-odds *on death*) (Doksum e Gasko, 1990) por

$\Gamma_\gamma(t|\boldsymbol{x}) = \frac{1-[1-F(t|\boldsymbol{x})]^\gamma}{\gamma[1-F(t|\boldsymbol{x})]^\gamma}$ se $\gamma > 0$ e $\Gamma_\gamma(t|\boldsymbol{x}) = -\log[1-F(t|\boldsymbol{x})]$ se $\gamma = 0$, (3)

e então podemos escrever o modelo (2) como

$$\Gamma_\gamma(t|\boldsymbol{x}) = \Gamma_{0\gamma}(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}),$$

em que $\Gamma_{0\gamma}(t) = \Gamma_\gamma(t|\boldsymbol{X} = \boldsymbol{0})$. Este modelo designa-se por modelo de γ-possibilidades proporcionais.

Tal como no modelo de possibilidades proporcionais, definimos, neste caso, a γ-razão de possibilidades de morte antes do instante t, de dois indivíduos a que estão associados os vectores $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$, através da expressão

$$\frac{\Gamma_\gamma(t|\boldsymbol{x_1})}{\Gamma_\gamma(t|\boldsymbol{x_2})} = \exp[(\boldsymbol{x}_1^T - \boldsymbol{x}_2^T)\boldsymbol{\beta}].$$

5. Dados censurados num intervalo

Como já vimos anteriormente, a censura intervalar ocorre quando não é possível observar diretamente o instante, mas apenas o intervalo de tempo aleatório, em que determinado acontecimento de interesse ocorreu (censura intervalar-caso II). Também referimos uma forma mais extrema de dados censurados num intervalo, conhecidos por dados do estado atual (censura intervalar-caso I), quando a única informação de que dispomos sobre o tempo de vida de cada indivíduo é se este tempo é maior ou menor que um determinado tempo de monitorização.

Como é habitual em estudos de sobrevivência, numa fase inicial pretende-se estimar a função de sobrevivência (ou, equivalentemente, a função de distribuição), sem ter ainda em conta o efeito das covariáveis e, posteriormente, analisa-se a influência que estas covariáveis exercem sobre o tempo de vida. Ora, quando estamos perante uma situação que envolve dados censurados num intervalo, sejam caso I ou caso II, os objetivos são os mesmos e, embora o nosso interesse recaia sobre dados do estado atual, deixaremos algumas referências à censura intervalar-caso II. Assim, para este tipo de censura, existem diversas abordagens que nos permitem estimar a função de distribuição da variável T, numa população homogénea:

Métodos não paramétricos: Peto (1973) e Turnbull (1976), foram pioneiros nesta área de investigação e o seu trabalho foi, sem dúvida, um ponto de partida para novos desenvolvimentos (Böhning *et al*., 1996; Gómez e Calle, 1999)

Métodos paramétricos: um dos primeiros artigos que aborda o estudo de observações censuradas num intervalo, no contexto de estudos de SIDA, é devido a Brookmeyer e Goedert (1989). Muñoz e Xu (1996) propõem modelar, também parametricamente, o tempo de incubação até ao aparecimento da SIDA.

Métodos Bayesianos não paramétricos: na presença de censura intervalar-caso II, salientamos o trabalho desenvolvido por Sinha e Dey (1997) e por Calle e Gómez (2001).

Ainda para o caso II da censura intervalar, o trabalho desenvolvido na área da análise de regressão é extenso e iremos apenas fazer algumas referências. Assim, salientamos os trabalhos sobre modelos de riscos proporcionais apresentados por Satten (1996) e Cai e Betensky (2003) e, na área dos modelos de possibilidades proporcionais, os trabalhos desenvolvidos por Shen (1998) e por Rabinowitz *et al*. (2000). No que diz respeito aos modelos de tempo de vida acelerado, referimos, entre outros, Betensky *et al*. (2001) e Tian e Cai (2004). Para um estudo mais detalhado sobre censura intervalar, pode ser consultada a referência Sun (2006).

Debrucemo-nos então sobre **dados censurados num intervalo-caso I** ou **dados do estado actual.**

Este tipo de dados surge em várias áreas de investigação, tendo sido, no entanto, em demografia que apareceram os primeiros estudos envolvendo dados do estado atual (Diamond e McDonald, 1991). De facto, em muitas aplicações demográficas, a principal variável de interesse é a idade em que determinado acontecimento de interesse ocorre, como por exemplo, a idade de desmame, a idade da menarca, a idade do primeiro casamento ou até a idade da morte. Acontece que estes dados provêm habitualmente de estudos rectrospectivos e, inevitavelmente, a sua recolha está sujeita a enviesamentos, dos quais salientamos os provocados por erros de memória. Assim sendo, vários autores (Bergsten-Brucefors, 1976; Quandt, 1987) tentaram perceber quais as consequências destes erros e chegaram à conclusão que os dados que envolvem datas fornecidas pelos entrevistados não são fiáveis, sendo por isso aconselhável a utilização de dados do estado atual. Outra forma de recolha alternativa, mais precisa, seria através de estudos prospetivos; no entanto, os elevados custos que habitualmente acompanham a sua implementação torna-os impeditivos na maioria dos casos. Juntando a estes factos a real inexistência de dados completos nalgumas situações, fica assim justificado o desenvolvimento sofrido pelas metodologias que utilizam dados do estado atual.

Existem outras áreas em que podemos encontrar dados do estado atual como, por exemplo, a epidemiologia. De facto, estes dados, revelam-se importantes no estudo de algumas características de doenças infecto-contagiosas, sobretudo quando não se consegue observar com exatidão o instante em que ocorre a infeção. Dos dados que se identificam com esta situação, salientamos os provenientes de estudos que envolvem doentes infetados com o VIH ou em risco de o estarem. Jewell e Shiboski (1990) e Shiboski (1998) abordam o grave problema de contágio pelo VIH entre parceiros sexuais. Neste tipo de estudos, são recrutados casais em que um dos parceiros está infetado com o VIH (por uma via alheia ao outro parceiro) e pretende-se estudar o tempo que decorre até que o outro também se infete, admitindo que o único meio de este contrair o vírus é através do contacto com o parceiro infetado. Estes estudos representam um marco na história dos dados do estado atual e não podem deixar de ser referenciados sempre que se aborda o tema.

Ainda em epidemiologia, outro trabalho que não poderíamos deixar de referir é o desenvolvido por Keiding (1991) e Keiding *et al*. (1996). De facto, é importante conhecer a distribuição da idade em que se contrai determinada doença para a qual existem testes de diagnóstico que nos permitem detetá-la. Assim, ao submetermos determinada população a um teste de rastreio, a presença/ausência de doença em indivíduos com determinada idade origina dados do estado atual sobre a idade em que é contraída a doença. Neste contexto, Keiding (1991) descreve um estimador de máxima verosimilhança para a distribuição da idade de ocorrência de infeção por hepatite A. Keiding *et al.* (1996) estudam ainda a distribuição da idade em que se contrai a rubéola, com base numa amostra de indivíduos do sexo masculino (população não vacinada). No entanto, nos casos em que a incidência da doença é baixa (como por exemplo a doença de Alzheimer), os dados devem ser obtidos através de um esquema de amostragem caso-controlo. Jewell e van der Laan (2002), baseando-se no trabalho de Scott e Wild (1997), propõem uma metodologia adequada à modelação de dados do estado atual provenientes de estudos caso-controlo, no âmbito do problema epidemiológico que acabamos de referir.

Dados do estado actual: estudo da distribuição do tempo de vida

No que se segue, denotaremos por V a variável aleatória que representa o tempo de vida e por T a variável aleatória que representa o tempo de monitorização.

Consideremos então uma amostra de n indivíduos a partir da qual se pretende estudar a distribuição da variável aleatória V. Com este tipo de dados, a informação de que dispomos consiste em observações (y_i, t_i) das variáveis aleatórias (Y_i, T_i), em que T_i representa o tempo de monitorização do i-ésimo indivíduo e Y_i é uma variável indicatriz definida da seguinte forma: $Y_i = 1$ se $V_i \leq T_i$ e $Y_i = 0$ se

$V_i > T_i$, em que V_i representa o tempo de vida do i-ésimo indivíduo. Isto significa que a única informação disponível sobre V_i se restringe à variável Y_i, observada em T_i. Na realidade, todas as observações de V_i são censuradas à esquerda $(Y_i = 1)$ ou censuradas à direita $(Y_i = 0)$, no i-ésimo instante de monitorização.

Assim, dadas as variáveis aleatórias V, T e Y, definidas anteriormente, podemos estabelecer a relação $E(Y|t) = P(V \leq T|t) = F(t)$, em que F representa a função de distribuição de V e, portanto, estimar F é equivalente a estimar a esperança matemática de Y para todo o instante de monitorização t, não esquecendo que F tem que ser uma função monótona crescente. Com efeito, suponhamos que temos uma amostra aleatória constituída pelos pares (y_i, t_i), $i = 1, \dots, n$. Então a verosimilhança é dada por

$$L = \prod_{i=1}^{n} F(t_i)^{y_i}\big(1 - F(t_i)\big)^{1-y_i} dG(t_i)\,,$$

em que G representa a função de distribuição de T. Se assumirmos que o tempo de monitorização T é independente do tempo de vida V, a estimação de F baseia-se na verosimilhança condicional dada por

$$L = \prod_{i=1}^{n} F(t_i)^{y_i}\big(1 - F(t_i)\big)^{1-y_i}.$$

Na ausência de covariáveis, salientamos o trabalho de Ayer *et al*. (1955) que demonstrou que um estimador de máxima verosimilhança não paramétrico (NPMLE) para a função de distribuição F pode ser facilmente calculado e estabeleceu a sua consistência. Groeneboom e Wellner (1992) mostraram que este estimador se pode identificar com a solução de um problema de regressão isotónica e apresentaram resultados assintóticos. No entanto, este estimador não paramétrico é uma função em escada com saltos num subconjunto dos instantes de monitorização $t_1, t_2, \dots, t_n$. É pois natural pensar em utilizar técnicas de suavização (Mukerjee, 1988; Mammen, 1991) incorporadas no algoritmo proposto (*pool-adjacent-violators algorithm*).

No caso de termos dados do estado atual duplamente censurados (em que também o instante inicial é desconhecido), podemos referir Jewell *et al*. (1994), van der Laan *et al*. (1997) e van der Laan e Jewell (2001).

Modelos de regressão para dados do estado atual

Para dados do estado atual, o estudo da relação entre as covariáveis e o tempo de vida foi abordado por vários autores que, na maioria dos casos, desenvolveram modelos de regressão semi-paramétricos. Com efeito, foram considerados os seguintes modelos:

- modelo de riscos proporcionais (Huang, 1996)
- modelo de possibilidades proporcionais (Rossini e Tsiatis, 1996)
- modelo de riscos aditivos semi-paramétrico (Martinussen e Scheike, 2002)
- modelo de tempo de vida acelerado (Rabinowitz *et al*., 1995)

Devemos ainda referir o trabalho de Burr e Gomatam (2002), que desenvolveram um modelo não paramétrico baseado no estimador NPMLE da função de distribuição, abordado por Ayer *et al*. (1955) e por Groeneboom e Wellner (1992).

Iremos agora descrever, com algum pormenor, o modelo semi-paramétrico desenvolvido por Shiboski (1998). De facto, utilizando a estrutura dos modelos aditivos generalizados (GAMs) para resposta binária (Hastie e Tibshirani, 1990), a metodologia desenvolvida por este autor permite uma estimação simultânea da função de distribuição subjacente correspondente ao tempo de vida, e dos parâmetros de regressão associados às covariáveis envolvidas.

Suponhamos então que temos n indivíduos em estudo e que desejamos estudar não só a distribuição F do tempo V mas também a possível influência que um conjunto de p covariáveis tem sobre V, com base na única informação disponível $(y_i, t_i, \boldsymbol{x}_i)$, $i = 1, \dots, n$. Então, podemos escrever que $P(Y_i = 1|t_i, \boldsymbol{x}_i) = F(t_i|\boldsymbol{x}_i)$ e $P(Y_i = 0|t_i, \boldsymbol{x}_i) = 1 - F(t_i|\boldsymbol{x}_i)$, donde $E(Y_i|t_i, \boldsymbol{x}_i) = F(t_i|\boldsymbol{x}_i) = \mu_i$, que, naturalmente, nos sugere a utilização de modelos lineares generalizados para relacionar as covariáveis com a variável resposta através de uma determinada função de ligação h, com base na relação $\mu_i = h(\eta_i)$.

Recordando a correspondência que existe entre os modelos de regressão em análise de sobrevivência e os modelos de regressão com resposta binária, temos que, no caso geral, $F(t_i|\boldsymbol{x}_i) = h[r(t_i) + \boldsymbol{x}_i^T\beta]$, $t_i \geq 0$, em que $r(t_i) = h^{-1}[F_0(t_i)]$, representa uma função contínua, não decrescente em $[0, +\infty)$ e F_0 representa a função de distribuição subjacente do tempo de vida V. Assim sendo, Shiboski (1998) propõe a utilização deste modelo para dados do estado atual.

Se assumirmos para V uma distribuição conhecida, a metodologia dos GLMs pode ser aplicada; caso contrário, estamos perante um modelo de regressão semi-paramétrico, na medida em que há que estimar F não parametricamente. A estimação baseia-se na verosimilhança que, neste caso, assume a forma

$$L = \prod_{i=1}^{n} F(t_i|\boldsymbol{x}_i)^{y_i}\big(1 - F(t_i|\boldsymbol{x}_i)\big)^{1-y_i} K\big(t_{i,}\boldsymbol{x}_i\big),$$

em que $K(t_i, \boldsymbol{x_i})$ representa a densidade conjunta de T_i e $\boldsymbol{X}_i$. Calculando o logaritmo da expressão anterior obtemos a log-verosimilhança

$$\log L = \sum_{i=1}^{n} y_i \log \mu_i + (1 - y_i)\log(1 - \mu_i) + \log K(t_{i,}\boldsymbol{x}_i), \qquad (4)$$

uma vez que $F(t_i|\boldsymbol{x}_i) = \mu_i$.

No entanto, a estimação faz-se com base em (4) omitindo o último termo, na medida em que estamos a supor que a censura é não informativa e que, portanto, o termo $K\big(t_{i,}\boldsymbol{x}_i\big)$ não fornece qualquer tipo de informação acerca da forma de $F(t_i|\boldsymbol{x}_i)$. A existência de censura informativa obriga a uma utilização de metodologias próprias (Zhang *et al*., 2005).

Um problema que pode complicar todo o processo de estimação tem a ver com o facto da ordenada na origem $r(t)$ ser uma função de infinitos parâmetros. Assim sendo, para ultrapassar este problema, Shiboski (1998) propôs uma diminuição da dimensão do espaço de parâmetros através de uma discretização da escala do tempo e/ou através da suavização de $r(t)$. Então, para a estimação do modelo, Shiboski (1998) utiliza uma versão modificada do algoritmo de *scores* local (Hastie e Tibshirani, 1990) em que $r(t)$ é sujeita a uma pré-suavização utilizando o LOWESS (*Locally Weighted Scatterplot Smoother*) (Cleveland, 1979), seguida por uma regressão isotónica em cada

iteração do algoritmo de *backfitting* (Hastie e Tibshirani, 1990), de forma a garantir, no final, uma estimativa da função $r(t)$ suave e monótona.

Como se pode constatar, o procedimento de estimação é bastante complexo. Foi então proposta, mais recentemente, uma nova abordagem para a modelação de dados do estado atual que se baseia em modelos de sobrevivência mais flexíveis (Papoila e Rocha, 2011):

Modelo Gama-logit aditivo

Este modelo de regressão, além de admitir que a função de ligação é desconhecida, também considera que a relação entre a variável resposta e as covariáveis pode ser uma relação não linear. Desta forma, existe seguramente uma maior flexibilidade no modelo definido pela expressão $h^{-1}[F(t|\boldsymbol{x})] = r(t) + \sum_{j=1}^{p} f_j(x_j)$, ou por

$$F(t|\boldsymbol{x}) = h\left[r(t) + \sum_{j=1}^{p} f_j(x_j)\right],$$

em que a função h é desconhecida, o preditor linear $\sum_{j=1}^{p} \beta_j x_j$ é substituído pelo preditor aditivo $\sum_{j=1}^{p} f_j(x_j)$ e as funções f_j refletem a relação funcional entre a variável resposta e as covariáveis.

O modelo correspondente em análise de regressão binária virá dado por $h^{-1}[\theta(\boldsymbol{x})] = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$, em que h é uma função de ligação desconhecida. Constatamos assim estar perante um GAM com função de ligação flexível.

Consideremos agora o caso particular em que a função h pertence à família proposta por Aranda-Ordaz (1981). O novo modelo de sobrevivência resultante virá dado por

$$\gamma\text{-logit}[F(t|\boldsymbol{x})] = \gamma\text{-logit}[F_0(t)] + \sum_{j=1}^{p} f_j(x_j). \qquad (5)$$

Este modelo é uma generalização natural do modelo definido em (2) e passaremos a designá-lo por modelo gama-logit aditivo. Utilizando a definição (3), podemos escrever o modelo (5) como $\Gamma_\gamma(t|\boldsymbol{x}) = \Gamma_{0\gamma}(t)\exp\left(\sum_{j=1}^{p} f_j(x_j)\right)$. Estamos pois perante um modelo de γ-possibilidades proporcionais.

Também agora, para o modelo proposto, podemos obter a γ-razão de possibilidades de morte antes do instante t de dois indivíduos caracterizados pelos vetores de covariáveis $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$ através da expressão

$$\frac{\Gamma_\gamma(t|\boldsymbol{x}_1)}{\Gamma_\gamma(t|\boldsymbol{x}_2)} = \exp\left\{\sum_{j=1}^{p}\left[f_j(x_{1j}) - f_j(x_{2j})\right]\right\},$$

a qual, obviamente, não depende de t.

No que diz respeito à estimação de β_0 e das funções parciais $f_1, \ldots, f_p$ foi utilizado o algoritmo iterativo *backfitting* e o algoritmo de *scores local*. Para estimar o parâmetro da função de ligação γ, foi utilizado um gráfico do perfil do desvio que se obtém calculando o desvio para uma grelha de valores do parâmetro. Interessar-nos-á o valor do parâmetro que minimize o desvio. Para mais detalhes sobre o modelo Gama-logit aditivo consultar a referência Papoila e Rocha (2011).

Referências Bibliográficas

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary regression data. *Biometrika* **68**, 357-363.

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T, e Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* **26**, 641-647.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273-277.

Bergsten-Brucefors, A. (1976). A Note on the accuracy of recalled age at menarche. *Annals of Human Biology* **3.1**, 71-3.

Betensky, R.A., Rabinowitz, D. e Tsiatis, A.A (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88**, 703-711.

Böhning, D., Schlattmann, P. e Dietz, E. (1996). Interval censored data: A note on the nonparametric maximum likelihood estimation of the distribution function. *Biometrika* **83**, 462-466.

Brookmeyer, R, e Goedert, J.J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* **45**, 325-335.

Burr, D. e Gomatam, S. (2002). On nonparametric regression for current status data. *Technical Report*, nº 2002-14, Dept. of Statistics, Stanford University.

Cai, T. e Betensky (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570-579.

Calle, M.L. e Gómez, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference* **98**, 73-87.

Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829-836.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2nd edition, Chapman & Hall/CRC, New-York.

Diamond, I.D. e Mcdonald, J.W. (1991). The analysis of current status data. Em Demographic Applications of Event History Analysis. J. Trussel, R. Hankinson e J. Tilton (eds.), Oxford University Press, Oxford.

Doksum, K.A. e Gasko, M. (1990). On a correspondence between models in binary regression and in survival analysis. *International Statistical Review* **58**, 243-252.

Gómez, G. e Calle, M.L. (1999). Nonparametric estimation with doubly censored data. *Journal of Applied Statistics* **26**(1), 45-58.

Groeneboom, P. e Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Boston.

Hastie, T. e Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, New-York.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* **24**, 540-568.

Jewell, N.P. e Shiboski, S.C. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics* **46**, 1133-1150.

Jewell, N.P., Malani, H.M. e Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of the American Statistical Association* **89**, nº 425, 7-18.

Jewell, N.P. e van der Laan, M.J. (2002). Case-control current status data. Working Paper **116**, Division of Biostatistics, School of Public Health, University of California, Berkeley.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series B* **154**, 371-412.

Keiding, N., Begtrup, K,, Scheike, T.H. e Hasibeder, G. (1996). Estimation from current-status data in continuous time. *Lifetime Data Analysis* **2**, 119-129.

Klein, J.P. e Moeschberger, M.L. (1997). *Survival Analysis - Techniques for Censored and Truncated data*. Springer-Verlag, New York.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Annals of Statistics* **19**, 724-740.

Martinussen, T. e Scheike, T.H. (2002). Efficient estimation in hazards regression with current status data. *Biometrika* **89**, 649-658.

McCullagh, P. e Nelder, J. (1989). *Generalized Linear Models*. 2nd edition, Chapman & Hall, London.

Mukerjee, R. (1988). Monotone nonparametric regression. *Annals of Statistics* **16**, 741-750.

Muñoz, A. e Xu, F. (1996). Models for the incubation of AIDS and variations according to age and period. *Statistics in Medicine* **15**, 2459-2473.

Papoila, A.L. e Rocha, C.S. (2011). Modelling current status data using Generalized Additive Models with flexible link: the additive gamma-logit model. *International Journal of Applied Mathematics & Statistics* **24**, Issue Nº. SI-11A, 2-19.

Peto, R, (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society, series C* **22**, 86-91.

Quandt, S. (1987). Material recall accuracy for dates of infant feeding transitions. *Human Organization* **46. 2**, 152-9.

Rabinowitz, D., Tsiatis, A.A. e Aragon, J. (1995). Regression with interval-censored data. *Biometrika* **82**, 501-513.

Rabinowitz, D., Betensky, R.A. e Tsiatis, A.A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics* **56**, 511-518.

Rossini, A.J. e Tsiatis, A.A. (1996). A semiparametric odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91** (434), 713-721.

Satten, G.A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83**, 355-370.

Scott e A.J., Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **85**, 165-177.

Shiboski, S. C. (1998). Generalized additive models for current status data. *Lifetime Data Analysis* **4**, 29-50.

Sinha, D. e Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195-1212.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer, New York.

Tian, L. e Cai, T. (2004). On the accelerated failure time model for current status and interval censored data. Working Paper **14**, Harvard University.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and data. *Journal of the Royal Statistical Society, series B* **38**, 290-295.

van der Laan, M.J., Bickel, P.J. e Jewell, N.P. (1997). Singly and doubly censored current status data: estimation, asymptotics and regression. *Scandinavian Journal of Statistics* **24**, 289-308.

van der Laan, M.J. e Jewell, N.P. (2001). The NPMLE for doubly censored current status data. *Scandinavian Journal of Statistics* **28** (3), 537-547.

Zhang, Z., Sun, J. e Sun, L. (2005). Statistical Analysis of current status data with informative observation times. *Statistics in Medicine* **24**, 1399-1407.

Análise de Sobrevivência – Modelos de Cura

Ana Maria Abreu, *abreu@uma.pt*

Universidade da Madeira, CCCEE, CCM

1. Introdução

A Análise de Sobrevivência consiste na análise de dados referentes a medições do tempo decorrido desde um instante inicial até à ocorrência de um determinado acontecimento, numa população. No entanto, nem sempre é possível observar o acontecimento de interesse devido, em geral, a certas restrições no processo de recolha dos dados, situação em que surgem as designadas observações censuradas.

Uma situação particular ocorre quando à maioria das observações censuradas correspondem os valores mais elevados. Este facto traduz-se graficamente na estabilização da curva da estimativa de Kaplan-Meier da função de sobrevivência (f.s.), como ilustra a Figura 1.

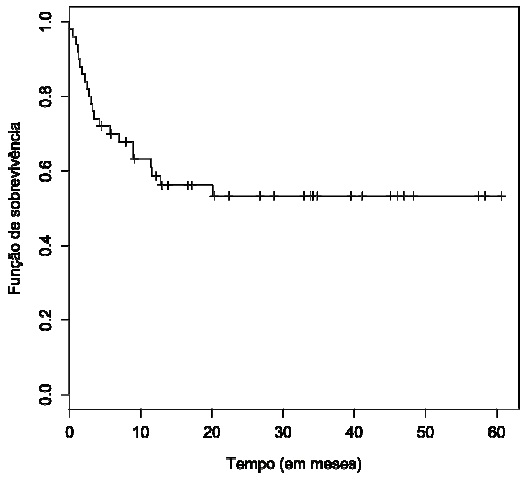


Figura 1: Dados de leucemia (Klein e Moeschberger, 1998, p.10)

Neste caso, suspeita-se da existência de indivíduos que nunca vão experimentar o acontecimento de interesse, ainda que o tempo em observação pudesse ser prolongado indefinidamente. Tais indivíduos são considerados imunes ao acontecimento de interesse e, nesta perspetiva, a função de sobrevivência que lhes corresponde, em qualquer instante, é sempre um (f.s. degenerada). Neste âmbito, a população em estudo é constituída por uma mistura de indivíduos imunes (ou curados ou não suscetíveis) e de indivíduos não imunes (ou doentes ou suscetíveis); a estes últimos corresponde uma f.s. não degenerada.

A observação do acontecimento de interesse permite que os indivíduos sejam classificados como suscetíveis. Quando esse acontecimento não é observado, em geral, não é possível identificar os indivíduos imunes, embora a sua presença seja admitida. Apenas no caso em que é definido algum critério de classificação para os indivíduos curados é que a identificação é inequívoca. Laska e Meisner (1992), Taylor (1995), Tamura *et al*. (2000) e Betensky e Schoenfeld (2001), fornecem dos poucos exemplos desta situação. Nos três primeiros, é definido um valor *a priori* e os indivíduos que tenham um tempo de vida superior a esse valor são considerados curados. Em Betensky e Schoenfeld (2001), a ocorrência de um certo acontecimento (alta hospitalar) distinto do acontecimento de interesse (morte pela doença) possibilita a identificação dos indivíduos curados.

Neste artigo, serão definidos os modelos de cura de mistura e de não-mistura. Será apresentado um modelo de cura de mistura baseado na distribuição de Chen (Chen, 2000, Abreu, 2005 e Abreu e Rocha, 2006) e ilustrada a estimação dos parâmetros através de dois exemplos de aplicação a dados reais.

2. Modelos de cura

Quando é admissível a existência de indivíduos curados na população, os modelos de cura, caracterizados por uma função de sobrevivência imprópria, constituem uma alternativa adequada aos modelos mais clássicos da Análise de Sobrevivência. Estes modelos dividem-se em dois grupos: os modelos de cura de mistura e os modelos de cura de não-mistura. Habitualmente designam-se apenas por modelos de cura, sempre que seja claro no contexto qual o tipo que está a ser considerado.

2.1 Modelo de cura de mistura

Sejam T uma variável aleatória (v.a.) que representa o tempo de vida de um indivíduo numa dada população e Y uma v.a. binária, onde Y=1 se o indivíduo é suscetível e Y=0 se o indivíduo é imune. Seja $S(t|Y=1)$ a função de sobrevivência (f.s.) do tempo de vida dos indivíduos suscetíveis e p a proporção de indivíduos imunes na população.
O modelo de cura definido à custa da f.s. é dado por

$$S(t) = p + (1-p)S(t \mid Y=1)$$

e à custa da função de risco é

$$h(t) = \frac{(1-p)f(t \mid Y=1)}{S(t)}.$$

Este caso particular dos modelos de mistura, designado por modelo de mistura não padrão, refere-se ao caso de uma mistura de duas distribuições, uma das quais degenerada. Note-se que $S(t)$ é uma f.s. imprópria pois $S(\infty)=p$, enquanto que $S(t|Y=1)$ é uma f.s. própria. Por outro lado, $h(t)$ tem integral finito em $(0, \infty)$, logo não se trata de uma função de risco no sentido usual.
Neste tipo de modelos de cura interessa estimar não só os parâmetros da função de distribuição dos indivíduos suscetíveis, como também a proporção de indivíduos imunes. Maller e Zhou (1992) propuseram uma estimativa não paramétrica desta proporção, que é obtida através do valor da estimativa de Kaplan-Meier da f.s. calculada na maior observação.

2.2 Modelo de cura de não-mistura

Uma vez que num modelo de cura a f.s. populacional é uma função imprópria e tendo em conta a relação existente entre a f.s. e a correspondente função de risco cumulativa, esta última vai ser uma

função limitada superiormente. Assim, seja $S(t)$ a f.s. de T e $H(t)$ é tal que $S(t)=\exp[-H(t)]$. Se $S(\infty)>0$, existe $\theta<\infty$, tal que $H(\infty)=\theta$. De acordo com Yakovlev *et al*. (1993), uma forma possível de caracterizar esta propriedade é considerar $H(t)=\theta F(t)$, onde $F(t)$ designa a função de distribuição (própria) de uma v.a. não negativa.
Deste modo, o modelo de cura de não-mistura pode ser escrito na forma

$$S(t)=\exp\left[-\theta F(t)\right],$$

sendo a correspondente função de risco dada por

$$h(t)=\theta f(t),$$

onde $f(t)$ é a função densidade correspondente a $F(t)$. Neste caso, $S(\infty)=e^{-\theta}$ corresponde à probabilidade de cura.
Neste contexto, existem duas hipóteses para obter uma f.s. imprópria. Ou se modifica o espaço de parâmetros de alguma distribuição própria como acontece, por exemplo, com a distribuição de Gompertz modificada (Cantor e Shuster, 1992), ou se considera uma variável fragilidade para caracterizar a suscetibilidade de cada indivíduo ao acontecimento de interesse. Nesta última situação, é a escolha da distribuição da fragilidade e não da distribuição do tempo de vida que vai dar origem a uma função de sobrevivência imprópria. Um exemplo deste caso é a distribuição Qui-quadrado não central com zero graus de liberdade (Rocha, 1995).

3. Modelo de cura baseado na distribuição de Chen

Esta secção é constituída por quatro partes. A primeira apresenta a distribuição de Chen, uma vez que ainda é algo desconhecida. Em seguida, surge o modelo de cura com base nesta distribuição, após o que é indicado a metodologia usada para a estimação dos parâmetros. Por último, é efetuada uma aplicação a dados reais.

3.1 Distribuição de Chen

A função de distribuição proposta por Chen (2000) é

$$F(t)=1-\exp\left[\lambda\left(1-\exp\left(t^{\beta}\right)\right)\right], \quad t>0,$$

onde λ e β são os parâmetros da distribuição, sendo λ o parâmetro de escala e β o parâmetro de forma. Assim, as correspondentes f.s. e função de risco são, respetivamente,

$$\overline{F}(t)=\exp\left[\lambda\left(1-\exp\left(t^{\beta}\right)\right)\right], \quad t>0 \qquad \text{e} \qquad h^{*}(t)=\lambda\beta t^{\beta-1}\exp\left(t^{\beta}\right) \quad t>0.$$

A função $h^{*}(t)$ é crescente quando $\beta\geq 1$ e, para $\beta<1$, é decrescente para $t \in \left]0,\left(\frac{1}{\beta}-1\right)^{\frac{1}{\beta}}\right[$ e é crescente para $t\geq\left(\frac{1}{\beta}-1\right)^{\frac{1}{\beta}}$.

3.2 Modelo de cura baseado na distribuição de Chen

Considere-se a distribuição de Chen para a distribuição do tempo de vida dos indivíduos suscetíveis. O modelo de cura, definido à custa da f.s. é dado por

$$S(t) = p + (1-p)\exp\left[\lambda\left(1-\exp\left(t^{\beta}\right)\right)\right], \quad t > 0.$$

Se o modelo for escrito à custa da função de risco, tem-se

$$h(t) = \frac{(1-p)\lambda\beta t^{\beta-1}\exp\left(t^{\beta}\right)\exp\left[\lambda\left(1-\exp\left(t^{\beta}\right)\right)\right]}{p + (1-p)\exp\left[\lambda\left(1-\exp\left(t^{\beta}\right)\right)\right]}.$$

Importa notar que, embora o parâmetro β continue a ser um parâmetro de forma do modelo populacional, o comportamento da função de risco populacional é completamente diferente do da função de risco dos indivíduos suscetíveis: $h(t)$ é decrescente para $\beta \leq 1$ e é unimodal para $\beta > 1$. Além disso, no caso em que $\beta > 1$, o valor da moda será tanto maior quanto menor for o valor de p, o que faz todo o sentido uma vez que p representa a proporção de indivíduos imunes. Relativamente ao parâmetro λ, de certa forma vai definir o valor de t a partir do qual a função de risco irá estabilizar num valor próximo de zero, ou seja, neste contexto, o valor de t a partir do qual se pode esperar que os indivíduos que sobrevivam até esse instante estejam curados. Mais especificamente, quanto mais pequeno for o valor de λ, maior o valor de t a partir da qual a função de risco é aproximadamente zero.

3.3 Estimação dos parâmetros

Caso geral

Recorde-se que, na Análise de Sobrevivência, se for admitido censura não informativa e censura à direita, a função de verosimilhança para uma amostra de dimensão n pode ser expressa por

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i},$$

onde δ_i é uma variável indicatriz tal que $\delta_i = 1$ se t_i corresponde a um tempo de vida e $\delta_i = 0$ se t_i é uma observação censurada.

Considere-se uma amostra de dimensão n e designe-se por $t_1, \ldots, t_n$ os tempos de vida. Sem perda de generalidade, suponha-se que os primeiros m ($m<n$) tempos de vida são censurados. Sejam $\delta_1, \ldots, \delta_n$ tais que

$$\delta_i = \begin{cases} 0 & \text{se} \quad 1 \leq i \leq m \\ 1 & \text{se} \quad m+1 \leq i \leq n \end{cases}$$

e $y_1, \ldots, y_n$ tais que

$$y_i = \begin{cases} 0 & \text{se o indivíduo é imune} \\ 1 & \text{se o indivíduo é suscetível} \end{cases}$$

Tendo em conta os valores possíveis que o par (δ_i, y_i) pode assumir, o tipo de contribuição para a verosimilhança da respetiva observação varia, como se pode observar na Tabela 1.

Tabela 1: Resumo das possibilidades para (δ_i, y_i).

Valores possíveis para (δ_i, y_i)	Descrição da situação	Contribuição para a verosimilhança
(0, 0)	Censurado, imune	p
(0, 1)	Censurado, suscetível	$(1-p)S(t\|Y=1)$
(1, 1)	Observado, suscetível	$(1-p)f(t\|Y=1)$

No caso do modelo de cura de mistura, a verosimilhança total observada é

$$L_O = \prod_{i=1}^{n} \left[(1-p) f(t \mid Y=1)\right]^{\delta_i} \left[p + (1-p) S(t \mid Y=1)\right]^{1-\delta_i}.$$

Se todos os y_i's fossem observados (em geral, não o são aqueles que correspondem a observações censuradas), pela Tabela 1 ter-se-ia a seguinte verosimilhança completa

$$L_C = \prod_{i=1}^{n} \left[\left[(1-p) f(t \mid Y=1)\right]^{y_i} \right]^{\delta_i} \left[p^{1-y_i} \left[(1-p) S(t \mid Y=1)\right]^{y_i} \right]^{1-\delta_i}.$$

Tendo em conta que $f(t \mid Y=1) = h(t \mid Y=1) S(t \mid Y=1)$ e que, dados os valores possíveis para (δ_i, y_i) tem-se $(1-\delta_i)(1-y_i) = 1-y_i$, a verosimilhança anterior pode ser fatorizada em

$$L_C = \prod_{i=1}^{n} (1-p)^{y_i} p^{1-y_i} \prod_{i=1}^{n} h(t \mid Y=1)^{y_i \delta_i} S(t \mid Y=1)^{y_i}.$$

O facto de nem todos os y_i's serem observados origina uma situação de dados incompletos, pelo que o algoritmo EM é um método iterativo adequado para obter as estimativas de máxima verosimilhança dos parâmetros da distribuição.

A etapa E deste algoritmo consiste em determinar o valor esperado do logaritmo da verosimilhança completa em relação à distribuição dos Y's não observados, condicional aos valores atuais dos parâmetros e aos dados observados $\mathcal{O}$, onde $\mathcal{O} = \{y_i$ observado, (t_i, y_i), $i=1, ..., n\}$. No entanto, como em relação às observações censuradas o logaritmo de L_C é linear em Y, para calcular o valor esperado de log L_C basta substituir, na verosimilhança completa, os valores não observados de Y pelos respectivos valores esperados, denotados por τ_i. Então, tem-se:

$$\tau_i = E(Y \mid \mathcal{O}) = P(Y_i = 1 \mid T_i > t_i, \delta_i = 0) = \frac{(1-p) S(t_i \mid Y_i = 1)}{S(t_i)}.$$

Assim, na verosimilhança completa, cada y_i é substituído por ω_i, onde ω_i é definido na forma que se segue:

$$\omega_i = \begin{cases} 1 & \text{se} \quad \delta_i = 1 \\ \tau_i & \text{se} \quad \delta_i = 0 \end{cases}.$$

Na etapa E, cada observação censurada i é atribuída à subpopulação $Y=1$ com probabilidade τ_i e à subpopulação $Y=0$ com probabilidade $1-\tau_i$.

Depois de substituir y_i por τ_i na verosimilhança completa, obtém-se a seguinte verosimilhança "esperada"

$$L_E = \prod_{i=1}^{n} q^{\omega_i} (1-q)^{1-\omega_i} \prod_{i=1}^{n} h(t \mid Y=1)^{\omega_i \delta_i} S(t \mid Y=1)^{\omega_i} = L_{E_1} L_{E_2},$$

onde $q=1-p$.

Para efetuar a etapa M, é necessário maximizar as duas componentes do logaritmo da função de verosimilhança esperada para, se possível, obter as expressões dos estimadores dos parâmetros. Relativamente ao parâmetro q, a expressão do seu estimador é dada por

$$\hat{q} = \frac{1}{n} \left[(n-m) + \sum_{i=1}^{m} \tau_i \right].$$

Em termos gerais, a estratégia a adotar para a implementação do algoritmo EM, resume-se a:

1. determinar os valores iniciais;
2. calcular ω_i com base nos valores atuais dos parâmetros;
3. maximizar $\log L_{E_1}$ e $\log L_{E_2}$ com base nos valores atuais de ω_i e dos parâmetros;
4. repetir 2. e 3. até à convergência.

Caso da distribuição de Chen

De acordo com a definição dos valores esperados dos valores não observados de *Y*, no caso de o tempo de vida dos indivíduos suscetíveis ter uma distribuição de Chen, tem-se

$$\tau_i = \frac{q\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]}{1-q+q\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]}.$$

Assim, como $\omega_i = 1$ se $\delta_i = 1$ e $\omega_i = \tau_i$ se $\delta_i = 0$, a referida verosimilhança "esperada" é

$$L_E = \prod_{i=1}^{n} q^{\omega_i}(1-q)^{1-\omega_i}\prod_{i=1}^{n}\left[\lambda\beta t_i^{\beta-1}\exp\left(t_i^{\beta}\right)\right]^{\omega_i\delta_i}\left[\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]\right]^{\omega_i} = L_{E_1}L_{E_2}$$

e os logaritmos de L_{E_1} e L_{E_2} são, respetivamente,

$$\log L_{E_1} = (n-m)\log q + m\log(1-q) + (\log q - \log(1-q))\sum_{i=1}^{m}\tau_i$$

e

$$\log L_{E_2} = \lambda\sum_{i=1}^{m}\tau_i\left[1-\exp\left(t_i^{\beta}\right)\right] + (n-m)(\log\lambda + \log\beta) + (\beta-1)\sum_{i=m+1}^{n}\log t_i + \\ + \sum_{i=m+1}^{n} t_i^{\beta} + \lambda\sum_{i=m+1}^{n}\left[1-\exp\left(t_i^{\beta}\right)\right],$$

tendo em conta que as *m* observações censuradas estão indexadas de 1 a *m* e os *n-m* tempos de vida observados estão indexados de *m*+1 a *n*.
Após derivar $\log L_{E_2}$, é possível obter uma expressão explícita apenas para o estimador de λ

$$\hat{\lambda} = \frac{n-m}{\sum_{i=1}^{m}\tau_i\left[\exp\left(t_i^{\beta}\right)-1\right] + \sum_{i=m+1}^{n}\left[\exp\left(t_i^{\beta}\right)-1\right]}$$

Assim, a estimativa de β é obtida recorrendo ao método de Newton-Raphson.
Para proceder à estimação dos parâmetros, ou seja, para usar o algoritmo EM, é necessário substituir τ_i pelo respectivo valor, donde obtém-se

$$\hat{q} = \frac{1}{n}\left[(n-m) + q\sum_{i=1}^{m}\frac{\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]}{1-q+q\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]}\right],$$

$$\hat{\lambda} = \frac{n-m}{q\sum_{i=1}^{m}\frac{\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]}{1-q+q\exp\left[\lambda\left(1-\exp\left(t_i^{\beta}\right)\right)\right]}\left[\exp\left(t_i^{\beta}\right)-1\right] + \sum_{i=m+1}^{n}\left[\exp\left(t_i^{\beta}\right)-1\right]}.$$

3.4 Aplicação a dados reais

A metodologia anterior é ilustrada com a aplicação a dois conjuntos de dados reais, que podem ser encontrados na literatura: Maller e Zhou (1996, p.82) e Klein e Moeschberger (1998, p.10). Os ajustamentos foram efetuados recorrendo a algoritmos programados na linguagem R (R, 2010).
O primeiro desafio que surge diz respeito à obtenção dos valores iniciais para os parâmetros. Note-se que existem dois tipos de parâmetros: o parâmetro q e os parâmetros da distribuição de Chen, logo duas abordagens distintas. Para q, com base em Maller e Zhou (1992), considera-se a estimativa da proporção de suscetíveis $q^{(0)} = 1 - \hat{S}\left(t_{(r)}\right)$, onde $t_{(r)}$ representa a maior observação não censurada e $\hat{S}(t)$ a estimativa de Kaplan-Meier da f.s. Para os parâmetros da distribuição, sugere-se duas opções:

1. efetuar um ajustamento apenas com os dados com valores até $t_{(r)}$ e as estimativas obtidas constituem os valores iniciais; este método requer a programação de um algoritmo para a distribuição de Chen;
2. determinar o valor da estimativa de Kaplan-Meier da f.s. em dois pontos afastados e resolver em ordem a λ para dois valores fixos de β; o valor de β que origine valores de λ mais próximos é a estimativa inicial de β; a estimativa inicial de λ é a média dos dois valores próximos.

Dados de leucemia (Maller e Zhou, 1996)

Num estudo efetuado ao longo de 5 anos, Kersey *et al*. (1987) compararam o efeito do tratamento da leucemia linfoblástica aguda através de transplante de medula óssea em 90 doentes, 46 dos quais com a sua própria medula (grupo 2) e 44 com medula de algum parente compatível (grupo 1). O tempo de vida foi definido como o tempo, em anos, desde o transplante até à recaída.
Ambos os grupos apresentam uma considerável quantidade de observações censuradas para além do maior tempo de vida observado. Como tal, optou-se por exemplificar o ajustamento apenas para o grupo 1. A partir dos valores iniciais dos parâmetros $q^{(0)}$=0.73662, $\lambda^{(0)}$=0.9930455 e $\beta^{(0)}$=1, obteve-se as estimativas que se seguem $\hat{q} = 0.7285978$, $\hat{\lambda} = 0{,}7611151$ e $\hat{\beta} = 0.6139747$, as quais deram origem à curva representada na Figura 2. Nesta mesma figura, encontra-se o ajustamento com o modelo de cura baseado noutras distribuições, nomeadamente a distribuição de Weibull (Maller e Zhou, 1996, p.108) e a distribuição de Burr de tipo XII (Shao e Zhou, 2004). Deste modo, é evidente que, para estes dados, o modelo de cura é adequado e que o modelo proposto (Abreu, 2005 e Abreu e Rocha, 2006) é uma boa alternativa aos modelos já existentes.

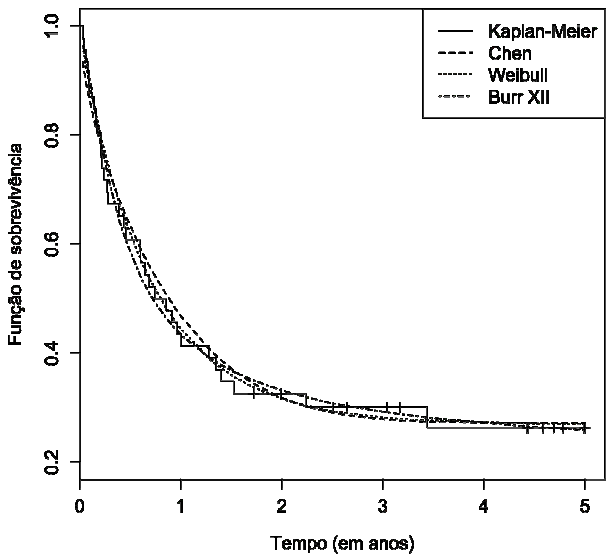


Figura 2: Dados de leucemia (Maller e Zhou, 1996, p.82)

Outros dados de leucemia (Klein e Moeschberger, 1998)

Os autores referem um estudo envolvendo 101 pacientes com leucemia mielógena aguda avançada. A todos estes pacientes foi feito transplante de medula óssea, 51 com a sua própria medula (os do grupo designado por *auto*) e 50 com medula de algum parente compatível (os do grupo designado por *allo*). Como o grupo *allo* tem uma apreciável quantidade de observações censuradas correspondentes aos valores mais elevados das observações, apresenta boas perspetivas de conter indivíduos imunes, daí que o ajustamento é realizado apenas para este grupo. De facto, o nivelamento da estimativa de Kaplan-Meier da f.s. ocorre após aproximadamente 20 meses (Figuras 1 e 3), num valor próximo de 0.5, o que é uma forte indicação da provável existência de indivíduos imunes.
Os valores iniciais dos parâmetros são $q^{(0)}$=0.46786, $\lambda^{(0)}$=0.183395 e $\beta^{(0)}$=0.5, dos quais resultaram as seguintes estimativas: $\hat{q}$ = 0.47060, $\hat{\lambda}$ = 0.12508, e $\hat{\beta}$ = 0.42100. A curva resultante está patente na Figura 3, onde se pode notar um ajustamento quase perfeito.

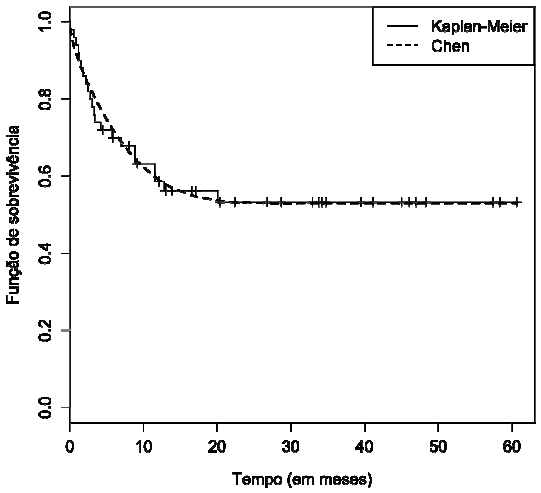


Figura 3: Dados de leucemia (Klein e Moeschberger, 1998, p.10)

4. Comentários finais

O desenvolvimento da medicina, nomeadamente no que diz respeito à cura de doenças do foro oncológico, tem originado um forte interesse na aplicação de modelos de cura, em detrimento ou como complemento dos modelos mais tradicionais. Assim, o foco passou a estar não só na estimação do tempo de vida dos indivíduos (onde se inclui o efeito das covariáveis relevantes) mas também na probabilidade de cura dos pacientes. Esta informação adicional é importante pois pode auxiliar os oncologistas na decisão da terapêutica a aplicar a cada doente; eventualmente, indivíduos com maior probabilidade de cura podem não ser sujeitos a tratamentos tão agressivos (Hunsberger *et al*., 2009).
Importa notar que os modelos de cura já estão numa fase de desenvolvimento mais avançada do que a aqui apresentada. Por exemplo, no que diz respeito aos modelos de cura de mistura, permitem a inclusão de covariáveis quer na f.s. dos indivíduos suscetíveis, quer na probabilidade de cura.

Referências Bibliográficas

Abreu, A.M. (2005). *Modelos de Sobrevivência para Populações com Indivíduos Imunes*. Dissertação de doutoramento. Universidade da Madeira.

Abreu, A.M. e Rocha, C.S. (2006). *Um novo modelo de cura paramétrico*. Em *Ciência Estatística*. Editores: Castro, L.C., Martins, E.G., Rocha, C., Oliveira, M.F., Leal, M.M., Rosado, F., p. 151-162. Edições SPE, Lisboa

Betensky, R.A. e Schoenfeld, D.A. (2001). Nonparametric estimation in a cure model with random cure times. *Biometrics*, Vol.57, p. 282-286.

Cantor, A.B. e Shuster, J.J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, Vol.11, p. 931-937.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtube shape or increasing failure rate function. *Statistics and Probability Letters*, Vol.49, p. 155-161.

Hunsberger, S., Albert, P.S. e London, W.B. (2009). A finite mixture survival model to characterize risk groups of neuroblastoma. *Statistics in Medicine*, Vol.28, p. 1301-1314.

Kersey, J.H., Weisdorf, D., Nesbit, M.E., Lebien, T.W., Woods, G.W., McGlave, P.B., Kim, T., Vallera, D.A., Goldman, A.I., Bostrom, B., Hurd, D. e Ramsay, N.K.C. (1987). Comparison of Autologous and Allogeneic Boné Marrow Transplantation for Treatment of High-Risk Refractory Acute Lymphoblastic Leukemia. The New England Journal of Medicine, Vol.317, p. 461-467.

Klein, J.P. e Moeschberger, M.L. (1998). *Survival Analysis. Techniques for Censored and Truncated Data*. 2ª impressão corrigida. Springer-Verlag, New York.

Laska, E.M. e Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, Vol.48, p. 1223-1234.

Maller, R.A. e Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, Vol.79, p. 731-739.

Maller, R.A. e Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*. John Wiley, New York.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (em 16/7/2011).

Rocha, C.S. (1995). *Modelos com fragilidade em Análise de Sobrevivência*. Dissertação de doutoramento. Faculdade de Ciências da Universidade de Lisboa.

Shao, Q. e Zhou, X. (2004). A new parametric model for survival data with long-term survivors. *Statistics in Medicine*, Vol.23, p. 3525-3543.

Tamura, R.N., Faries, D.E. e Feng, J. (2000). Comparing time to onset of response in antidepressant clinical trials using the cure model and the Cramer-von Mises test. *Statistics in Medicine*, Vol.19, p. 2169-2184.

Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, Vol. 51, p. 899-907.

Yakovlev, A.Y., Asselain, B., Bardou, V.J., Fourquet, A., Hoang, R., Rochefediere, A. e Tsodikov, A.D. (1993). *A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer*. Em *Biometrie et Analyse de Donnees Spatio-Temporelles*. Editores: Asselain, B., Boniface, M., Duby, C., Lopez, C., Masson, J.P.E Tranchefort, J., Vol.12, p. 66-82. Societé Française de Biométrie, ENSA Rennes, France.

O estimador de Kaplan-Meier: Novos desenvolvimentos e aplicações no contexto da análise de sobrevivência multiestado

Luís F. Meira Machado, *lmachado@math.uminho.pt*

Departamento de Matemática e Aplicações
Universidade do Minho

1. Introdução

A estimação da função de sobrevivência é um tópico que tem recebido muita atenção, quer na literatura estatística, quer na investigação médica. O estimador produto-limite de Kaplan-Meier tem sido considerado como um método padrão para obter estas estimativas. Este estimador, não assume nenhuma suposição sobre a distribuição de probabilidade do tempo de vida, razão pela qual é referido como um estimador não-paramétrico. A representação mais conhecida do estimador de Kaplan-Meier é baseada num produto de probabilidades elementares cuja ideia subjacente é a computação de probabilidades de sobrevivência condicionais. Neste trabalho, apresentaremos representações alternativas deste estimador, bem como, aplicações e vantagens da sua utilização. Uma destas representações consiste na definição do estimador como uma soma de pesos para ponderar os dados; numa outra abordagem, o estimador de Kaplan-Meier pode ainda ser representado como uma média ponderada de termos identicamente distribuídos, onde os pesos são obtidos por utilização de probabilidade inversa de censura. Apresentaremos novos desenvolvimentos e algumas aplicações destas abordagens no contexto dos modelos de multiestado, mais concretamente na estimação das probabilidades de transição.

A análise de dados censurados tem várias aplicações em estudos médicos longitudinais. Neste tipo de dados é frequente dispor de observações incompletas, devido à perda de *follow-up*, abandono do estudo, etc. O estimador produto-limite de Kaplan-Meier (Kaplan e Meier 1958) foi desde sempre considerado como um método padrão para obter sínteses estatísticas para esses dados. Este método pode ter em consideração vários tipos de censura, o que pode ser considerado como uma vantagem para a sua utilização. Além disso, ele está implementado em quase todos os *softwares* estatísticos tais como R, SPSS, STATA, etc. A livraria "*survival*" do *software* estatístico R é provavelmente a função mais utilizada para implementar o método de Kaplan-Meier. Na Secção 2 descrevemos o estimador de Kaplan-Meier usando representações alternativas. Numa destas representações, o estimador é apresentado na forma de pesos Kaplan-Meier que, como demonstraremos, é conveniente para introduzir 'pré-suavização' (veja-se a ideia subjacente em baixo ou veja-se o trabalho de Dikta (1998) para mais detalhes). Num trabalho recente, Meira-Machado, de Uña-Álvarez e Cadarso-Suárez (2006) apresentaram um estimador para as probabilidades de transição no caso de um modelo doença-morte não Markoviano. A ideia subjacente é a utilização de pesos Kaplan-Meier relativos à distribuição do tempo total para ponderar os dados. Posteriormente, foi proposta uma modificação deste estimador com base no princípio da utilização de pré-suavização (Amorim, de Uña-Álvarez e Meira-Machado 2011).

O estimador de Kaplan-Meier pode ser generalizado a processos não homogéneos de Markov com um número finito de estados. Essa generalização foi considerada por Aalen (1978) para o modelo de riscos competitivos e de forma independente por Aalen e Johansen (1978) e Fleming (1978) para o caso geral. Aalen e Johansen (1978) mostram como o chamado estimador de Aalen-Johansen pode ser visto

como uma versão matricial do estimador de Kaplan-Meier. Esta relação permite apresentar uma versão pré-suavizada do estimador de Aalen-Johansen para as probabilidades de transição.
Na Secção 2, o estimador de Kaplan-Meier é introduzido e novas contribuições com diferentes representações são apresentadas e posteriormente utilizadas na Secção 3, onde apresentamos novos estimadores para as probabilidades de transição no âmbito dos modelos de multiestado.

2. O estimador de Kaplan-Meier

A representação mais conhecida do estimador de Kaplan-Meier é baseada num produto de probabilidades elementares, passando de valores elevados para valores baixos de acordo com as observações detectadas. Esta representação é discutida na maioria dos livros de análise de sobrevivência (veja-se, por exemplo, Kalbfleisch e Prentice 1980; Cox e Oakes 1984). Consideremos que dispomos de dados de sobrevivência e que o evento de interesse é a morte. Então, a expressão do estimador de Kaplan-Meier envolve o cálculo do número de indivíduos que falecem em determinado momento, dividido pelo número de indivíduos em risco (ou seja, que ainda estavam vivos e no estudo antes do tempo observado). Suponha-se que temos uma amostra de *n* indivíduos de uma população homogénea. Considere-se também que o tempo de sobrevivência desses indivíduos, *T*, está potencialmente censurado (que assumimos independente do processo). Denotemos os tempos dos eventos por $t_1 < t_2 < \cdots < t_k$; e seja d_i o número de indivíduos que falecem no tempo t_i, e r_i o número de indivíduos em risco justamente antes desse tempo. A ideia subjacente ao estimador de Kaplan-Meier é a computação das probabilidades de sobrevivência condicionais, $P(T \geq t_i | T \geq t_{i-1}) = 1 - d_i/r_i$. Então, o estimador de sobrevivência Kaplan-Meier é a probabilidade incondicional de sobrevivência que é justamente o produto cumulativo das probabilidades condicionais

$$\hat{S}^{KM}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

Quando não existir censura ou truncatura, o estimador de Kaplan-Meier é equivalente ao estimador empírico.

Novas representações

O estimador de Kaplan-Meier também pode ser representado como uma soma de pesos (pesos Kaplan-Meier). Os pesos, correspondem aos 'saltos' da distribuição do estimador estando definidos apenas para os tempos de evento. Estes pesos podem ser obtidos pela expressão $W_i = \hat{S}^{KM}(t_{i-1:k}) - \hat{S}^{KM}(t_{i:k})$, para $i = 2, \ldots, k$, com $W_1 = 1 - \hat{S}^{KM}(t_{1:k})$ e onde $t_{1:k} \leq \cdots \leq t_{k:k}$ são as estatísticas ordinais obtidas a partir dos tempos dos eventos. Em seguida, introduziremos uma notação alternativa que nos permite obter uma expressão empírica dos pesos Kaplan-Meier.
Seja *T* o tempo de sobrevivência que assumimos sujeito a censura aleatória pela direita e denotemos por *C* a variável de censura, que assumimos independente de *T*. Devido à censura, em vez de *T*, observamos o par (Y, Δ), onde $Y = min(T, C)$ e $\Delta = I(T \leq C)$ é o indicador de censura. Sejam (Y_i, Δ_i), $1 \leq i \leq n$, observações independentes e identicamente distribuídas com a mesma distribuição que (Y, Δ). Seja W_i o peso Kaplan-Meier associado a Y_i quando estimamos a distribuição marginal de *Y* a partir dos pares (Y_i, Δ_i)'s. Ou seja,

$$W_i = \frac{\Delta_{[i:n]}}{n - i + 1} \prod_{j=1}^{i-1} \left(1 - \frac{\Delta_{[j:n]}}{n - j + 1}\right)$$

é o peso associado a $Y_{i:n}$, onde $Y_{1:n} \leq \cdots \leq Y_{n:n}$ são as estatísticas ordinais obtidas a partir dos tempos Y_i's e onde $\Delta_{[i:n]}$ denota o i-ésimo concomitante. Os empates entre as observações censuradas ou entre os eventos são ordenados arbitrariamente. Os empates entre os eventos e os tempos de censura são tratados de modo que os primeiros (eventos) precedem os últimos (censura).
No caso de não termos censura temos $W_i = 1/n$ para cada *i*. O estimador da função de sobrevivência de Kaplan-Meier pode então ser definido como

$$\hat{S}^{KM}(t) = 1 - \sum_{i=1}^{n} W_i I(Y_i \leq t)$$

Como se pode ver facilmente, o estimador é uma função em escada com pontos de salto localizados nos tempos de evento. O tamanho do 'salto' é uma função não decrescente. Como teremos oportunidade de verificar em seguida, esta forma de representação do estimador de Kaplan-Meier é conveniente para a introdução de 'pré-suavização'. Pré-suavização foi usada anteriormente pelo menos por Dikta (1998) e por vários autores mais tarde, e a ideia subjacente consiste em substituir cada indicador de censura por um ajuste suave de um modelo de regressão com resposta binária. Esta substituição resulta numa redução da variabilidade do estimador.

Seja $m(y) = P(\Delta = 1|Y = y)$, a probabilidade condicional de ser observado o evento dado $Y = y$. Esta função pode ser estimada paramétricamente (por exemplo, recorrendo ao modelo de regressão logístico) ou não-paramétricamente (utilizando funções do tipo núcleo ou splines). O estimador pré-suavizado da função de sobrevivência é simplesmente uma versão modificada do estimador de Kaplan-Meier com pesos suavizados que são obtidos por substituição das variáveis indicadoras de censura. O termo 'pré-suavização' vem do facto de que a suavização é apenas utilizada para obter uma versão modificada do estimador de Kaplan-Meier, mas a estimativa não é suavizada. O novo estimador é dado por

$$\hat{S}^{PKM}(t) = 1 - \sum_{i=1}^{n} {W_i}^{PKM} I(Y_i \leq t)$$

onde

$${W_i}^{PKM} = \frac{m_n(Y_{i:n})}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{m_n(Y_{j:n})}{n-j+1}\right)$$

onde $m_n(y)$ é um estimador de $m(y)$ baseado nos pares (Y_i, Δ_i)'s. Esta função pode ser estimada recorrendo a um estimador paramétrico (Dikta 1998) ou recorrendo ao estimador do tipo núcleo de Nadaraya-Watson (Cao 2005). Contribuições recentes recorrendo a pré-suavização paramétrica (de Uña-Álvarez e Rodríguez-Campos 2004; Iglesias-Pérez, de Uña-Álvarez 2008) sugerem que o estimador resultante tem uma menor variabilidade na estimação, em particular na cauda à direita, quando comparado com o estimador original (não-paramétrico). Uma abordagem usual consiste em assumir que esta função será estimada utilizando um modelo de regressão logístico. Assim, $m_n(y) = m_n(y;\beta)$ representa um modelo de regressão logístico onde β é substituido por um estimador consistente β_n, que será obtido por maximização da função de verosimilhança condicional de Δ dado Y. Deste modo, o estimador de Kaplan-Meier com pesos pré-suavizados é dado por

$${W_i}^{PKM}(\beta_n) = \frac{m_n(Y_{i:n};\beta_n)}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{m_n(Y_{j:n};\beta_n)}{n-j+1}\right)$$

O estimador de Kaplan-Meier pode ainda ser representado como uma média ponderada de termos identicamente distribuídos, onde os pesos são obtidos por utilização de ponderação por probabilidade inversa de censura (PPIC). Satten e Datta (2001) mostraram que esta representação, por meio de uma média ponderada, é conveniente para o desenvolvimento de teoria assimptótica e leva a uma decomposição da variância interessante do estimador de Kaplan-Meier. O estimador resultante escreve-se

$$\hat{S}^{KM-PPIC}(t) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i \leq t)\Delta_i}{\hat{G}(Y_i^-)}$$

onde $P(Y \geq y) \equiv G(y^-)$ e $\hat{G}$ denota o estimador de Kaplan-Meier da variável de censura C que é baseado nos pares $(Y_i, 1-\Delta_i)$'s; ou seja, $\hat{G}$ é o estimador da função de sobrevivência para dados censurados usando $\hat{S}^{KM}(\cdot)$ mas considerando tempos de falha como observações "censuradas" e tempos censurados como "falhas".

3. Estimadores para as probabilidades de transição

Em estudos longitudinais médicos, os doentes podem observar vários eventos num determinado período de *follow-up*. A análise destes estudos pode ser realizada com sucesso pelos modelos de multiestados (Andersen et al. 1993; Meira-Machado et al. 2008). Um desses modelos é o modelo de doença-morte que é totalmente caracterizado por três estados e três transições entre eles (Figura 1). Um dos objectivos principais em aplicações clínicas de modelos de multiestado é a estimação de probabilidades de transição. Estas quantidades têm proporcionado um crescente interesse pois elas permitem efectuar previsões a longo prazo do processo.

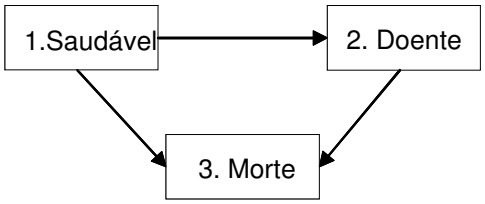


Figura 1: Modelo de Doença-morte.

Aalen e Johansen (1978) introduziram um estimador não paramétrico das probabilidades de transição para os modelos Markovianos. O estimador de Aalen-Johansen é o estimador usual para estimar as probabilidades de transição em processos de Markov. O pressuposto de Markov afirma que a evolução futura do processo é independente dos estados já visitados e dos tempos de transição entre eles, dado o estado actual do processo. Esta hipótese simplificadora permite a construção de estimadores simples. Contudo, este pressuposto é violado em algumas aplicações (veja-se, por exemplo, Andersen et al., 2000). Esta é uma observação relevante, uma vez que o estimador de Aalen-Johansen pode ser inconsistente se o processo é não Markoviano (Meira-Machado et al., 2006). Nesta secção descrevemos o estimador de Aalen-Johansen usando representações alternativas e apresentamos também estimadores alternativos que não assumem o pressuposto de Markov.

3.1 O estimador de Aalen-Johansen

Os processos de Markov não homogéneos podem ser modelados de modo não-paramétrico. Isso pode ser pensado como a generalização da abordagem de Kaplan-Meier para o modelo de mortalidade simples (modelo com dois estados e uma única transição) e foi proposta por Aalen e Johansen (1978) para modelos de multiestados gerais com um número finito de estados.

Para simplificar, consideremos o modelo de doença-morte com estados "saudável", "doente" e "morte", representado na Figura 1. Consideremos que temos uma amostra de *n* indivíduos com tempos de doença e de morte, $t_1, t_2, \dots, t_n$, eventualmente censurados. Se assumirmos que possuímos *k* eventos ("mortes") e $n-k$ observações censuradas, podemos escrever $t_{1:k} < t_{2:k} < \cdots < t_{k:k}$ para os k tempos de evento ordenados. Denotemos por r_{1i} e r_{2i} o número de indivíduos saudáveis e doentes, respectivamente, justamente antes do tempo de evento $t_{i:k}$. Denotemos por d_{12i} o número de indivíduos que adoeceram no tempo $t_{i:k}$, enquanto d_{13i} e d_{23i} representam, respectivamente, o número de indivíduos saudáveis e doentes que falecem no respectivo tempo. Então, as probabilidades de transição $p_{11}(s,t)$ e $p_{22}(s,t)$ podem ser estimadas pelos estimadores Kaplan-Meier apresentados em baixo

$$\hat{p}_{11}(s,t) = \prod_{s<t_{i:k}\leq t} \left[1 - \frac{d_{12i} + d_{13i}}{r_{1i}}\right]$$

$$\hat{p}_{22}(s,t) = \prod_{s<t_{i:k}\leq t} \left[1 - \frac{d_{23i}}{r_{2i}}\right]$$

enquanto que um estimador para $p_{12}(s,t)$ é dado por

$$\hat{p}_{12}(s,t) = \sum_{s<t_{i:k}\leq t} \hat{p}_{11}(s,t_{i:k}) \frac{d_{12i}}{r_{1i}} \hat{p}_{22}(t_{i+1:k},t)$$

é um estimador *plug-in*, obtido por substituição de $p_{11}(s,u) = p_{11}(s,u-)$ por $\hat{p}_{11}(s,u-)$, $p_{22}(u,t)$ por $\hat{p}_{22}(u,t)$ e $\alpha_{12}(u)$ por $d\hat{A}_{12}(u)$ o incremento do estimador de Nelson-Aalen $\hat{A}_{12}(u) = \sum_{t_{i:k}\leq t} \frac{d_{12i}}{n_{1i}}$ para a intensidade de doença cumulativa $A_{12}(t) = \int_0^t \alpha_{12}(u)du$.
Estes estimadores estão implementados no *software* estatístico R. As livrarias "change LOS" e "etm" podem ser utilizadas para obter estimativas das probabilidades de transição em qualquer modelo de multiestados.

Nova representação

O processo de um modelo doença-morte é completamente caracterizado pelas três transições representadas na Figura 1. Seja T_{hj} o tempo potencial de transição do estado *h* para o estado *j*. Neste modelo temos duas transições competitivas, $1 \to 2$ e $1 \to 3$. Denotemos por $\rho = I(T_{12} \leq T_{13})$ a função indicadora de transição para o estado 2 num determinado tempo e consideremos $U = min(T_{12}, T_{13})$ o tempo de permanência no estado 1. Finalmente, denotemos por $T = U + \rho T_{23}$ o tempo de sobrevivência total do processo. Contudo, devido a limitações de *follow-up*, perdas de seguimento, e por outras causa, em vez de (U,T,ρ) observamos o vector $(Z,Y,\Delta_1,\Delta,\rho)$, onde $Z = min(U,C)$ denota o tempo de permanência no estado 1 e $\Delta_1 = I(U \leq C)$ o respectivo indicador de censura; $Y = min(T,C)$ denota o tempo de vida global e $\Delta = I(T \leq C)$ o respectivo indicador de censura. Sejam $(Z_i, Y_i, \Delta_{1i}, \Delta_i, \rho_i)$, $1 \leq i \leq n$, observações independentes e identicamente distribuídas com a mesma distribuição que $(Z,Y,\Delta_1,\Delta,\rho)$. *C* denota o tempo potencial de censura que assumimos independente do processo, ou seja que *C* e (Z,Y) são independentes.
Consideremos $Z_{1:n} \leq \cdots \leq Z_{n:n}$ as estatísticas ordinais obtidas a partir dos tempos Z_i's e seja $\Delta_{1[i:n]}$ o i-ésimo concomitante. De acordo a nova notação, o estimador de Aalen-Johansen para a probabilidade de transição $p_{11}(s,t) = P(U > t | U > s)$ pode escrever-se como

$$\hat{p}_{11}(s,t) = \prod_{s<Z_{i:n}\leq t} \left[1 - \frac{\Delta_{1[i:n]}}{nM_{0n}(Z_{i:n})}\right]$$

onde $M_{0n}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \geq y)$. Similarmente, consideremos $Y_{1:n} \leq \cdots \leq Y_{n:n}$ as estatísticas ordinais obtidas a partir dos tempos Y_i's ordenados e sejam $Z_{[i:n]}$ e $\Delta_{[i:n]}$ os respectivos concomitantes. Então, a probabilidade de transição $p_{22}(s,t) = P(T > t | U \leq s, T > s)$ pode escrever-se como

$$\hat{p}_{22}(s,t) = \prod_{s<Z_{[i:n]}\leq t, Z_{[i:n]}\leq Y_{i:n}} \left[1 - \frac{\Delta_{[i:n]}}{nM_{1n}(Y_{i:n})}\right]$$

onde $M_{1n}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i < y \leq Y_i)$. Finalmente, a probabilidade de transição $p_{12}(s,t)$ pode ser estimada por métodos *plug-in*.

A introdução desta nova representação do estimador de Aalen-Johansen permite, de modo análogo ao introduzido na Secção 2, a introdução de pré-suavização no estimador de Aalen-Johansen. Os novos estimadores são obtidos por substituição das variáveis indicadoras de censura por funções que podem

ser estimadas paramétricamente ou de não-paramétricamente. Este tópico está a ser investigado e esperamos que venha a ser parte de um artigo para publicação nos próximos tempos.

3.2 Estimadores não Markovianos

Meira-Machado et al., (2006) introduziu um substituto para o estimador de Aalen-Johansen, no caso de um modelo de doença-morte para processos não Markovianos. Com este trabalho, mostraram que quando o pressuposto de Markov não é válido, o novo estimador pode comportar-se muito melhor que o de Aalen-Johansen, que pode ser sistematicamente enviesado.
Considerando que o processo de um modelo doença-morte pode ser representado pelo vector $(Z, Y, \Delta_1, \Delta, \rho)$, a probabilidade de transição $p_{11}(s,t) = P(U > t | U > s)$ pode ser estimada recorrendo aos pesos Kaplan-Meier associados a Z_i quando estimamos a distribuição marginal de Z a partir dos pares (Z_i, Δ_{1i})'s. O estimador proposto por Meira-Machado et al., (2006) é dado por

$$\hat{p}_{11}(s,t) = \frac{1 - \frac{1}{n}\sum_{i=1}^{n} W_{0i} I(Z_i \leq t)}{1 - \frac{1}{n}\sum_{i=1}^{n} W_{0i} I(Z_i \leq s)}$$

onde

$$W_{0i} = \frac{\Delta_{1[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{\Delta_{1[j:n]}}{n-j+1}\right)$$

é o peso associado a $Z_{i:n}$.
Analogamente, a probabilidade de transição $p_{22}(s,t) = P(U \leq s, T > t | U \leq s, T > s)$ pode ser estimada recorrendo aos pesos Kaplan-Meier associadas ao par (Y, Δ):

$$\hat{p}_{22}(s,t) = \frac{\frac{1}{n}\sum_{i=1}^{n} W_i I(Z_i \leq s, Y_i > t)}{\frac{1}{n}\sum_{i=1}^{n} W_i I(Z_i \leq s, Y_i > s)}$$

e $p_{12}(s,t) = P(s < U \leq t, T > t | U > s)$ pode ser estimada por

$$\hat{p}_{12}(s,t) = \frac{\frac{1}{n}\sum_{i=1}^{n} W_i I(s < Z_i \leq t, Y_i > t)}{1 - \frac{1}{n}\sum_{i=1}^{n} W_{0i} I(Z_i \leq s)}$$

onde W_i é o peso Kaplan-Meier associado a Y_i quando estimamos a distribuição marginal de *Y* a partir dos pares (Y_i, Δ_i)'s (veja-se a Secção 2).

Ao retirar a condição de Markov, o estimador proposto por Meira-Machado et al (2006) proporciona erros padrão elevados. Esse problema pode agravar-se em situações onde existe uma grande proporção de dados censurados. Para resolver este problema, Amorim et al., (2011) propôs uma modificação do estimador de Meira-Machado et al., (2006), baseados em pré-suavização. Estes autores demonstraram que a introdução de pré-suavização permite uma redução de variância, na presença de censura. Os novos estimadores são baseados em uma estimativa preliminar (pré-suavização) da probabilidade de censura para o tempo total, dadas as informações disponíveis. A ideia subjacente já foi usada antes em análise de sobrevivência univariada (veja-se a Secção 2), mas a sua aplicação em modelos de multiestados levanta alguns novos problemas. Para mais detalhes, consulte-se o artigo de Amorim et al. (2011).

4. Conclusão

Neste trabalho, foram introduzidos diferentes representações para o estimador de Kaplan-Meier. Com base numa destas representações, baseada nos pesos Kaplan-Meier, novos estimadores foram introduzidos para quantidades de interesse no contexto da análise de sobrevivência de multiestados.

Em concreto introduzimos estimadores para as probabilidades de transição, embora a mesma abordagem possa ser utilizada para estimar outras quantidades relevantes em aplicações médicas tais como a função de distribuição bivariada para tempos sequenciais censurados, a função de incidência cumulativa, etc.

Os novos estimadores para as probabilidades de transição são consistentes independentemente da condição de Markov. Este facto é interessanteporque os problemas reais são, muitas vezes não Markovianos e, portanto, a consistência do estimador de Aalen-Johansen (tradicionalmente considerado) não pode ser assegurada nessas situações.

Agradecimentos

Este trabalho recebeu apoio financeiro do Ministério Português da Ciência, Tecnologia e Ensino Superior, pelo projecto PTDC/MAT/104879/2008. A investigação também foi parcialmente financiada pela FCT e CMAT sob o programa POCI 2010.

Bibliografia

Aalen, O.O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *Annals of Statistics*, 6, 534-545.

Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5, 141-150.

Amorim, A.P., de Uña-Álvarez. J. and Meira-Machado, L. (2011). Presmoothing the transition probabilities in the illness-death model, *Statistics & Probability Letters*, 81, 797-806.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Andersen, P.K., Esbjerg, S. and Sorensen, T.I.A. (2000). Multistate models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine*, 19: 587-599.

Cao, R., López, de Ullibarri, I., Janssen, P. and Veraverbeke, N. (2005). Presmoothed Kaplan-Meier and Nelson-Aalen estimators. *Journal of Nonparametric Statistics*, 17, 31-56.

Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data, Chapman & Hall.

de Uña-Álvarez, J. and Rodríguez-Campos, C. (2004). Strong consistency of presmoothed Kaplan-Meier integrals when covariables are present. *Statistics*, 38(6), 483–496.

Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference*, 66, 253–279.

Fleming, T.R. (1978). Nonparametric estimation for nonhomogeneous Markov processes in the problem of competing risks. *Annals of Statistics*, 6, 1057-1070.

Iglesias-Pérez, M.C. and de Uña-Álvarez, J. (2008). Nonparametric estimation of the conditional distribution function in a semiparametric censorship model. *Journal of Statistical Planning and Inference*, 138, 3044-3058.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data*. John Wiley & Sons.

Meira-Machado, L., de Uña-Álvarez, J. and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 12, 325-344.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., Andersen, P.K. (2009) Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, 18, 195-222

Satten, G.A., and Datta, S. (2001). The Kaplan–Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average. *The American Statistician*, 55(3), 207-210.

Análise Bayesiana de Modelos de Sobrevivência Baseados em Processos de Contagem

Giovani Loiola da Silva, *gsilva@math.ist.utl.pt*

Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)

Departamento de Matemática – IST, Universidade Técnica de Lisboa

1. Introdução

O estudo da análise de sobrevivência centra-se num conjunto de unidades (indivíduos) que são observadas até à ocorrência de algum evento de interesse, por exemplo, a morte do indivíduo (tempo de sobrevivência). Frequentemente este evento não chega a ocorrer para algumas dessas unidades durante o período de observação (censura), o que torna a análise estatística desses dados distinta das análises usuais. Daí o surgimento da análise de sobrevivência como metodologia estatística apropriada para o estudo de unidades sujeitas a censura. Para maiores detalhes sobre análise de sobrevivência, consulte e.g. Kalbfleisch e Prentice (2002) para análise semi paramétrica, Lawless (2003) para análise paramétrica, Andersen *et al.* (1993) e Fleming e Harrington (1991) para análise baseada em processos de contagem, Ibrahim, Chen e Sinha (2001) para uma abordagem bayesiana e Silva (2001) para análise bayesiana de modelos de sobrevivência com fragilidade.

Algumas funções de interesse em análise de sobrevivência para uma população de unidades cujo tempo de sobrevivência é representado por T são a função de sobrevivência (f.s.) que descreve a forma distribucional dos tempos de sobrevivência através da probabilidade de uma unidade sobreviver pelo menos até ao instante *t*, $S(t) \equiv P(T \geq t)$, onde T é uma variável aleatória (v.a.) não negativa, a função risco, i.e., a taxa de ocorrência do evento de interesse no instante $t \in J = [0, \infty)$, definida por

$$\lambda(t) = \lim_{dt \to 0^+} P(t \leq T < t + dt | T \geq t)/dt \qquad (1)$$

e a função risco cumulativa ou integrada $\Lambda(t) = \int_o^t \lambda(u)du$, que é finita para algum t>0 e $\int_o^\infty \lambda(u)du = \infty$.

Aalen (1978) apresentou uma nova abordagem para análise de sobrevivência não paramétrica com base na teoria de processos de contagem (processos pontuais). Esse trabalho serviu de inspiração a vários modelos de sobrevivência definidos a partir de intensidades de processos de contagem quantificando ocorrências de um evento de interesse ao longo do tempo. O carácter temporal desses modelos traduz a característica dinâmica das unidades em análise de sobrevivência que se movem entre um número finito de estados. Por exemplo, a mudança de um indivíduo do estado vivo para o estado morto que é uma situação simples neste tipo de análise.

Um processo de contagem $N(t)$ é o número de eventos de um processo pontual em $[0, t]$, i.e., $N(t)$ é um processo estocástico que assume os valores 0,1,2,… registando um salto do valor $i - 1$ para o valor i quando ocorre o evento de interesse pela i-ésima vez, $i = 1,2, \ldots$. Em análise de sobrevivência, o carácter temporal dos processos de contagem requer um conhecimento das unidades em cada instante

t, $t \in J = [0, \infty)$. Este conhecimento é a história completa das unidades até ao instante t representada por H_t.
Para um processo de contagem $N(t)$ munido de uma história H_t por vezes existe o seguinte limite

$$I(t) = \lim_{dt \to 0^+} E[N(t + dt) - N(t)|H_t]/dt, \qquad (2)$$

onde $I(t)$ é chamado de processo de intensidade ou, simplesmente, intensidade do processo $N(t)$ no intervalo infinitesimal $[t, t + dt), \ \forall t \in J$.
Sejam $T_1, \dots, T_n$ v.a. contínuas que representam tempos de sobrevivência distintos de n unidades independentes sujeitas à censura e oriundas de uma população homogénea com função densidade de probabilidade (f.d.p.) $f(t)$ e f.s. $S(t)$ de suporte J. O número de ocorrências do evento de interesse até ao instante t, $t \in J$, nas n unidades é um processo de contagem, definido por $N(t) = \sum_{i=1}^{n} \mathrm{I}_{(T_i \leq t, \gamma_i = 1)}$, onde γ_i é a função indicadora de não censura da unidade i. Neste contexto, a função risco (1) pode ser apresentada heuristicamente como

$$P(t \leq T_i < t + dt, \gamma_i = 1 | H_t) = \begin{cases} \lambda(t)dt, & T_i \geq t; \\ 0 & T_i < t, \end{cases}$$

onde H_{t^-} é a história do processo estocástico $N(t)$ representando os dados disponíveis até ao instante imediatamente antes de t, $t \in J$. Na expressão acima, substituindo a probabilidade à esquerda pelo valor esperado de uma v.a. indicadora e somando em i, $i = 1, \dots, n$, obtém-se

$$E[\sum_{i=1}^{n} \mathrm{I}_{(T_i \leq t, \gamma_i = 1)} | H_t] = Y(t)\lambda(t)dt \qquad (3)$$

onde $Y(t) = \sum_{i=1}^{n} \mathrm{I}_{(T_i \geq t)}$ é o número de unidades no conjunto de risco no instante t. Note-se que a soma de variáveis indicadoras referida no valor esperado em (3) é precisamente o incremento do processo de contagem $N(t)$ em $[t, t + dt)$, denotado por $dN(t) = N((t + dt)^-) - N(t^-)$, sendo este valor esperado essencialmente a intensidade do processo $N(t)$, definida em (2). Assim, a relação (3) pode ser rescrita como

$$E(dN(t)|H_{t^-}) = I(t)dt$$

com a intensidade do processo $N(t)$ dada por

$$I(t) = Y(t)\lambda(t) \qquad (4)$$

que é conhecida como o modelo de intensidade multiplicativo de Aalen (1978).

2. Função de verosimilhança

Nesta secção faz-se uma construção (heurística) da função de verosimilhança constatando-se que ela é essencialmente a função de verosimilhança usual em análise de sobrevivência. Para populações heterogéneas o conjunto de dados de sobrevivência (usual) $D = \{(T_i, \gamma_i), i = 1, \dots, n\}$, onde T_i e γ_i, $i = 1, \dots, n$, são respectivamente os tempos de sobrevivência e as funções indicadoras da ausência de censura, é definido para cada instante $t \in J$ da seguinte forma $D = \{(N_i(t), Y_i(t)), i = 1, \dots, n\}$, onde os elementos de D são os valores observados das seguintes funções indicadoras

$$N_i(t) = \mathrm{I}_{(T_i \leq t, \gamma_i = 1)} \ \text{ e } \ Y_i(t) = \mathrm{I}_{(T_i \geq t)} \qquad (5)$$

visto que há no máximo uma ocorrência do evento de interesse para i-ésima unidade com tempo de sobrevivência T_i, $i = 1, \dots, n$. Estes processos estão relacionados com uma história H_t, que resume os dados disponíveis até ao instante t.

Considere-se um processo de contagem multivariado (vide Aalen, 1978), i.e., as componentes são os processos de contagem univariados $N_i(t)$, $i = 1, \dots, n$, definidos em (5). As intensidades individuais destes processos $I_i(t)$, $i = 1, \dots, n$, são supostas ser dependentes de um vector de parâmetros θ. Note-se que a intensidade do processo $N_i(t)$ acima está definida no intervalo infinitesimal $[t, t + dt)$ com $I_i(t) \cong P(T_i \in [t, t + dt) | T_i \geq t)$, quando $dt \to 0$. Dada uma história H_t, o incremento do processo $N_i(t)$ em $[t, t + dt)$, $dN_i(t)$, tem aproximadamente uma distribuição de Bernoulli com probabilidade $I_i(t)dt$ de ocorrer o acontecimento $dN_i(t) = 1$. Além disso, como as n unidades do conjunto de dados de sobrevivência D são estatisticamente independentes, a respectiva função de verosimilhança num dado instante t é dada por

$$L_t(\theta|D) = \prod_{i=1}^{n} I_i(t)^{dN_i(t)} \ (1 - I_i(t)dt), \qquad (6)$$

onde $f_i(.)$ e $S_i(.)$ são, respectivamente, a f.d.p. e a f.s. da i-ésima unidade, i=1,…,n, dado o vector paramétrico $\theta \in \Theta$.
Entretanto, a função de verosimilhança do processo de contagem multivariado abrange toda a informação para todos os instantes u, $u \in J$. Isso implica a "integração" da função (6) sobre o intervalo $[0, \tau)$ com $\tau \leq \infty$. O integral em causa é o integral produto que redefine a função de verosimilhança a partir de (6), sendo frequentemente apresentada na forma mais vulgar do integral produto como

$$L(\theta|D) \propto \prod_{i=1}^{n} \left[\prod_{t \geq 0} I_i(t)^{dN_i(t)}\right] \ exp\left(\left[-\int_o^t I_i(u)du\right]\right), \ \forall\, t \in J. \quad (7)$$

Note-se que a função (7) é essencialmente a função de verosimilhança de n v.a. independentes com distribuição de Poisson de médias iguais aos incrementos das funções cumulativas $\Lambda_i^*(t)$, denotados por $d\Lambda_i^*(t) = I_i(t)dt$, $i = 1, \dots, n$. Ou seja, os incrementos dos processos de contagem no intervalo infinitesimal $[t, t + dt)$, $dN_i(t)$, comportam-se como se seguissem distribuições de Poisson independentes com médias $I_i(t)dt$, $i = 1, \dots, n$.

3. Modelos de regressão

As unidades dos dados de sobrevivência são frequentemente provenientes de populações heterogéneas implicando a observação de um conjunto de covariáveis juntamente com os tempos de sobrevivência. Neste cenário, uma questão de interesse é investigar a influência das covariáveis nos tempos de sobrevivência, o que justifica a consideração de modelos de regressão para o efeito. A investigação da influência de covariáveis nos tempos de sobrevivência faz-se usualmente através da função risco. A estruturação desta função é feita sobretudo em termos multiplicativos ou aditivos, com as distribuições dos tempos de sobrevivência a pertencer a família de distribuições absolutamente contínuas indexadas por um vector de parâmetros $\theta \in \Theta$.
Dado um conjunto de dados de sobrevivência D com n unidades independentes e p covariáveis observadas, $D = \{(T\_i, \gamma_i, Z_i), i = 1, \dots, n\}$, os modelos de regressão multiplicativos definem-se pela seguinte estrutura para a função risco da i-ésima unidade de D num instante $t \in J$

$$\lambda(t|Z_i, \theta) = \lambda_0(t) \ \psi(Z_i(t)'\beta), \qquad (8)$$

onde $\theta = (\lambda_0(.), \beta)$, podendo ser algumas vezes ampliado com novas componentes inerentes à forma da função ψ, $\lambda_0(.)$ é a função risco subjacente, $\beta \in \mathcal{R}^p$ é o vector de coeficientes de regressão, $\psi(.)$ é uma função positiva escolhida usualmente como $\psi(x) = \exp(x)$ e $Z_i(t) = \left(Z_{i1}(t), \dots, Z_{ip}(t)\right)'$ é o vector de covariáveis para a unidade i que pode variar com o tempo, $i = 1, \dots, n$. O modelo multiplicativo mais conhecido é o modelo de Cox (1972), onde $\psi(.)$ é a função exponencial, $\lambda_0(.)$ é uma função não especificada e as covariáveis não dependem do tempo.
Por outro lado, quando o efeito das covariáveis é expresso aditivamente na função risco tem-se a segunda principal classe de modelos de regressão em análise de sobrevivência, conhecida por modelos aditivos. Analogamente aos modelos (8), os modelos de regressão aditivos são definidos pela seguinte função risco da unidade i de D no instante $t \in J$

$$\lambda(t|Z_i,\theta) = a_0\,(t) + \zeta(Z_i(t)'a(t))\,, \qquad (9)$$

onde $\theta = (\alpha_0\,(.), a(.))$, podendo também ser ampliado com novas componentes associadas a forma funcional de ζ, $\alpha_0(.)$ é a função risco subjacente, $\zeta(.)$ é uma função positiva escolhida usualmente como $\zeta(\mathrm{x}) = \mathrm{x}$, $\alpha(.) = \left(\alpha_1(.), \dots, \alpha_p(.)\right)'$ e $\alpha_q(.)$ é a função de regressão associada à q-ésima covariável de $Z_i = Z_i(t) = \left(Z_{i1}(t), \dots, Z_{ip}(t)\right)'$, $q = 1, \dots, p$, $i = 1, \dots, n$. O modelo de Aalen (1980) é o mais conhecido dos modelos aditivos, onde $\zeta(\mathrm{x}) = \mathrm{x}$ e $\alpha_q(t)$, $q = 0, \dots, p$, são funções não especificadas. Os modelos de sobrevivência aditivos parecem ser mais adequados do que os modelos multiplicativos quando as funções de risco são mais paralelas (riscos absolutos) do que proporcionais (riscos relativos).

No contexto de processos de contagem, estes modelos podem ser igualmente definidos pela intensidade de um processo de contagem $N_i(t)$ associado à unidade i no instante $t \in J$, como em (4),

$$I_i(t) = Y_i(t) + \lambda_i(t|Z_i(t),\theta)\,, \qquad (10)$$

onde $Y_i(t)$ é um processo predizível no instante imediatamente antes de t e $\lambda_i(t|Z_i(t),\theta)$ é uma das funções risco acima, $i = 1, \dots, n$.

Para os modelos de sobrevivência (10) considerou-se que os tempos de sobrevivência são independentes, condicionalmente aos valores das covariáveis observadas, e que existe homogeneidade entre as unidades com o mesmo valor das covariáveis observadas. Essas suposições estão usualmente subjacentes aos modelos de sobrevivência em populações heterogéneas, podendo não ser satisfeitas em algumas situações. Por exemplo, no estudo da incidência de determinadas doenças pode-se verificar uma tendência na ocorrência da doença em indivíduos de uma mesma família devido a predisposição genética (não observada) e, portanto, a hipótese de independência acima pode ser posta em causa. Os modelos que visam investigar variações não observadas entre as unidades de um conjunto de dados censurados são denominados modelos de sobrevivência com fragilidade, ou simplesmente, modelos de fragilidade.

Os modelos de fragilidade são aqui caracterizados pela introdução de um efeito aleatório (fragilidade) na função risco ou intensidade com o objectivo de controlar a heterogeneidade não observável das unidades em estudo, inclusivamente a dependência das unidades que partilham factores de risco. Vários trabalhos têm sido publicados sobre esses modelos, nomeadamente, Vaupel *et al.* (1979) com a introdução do termo *frailty*, Clayton (1991) com a primeira abordagem bayesiana desses modelos e Rocha (1995) com um amplo estudo de modelos probabilísticos com estrutura de fragilidade. Em geral, os modelos de fragilidade são extensões dos modelos de regressão multiplicativos. Andersen e tal. (1993, cap. 9) e Hougarad (1995) apresentam uma revisão dos modelos de fragilidade (multiplicativos) numa perspectiva frequencista, enquanto Silva e Amaral-Turkman (2004) elaboram uma revisão completa dos modelos de fragilidade (aditivos) sob a perspectiva bayesiana.

Para populações com p covariáveis observadas, o modelo de fragilidade pode ser uma extensão dos modelos de regressão multiplicativos com o termo de fragilidade a actuar multiplicativamente na função risco (8). Por outras palavras, a função risco de uma unidade com vector de covariáveis Z e fragilidade w é dada por

$$\lambda(t|Z,w,\theta) = \lambda_0(t)\ \psi(Z_i(t)'\beta)\ w, \qquad (11)$$

onde $\theta = (\lambda_0\,(.), \beta)$, $\lambda_0\,(.)$ é a função risco subjacente com fragilidade "nula" ($w = 1$), w é um efeito aleatório e os demais termos encontram-se definidos em (8). Os modelos de sobrevivência com função risco (11) são denominados modelos multiplicativos com fragilidade.

Os termos de fragilidade podem ser também introduzidos aditivamente e a relação entre a função risco e as p covariáveis pode ser também aditiva como nos modelos de regressão aditivos (9). Por questões de conveniência, os modelos aditivos (9) são aqui usados com $\zeta\,(\mathrm{x}) = \mathrm{x}$. Neste cenário, a junção destes modelos dá origem aos modelos aditivos com fragilidade para dados univariados, em que a função risco da i-ésima unidade de D com vector de covariáveis Z_i e fragilidade w_i é dada por

$$\lambda_i(t|Z_i, w_i, \theta) = a_0(t) + Z_i(t)'a(t) + w_i, \qquad (12)$$

onde $\theta = (\alpha_0(.), \alpha_1(.), \ldots, \alpha_p(.))'$, $\alpha_0(.)$ é a função risco subjacente com fragilidade nula e $\alpha_q(.)$ é a função de regressão associada à covariável $z_q(.)$, $q = 1, \ldots, p$, $i = 1, \ldots, n$. Observe-se que estes modelos de fragilidade são duplamente aditivos, uma vez que tanto as covariáveis observadas (Z) como a fragilidade (w) são introduzidas na intensidade (12) aditivamente.

4. Modelos de sobrevivência aditivos bayesianos

Sob a abordagem de processos de contagem, os modelos aditivos com fragilidade (12), onde as covariáveis observadas não variam com o tempo por questões de simplicidade, são redefinidos usando a intensidade do processo de contagem $N_i(t)$ associado à i-ésima unidade de D no instante $t \in J$, i.e.

$$I_i(t|W) = I_i(t|Z_i, W, \theta) = Y_i(t)[a_0(t) + Z_i(t)'a(t) + A_i''W], \qquad (13)$$

onde $Y_i(t)$ é o processo indicador se a unidade i está ou não em risco imediatamente antes do instante t, $W = (w_1, \ldots, w_k)'$ é o vector de termos de fragilidade e $Z_i = (z_{i1}, \ldots, z_{ip})'$ e $A_i = (a_{i1}, \ldots, a_{ik})'$ são, respectivamente, os vectores de covariáveis e de coeficientes de fragilidade para a unidade i com $a_{il} = 1$, se a unidade i tem fragilidade w_l, e 0 no caso contrário, $i = 1, \ldots, n$, $l = 1, \ldots, k$. Quando $\sum_{l=1}^{k} a_{il} = 1 \ \forall\, i$, os modelos de fragilidade (13) abrangem os modelos de fragilidade para dados univariados (12).
Para os modelos de fragilidade (13), a f.s. (condicional) para uma unidade, com vector de covariáveis Z e vector de coeficientes de fragilidade A, é dada por

$$S(t|W) = \exp\left[-\int_0^t (a_0(u) + Z_i' a(u) + A_i' W)du = \prod_{q=0}^{p} \exp\left[-z_q \Omega_q(t)\right] \prod_{q=0}^{p} \exp[-a_l w_l t]\right., \quad (14)$$

onde $z_0 \equiv 1$ e $\Omega_q(t) = \int_0^t \alpha_q(u)du$ é a função cumulativa de $\alpha_q(t)$, $q = 0, \ldots, p$.
Se as k fragilidades $W \in \Theta_W$ são v.a. independentes com função de distribuição conjunta $G(W)$, a f.s. marginal dos modelos de fragilidades (13) deduzida a partir de (14) é dada por

$$S(t) = \int_{\Theta_W} S(t|W)dG(W) = \prod_{q=0}^{p} \exp[-z_q \Omega_q(t)] \prod_{l=1}^{k} Q_l(a_l t), \qquad (15)$$

onde $Q_l(.)$ é a transformada de Laplace da fragilidade w_l, $l = 1, \ldots, k$.
Se os k termos de fragilidade dos modelos aditivos (13) são v.a. independentes com distribuição gama de parâmetros de forma δ_l e de escala η para o termo w_l, $l = 1, \ldots, k$, então a distribuição conjunta dos termos de fragilidade, denotada por $g(W|\delta)$, é dada por

$$\eta^{\sum_{l=1}^{k} \delta_l} \prod_{l=1}^{k} \frac{w_l^{\delta_l - 1}}{\Gamma(\delta_l)} \exp[-\eta \sum_{l=1}^{k} w_l],$$

onde $\delta = (\delta_1, \ldots, \delta_k, \eta)'$ é o parâmetro de fragilidade. Quando $\delta_l = \eta, \forall\, l = 1, \ldots, k$, o parâmetro η mede o grau de heterogeneidade não observada nas unidades dos dados de sobrevivência D. Neste caso, os valores esperados das fragilidades são finitos (iguais a um), satisfazendo assim uma das suposições de identificabilidade para os modelos multiplicativos com fragilidade.
A função de verosimilhança para os modelos aditivos com fragilidade, supondo mecanismos de censura independentes e não informativos, é dada por

$$L(\theta|D, W) \propto \prod_{i=1}^{n} \left[\prod_{t \geq 0} I_i(t|W)^{dN_i(t)}\right] exp\left(\left[-\int_o^t I_i(u|W)du\right]\right), \qquad (16)$$

onde $dN_i(t)$ é o incremento no intervalo infinitesimal $[t, t + dt)$ do processo de contagem $N_i(t)$ com intensidade $I_i(t|W)$ definida em (13), $i = 1, \ldots, n$. Supõem-se aqui que os coeficientes de fragilidade

A_i, $i = 1, \dots, n$, fazem parte dos dados D, visto que são quantidades conhecidas como os valores das covariáveis Z para cada unidade.
Perante a dimensionalidade infinita dos parâmetros dos modelos aditivos com fragilidade, opta-se aqui pela discretização dos mesmos nos intervalos disjuntos consecutivos $B_1, \dots, B_m$. Por questões de conveniência, as funções $\alpha_q(t)$, $q = 0, \dots, p$, são preteridas em relação às suas funções cumulativas $\Omega_q(t)$ e portanto os incrementos dessas funções cumulativas no intervalo $B_j = [u_{j-1}, u_j)$, $j = 1, \dots, m$, são denotados por

$$\Omega_{qj} \equiv d\Omega_{\mathrm{q}}(t) = \alpha_{\mathrm{q}}(t)dt, \ \forall \ u_{j-1} \leq t < u_j, \tag{17}$$

dando origem ao vector $\Omega = (\Omega_0, \dots, \Omega_p)'$ com $\Omega_q = (\Omega_{q1}, \dots, \Omega_{qm})'$, $q = 0, \dots, p$. Assim, esta discretização das funções $\Omega_q(t)$, $q = 0, \dots, p$, proporciona um vector paramétrico Ω com dimensão $m \times (p + 1)$ para os modelos aditivos com fragilidade (13).
Com a discretização das funções cumulativas $\Omega_q(t)$, $q = 0, \dots, p$, a verosimilhança (16) para os modelos de fragilidade (13) é reexpressa como

$$L(\theta|D, W) \propto \prod_{i=1}^{n} \prod_{j=1}^{m} {I_{ij}}^{N_{ij}} exp\left(-I_{ij}\right), \tag{18}$$

onde os incrementos do processo de contagem $N_i(t)$ e de sua intensidade no intervalo B_j são denotados, respectivamente, por $N_{ij} = dN_i(t_j)$ e

$$I_{ij} = Y_{ij}\left[Z_i' \, \Omega_{\mathrm{j}} + A_i' \, W dt_j\right], \tag{19}$$

com $Y_{ij} \equiv Y_i(t_j)$ a indicar se a unidade i pertence ao conjunto de risco no intervalo B_j, $\Omega_j = (\Omega_{0j}, \dots, \Omega_{pj})'$ e Ω_{qj}, $q = 0, \dots, p$, definidos em (17), $dt_j = u_j - u_{j-1}$, $W = (w_1, \dots., w_k)$, $Z_i = (1, z_{i1}, \dots, z_{ip})'$ e $A_i = (a_{i1}, \dots, a_{ik})'$, $i = 1, \dots, n$, $j = 1, \dots, m$. Esta verosimilhança é essencialmente a verosimilhança de um produto de v.a. independentes com distribuição de Poisson de médias I_{ij} em (19), $i = 1, \dots, n$, $j = 1, \dots, m$.
Os processos de Lévy são potenciais candidatos para os processos *a priori* das funções cumulativas discretizadas $\Omega_q(t)$, $q = 0, \dots, p$, visto que estes atendem às características destas funções sendo não negativos e não decrescentes em $t \in J$. O mais popular dos processos de Lévy para modelar a função risco subjacente é o processo gama (vide e.g. Silva, 2001). Se este processo com incrementos independentes for adoptado como processo *a priori* para cada uma das versões discretizadas das funções cumulativas $\Omega_q(t)$, $q = 0, \dots, p$,, associadas aos modelos aditivos vigentes, a distribuição *a priori* dos incrementos de uma função cumulativa $\Omega_q(t)$, definidos em (17) e denotados por $\Omega_q = (\Omega_{q1}, \dots, \Omega_{qm})'$, é dada por

$$\phi\left(\Omega_{\mathrm{q}}\right) \propto \prod_{j=1}^{m} \frac{{c_q}^{c_q \, \Omega^*_{\mathrm{qj}}}}{\Gamma(c_q \, \Omega^*_{\mathrm{qj}})} {\Omega_{\mathrm{qj}}}^{c_q \, \Omega^*_{\mathrm{qj}} - 1} \exp\left[-c_q \, \Omega_{\mathrm{qj}}\right], \tag{20}$$

onde c_q é interpretado como uma medida de precisão da conjectura inicial Ω^*_{qj} da função cumulativa $\Omega_{\mathrm{q}}(\mathrm{t})$ e $\Omega^*_{\mathrm{qj}} = \Omega^*_{\mathrm{q}}(u_j) - \Omega^*_{\mathrm{q}}(u_{j-1})$, $j = 1, \dots, m$, são as respectivas conjecturas para as componentes de Ω_{q}, $q = 0, \dots, p$. Note-se que a média e a variância de Ω_{qj}, $j = 1, \dots, m$, são dadas, respectivamente, por $\mathrm{E}\left(\Omega_{\mathrm{qj}}\right) = \Omega^*_{\mathrm{qj}}$ e $\mathrm{Var}(\Omega_{\mathrm{qj}}) = \Omega^*_{\mathrm{qj}}/\mathrm{c_q}$. Por vezes, assume-se que $\Omega^*_{\mathrm{qj}} = \mathrm{r_q} \, dt_j$, onde $\mathrm{r_q}$ é um valor proposto para a função $\alpha_{\mathrm{q}}(\mathrm{t})$ por unidade de tempo, e.g., $\mathrm{r_0}$ pode ser uma conjectura da taxa de falha por unidade de tempo, e $dt_j = u_j - u_{j-1}$, $q = 0, \dots, p, j = 1, \dots, m$. As quantidades $\mathrm{c_q}$ e $\mathrm{r_q}$, $q = 0, \dots, p$, são hiperparâmetros dos modelos vigentes possivelmente desconhecidos.
Considerando que as $p + 1$ funções cumulativas $\Omega_{\mathrm{q}}(\mathrm{t})$ são independentes *a priori*, a distribuição *a priori* dos vectores de incrementos $\Omega = (\Omega_0, \dots, \Omega_p)'$ é dada simplesmente por

$$\phi(\Omega) = \prod_{q=0}^{p} \phi(\Omega_q), \qquad (21)$$

onde a distribuição *a priori* $\phi(\Omega_q)$ pode ser a distribuição (20) se for adoptado o processo *a priori* com incrementos independentes gama para Ω_q, $q = 0, \dots, p$.
Logo, a distribuição *a posteriori* dos modelos aditivos com fragilidade é dada por

$$\pi(\Omega, W, \delta|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{m} {I_{ij}}^{N_{ij}} exp(-I_{ij}) \ \phi(\Omega)\, \phi(W|\delta)\, \phi(\delta), \qquad (22)$$

onde I_{ij} é a intensidade do processo de contagem $N_i(t)$ associado à unidade i no intervalo de tempo B_j, $i = 1, \dots, n$, $j = 1, \dots, m$, definida em (19), e $\phi(\Omega)$, $\phi(W|\delta)$ e $\phi(\delta)$ são, respectivamente, as distribuições *a priori* dos incrementos das funções cumulativas Ω em (21), dos termos de fragilidade W e dos hiperparâmetros de fragilidade δ.
A distribuição *a posteriori* conjunta (22) não permite a obtenção de distribuições *a posteriori* marginais explicitamente. Assim, as inferências sobre os parâmetros de interesse nos modelos aditivos com fragilidade (13) ficam dificultadas devido aos problemas na integração de (22) com respeito aos parâmetros perturbadores (termos de fragilidade). Contudo, as quantidades de interesse podem ser avaliadas por métodos de simulação tais como os métodos de Monte Carlo via cadeias de Markov (MCMC) (vide e.g. Paulino, Amaral-Turkman e Murteira, 2003).
Em análise de sobrevivência uma questão de interesse é conhecer a f.s. ao nível "populacional", i.e., a f.s. marginal que tem uma forma analítica, expressa em (15). Por outras palavras, a f.s. marginal *a posteriori* dos modelos aditivos com fragilidade para unidades com vectores de covariáveis $Z = (1, z_1, \dots, z_p)'$ e de coeficientes de fragilidade $A = (a_1, \dots, a_k)'$, é dada por

$$S(t|\Omega, \delta) = \exp\left[-\sum_{q=0}^{p} z_q \sum_{j:u_j \le t}^{p} \Omega_{qj}\right] \prod_{l=1}^{k} Q_l\,(a_l t|\delta), \qquad (23)$$

onde $Q_l(.|\delta)$ é a transformada de Laplace da fragilidade w_l, $l = 1, \dots, k$, e Ω e δ são quantidades *a posteriori* de interesse nestes modelos de fragilidade designando, respectivamente, os parâmetros de regressão e os hiperparâmetros de fragilidade. Se os termos de fragilidade são identicamente distribuídos com distribuição gama de parâmetros de forma δ_1 e de escala δ_2 e a unidade em causa tem somente um termo de fragilidade, então a f.s. marginal (23) fica reduzida a $S(t|\Omega, \delta_1, \delta_2) = \exp\left[-\sum_{q=0}^{p} z_q \sum_{j:u_j \le t}^{p} \Omega_{qj}\right] \left(\frac{\delta_2}{\delta_1 + t}\right)^{\delta_1}$. Quando avaliada em cada uma das s amostras dos métodos MCMC, a f.s. marginal (23) é aproximada por $\frac{1}{s}\sum_{i=1}^{s} S(t|\,\Omega^{(i)}, \delta^{(i)})$, onde $\Omega^{(i)}$ e $\delta^{(i)}$ são, respectivamente, os valores simulados para Ω e δ relativamente à i-ésima amostra da distribuição *a posteriori* conjunta (22).

Ilustração: Dados de cancro de laringe

Os dados de cancro de laringe foram analisados em Kardaun (1983) e reportam-se a 90 pacientes com este tipo de cancro tratados num hospital holandês entre 1970 e 1981. Apesar de haver outras covariáveis observadas neste conjunto de dados, a análise que se segue leva em conta somente a covariável estádio da doença, uma vez que as demais covariáveis foram consideradas pouco influentes em análises anteriores. O estádio da doença foi observado no momento de diagnóstico do cancro, sendo os seus quatro níveis ordenados do menos grave (estádio 1) ao mais grave (estádio 4). Dos 90 tempos de sobrevivência observados, 40 tempos foram censurados à direita e o último tempo de sobrevivência foi 10.7 anos (censurado).
Como os dados em questão são univariados, um termo de fragilidade é introduzido para cada paciente com objectivo de controlar a heterogeneidade não observada em termos individuais. Portanto, os modelos aditivos com fragilidade (13) são aqui usados para modelar a intensidade no instante $t \in J$ dos processos de contagem associados aos 90 pacientes, reexpressa por

$$I_i(t|W) = Y_i(t)[a_0(t) + \sum_{q=1}^{3} z_{iq}\, a_q(t) + w_i], \qquad (24)$$

onde $Y_i(t)$ indica se o paciente i está vivo imediatamente antes do instante t, $W = (w_1, \dots, w_{90})'$ é o vector de fragilidades individuais e $Z_i = (z_{i1}, z_{i2}, z_{i3})'$ é o vector de variáveis indicadoras do paciente i, que são definidas por $z_{iq} = 1$ se o paciente i tem estádio $q + 1$ e $z_{iq} = 0$, no caso contrário, com as respectivas funções de regressão $a_q(t)$, $q = 1,2,3$, $i = 1, \dots, 90$. Estas funções de regressão permitem avaliar a influência dos estádios 2, 3 e 4 ao longo do tempo.

Os termos de fragilidade em (24) são considerados independentes e identicamente distribuídos com distribuição de fragilidade pertencente à família de distribuições gama com hiperparâmetro δ. Neste cenário, três modelos aditivos com fragilidade podem ser ajustados aos dados de cancro de laringe formados a partir das seguintes distribuições de fragilidade: i) distribuição exponencial com $\delta = (1, \delta^*)$ - modelo MAF1; ii) distribuição gama com média 1 e $\delta = (\delta^*, \delta^*)$ - modelo MAF2; iii) distribuição gama irrestrita com $\delta = (\delta_1, \delta_2)$- modelo MAF3. Nestes modelos de fragilidade o eixo do tempo J foi particionado em $m = 20$ intervalos disjuntos $B_j = [u_{j-1}, u_j)$ com o mesmo comprimento e limitados por $u_0 = 0$ e $u_{20} = 10.7$. Este total de intervalos não proporciona posteriormente a definição de um número elevado de incrementos para as funções de regressão. Além disso, os processos *a priori* gama com incrementos independentes (20) são adoptados para os quatro parâmetros de regressão com $c_q = 0.001$ e $r_q = 0.1$, $q = 0, \dots, 3$. Devido às dificuldades de implementação dos modelos de sobrevivência aditivos com fragilidade, faz-se aqui o uso de métodos de simulação, nomeadamente os métodos de Monte Carlo via cadeias de Markov (MCMC) implementados no *software* WinBUGS (Spiegelhalter *et al*., 2003).

Uma medida do grau de heterogeneidade não observada nos pacientes é o desvio padrão da fragilidade, denotado por σ_{fr}. Nos modelos de fragilidade MAF1, MAF2 e MAF3 esta medida é dada, respectivamente, por $\frac{1}{\delta^*}$, $\frac{1}{\sqrt{\delta^*}}$, e $\frac{\sqrt{\delta_1}}{\delta_2}$, onde δ^*, δ_1 e δ_2 são os correspondentes hiperparâmetros de fragilidade. Algumas medidas sumariadas do parâmetro σ_{fr} encontram-se na tabela 1 para os modelos de fragilidade MAFq, q=1,2,3. Nomeadamente, as médias *a posteriori*, os desvios padrões e os respectivos limites dos intervalos de credibilidade a 95%. Em todos os modelos MAF existe alguma heterogeneidade não observada entre os pacientes devido à grandeza dos valores estimados para o parâmetro de fragilidade σ_{fr}.

Table 1: Medidas sumárias do desvio padrão da fragilidade σ_{fr}

Modelo	Média	D.Padrão	LIC(2.5%)	LIC(97.5%)
MAF1	0.163	0.0307	0.1109	0.2313
MAF2	0.007	0.0024	0.0049	0.0137
MAF3	0.016	0.0082	0.0062	0.0036

Durante o ajustamento dos modelos de fragilidade MAFq, q=1,2,3, os valores *a posteriori* da medida de validação cruzada negativa, discutida em Gelfand e Dey (1994), foram calculados para cada um dos modelos, dados por 211.4, 503.5 e 209.3 relativamente aos modelos MAF1, MAF2 e MAF3, respectivamente. Por conseguinte, o modelo escolhido deveria ser o modelo MAF3 mas este modelo apresentou problemas graves de convergência referidos a seguir e, portanto, o modelo mais apropriado entre os três modelos em consideração é o modelo MAF1. Note-se que este modelo apresentou um maior grau de heterogeneidade não observada entre os pacientes (reveja-se a média *a posteriori* de σ_{fr} na tabela 1).

Baseando-se nas quantidades *a posteriori* dos parâmetros de interesse no modelo MAF1 podem-se calcular as estimativas das funções de regressão cumulativas $\Omega_q(t)$, $q = 1,2,3$, que se encontram na figura 1. A interpretação destas curvas faz-se a partir de sua inclinação recordando que uma função de regressão $\alpha_q(t)$, sem influência ao longo do tempo deveria proporcionar uma função cumulativa do tipo $\Omega_q(t)$=t.

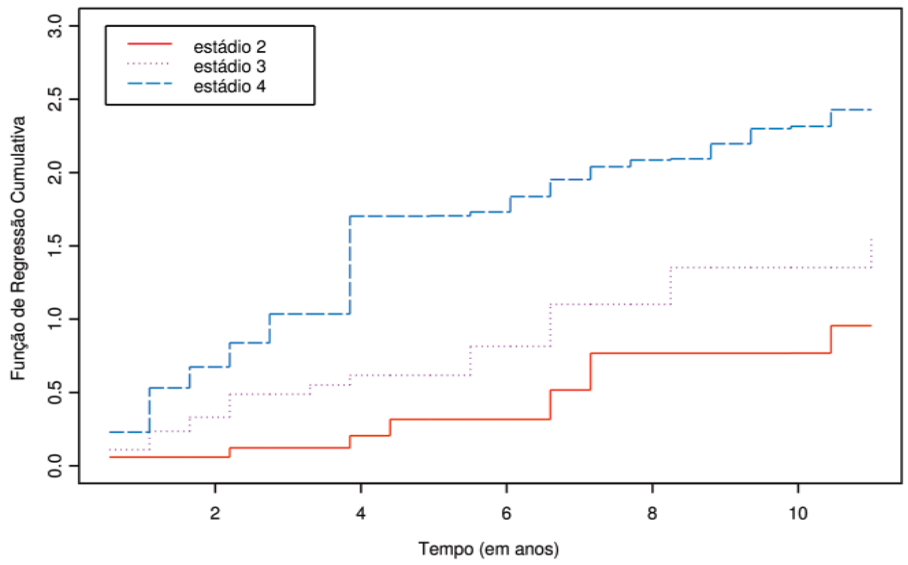


Figure 1: Função de regressão cumulativa – MAF1

As quantidades *a posteriori* das funções de regressão e dos hiperparâmetros de fragilidade com respeito ao modelo MAF1 foram usadas para avaliar as funções de sobrevivência marginal por estádio da doença. Na figura 2, estas funções revelam que o tempo de sobrevivência de pacientes com cancro de laringe é menor para aqueles pacientes com um maior avanço do estádio da doença.

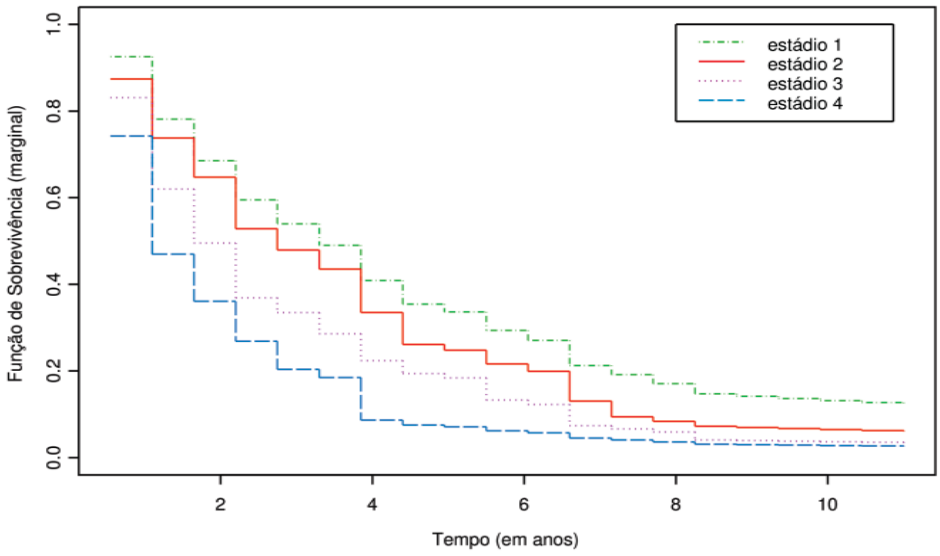


Figure 2: Função de sobrevivência – MAF1

Referências

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6, 701–726.

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Lecture Notes in Statistics* 2, 1–25.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models based on Counting Processes*. New York: Spring-Verlag.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* 47, 467–485.
Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* 34, 187–220.
Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: John Wiley and Sons.
Gelfand, A. E. G. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society B* 56, 501–514.
Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis* 1, 255–273.
Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.
Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). New York: John Wiley and Sons.
Kardaun, O. (1983). Statistical survival analysis of male larynx-cancer patients - a case study. *Statistica Neerlandica* 37, 103–125.
Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). New York: John Wiley and Sons.
Paulino, C. D., Amaral-Turkman, M. A, and Murteira, B. (2003). *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian.
Rocha, C. S. (1995). *Modelos com Fragilidade em Análise de Sobrevivência*. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa.
Silva, G. L. (2001). *Análise Bayesiana de Modelos de Sobrevivência com Fragilidade*. Tese de doutoramento, Departamento de Matemática - IST, Universidade de Técnica de Lisboa.
Silva, G. L. and Amaral-Turkman, M. A. (2004). Bayesian analysis of an additive survival model with frailty. *Communications in Statistics – Theory and Methods* 33, 2517–2533.
Spiegelhalter, D. Thomas, A., Best, N. and Lunn, D. *WinBUGS User Manual* (version 1.4). Department of Epidemiology and Public Health, Imperial College, St Mary's Hospital London, 2003 (http://www.mrc-bsu.cam.ac.uk/bugs/).
Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16, 439–454.

Sobrevivência de múltiplos eventos

Valeska Andreozzi[1], *valeska.andreozzi@fc.ul.pt*
Marilia Sá Carvalho[2], *carvalho@fiocruz.br*

[1]*Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal* (CEAUL)
[2]*Centro de Métodos Quantitativos da Fundação Oswaldo Cruz, Brasil* (CEMEQ)

1 Introdução

O presente texto tem como principal objetivo tornar acessível à comunidade aplicada a linguagem estatística dos modelos de sobrevivência aplicados aos eventos múltiplos. Para tal, complementa-se a componente teórica com um conjunto de exemplos reais estudados com o software estatístico R e com os respectivos códigos de programação. Este trabalho decorre de um outro mais vasto, que incluiu a revisão de diversos livros de análise de sobrevivência. O resultado final dessa compilação será publicado brevemente na segunda edição do livro *Análise de Sobreviviência: teoria e aplicações em Saúde*; Carvalho MS, Andreozzi V, Codeço CT, Barbosa MTS, Shimakura SE, Campos D; Editora Fiocruz, Rio de Janeiro, Brasil, que será lançado até ao final de 2011, no Brasil.

Muitas vezes o interesse na análise de sobrevivência é estudar o tempo até eventos que podem acontecer mais de uma vez para um mesmo indivíduo: gestações, internações, cáries, infartos do miocárdio e fraturas são alguns exemplos. Em outras situações, o interesse não está relacionado a um único tipo de evento, mas a diferentes tipos de eventos decorrentes de um mesmo fator de risco em estudo. Por exemplo, efeitos adversos de medicamentos, doenças oportunistas da Aids, ou o desfecho de um paciente em diálise (que pode ser óbito decorrente da insuficiência renal *versus* óbito por outras causas). As duas perguntas básicas quando analisamos a ocorrência de eventos múltiplos são: (1) Quais são os fatores de risco associados aos vários tempos até ocorrências de eventos em um mesmo indivíduo? e (2) Como analisar diferentes eventos como desfecho de uma mesma situação de risco?

A principal característica da análise de sobrevivência de eventos múltiplos é que para cada indivíduo registra-se mais de um tempo. Nesse caso, a utilização direta do modelo de riscos proporcionais de Cox não é adequada, mesmo se utilizarmos a formulação por processo de contagem, pois os intervalos de tempo de um mesmo indivíduo podem se sobrepor. Além disso, como a mesma pessoa sofre diversos eventos, esses podem ser correlacionados. O mecanismo pelo qual vários eventos ocorrem deve ser examinado cuidadosamente, pois a partir daí se define o modelo de análise, incorporando, caso necessário uma estrutura de dependência entre os tempos.

Outra particularidade da análise de eventos múltiplos é a composição do grupo de risco. Ao contrário do que ocorre no modelo de Cox clássico, e mesmo nos modelos com covariáveis mudando no tempo, quando os eventos são múltiplos, os indivíduos se mantêm no grupo de risco após a ocorrência de algum evento.

Entre as diversas abordagens possíveis, se os eventos são do mesmo tipo, pode-se considerar o número de eventos por indivíduo durante o período de estudo como a variável resposta. Nesse caso, considera-se que a variável dependente é uma contagem, e o modelo de Poisson pode ser utilizado. Contudo, o tempo entre os eventos não é incorporado nessa análise e o modelo não é capaz, por exemplo, de tratar de forma diferenciada dois eventos com um ano de intervalo de outro par com dez dias de intervalo.

Outra estratégia de modelagem consiste na utilização do modelo de riscos proporcionais, considerando como variável resposta apenas o tempo até um dos eventos analisados. Modela-se somente o tempo até o 1º evento, ou o tempo entre o 1º e o 2º eventos, ignorando-se a multiplicidade. Ou, caso se esteja trabalhando com diferentes desfechos, pode-se escolher apenas um deles, isoladamente. Nesse caso, a estimativa do efeito das covariáveis será válida somente para o evento analisado. Essa é uma abordagem simples de implementar, mas não permite conclusões muito abrangentes.

Um caso particular de eventos múltiplos é quando a ocorrência de um deles exclui a possibilidade de ocorrência do outro. São os denominados eventos competitivos. Novamente, pode-se analisar cada um em separado, mas a análise conjunta permite estimar o efeito de fatores de risco para diferentes desfechos.

Uma terceira estratégia é o ajuste de modelos de efeitos aleatórios, denominados de modelos de fragilidade na área de análise de sobrevivência. Condicionados ao efeito aleatório de cada indivíduo, considera-se que os eventos são independentes. É aplicável a eventos do mesmo tipo, recorrentes, mas não se adapta a eventos de tipos diferentes.

A abordagem que será aqui apresentada é denominada **modelos marginais**, que são também extensões do modelo de Cox. Identificar qual modelo marginal aplicar em cada situação é um dos aspectos fundamentais da modelagem. Para auxiliar nessa decisão, será apresentado na próxima seção uma classificação para os eventos múltiplos.

2 Classificação de eventos múltiplos

A classificação apresentada a seguir não é rígida nem definitiva. Por um lado, porque a evolução de métodos nesta área é grande e novos modelos são desenvolvidos, sempre buscando expressar melhor as características das várias situações analisadas. Por outro lado, o problema em si, que gera o aparecimento dos eventos, e o próprio desenho do estudo pode permitir diversas interpretações sobre qual o mecanismo gerador dos dados observados e consequentemente qual o modelo mais adequado. Assim, para orientar a escolha do modelo, as principais questões que devem ser observadas na análise do tempo para ocorrência de eventos múltiplos são:

- a população em risco;
- o risco basal (λ_0);
- a estrutura temporal (ordenação);
- a estrutura de dependência entre os eventos.

A questão de fundo mais importante é a identificação de quem está em risco a cada momento em que ocorre um dos possíveis eventos. Por exemplo, em estudo de fatores associados à paridade, somente estará em risco de nova gestação mulheres não grávidas. Durante o período de gravidez a mulher não pode ser incluída na população em risco. Além disso, o risco de base (λ_0) pode ser constante para todos os eventos múltiplos ou pode variar, se considerarmos que para cada evento o risco de base é diferente.

A estrutura temporal muitas vezes também não é de fácil definição. A adoção de estimação por processo de contagem impõe necessariamente uma ordem aos eventos, que pode ser evitada com o uso da abordagem mais simples do tempo até o evento ou censura. Também é necessário avaliar se os tempos até os eventos se sobrepõem. Essa estrutura temporal é definida pela indicação de ordenação entre os estratos de riscos relacionados a cada evento. Além disso, a estrutura de dependência gerada pelas múltiplas observações de cada indivíduo define a necessidade de métodos robustos para corrigir a variância dos estimadores.

Um primeiro tipo de eventos múltiplos são os desfechos ou **eventos paralelos**, em que a ocorrência de um evento não exclui a ocorrência de outro evento, e não há qualquer ordem preferencial. É o caso do estudo de doenças oportunistas relacionadas à Aids, suponhamos tuberculose, pneumocistose e diarreia. Todas podem ocorrer, somente algumas ou mesmo nenhuma. Também podem ocorrer ao mesmo tempo ou sucessivamente.

Um segundo tipo de problema ocorre quando os eventos são ordenados. Nos eventos ordenados a sucessão de tempos segue obrigatoriamente uma ordem, que é dada seja pela estrutura de datas de início e fim de cada evento, seja por assumir uma ordem nos estratos de risco relacionados a cada desfecho, ou ambos.

A combinação entre definição de grupo sob risco e estrutura do risco basal (λ_0) permite identificar claramente dois tipos de problemas em que os eventos são ordenados: **eventos ordenados independentes** e **eventos ordenados estruturados**. O primeiro assume que o indivíduo sempre retorna ao grupo de risco após sofrer cada evento, e que o momento de ocorrência de cada evento não depende dos tempos anteriores. No segundo, assume-se que o indivíduo só entra em risco de sofrer o enésimo evento depois que o evento de ordem $n-1$ tiver ocorrido. Isto é, o risco basal para um segundo evento é zero até que o primeiro ocorra, enquanto que o risco do terceiro evento é zero até que o segundo ocorra.

Um exemplo de **eventos ordenados independentes** é a ocorrência de infecções respiratórias agudas. Os episódios são ordenados, e na vigência de um episódio o indivíduo não está na população em risco. A hipótese de independência nesse contexto significa que o momento de ocorrência de cada evento é totalmente independente do tempo decorrido anteriormente ou do número de eventos até então. O risco basal não varia entre os eventos, o que é razoável supor, ainda que as crianças (ou idosos) com muitos episódios possam ter perfis diferentes, seja por concomitância de outras doenças ou processos alérgicos. Esses diferentes perfis devem ser modelados por meio da inclusão de covariáveis no modelo. Esse tipo de problema dá origem ao **modelo de eventos ordenados independentes** ou **AG** (Andersen-Gill), discutido na Seção 5.1.

No caso dos **eventos ordenados estruturados** o infarto agudo do miocárdio é um exemplo no qual é razoável supor que o risco de sofrer o primeiro infarto é diferente do risco de sofrer o segundo. Reinternações hospitalares também se encaixam nesta classificação, se o risco de reinternação depender do número de internações já sofridas. Nesses casos, considera-se que o risco de base se altera à medida que o indivíduo sofre novos eventos, ou seja, os eventos, além de ordenados no tempo, são ordenados também segundo o risco basal, podendo-se considerar que o indivíduo somente estará em risco do segundo evento se o primeiro tiver ocorrido. Em outras palavras, o segundo evento está condicionado à ocorrência do primeiro evento e assim sucessivamente. Esse problema dá origem ao **modelo estruturado** ou **PWP** (Prentice, Williams & Peterson).

Os **eventos com risco concomitante**, também são exemplos de eventos ordenados em que o indivíduo, ao iniciar o acompanhamento, está concomitantemente em risco de sofrer o primeiro, segundo, terceiro, enésimo evento. O risco só cessa quando o evento de ordem n, predefinida, acontece. Considera-se que cada evento tem um risco basal diferente, dependendo de sua posição na sequência de eventos, mas independente do tipo de evento. A ocorrência de reações adversas a medicamentos pode ser tratada como um exemplo teórico desse tipo. Entretanto, a suposição de ordenamento com risco concomitante significa que o tempo para a ocorrência de cada evento é sempre contado a partir do tempo zero, não sendo indicado o uso de estrutura de contagem. A ordem, neste caso, é dada pela enumeração do evento, sem condicionar a ocorrência de um evento à ocorrência do anterior, mas considerando um risco basal diferente conforme a ordem de ocorrência. Essa abordagem é denominada modelo marginal ou **WLW** (Wei, Lin & Weissfeld).

3 Modelos marginais

Os modelos marginais são utilizados em análise de sobrevivência quando os tempos são correlacionados, ou seja, quando existem várias observações de um mesmo indivíduo. A estratégia dessa abordagem é assumir que apenas a resposta média da população, modelada como uma função das covariáveis, é o foco de interesse. E por isso, os valores dos parâmetros de regressão β mantêm a mesma interpretação de um modelo para problemas com observações independentes. Nos modelos de regressão marginais as estimativas pontuais dos parâmetros de regressão são calculadas utilizando métodos baseados na verossimilhança parcial, assumindo que as observações são não correlacionadas.

A variância das estimativas dos parâmetros, no entanto, precisa ser calculada levando-se em consideração a estrutura de correlação dos dados intraindivíduo. A derivação de estimativas robustas de variância para o modelo de Cox na presença de dependência utiliza a mesma forma do cálculo dos resíduos escore.

Para a modelagem de eventos múltiplos através do ajuste de modelos marginais, é necessário percorrer algumas etapas. A primeira delas, e a mais importante, é identificar conceitualmente, com base na classificação dos eventos apresentada na Seção 2, qual tipo de evento se aproxima mais do problema em questão. Por exemplo, vamos analisar os tempos até todos os eventos ou apenas até o primeiro evento? Os eventos são ordenados ou não? Em alguns casos, só existe uma forma de analisar o problema. Eventualmente, mais de uma abordagem é possível e até mesmo necessária. Além disso, além dos modelos marginais, outras abordagens também podem ajustar eventos múltiplos, sendo muito usados os modelos de fragilidade.

Com base na classificação do tipo de evento, a pergunta seguinte é: quem está sob risco em cada momento? A partir dessa definição, se organiza o banco de dados apropriadamente, pois esta formatação muda conforme a definição do grupo de risco. Com frequência esta é a parte mais trabalhosa de todo o processo. Depois de organizado o banco, o modelo de Cox deve ser ajustado da forma usual, ignorando-se a estrutura de dependência dos dados. Os resultados servirão para comparação. Espera-se que as estimativas pontuais se mantenham razoavelmente similares e por isso compará-las com efeitos estimados por modelos mais complexos ajuda a detectar eventuais (e frequentes) erros na formatação do banco de dados.

Por fim, caso seja razoável para a análise em questão, pode ser interessante experimentar duas ou mais abordagens. Muitas vezes os parâmetros estimados por diversas abordagens são semelhantes, e nesse caso o modelo mais simples é o mais indicado.

Apresentaremos nas seções seguintes exemplos de como modelar os múltiplos eventos, ordenados ou não. Para facilitar a apresentação dos diferentes modelos para eventos múltiplos, faremos uso de uma representação esquemática por meio de modelos multiestados (Figura 1). Esta representação consiste em identificar os estados que cada indivíduo em observação pode assumir e as possíveis transições entre tais estados. Por exemplo, o modelo clássico de sobrevivência (com uma ocorrência de um único desfecho) tem apenas dois estados: vivo (em risco) e morto (fora de risco); ou, sem Aids (em risco) e com Aids (fora de risco); amamentando (em risco) ou não amamentando (fora de risco). Nesses casos, todos os indivíduos entram no estudo no primeiro estado e são observados até passarem para o outro estado. Modelos de eventos múltiplos, por sua vez, possuem mais do que dois estados possíveis. A forma como os estados se relacionam é o que define os diferentes tipos de modelos para eventos múltiplos. O que se busca estimar são os fatores associados ao risco de transição entre estados.

A seguir, serão abordados os eventos paralelos, ordenados independentes e ordenados estruturados. Para os eventos competitivos e ordenados com risco concomitante sugerimos, respectivamente, a leitura de Putter *et al* (2007)[1] e Signorovitch and Wei (2008)[2].

[1] Putter, H., Geskus, R. B. & Fiocco, M. Tutorial in biostatistics: Competing risks and multi-state models.*Statistics in Medicine 26* (2007), 2389–2430.

[2] Signorovitch, J. E. & Wei, L.-J. *Wei-Lin-Weissfeld Method for Multiple Times to Events*.Wiley Encyclopedia of Clinical Trials (2008)

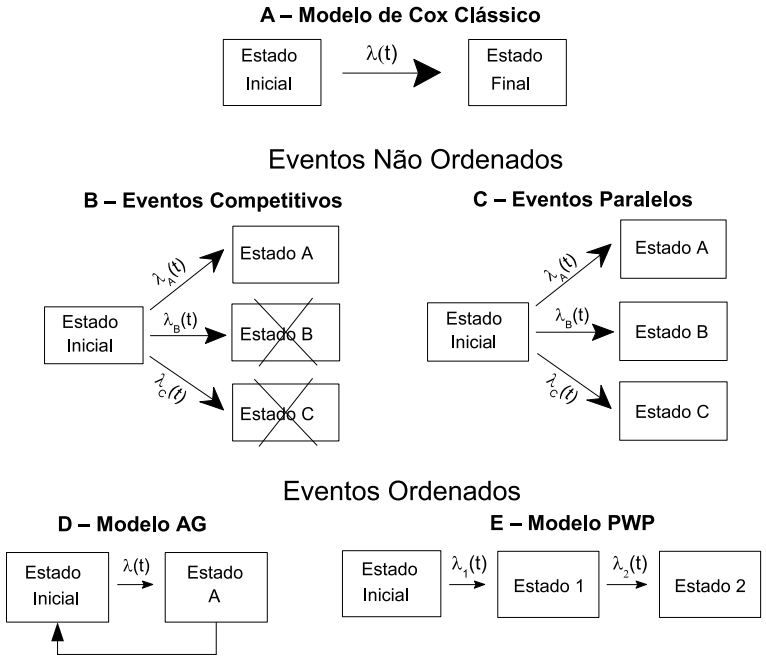


Figura 1: Representação gráfica de modelos de sobrevivência por meio de modelos multiestados

4 Eventos paralelos

O tempo até o aparecimento de doenças oportunistas em pacientes portadores de HIV pode ser investigado para se estudar, por exemplo, quais covariáveis estão mais associadas a uma ou outra complicação, se o tipo de doença muda conforme a terapia antirretroviral em curso, ou ainda para verificar se a linha de base do risco é a mesma, qualquer que seja a doença. Os eventos não são competitivos, pois a ocorrência de um não elimina o risco de ocorrência de outro. Eles também não são ordenados, pois podem ocorrer em qualquer ordem e mesmo simultaneamente. O objetivo é estimar o efeito de diversos fatores de risco para diferentes doenças oportunistas entre indivíduos que apresentaram alguma delas.

Para cada paciente HIV positivo, registra-se o tempo desde o início do acompanhamento até a ocorrência de algumas doenças oportunistas de interesse para o estudo, selecionadas entre as mais frequentes e incluindo alterações hematológicas que podem estar relacionadas ao uso dos antirretrovirais. Se um paciente experimenta n doenças, então n tempos são registrados (e n linhas são colocadas no banco de dados). A Figura 2 mostra os tempos até a ocorrência de doenças oportunistas para quatro pacientes. O paciente 2, por exemplo, teve herpes e pneumocistose 249 dias após sua entrada na coorte e posteriormente apresentou candidíase no dia 459. Como um evento não elimina o risco de ocorrência do outro, as linhas para cada evento possível correm paralelamente. O paciente 4, que teve um único evento (alterações hematológicas no dia 184 após sua entrada no estudo), é representado por apenas uma linha no gráfico.

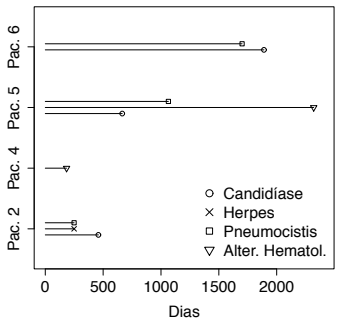


Figura 2: Tempo até a ocorrência de doenças oportunistas em 4 pacientes HIV+

Observe que os tempos de sobrevivência de um mesmo paciente apresentam sobreposição. Esta sobreposição é realista, pois o paciente está em risco, simultaneamente, para todas as doenças consideradas. Uma questão importante é a ausência de censura nesse banco de dados. Um paciente entra no estudo somente se apresenta ao menos uma complicação associada à Aids, e sai inteiramente do banco

no momento da última doença registrada. Ou seja, somente se compara indivíduos que apresentaram doenças oportunistas, e não os que potencialmente poderiam apresentar doenças.

4.1 Modelo para eventos paralelos

Na Figura 1, o quadro C indica que a partir do estado original, qualquer das ocorrências é possível, inclusive mais de uma simultaneamente. Haverá tantas transições para cada indivíduo quantas doenças aparecerem, sem número fixo. O risco $\lambda(t)$ pode ser constante para todos os desfechos, ou variar, e neste caso haveria um $\lambda_A(t)$ para a doença oportunista **A**, um $\lambda_B(t)$ para a doença **B**, e assim por diante.

O modelo marginal para eventos paralelos tem a mesma formulação do modelo de Cox clássico. A estratificação pode ser necessária se houver diferenças no risco basal entre as várias doenças. O que caracteriza o modelo como de eventos paralelos não ordenados é a itilização do tempo transcorrido desde a entrada do paciente na coorte.

É importante verificar se o risco de base do risco difere entre as doenças, indicando a necessidade de considerar as doenças como estratos através da estimativa não paramétrica das curvas de sobrevivência de cada doença oportunista.

Modelando-se o efeito de diversos fatores de risco para a ocorrência de qualquer uma das doenças, a única novidade será a inclusão de um termo que permite identificar para um mesmo indivíduo seus diversos desfechos, permitindo assim a correção da variância dos parâmetros da regressão. No R isto é feito incluíndo o argumento `cluster()`.

```
> summary(coxph(Surv(tempo,status)~var1+var2+cluster(registro),data=dados))
```

Se houver indicação de diferença no risco basal segundo o evento, inclui-se o argumento `strata()`, assim:

```
> summary(coxph(Surv(tempo,status)~var1+var2+cluster(registro)+strata(oport),
  data=dados))
```

5 Eventos ordenados independentes

Um estudo de coorte com crianças de 6 a 48 meses foi conduzido para testar a suplementação de vitamina A na proteção contra a diarreia infantil. O desfecho de interesse é o episódio de diarreia, e o tempo de sobrevivência é definido como tempo até a ocorrência de um episódio de diarreia. A Figura 3 mostra os tempos observados para as primeiras 10 crianças da coorte. O símbolo | indica o início do episódio (evento) e o símbolo ○ indica censura, que neste caso foi sempre ao término do estudo. Como uma mesma criança pode ter mais de um episódio de diarreia, trata-se de um exemplo de eventos recorrentes: a criança volta a fazer parte do grupo de risco, ao recuperar-se de um episódio de diarreia. Observe que logo após o início de cada episódio de diarreia há um pequeno intervalo que representa sua duração, durante o qual a criança não está em risco de um novo episódio.

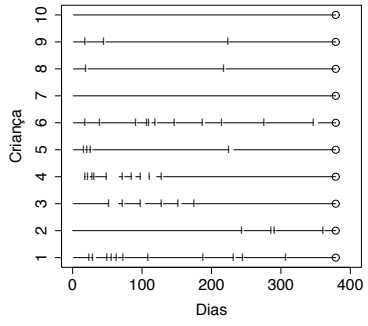


Figura 3: Tempos observados entre episódios recorrentes de diarreia

5.1 Modelo AG

O quadro D da Figura 1 mostra, de forma esquemática, o modelo de eventos ordenados independentes. O indivíduo no estado inicial tem um risco $\lambda(t)$ de sofrer um evento. Após a ocorrência do evento, ele retorna ao estado inicial, com o mesmo risco $\lambda(t)$ de sofrer um novo evento. O pressuposto de risco igual para qualquer evento é bastante forte, pois assume que o histórico do indivíduo não afeta seu risco presente. Outro tema que se pode modelar da mesma forma é o risco de engravidar, desde que os hábitos contraceptivos não mudem, a fertilidade não seja alterada pelo número de gestações e o período de observação seja curto o suficiente para não ser afetado pela idade.

Observe que no esquema da Figura 1 existem duas setas (em outras palavras, duas transições), mas só uma é associada a um risco $\lambda(t)$. Isso indica que o interesse está focado apenas em um tipo de transição. No exemplo da diarreia, a pergunta se refere ao risco de ter um episódio de diarreia e **não** aos fatores associados à duração do episódio de diarreia. Assim, o tempo que o indivíduo se encontra no segundo estado não é incorporado na análise. Observe na Figura 3 que os intervalos em que a criança está com diarreia, ou seja, o tempo em que ela está no estado A, é omitido do banco de dados. Esse tempo é considerado um tempo censurado. Nada impede porém que se estude o risco de terminar o episódio de diarreia, ou o risco de receber alta. O banco de dados reflete a pergunta e deve ser construído de acordo com o objetivo do estudo. Também é possível modelar a probabilidade associada à transição entre estados, nos chamados modelos multiestados.

O modelo AG tem a seguinte expressão, utilizando a notação de processo de contagem:

$$\lambda_i(t|x_i(t)) = Y_i(t)\lambda_0(t)e^{\beta x_i(t)}, \tag{1}$$

em que $Y_i(t)$ é igual a 1 quando os indivíduos estão em risco, e igual a 0 nos períodos de duração da doença ou no fim do evento.

A sintaxe do modelo AG no R é apresentada a seguir:

```
> modeloAG<-coxph(Surv(ini,fim,status)~var1+var2+cluster(registro),data=dados)
> summary(modeloAG)
```

Retornando à definição do modelo AG, lembre que um de seus pressupostos fundamentais é de que os tempos entre eventos de um mesmo indivíduo são independentes entre si. Como testar esse pressuposto? Bem, supor tempos independentes significa, em outras palavras, que não há uma estrutura de correlação das observações de um mesmo indivíduo, desde que as covariáveis incluídas no modelo expliquem as diferenças entre indivíduos. Se isso é correto, espera-se que a variância calculada assumindo-se independência entre todas as observações (modelo de Cox clássico) seja apenas um pouco menor que a variância robusta. Caso isso não aconteça, é factível pensar que o risco de sofrer novos episódios de diarreia é de alguma forma correlacionado aos episódios anteriores. Nesse caso, o modelo AG talvez não seja o mais indicado. Duas alternativas de modelagem devem ser então experimentadas: o modelo estruturado PWP, apresentado a seguir, e o modelo de fragilidades

6 Eventos ordenados estruturados

Os episódios de diarreia poderiam ter uma estrutura na qual o risco de um episódio aumentaria a cada novo episódio. Para verificar essa hipótese ajustamos um novo modelo aos dados de diarreia apresentados no Seção 5. Para isso uma variável de ordenação é criada no banco de dados, e o modelo é ajustado estratificando-se por essa variável, de forma que riscos basais diferentes e ordenados sejam associados a cada evento.

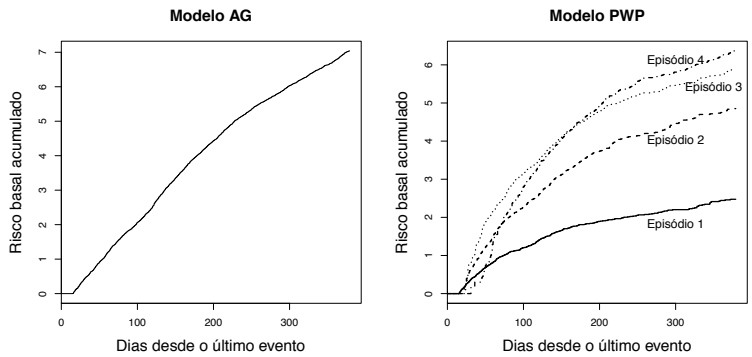


Figura 4: Risco basal acumulado para ocorrência de novo episódio de diarreia segundo o modelo AG e PWP

6.1 Modelo PWP

A Figura 1, quadro E mostra a lógica conceitual do modelo estruturado ou PWP. Ao contrário do modelo anterior, nesse se assume que existe uma estrutura de dependência entre os tempos de um indivíduo. Isto é, considera-se que o risco $\lambda(t)$ do indivíduo sofrer o primeiro evento é diferente do risco de sofrer um segundo evento (dado que sofreu o primeiro) e assim por diante, ou seja, $\lambda_1(t) \neq \lambda_2(t)$, e que o evento j somente ocorre após o evento $j-1$ ter ocorrido.

O modelo de riscos proporcionais, neste caso, assume a forma:

$$\lambda_{ij}(t|x_i(t)) = Y_{ij}(t)\lambda_{0j}(t)e^{\beta x_i(t)}, \tag{2}$$

em que $\lambda_{ij}(t)$ é o risco do indivíduo i sofrer o evento j. Observe que neste caso, o índice j do evento é importante.

Na prática, o modelo estruturado assume que um indivíduo só estará em risco para o evento j depois de ter ocorrido o evento $j-1$. Na fórmula, isso implica dizer que $Y_{ij}(t) = 0$ até que o evento j ocorra e então se torne 1.

A incorporação da estrutura nos eventos ordenados é feita através da inclusão da variável de ordenação no modelo. Essa variável tem que respeitar a ordenação dos tempos de ocorrência da diarreia. No R, o modelo PWP tem a seguinte sintaxe:

```
> modeloPWP<-coxph(Surv(ini,fim,status)~var1+var2+strata(enum)+cluster(registro),
  data=dados)
> summary(modeloPWP)
```

Observe na Figura 4 as estimativas do risco basal acumulado para diarreia, segundo os modelos AG e PWP. O modelo AG estima uma única curva de risco basal, que pode ser interpretada como uma média dos riscos basais de cada evento. Já o modelo PWP ajusta uma curva para cada episódio de diarreia. Veja como o risco basal de um novo episódio aumenta conforme o número de episódios já sofridos. Parece, a partir desta figura, que o pressuposto de eventos ordenados independentes não é o mais adequado para o problema da diarreia.

Os dados dos exemplos mencionados no texto encontram-se no site da primeira versão do livro *http://sobrevida.fiocruz.br/*.

Introdução de uma componente de *e-learning* no ensino da Bioestatística: uma oportunidade de reformulação dos métodos de ensino e aprendizagem[1]

Margarida Fonseca Cardoso, *mcard@icbas.up.pt*

ICBAS – Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto

Teresa Correia, *tcorreia@reit.up.pt*

Universidade do Porto, Universidade Digital - Novas Tecnologias na Educação

1. Contextualização

Atualmente, a importância da bioestatística na investigação e prática das ciências biológicas, biomédicas e matemáticas é notória. A sua aplicação efetiva requer o envolvimento de equipas multidisciplinares, ou seja, constituída não só por investigadores/profissionais de áreas biológicas/biomédicas mas, também por investigadores/profissionais de áreas mais quantitativas. Sendo grande parte da evidência científica das áreas biológicas/biomédicas baseada em estudos que envolvem a aplicação de métodos estatísticos, torna-se necessário que o futuro profissional – atual estudante - saiba interpretar resultados estatísticos (DeMets et al. 2006; Freeman et al. 2008; Zelen 2006).

Assim, é objetivo comum aos Cursos com unidades curriculares de Bioestatística que todos os estudantes adquiram conhecimentos dos princípios gerais da estatística (descritiva e inferencial) assim como conhecimento dos métodos de estatística básica. Todos os estudantes deverão ainda adquirir competências de comunicação oral e escrita, de forma a serem produtivos e eficientes em equipa. Uma primeira exposição a aplicações da bioestatística deverá levar o estudante a ter um contacto mais próximo com a realidade profissional que encontrará e deverá também servir para motivá-lo para o processo de investigação (Aliaga et al. 2010; DeMets et al. 2006; Freeman et al. 2008; Zelen 2006), ou seja, todos os estudantes devem ser preparados quer em termos teóricos quer para a prática da estatística, mesmo que apenas uma parte deles prossiga para estudos mais avançados.

Após a frequência da unidade curricular todos os estudantes deverão ser capazes de explorar dados utilizando medidas estatísticas e métodos gráficos; saber aplicar procedimentos estatísticos simples autonomamente; identificar os tipos de testes estatísticos que podem ser aplicados num dado conjunto de dados, conhecer os pressupostos e saber testá-los; conhecer as limitações e a força das conclusões obtidas a partir da análise estatística efetuada; participar ativamente na resolução cooperativa de problemas, principalmente em exercícios realizados em grupos pequenos; demonstrar capacidades de pesquisa, análise e sistematização da informação recolhida e sua comunicação escrita e oral (Aliaga et al. 2010).

De acordo com informação fornecida pela Direção Geral do Ensino Superior, em Portugal a concretização do Processo de Bolonha foi globalmente concluída no ano letivo de 2009/2010, como resultado da aplicação dos Decretos-Lei nº 42/2005, nº 74/2006 e nº 107/2008. Estes Decretos-Lei consideram não só aspetos formais da reestruturação do Ensino Superior, como também a "transição

[1] Versão reduzida do artigo apresentado no *VII Workshop de e-learning* da Universidade do Porto [http://elearning.up.pt].

de um sistema de ensino baseado na transmissão de conhecimentos para um sistema baseado no desenvolvimento das competências dos estudantes, em que as componentes de trabalho experimental ou de projeto, entre outras, e a aquisição de competências transversais devem desempenhar um papel decisivo" (Ministério da Ciência Tecnologia e Ensino Superior 2006). Em qualquer um dos três ciclos do Ensino Superior - Licenciatura, Mestrado e Doutoramento - o grau deve ser atribuído àqueles que demonstrem capacidade de aplicar os conhecimentos adquiridos, resolver problemas, interpretar informação relevante e comunicar com públicos especialistas e não especialistas.

2. Modelo/Estratégia

Tal como o plano de estudos, a metodologia adotada no ensino teórico e prático apresentou alguma flexibilidade de acordo com o grau do Curso, a formação de base dos estudantes, a carga horária e o número de ECTS da unidade curricular. Assim, os métodos de ensino teórico-práticos definidos para lecionar os conteúdos de Bioestatística em diferentes unidades curriculares compreendiam três vertentes complementares: aulas teóricas, práticas e a utilização da plataforma de *e-learning*.

Nas aulas presenciais deu-se prioridade à exposição detalhada dos conteúdos programáticos, à realização de exercícios que estimulassem a capacidade de raciocínio e ao esclarecimento de dúvidas que implicassem um "atendimento" mais personalizado. A componente *on-line* funcionou como uma "extensão" da sala de aula, dando a possibilidade de treinar mais intensamente as capacidades dos estudantes, fazer uma avaliação contínua e disponibilizar um conjunto de materiais de apoio ao estudo das matérias lecionadas nas aulas teóricas e práticas.

A plataforma Moodle U.PORTO foi o sistema utilizado para disponibilizar a componente *on-line* de suporte às aulas presenciais. As principais ferramentas interativas do Moodle utilizadas para lecionar os conteúdos de Bioestatística foram: o fórum, a base de dados, os trabalhos e os testes *on-line*. Para além destas ferramentas a plataforma serviu também para disponibilizar diversos materiais de apoio ao estudo, em formato digital.

A introdução de uma base de dados na página *on-line* de cada unidade curricular, que possa ser construída pelos estudantes e docentes, foi uma forma simples de complementar a bibliografia inicialmente fornecida.

Para cada tópico do programa, foi proposto um trabalho semanal, a realizar em grupo, que seria resolvido no formato de um curto artigo científico. O objetivo dos trabalhos semanais era que os grupos construíssem um hipotético "estudo de investigação". No enunciado eram identificadas as variáveis consideradas, o estudo efetuado e fornecido um conjunto de dados; a partir desta informação os grupos deveriam construir o âmbito e os objetivos do estudo. As instruções para os trabalhos fornecidos dividem-se em duas partes. A primeira parte correspondia em linhas gerais às instruções para os autores de qualquer revista científica na área das ciências biológicas/biomédicas, em que os artigos são divididos em Introdução, Métodos, Resultados e Discussão. A segunda parte, denominada Apêndice, é mais académica e corresponde à apresentação formal dos testes estatísticos efetuados, assim como dos *outputs* correspondentes ao programa estatístico utilizado na análise dos dados.

Em todas as unidades curriculares lecionadas foi necessário avaliar individualmente todos os estudantes, de modo a ter uma perspetiva mais próxima dos conhecimentos assimilados por cada um, uma vez que os trabalhos de grupo podem, por vezes, não ser comprovativos dos reais conhecimentos e capacidades de cada um dos estudantes. Os estudantes tiveram acesso ao enunciado dos exames e submeteram-nos através da página *on-line.*

Foram consideradas duas formas distintas de avaliação, utilizadas isoladamente ou em conjunto: resolução de problemas estatísticos e testes de escolha múltipla. Nos problemas foram avaliadas duas componentes: conhecimentos adquiridos sobre os métodos estatísticos lecionados; e interpretação e transmissão correta dos resultados obtidos numa análise estatística.

Na formulação das questões de escolha múltipla considerou-se algumas questões simples de resposta imediata e questões mais complexas. As questões mais complexas corresponderam a problemas biológicos/biomédicos, em que o estudante teria de identificar o método estatístico adequado à sua resolução, decidir a formulação da hipótese estatística associada ao problema ou interpretar os resultados obtidos através da aplicação de um método estatístico. Na construção de distratores plausíveis, considerou-se erros comuns cometidos geralmente pelos estudantes, e conteúdos próximos da resposta correta credíveis para os estudantes com fracos conhecimentos (Haladyna and Downing 1989; Moreno et al. 2006). Visto que o teste havia sido elaborado de modo a que não houvesse

penalizações para respostas erradas, os estudantes foram incentivados a responder a todas as questões, sendo que no fim foi feita uma correção da classificação para o acerto ao acaso (Prihoda et al. 2006).

3. Resultados

Em todas as unidades curriculares lecionadas em 2009/2010 o modelo de ensino utilizado foi o mesmo, ou seja, a estrutura da página *on-line*, conteúdos programáticos, recursos e atividades propostas foram muito semelhantes, tendo-se apenas feito algumas adaptações consoante o grau do Curso e as especificidades da área de estudo em questão. As diferentes unidades curriculares lecionadas correspondem a: Bioestatística da Licenciatura em Ciências do Meio Aquático; módulo dado em conjunto aos Mestrado em Saúde Pública, Programa de Doutoramento em Matemática Aplicada e Mestrado Profissionalizante em Melhoramento Genético; e módulo de métodos estatísticos da unidade curricular de Métodos Numéricos e Estatísticos do Mestrado Integrado em Bioengenharia.

Os estudantes - sobretudo os estudantes trabalhadores - que não tiveram oportunidade de comparecer a algumas das aulas presenciais, após consultarem o material de apoio, puderam através do fórum colocar as suas questões à docente. O fórum permitiu ainda o contacto com a docente sem restrição de dia e hora.

Apesar de a participação no fórum não ter sido explorada por todos os estudantes, quase todos os grupos das diferentes unidades curriculares lecionadas participaram no fórum de dúvidas, com maior ou menor frequência, tendo sido a maior parte das questões colocadas relativas aos trabalhos semanais. Os períodos de comunicação mais intensos eram na véspera da entrega dos trabalhos e o fórum permitiu uma rápida resolução de dúvidas e pedidos de esclarecimentos durante toda a semana (incluindo ao fim de semana). De salientar que este acompanhamento dos trabalhos por parte da docente, não seria possível se o estudante tivesse de aguardar pela marcação de uma reunião com consequente deslocação ao campus.

Muitas das questões colocadas constituíram uma oportunidade para discutir conceitos teóricos e aplicação prática da Bioestatística, sem a pressão imposta pelo tempo como acontece na sala de aula. Esta comunicação (assíncrona) mais ativa com os estudantes, apesar de constituir um acréscimo de trabalho para a docente, permitiu também fazer quando necessário alguma pesquisa de forma a poder fornecer uma resposta mais adequada e clara (Bruce et al. 2002).

Nas unidades curriculares lecionadas, a maioria dos registos adicionados à base de dados resultaram de contribuições dos estudantes. Essas contribuições incluíram sobretudo artigos discutindo a aplicação de métodos estatísticos, artigos que exemplificavam a aplicação de métodos estatísticos e páginas *web* com conteúdos de estatística. Foram ainda colocados alguns vídeos de palestras de bioestatísticos reconhecidos e vídeos com exemplificação da utilização de métodos estatísticos lecionados. No início, e em geral, os estudantes de todas as unidades não estavam muito familiarizados com a pesquisa de recursos em revistas científicas internacionais, nomeadamente da área biológica/biomédica, e a *Wikipedia* constituía a principal fonte de informação. Ao longo do período letivo, esta tendência alterou-se e os estudantes foram-se habituando a utilizar estas fontes bibliográficas não só no sentido de partilharem a informação na base de dados como em proveito próprio, utilizando-as na resolução dos seus trabalhos.

Como já foi referido anteriormente, os trabalhos semanais, elaborados em grupos, foram a base do ensino da Bioestatística em todas as unidades curriculares. Os temas dos trabalhos semanais foram sempre que possível escolhidos de acordo com a área científica do Curso a que pertenciam os estudantes. Apesar de o enunciado e os respetivos dados serem únicos para cada grupo, todos os trabalhos diziam respeito ao mesmo conteúdo programático, o que permitiu aos grupos/estudantes discutir entre si as diferentes abordagens possíveis.

Nos trabalhos semanais, em forma de artigo, os grupos deveriam apresentar um hipotético "estudo de investigação". Apesar de se ter inserido uma componente criativa, tal como num "verdadeiro" estudo de investigação, os grupos deveriam saber fazer uma ligação coerente entre os objetivos definidos, a metodologia estatística utilizada, e a discussão dos resultados obtidos.

Nos primeiros trabalhos, os grupos apresentaram-se muito heterogéneos entre si. Enquanto uns pareciam mais familiarizados com a escrita de artigos em formato científico e com os métodos estatísticos inicialmente utilizados, os restantes mostravam o oposto, evidenciavam grandes dificuldades e apresentaram trabalhos com conteúdos muito afastados do pretendido. No entanto, gradualmente o desempenho dos grupos foi-se aproximando e a qualidade dos trabalhos apresentados

melhorou significativamente. Esta melhoria deveu-se em parte às observações sobre o trabalho efetuado, fornecidas pela docente em reunião presencial com o grupo, que no caso de verificar graves erros nos trabalhos elaborados optou por solicitar a reformulação dos mesmos, de modo a assegurar que todos os grupos tivessem a capacidade de elaborar um relatório com a estrutura definida e com uma análise estatística adequada. Por outro lado, essa melhoria também se deveu à crescente familiarização dos estudantes com a escrita científica na área biológica/biomédica.

No que se refere ao trabalho final, realizado em uma das unidades (Bioestatística da Licenciatura em Ciências do Meio Aquático), os trabalhos propostos pelos diferentes grupos foram na maioria resultado de colaboração prestada por investigadores e docentes ligados à área.

A avaliação individual dos estudantes foi considerada em todas as unidades curriculares lecionadas, no entanto as componentes de avaliação consideradas variaram. Em todos as unidades, uma parte dos estudantes não conseguiu aprovação em época normal, revelando agora na avaliação individual alguma heterogeneidade de conhecimentos entre membros do mesmo grupo. No entanto, após o exame de recurso a taxa de aprovação atingida foi elevada. Assim, no ano letivo 2009/2010, nos 23 estudantes avaliados na Época Normal e/ou Recurso da Licenciatura em Ciências do Meio Aquático nenhum estudante reprovou; dos 38 estudantes avaliados no módulo dado em conjunto aos Mestrados e Programa de Doutoramento apenas 1 estudante reprovou; no Mestrado Integrado em Bioengenharia apenas 1 dos estudantes avaliados reprovaria com base nas classificações obtidas na componente de estatística da unidade curricular lecionada. Com exceção do módulo dado em conjunto aos Mestrados e Programa de Doutoramento, na Licenciatura em Ciências do Meio Aquático e Mestrado Integrado em Bioengenharia verificou-se um desfasamento entre o número de estudantes inscritos para exame e o número de estudantes avaliados. Alguns destes estudantes que não compareceram a exame faziam parte de grupos com aprovação nos trabalhos semanais, no entanto, de acordo com a perceção da docente seriam elementos sem uma participação efetiva. Esta ideia baseia-se sobretudo na não participação destes estudantes no fórum, assim como na não participação em geral nas reuniões do grupo com a docente.

4. Conclusão

Há já alguns anos que se sente a necessidade de adotar um plano inovador no modo de ensino da bioestatística e motivador no modo de aprendizagem. Um plano que vá de encontro às necessidades profissionais futuras do estudante As novas tecnologias de informação e comunicação oferecem, atualmente, uma grande oportunidade de inovação, quer ao nível da transmissão dos conhecimentos quer ao nível da avaliação e comunicação entre os docentes e os estudantes (Aliaga et al. 2010; Bruce et al. 2002; Duarte 2008; Heller et al. 2008; Simpson et al. 2009).

Neste artigo foi apresentado o ensino de conteúdos de Bioestatística básica, incluindo o programa, os conteúdos e os métodos de ensino feito de forma semelhante em diferentes unidades curriculares, de diferentes graus (Licenciatura, Mestrado e Doutoramento) e áreas científicas (Ciências do Meio Aquático, Saúde Pública, Matemática, Bioengenharia, etc.). Este método passou então pela introdução de uma componente de *e-learning*, complementando assim as tradicionais aulas presenciais (teóricas e práticas), ou seja, foi adotado o denominado modelo de *blended-learning*. O balanço global da utilização desta metodologia, no ano letivo de 2009/2010, foi sem dúvida positivo.

A utilização de um modelo de *blended-learning* facilitou a introdução de uma forte componente de resolução de problemas e um melhor acompanhamento do desempenho dos estudantes, através da realização de trabalhos semanais. A resolução desses problemas levou os estudantes a efetuarem mais pesquisas não só como forma de perceberem melhor o problema biológico/biomédico em causa e os métodos estatísticos necessários para a sua resolução, mas também para verem exemplos de análises escritas (que lhes permitissem perceber melhor como escrever/elaborar o trabalho) (Aliaga et al. 2010).

A taxa de sucesso obtida nos estudantes avaliados em 2009/2010 foi muito próxima dos 100%. A alta taxa de sucesso obtida não se pode confundir com facilitismo, mas resultou sem dúvida de um grande envolvimento dos estudantes durante todo o processo de ensino e aprendizagem. Independentemente das aulas presenciais, que são uma necessidade em qualquer unidade curricular de qualquer Curso, a utilização de uma plataforma de *e-learning* faz todo o sentido e constitui uma mais-valia para docentes e estudantes. Todavia, é importante sublinhar que estes ambientes de ensino virtuais não garantem automaticamente a melhoria de *performance* e aprovação dos estudantes nem o

sucesso de uma unidade curricular (Bruce et al. 2002; Duarte 2008), assim como também não substituem a experiência de uma equipa docente.

Os métodos utilizados no ensino e aprendizagem devem acompanhar a constante alteração da sociedade. Neste momento, a sociedade é marcada pela utilização das tecnologias a nível profissional ou particular. Se os métodos de ensino utilizados hoje podem estar completamente desatualizados num futuro próximo, essa não deve ser a preocupação do professor, mas sim a revisão, melhoria e adaptação das metodologias de ensino às necessidades dos estudantes.

5. Referências

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., e Witmer, J. (2010). *Guidelines for assessment and instruction in statistics education - college report*. American Statistical Association, San Francisco, California.

Bruce, J. C., Bond, S. T., e Jones, M. E. (2002). "Teaching epidemiology and statistics by distance learning." *Stat Med*, 21(7), 1009-20; discussion 1021-2.

DeMets, D. L., Stormo, G., Boehnke, M., Louis, T. A., Taylor, J., e Dixon, D. (2006). "Training of the next generation of biostatisticians: a call to action in the U.S." *Stat Med*, 25(20), 3415-29.

Duarte, A. M. (2008). "E-learning e abordagens à aprendizagem no ensino superior." *Sísífo / Revista de Ciências da Educação*, 7, 39-50.

Freeman, J. V., Collier, S., Staniforth, D., e Smith, K. J. (2008). "Innovations in curriculum design: A multi-disciplinary approach to teaching statistics to undergraduate medical students." *Bmc Medical Education*, 8(28).

Haladyna, T. M., e Downing, S. M. (1989). "Taxonomy of multiple choice item-writing rules." *Applied Measurement in Education*, 37-50.

Heller, G. Z., Forbes, A. B., Dear, K. B. G., e Jobling, E. (2008). "Biostatistics @ distance: a model for successful multi-institutional delivery." *The American Statistician*, 62(4), 1-8.

Ministério da Ciência Tecnologia e Ensino Superior. (2006). "Diário da Républica I Série A - No. 60 - 24 de Março de 2006, Decreto-Lei nº 74/2006.". pp. 2242-57.

Moreno, R., Martinez, R. J., e Muniz, J. (2006). "New Guidelines for Developing Multiple-Choice Items." *Methodology*, 2(2), 65-72.

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., e Jones, A. C. (2006). "Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course." *J Dent Educ*, 70(4), 378-86.

Simpson, J. M., Ryan, P., Carlin, J. B., Gurrin, L., e Marschner, I. (2009). "Training a New Generation of Biostatisticians: A Successful Consortium Model." *Journal of Statistical Education*, 17(2), 1-11.

Zelen, M. (2006). "Biostatisticians, biostatistical science and the future." *Stat Med*, 25(20), 3409-14.

Pós-Doc

Tese de Doutoramento: Conjuntos Difusos: Uma Abordagem Estatística

(Boletim SPE Outono de 2007, p. 48)

Abdul Suleman, *abdul.suleman@iscte.pt*

ISCTE - IUL Instituto Universitário de Lisboa

A área de investigação científica em que me encontro inserido é a mesma que deu origem à minha dissertação de doutoramento, ou seja, aplicação estatística de conjuntos difusos. No essencial, tenho procurado uma abordagem classificatória baseada em conjuntos difusos na resolução de problemas comuns, onde as modelações baseadas na classificação tradicional revelam fraca proficuidade, particularmente em presença de populações heterogéneas. A investigação incorpora projectos de duas naturezas distintas: uma exploratória, em que se procura avaliar a pertinência da abordagem difusa em temáticas específicas, e implica uma discussão alargada em fóruns especializados, e outra de continuidade, que se traduz numa nova abordagem de trabalhos anteriormente desenvolvidos ou ainda o seu aprofundamento a preocupações emergentes. O contexto da aplicação inclui a resolução de problemas concretos de três áreas do saber - Mercados de Trabalho, Economia das Competências e Saúde Pública -, e tem por objectivo não só identificar a estrutura que governa o universo em análise, como ainda criar instrumentos analíticos que permitam quantificar grandezas específicas de cada uma dessas áreas. A contextualização de resultados estabelecidos na teoria dos conjuntos difusos é um recurso a ter em conta na definição de tais instrumentos.

A face visível da minha investigação pode ser aferida por quatro acontecimentos relevantes:

i) a aceitação de um artigo, sem revisões, pela revista *Quality & Quantity* (DOI: http://www.springerlink.com/content/y003450501357rw0/) onde se apresenta um resultado inovador, sob a forma de um teorema, o qual permite, sob certas condições, ordenar trabalhadores por competência;

ii) participação num projeto internacional, financiado pela FCT - Fundação para Ciência e Tecnologia, que tem como proposta de análise quantitativa um modelo baseado em conjuntos difusos;

iii) participação num projeto nacional, intitulado "Classificação de doentes de Medicina Física e de Reabilitação" cujo objectivo principal é a classificação de doentes em regime de internamento e a identificação de perfis homogéneos, a partir de um conjunto de variáveis quantitativas e qualitativas;

iv) orientação de uma tese de doutoramento sobre o risco do crédito ao consumidor, risco esse (também) avaliado por um instrumento analítico suportado na teoria dos conjuntos difusos.

Tese de Doutoramento: Testes de tendência com Aplicação à Avaliação da Qualidade da Água

(Boletim SPE Primavera de 2008, p. 50)

Maria do Rosário Ramos, *marosram@univ-ab.pt*

Departamento de Ciências e Tecnologia, Universidade Aberta

Caros colegas

Devo confessar-vos que no final do dia das provas públicas realizadas na Universidade de Lisboa no dia 13 de Novembro de 2006, a óbvia sensação de felicidade e contentamento estava contaminada por uma pequena angústia. De facto, depois de explorar e testar algumas vias para abordar o problema principal definido na tese e de aplicar as metodologias em séries de dados reais, outras questões surgiram, tanto relativamente à escolha das metodologias teóricas como a situações sugeridas pelos próprios dados. A minha pequena angústia era que eu queria continuar a desenvolver mais os estudos, colocá-los na tese, mas não tinha tempo para o fazer dentro dos limites dos prazos definidos no estatuto da carreira docente universitária. Sossegou-me um pouco a minha orientadora, a Professora Teresa Alpuim que me trouxe à terra e me recordou que a carreira de investigação estava apenas a começar e que tinha o resto da vida…

Na tese foi abordado problema da análise da tendência monótona numa série temporal em algumas situações que põem em causa a fiabilidade dos resultados dos testes mais usuais. Foi considerado o problema da autocorrelação na série, com incidência nos processos de médias móveis MA e AR(1), a situação em que o número disponível de observações é reduzido e quando a distribuição dos valores da série apresenta uma assimetria acentuada. Contemplou-se ainda um possível efeito sazonal.

Este tipo de dados surge frequentemente em séries de variáveis de qualidade da água, tanto de tipo físico-químicas como bacteriológicas e daí o seu interesse prático. Investigaram-se as consequências destas características específicas sobre o teste ao estimador dos mínimos quadrados do declive (que nos dará a direção da tendência) e sobre uma alternativa não paramétrica, o teste de Mann-Kendall, o

qual tem um excelente comportamento quando usado em séries de observações independentes não normais e de dimensão reduzida. Concluiu-se que ambos os tipos de testes são fortemente afetados pela existência de correlação serial, que é necessário adaptar a estatística de teste, e que a própria estimação do parâmetro autorregressivo (por exemplo) é fraca e influencia quase igualmente qualquer uma das estatísticas de teste. Paralelamente a estes resultados de caráter mais teórico, deparámo-nos com ocorrências nos dados, como dados em falta na série e a existência de valores discrepantes (análise de outliers?). Obviamente que este é um problema suplementar e que vai afetar não só a estimação do parâmetro autoregressivo como a do próprio declive.

Propus-me então continuar com o estudo destas situações, procurar melhores estimadores para a autocorrelação, investigar outro tipo de modelos para modelar a tendência, considerar variação espacial e esperar por mais dados das séries de qualidade da água. Entretanto passei a membro integrado do centro de Matemática e Aplicações Fundamentais da UL.

A Estatística Ambiental é uma área que me agrada bastante, assim como a Estatística aplicada à Saúde e à Medicina.

Nos anos mais recentes fui desafiada por uma colega e amiga de longa data para fazer uma pequena colaboração em estudos sobre a identificação de preditores genéticos, físicos e demográficos da obesidade e de doenças cardiovasculares. Os estudos estão no âmbito de projetos do Laboratório de Genética da Faculdade de Medicina da Universidade de Lisboa, são aliciantes mas têm associados alguns obstáculos. Não é fácil obter uma "boa amostra". Os indivíduos são os que estão disponíveis nos registos médicos, por vezes não há registos informatizados de todas as análises médicas realizadas, faltam valores de variáveis e as dimensões dos grupos são bastante desequilibradas. Encontrar um modelo de regressão categorial com um bom poder discriminatório e explicativo nem sempre é conseguido. Mas uma conclusão válida, pequena que seja, tem um efeito muito positivo sobre os colegas da Medicina e isso também nos realiza, mesmo não sendo uma investigação em Estatística propriamente dita.

No ano de 2008 integrei a equipa do Projecto Determinantes Sociais e Psicológicos do Comportamento Alimentar e Infantil da Universidade de Lisboa, faculdade de Psicologia, sendo a estatística da equipa. O projeto foi financiado pela FCT e pela Nestlé.

O objetivo principal é contribuir para o conhecimento da relação entre os determinantes parentais e as preferências e comportamentos alimentares dos filhos de 5 e 6 anos de idade. Neste contexto investigam-se os determinantes psicológicos e comportamentais e possíveis interações entre eles.

A minha participação individual no projeto diz respeito à análise estatística dos dados após todo o trabalho de campo realizado por psicólogas e técnicas de saúde na recolha de dados, através de entrevistas e de questionários. O trabalho envolve métodos estatísticos descritivos, Regressão Multinomial, um estudo mais alargado de Modelos Lineares Generalizados e alguma pesquisa e aplicação de métodos de imputação de dados para respostas do tipo qualitativo ou ordinal.

Como acontece com qualquer colega que está na carreira docente, estes trabalhos têm de ser articulados com mais uma atividade principal, a docência, e com as componentes administrativas referentes à coordenação de cursos formais e de outras ofertas pedagógicas.

A situação financeira atual não permite a contratação de novos docentes, pelo menos no número desejado, todos temos de elaborar novos "produtos" que sejam atrativos e tragam novos públicos à Universidade. Uma consequência é a carga letiva de formação inicial excessiva, dificultando a dedicação regular à investigação, à organização e escrita de um artigo e à colaboração, a qual de modo indireto beneficiaria os alunos de mestrado que existem e os de doutoramento que venham a existir.

Acresce-se o facto da particularidade da Universidade Aberta, na qual praticamente todo o ensino é realizado através de uma plataforma de e-Learning desde o ano letivo 2007/2008, ano em que os cursos foram adequados a Bolonha. Este método de ensino colocou-nos o desafio de ensinar Matemática e Estatística Online, duas áreas com as dificuldades que já conhecemos no ensino presencial. Foi necessária uma adaptação às ferramentas de comunicação da matemática *online*, produção de novos materiais, uma elevada disciplina na gestão de datas e distribuição das matérias, e lidar de modo assíncrono com alunos que estão em qualquer lugar do mundo, em vários fusos horários.

Disse-me um colega dois dias depois das provas que depois do doutoramento o céu deixaria de ser azul tais seriam as novas tarefas que me esperavam, no entanto, as partes positivas existem e são muito gratificantes, seja devido a um resultado obtido, seja pelo sucesso dos nossos alunos e, porque não assumir, pelos momentos de partilha e recreativos que são vividos nos encontros científicos da comunidade académica.

Até um próximo encontro

Maria do Rosário Ramos

Tese de Doutoramento: Incorporating measurement error and density gradients in distance sampling surveys

(Boletim SPE Primavera de 2008, p. 49)

Tiago A. Marques, *tiago@mcs.st-and.ac.uk*

University of St Andrews, Centre for Research into Ecological and Environment Modelling
Centro de Estatística e Aplicações da Universidade de Lisboa

Terminei o meu doutoramento em fins de Junho de 2007, 15 dias depois nasceu o Filipe, o nosso primeiro filho, e mais quinze dias depois, a 22 de Julho, por coincidência no meu dia de anos (bom, os mais cínicos verão aqui apenas uma tentativa descarada e desesperada de receber mais prendas para o ano que vem!), comecei o meu pos-doc (uma Research Fellow, que é mais pomposo!). Foram de facto dias loucos, e se tenho um conselho a dar para quem esteja nesta vida de docs-pos-docs é que, se puderem, tirem umas boas férias a seguir ao doutoramento. A vida não é só trabalho, mas se deixarmos, parece! Agora já me deixei disso...

Apesar de ter começado a trabalhar na Universidade de St Andrews, no Reino Unido, o facto de ter tido o Filipe fez com que optasse por viver em Portugal, indo a St Andrews apenas "quando fosse preciso". Essa abertura para trabalhar remotamente surgiu num contexto específico: já tinha feito lá o doutoramento a tempo parcial, com uma bolsa mista da FCT, passando metade do tempo em Portugal. Por isso, as pessoas com quem trabalhava já sabiam que os períodos passados em Portugal não eram apenas férias: voltava a terras de Sua Majestade sempre com trabalho feito. Na realidade, também fruto de um trágico acidente com a minha mulher em Setembro de 2009, a verdade é que nos últimos 4 anos só estive em St Andrews duas vezes, uma para substituir uma professora durante a sua licença de maternidade, e outra para um congresso. E aqui tenho de tirar o chapéu ao meu "chefe", colega e amigo, Len Thomas, que me permitiu sempre gerir o meu tempo sem restrições, chegando de facto a, por vezes, ser ele a insistir para eu tirar férias. E a propor-me assinar um segundo contrato de 3 anos para outra Reserch Fellow, mesmo depois de ter passado um ano quase sem trabalhar, num momento em que do ponto de vista pessoal me encontrava uma lástima. Agora sinto-me melhor, e se puder e depender de mim, enquanto me quiserem por lá, por lá estarei.

O meu primeiro contrato foi realizado no âmbito de um grande projecto, o DECAF - *Density Estimation for Cetaceans from passive Acoustic Fixed sensors*. O DECAF era um projecto ambicioso, porque pretendia desenvolver metodologias de estimação de abundância e densidade usando dados acústicos, algo que em 2007 ainda não tinha sido tentado na prática, mas para o qual todos os ingredientes já estavam disponíveis. Felizmente, o projecto foi um sucesso, com mais de uma dezena de artigos publicados em revistas internacionais e uma série de cursos e workshops implementados. Um acontecimento importante neste primeiro contrato foi o convite recebido por parte de um colega que se tornou um amigo, o Yorgos Stratoudakis, para partilhar com ele um gabinete no IPIMAR, permitindo-me assim ter um "local de trabalho" eem Portugal, em vez de trabalhar em casa. Estive por lá quase dois anos, e foi uma excelente experiência. Ao longo deste primeiro contrato continuei a

trabalhar em metodologias de estimação de abundância em populações animais, e alem da metodologia de amostragem por distâncias, que já conhecia bem do doutoramento, comecei a contactar com metodologias de captura recaptura espacialmente explicita, uma outra área de métodos para a estimação da densidade animal, que antevejo vir a desenvolver-se muito nos anos que ai vêm. Foi também o período em que pela primeira vez implementei uma análise Bayesiana, um facto relevante na carreira de qualquer "estatístico" que se prese! Depois do acidente da minha mulher e um período complexo a trabalhar em casa e em hospitais não voltei ao IPIMAR, não porque não continuasse a ser bem-vindo, mas porque me parecia mais natural tentar arranjar um local mais próximo de outros estatísticos. Foi nesse sentido que falei com o Dinis Pestana, responsável por me iniciar nestas lides, professor, mentor e amigo, que sempre me apoiou, para perceber se haveria a possibilidade de arranjar um gabinete na Faculdade de Ciências (da Universidade de Lisboa, FCUL). Afinal, sendo membro integrado do Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), seria natural trabalhar mais próximo de outros membros do CEAUL. E assim foi, passei a trabalhar na FCUL, num gabinete que partilhava com mais algumas pessoas, mas cujo único habitante quase permanente era eu. Um outro facto relevante foi passado uns meses ter vindo ocupar um outro lugar na mesma sala a Regina Bispo, mais uma vez uma colega, aluna (de doutoramento, mas daquelas com quem pouco se ensina e tudo se partilha), professora e agora tambem amiga. Felizmente, sempre vou fazendo amigos por onde passo. E isso foi bom porque passei a ter um interlocutor de trabalho não só em Portugal, mas no meu gabinete. Falamos a mesma linguagem, já não é a língua dos p's das crianças, é mais a linguagem dos r's, ou do R. E este facto parece determinante para a minha evolução futura, porque com a excepção do Russel Alpizar-Jara, da Universidade de Évora, com quem tenho vindo a tentar colaborar mais recentemente, e que de certa forma assistiu ao meu "nascimento estatístico" (foi meu arguente na tese de mestrado) e contribuiu para o meu crescimento inicial, a Regina é das poucas pessoas que trabalha na parte estatística da estimação de populações animais (no caso dela, os animais estão mortos... mas é conceptualemente a mesma coisa...Humm... A vida e a morte não são conceptualmente diferentes, um digno paradoxo!).

Durante estes primeiros anos pós-doc, de certa forma comecei a amadurecer um pouco as minhas ideias quanto ao que sei e não sei. Sei um pouco de muita coisa, e não sei muito de quase nada. Gosto de ver dados novos, gosto de pensar em problemas novos, interesso-me por quase tudo o que me passa pelas mãos, ler trabalhos de amigos e colegas e fazer sugestões, rever trabalhos de outros, pensar em problemas práticos e tentar resolve-los. Mas gosto assim, acho que é um luxo poder faze-lo. E aqui acho que vou deambular um pouco sobre aquilo em que me estou a tornar. Se há 4 anos, apesar do doutoramento em estatística, achava que nunca seria um estatístico, neste momento o que tenho a crteza é que já não sou bem um biólogo. Que raio, acho que sou só o pai do Filipe e o marido da Luisa, como rótulos chega. De resto, vou felizmente trabalhando naquilo que gosto e sem ter de me definir mais do que alguém que trabalha em estatística aplicada à biologia. Serei um bio - estatístico?

O DECAF foi um precursor natural do projecto em que trabalho atualmente, o LATTE (*Linking Acoustic Tests and Tagging using statistical Experimentation*) para os amigos, também conhecido por "*Modeling the Behavior of Beaked Whales in Response to Medium Frequency Active Sonar: Integrating Experimental Tagging Studies with Opportunistic Passive Acoustic Monitoring*". DECAF, LATTE, não percebo bem estes acrónimos, eu nem gostava muito de café e seus derivados - embora agora lhe esteja a tomar o gosto. Temo que o proximo projecto se venha a chamar bica, garoto ou capucino. A ideia aqui é propor e, ainda mais difícil, ajustar a dados reais, modelos de movimento animal que permitam integrar informação acustica de cetáceos obtida a várias escalas. Estas vão desde uma escala muito fina, ao nível de animais com sensores acústicos acoplados que fornecem posições em 3D em tempo real, até a uma escala grosseira, ao nível de um campo de hidrofones montados no fundo do oceano e os sons neste detetados. A ideia é conseguir depois detectar alterações nos

parãmetros dos modelos na presença de perturbações externas, por exemplo uso de SONARes navais, e assim permitir perceber e quantificar melhor o impacto destes nas populações de cetáceos. Se o DECAF era ambicioso mas praticável, o LATTE é potencialmente megalómano e impraticável. É um a ver vamos...

Em relação ao meu percurso profissional, parece-me que se avizinham aqueles que serão os anos do fracasso ou da consolidação. Um jogador sozinho nada faz, e a investigação é na minha opinião, um trabalho de equipa. Penso que tanto como eu próprio, as pessoas acima mencionadas serão fundamentais para que o pêndulo se incline mais para um lado ou para o outro. Já desisti de ganhar um prémio Nobel, ou até mesmo o prémio da SPE. Já fico feliz se alguém usar na prática alguma das coisas que eu desenvolver!

Tiago Marques

Ciência Estatística

• Artigos Científicos Publicados

Amorim, A. P., Jacobo de Uña-Álvarez, Luís Meira-Machado (2011). Presmoothing the transition probabilities in the illness-death model, *Statistics & Probability Letters*, 81(7), 797-806.Bastos, J. A. and Caiado, J. (2011). Recurrence quantification analysis of global stock markets. *Physica A: Statistical Mechanics and its Applications*, 390, 1315-1325.

Barreto-Hernandez, E., Gama-Carvalho, M. and Sousa, L. (2010). Pre-processing Optimization of RNA Immunoprecipitation Microarray Data. *Journal of Computational Biology* 18(10): 1319-1328.

Borchers, D. L. , Marques, T., Gunnlaugsson, T. & P. Jupp (2010). Estimating distance sampling detection functions when distances are measured with errors. *JABES*. 15: 346-361.

de Carvalho, M. (2011). Confidence Intervals for the Minimum of a Function Using Extreme Value Statistics, *International Journal of Mathematical Modeling and Numerical Optimization*, 2, 288–296.

Duarte Silva, A. P. (2011) Two-group classification with high-dimensional correlated data: A factor model approach, Computational Statistics and Data Analysis 55, 2975–2990.

Faustino, C.E.S., M.A. Silva, T.A. Marques & L. Thomas (2010). Designing a shipboard line transect survey to estimate cetacean abundance off the Azores archipelago. *Arquipelago - Life and Marine Sciences*. 27: 49-58.

Grilo, L. M. e Coelho, C. A. (2010). The exact and near-exact distribution for the Wilks Lambda statistic used in the test of independence of two sets of variables. *American Journal of Mathematical and Management Sciences*, **30** (2) 111-140.

Heide-Jørgensen, M.; Laidre, K.; Borchers, D.; Marques, T. A.; Stern, H. & Simon, M. (2010). The effect of sea-ice loss on beluga whales (delphinapterus leucas) in west greenland. *Polar Research*. 29: 198-208.

Heide-jørgensen, M. P., K. L. Laidre, M.L. Burt, D.L. Borchers, T. A. Marques, R. G. Hansen, M. Rasmussen and S. Fossette (2010). Abundance of narwhals (Monodon monoceros L.) on the hunting grounds in Greenland. *Journal of Mammalogy* 91: 1135-1151.

Kusel, E., Mellinger, D. K., Thomas, L., Marques, T.A., Moretti, D. & Ward, J. (2011). Cetacean population density estimation from single fixed sensors using passive acoustics. *The Journal of the Acoustical Society of America*. 129: 3610-3622.

Marçalo, A., Marques, T. A.; Araújo, J., Pousão-Ferreira, P, Erzini, K. & Stratoudakis, Y. (2010). Fishing simulation experiments for predicting effects of purse seine capture on sardines (sardina pilchardus). *ICES Journal of Marine Sciences*. 67: 334-344.

Marques, T. A., Buckland, S. T., Borchers, D. L., Tosh, , D. & Mcdonald, R. A. (2010). Point transect sampling along linear features. *Biometrics*. 66: 1247-1255.

Marques, T. A.; Munger, L.; Thomas, L.; Wiggins, S. & Hildebrand, J. A. (2011). Estimating North Pacific right whale (Eubalaena japonica) density using passive acoustic cue counting. *Endangered Species Research*. 13: 163-172.

Marques, T. A., Thomas, L. and J. A. Royle (2011). A hierarchical model for spatial capture-recapture data: *Comment. Ecology*. 92: 526-528.

Meira-Machado, L. and Javier Roca-Pardiñas. (2011). p3state.msm: Analyzing survival data from an illness-death model. *Journal of Statistical Software*, 38(3), 1-18.

Moretti, D., T.A. Marques, L. ThomaS, N. Dimarzio, A. Dilley, R. Morrissey, E. Mccarthy, J. Ward and S. Jarvis (2010). A dive counting density estimation method for Blainville's beaked whale (Mesoplodon densirostris) using a bottom-mounted hydrophone field as applied to a Mid-Frequency Active (MFA) sonar operation. *Applied Acoustics.* 71: 1036-1042.

Noirhomme-Fraiture, M., Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining*, Volume 4, Issue 2, pp.157-170.

van Keilegom, I., Jacobo de Uña-Álvarez and Luís Meira-Machado (2011). Nonparametric location-scale models for successive survival times under dependent censoring. *Journal of Statistical Planning and Inference.* 141: 1118-1131.

Ward, J., Jarvis, S., Moretti, D., Morrissey, R., Dimarzio, N., Thomas, L. & Marques, T. A. (2011). Beaked whale (Mesoplodon densirostris) passive acoustic detection with increasing ambient noise. *The Journal of the Acoustical Society of America*. 129: 662-669.

• Teses de Mestrado

Título: Modelos Estocásticos de Crescimento Individual e Desenvolvimento de Software de Estimação e Previsão
Autor: Nuno Miguel Baptista Brites, *d7488@alunos.uevora.pt*
Orientador: Carlos Alberto dos Santos Braumann

Título: Problemas Mal Resolvidos em Amostragem
Autor: António Alberto Oliveira, *a.a.oliveira@hotmail.com*
Orientadores: Dinis Pestana e Fernanda Diamantino

• Capítulos de Livros

Caeiro, F., **Gomes, M.I.** and **Vandewalle, B.** (2010). Semi-Parametric Probability-Weighted Moments Estimation Revisited. *IWAP* 2010: *International Workshop on Applied Probability*. On-line publication.

Cordeiro, C., Machás, A. and **Neves, M.M.** (2010). "A Case Study of a Customer Satisfaction Problem: Bootstrap and Imputation Techniques", *Handbook of Partial Least Squares Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics, Esposito Vinzi, V.; Chin, W.W.; Henseler, J.; Wang, H. (Eds.), 279-288.

Fraga Alves, M.I., Neves, C. and Cormann, U. (2010): Heavy and Super-Heavy Tail Analysis. In Falk, M., Hüsler, J. and Reiß, R.-D., *Laws of Small Numbers: Extremes and Rare Events*, Third Edition, Springer-Basel, ISBN: 978-3-0348-0008-2, Chapter 2, Section 2.7, pgs 75-101

Fraga Alves, M.I. and Neves, C. (2010). Extreme Value Distributions. In Lovric, M. (Ed.), *International Encyclopedia of Statistical Science*, Springer-Verlag, 1st Edition, ISBN: 978-3-642-04897-5. Encyclopedia released on December 19, 2010.

Gomes, M.I. (2010). Statistical Process Control. In Lovric, M. (Ed.), *International Encyclopedia of Statistical Science*, Springer-Verlag, 1st Edition, ISBN: 978-3-642-04897-5. Encyclopedia released on December 19, 2010.

Gomes, M.I. (2010). Acceptance Sampling. In Lovric, M. (Ed.), *International Encyclopedia of Statistical Science*, Springer-Verlag, 1st Edition, ISBN: 978-3-642-04897-5. Encyclopedia released on December 19, 2010.

Marques, T. A., Buckland, S. T., Borchers, D. L., Rexstad, E. & Thomas, L. (2011), Distance Sampling, Pages 398-400 in International Encyclopedia of Statistical Science (ed. L. Miodrag), Springer.

• Livros

Título: *Análise de Dados Longitudinais*
Autoras: Maria Salomé Cabral e Maria Helena Gonçalves
Ano: 2011. Edições SPE. ISBN: 978-972-8890-24-7

Título: *A Matemática, a Estatística e o Ensino nos Estabelecimentos de Formação de Oficiais do Exército Português no Período 1837-1926: Uma Caracterização.*
Autor: Filipe Papança
Ano: 2011. Editora: ediumeditores. ISBN: 978-989-701-035-4

Título: *ANÁLISE DE DADOS, Aplicações às Ciências Económicas e Empresariais.*
Autores: António St.Aubyn e Nuno Venes
Ano: 2011. Editora: Verlag Dashöfer. ISBN: 978-989-642-185-4

Título: *Controlo Estatístico da Qualidade.*
Autoras: Maria Ivette Gomes, Fernanda Figueiredo e Maria Isabel Barão
Ano: 2010. Editora: Edições INE. 2ª Edição, revista. ISBN: 978-972-8890-23-0

• Teses de Doutoramento

Título: Modelação e Avaliação de Desempenho de Redes Móveis Ad Hoc
Autor: Gonçalo João Costa Jacinto, *gjcj@uevora.pt*
Orientadores: António Manuel Pacheco Pires e Nelson Gomes Rodrigues Antunes

Na minha tese procurou-se modelar as redes de telecomunicações móveis *ad hoc* quer ao nível da mobilidade dos nós, quer ao nível da conectividade existente entre um nó emissor e um nó receptor.
As redes de telecomunicações móveis *ad hoc* são constituídas por nós que se organizam de forma autónoma e sem qualquer infraestutura, sendo uma das mais promissoras modernizações das actuais redes de telecomunicações sem fios. A mobilidade e a possibilidade de comunicação por rotas com múltiplos passos torna a topologia destas redes dinâmica e imprevisível, sendo necessário desenvolver modelos que descrevam a conectividade e a dinâmica dessas rotas.
A investigação inicia-se com o estudo da conectividade para redes unidimensionais e bidimensionais. É derivada a distribuição de probabilidade do número de passos duma rota quando a distribuição espacial dos nós provém de um processo de Poisson ou, utilizando o método de aleatorização de Poisson, quando um número fixo de nós está uniformemente distribuído numa dada região. Resultados numéricos ilustram o comportamento da distribuição de probabilidade do número de passos duma rota.
De seguida é desenvolvido um modelo para caracterizar a dinâmica das rotas através de um processo de Markov determinístico por troços. A distribuição e o tempo médio de duração das rotas são derivados, sendo estes resultados obtidos através de um sistema de equações integro-diferenciais. Um método recursivo é proposto para sua computação. Resultados numéricos ilustram o cálculo destas medidas, os quais são comparados com os obtidos quando se assumem rotas com passos independentes.

Gonçalo Jacinto

Título: Modelo combinado captura-recaptura e transectos lineares: Uma abordagem bayesiana
Autor: Filipe Monteiro, *jfgm@uevora.pt*
Orientador: Russell Alpizar

Na minha tese apresento uma abordagem bayesiana para estimar a probabilidade de detectar um animal/objecto na distância zero, conhecida como g_0, utilizando o modelo combinado de captura-recaptura e transectos lineares (Alpizar-Jara e Pollock, 1999, Em *Marine Mammal Survey and Assessment Methods*, 99-114 pp). Um estimador para o tamanho da população pode ser enviesado se a heterogeneidade não for considerada na modelação das probabilidades de captura, relativa às características inerentes dos indivíduos que são difíceis de medir ou não observáveis. Este tipo de problema tem sido tradicionalmente abordado mediante os modelos de captura-recaptura para populações fechadas, designados por M_h e M_{th}. Nesta tese formulo um modelo generalizado combinado de captura-recaptura e transectos lineares para populações fechadas, que incorpora heterogeneidade nas probabilidades de detecção relativa às características inerentes dos indivíduos. A probabilidade de detectar um indivíduo em cima da linha do transecto percorrido, é estimada admitindo que é menor ou igual a 1. Assume-se que a probabilidade de avistar um animal depende de características individuais. A heterogeneidade observável nas probabilidades de captura dos indivíduos na população é modelada através da regressão logística utilizando covariáveis, tais como o sexo, a idade, ou o tamanho do grupo em que o animal se encontra. A heterogeneidade não observável é modelada através de um efeito aleatório, utilizando a inferência bayesiana. O parâmetro g_0 é estimado como sendo uma média baseada na informação dos indivíduos como se estivessem na linha do transecto. O desempenho dos estimadores da probabilidade de um indivíduo ser observado na distância zero é analisado através de simulações, realizadas no programa R, e comparada com as situações em que apenas é modelada a heterogeneidade observável ou quando são modeladas ambas as heterogeneidades, observável e não observável. As distribuições *a posteriori* dos parâmetros que determinam a função de detecção são obtidas através do método de amostragem Gibbs através do método de Monte Carlo baseado em cadeias de Markov implementado no WINBUGS. Os resultados são ilustrados com dados reais da população de ungulados de montanha (*Rupicapra p. pyrenaica*) do Parque Nacional dos Pirenéus (sul da França).

Filipe Monteiro

Título: Análise de dados categorizados com omissão em variáveis explicativas e respostas
Autor: Frederico Poleto, *fpoleto@gmail.com*
Orientadores: Julio Singer e Carlos Daniel Paulino

Nesta tese apresentam-se desenvolvimentos metodológicos para analisar dados com omissão e também estudos delineados para compreender os resultados de tais análises, sempre ilustradas com conjuntos de dados reais.

Escrutinam-se análises de sensibilidade bayesiana e clássica para dados com respostas categorizadas sujeitas a omissão. Mostra-se que as componentes subjetivas de cada abordagem podem influenciar os resultados de maneira não trivial, independentemente do tamanho da amostra, e que, portanto, as conclusões devem ser cuidadosamente avaliadas. Especificamente, demonstra-se que distribuições *a priori* comumente consideradas como não-informativas ou levemente informativas podem, na verdade, ser bastante informativas para parâmetros inidentificáveis, e que a escolha do modelo sobreparametrizado também tem um papel importante.

Quando há omissão em variáveis explicativas, também é necessário propor um modelo marginal para as covariáveis mesmo se houver interesse apenas no modelo condicional. A especificação incorreta do modelo para as covariáveis ou do modelo para o mecanismo de omissão leva a inferências enviesadas para o modelo de interesse. Trabalhos anteriormente publicados têm-se dividido em duas vertentes: ou utilizam distribuições semiparamétricas/não-paramétricas, flexíveis para as covariáveis, e identificam o modelo com a suposição de um mecanismo de omissão não-informativa, ou empregam distribuições paramétricas para as covariáveis e permitem um mecanismo mais geral de omissão informativa. Neste trabalho analisam-se respostas binárias, combinando um mecanismo de omissão informativa com um modelo não-paramétrico para as covariáveis contínuas, por meio de uma mistura induzida por um processo de Dirichlet.

No caso em que o interesse recai apenas em momentos da distribuição das respostas, propõe-se uma nova análise de sensibilidade sob o enfoque clássico para respostas incompletas que evita suposições distribucionais e utiliza parâmetros de sensibilidade de fácil interpretação. O procedimento tem, em particular, grande apelo na análise de dados contínuos, campo que tradicionalmente emprega suposições de normalidade e/ou utiliza parâmetros de sensibilidade de difícil interpretação.

Frederico Poleto

PRÉMIOS "ESTATÍSTICO JÚNIOR 2011"

Ensino Básico- Trabalho classificado em 1º lugar

Título: Uma questão de opções...
Autores: Ana Beatriz Sousa Pinto, Beatriz Cecília Santos Leite, Ângela Sofia Santos Leite
Professor orientador: Vasco Filipe de Magalhães Ribeiro
Estabelecimento de Ensino: Escola Básica e Secundária de Pinheiro – Pinheiro - Penafiel

Ensino Básico- Trabalho classificado em 2º lugar

Título: A crise económica na vida dos torrejanos
Autores: Mariana Sofia das Neves Cruz, Mariana Branco Farinha, Henrique Manuel Vinhas Nunes
Professora orientadora: Maria Alice da Silva Martins
Estabelecimento de Ensino: Agrupamento de Escolas Artur Gonçalves – Torres Novas

Ensino Básico- Trabalho classificado em 3º lugar

Título: 8º A: Quem e como somos
Autores: Petra Bebiana Miranda Carneiro, Cláudia Sofia Almeida Oliveira, Ana Margarida Freitas Fernandes
Professora orientadora: Carina Maria Martins Duarte da Silva
Estabelecimento de Ensino: Escola EB 2,3 Gil Vicente – Urgezes - Guimarães

Ensino Secundário- Trabalho classificado em 2º lugar

Título: Sociedade Online ou In loco
Autores: Carlos Moura Pereira Lucas Teixeira
Professor orientador: Nuno Duarte Veríssimo Rodrigues
Estabelecimento de Ensino: Escola Secundária 3 EB Dr. Jorge Correia - Tavira

Ensino Secundário- Trabalho classificado em 3º lugar

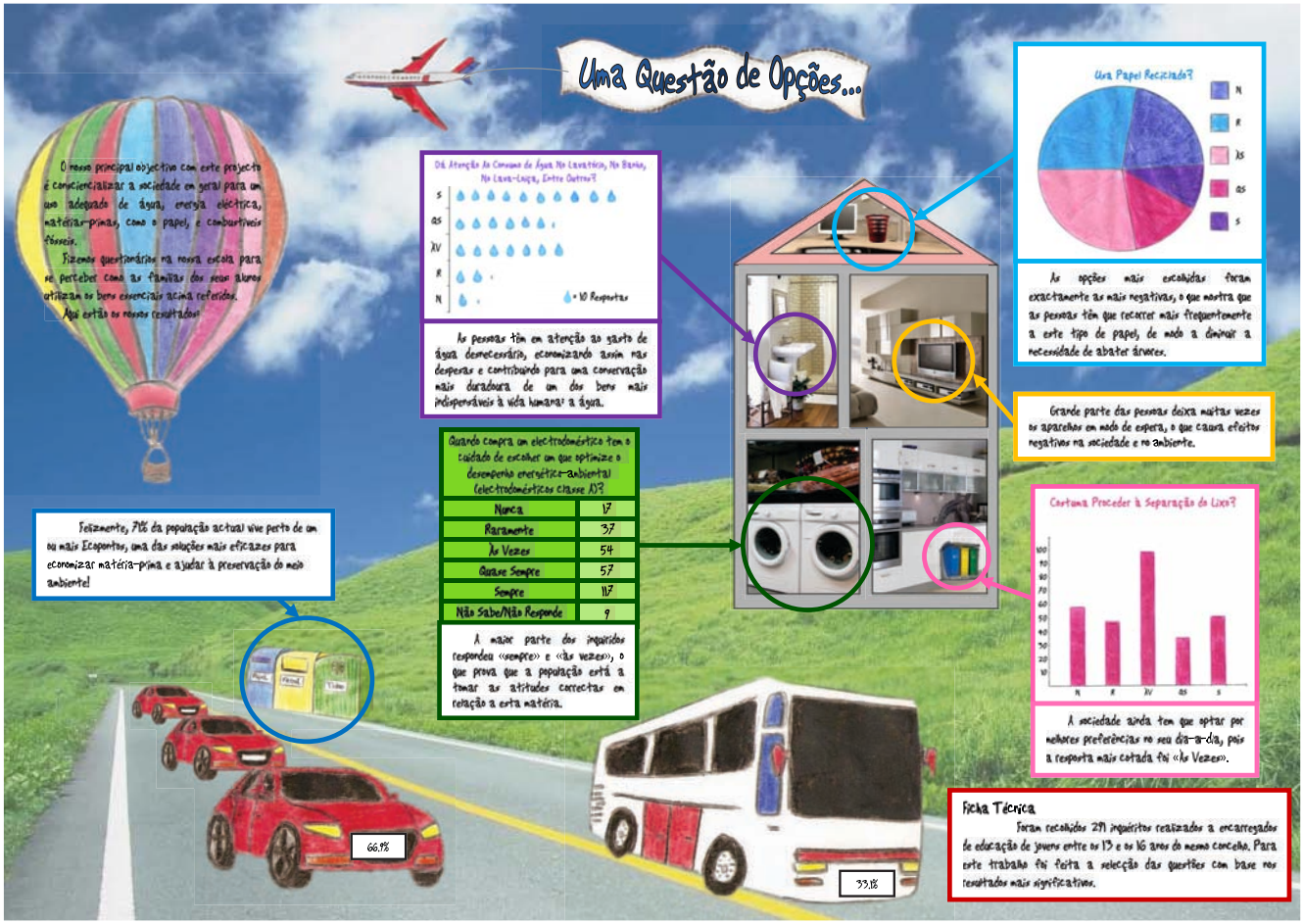
Título: Totoloto e Euromilhões
Autores: Pascoal Sebastião Matiue, Eurico Daniel Leite Teixeira, Fernando Daniel Magalhães de Abreu
Professora orientadora: Patricia Alexandra da Silva Ribeiro Sampaio
Estabelecimento de Ensino: Escola Profissional de Fermil – Celorico de Bastos

Cursos EFA/CEF- Trabalho premiado

Título: Utilização de redes sociais
Autores: Nádia Filipa Ferreira Ramos, Fábio André Fernandes Martins, Helder Emanuel Fernandes Ferreira
Professora orientadora: Maria Manuela R. Matos Alves Neto
Estabelecimento de Ensino: Escola EB 2,3 de Maria Lamas, Ramalde - Porto

PRÉMIOS "ESTATÍSTICO JÚNIOR 2011"

Trabalho Classificado em 1.º lugar (Ensino Básico)



PRÉMIOS "ESTATÍSTICO JÚNIOR 2011"

Cursos EFA/CEF - Trabalho Premiado

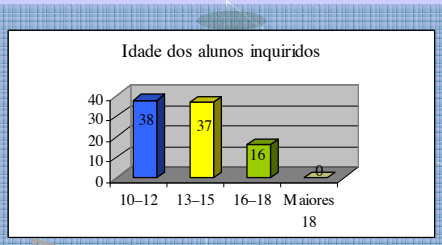
PRÉMIO ESTATÍSTICO JUNIOR 2011

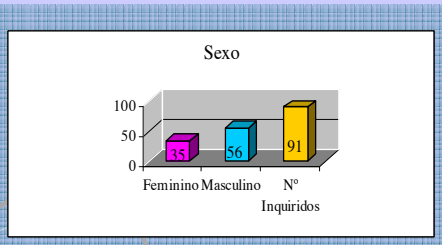
UTILIZAÇÃO DE REDES SOCIAIS

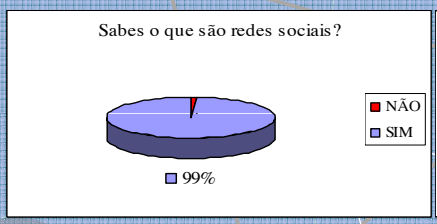
Escola EB 2,3 de Maria Lamas

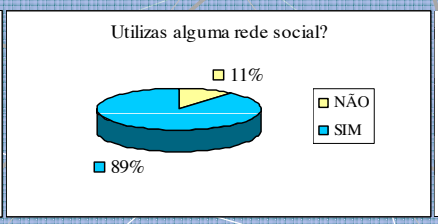
Nádia Ramos, Fábio Martins e Hélder Ferreira, Curso de Educação e Formação

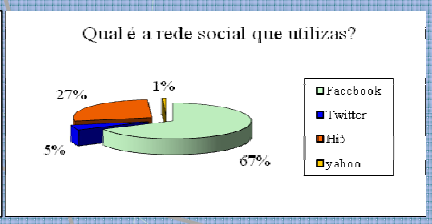
Orientação da Professora Manuela Neto

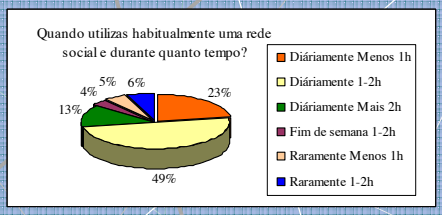


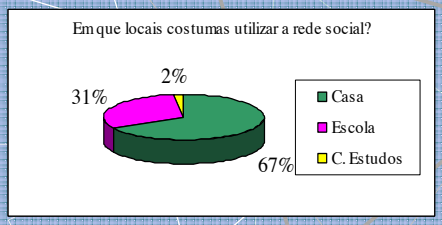


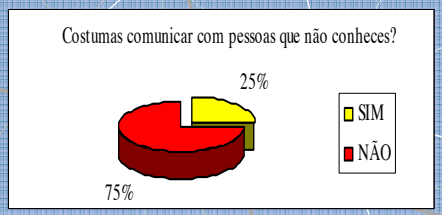


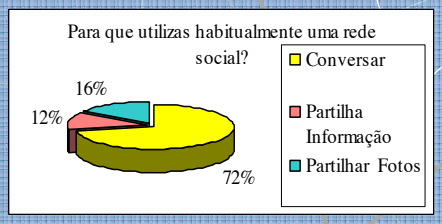


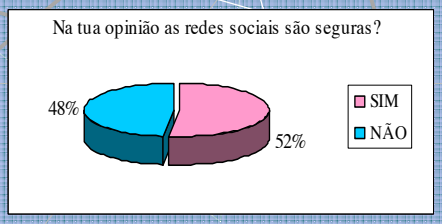












Resumo

Fases do planeamento do trabalho

1ª Selecção do tema (articulação com Língua Portuguesa)
2ª Pesquisa de informação (articulação com Língua Portuguesa)
3ª Elaboração dos inquéritos (Matemática)
4ª Selecção da amostra (Matemática) (91 inquéritos: 5.º - 9.º ano/CEF)
5ª Aplicação dos inquéritos aos alunos da escola
6ª Recolha e organização dos dados (articulação com TIC)
7ª Tratamento e análise dos dados (Matemática)
8ª Apresentação do trabalho (Matemática/TIC)
Recursos utilizados: Software Excel/Word/PowerPoint

Síntese dos resultados

- **Amostra: 91 inquéritos: 5.º - 9.º ano/CEF**
- **Só 1% dos alunos não sabe o que são redes sociais e a maioria utiliza-as.**
- **O Facebook é a rede social mais utilizada e a maioria dos alunos questionados está ligada a uma das redes menos de 1h por dia.**
- **A maioria dos alunos desta amostra utiliza as redes sociais em casa mas há 31% que utiliza na escola.**
- **A maioria utiliza as redes sociais para conversar.**
- **75% dos alunos conversa com desconhecidos e 42% coloca fotografias pessoais nas redes sociais.**
- **Quase metade dos inquiridos considera que as redes sociais são seguras.**

ESTATÍSTICOS JÚNIOR 2011

PRÉMIOS "ESTATÍSTICO JÚNIOR 2012"

Está aberto, até 25 de Maio de 2012, o concurso para atribuição de prémios **"Estatístico Júnior 2012"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2012"** é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística.

2. Os candidatos aos prémios **"Estatístico Júnior 2012"** devem ser alunos do 3.° Ciclo do Ensino Básico, do Ensino Secundário, dos Cursos de Educação e Formação (CEF), ou dos Cursos de Educação e Formação de Adultos (EFA), no ano lectivo 2011-2012.

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

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

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

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2012**" a sete trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.° Ciclo do Ensino Básico, aos três primeiros classificados de entre os trabalhos candidatos do Ensino Secundário e um primeiro classificado de entre os trabalhos candidatos dos Cursos CEF-EFA. Os prémios são constituídos por produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 600 euros, 300 euros e 200 euros, a atribuir, respectivamente, aos grupos cujos trabalhos sejam classificados em 1.°, 2.° e 3.° lugares, para as categorias Ensino Básico e Secundário, e 600 euros para a categoria Cursos CEF-EFA.

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

8. Aos grupos proponentes dos trabalhos classificados em 1º lugar será também oferecida uma ampliação do correspondente poster que será colocado na Sessão de Posters do XIX Congresso Anual da SPE.

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

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

O boletim de candidatura e este regulamento podem ser obtidos em
http://www.spestatistica.pt/BoletimCandidaturaPEJ12.pdf
http://www.spestatistica.pt/RegulamentoPEJ12.pdf

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

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

12. A atribuição dos prémios **"Estatístico Júnior 2012"** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será realizada no XIX Congresso Anual da SPE.

13. Os prémios **"Estatístico Júnior 2012"** poderão não ser atribuídos.

Apoio Porto Editora

Edições SPE - Minicursos

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Sociedade Portuguesa de Estatística
tem o prazer de o convidar a participar no novo projeto

Radical Estatística

Radical Estatística é um projeto "hands-on" dirigido a todos os alunos que frequentem no ano lectivo de 2011/2012 o 10º ou o 11º ano.

O Objetivo geral desta iniciativa é promover o interesse pela estatística junto dos mais jovens, culminando na organização do "Campo SPE Clube Júnior – Radical Estatística", a realizar durante o período da Páscoa de 2012, **totalmente gratuito** para os alunos e professores selecionados para esta grande final.

A inscrição é feita on-line em **www.radicalestatistica.net** e consiste no registo de uma equipa de 4 a 5 alunos e um professor responsável. No momento da inscrição, cada aluno terá de responder a um questionário on-line com vista a avaliar a sua apetência pela estatística, bem como a perceção que tem da importância desta ciência em fornecer métodos e técnicas necessárias para lidar com situações de incerteza.

As equipas inscritas participarão numa competição on-line em atividades ligadas a diversos temas da estatística como, por exemplo, interpretação de diversas representações gráficas, medidas de localização e dispersão e regressão linear simples.

Durante esta competição on-line (acessível em **www.radicalestatistica.net**), as **15 melhores equipas** serão então selecionadas para a **semifinal** onde serão selecionadas as **10 equipas finalistas** para o "**Campo SPE Clube Júnior – Radical Estatística**".

Na final, a recolha, o tratamento e a análise de dados serão realizados pelos alunos participantes, sendo convidados a participar em diversas atividades radicais no Campo de Aventuras. Para além das despesas totalmente pagas neste fim-de-semana **Radical Estatística** no Campo de Aventuras, as três primeiras equipas receberão ainda prémios bastante aliciantes.

Os **Professores** responsáveis pelas equipas que estarão na meia-final terão acesso a uma ação de formação acreditada pelo Conselho Científico Pedagógico da Formação Contínua, envolvendo os conteúdos programáticos de Estatística lecionados no ensino básico e no ensino secundário.

Os seus alunos não vão querer perder!

www.radicalestatistica.net

destinado a todos os alunos de 10º e 11º ano

+ ação de formação acreditada para professores

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

SPE CLUBE JÚNIOR

RADICAL estatística

janeiro / fevereiro 2012 // competição online
actividades ligadas a temas de estatística

março 2012 // semifinal
em Lisboa com as 15 melhores equipas

páscoa 2012 // campo SPE Clube Júnior – Radical Estatística
actividades estatísticas em ambiente radical (Campo Aventura)

www.radicalestatistica.net
para mais informações geral@radicalestatistica.net

Ciência Viva 2010 · Inovação 2010 · Universidade de Évora · cima-UE · CMDT

Índice