# Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral — Primavera de 2011

## Sondagens e Censos



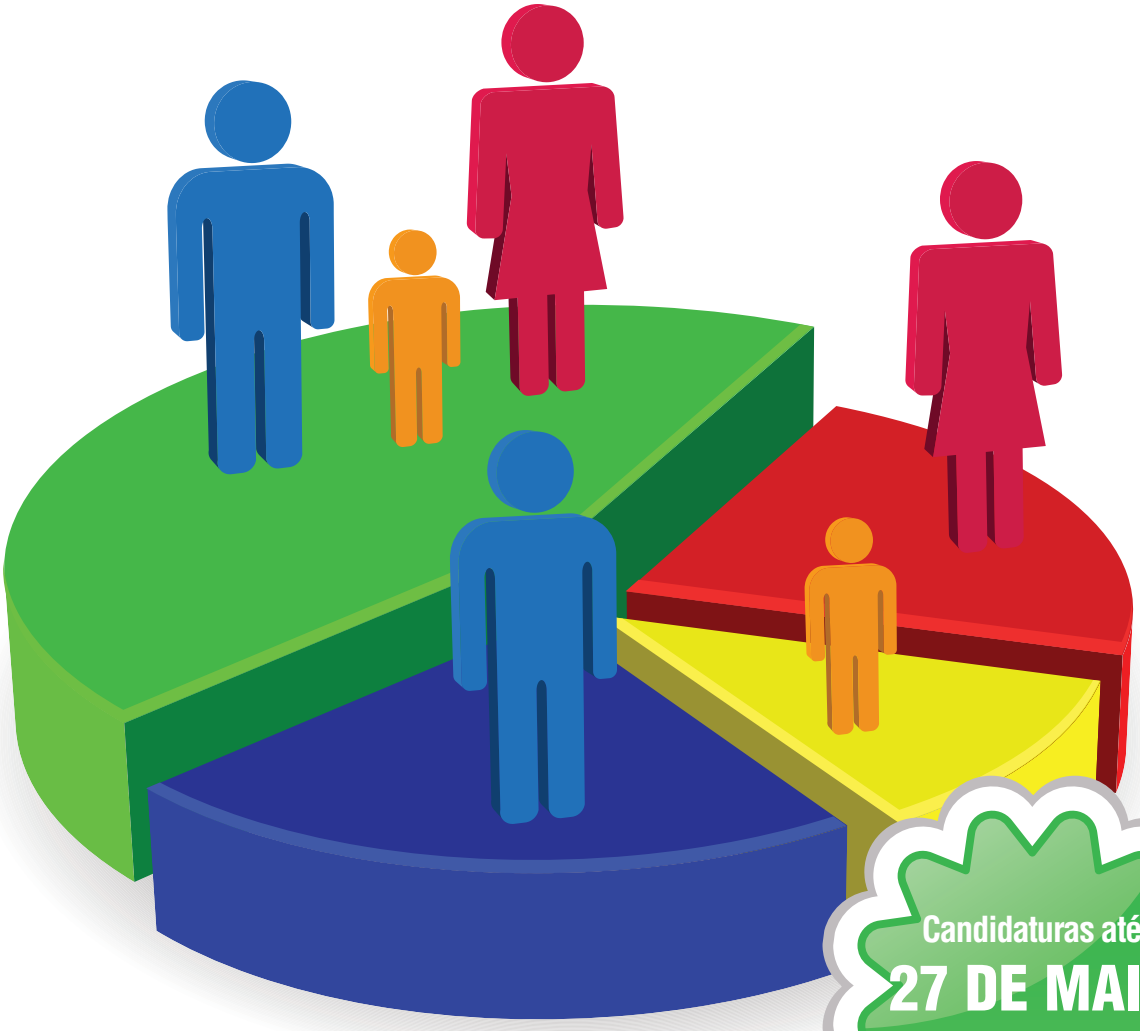
**Informação Editorial**

**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 2011



**Candidaturas até 27 DE MAIO DE 2011**


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

# Sociedade Portuguesa de Estatística

XIX Congresso - Nazaré 2011

Setembro 28 - Outubro 1

www.spe2011.ipleiria.pt

# Editorial

**… (com um pouco de história) da Estatística em Portugal e no mundo…**

No passado dia 20 de Outubro, conforme anunciado no Boletim anterior e de que damos o devido destaque nesta edição, celebrou-se o I Dia Mundial da Estatística. Essa foi uma data importante para a Estatística em 2010. A criação do Dia Mundial da Estatística é um pequeno passo no avanço científico mas, simbolicamente, um grande passo no reconhecimento da Estatística como Ciência que produz estatísticas que, cada vez mais, são importantes instrumentos de decisão.
E, esta feliz iniciativa decorreu no ano em que a SPE comemorou 30 anos de actividade cujo principal objectivo é a sedimentação da Estatística. Nessa longínqua data não existiam estatísticos portugueses. Assim, foi há trinta anos que se iniciou a "geração Estatística de Portugal" e ela tem produzido frutos nas mais diversas áreas científicas em todo o Mundo. Há pois uma dupla razão para celebrar!
De facto, a Sociedade Portuguesa de Estatística e Investigação Operacional foi fundada em 1980 usando a sigla SPEIO. Inicialmente era uma organização com interesses científicos em Estatística e Investigação Operacional.
Durante os primeiros dez anos consolidou-se e melhor se definiram os diferentes rumos científicos conduzindo em termos práticos na separação do grupo de investigação operacional que se juntou à então recém-criada Associação Portuguesa de Investigação Operacional (APDIO).
Desde 1985 até 1990 foi experimentada alguma inactividade. Em Junho de 1990 houve eleições. Ivette Gomes foi eleita Presidente e durante os tempos seguintes foi-se consolidando o projecto que conduziu à actual SPE.
A Sociedade Portuguesa de Estatística não foi um objectivo inicial dos primeiros cientistas que, em Portugal nos anos 70, se congregaram em torno da Estatística e da Investigação Operacional. No Editorial do Boletim Informativo de Estatística e Investigação Operacional publicado em Novembro de 1980 pelo Centro de Estatística e Aplicações (INIC) e que actualmente está integrado na Universidade de Lisboa, escreve o editor: "Antes de concluir este editorial não queremos deixar de assinalar a fundação da "Sociedade Portuguesa de Estatística e Investigação Operacional" de que damos notícia mais detalhada na rubrica "Noticiário"…" De facto, essa notícia é apresentada na página 29 do referido Boletim. Embora com uma modesta notícia, de facto, tinha sido dado um grande passo científico nos domínios da estatística e da investigação operacional fruto do dinâmico grupo de cientistas daquelas áreas que integravam o Centro de Estatística e Aplicações da Universidade de Lisboa (aliás em cujas instalações se encontra ainda a sede da SPE – art 2 dos estatutos).
Resumidamente, este é o ponto da situação de um ciclo que começou há 30 anos e que revisitamos neste ano em que se comemoram os 100 anos da Universidade de Lisboa e da FCUL – Faculdade de Ciências da Universidade de Lisboa, instituição de acolhimento mas também o berço da Sociedade Portuguesa de Estatística.

Um "simbólico dado" continua como bandeira desta associação científica. No início com uma sigla (5 pintas) SPEIO e, desde 1991 - há 20 anos - "reequilibrado" com SPE.


*Alea jacta est!*

O tema central do próximo *Boletim* será *Análise de Sobrevivência.*

# Mensagem do Presidente

Caros Colegas:

É altura de vos fazer um breve balanço do ano de 2010.

Pela primeira vez, o nosso Congresso Anual, o XVIII, foi organizado conjuntamente por uma instituição universitária e uma instituição politécnica, a Universidade de Coimbra e o Instituto Politécnico de Viseu, que nos juntaram em S. Pedro do Sul para a festa anual da Estatística em Portugal e de que vos dei notícia no Boletim anterior. Os nossos reiterados parabéns à Comissão Organizadora, sob a presidência do Paulo Oliveira e vice-presidência da Carla Henriques, bem como os agradecimentos a todos os que participaram nas mais variadas funções. As Actas do XVII Congresso, pela primeira vez na série internacional "Selected Papers in Statistics" do grupo Springer estão, depois da demora inicial para concretização dos acordos envolvidos nesta fase de arranque, em estado muito adiantado, esperando-se a sua conclusão dentro de poucos meses. As do XVIII Congresso seguem o mesmo modelo e estão já na fase de submissão dos artigos, mas, naturalmente, já não terão de passar pelas vicissitudes de desbravar terreno virgem.

O XIX Congresso, a decorrer de 28 de Setembro a 1 de Outubro de 2011, vai na mesma senda do anterior, com organização conjunta do Instituto Superior Técnico e do Instituto Politécnico de Leiria, sendo o colega António Pacheco o Presidente e a colega Alexandra Seco a Vice-Presidente. Promete vir a ser um grande Congresso, com excelentes convidados e com não menos excelentes participantes. Lá estaremos certamente todos para concelebrar mais um ano de grande actividade científica da comunidade estatística nacional e para mais um reencontro de amigos e o acolhimento de novos colegas.

Prosseguimos com as nossas actividades habituais, incluindo, entre outras, o apoio a diversas iniciativas, os Encontros CIM-SPE, o Prémio SPE 2010, os Prémios Estatístico Júnior 2010 com o apoio da Porto Editora e, obviamente, o infalível e cada vez melhor Boletim e os seus dois números anuais. O trabalho das nossas Comissões Especializadas tem continuado a bom ritmo. Tivemos a celebração, organizada pela Maria Antónia Amaral Turkman, do Dia Mundial da Estatística em colaboração com o Centro de Estatística e Aplicações da Universidade de Lisboa. Participámos numa reunião de 14 sociedades estatísticas europeias (representados pela Luísa Loura). Integrámos pela primeira vez a Comissão Nacional de Matemática, sinal de reconhecimento do papel da Sociedade, e tive a honra de, em Novembro de 2010, representar a Direcção da SPE na primeira reunião da Comissão em que participámos como membros de pleno direito (já tínhamos participado numa ou noutra reunião como observadores). Foi criada a jSPE, Secção dos Jovens Estatísticos da SPE, que teve a sua primeira actividade pública no Congresso de S. Pedro do Sul. Novas iniciativas se preparam para 2011, mas delas lhes daremos notícia quando estiverem mais maduras.

A todas as pessoas e instituições que connosco colaboraram em 2010, o nosso profundo agradecimento. Aos sócios, o nosso muito obrigado pela vossa participação activa na vida da Sociedade. É devido um agradecimento muito especial àqueles que nela exerceram as mais diversas actividades e funções e que assim permitiram um ano pleno de realizações.

Saudações cordiais

# Notícias

## • XIX Congresso SPE


A Direcção da Sociedade Portuguesa de Estatística (SPE) convidou a Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e o Instituto Superior Técnico da Universidade Técnica de Lisboa para colaborarem com a SPE na organização do XIX Congresso Anual da SPE - SPE2011. É com prazer que informamos que a SPE2011 decorrerá na Nazaré de 28 de Setembro a 01 de Outubro de 2011.

O Congresso será presidido pelo Professor Carlos Braumann, Presidente da SPE, sendo a Comissão Organizadora do Congresso constituída pelos professores: António Pacheco e M. Rosário de Oliveira, do Instituto Superior Técnico, e Alexandra Seco, Helena Ribeiro, Miguel Felgueiras e Rui Santos, da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria. Por sua vez, a Comissão Científica do Congresso será formada pelos professores: António Pacheco, António St. Aubyn, Carlos Braumann, Carlos Tenreiro e M. Ivette Gomes.

SPE2011 manterá a estrutura dos congressos recentes da SPE, incluindo a contribuição de comunicações orais e posters, sessões plenárias com ilustres convidados portugueses e estrangeiros, e um mini-curso precedendo o congresso propriamente dito. O mini-curso deste congresso terá por tema a Análise de Dados Longitudinais e será ministrado pela professora M. Salomé Cabral, da Faculdade de Ciências da Universidade de Lisboa. É com grande satisfação que anunciamos que a SPE pretende homenagear durante a SPE2011 os seus dois primeiros presidentes, os professores M. Ivette Gomes e João A. Branco.

A cidade da Nazaré é um local de grande interesse turístico, com fortes tradições piscatórias, pussuidora de uma beleza ímpar, um clima ameno e uma grande sabedoria na arte de bem receber. O Programa Social do Congresso, que se encontra em elaboração, procurará pôr em evidência os encantos da Nazaré e da sua região; aproveitamos desde já para convidar os congressistas e seus acompanhantes a participar no mesmo.

O Congresso terá lugar no Hotel Miramar Sul, a partir do qual é possível desfrutar da imensidão do mar e da beleza do casario da Nazaré. Este hotel possui restaurante e bar panorâmico, piscina interior e exterior, e quartos e salas de reuniões confortáveis. O Hotel está localizado numa zona previlegiada a cerca de 1km da praia, e dispõe de transporte gratuito para o centro da Nazaré, disponível para todos os congressistas e seus acompanhantes registados.

Informação actualizada mais detalhada sobre os programas científico e social do XIX Congresso Anual da SPE e outros detalhes úteis podem ser consultados na página web do congresso: http://www.spe2011.ipleiria.pt. Em particular, podem ser consultadas nessa página instruções para a submissão de resumos de comunicações e efectuada a inscrição no congresso. É conveniente alertar desde já que a data limite para envio de resumos de comunicações é 15 de Abril de 2011.

Contamos com a vossa presença na Nazaré no XIX Congresso Anual da SPE para que este seja, à semelhança dos congressos anteriores da SPE, um sucesso científico e uma grande Festa Estatística.

A Comissão Organizadora do XIX Congresso Anual da SPE

## • I Dia Mundial da Estatística


Realizou-se no dia 20 de Outubro de 2010 o I Dia Mundial da Estatística. A SPE associou-se ao CEAUL nestas celebrações com um encontro sob o lema "O papel da Estatística na Investigação Científica ao Serviço da Sociedade".

O encontro, com cerca de 250 participantes, teve lugar no grande anfiteatro da FCUL e contou com a presença de Pinto Paixão, Director da FCUL, Alda Carvalho, Presidente do INE, Paula Brito em representação do Presidente da SPE, Antónia Turkman, Coordenadora do CEAUL, os quais, durante a sessão de abertura teceram considerações relativamente à importância e significado deste dia para a Estatística.

A abrir a sessão, Lisete Sousa leu a seguinte mensagem do Presidente e do Director do ISI a todos os Estatísticos:

*Hoje, 20-10-2010 estatísticos de todo o mundo celebram o Dia Mundial da Estatística. As celebrações revestem-se de formas diversas. Uns organizam conferências e seminários, outros lançam publicações especiais nessa ocasião.*

*Em nome do Instituto Internacional de Estatística (ISI), queremos apresentar os nossos mais calorosos parabéns a todos os estatísticos do mundo, que trabalham em todos os tipos de disciplinas de estatística. Queremos também dar os parabéns à Divisão de Estatística das Nações Unidas por terem lançado esta iniciativa do Dia Mundial da Estatística. Esperemos que muitos mais se repitam no futuro.*

*A celebração do Dia Mundial da Estatística será uma oportunidade para atrair a atenção positiva dos governos, instituições e meios de comunicação social para os avanços da estatística como instrumentos indispensáveis para uma boa governação, para tomadas de decisão baseadas na evidência e para o desenvolvimento humano.*

*O Instituto Internacional de Estatística irá contribuir para esta celebração com o lançamento da revisão da DECLARAÇÃO DE ÉTICA PROFISSIONAL. Esta declaração contém um conjunto coerente de valores e princípios a serem usados por estatísticos que trabalham em diversas organizações por todo o Mundo.*

*Desejo-vos um festivo Dia Mundial da Estatística!*

*Jef Teugels, Presidente do ISI*
*Ada van Krimpen, Director do ISI*

A palestra convidada "*Statistical Methods for Cost-Effective Health Care*", a cargo de Tony O'Hagan, foi muito aplaudida pelos participantes. Seguiu-se uma exposição de painéis de diversos grupos de investigação na área de Estatística, acompanhada por um pequeno Porto de Honra. A atmosfera foi de convívio e muito agradável.

A palestra e o texto dos painéis expostos encontram-se em http://www.ceaul.fc.ul.pt/.

Antónia Turkman

## • Professora Fátima: A Sócia n.º 1, faleceu.

No passado dia 15 de Outubro de 2010 faleceu a Professora Fátima.

Maria de Fátima Fontes de Sousa era Licenciada em Matemática e Doutorada em Matemática Aplicada pela Universidade de Lisboa. Foi Professora Catedrática na Faculdade de Ciências de Lisboa, onde fez toda a sua carreira universitária.

No decurso dessa carreira foi co-fundadora, juntamente com o Prof. J. Tiago de Oliveira, do 1º Departamento de Estatística e Investigação Operacional em Portugal, em 1981, do 1º Centro de Estatística e Aplicações da F.C.U.L. (1979) e da Sociedade Portuguesa de Estatística.

Inaugurou, em Portugal, o ensino regular de disciplinas do âmbito da estatística, tais como Processos Estocásticos, Simulação, Planeamento de Experiências, Programação Matemática e Amostragem entre outras.

No *Memorial da Sociedade Portuguesa de Estatística*, em 2005, publicou "O Ensino da Estatística em Portugal nos últimos 150 Anos".

Na ocasião, o Presidente da SPE enviou aos sócios a seguinte mensagem:

"Lamento informá-los do falecimento da Professora Doutora Maria de Fátima Fontes de Sousa, sócia nº 1 da Sociedade Portuguesa de Estatística a quem a Sociedade e a Estatística muito devem. Muito lhe devemos também pela sua dedicação e pelos seus contributos, ao longo de muitos anos, para o ensino e a investigação na área da Estatística, de que tantos de nós, como é o meu caso, beneficiámos e que tanto contribuíram para o extraordinário desenvolvimento e relevância internacional da comunidade estatística nacional.

Recordar, numa hora de luto para a nossa comunidade, esse labor de uma vida cujos frutos permanecerão é o singelo tributo de uma comunidade reconhecida. Renovar neste momento o compromisso da nossa comunidade no progresso da Estatística é a homenagem devida à Professora Maria de Fátima Fontes de Sousa e ao projecto de futuro que ela tanto ajudou a construir".

Em 2008, a Prof. Fátima foi homenageada por ocasião do seu 80º aniversário (notícia no *Boletim SPE Outono de 2008*).

FR

## • jSPE - Secção de Jovens Estatísticos da SPE

A recém-criada jSPE — Secção de Jovens Estatísticos da SPE — está neste momento a dar os primeiros passos. A jSPE foi estabelecida com o objectivo de promover, cultivar e incentivar um intercâmbio de informação e conhecimento entre estatísticos em início de carreira. Não é obrigatório ser "jovem" para pertencer à jSPE, basta somente ser "jovem" neste interesse.

Para que este projecto seja um sucesso é necessário o apoio de todos os sócios da SPE. Aos Jovens Estatísticos apelamos que se juntem à jSPE; basta que nos enviem um e-mail. 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 projecto.

Em https://sites.google.com/site/jspespe podem encontrar mais informações na página provisória da jSPE

Pela Comissão Instaladora,
Paulo Canas Rodrigues (*paulocanas@gmail.com*)
Miguel de Carvalho (*mb.carvalho@fct.unl.pt*)

## • Simpósio sobre análise de dados em painel


O Departamento de Métodos Quantitativos e a UNIDE, ISCTE-IUL - Instituto Universitário de Lisboa, organizam nos dias **15 e 16** de **Junho** de 2011 um **Simpósio sobre desenvolvimentos recentes em métodos para análise de dados em painel** (Comissão organizadora: Mª de Fátima Salgueiro, José G. Dias, Elizabeth Reis, Paula Vicente, Catarina Marques e Patrícia Serra).

O simpósio tem como objectivo principal discutir o estado da arte de diferentes abordagens à modelação de medidas repetidas e dados longitudinais ou em painel. Para tal reunir-se-ão especialistas em diferentes abordagens à modelação destes dados, destacando-se os seguintes oradores convidados: Kenneth Bollen (North Carolina, USA); Mª Eugénia Ferrão (UBI, Portugal); Ronja Foraita (BIPS – Bremen University, Alemanha); Rui Menezes (ISCTE-IUL, Portugal); Irini Moustaki (LSE, UK); Stephen Pudney (MISOC-ISER, UK) e Fiona Steele (Bristol, UK).

Convidamos deste já todos os sócios da SPE a submeter os seus trabalhos, para comunicação oral ou em poster, em Inglês (língua oficial do encontro), até ao dia 27 de Março. O encontro tem o patrocínio da SPE, beneficiando os seus sócios de uma redução no valor da inscrição. No dia **14** de **Junho** decorrerá um **mini-curso** sobre métodos de análise de dados em painel. Informação mais detalhada está disponível em **http://sympaneldata.dmq.ibs.iscte.pt/**, podendo ser igualmente solicitada por email: **SymPanelData.dmq.ibs@iscte.pt .**

Fátima Salgueiro

## • Lembrete da Directoria - Quotas SPE

Relembramos a todos os sócios que se encontra a pagamento a quota de 2011 - no valor de EUR 30,00 (ou de EUR 15,00 caso seja estudante e junte prova).
Se tem quotas em atraso agradecemos que o seu pagamento seja efectuado o mais rápido possível.
Da sua colaboração depende o progresso da SPE.

A Direcção

## • EQS 2011 - VII Conferência Estatística e Qualidade na Saúde


Na sequência das conferências organizadas em anos anteriores, terá lugar nos dias **19 e 20 de Maio** a VII Conferência Estatística e Qualidade na Saúde – **EQS 2011**.

Esta reunião científica é organizada pela **Escola Superior de Tecnologia da Saúde de Lisboa**, em parceria com o **Instituto de Higiene e Medicina Tropical**, sendo este ano esta a instituição acolhedora.

1. A conferência decorre ao longo de dois dias.
2. A EQS 2011 integra:
    a. **2 cursos satélite** (manhã do 1º dia)
        i. " Sistemas de Informação Geográfica para a Saúde"
        ii. "Sistemas de Gestão da Qualidade na Saúde"
    b. **4 painéis temáticos:**
        i. Medicina Baseada na Evidência; novas tendências
        ii. Bioinformática
        iii. Este país (não) é para velhos
        iv. Avaliação de Tecnologias em Saúde
    c. **2 conferências:**
        i. Marília Sá de Carvalho (abertura)
        ii. Pedro Dias Alves (encerramento)
    d. 2 momentos especialmente dedicados a **comunicações livres** com o objectivo de estimular a apresentação de trabalhos.

Poderá acompanhar toda a informação da EQS 2011 através da página **http://eqs.estesl.ipl.pt**.

Gilda Cunha

# Sondagens e seus Desenvolvimentos

Manuela Magalhães Hill, *mmmh@iscte.pt*

*UNIDE-IUL, ISCTE - Instituto Universitário de Lisboa*

Paula Vicente, *pbcv@iscte.pt*

*UNIDE-IUL e Departamento de Métodos Quantitativos – ISCTE - Instituto Universitário de Lisboa*

## 1. Introdução

Quase todos os dias, ao abrir um jornal ou ao ouvir os noticiários na TV somos confrontados com informação em forma de números, gráficos ou tabelas. Alguma desta informação é obtida a partir de uma *sondagem probabilística* ou simplesmente *sondagem*.

O termo sondagem tem origem na palavra francesa «*sondage*» que significa o acto de investigar a profundidade da água com uma sonda. Actualmente esta palavra é utilizada em diversos domínios, entre os quais a estatística, para expressar a ideia de pesquisa ou de investigação. No caso concreto do domínio da estatística, sondagem refere-se a um processo de recolha de informação a partir de uma amostra representativa da população.

Na língua portuguesa, tal como na língua francesa, o termo sondagem aplica-se a estudos que envolvem operações de amostragem, independentemente do seu domínio. Já a língua inglesa diferencia todas as formas de sondagens, nomeadamente, com o termo «*poll*» para sondagens de opinião e «*survey sampling*» para outros tipos de sondagens estatísticas.

Historicamente o uso das sondagens tem tido os seus altos e baixos, em grande parte devido a juízos errados acerca do uso de uma amostra. O cerne disso parece ser uma crença de que para conhecer algo sobre uma população, é melhor contactar a população toda (recenseamento ou censo) do que apenas uma parte. Kish (1979) refere que um recenseamento efectuado correctamente tem a vantagem de fornecer informação precisa, detalhada e credível sobre todos os elementos da população; uma sondagem tem a vantagem de fornecer informação mais rica, mais complexa, mais barata e em menos tempo, mas importa que seja possível extrapolar os resultados à população, o que só é teoricamente válido se as amostras forem representativas. A fiabilidade dos resultados de uma sondagem reside, assim, em boa parte, na forma como a amostra é seleccionada, mas também na forma como o instrumento de recolha de dados, geralmente o questionário ou, em alguns casos, a entrevista directa, é também construído. A vantagem principal das sondagens é, assim, a possibilidade de obter conclusões para a totalidade da população cujo conhecimento seria, de outra forma, impraticável. Colocam-se, por outro lado, problemas particulares de interpretação da informação recolhida relacionados com a própria composição e selecção da amostra, e também com as características das técnicas de

investigação normalmente usadas, além de não ser uma técnica de estudo adequada a todo o tipo de variáveis.
Situações em que a população seja de pequena dimensão ou o custo de errar seja muito elevado poderão justificar a realização de um censo, mas a maioria dos casos podem ser resolvidos por uma sondagem.
As áreas de aplicação das sondagens estatísticas são muito diversas, nomeadamente, nas Ciências Sociais, onde se estudam as atitudes, hábitos ou preferências das populações humanas. Como exemplos mais conhecidos temos os estudos pré-eleitorais ou os estudos de opinião e de mercado.
Em Portugal, a primeira publicação de uma sondagem eleitoral aconteceu em 1973, tendo o primeiro número do semanário Expresso divulgado na primeira página que "63 por cento dos portugueses nunca votaram" (Lima, 1995). No panorama internacional e a título de curiosidade, podemos referir regimes distintos relativamente às sondagens eleitorais. França continua a proibir a publicação de sondagens no período imediatamente anterior ao acto eleitoral. No Reino Unido existe uma prática voluntária largamente respeitada de não publicar resultados de sondagens no dia das eleições. A República do Montenegro apresenta uma das situações mais curiosas dado que além de proibir nos órgãos de comunicação social públicos a divulgação de resultados de sondagens ou de qualquer tipo de projecção dos resultados da eleição proíbe, ainda, no dia da eleição a publicação de resultados de anteriores actos eleitorais. Já nos Estados Unidos, a publicação pelos meios de comunicação social de resultados de sondagens sobre actos eleitorais é vista como parte integrante da liberdade de expressão.
A primeira sondagem realizada pelo Gabinete de Recenseamento americano foi efectuada em 1937 como teste para analisar o desemprego nos E.U. durante a Grande Depressão, seguindo-se o seu uso nos censos da população a partir de 1940. Foi assim possível incluir questões adicionais (aplicadas a um em cada cinco respondentes) sobre ocupação, parentesco, fertilidade e gostos, sem aumento de custos nem sobrecarga para os respondentes. A amostragem passou então a fazer parte dos recenseamentos, tendo estes questionários cada vez mais extensos.
Actualmente, o Gabinete de Recenseamento americano implementa por ano mais de 200 sondagens sobre dados demográficos e económicos para produzir valores nacionais.
Vejamos, de seguida, alguns momentos importantes no desenvolvimento teórico e prático das sondagens.

## 2. A primeira metade do século XX

A origem das sondagens remonta, tanto quo se conhece, a 24 de Julho de 1824 e foi efectuada pelo jornal americano "*The Harrisburg Pennsylvanan*" com inquéritos de opinião aos leitores naquilo que ficou conhecido como o voto de palha ("*straw vote*" – voto não oficial).
No entanto, só nos finais do século XIX é que surgiu a ideia de *amostra representativa* desenvolvida pelo norueguês Kaier em 1895 e que foi muito bem recebida pela comunidade estatística numa conferência realizada em Estocolmo em 1897. O autor sugeriu a obtenção de uma amostra que fosse aproximadamente uma miniatura da população e que pudesse ser utilizada em estudos mais especializados. Isto faria diminuir consideravelmente o custo. Contudo, a amostra não era probabilística.
Paralelamente, dadas a necessidade apontada pelo Gabinete de Recenseamento americano, Herman Hollerit criou a primeira máquina de processamento de dados (cartão perfurado), que foi utilizada no tratamento do censo de 1890. Verificou-se então uma significativa redução de tempo, pois enquanto que o processamento dos dados do censo de 1880 levou 7 anos a ficar concluído, o de 1890 precisou de 4 anos e meio. Foram utilizadas 180 toneladas de cartões, que foram processados à velocidade de 6900 cartões por dia (6 horas e meia de trabalho).
A máquina de Hollerit não foi inicialmente muito bem aceite, tendo sido utilizada pelo Canadá no recenseamento de 1891, mas não tendo tido aceitação por parte do Reino Unido. A justificação era que a suposta poupança de tempo não compensava o tempo usado na perfuração de cartões. Em 1946, Haltley mostrou que estas máquinas podiam ser utilizadas no cálculo de médias móveis, correlações e resolução de equações simultâneas.
Entretanto, outros desenvolvimentos foram surgindo. Bowley (1926) recomendou o uso de amostras aleatórias. Num encontro no Instituto Internacional de Estatística em 1925, apresentou uma

monografia teórica resumindo os resultados até aí conhecidos sobre amostragem aleatória e intencional. Esta monografia desenvolvia ainda uma amostragem estratificada com afectação proporcional e um desenvolvimento teórico da escolha intencional através da correlação entre variáveis de controlo e variáveis de interesse. No entanto, esta sugestão pouco efeito teve na maioria das amostragens desenvolvidas na primeira metade do século XX.

Neyman (1891-1981) foi, talvez, a figura cimeira desta primeira fase do desenvolvimento das sondagens, evidenciando a superioridade das sondagens aleatórias em universos estratificados. Em 1934 Neyman argumentou contra a amostragem intencional até então considerada por muitos como o modo mais razoável de seleccionar uma amostra representativa. Neyman introduz a amostragem em duas etapas e a utilização conjunta de uma função de custo e de variância, ou seja, o tamanho da amostra em cada estrato deve ser escolhido de forma óptima (menor variância sem ultrapassar o orçamento disponível).

Por necessidade, era importante assumir que se dispunha de uma amostra representativa. Por vezes, as amostras eram validadas por comparação com os resultados populacionais mas nenhuma ferramenta probabilística era usada para quantificar a incerteza dos resultados. Por exemplo, nas eleições americanas em 1936, momento em que já se havia passado à fase dos grandes inquéritos de opinião, a revista *Literary Digest* perguntou a 2 milhões de cidadãos americanos qual a sua intenção de voto entre dois candidatos, Landon e Roosevelt e concluiu que Landon ganharia confortavelmente, indo de encontro ao sentimento generalizado dos americanos. Ao mesmo tempo, Georges Gallup inquiriu 4 mil americanos e concluiu que Roosevelt ganharia com 56% dos votos. Ora, quem triunfou foi Roosevelt com cerca de 61% dos votos. A amostra da revista *Literary Digest*, apesar de estratificada por sexo e idade, estava claramente enviesada porque era constituída apenas por leitores da revista.

Este facto revelou que a exactidão dos inquéritos ou sondagens não dependia da dimensão da amostra mas duma correcta selecção da mesma, isto é, passou também a ser determinante a previsão da margem de erro.

Uma vez que o trabalho da análise dos dados aumenta consideravelmente com o tamanho da amostra, poucas vezes eram calculados desvios padrão e quando calculados, raramente eram utilizadas as fórmulas correctas, pois qualquer que fosse o método de amostragem utilizado, a fórmula utilizada era sempre a desenvolvida para uma amostra aleatória simples.

Mahalanobis (1946) sugeriu que se retirassem duas ou mais amostras, de acordo com o desenho amostral. A variação entre as estimativas das subamostras da população fornece uma estimativa não enviesada da variância do estimador final para a população. Computacionalmente, este método traz vantagens pois em ambiente de cartões perfurados é mais fácil calcular somas do que variâncias.

Com base neste método, novos desenvolvimentos surgiram utilizando «amostragem replicada». Nos Estados Unidos esta ideia evoluiu para uma *pseudo-replicação* na estimação da variância em amostras estratificadas (McCarthy, 1969). Em 1943, Hansen e Hurwitz apresentam os primeiros resultados de estimação e precisão dos estimadores quando a selecção amostral é feita com probabilidades desiguais e no caso particular dos planos de sondagens com reposição.

É neste contexto que Hansen, Hurwitz e Madow escreveram em 1953 o livro *Sample Survey Methods and Theory*. O livro começa com uma discussão sobre viés e erro amostral em sondagens. Descreve em seguida as noções básicas sobre planeamento amostral de amostras aleatórias simples, estratificadas, por *clusters* e multi-etapas. Apresenta vários métodos de estimação para melhorar a eficiência dos estimadores, introduz o método de grupos aleatórios para estimar variâncias, que mais tarde conduziu aos métodos *jackknife* e *bootstrap*, e termina com três estudos de caso.

Assiste-se de seguida ao aparecimento dos primeiros Institutos de Sondagem de Opinião Europeus, nomeadamente em Inglaterra e França; neste país, ocorre em 1947 o primeiro congresso internacional sobre esta temática.

Nos anos 1960s aprofundam-se as relações entre as sondagens e o modelo de estatística inferencial. Vários autores analisam a escolha dos estimadores e do plano de sondagem considerando que a população finita estudada é uma amostra extraída de uma superpopulação de dimensão infinita. Destacam-se os trabalhos de Royall (1970) e Särndall (1991) na estimação *model-assisted* e *model based*.

## 3. A era do computador digital

O primeiro computador digital foi produzido durante a Segunda Guerra Mundial para uso exclusivamente militar. Só nos anos 50 é que os computadores foram construídos para uso comercial e é aqui que a aplicação prática das sondagens tem o seu maior desenvolvimento. A primeira geração de computadores incluiu o UNIVAC seguido pela série IBM 700. O recenseamento de 1961 no Reino Unido sublinha o papel central dos militares na computação. O recenseamento foi processado num computador IBM 705 pertencente ao *War Office* e era usado pela *Royal Army Pay Corps*. Os funcionários do Gabinete do Recenseamento podiam utilizar o computador fora das horas de serviço, sendo os cartões perfurados num outro local e posteriormente transportados para o computador.

O Gabinete do Recenseamento americano, não só recebeu o primeiro UNIVAC produzido, como alguns dos seus colaboradores participaram no seu desenho e construção (Hansen, 1987). Este computador foi usado para processar os dados do recenseamento de 1950, trabalhava todos os dias da semana, 24 horas por dia, e logo que terminou o tratamento dos dados, foi utilizado para o tratamento de outros censos e sondagens.

Talvez devido ao elevado custo e algumas dúvidas sobre as vantagens dos computadores digitais, a sua adopção por parte dos departamentos de estatística foi bastante lenta. Canadá foi o país que logo a seguir adquiriu um UNIVAC para o tratamento do recenseamento de 1961.

Nos últimos 40 anos assistiu-se a um desenvolvimento rápido na tecnologia dos computadores. Os computadores modernos são fisicamente mais pequenos, muito mais rápidos e com muito mais capacidade de armazenamento. O avanço tecnológico fez com que diminuísse consideravelmente o desfasamento entre o desenvolvimento da teoria das sondagens e a sua prática. Até então usavam-se aproximações no cálculo das estimativas da variância para contornar a limitação de tempo e capacidade do equipamento disponível. O UNIVAC e o IBM permitiram melhorar consideravelmente o cálculo dessas estimativas.

Os anos 1960s foram também anos de mudanças societais. Sondagens por telefone começaram a ser cada vez mais utilizadas, originando o desenvolvimento de novos desenhos amostrais tal como o desenho amostral por clusters Mitofsky-Waksberg.

## Referências bibliográficas


Bellhouse, D.R. (1988). A Brief History of Random Sampling Methods, em (P.R. Krishnaiah e C. R. Rao, Editores) *Handbook to Statistics*, Vol. 6, New York:Elsevier Science Publishers B.V., 1-14.

Bellhouse, D.R. (2000). Survey sampling theory over the twentieth century and its relation to computing technology. *Survey Methodology*, 28, nº 1, 11-20.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22(1), 1-62.

Gomes, P. (1998). *Tópicos de Sondagens*. Sociedade Portuguesa de Estatística.

Hansen, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.

Hansen, M.H., Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hansen, M.H, Hurwitz, W.N. e Madow, W.O. (1953). Sample Survey Methods, vols. I e II. New York: Wiley.

Hartley, H.O. (1946). The application of some commercial calculating machines to certain statistical calculations. Suplemento do *Journal of the Royal Statistical Society*, 8, 154-183.

Kaier, A.N. (1897). *The Representative Method of Statistical Surveys* (1976, Tradução inglesa do texto original Norueguês). Oslo, Gabinete Central de Estatística da Noruega.

Kish, L. (1979). Sampling and Censuses. *International Statistical Review*, 47, 99-109.

Lima, R. P. (1995). A Arte de Saber Usar as Sondagens, *Expresso*, 25 Fevereiro, 14.

McCarthy, P.J. (1969). Pseudo-replication: half samples. *Review of International Statistical Institute*, 37, 239-264.

Neyman, J. (1934). On the two different aspects of the representative method: stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
Royall, R.M. (1970). On Finite Population Sampling Theory Under. Certain Linear Regression Models. *Biometrika*, 57, 377-387
Särndall, C.E., Swensson, B. Wretman, J. (1991). Model Assisted Survey Sampling, new York, Springer-Verlag.
Vicente, P., Reis, E., Ferrão, F. (2001). *Sondagens, 2ª Edição*. Ed. Sílabo.

# Sondagens: perspectivas para o séc. XXI

Paula Vicente, *pbcv@iscte.pt*

*UNIDE-IUL e Departamento de Métodos Quantitativos, ISCTE - Instituto Universitário de Lisboa*

Manuela Magalhães Hill, *mmmh@iscte.pt*

*UNIDE-IUL, ISCTE - Instituto Universitário de Lisboa*

## Introdução

O uso generalizado do telefone fixo como meio telecomunicação desde o último quartel do séc. XX foi determinante na actividade de realização das sondagens. De facto, as sondagens *CATI-Computer Assisted Telephone Interviewing* tornaram-se o modo principal de recolha de informação quer em estudos sobre agregados familiares e indivíduos quer em estudos sobre empresas e organizações, especialmente nos países da Europa Ocidental e da América do Norte onde a taxa de cobertura telefónica atingiu valores elevados.

Centenas de trabalhos, entre artigos científicos e livros, foram publicados abordando a multiplicidade de aspectos metodológicos e não-metodológicos envolvidos nas sondagens telefónicas. Uma compilação das referências bibliográficas mais relevantes pode ser encontrada em Khurshid e Sahai (1995) e Survey Research Center (2000).

Os meios "tradicionais" de sondagem – a entrevista pessoal e o questionário por correio postal – têm vindo a perder relevo ficando a sua utilização circunscrita a estudos com características muito específicas. As sondagens com entrevistas pessoais são comuns nos estudos pré-eleitorais com simulação de voto em urna, ou em estudos que requerem a utilização de adereços durante a aplicação do questionário. O envio do questionário por correio postal é adoptado sobretudo em estudos de satisfação de clientes nos quais a disponibilidade de uma listagem de nomes e endereços das unidades da população-alvo a contactar permite personalizar o envio do questionário.

O desenvolvimento tecnológico nas telecomunicações tem permitido inovar e alargar as opções de recolha de informação nas sondagens. De facto, meios de telecomunicação como o e-mail, a Web, a vídeo/tele-conferência, o fax ou o telemóvel estão a tornar-se importantes complementos ou mesmo alternativas ao telefone fixo. Mas o desenvolvimento tecnológico coloca, simultaneamente, algumas dificuldades às *tele-sondagens*. A sofisticação crescente dos dispositivos que filtram ou reencaminham chamadas telefónicas, os atendedores automáticos e as caixas de *voice-mail* e a identificação do emissor da chamada tornam mais fácil do que nunca "escapar" à participação em sondagens.

Neste início de século, é evidente que se colocam desafios às sondagens que obrigam a repensar a sua metodologia e contextos de utilização, tendo como pano de fundo os novos meios de telecomunicação que se afiguram como meios potenciais de recolha de informação. Neste artigo faz-se uma análise breve do impacto da utilização da internet e do telemóvel na realização das sondagens em Portugal. Estes dois meios de telecomunicação são já utilizados para recolha de informação mas afiguram-se como vindo a tornar-se meios predominantes de sondagem num curto/médio prazo.

## E-mail e Web

A sondagem por e-mail é muito semelhante a uma sondagem por correio postal na medida em que se baseia no envio, para um conjunto de endereços, de um e-mail contendo a solicitação de participação e o *link* de acesso ao questionário. A sondagem na Web consiste em disponibilizar num sítio da internet o questionário ou um "aviso" de sondagem com a expectativa de que os utilizadores desse sítio adiram a participar.

As vantagens da internet como meio de sondagem são evidentes: baixo custo, rapidez na recolha, no processamento e na análise dos dados, flexibilidade na concepção dos questionários permitindo a implementação de funcionalidades que os métodos em papel não permitem e possibilidade de realizar grandes amostras. Além deste aspectos, a internet parece ter "democratizado" a actividade de realização de sondagens pois qualquer pessoa/empresa com acesso à Internet e com razoáveis conhecimentos informáticos cria sem dificuldade um questionário e torna-o acessível a um número infindável de potenciais respondentes, quer enviando-o por e-mail quer colocando-o num sítio da internet.

As primeiras sondagens via internet (quer por e-mail quer na Web) surgiram pela mão de especialistas informáticos, que dominavam muito bem a tecnologia informática da internet mas que pouco sabiam de metodologia das sondagens. As críticas/alertas por parte dos metodologistas das sondagens (e.g. Couper 2000) sobre factores potencialmente comprometedores da qualidade dos estudos realizados por este meio não se fizeram esperar: (a) sub-cobertura da população em geral, (b) enviesamento de selecção na obtenção das amostras, (c) baixas taxas de resposta, (d) dificuldades de *hardware* e *software*, que impedem os respondentes de aceder à sondagem em igualdade de condições e (e) fraco controlo sobre o processo de resposta ao questionário.

O problema da sub-cobertura é aquele que, por ora, constituiu o maior obstáculo à massificação das sondagens via internet. Em Portugal, a percentagem de agregados com acesso à internet cresceu de 32% em 2005 para 46% em 2008 (INE 2009), o que, apesar de ser uma variação apreciável, está longe de traduzir uma cobertura total da população. Acresce o facto de a taxa de cobertura não ser uniforme entre segmentos da população. Por exemplo, cerca de 45% dos agregados cujo membro responsável tem idade entre 25-39 anos possuem acesso à internet em casa enquanto que no grupo etário de 55 ou mais anos essa percentagem não chega a atingir 15% (CE 2008).

A sub-cobertura da internet é causa potencial de erro de cobertura em sondagens que pretendam estudar a população geral portuguesa pelo que a utilização da internet para esse fim é arriscado. Porém, em contextos específicos, onde seja legítimo assumir uma taxa de cobertura (quase) completa da população, como é o caso dos alunos e professores das universidades, dos trabalhadores de empresas ou organizações, ou dos clientes de um serviço online, as sondagens via internet podem ter um bom desempenho. É claro que a qualidade de uma sondagem via internet não é determinada somente pela taxa de cobertura da internet mas também por outros factores metodológicos como a selecção das amostras ou a taxa de respostas. Porém, só depois de "assegurada" uma boa taxa de cobertura vale a pena pensar no melhor desenho da sondagem à luz dos respectivos objectivos.

Nos últimos anos, a investigação na área da metodologia das sondagens tem tido uma incidência particular nas sondagens via internet com contribuições importantes que clarificaram as potencialidades e limitações deste meio. No sítio **http://www.websm.org/** encontra-se um repositório extenso e actual das referências bibliográficas mais relevantes sobre metodologia das sondagens na internet.

## Telemóveis

Em Portugal, a percentagem de agregados familiares com pelo menos um telemóvel cresceu de 11% em 1997 para 87% em 2008 (INE 2009) e a percentagem de agregados familiares sem telefone fixo mas possuindo um ou mais telemóveis situa-se perto dos 40% (CE 2008). Desde 2004 que a tendência de crescimento da posse de telemóvel vem sendo acompanhada pelo decréscimo continuado da percentagem de agregados familiares com telefone fixo, actualmente situada nos 70% (INE 2009). Esta situação representa um revés face àquele que foi o cenário que proporcionou a consolidação do telefone fixo como meio de sondagem. Acresce o facto de os utilizadores exclusivos de telemóveis (que representam a maior parcela do segmento "sem telefone fixo") serem diferentes da generalidade da população portuguesa: tendem a ser mais jovens, com maior probabilidade pertencem à população

activa, e em termos de ciclo de vida, tendem a viver sozinhos ou a ser casados sem filhos (Vicente e Reis 2009). Esta conjugação de factores coloca em risco a "capacidade" do telefone fixo para representar adequadamente a população geral portuguesa e faz antecipar a perda de qualidade das sondagens telefónicas (com telefone fixo) em estudos à população em geral.

Face aos actuais níveis de cobertura, a população geral portuguesa seria melhor sondada através do telemóvel do que através do telefone fixo, mas por ora existem obstáculos importantes à adopção generalizada do telemóvel como meio de sondagem. Destaca-se, em primeiro lugar, a dificuldade de constituição de bases de sondagem de números de telemóvel. Enquanto que para os telefones fixos existem listagens publicadas dos números de telefone atribuídos (a mais facilmente acessível das quais é a *Páginas Brancas* da Portugal Telecom) o mesmo não sucede com os números de telemóvel. Esta limitação pode ser contornada pela geração aleatória de números de telemóvel mas este é um procedimento que acarreta alguma ineficiência pelo risco de se gerarem muitos números não atribuídos. Num estudo comparativo entre uma sondagem por telefone fixo e uma sondagem por telemóvel, Vicente et al (2009) reportam uma percentagem de 59% de números de telemóvel marcados e não atribuídos contra apenas 26% no caso de números de telefone fixo. Mesmo optando pela geração aleatória de números de telefone subsiste um problema, não menos importante, que se prende com a dificuldade de implementar um controlo geográfico dos números seleccionados. De facto, um número de telefone fixo "contém" informação sobre a localização geográfica do endereço que lhe corresponde, mas um número de telemóvel não dá qualquer informação a este respeito. O controlo geográfico é imprescindível na maioria das sondagens, mas a utilização dos telemóveis torna a sua implementação difícil quer em estudos locais ou regionais quer em estudos à escala nacional.

Em segundo lugar, existe o problema do custo das chamadas telefónicas. As chamadas na rede móvel têm um tarifário mais caro do que as chamadas na rede fixa, pelo que a adopção plena do telemóvel em detrimento do telefone fixo só deverá acontecer quando o ganhos esperados na precisão dos resultados compensarem o acréscimo de custos que a transição de modos acarreta.

A utilização do telemóvel como meio de sondagem coloca ainda preocupações ao nível da qualidade dos dados uma vez que as circunstâncias em que se atende uma chamada no telemóvel por ser muito diversa, e afectar o grau de concentração e atenção que o inquirido dispensa à entrevista. Este aspecto é ainda mais importante em sondagens em que o telemóvel seja utilizado em combinação com outro modo de recolha de informação, pois importa assegurar a comparabilidade dos dados obtidos pelos diferentes modos.

Como meio potencial de sondagem, o telemóvel não deve ser visto apenas como um dispositivo de comunicação por voz. Quando o acesso à internet via telemóvel se generalizar, o telefone móvel será um meio propício para administrar questionários por auto-resposta. Até há pouco tempo a percentagem de utilizadores de telemóvel em Portugal que acediam à internet através do seu telefone móvel era de apenas 4% (Cardoso et al 2007), mas estimativas apontam para que o telemóvel seja o principal meio de acesso à Internet em 2013 (Pinto 2010). Acompanhando a tendência mundial, o telemóvel será, para cada vez mais utilizadores, um computador "de mão", que possibilita a realização de chamadas telefónicas, mas também o acesso a dados, à internet, ao pagamento de serviços, a uma infinidade de funcionalidades. A concepção de questionários electrónicos, que podem ser respondidos quer através de "teclado e/ou rato + monitor" de um computador ou através de "teclado e/ou ecrã" de um telemóvel, requer preocupações adicionais devido às diferenças técnicas inerentes aos equipamentos (por exemplo ecrãs de diferentes dimensões, diferenças de *software*) que podem comprometer o acesso e a visualização do questionário com repercussões na qualidade das respostas (Couper 2010).

## Futuros desenvolvimentos nas sondagens

A dificuldade crescente em contactar e conseguir a colaboração das pessoas para as sondagens, a facilidade com que se colocam sondagens na internet (por vezes ignorando os princípios metodológicos de qualidade), o excesso de sondagens e de solicitações de colaboração junto das pessoas, são factores capazes de gerar um decréscimo da confiança nas sondagens como meio de recolha de informação. Poynter (2000) refere que, neste enquadramento, técnicas de prospecção, destinadas a explorar dados recolhidos por outras vias, podem tornar-se predominantes em pesquisa de mercados. Mesmo na elaboração de estatísticas oficiais o recurso a registos administrativos de diversas fontes afigura-se como provável num futuro breve (Casimiro 2010). Persistirão, contudo, áreas, como

as do estudo de opinião ou do comportamento não observável, onde as sondagens continuarão a ser o modo de recolha de dados por excelência.

Os desenvolvimentos tecnológicos na área das telecomunicações trazem oportunidades e estabelecem novas fundações para a realização das sondagens. Uma das transformações previstas é a passagem das sondagens via internet de novidade para rotina. Para isto contribuirá a utilização cada vez mais frequente da internet para resolver muitos e diversos assuntos (desde a compra de livros, a marcação de viagens, a reserva de hotéis, tudo parece ser resolúvel através da internet), conferindo inevitavelmente às pessoas "competências" para lidar, sem dificuldade, com a resposta a questionários *online*. Em algumas populações-alvo o acesso e a capacidade de utilização da internet são uma realidade para a totalidade dos seus elementos. Estudantes universitários, professores, membros de associações profissionais, empregados de muitas empresas ou organismos são exemplo dessas populações. O acesso à internet continuará a crescer e a estender-se onde a penetração actual da internet é baixa.

As sondagens *mixed-mode* tinham já alguma popularidade no final do séc. XX, mas tenderão a tornar-se uma necessidade neste início de séc. XXI. Numa sondagem *mixed-mode* dois (ou mais) métodos são combinados para recolher informação (De Leeuw 2005). A crescente taxa de não-respostas, que afecta todos os modos de sondagem, está a forçar a combinação de modos para que se consigam alcançar níveis razoáveis de resposta. Poderá mesmo ser dada aos indivíduos a possibilidade de escolha do método de resposta que preferem. De facto, as sondagens serão cada vez mais dependentes do voluntarismo dos indivíduos e oferecer a possibilidade de responder segundo o modo mais conveniente pode resultar favoravelmente na captação de inquiridos (Nathan 2001). O problema da sub-cobertura, quer nas sondagens via internet quer nas sondagens telefónicas, torna também necessária uma abordagem *mixed-mode* quando se estudar a população em geral.

As sondagens *mixed-mode* requerem adaptações nos procedimentos de estimação e colocam problemas ao nível da comparabilidade da informação. A combinação de modos numa mesma sondagem obrigará à constituição/obtenção de bases de sondagem adequadas a cada modo e à alocação da amostra pelos modos utilizados. Importaria também que as bases de sondagem contivessem informação auxiliar, sobretudo relativa ao endereço permanente do indivíduo, por forma a possibilitar a implementação de algum tipo de controlo geográfico. A duplicação de elementos é inevitável quando se combinam bases de sondagem, mas não é desejável. Assim, é provável que as metodologias de estimação *multiple frame* se tornem predominantes para dar resposta à calibração da amostra no contexto específico de cada combinação de modos (e.g. Maia e Vicente 2010), o que exigirá que se conheça para cada indivíduo os meios de comunicação de que dispõe.

Os problemas ao nível da comparabilidade da informação surgem por não ser garantido que o inquirido responda da mesma forma a uma questão pelo telefone fixo, pelo telemóvel, por e-mail, ou por qualquer outro modo presente numa combinação de modos. A taxa de respostas pode também ser diferente entre modos. Alguns trabalhos entretanto realizados com sondagens *mixed-mode* apontam para padrões de resposta e taxas de resposta com diferenças não negligenciáveis entre modos de sondagem (e.g. Roster et al. 2004, Jäckle e Roberts 2007, Vicente et al. 2009, Lynn e Kaminska 2010). A estimação deverá portanto contemplar mecanismos de ajustamento que podem passar pela modelação do processo de resposta e pela utilização de informação auxiliar específica de cada modo de sondagem, no sentido de "compatibilizar" a informação obtida pelos diferentes modos (e.g. Cobben et al. 2007).

Até agora, o maior desenvolvimento nas sondagens do séc. XXI é o reconhecimento de que o leque de modos de sondagem se alargou para lá dos "tradicionais" entrevista pessoal, entrevista telefónica e questionário postal, e que modos "tradicionais" e modos "novos" co-existirão no futuro. Cada modo tem as suas especificidades, e as respectivas vantagens residem na capacidade de fazer o que os outros modos não conseguem fazer. Importa descobrir para os "novos" modos o que já se conhece para os modos "tradicionais". Enquadrada pelo conceito de Erro Total da Sondagem (Groves 1989) a investigação em metodologia das sondagens deverá seguir explorando os erros típicos de qualquer modo de sondagem - cobertura, amostragem, não-resposta e medição – quer na aplicação exclusiva quer em contextos de combinação de modos.

## Referências bibliográficas


Cardoso, G., Gomes, M., Espanha, R., Araújo, V. (2007) *Portugal móvel: utilização do telemóvel e transformação da vida social*. Research Report Obercom.

Casimiro, F. (2010) A transição censitária em Portugal. *Comunicação apresentada no XVIII Congresso da Sociedade Portuguesa de Estatística*, S. Pedro do Sul, 29 Setembro – 2 Outubro.

Cobben, F., Janssen, B., Berkel, K., Brakel, J. (2007) Statistical Inference in a mixed-mode data collection setting, *Paper presented at the 56th Session of the International Statistical Institute*, Lisbon, 22-29 August.

Comissão Europeia (2008) *Eurobarometer 66.3*. Bruxelas: Comissão Europeia.

Couper, M. (2000) Web surveys: a review of issues and approaches, *Public Opinion Quarterly*, 64, 464-494.

Couper, M. (2010) Visual design in online surveys: lessons for the mobile world. *Paper presented at the Mobile Research Conference 2010*, London, 8-9 March.

De Leeuw, E. (2005) To mix or not to mix data collection modes in surveys, *Journal of Official Statistics*, 21, 2, 233–255.

Groves, R. (1989) *Survey errors and survey costs*. New York: Wiley.

Instituto Nacional de Estatística (2009) *Indicadores sociais 2008*. Instituto Nacional de Estatística.

Jäckle, A., Roberts, C. (2007) Assessing the effect of data collection mode on measurement. *Paper presented at the 56th Session of the International Statistical Institute*, Lisbon 22-29 August.

Kurshid, A., Sahid, H. (1995) A bibliography on telephone survey methodology, *Journal of Official Statistics*, 11, 325-367.

Lynn, P., Kaminska, O. (2010) The impact of mobile phones on survey measurement error. *Paper presented at the Mobile Research Conference 2010*, London, 8-9 March.

Maia, M., Vicente, P. (forthcoming) Optimal estimator in Indirect Sampling using dual frames. *Proceedings of the Statistics Canada's 2010 International Methodology Symposium*.

Nathan, G. (2001) Methodologies for internet surveys and other telesurveys. *Proceedings of ETK 2001 International Seminar on the Exchange of Technology and Know-how and the fourth NTTS seminar, New Techniques and Technologies for Statistics,* Crete, June, 123-132.

Pinto, P. (2010) Internet – Acesso no telemóvel ultrapassa PC em 2013. Disponível em http://pplware.sapo.pt/informacao/internet-acesso-no-telemovel-ultrapassa-pc-em-2013/ (consultado em 16 de Dezembro 2010).

Poynter, R. (2000) We've got five years. *Paper presented at the Association for Survey Computing's Meeting on Survey Research on the Internet*, London, September.

Roster, C., Rogers, R., Albaum, G., Klein, D. (2004) A comparison of response characteristics from web and telephone surveys, *International Journal of Market Research*, 46, 3, 359-373

Survey Research Center (2000). *Sample design for household telephone surveys: a bibliography 1949-1996.* College Park, MD: University of Maryland. Disponível em http://www.musc.edu/bmt738/German/sampbib.html. (4 Outubro 2010).

Vicente, P., Reis, E. (2009) The mobile-only population in Portugal and its impact in a dual frame telephone survey, *Survey Methods Research*, 3, 2, 105-111.

Vicente, P., Reis, E., Santos, M. (2009) Using mobile phones for survey research: a comparison with fixed phones, *International Journal of Market Research*, 51, 5, 613-633.

# O SISTEMA ESTATÍSTICO EUROPEU

Maria Lucília Carvalho, *mlucilia.carvalho@gmail.com*

*DEIO e CEAUL – Universidade de Lisboa*

*Membro do ESAC*

## 1. Introdução

Este texto tem por objectivo fornecer uma breve descrição da constituição e funcionamento do Sistema Estatístico Europeu (ESS)[1] baseada essencialmente nos regulamentos e decisões comunitárias que o estabelecem. Para informação mais detalhada pode consultar-se o site do Eurostat onde se encontram também os documentos legais referidos ao longo do texto.

## 2. Sistema Estatístico Europeu

A necessidade de informação estatística imparcial e objectiva sobre um conjunto muito vasto de fenómenos é uma constante no mundo de hoje em que a internacionalização e interdependência entre os países, políticas, sectores de negócios e, claro, entre os indivíduos são o pano de fundo da globalização.

Evidentemente que, a nível da União Europeia (EU) a existência de estatísticas comparáveis e fiáveis é essencial para a definição de objectivos estratégicos, para o planeamento e implementação das políticas comunitárias e consequente tomada de decisões.

Para responder a essa necessidade o Sistema Estatístico Europeu foi sendo criado gradualmente ao longo dos anos em que a União Europeia se foi consolidando, mas só em 2009 foi estabelecida uma base legal transparente e estável para garantir a independência, integridade e responsabilidade das autoridades estatísticas.

De facto, o Regulamento (CE) Nº 223/2009 do Parlamento Europeu e do Concelho de 11 de Março de 2009 sobre as estatísticas europeias, que vigora actualmente, é uma verdadeira "lei-quadro" que estabelece os princípios básicos e as regras que governam o funcionamento do ESS determinando quem deve fazer o quê, e quem e como se deve decidir nesta área.

Nela se define que o ESS é uma parceria entre a autoridade estatística comunitária, o Eurostat, que é parte da Comissão Europeia, e os Institutos Nacionais de Estatística (NSIs) e outras autoridades nacionais dos Estados Membros (MS), responsável pelo desenvolvimento, produção e disseminação das estatísticas europeias. A parceria também inclui outros países do Espaço Económico Europeu (EEA).

[1] Por ser muito frequente o uso das iniciais em língua inglesa de muitos dos organismos referidos no texto optou-se por inserir as mesmas entre parêntesis.

Desta forma, o ESS funciona como uma rede na qual o Eurostat tem o papel de liderar a harmonização das estatísticas em colaboração estreita com as autoridades estatísticas nacionais que recolhem os dados e compilam estatísticas para fins relacionados tanto com as políticas nacionais como com as europeias.
O ESS também coordena o seu trabalho com os países candidatos, com a Suíça e ao nível europeu com outros serviços da Comissão, Agências e Banco Central Europeu (ECB) e organizações internacionais como a Organização de Cooperação e de Desenvolvimento Económico (OECD); as Nações Unidas (UN), o Fundo Monetário Internacional (IMF) e o Banco Mundial (WB).
O mesmo regulamento fornece também a base legal para a preparação do Programa Estatístico Europeu Multianual que determina o desenvolvimento, produção e disseminação das estatísticas europeias, os campos principais e os objectivos das acções previstas para um período de 5 anos. O programa corrente cobre os anos de 2008-2012.
Os programas quinquenais são acompanhados por Programas Anuais de Trabalho que detalham os objectivos relacionados com a política da Comissão Europeia de cada ano.

## 3. Comitologia

Para exercer as competências de execução que lhe estão atribuídas no âmbito do Sistema Estatístico Europeu a Comissão conta com o apoio de comités de comitologia cujos nomes e atribuições se descrevem brevemente.

### Comité do Sistema Estatístico Europeu (ESSC)

O coração do sistema estatístico europeu é o Comité do Sistema Estatístico Europeu cujo fim é o de garantir "a orientação profissional ao ESS para desenvolver, produzir e disseminar as estatísticas europeias".
O Comité é dirigido pelo Director Geral do Eurostat e é composto pelos Presidentes dos Institutos Nacionais de Estatística[2] dos estados membros e dos estados do Espaço Económico Europeu e da Associação Europeia do Comércio Livre (EEA-EFTA), estes últimos participando como observadores. Observadores do BCE e da OCDE, etc, podem também participar nas reuniões do CSEE.
Reúne-se 4 vezes por ano e a primeira reunião deu-se a 14 de Maio de 2009.

### Partnership Group

O órgão de coordenação estratégica do ESSC é o Partnership Group pois tem como missão promover o desenvolvimento do ESS ao mais alto nível nomeadamente estabelecendo a agenda do Comité do Sistema Estatístico.
É liderado pelo Eurostat e é composto por um subgrupo dos Presidentes dos NSIs dos estados membros escolhidos rotativamente.
Reúne-se normalmente quatro vezes por ano entre as reuniões do ESSC.

### Directores Gerais dos Institutos Nacionais de Estatística (DGINS)

A Conferência dos DGINS foi criada a 15 de Julho de 1953 em Luxemburgo.
Reúne uma vez por ano com o objectivo de discutir tópicos relacionados com o programa estatístico e com os métodos e processos para a produção das estatísticas comunitárias. É acolhida de forma rotativa pelos diferentes estados membros sendo dirigida pelo presidente do NSI hospedeiro.
A última conferência realizou-se em Sofia em 2010 e tratou o tema: Medir o progresso, o bem-estar e o desenvolvimento sustentável.

[2] Embora a legislação europeia designe por Director Geral o dirigente máximo dos NSIs optou-se por usar a designação portuguesa de Presidente.

**Grupos sectoriais**

Os grupos sectoriais tratam temas relacionados com áreas particulares da estatística comunitária antes de serem formalmente apreciados pelo ESSC. São exemplo o Grupo dos Directores das Estatísticas das Empresas, Directores das Estatísticas Sociais, Directores das Contas Nacionais e Directores da Metodologia.

Em 2008 foram também criados dois comités externos de supervisão.

**Conselho Consultivo Europeu para a Governação Estatística (ESGAB)**

Este conselho tem como objectivo providenciar uma supervisão independente do ESS no que diz respeito ao Código de Prática das Estatísticas Europeias. Este código foi adoptado em 2005 por todas as autoridades estatísticas da EU e é baseado em 15 princípios que cobrem o ambiente institucional, os processos estatísticos e os outputs.
O ESGAB aconselha o Eurostat sobre as medidas apropriadas para facilitar a aplicação do Código de Prática, preparando um relatório anual sobre o assunto para o Parlamento Europeu e para o Conselho.
O ESGAB compreende sete membros independentes com uma elevada competência no domínio da estatística e o seu primeiro encontro realizou-se em Março de 2009.

**Comité Consultivo Europeu da Estatística (ESAC)**

O ESAC tem 24 membros e é a única instituição que, junto da Comissão, representa os utilizadores, os respondentes a inquéritos e outras partes interessadas nas Estatísticas Europeias (incluindo a comunidade científica, os parceiros sociais e a sociedade civil) bem como os utilizadores institucionais (e.g, o Conselho e o Parlamento Europeu).
O Comité tem o importante papel de assegurar que são tomadas em consideração as necessidades dos utilizadores, bem como a carga estatística sobre os respondentes e produtores aquando do desenvolvimento dos Programas Estatísticos.
Deverá dar a sua opinião sobre o Programa Estatístico Europeu Multianual, em particular no que diz respeito à sua relevância para as necessidades da integração europeia. Também deve dar a sua opinião sobre o equilíbrio (prioridades e recursos) entre as diferentes áreas do mesmo Programa, bem como sobre o Programa Anual de Trabalho da Comissão.
A primeira reunião do ESAC foi realizada a 30 de Junho de 2009 e desde essa altura até hoje foram realizadas mais 4 reuniões.

Realça-se que este é o comité que mais directamente interessa à comunidade científica que nele pode fazer ouvir a sua voz sobre as necessidades não preenchidas actualmente pelo ESS no que diz respeito ao desenvolvimento da sua actividade de pesquisa.

# Censos 2011: Relevância, pertinência e perspectivas para o futuro

Paulo Gomes, *paulo.gomes@ccdr-n.pt*

*Vice-Presidente da Comissão de Coordenação de Desenvolvimento Regional do Norte*

## 1 – Introdução

A realização de mais um recenseamento geral simultâneo da população e da habitação em Portugal (Censos 2011) constitui uma boa oportunidade para revisitarmos três dimensões de análise subjacentes a esta operação de grande envergadura e que faz parte integrante e central do Programa das Estatísticas Oficiais Portuguesas - a relevância dos Censos, o papel da cooperação e do processo de comunicação e por fim a pertinência da informação censitária. Tais dimensões estão presentes na resposta à seguinte pergunta–base: "o que são os censos 2011, com que meios se vão realizar e para que servem?", o que nos remete, desde já, à consideração dos Censos representarem uma operação exaustiva de recolha e tratamento dos dados relativos à população e ao parque habitacional (dos alojamentos destinados à habitação e dos edifícios que contenham, pelo menos, um alojamento), envolvendo por isso uma multiplicidade de actores, desde logo toda a população residente (ou presente) em Portugal, vários milhares de agentes recenseadores, coordenadores ou controladores do trabalho de recolha de dados, e uma série de entidades para além do próprio INE, tais como, os Serviços regionais dos Açores e Madeira, as Câmaras Municipais e as Juntas de Freguesia. Ao nível dos meios destaca-se ainda o recurso à leitura óptica dos questionários, à semelhança do já ocorrido na edição dos Censos de 2001, mas também a possibilidade à resposta via Internet com mecanismos que garantem a segurança e a confidencialidade da informação prestada e incrementam a acessibilidade ao universo dos respondentes.

A questão colocada sinaliza também o quanto é importante todos percebermos o valor da informação censitária recolhida enquanto fonte imprescindível e suficientemente rigorosa para o conhecimento em termos estatísticos da realidade demográfica, social e económica do país, a nível nacional, regional e local. A este propósito, sublinha-se a singularidade dos censos produzirem, harmoniosamente, informação ao nível de unidades geográficas muito finas que se alicerçam em delimitações definidas a partir das coordenadas geográficas – o facto dos Censos 2011 geo-referenciarem cada edifício com tais coordenadas permitirá uma total adaptabilidade da informação numérica disponível à geografia em análise e, por conseguinte, tornam-no um instrumento coadjuvante na avaliação do quadro local em matéria de apetrechamento de equipamentos ou serviços independentemente das delimitações administrativas e consequentemente proporcionam uma melhor análise custo-benefício na equação decisória de um concreto equipamento ou serviço a instalar no território – aliás, todo o caminho percorrido nos últimos anos na reanálise da provisão de equipamentos e serviços públicos à escala eminentemente local, nomeadamente com a redução drástica do número de unidades escolares do ensino básico ou preparatório ou com a reestruturação da distribuição das valências dos serviços de saúde pelo território, é tributário de informação estatística fina e geograficamente referenciada e que se tiver um bom nível de actualização, proporcionará o necessário apoio à tomada de decisão ou de partilha nesse processo para além de facultar os ingredientes para uma boa compreensão, da parte da população, das opções tomadas em nome do interesse colectivo e não propriamente da mera geografia administrativa do País. Complementarmente, os resultados dos Censos proporcionam a constituição de uma base de referência, indispensável para a extracção de amostras de suporte aos inquéritos realizados junto das famílias no âmbito do Sistema Estatístico Nacional, nomeadamente actualizando a

chamada amostra-mãe. Também as sondagens de opinião recorrem intensamente à informação censitária, na razão directa do seu nível de actualização, para efeitos da definição da arquitectura amostral e de processos de melhoria da estimação tais como a pós-estratificação multi-critério (algoritmo do "Raking Ratio") a partir dos dados censitários que disponibilizam a "verdadeira população" para cada um dos cruzamentos dos critérios de pós-estratificação.

Por último sublinha-se o conteúdo da alínea c) do artigo 3º do Decreto-Lei 226/2009 sobre os Censos 2001, o qual inclui na lista dos seus grandes objectivos "a organização de uma base de dados de natureza individualizada, para edifícios, alojamento, famílias e indivíduos, que permita a futura integração de dados com os provenientes de fontes administrativas, de modo a implementar a transição censitária para um modelo de produção de dados censitários, sobre a população e a habitação, de forma mais frequente e com menos custos".

Para além destas dimensões aqui muito sinteticamente abordadas procuraremos neste artigo cruzar os *inputs* gerados pela informação censitária da população e habitação com temas da maior actualidade para a Coesão e Competitividade do território, os quais observando orientações estratégicas de raiz comunitária ou nacional encontram na região ou nos municípios o palco natural de operacionalização e concretização.

Destacaremos aqui cinco temas nucleares ao nível da territorialização de políticas públicas:

1 – A nova disciplina de uso e ocupação do solo e a monitorização dos planos regionais de ordenamento do território;
2 – Monitorização do território-cidade enquanto espaço motor do desenvolvimento e espelho das maiores fragilidades do tecido social;
3 – A eficiência energética;
4 – A mobilidade sustentável;
5 – Geração de políticas assertivas e persistentes ao nível da economia social.

Em cada um dos temas procuraremos ilustrar a indispensabilidade da informação fornecida pelos Censos 2001 e concomitantemente a necessidade de se incrementar a usabilidade de tal informação ao serviço de diagnósticos prospectivas que fundamentem prioridades de investimento em infra-estruturas ou em serviços a criar por parte de actores públicos ou privados.

Por último e como corolário das problemáticas directa ou indirectamente aqui suscitadas, afloraremos a questão incontornável dos recenseamentos do futuro, não tanto como mero caminho para reduzir os avultados custos de uma operação censitária decenal mas essencialmente enquanto oportunidade para avaliarmos, no início desta década, as potencialidades e os limites da informação de carácter administrativo, hoje disponíveis em Portugal, cruzando-a ao nível de variáveis-chave fornecidas pelos Censos 2011 e para escalões geográficos progressivamente mais finos – o balanceamento crítico entre conteúdo informacional e valor da actualização da informação determinará, por recurso, ao *know-how* estatístico e às tecnologias de inter-operabilidade disponíveis, o melhor compromisso para os recenseamentos futuros, indiscutivelmente alicerçados no contínuo *upgrade* das fontes estatísticas administrativas.

## 2 – A informação censitária e os desafios da década

A agenda europeia e por mera declinação a agenda dos 27 países que constituem este espaço alargado enunciaram recentemente três grandes princípios que projectam a estratégia Europa 2020 e que se consubstanciam em politicas indutoras de **crescimento inteligente**, **crescimento sustentável** e **crescimento inclusivo** – compreende-se que haja um elevado consenso em torno da tais princípios. Contudo, as intervenções à escala nacional, regional e local resultarão do permanente equilíbrio entre coesão e competitividade territoriais e não podem naturalmente prescindir do diagnóstico exigente do ponto de partida, orientador das efectivas prioridades, isto é, de informação estatística disponível a nível geograficamente desagregada e adaptativa a áreas geográficas de geometria variável – compreende-se bem o facto das instâncias europeias terem definido um patamar comum e prioritário

de informação harmonizada à escala europeia (*cor topics*) disponibilizada pelas operações censitárias de tais países, independentemente dos modelos de recenseamento aí estabelecidos.

A título meramente exemplificativo abordaremos agora alguns temas incontornáveis da agenda política portuguesa, alinhados com as grandes orientações europeias, com repercussão a diferentes escalas territoriais e, por conseguinte, tributárias de informação estatística assim georreferenciada.

### 2.1 – Monitorização dos Planos Regionais do Ordenamento do Território

O Plano Nacional para o Ordenamento do Território publicado em 2007 e as respectivas declinações regionais (Planos Regionais de Ordenamento do Território), já publicados ou em fase final de aprovação pelo Governo, preconizam uma maior disciplina ao nível do uso e ocupação do solo plasmada em recomendações e directrizes que os Instrumentos de Gestão do Território e, em particular, os planos municipais de Ordenamento do Território (PMOT) respeitam:

- o carácter excepcional de reclassificação do solo rural em urbano com vista a travar a actual prática de aumento indiscriminado dos perímetros urbanos;

- a contenção da expansão do solo urbano com critérios de economia dos recursos territoriais e de infra-estruturas, equipamentos e serviços, em articulação com a rede de acessibilidades e transportes colectivos;

- o estabelecimento de modelos de ocupação do solo e de disciplina de edificabilidade que promovam a concentração de edificabilidade em aglomerados urbanos ou rurais "delimitados" para o efeito, de forma a contrariar padrões de povoamento disperso ou linear.

A gestão deste processo, na fase de transição pós-aprovação dos Planos Regionais de Ordenamento do Território, faz apelo a uma clara caracterização da situação de referência do solo urbano ou urbanizável e de uma contínua monitorização desses espaços através do desenvolvimento de uma base cartográfica para cada Região que permita a caracterização do edificado segundo a distribuição e concentração no território:

- Áreas edificadas consolidadas/em consolidação, correspondentes ao solo urbano/em consolidação;
- Núcleos edificados com funções residenciais e outras, correspondentes aos aglomerados rurais;
- Áreas de edificações dispersas.

Ora, os dados censitários e muito particularmente a informação georreferenciada, proveniente dos Censos 2011 da população e da habitação, assumirão particular pertinência enquanto *inputs* actualizados para essa caracterização.

### 2.2 – Monitorização do "Território Cidade" enquanto espaço motor do desenvolvimento e espelho das maiores fragilidades do tecido social

A melhoria da qualidade de vida urbana nas suas múltiplas dimensões constitui um factor determinante para atrair e manter nas cidades a força de trabalho qualificada, empresários e investigadores, estudantes, turistas e principalmente os próprios residentes. Por isso mesmo a Comissão Europeia tem vindo a enfatizar que um diagnóstico periodicamente actualizado da evolução da situação das principais cidades europeias constitui um pré-requisito para a monitorização dos resultados atingidos ou de metas pré-estabelecidas no âmbito de políticas regionais e urbanas.

O programa europeu "Urban Audit" procurou responder à procura de informação, quantitativa e qualitativa, que permitisse avaliar a qualidade de vida das cidades europeias – independentemente da abrangência dos indicadores considerados neste programa, a informação censitária "Censos 2011" proporcionará, pelo menos para as principais cidades portuguesas, uma informação de base ao nível da população e do parque habitacional susceptível de impulsionar finalmente uma presença significativa do nosso país nesse programa europeu. Complementarmente faz hoje todo o sentido que se construa

informação de 2º nível a partir dos dados relativos a cada quarteirão do *espaço cidade* que permita elaborar tipologias socio-económicas das áreas metropolitanas de Lisboa e Porto (tal como foi feito no passado, a partir dos Censos 91) mas alargando a outros aglomerados urbanos ou a redes de cidades que desempenhem um papel chave no modelo territorial de cada região.

Esta abordagem proporcionará, também, procedimentos de observação da evolução anual de Áreas Urbanas Sensíveis em dimensões como o envelhecimento, precariedade e exclusão social, serviços de proximidade, emprego de proximidade ou segurança pública, assim como em dimensões mais relacionadas com o parque habitacional que, cruzadas com as primeiras, influenciem prioridades ao nível das "parcerias para a Regeneração Urbana" ou nos "Contratos de Desenvolvimento Social".

Escreveu Ítalo Calvino, em "As Cidades Invisíveis", que a "cidade aparece-nos como um todo em que nenhum desejo se perde e de que nós fazemos parte"…

### 2.3 – Mobilidade Sustentável

Em todo o Mundo o tema da mobilidade encontra-se num patamar prioritário da agenda política não só pelas crescentes preocupações em matéria de emissão de gases com efeito de estufa e inevitáveis consequências para a saúde pública e para a própria sustentabilidade do planeta, mas também porque é por de mais óbvio que não podemos aumentar a nossa dependência dos combustíveis fósseis, cujos preços no mercado global não pararão de crescer ao longo das duas próximas décadas – isto é, no "virar de página" em que o Mundo hoje se encontra, vai ser necessário travar o consumo não reprodutivo de combustíveis fósseis, acelerando políticas públicas que reestruturem o espaço urbano como forma de promover a mobilidade dos cidadãos por via de diversificação das soluções de transporte e através de intervenções que apontem numa reforma dos espaços – canal conducente ao desenvolvimento de infra-estruturas para modos alternativos estimuladoras da multimodalidade urbana. Complementarmente, e mormente nas áreas metropolitanas, é urgente concretizar-se medidas de articulação entre políticas de uso do solo e políticas de transporte combinando as questões do urbanismo, do meio ambiente e da energia e integrando-as.

Esta muito sintética afloração ao tema evidencia o papel crucial dos Censos 2011 na informação que proporciona nas deslocações casa-trabalho ou casa-escola e sobretudo pela base de sondagem ao nível do território para o desenvolvimento de um novo inquérito à mobilidade nomeadamente representativo ao nível das freguesias das áreas metropolitanas de Lisboa e Porto ou de outros principais aglomerados urbanos.

### 2.4 – A eficiência energética

A gestão de procura de energia é um aspecto decisivo para o futuro do País na medida em que, enquanto não se vislumbra uma forma clara e definitiva de se poder dispor de electricidade proveniente de fontes limpas, a problemática da electricidade de origem fóssil reclama quer a promoção do uso racional e eficiente da electricidade, valorizando a sua qualidade energética, como o combate ao seu desperdício. De facto só com moderação do crescimento do consumo de electricidade é que o esforço de produção de electricidade a partir das energias renováveis se repercutirá num continuado aumento da electricidade de origem renovável no "mix" nacional.

Este tipo de preocupações e consequentes medidas de atenuação do consumo estão presentes no Plano Nacional de Acção para a Eficiência Energética no qual destacamos iniciativas que procuram estimular a excelência energetico-ambiental no parque edificado nomeadamente, através de uma melhor caracterização dos equipamentos que hoje produzem calor, frio ou água quente em Portugal. A informação que os Censos 2011 introduzem pela primeira vez relativa ao ar condicionado e a informação sobre a principal fonte de energia utilizada para o aquecimento, suscitará conhecimento demonstrativo da dimensão do trabalho que temos pela frente e por conseguinte da necessidade de se proporcionar às famílias novos estímulos a um desempenho menos "energívoro" dos seus habitats e sem perda de conforto.

**2.5 – A geração de políticas assertivas e persistentes ao nível da economia social**

Aqui reside inexoravelmente o tema nuclear da agenda política desta década porque as mutações rápidas do modo quotidiano de viver das pessoas e das famílias, em particular com o aumento significativo das famílias idosas, a par de uma forte instabilidade no mercado de trabalho, também com reflexos preocupantes na inserção dos jovens, exige uma revisitação do leque de iniciativas que podem reequilibrar o tecido social e por isso iniciativas construídas a partir do conhecimento de proximidade e de uma prévia selecção das áreas geográficas mais sensíveis, na fronteira de exclusão social ou abaixo efectivamente do limiar da pobreza – a fotografia censitária proporcionará uma "TAC" actualizada das potenciais áreas problemáticas e por essa via uma reanálise da equação social em tempos de recursos financeiros tendencialmente escassos e da inevitabilidade de investirmos dinamicamente em tudo aquilo que se revele demonstradamente prioritário.

**3 – Quais as expectativas para o futuro?**

Em todos os países com Sistemas Estatísticos mais avançados tem-se vindo a considerar novas abordagens para a realização dos recenseamentos em alternativa ao "recenseamentos clássicos", os quais, designadamente na Europa, observam hoje um quadro relativamente coerente e harmonizado de recomendações emanadas da Comissão Económica das Nações Unidas para a Europa em articulação com o Serviço Estatístico Europeu, EUROSTAT.

Entre esses países, alguns evoluíram, há vários anos, para o uso intensivo de fontes administrativas inter-operáveis e em boa parte gerando informação estatística geograficamente desagregada calculada por estimação a partir de tais registos administrativos. Contudo tal evolução é um trabalho de décadas pois exige, caso a caso, que os dados produzidos através de ficheiros administrativos sejam o mais próximo possível daqueles que poderiam ser obtidos através do recenseamento clássico.

Adicionalmente, o caminho crítico para este novo paradigma inclui um modelo comunicacional a longo prazo sensibilizando progressivamente a população sobre as vantagens na utilização de registos administrativos, e consequentemente a possibilidade dos serviços de estatística acederem a dados administrativos a nível individual.

Portugal tem vindo a fazer paulatinamente esse caminho e os próximos dois anos serão decisivos para cruzarmos informação proveniente de diferentes fontes e tomarmos uma decisão de gestão em fase de transição.

Enquanto utilizador de informação é sedutor podermos recorrer a uma abordagem mista e que inclua inquéritos por amostragem ou mini-operações censitárias ao longo de década permitindo uma contínua actualização de pelo menos um subconjunto das variáveis dos censos 2001[1] com efeitos na melhor usabilidade e actualidade da informação a nível regional e local mas também com efeitos importantes na qualidade da base de sondagem que alimenta a produção de estatísticas oficiais portuguesas sobre as pessoas, famílias e os alojamentos ou as próprias sondagens de opinião.

Por outro lado, tal via facilitará também o tratamento e actualização das séries estatísticas tributárias da informação proveniente dos censos.

Em conclusão, estamos em presença de um triplo desafio – produzir informação estatística relativamente equivalente à fornecida hoje pelos Censos, reduzindo os custos globais e a respectiva concentração num só ano da década, diminuindo a carga estatística sobre os respondentes e proporcionando níveis acrescidos de usabilidade de informação, isto é, maximizando a **pertinência da informação censitária**.

[1] O caso da França e o seu "recenseamento anual" recorre, de facto, a um mix metodológico e, no plano de inferência estatística, a técnicas de estimação combinadas por recurso a informação auxiliar disponível ou através dos métodos de estimação em pequenos domínios, sendo que 70% da população francesa é recenseada de cinco em cinco anos através de uma amostra geograficamente rotativa.

# Os censos 2011 e o futuro[1]

Fernando Casimiro, *fernando.casimiro@ine.pt*

*Coordenador do Gabinete dos Censos do Instituto Nacional de Estatística*

## 1 Introdução

Censos 2011 (XV Recenseamento Geral da População e V Recenseamento Geral da Habitação) é a abreviatura utilizada para designar uma operação estatística que engloba em simultâneo a realização do que tradicionalmente chamamos recenseamentos da população e da habitação. Estes recenseamentos realizam-se, em Portugal, de forma regular e harmonizada em termos nacionais e internacionais: desde 1864 o da população e desde 1970 o da habitação. Daí que assumam ordenações diferentes na respectiva série e também porque a componente da habitação desde 1970 que é bastante mais abrangente, em termos de conteúdo, do que anteriormente àquela data.

Os recenseamentos da população e da habitação constituem sempre um "acontecimento estatístico" importante na actividade estatística, tanto por serem as operações estatísticas mais complexas que cada país realiza, como por abrangerem toda a população e envolverem toda a Administração Pública, com especial destaque para a regional e local; os seus resultados são sempre aguardados com muita expectativa por parte dos respectivos utilizadores. Para além de constituírem o "benchmark" de outras actividades estatísticas, os resultados censitários constituem a maior fonte de informação estatística harmonizada e a mais desagregada geograficamente pelo que o seu potencial de utilização é o maior em termos estatísticos e, designadamente a nível das autarquias locais, constitui-se como a fonte privilegiada deste tipo de informação.

Tanto o conteúdo como a realização dos Censos 2011 estão fortemente ancorados na experiência censitária mais recente (2001) e nas novas exigências da sociedade actual. Assim o conteúdo foi ajustado o suficiente para avaliar novas realidades demográficas e sociais mas mantém o essencial das séries censitárias. O enquadramento legal tanto nacional como comunitário sofreu fortes alterações tanto nos objectivos como nos procedimentos. Para além da habitual disponibilização de informação censitária, há um objectivo nacional novo que consiste em potenciar a utilização dos resultados dos Censos 2011 para implementar um novo modelo censitário que deverá passar pela utilização futura dos ficheiros administrativos. O enquadramento comunitário da União Europeia (UE) foi fortemente alterado com a publicação de quatro regulamentos de implementação imediata que cobrem o conteúdo, a organização e difusão dos resultados bem como a preparação de relatórios de qualidade.

A transição de modelo censitário em Portugal exige um amplo e complexo trabalho de análise e interligação de ficheiros administrativos que existem mas que não contêm toda a informação que estes recenseamentos normalmente disponibilizam. Assim, para além da informação em falta e que deverá ser colmatada por outras vias, é necessário garantir que a informação existente é consistente e fiável ao nível de cada ficheiro e inter-ficheiros, pelo que é inevitável partir de uma referência exaustiva e actualizada para servir como padrão de referência no tempo e no conteúdo, na medida do possível.

Assim, neste artigo, procura-se desenvolver cada um destes temas na medida em que o conhecimento hoje o permite e com os limites que o aprofundamento destas matérias não se transforme demasiado entediante para quem o vai ler.

[1] As opiniões expressas neste documento são da exclusiva responsabilidade do seu autor e não reflectem necessariamente a opinião do INE. Para informações mais desenvolvidas sobre os temas abordados neste documento, sugere-se a consulta de http://censos.ine.pt

## 2 A relevância da informação censitária na actualidade

A informação censitária da população e habitação, independentemente da forma como é produzida, assume sempre uma relevância muito significativa porque cobre duas das áreas mais importantes da governação do País. Conhecer a população e as suas condições de habitação são elementos essenciais para apoiar uma boa governação a todos os níveis da Administração porque esta informação não só traça um retrato bastante fiel da realidade como dá seguimento a uma longa série estatística sobre a população e a habitação. Por outro lado, a última informação equivalente já tem dez anos, o que a torna ainda mais esperada, em especial pelo facto de ser informação exaustiva e desagregada a níveis geográficos muito finos.

Os grandes utilizadores da informação censitária podem ser agrupados de acordo com o seguinte: administração pública, actividade privada, ensino e investigação e suporte para outra produção estatística. Seguidamente apresentam-se alguns exemplos de utilizações destes dados.

Na administração pública podemos identificar utilizações que vão desde a identificação e localização das necessidades de infra-estruturas escolares e sociais como escolas, hospitais e lares de idosos até à distribuição de fundos financeiros da Administração Central para a Local, bem como a elaboração e actualização dos planos directores municipais e de áreas específicas de actuação a nível regional.

Na actividade privada, os estudos de localização de superfícies comerciais e industriais, os estudos de mercado e as sondagens de opinião fazem uma ampla utilização desta informação.

No ensino e investigação, destacam-se os trabalhos escolares dos alunos dos vários níveis de ensino bem como as teses de mestrado e doutoramento em diversas áreas da população e habitação que fazem uso destes resultados, tanto na perspectiva analítica como enquanto suporte para a realização de investigações específicas com base em inquéritos. Para dar um exemplo não muito conhecido, a actual carta de vulnerabilidade sísmica do edificado para habitação conta com uma ajuda preciosa dos resultados de 2001 que terão continuidade em 2011.

No suporte à outra produção estatística releva-se a preparação da chamada "amostra-mãe" que é sempre actualizada na sequência de cada recenseamento e que permite a extracção de amostras para todos os inquéritos às famílias. Para além disso, as estimativas inter-censitárias anuais da população, que suportam um crescente número de decisões, desde a produção estatística corrente enquanto ponderadores de resultados de inquéritos até aos cálculos da sustentabilidade da segurança social, permanecem provisórias até serem "amarradas" entre censos, passando então a definitivas.

Deste modo, os resultados censitários destacam-se sempre como um elemento de referência determinante não só pela forma como permitem avaliar a realidade da população e habitação no momento em que são disponibilizados, mas também pelo contributo que dão para outras actividades estatísticas de importância crescente no dia a dia de todos os cidadãos.

## 3 Enquadramento nacional e internacional destes recenseamentos

Estes recenseamentos são objecto de legislação específica nacional e internacional desde longa data, bem como de recomendações internacionais das Nações Unidas. As recomendações internacionais são revistas todas as décadas e a legislação da União Europeia foi significativamente aumentada e reorganizada, com aplicação a partir de 2011.

### *3.1 Enquadramento nacional*

A realização destes recenseamentos sempre foi objecto de legislação nacional específica, tendo em conta as suas especificidades e a necessidade de envolver todas as estruturas da administração pública.

Para os Censos 2011 foi publicado o Decreto-Lei 226/2009 de 14 de Setembro que estabelece as respectivas regras organizativas e executivas. Para além do enquadramento organizativo, tem particular importância o facto desta legislação definir uma maior abrangência dos objectivos destes recenseamentos, com especial destaque para a alínea c) do seu artigo 3º:

*Artigo 3.º*
***Objectivos***
*Os Censos 2011 têm por objectivos:*
*a) A recolha, o apuramento, a análise e a divulgação de dados estatísticos oficiais referentes às características demográficas e socioeconómicas da população abrangida e do parque habitacional;*
*b) A criação de uma base de informação de referência, fundamental para a selecção e extracção de amostras, garantindo o suporte aos inquéritos realizados no quadro do sistema de informação estatística para as famílias;*
*c)* ***A organização de uma base de dados de natureza individualizada, para edifícios, alojamentos, famílias e indivíduos, que permita a futura integração de dados com os provenientes de fontes administrativas, de modo a implementar a transição censitária para um modelo de produção de dados censitários, sobre a população e a habitação, de forma mais frequente e com menores custos.***

Enquanto as duas primeiras alíneas têm sido uma constante dos últimos recenseamentos, a alínea c) determina claramente a utilização dos resultados dos Censos 2011 para a implementação de um processo de transição censitária apoiado em ficheiros administrativos.
Este enquadramento legal nacional também está subjacente na Lei 22/2008 de 13 de Maio, que regula o Sistema Estatístico Nacional, designadamente o nº 2 do artigo 4º, o qual estabelece o princípio da utilização dos dados administrativos para fins estatísticos oficiais.
Assim, a organização legal nacional, para além de dar seguimento a processos idênticos anteriores também se enquadra no modelo internacional vigente, nomeadamente a nível da UE e das recomendações das Nações Unidas para a Europa.

### *3.2 Enquadramento internacional*

O enquadramento internacional consubstancia-se nas recomendações da UNECE (United Nations Economic Commission for Europe) e num conjunto de regulamentos que a UE fez publicar para vigorarem a partir de 2011.
As Recomendações da UNECE, cuja última edição é de 2006, estabelecem os conceitos e definições inerentes a todas as unidades estatísticas e variáveis que devem ser observadas nestes recenseamentos. Para além disso, definem um conjunto essencial de regras básicas às quais devem obedecer estas operações estatísticas, a saber:

- Recolha individualizada dos dados referentes a cada unidade estatística (edifício, alojamento, família e indivíduo)
- Simultaneidade/momento censitário definido, em relação ao qual se deve recolher todos os dados;
- Universalidade da observação no respectivo território;
- Disponibilização de resultados para as pequenas áreas estatísticas (introduzida nesta última versão das recomendações);
- Periodicidade definida, correspondente à realização regular destes recenseamentos, no mínimo decenal.

No conjunto das variáveis e classificações definidas nestas recomendações faz-se uma clara distinção entre as que são prioritárias e as que não o são (*core topics* e *non core topics),* assumindo-se que as prioritárias devem constituir a matriz mínima comum para todos os países da região.
Adicionalmente, estas recomendações também estabelecem os quatro modelos fundamentais de recenseamento baseados no modelo de recolha de dados: método clássico com ou sem amostragem para algumas variáveis, método clássico com actualizações anuais de dados sobre algumas variáveis com utilização de amostragem, utilização de ficheiros administrativos e combinação de registos administrativos com inquéritos por amostragem.
Na União Europeia as condições legais de realização destes recenseamentos alteraram-se radicalmente em relação ao passado recente. Até 1991 inclusive era publicada uma Directiva que reportava para um programa de apuramentos comum a todos os Estados Membros (EMs) mas que levantava sempre dificuldades de aplicação mercê das diferenças entre EMs na utilização de métodos censitários. Em 2001 optou-se por um "gentlemen's agreement" para o cumprimento de um programa de apuramentos

comum que resultou na disponibilização tardia de resultados e com fortes inconsistências na comparação de resultados entre os vários países.
Face à situação descrita anteriormente, foi decidido implementar um processo legislativo mais abrangente, exaustivo e rigoroso no sentido de garantir maior consistência, qualidade e rapidez na disponibilização dos resultados censitários a partir de 2011. Para isso foram publicados quatro regulamentos de acordo com o seguinte:

- Regulamento de enquadramento nº 763/2008 que define os modelos censitários da UE e o conteúdo obrigatório destes recenseamentos na base das variáveis prioritárias da UNECE, bem como a desagregação geográfica mínima para os resultados de cada variável; além disso define as regras gerais a observar na transmissão dos resultados ao Eurostat, bem como os requisitos de qualidade a que eles devem obedecer;
- Regulamento de aplicação nº 1201/2009 que define as especificações técnicas das variáveis estatísticas e a respectiva desagregação geográfica e de modalidades;
- Regulamento de aplicação nº 519/2010 que define o formato (hipercubo) dos resultados a produzir bem como a respectiva metainformação;
- Regulamento de aplicação nº 1151/2010 que define as formas e a estrutura dos relatórios de qualidade.

Como se pode deduzir da descrição resumida anteriormente, a legislação comunitária sobre estes recenseamentos passou a ser de um pormenor nunca visto até agora.
De salientar que Portugal não deverá ter quaisquer problemas no seu cumprimento uma vez que o Programa de Acção dos Censos 2011 foi sendo sucessivamente adaptado aos requisitos comunitários, à medida que eles assumiam a forma final de regulamento.

## 4 As alterações mais importantes dos Censos 2011

Todo o processo de preparação e execução dos Censos 2011 foi conduzido no sentido de adaptar o seu conteúdo e o modelo organizativo à realidade actual, tanto em termos nacionais como internacionais. Assim, foi necessário proceder a alterações no conteúdo a observar (variáveis das unidades estatísticas) de modo a garantir as necessidades nacionais e a responder às recomendações da UNECE e dos regulamentos da UE.

### *4.1 Alterações de conteúdo*

A avaliação do conteúdo censitário é sempre um exercício complicado de equilíbrio entre o que os principais utilizadores consideram como necessidades fundamentais e as limitações inerentes a uma operação estatística desta dimensão, cujo conteúdo deverá ser restringido às variáveis que evitem a sobrecarga estatística dos respondentes. Procurou-se fazer este equilíbrio, embora tal seja sempre difícil como se pode verificar pelo quadro 1, com um aumento sistemático nos últimos quatro recenseamentos.

**Quadro 1 - Número de variáveis por unidade estatística nos censos de 1981 a 2011**

| | Censo | | | |
|---|---|---|---|---|
| | **1981** | **1991** | **2001** | **2011** |
| **Edifício** | 10 | 15 | 16 | 17 |
| **Alojamento** | 15 | 14 | 17 | 18 |
| **Família** | 5 | 5 | 5 | 5 |
| **Indivíduo** | 27 | 30 | 33 | 43 |
| **Total** | **57** | **64** | **71** | **83** |

As alterações de conteúdo afectaram todas as unidades estatísticas, embora de forma diferente, de acordo com o seguinte:

- No **edifício** apenas se reformularam algumas variáveis já existentes no sentido de as adaptar às recomendações internacionais, tornar algumas modalidades das variáveis mais facilmente identificáveis e reduzir a carga estatística noutros casos. De salientar o facto das variáveis sobre a vulnerabilidade sísmica apenas se destinar a edifícios construídos para terem 3 ou mais alojamentos.
- No **alojamento** salienta-se a saída das variáveis electricidade e cozinha, que cobrem a quase totalidade dos alojamentos recenseados em 2001, e a entrada de 4 variáveis novas: ar condicionado, principal fonte de energia utilizada para aquecimento, área útil e lugar de estacionamento.
- Na **família** há sobretudo a reformulação da relação de parentesco com o representante da família e o tipo de família na base nos núcleos familiares, pelo facto de passarmos a incluir os núcleos homossexuais e a excluir os núcleos familiares "avoengos", estes últimos decorrentes da harmonização comunitária.
- No **indivíduo** existem alterações de conteúdo mais importantes: foram introduzidas novas variáveis sobre o estado civil legal, a união de facto, o tipo de incapacidade, a residência anterior no estrangeiro e o ano de chegada ao país, o país de proveniência, a utilização de um segundo meio de transporte na deslocação casa-trabalho/local de estudo; por outro lado foram excluídas as variáveis estado civil de facto, tipo de deficiência e grau de incapacidade atribuído.

O balanço final acabou por ficar num acréscimo de variáveis em relação a 2001.

### *4.2 Cartografia de maior qualidade*

Como tem vindo a acontecer em relação aos últimos recenseamentos, a cartografia censitária para 2011 tem uma qualidade superior à do anterior recenseamento, tanto em termos de harmonização e pormenor, como na sua actualidade. Vamos trabalhar com cartografia digital em escalas que variam entre 1/2000 e 1/10000, com a inserção dos limites administrativos oficiais e com toponímia muito actualizada. O país ficou dividido em cerca de 18.100 secções estatísticas (área de trabalho de um recenseador) e cerca de 265.000 subsecções estatísticas (quarteirões e partes de lugares) contra cerca de 16.100 e cerca de 178.000, respectivamente, em 2001. Este pormenor quantitativo vai reflectir-se na maior qualidade da informação estatística e geográfica a disponibilizar.
Por outro lado, e pela primeira vez, os Censos 2011 vão georreferenciar cada edifício com as respectivas coordenadas geográficas o que permitirá garantir uma maior maleabilidade na utilização da informação numérica adaptada à geografia (no passado apenas se podia utilizar a informação georreferenciada até ao nível da subsecção estatística). As coordenadas dos edifícios vão permitir fazer as agregações da informação numérica e geográfica de um modo geográfico completamente livre, constituindo, por exemplo, frentes de quarteirões, segmentos de arruamentos, ou outras quaisquer unidades geográficas que se apoiem em delimitações com base em coordenadas geográficas, independentemente das delimitações administrativas ou físicas.

### *4.3 Novas formas de recolha de dados*

Também, pela primeira vez, vamos utilizar a resposta, pela internet, aos questionários dos Censos 2011 que são preenchidos pelas pessoas (alojamento, família e individual).
Durante a fase de distribuição dos questionários, o recenseador deixa nos alojamentos o questionário de alojamento com o código que identifica o respectivo alojamento e um envelope onde se encontra um par de códigos chamado ID/PIN. Por fora do envelope está um código que o recenseador transcreve para o controlo do edifício e alojamento onde faz a distribuição; este código ID está

associado a um código PIN que se encontra no interior do envelope e que só pode ser aberto pelo ocupante do alojamento ou alguém a quem ele peça ajuda para responder pela internet. Com estes códigos (alojamento e par ID/PIN) a pessoa pode aceder à área de resposta pela internet e responder a todos os questionários referentes ao respectivo alojamento.
Sempre que num determinado alojamento é concluída a resposta via internet, o recenseador recebe uma mensagem *sms* a indicar o código do alojamento ao qual a resposta se refere, pelo que já não necessita de lá voltar para recolher os questionários.

### *4.4 Novas formas de difusão dos resultados*

A quantidade de informação censitária difundida também tem vindo a aumentar significativamente, tanto pela via da maior desagregação geográfica como pela maior quantidade de dados recolhidos; outro factor determinante tem sido a maior capacidade tecnológica de tratamento e difusão. Assim, como se pode verificar no quadro 2 abaixo, o forte crescimento dá-se sobretudo a partir de 1991, pelo facto de se ter decidido desagregar toda a informação a disponibilizar até freguesia.

**Quadro 2 - Número de células de informação disponibilizadas**

| Ano dos Censos | 1981 | 1991 | 2001 | 2011 (est.) |
|---|---|---|---|---|
| Nº de células de informação disponibilizadas com os resultados definitivos (milhões) | 20,6 | 314,9 | 613,2 | 648,3 |

Para 2011 a disponibilização dos resultados foi organizada em três fases de acordo com o seguinte calendário:

- **Preliminares**: até 4 meses após o momento censitário, o que corresponde à disponibilização dos respectivos resultados até Julho de 2011; estes resultados são constituídos totalizadores das unidades estatísticas provenientes dos controlos de trabalho no terreno
- **Provisórios**: até 11 meses após o momento censitário, o que corresponde à sua divulgação até Fevereiro de 2012; estes resultados são formados por 8 quadros correspondentes a um conjunto de 18 indicadores estatísticos e que serão apurados numa fase intermédia do processo de tratamento dos dados;
- **Definitivos**: a disponibilizar no quarto trimestre de 2012 e englobam uma série de sete tipos diferentes de produtos (publicações, informação com suporte em quadros, Censos 2011 em números com informação dos tês últimos censos, indicadores estruturados para utilização nacional e comunitária, informação alfanumérica apoiada na cartografia e uma amostra de microdados anonimizados para utilização especializada).

Com este leque alargado de fases de disponibilização e de produtos dos resultados definitivos procuramos responder a uma procura cada vez mais exigente e variada.
Por outro lado, todo o programa de difusão privilegia a disponibilização de resultados através da internet, nomeadamente as publicações que deverão ficar disponíveis em suporte electrónico, para além do papel.

## 5 A transição censitária

A transição censitária foi assumida como um dos principais objectivos dos Censos 2011, não necessariamente na perspectiva da sua implementação imediata, mas sim enquanto utilização dos seus resultados para "enquadrar" o futuro imediato deste objectivo.

Há dois factores determinantes para esta alteração de modelo censitário: os custos são crescentes e as alternativas são cada vez mais consistentes. Os custos são suportados pelos recursos nacionais que são cada vez mais escassos e os ficheiros administrativos são uma presença cada vez mais constante e consistente na vida dos cidadãos.

Além disso, existe uma quantidade crescente de população que se desloca frequentemente "entre residências", sendo muitas vezes difícil determinar qual a residência efectiva das pessoas que vivem alternadamente em locais diferentes. Esta situação constitui uma dificuldade adicional de encontrar essas pessoas no "local certo" no momento censitário e provoca inconsistências estatísticas na comparação de resultados entre fontes diferentes de dados, nomeadamente com os resultados de ficheiros administrativos.

Um outro argumento adicional relaciona-se com o facto de haver um número crescente de utilizadores estatísticos que necessita de informação consistente e com regularidade muito maior (mais frequente) do que a decenal, pois a realidade populacional e habitacional evolui mais rapidamente e a todos os níveis da estrutura geográfico-administrativa. Esta situação não é passível de ser resolvida apenas com inquéritos de grande escala, porque existem muitos utilizadores à escala local e regional que necessitam de suportar as suas decisões em informação cada vez mais recente e mais consistente do que previsões cujas confirmações só podem ocorrer em largos períodos de tempo.

Por outro lado, não existem limitações metodológicas nas recomendações internacionais nem impedimentos legais nacionais ou comunitários que impeçam esta mudança. Os tempos são de mudança e implementação de alternativas a nível internacional como se pode verificar no quadro seguinte:

**Quadro 3 – Evolução na utilização dos modelos censitários na UE (fonte: UNECE 2010)**

| | Clássico/Tradicional | Combinado | Registos | Outros métodos | Sem Censo | Total |
|---|---|---|---|---|---|---|
| **UE 2011** | 11 | 11 | 4 | 1 | - | 27 |
| **UE 2001** | 18 | 5 | 2 | - | 2 | 27 |

Assim, a transição está a fazer-se, sobretudo, para formas combinadas entre registos administrativos e inquéritos, o que pressupõe sempre uma utilização crescente dos registos administrativos existentes.

### *5.1 O programa de trabalhos*

O programa de trabalhos para a preparação do modelo de transição censitária decorre em paralelo com a preparação e execução dos Censos 2011. Este programa está orientado no sentido de fazer a avaliação dos principais ficheiros administrativos e cruzar os dados equivalentes entre vários ficheiros, de modo a detectar e analisar as eventuais inconsistências e comparar os respectivos resultados com os dos Censos 2011 ao nível micro das unidades geográficas.

Os trabalhos dos Censos 2011 no terreno vão utilizar um sistema de indicadores de alerta, com o objectivo de detectar situações de forte inconsistência entre os dados administrativos e estimativas sobre população e alojamentos a nível de freguesia. Para ajudar na estimação destes indicadores estamos a utilizar alguns dados de ficheiros que estão a ser analisados para a transição censitária.

Assim, a calendarização das fases de preparação integrada é a seguinte:

- Utilizar a informação administrativa existente para apoiar o Sistema de Indicadores de Alerta;
- Análise dos ficheiros administrativos existentes com informação sobre a população no período de 2009 e 2010;
- Sistematização do modelo de integração e interligação da informação em 2009, 2010 e 2011;
- Criação de uma estrutura administrativa de teste em 2011 de modo a controlar mais exaustivamente a relação entre a informação censitária recolhida e a informação administrativa existente e já analisada;
- Comparação entre os resultados censitários e os equivalentes administrativos, no primeiro semestre de 2012;
- Relatório das observações e propostas de passos futuros, no segundo semestre de 2012.

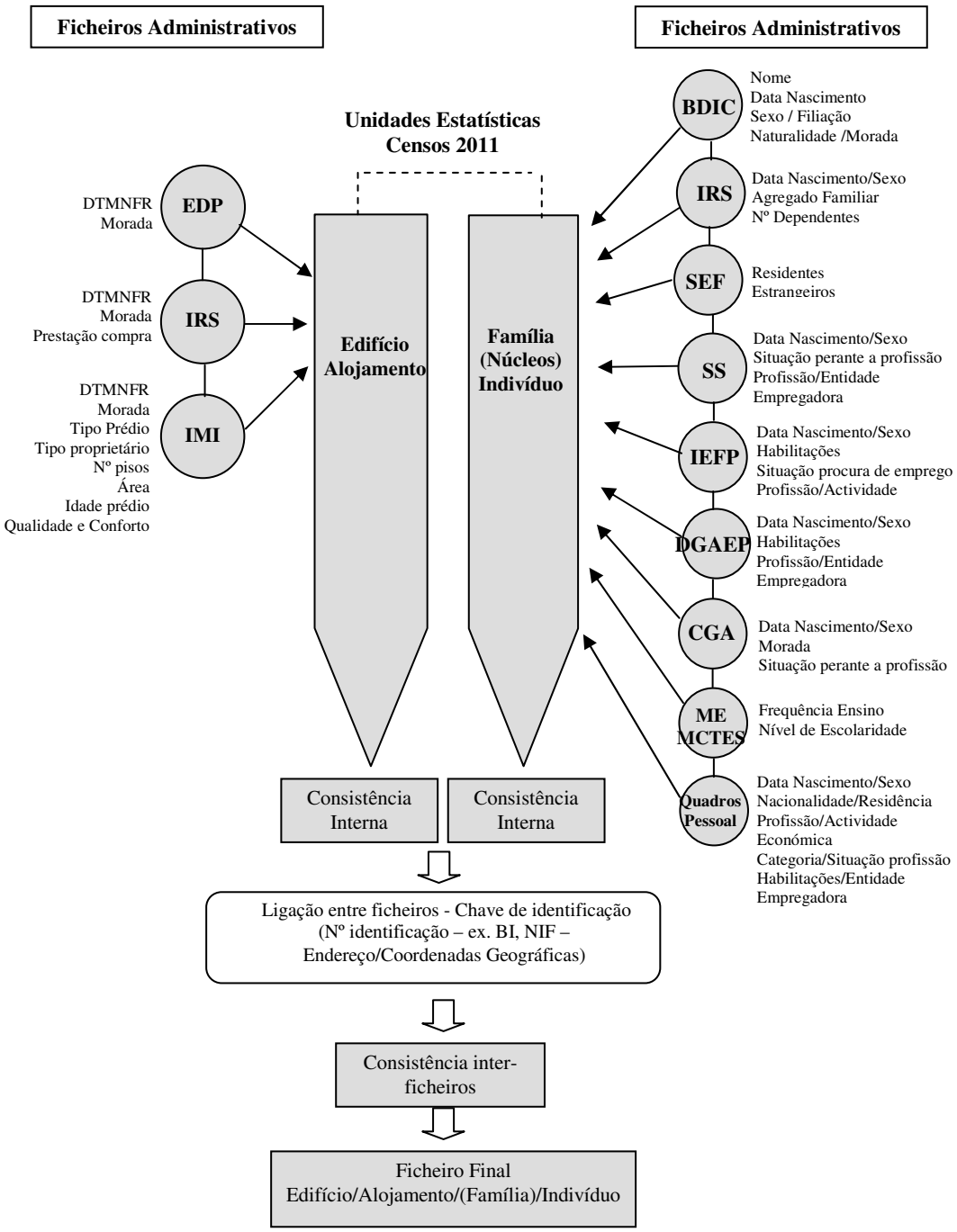
Esperamos ter, no final de 2012, uma perspectiva consistente sobre o modelo a seguir no qual a utilização dos ficheiros administrativos nos parece incontornável.

### *5.2 O modelo de análise e interligação dos ficheiros administrativos*

A transição censitária baseada na utilização de registos administrativos sempre se fez em várias fases e apoiada numa análise cuidada dos ficheiros existentes. Dificilmente se consegue compatibilizar informação administrativa e censitária ou de inquéritos sem um amplo trabalho prévio de análise e avaliação da consistência da informação existente nos ficheiros administrativos e mesmo inter-ficheiros, directamente associada aos conceitos que lhe estão subjacentes.
Este trabalho tem deparado com algumas dificuldades no acesso a esses ficheiros, mas tal tem vindo a ser ultrapassado com a análise bilateral dessas dificuldades e com a celebração de protocolos entre o INE e as entidades em causa.
Na figura 1 que se apresenta a seguir, descreve-se o fluxo e o conteúdo dos ficheiros que estamos a solicitar e a analisar.

Como se pode verificar na figura 1, há variáveis que se encontram em quase todos os ficheiros enquanto outras não existem em qualquer deles. Quanto à primeira situação, não causa qualquer problema significativo uma vez que a ligação da informação deverá permitir escolher os dados mais correctos. Quanto à ausência de informação censitária fundamental como as migrações internacionais, teremos de encontrar alternativas adequadas de a recolher, ou por via da melhoria do conteúdo dos ficheiros existentes ou através de outra metodologia estatística adequada. Contudo, em todos os países onde se avançou para a transição censitária nunca foi possível fazê-la sem ter alguns "custos de conteúdo" o que será inevitável sobretudo nas unidades estatísticas edifício e alojamento onde existe uma componente significativa de variáveis cuja relação entre conteúdo e desagregação geográfica fina apenas é possível no modelo clássico destes recenseamentos.
As análises de consistência interna de cada ficheiro e da consistência inter-ficheiros deverá permitir constituir uma base de dados de referência que deverá ser confrontada com a informação equivalente dos Censos 2011, no sentido de determinar a eventual existência de diferenças significativas nos respectivos resultados e uma aproximação às causas dessas diferenças.

**Figura 1**




**Descrição de siglas menos conhecidas**: DTMNFR (DistritoMunicípioFreguesia);BDIC (Base de Dados de Identificação do Cidadão); SS (Segurança Social).

### *5.3 Consequências desta alternativa*

Esta alternativa de modelo censitário levanta naturalmente problemas técnicos de implementação e a necessidade de garantir a inviolabilidade dos dados pessoais que se encontram nestes ficheiros. Quanto aos problemas técnicos há que relevar o seguinte:

- Outros países seguiram este caminho e conseguiram encontrar soluções técnicas para os problemas encontrados, a ponto de não haver qualquer exemplo de "retorno ao passado"; um dos resultados mais relevantes nesta matéria é o ganho de experiência e de melhoria na qualidade da informação administrativa existente;
- Custos reduzidos na produção anual de resultados estruturais sobre a população e a habitação;
- Necessidade de uma solução integrada, para observação das variáveis fundamentais que não constam dos ficheiros; para além das que são recomendadas internacionalmente, pode haver também necessidades nacionais que tenham de ser incluídas nesta solução;

Quanto ao problema da inviolabilidade dos dados pessoais, salienta-se que:

- Todos os intervenientes nestes dados estão abrangidos pelos princípios do segredo estatístico e profissional, incorrendo em penas se o violarem;
- Os ficheiros são guardados em condições de máxima segurança;
- Este modelo censitário já existe em países com populações especialmente sensíveis aos problemas da privacidade e nunca esta razão foi motivo para não seguir este modelo, obviamente com as garantias indicadas anteriormente.

A avaliação da experiência internacional nesta matéria aponta claramente no sentido de as vantagens superarem claramente os inconvenientes desta abordagem da transição censitária. Todos os países mais desenvolvidos possuem hoje um conjunto alargado de informação administrativa que os estaticistas têm a obrigação de utilizar para fins estatísticos e contribuírem também para o desenvolvimento deste tipo de informação. Atingem-se dois grandes objectivos: reduz-se o custo desta actividade estatística e diminui-se a sobrecarga estatística sobre os cidadãos.

# As sondagens e os resultados eleitorais em Portugal

Pedro Magalhães, *pedro.magalhaes@ics.ul.pt*
*Instituto de Ciências Sociais da Universidade de Lisboa*

Luís Aguiar-Conraria, *lfaguiar@eeg.uminho.pt*
*Núcleo de Investigação em Políticas Económicas da Universidade do Minho*

Miguel Maria Pereira, *miguelmaria.gp@gmail.com*
*Instituto de Ciências Sociais da Universidade de Lisboa*

## 1. Introdução

Os resultados das sondagens pré-eleitorais, que medem intenções de voto junto de amostras e as inferem para o eleitorado, são habitualmente confrontados com os resultados das eleições subsequentes. Verificando-se discrepâncias entre umas e outros, segue-se controvérsia pública, com alusões quer a "erros metodológicos" não especificados quer a tentativas de manipulação da opinião pública. Contudo, sabemos ainda pouco no caso português sobre a real dimensão dessas discrepâncias, sobre quaisquer tendências sistemáticas e ainda menos sobre o mais importante de tudo: os factores que podem afectar esses fenómenos.

Este artigo procura responder a estas perguntas. Depois desta breve introdução, inventariamos o conjunto de condições necessárias para que inquéritos amostrais sobre intenções de voto pudessem produzir resultados iguais entre si e iguais aos de uma eleição subsequente. Na secção seguinte, apresentamos dois indicadores que captam as diferenças entre resultados de sondagens e resultados eleitorais. Na quarta parte do artigo, testamos algumas hipóteses de explicação dos desvios absolutos médios verificados entre as estimativas de intenções de voto que resultaram de 287 sondagens conduzidas entre 1991 e 2009 e os resultados das eleições subsequentes. Finalmente, na quinta parte, damos atenção ao que passa em relação a partidos concretos, e testamos algumas hipóteses a esse respeito.

## 2. Intenções de voto e resultados eleitorais

Que condições teriam de se verificar para que não houvesse variações entre diferentes sondagens e diferenças entre estas e os resultados das eleições respectivas?

A primeira condição necessária seria que *essas estimativas não se baseassem em amostras da população*. Por outras palavras, não lhes poderia estar associado erro amostral. Admitindo que é impossível usar a totalidade da população como amostra, e presumindo que as amostras são extraídas aleatoriamente, dando a cada um dos membros do universo a mesma probabilidade de serem seleccionados, as Leis dos Grandes Números dizem-nos que se a amostra crescer indefinidamente esse erro tende a desaparecer. A alternativa, extrair um número infinito de amostras e calcular a média das estimativas, também não é viável. O melhor que podemos fazer é basearmo-nos numa implicação do Teorema Limite Central: as médias amostrais baseadas em amostras aleatórias (ou probabilísticas) tendem a ser normalmente distribuídas. Partindo de uma estimativa baseada numa dessas amostras e

conhecendo a sua dimensão, podemos estimar intervalos de confiança, de 95%, por exemplo. Fazendo este exercício diversas vezes, os intervalos de confiança deverão incluir o valor real do parâmetro de interesse 95% das vezes. Tudo isto significa que é extremamente improvável que exactamente a mesma estimativa resulte de várias amostras aleatórias obtidas de uma mesma população. Logo, discrepâncias entre sondagens e discrepâncias entre essas sondagens e os resultados eleitorais são um preço inevitável a pagar pelo recurso a amostras.

A segunda condição necessária para que as estimativas das sondagens fossem iguais entre si e iguais aos resultados eleitorais seria a *inexistência de erros sistemáticos de cobertura, de não-contacto ou de não-resposta* na constituição da amostra. Teria de haver uma lista – a chamada base de sondagem - contendo todos os membros da população eleitora e da qual se extraísse a amostra dando a cada membro do universo a mesma probabilidade (ou pelo menos uma probabilidade conhecida) de ser seleccionado. Essa lista não existe. Na sua ausência, uma alternativa comum consiste em recorrer a listas nacionais dos números de telefones fixos ou até a geração aleatória de números de telefone. Contudo, isto implica a impossibilidade de que um eleitor que não resida em domicílios com telefone seja incluído na amostra. Outra alternativa comum consiste em recorrer a selecção aleatória de localidades, de edifícios, de domicílios e de residentes nesses domicílios para constituir a amostra, evitando os erros de cobertura criados pelo recurso ao telefone. Contudo, continua a haver aqui a possibilidade de erros. Para os evitar, teríamos de garantir que seria sempre possível contactar e inquirir aqueles que são seleccionados através deste procedimento aleatório. Mas mesmo que o contacto com os indivíduos seleccionados fosse sempre possível, nada garante que esses contactados responderiam ao inquérito. Naturalmente, estes erros não teriam consequências relevantes se os indivíduos que fazem parte da base de amostragem e os que não fazem, os que são contactados e os que não são, e os que respondem e não respondem ao inquérito se distribuíssem aleatoriamente. Contudo, há boas razões para supor que as características sociais e as atitudes políticas dos eleitores não se distribuam dessa forma. O resultado é erro sistemático de amostragem.

A terceira condição necessária para que as estimativas das sondagens fossem iguais entre si e iguais aos resultados eleitorais seria a *inexistência de erros sistemáticos de medição.* Contudo, não é possível garantir a ausência desse tipo de erros. Por exemplo, intenções comportamentais vistas como socialmente indesejáveis ou claramente minoritárias – a abstenção, o voto em pequenos partidos ou em partidos situados em posições ideologicamente extremas – podem ser sistematicamente omitidas pelos inquiridos numa sondagem, e isso implicará discrepâncias entre as inferências feitas na base na medição junto da amostra e as reais intenções dos indivíduos. Note-se que, no caso da medição da intenção de abstenção, a sua subestimação sistemática é compatível, apesar de tudo, com a ausência de discrepâncias sistemáticas entre as intenções de voto estimadas e as intenções de voto reais. Mas para tal seria necessário que as intenções de voto captadas junto dos que acabam por se abster fossem iguais às intenções dos que acabam por votar realmente. Se não for esse o caso, os resultados de sondagens serão necessariamente diferentes das reais intenções dos indivíduos e, por extensão, dos resultados eleitorais.

A quarta condição necessária seria a *estabilidade das intenções de voto* ao longo de todo o período em que são medidas e até às eleições. Se fosse possível conjugar esta condição com todas as anteriores, todas as sondagens, feitas a três meses ou a três dias das eleições, produziriam exactamente os mesmos resultados, porque o valor verdadeiro dessas intenções permaneceria o mesmo. Contudo, esta pressuposição é pouco plausível. Muita coisa pode ocorrer entre diferentes medições das intenções comportamentais de um conjunto de indivíduos que pode contribuir para as modificar. Basta lembrar o atentado terrorista de 11 de Março de 2004 que vitimou Madrid e minou a credibilidade do então presidente de governo José Maria Aznar. Este atentado, ocorrido a três dias das eleições, terá tido influência nas intenções de voto.

Suponhamos, contudo, que todas as quatro condições referidas até agora se verificavam. Sem erro amostral, sem enviesamentos trazidos por erros de cobertura, não-contacto e não-resposta, sem

enviesamentos causados por erros de medida e com total estabilidade das intenções comportamentais dos eleitores, todas as sondagens produziriam iguais estimativas sobre as intenções de voto dos eleitores ao longo de uma campanha. Mesmo assim, para que não existissem discrepâncias entre os resultados de sondagens e os resultados das eleições, uma quinta e última condição seria necessária: que as *intenções comportamentais dos eleitores estivessem perfeitamente correlacionadas com os seus comportamentos futuros*. Contudo, essa relação não é perfeita, e torna-se compreensivelmente mais fraca quando se trata da relação entre a intenção de votar e a real participação em eleições, aspecto onde recursos, oportunidades e circunstâncias imprevistas tendem a fazer com que a conversão entre intenções e comportamentos não dependa exclusivamente da vontade dos indivíduos (Ajzen 1991). Logo, o que vai inevitavelmente suceder - especialmente se esse grau de controlo não estiver distribuído de forma homogénea entre os diferentes grupos de eleitores definidos pelas suas intenções de voto - é uma discrepância entre as intenções e comportamentos de voto registados em sondagens e em resultados eleitorais.

Verificamos assim que, na prática, nenhuma das condições necessárias para uma identidade sistemática entre os resultados de sondagens e os resultados eleitorais se verifica. Mas isso não implica a ausência de variações consideráveis, de sondagem para sondagem e de eleição para eleição, no grau de semelhança entre os resultados das sondagens e aqueles que vêm a ser os resultados eleitorais. O que explica essas variações? E antes ainda de buscar explicações, como se descrevem? Este é o tema da secção seguinte.

## 3. Medir discrepâncias entre resultados de sondagens e de eleições

Como se pode medir a discrepância geral entre os resultados de cada sondagem e os resultados eleitorais? Num relatório para o *Social Science Research Council* (Mosteller et al. 1949), Mosteller e os seus colegas examinaram oito medidas diferentes. Uma revisão mais recente destas medidas, aplicada às eleições americanas, isolou duas com as propriedades mais desejáveis: os chamados métodos 3 e 5 de Mosteller (Mitofsky 1998). O método 3 consiste na média dos desvios absolutos, em pontos percentuais, para cada um dos partidos/candidatos, entre os resultados da sondagem e os resultados eleitorais. O método 5 consiste numa diferença entre duas diferenças: primeira é a diferença em pontos percentuais entre as intenções de voto estimadas para os dois primeiros candidatos/partidos; a segunda é a diferença entre os resultados eleitorais verificados para os dois primeiros partidos/candidatos. Por outras palavras, o método 3 fornece-nos um indicador genérico da "precisão" das sondagens utilizando informação sobre estimativas de intenção de voto e resultados eleitorais para os vários partidos, enquanto o método 5 se concentra nos dois principais partidos em cada eleição. Tendo em conta a natureza multipartidária do sistema político português – em contraste com o bipartidarismo dos Estados Unidos, de onde estas medidas são originárias – a nossa opção neste trabalho vai para o método 3.

Como se calcula? Cada sondagem fornece estimativas de intenções de voto para vários partidos, expressas em percentagens. Tipicamente, há variações entre os resultados de sondagens no que respeita à base de cálculo dessas percentagens. Por exemplo, elas podem ser calculadas em relação à totalidade da amostra, à totalidade da amostra menos aqueles que declaram não tencionar votar ou à totalidade da amostra menos aqueles que declaram não tencionar votar ou não saber em que partido ou candidato tencionam votar. Logo, um primeiro passo indispensável consiste em dar a todas estas estimativas uma mesma base. Se $e_{ps}$ representar a estimativa de intenções de voto facultada por uma sondagem para o partido $p$ (com $p = 1,...,n$) na sondagem $s$ numa dada eleição, $e'_{ps}$ resulta de um novo cálculo dessa estimativa, tendo desta vez por base apenas a soma das estimativas de intenções de voto em partidos facultadas pela sondagem:

$$e'_{ps} = \frac{e_{ps}}{\sum_{p=1}^{n} e_{ps}}$$

A comparação de $e'_{ps}$ com os resultados eleitorais exige que, para cada sondagem, esses resultados sejam calculados na mesma base das estimativas. Assim, se $r_p$ for o resultado eleitoral de um partido $p$ numa dada eleição, e $r_{ps}$ o resultado eleitoral de um partido $p$ nessa eleição para o qual a sondagem $s$ tenha fornecido estimativas de intenções de voto, temos de o transformar, para cada sondagem, de forma a que seja comparável com o correspondente $e'_{ps}$.

$$r'_{ps} = \frac{r_{ps}}{\sum_{p=1}^{n} r_{ps}}$$

Feita a transformação, podemos estimar $D_s$, que designaremos de *desvio absoluto médio*: a média dos desvios absolutos em pontos percentuais, para cada um dos partidos/candidatos, entre os resultados da sondagem e os resultados eleitorais.

$$D_s = \frac{\sum_{p=1}^{n} \left| e'_{ps} - r'_{ps} \right|}{n}$$

Por construção, $D_s$ terá sempre um valor não negativo. Quando maior for esse valor, maior a discrepância genérica entre as estimativas de intenções de voto obtidas para os diferentes partidos numa sondagem e os resultados eleitorais que cada um desses partidos obteve.

Estamos também interessados num segundo tipo de fenómeno: a tendência sistemática para subestimar, ou sobrestimar, os resultados eleitorais de partidos específicos. Esse indicador – que designaremos de *enviesamento* – pode ser facilmente calculado com base nos já estimados $e'_{ps}$ e $r'_{ps}$. $B_{ps}$ representa o enviesamento na estimação do resultado do partido $p$ na sondagem $s$ :

$$B_{ps=} e'_{ps} - r'_{ps}$$

$B_{ps}$ poderá ter valores positivos ou negativos. Quando positivo, isso significa que o resultado eleitoral do partido $p$ numa das eleição por referência à sondagem $s$ acabou por ser inferior ao valor de intenção de voto estimado pela sondagem .

## 4. O que determina as discrepâncias entre os resultados de sondagens e de eleições?

A nossa base de dados é constituída por todas as sondagens tornadas públicas através dos órgãos de comunicação social cujo trabalho de campo tenha terminado nos últimos 100 dias antes de 65 eleições realizadas em Portugal, entre 1991 e 2009. Estas 65 eleições incluem todas as legislativas (seis) e europeias (quatro) realizadas no período, assim como 55 eleições para a presidência de câmaras municipais, realizadas nas eleições autárquicas de 2005 e 2009.[1] Estas sondagens foram conduzidas por 19 empresas diferentes ao longo destas quase duas décadas.

**Quadro 1. Eleições e sondagens**

| Eleição | Nº de sondagens publicadas | Institutos/empresas responsáveis | Distância média (dias) entre ultimo dia de trabalho de campo e eleição | Dimensão amostral média |
|---|---|---|---|---|
| Legislativas 1991 | 22 | 5 | 41,4 | 1364 |
| Europeias 1994 | 12 | 5 | 29,8 | 1104 |
| Legislativas 1995 | 25 | 7 | 45,4 | 1519 |
| Europeias 1999 | 13 | 6 | 32,9 | 1082 |
| Legislativas 1999 | 20 | 7 | 39,8 | 986 |
| Legislativas 2002 | 22 | 7 | 28,7 | 1242 |
| Europeias 2004 | 14 | 6 | 22,0 | 1212 |
| Legislativas 2005 | 26 | 7 | 30,8 | 1177 |
| Autárquicas 2005 | 66 | 9 | 29,2 | 655 |
| Europeias 2009 | 13 | 5 | 20,9 | 1227 |
| Legislativas 2009 | 16 | 5 | 27,5 | 1159 |
| Autárquicas 2009 | 38 | 7 | 21,8 | 715 |
| **Total** | **287** | **19** | **30,8** | **1027** |

Recordemos a primeira das cinco condições necessárias para que não existissem discrepâncias entre resultados destas sondagens e os resultados eleitorais: a inexistência de erro amostral. Uma implicação desta condição pode ser formulada como uma hipótese:

*H1: Quanto maior o erro amostral máximo associado a uma sondagem, maior deverá ser* $D_s$.

Resultados empíricos que apoiam com esta hipótese têm sido encontrados quer em estudos internacionais (DeSart e Holbrook 2003) quer em estudos sobre as sondagens em Portugal (Magalhães 2005; Magalhães e Moreira 2007; Pereira 2011). Note-se, contudo, que são de esperar algumas perturbações nesta relação. A informação disponível sobre a dimensão amostral, recolhida nas fichas técnicas divulgadas pela comunicação social aquando da divulgação dos resultados, refere-se tipicamente à totalidade da amostra, e não à dimensão da sub-amostra que manifestou uma intenção de voto num determinado partido. Imaginemos que, numa amostra de 1000 indivíduos, 500 declararam não tencionar votar, não saber em que partido tencionam votar, que tencionam votar em branco ou nulo ou recusaram responder à pergunta sobre intenção de voto, e que, dos restantes 500, 250 manifestaram tencionar votar no partido A e 250 no partido B. Por outras palavras, a estimativa de intenção de voto válido para os dois partidos é de 50%, e o erro amostral que lhe está associado é de +/- 4,38 pontos percentuais. Imaginemos agora que, numa amostra igualmente com 1000 indivíduos, são manifestadas 750 intenções de voto válidas, repartindo-se igualmente pelo partido A e pelo partido B (375 cada). A estimativa de intenção de voto válido para os dois partidos continua a ser de 50%, mas o erro amostral que lhe está associado é de +/- 3,58 pontos percentuais, inferior à situação anterior. Contudo, se a única informação de que dispomos é a dimensão total da amostra - ou seja 1000, em ambos os casos - esta diferença não vai ser captada.

Enviesamentos causados por erros de cobertura, de não-contacto ou de não-resposta constituem as violações da segunda condição para a inexistência de discrepâncias entre resultados de sondagens e de eleições. Até muito recentemente, a prática de reporte de taxas de contacto e de resposta nas fichas técnicas em Portugal encontrava-se bastante desregrada, sendo inconsistente entre sondagens e muitas vezes ausente das próprias fichas técnicas. Desta forma, esses valores não podem ser usados numa

[1] Para o caso das autárquicas, toda a informação foi extraída do site da Marktest e dos dois "dossiers autárquicas" que dedicou às sondagens realizadas em 2005 e 2009. Disponível em: http://www.marktest.com/wap/a/p/id~cd.aspx e http://www.marktest.com/wap/a/p/id~f9.aspx. Para as restantes eleições, a informação foi directamente recolhida dos jornais.

análise das sondagens conduzidas desde 1991. Contudo, há pelo menos uma hipótese que deriva desta condição. Por um lado, as sondagens telefónicas deixam de fora das amostras a população sem telefone fixo, ao contrário do que sucede em sondagens presenciais. Por outro lado, as taxas de resposta dessas sondagens tendem a ser inferiores às das sondagens conduzidas presencialmente junto dos eleitores nos seus domicílios (Asher 1992; Miller 2002). Logo, na medida em a população não coberta e com maior propensão a não responder tenda a exibir intenções de voto distintas da população representada na amostra, as sondagens telefónicas tenderão a ser caracterizadas por maiores erros sistemáticos de amostragem.

*H2: Sondagens telefónicas causam maiores valores de* $D_s$.

Estudos anteriores tendem, contudo, a não dar apoio empírico a esta hipótese (Crespi 1988; Crewe 1993; Moon 1999; Magalhães 2005; Magalhães e Moreira 2007). Por um lado, a utilização de inquirição presencial pode ter também consequências negativas a nível da homogeneização dos procedimentos de aplicação dos inquéritos e de capacidade de monitorização do trabalho dos inquiridores, potenciais fontes de erros sistemáticos na construção da amostra e medição das intenções dos indivíduos. Por outro lado, categorias como "sondagem telefónica" podem ocultar enorme diversidade de procedimentos destinados a corrigir erros de cobertura, de não-contacto e não-resposta, tais como o recurso a ponderações pós-amostrais que reequilibrem a amostra do ponto de vista de características conhecidas da população. Na ausência de informação que nos permita construir variáveis que captem essa diversidade, quaisquer efeitos das sondagens telefónicas na relação entre resultados de sondagens e resultados de eleições serão mais difíceis de detectar.

Erros sistemáticos de medição na intenção de participar no sufrágio e nas intenções de voto dos eleitores constituem violações da terceira condição para a congruência entre sondagens e resultados eleitorais. Há algumas características do instrumento de medida – o questionário – que se encontram relativamente bem identificadas como podendo reduzir erros sistemáticos de medição das intenções de voto dos inquiridos, tais como a introdução de perguntas-filtro para a detecção de indivíduos cuja probabilidade de votar seja baixa (Crespi 1988; Visser et al. 2000) ou a omissão de uma categoria "não sabe" como uma opção de resposta explícita (Gilljam e Granberg 1993; Visser et al. 2000). Contudo, uma vez mais, as fichas técnicas disponíveis sobre as sondagens conduzidas em Portugal são, na sua esmagadora maioria para o período sob análise, insuficientes para medir estes atributos dos inquéritos. Há, no entanto, duas hipóteses que podem ser avançadas deste ponto de vista. A primeira é que as sondagens telefónicas sejam caracterizadas por maiores erros sistemáticos de medição. Holbrook et al. (2003) mostram que, em comparação com inquéritos realizados face-a-face, sondagens telefónicas fazem com que os indivíduos tendam a ocultar mais frequentemente atitudes e comportamentos socialmente indesejáveis e a dar respostas irreflectidas. Semelhante efeito na discrepância entre resultados de sondagens e resultados de eleições tenderia a ser captada novamente através da hipótese 2.

*H2: Sondagens telefónicas têm maiores valores de* $D_s$.

Uma outra implicação observável da existência de erros sistemáticos de medida está relacionada com a abstenção. Na medida em que fenómenos de desejabilidade social façam com que a intenção de votar tenda a ser sobrestimada nas sondagens, então o efeito desse erro sistemático na discrepância entre sondagens e resultados eleitorais será tanto maior quanto maior for a abstenção numa dada eleição. Esta hipótese foi confirmada em estudos sobre eleições nos Estados Unidos (Crespi 1988) e em eleições nacionais em Portugal (Magalhães 2005), mas não em eleições locais (Magalhães e Moreira 2007), pelo que merece especial atenção neste estudo.

*H3: Quanto maior a abstenção na eleição, maior deverá ser* $D_s$.

Na medida em que as intenções de voto não sejam estáveis – violação da quarta condição – e que a relação entre intenções de voto e comportamentos não seja perfeita – quinta condição - o momento em que cada sondagem é realizada há-de afectar a discrepância dos seus resultados com os resultados eleitorais. A hipótese mais frequentemente testada a este nível consiste na existência de uma relação positiva entre a distância temporal entre o trabalho de campo e o dia da eleição e $D_s$ (ou outros indicadores de desvio). Como assinala Ajzen (1991), a relação entre medidas de intenções e medidas de comportamentos é tanto mais forte quanto menor o espaço de tempo que medeia entre as duas, diminuindo a probabilidade de que eventos imprevistos intervenham para modificar intenções. E na medida em que as campanhas contribuam para actualizar as crenças dos eleitores quanto às consequências das suas decisões, as suas atitudes e as suas intenções (Gelman e King 1993), quanto mais longe do dia da eleição forem medidas das intenções dos eleitores, menos actualizada há-de ser essa medição em relação ao estado da opinião pública que vigora no dia da eleição. Finalmente, cientes de que as comparações entre os resultados das últimas sondagens e das eleições vão ser feitas publicamente, as empresas podem, por razões reputacionais, investir mais nestes sondagens que nas anteriores em termos de rigor de procedimentos e controlo de qualidade.

*H4a: Quanto maior a distância entre o trabalho de campo e a eleição, maior deverá ser* $D_s$.

Esta hipótese recebeu apoio empírico em vários estudos sobre o tema, seja em Portugal (Magalhães 2005; Magalhães e Moreira 2007) seja noutros países (Crespi 1988; DeSart e Holbrook 2003). Contudo, é possível formular uma hipótese alternativa a este respeito. Ela consiste em supor que a proximidade temporal das sondagens em relação às eleições tenha um efeito *negativo* sobre a sua relação com os resultados eleitorais. Goodin e Rice (2009) mostram que, pelo menos para algumas eleições e países, as intenções de voto tal como medidas antes da campanha acabam estar mais próximas dos resultados eleitorais do que quando medidas durante a campanha. Dois argumentos parecem sustentar este resultado. Por um lado, os indivíduos cujas intenções de voto são mais instáveis e afectadas pelos eventos da campanha são também aqueles com menor probabilidade de votar. Por outro lado, este fenómeno pode estar ligado a erros de medição: confrontados com uma pergunta numa sondagem, os eleitores tendem a economizar esforços e a fornecer respostas pouco reflectidas, baseadas nos eventos do dia-a-dia da campanha. Contudo, ao chegarem à cabine de voto, os eleitores tomam decisões ponderadas recorrendo a toda a informação de que dispõem (Goodin e Rice 2009: 905; ver também Martin, Traugott e Kennedy 2005).

*H4b: Sondagens realizadas durante a campanha eleitoral deverão ter maiores valores de* $D_s$ .

Uma hipótese adicional a este respeito pode combinar elementos das anteriores. Por um lado, é possível que sondagens conduzidas antes da campanha, ao captarem as intenções de voto mais estáveis entre a população, acabem por estar mais próximas dos resultados eleitorais que outras que são conduzidas durante os momentos de "perturbação" trazidos pela campanha. Por outro lado, isto não é incompatível com a noção de que, *durante a campanha* e até ao seu final, ocorra uma crescente cristalização das intenções de voto e captação dos factores que as vão determinar no dia da eleição, assim como um crescente esforço das empresas de sondagens para obter medidas mais precisas. Logo:

*H4c: A relação entre a distância entre o trabalho de campo e a eleição e* $D_s$ *deverá ser não monotónica*

Uma quinta hipótese está ligada ao grau de competitividade da eleição. Eleições pouco competitivas, em que o vencedor parece garantido à partida, podem perturbar a correspondência entre intenções e comportamentos. Por um lado, a percepção de uma opção de voto claramente dominante pode reforçar

erros sistemáticos de medida ligados a desejabilidade social, fazendo com que os eleitores declarem intenções de voto num partido claramente dominante quando, na realidade, tencionam votar noutros partidos (Noelle-Neumann 1993). Por outro lado, grandes margens de vitória podem constituir incentivos para a mudanças de última hora das intenções de votar e das opções de voto. Elas estimulam a desmobilização diferencial dos eleitores, especialmente daqueles que, em sondagens, teriam manifestando intenções de voto num partido que adquire grande favoritismo, assim como mudanças de última hora de intenções de voto em grandes partido para intenções de voto em pequenos partidos, ou seja, de voto estratégico para voto sincero (Crespi 1988; Magalhães 2005).

*H5: Quanto maior a margem de vitória do primeiro sobre o segundo partidos, maior deverá ser* $D_s$.

Um conjunto final de hipóteses está relacionado com aspectos não directamente mensuráveis da realização das sondagens mas que poderão ter consequências relevantes. Mencionámos já a possibilidade de que uma aproximação final das sondagens aos resultados eleitorais possa resultar do facto de as empresas empregarem maiores recursos para a constituição de amostras representativas da população e na adopção de procedimentos de maior exigência na correcção de erros de medição. Três factores adicionais podem ajudar a captar as diferentes capacidades empregadas pelas empresas para a constituição de "boas" amostras e para a correcta medição das intenções de voto. Primeiro, esperamos que eleições mais importantes tendam a gerar menores discrepâncias entre os resultados das sondagens e das eleições. Segundo, empresas com maior experiência na realização deste tipo de trabalhos tenderão a gerar resultados com menores desvios em relação às eleições. Finalmente, eleições onde mais do que uma empresa esteja a realizar sondagens tenderão a ser caracterizadas, por efeito da competição entre empresas, por resultados mais precisos.

*H6: Sondagens realizadas em eleições para as quais foi realizado um maior número de sondagens (importância) deverão ser caracterizadas por um valor menor de* $D_s$.

*H7: Quanto maior o número acumulado de sondagens prévios realizado por uma empresa , menor deverá ser* $D_s$.

*H8: Sondagens em eleições para as quais mais do que uma empresa se encontra a realizar estudos deverão ser caracterizadas por valor menor de* $D_s$.

O quadro 2 mostra os resultados de regressões lineares através das quais testamos as nossas hipóteses. A variável dependente é $D_s$. Para além das variáveis que servem para testar as hipóteses anteriores, adicionamos algumas variáveis de controlo: a presença de candidatos independentes em eleições autárquicas, que estudos anteriores (Magalhães e Moreira 2007) sobre as sondagens em eleições autárquicas mostraram estar associada a valores superiores de $D_s$; o número de anos ou suas fracções decorridos desde o início da série, tomando em conta a possibilidade de tendências seculares de aumento ou diminuição de $D_s$; uma variável muda com valor 1 se a eleição era autárquica, controlando o facto de as nossas sondagens sobre eleições autárquicas se concentrarem exclusivamente nos anos de 2005 e 2009, ao contrário do que sucede com as outras eleições; variáveis mudas para cada uma das empresas de sondagens que conduziram mais do que 5 sondagens, captando *house effects* que não são medidos através de variáveis propriamente ditas (com o grupo das restantes empresas a servirem de categoria de referência); e o valor de *n*, ou seja, o número de partidos sobre o qual cada sondagem fornece informação. Como assinala Mitofsky (1998), a principal desvantagem do método 3 é o facto de reduzir o valor do erro quanto maior for o número de partidos e candidatos tomados em conta: à medida que esse valor aumenta, aumenta também a probabilidade de que a sondagem contemple partidos cujos resultados eleitorais são percentualmente baixos. Nesses casos, o erro amostral associado será sempre mais reduzido do que sucede na estimação de partidos cujos resultados esteja próximos dos 50%. Logo, qualquer análise dos factores que determinam $D_s$ terá de ter *n* em conta como variável de controlo.

O modelo 1 estima uma relação linear entre o tempo (em semanas) decorrido entre o trabalho de campo e a eleição e o desvio absoluto médio. O modelo 2 testa a hipótese 4b, introduzindo uma variável muda para todas as sondagens realizadas nas últimas três semanas antes da eleição. O modelo 3 testa a hipótese de uma relação quadrática entre as semanas antes da eleição e o desvio absoluto médio. Como não temos informação sobre o modo de inquirição para 9 das 287 sondagens, o número de observações fica reduzido a 279. Análises iniciais revelaram a presença de heterocesdasticidade, pelo que estimamos erros-padrão robustos de White.

**Quadro 2: Os determinantes de $D_s$**

| | Modelo 1 | Modelo 2 | Modelo 3 |
|---|---|---|---|
| Erro amostral máximo | 0,47 (0,44) | 0,46 (0,43) | 0,40 (0,44) |
| Sondagem telefónica | -0,23 (0,43) | -0,19 (0,43) | -0,27 (0,43) |
| Taxa de abstenção | -0,02 (0,02) | -0,02 (0,02) | -0,02 (0,02) |
| Semanas entre eleição e trabalho de campo | 0,12 (0,04)*** | - | 0,44 (0,11)*** |
| Quadrado de semanas entre eleição e trabalho de campo | - | - | -0,02 (0,01)*** |
| Sondagens conduzidas nas últimas três semanas da campanha | - | -1,03 (0,26)*** | - |
| Margem de vitória em pontos percentuais | -0,05 (0,03) | -0,05 (0,03) | -0,05 (0,03) |
| Número de sondagens realizado na eleição | -0,09 (0,04)** | -0,09 (0,04)** | -0,09 (0,04)** |
| Número acumulado de sondagens realizado pela empresa na base | -0,05 (0,02)** | -0,05 (0,02)** | -0,05 (0,02)** |
| Mais do que um instituto a conduzir sondagens | -2,02 (0,90)** | -1,92 (0,88)** | -1,87 (0,88)** |
| Candidatos independentes | 4,97 (1,45)*** | 4,94 (1,43)** | 4,93 (1,42)*** |
| Anos desde início da série | -0,12 (0,06)** | -0,11 (0,06)* | -0,12 (0,06)* |
| Eleição autárquica | -2,88 (0,89)*** | -2,78 (0,90)*** | -2,55 (0,86)*** |
| Nº de partidos considerados | -2,17 (0,57)*** | -2,12 (0,56)*** | -2,10 (0,55)*** |
| Eurosondagem | 0,84 (0,63) | 0,77 (0,62) | 0,65 (0,62) |
| Aximage/SIC-Visão | 0,87 (0,57) | 0,80 (0,57) | 0,78 (0,57) |
| Marktest | 1,24 (0,53)** | 1,14 (0,53)** | 1,02 (0,53)** |
| Católica-CESOP | 0,96 (0,71) | 0,98 (0,70) | 0,83 (0,71) |
| Intercampus | 0,43 (0,70) | 0,43 (0,69) | 0,45 (0,69) |
| Euroteste | 0,73 (0,59) | 0,68 (0,57) | 0,70 (0,58) |
| Gemeo-IPAM/IPAM | -0,31 (1,32) | -0,39 (1,31) | -0,34 (1,31) |
| Euroexpansão | 1,44 (0,61)** | 1,26 (0,61)** | 1,30 (0,61)** |
| IPOM | 0,87 (1,01) | 1,08 (1,03) | 1,01 (1,01) |
| REGIPOM | 2,67 (1,04)** | 2,79 (1,05)*** | 2,78 (1,02)*** |
| Metris | -1,41 (0,57)** | -1,31 (0,54)** | -1,25 (0,54)* |
| Constante | 16,94 (3,86)*** | 17,63 (3,79)*** | 15,87 (3,66)*** |
| N | 279 | 279 | 279 |
| $R^2$ | 0,53 | 0,53 | 0,53 |

*p<0,10; **p<0,05; ***p<0,01 (bilateral); erros-padrão robustos entre parêntesis

Os modelos explicam cerca de metade da variância na variável dependente. As variáveis de controlo fornecem desde logo resultados relevantes. Previsivelmente, o desvio absoluto médio é menor quanto maior for o número de partidos considerado para o seu cálculo. Da mesma forma, eleições com candidatos independentes geram maior desvio absoluto médio. Menos previsíveis eram os resultados das duas outras variáveis de controlo. Por um lado, eleições autárquicas parecem caracterizar-se por menores discrepâncias entre sondagens e resultados eleitorais. Note-se que a mera correlação entre eleição autárquica e $D_s$ é positiva, mas essa relação inverte-se assim que introduzimos controlos para factores como a abstenção, a presença de candidatos independentes ou variáveis mudas para as várias empresas. Por outro lado, a variável "Anos desde o início da série" tem uma relação negativa sobre o desvio absoluto médio. Por outras palavras, esse desvio tem aumentado com o tempo decorrido desde 1991, a um ritmo de cerca de 1,2 pontos percentuais a cada dez anos.

Não encontramos confirmação empírica para as hipóteses 1, 2, 3 ou 5. Em três dos casos – efeitos das sondagens telefónicas, da abstenção e da margem de vitória – os sinais dos coeficientes são até opostos aos previstos, mas estão muito longe de significância estatística a níveis convencionais. A relação entre erro amostral máximo e o desvio absoluto médio é positiva, como antecipámos, mas carece também de significância estatística. Uma explicação possível para estes resultados seria a possibilidade de que estivessem fortemente correlacionadas com outras variáveis no modelo. Contudo, testes de multicolinearidade mostram que o factor de inflação da variância mais elevado para qualquer um das quatro variáveis é de 3,4, sugerindo que esse problema não é relevante neste caso.

Já a hipótese 4b é claramente refutada: como vemos no modelo 2, sondagens realizadas nas últimas três semanas antes das eleições caracterizam-se por *menor* (e não maior) desvio absoluto médio. Contudo, apesar de o modelo 1 sugerir que o desvio absoluto médio diminui à medida que nos aproximamos da eleição, o modelo 3 fornece uma leitura mais fina do fenómeno. Ao longo dos 100 dias antes da eleição sobre os quais temos observações, o desvio absoluto médio começa por *aumentar* à medida que o tempo passa: as sondagens realizadas inicialmente estão menos distantes daqueles que vêm a ser os resultados eleitorais do que algumas das realizadas em momentos posteriores. Contudo, os valores máximos desse desvio são atingidos perto de nove semanas antes da eleição, começando depois a diminuir de novo até muito perto da eleição. Por outras palavras, os resultados apoiam a hipótese 4c: uma relação não monotónica entre o desvio absoluto médio e a distância temporal da eleição em que a sondagem é realizada. Finalmente, as hipóteses 6, 7 e 8 são confirmadas. Por um lado, eleições às quais foi dada mais atenção pelas empresas de sondagens e onde há competição entre diferentes empresas são caracterizadas por estimativas de intenções de voto mais próximas dos resultados das eleições. Por outro lado, sondagens feitas por empresas com maior experiência acumulada na medição de intenções de voto tendem a exibir menores desvios absolutos médios em relação aos resultados eleitorais.

## 5. O que determina enviesamentos na estimação das intenções de voto em determinados partidos?

Independentemente das discrepâncias genéricas entre as intenções de voto e os resultados eleitorais, há um segundo fenómeno de potencial interesse: a tendência das intenções de voto para constituírem sobrestimações ou subestimações sistemáticas dos resultados que determinados partidos acabam por obter nas eleições. O gráfico 1 mostra o valor médio, para todas as sondagens realizadas, de $B_{ps}$ para os cinco principais partidos (PS, PSD, CDS-PP, BE e CDU) e para listas de coligação entre o PSD e o CDS-PP (que se verificou nas eleições europeias de 2004 e em várias eleições locais em 2005 e 2009). Claramente, tem havido uma tendência para que as intenções de voto captadas para o PS e (em menor grau) o PSD tenham sido superiores àqueles que vêm a ser os seus resultados eleitorais, e uma tendência oposta para o CDS-PP e para a CDU. Nos restantes casos, o valor do enviesamento não é estatisticamente diferente de zero.

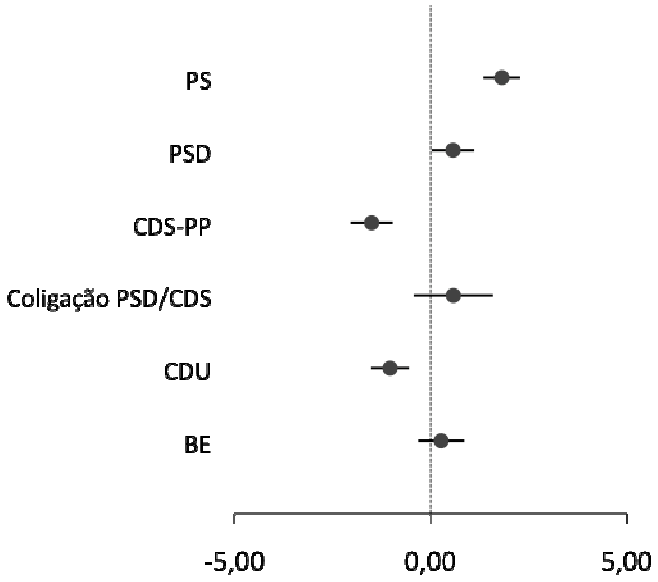



**Gráfico 1: Valores médios de $B_{ps}$ por partido**

O que pode estar por detrás deste padrão? Uma primeira ordem de factores pode estar relacionada com erros de medição, particularmente quando ligados a questões de desejabilidade social no fornecimento de respostas a perguntas de intenção de voto. Opções vistas como minoritárias num dado contexto podem tender a ser omitidas pelos eleitores quando questionados, em favor de não-respostas, manifestações de indecisão ou de intenções de voto em partidos maioritários ou dominantes. Daqui decorrem duas hipóteses. A primeira é que partidos vistos como favoritos à vitória numa dada eleição (definidos aqui como liderando todas as sondagens de intenção de voto para uma dada eleição) tenderão a receber uma percentagem maior de intenções de voto do que a percentagem dos votos que acabam por receber. A segunda é que o mesmo tenderá a suceder com grandes partidos (definidos como obtendo pelo menos 20 por cento) numa dada eleição, independentemente da sua condição de favoritos ou não.

*H9: As votações nos partidos favoritos à vitória tenderão a ser sobrestimadas pelas sondagens.*

*H10: As votações nos grandes partidos tenderão a ser sobrestimadas pelas sondagens.*

Na secção anterior, testámos a hipótese de que altos níveis de abstenção fossem responsáveis por maiores discrepâncias gerais entre os resultados de eleições e os resultados de sondagens. Essa hipótese não se confirmou. Contudo, uma hipótese adicional é que diferentes níveis de abstenção afectem de forma diferente a estimação das intenções de voto em diferentes partidos. Ao contrário do que sucede com os pequenos partidos, que tendem a contar muito mais com o apoio de votantes "sinceros" – próximos desses partidos em termos ideológicos e de simpatia partidária – os grandes partidos têm de contar necessariamente com o apoio de votantes "estratégicos", preocupados não apenas com a expressão das suas primeiras preferências mas também com as consequências da sua acção para quem sai vitorioso na eleição. Contudo, eleições de elevada abstenção são, quase por definição, vistas pelos eleitores como menos importantes e decisivas. Nesses casos serão menores os incentivos para o voto estratégico, criando assim a possibilidade de uma particular desmobilização entre os apoiantes dos grandes partidos anteriormente captados por sondagens. Por outras palavras, na medida em que desvios entre intenções e resultados eleitorais decorram da inquirição de indivíduos que acabam por não ir às urnas, esperamos que os grandes partidos sejam os mais prejudicados pela desmobilização. Pelo contrário, nessas eleições com elevada abstenção, devemos esperar que os

pequenos partidos tenham um melhor desempenho eleitoral do que aquilo que as sondagens lhes atribuem.

*H11: Os grandes partidos deverão ser particularmente sobrestimados (e os pequenos partidos subestimados) pelas sondagens em eleições com elevada abstenção.*

Finalmente, devemos esperar que diferentes tipos de sondagens causem diferentes tipos de enviesamentos na estimação das intenções de voto para diferentes partidos. Indivíduos que vivem em domicílios com telefone fixo cujo número se encontra em listas telefónicas tendem a ter características diferentes daqueles que não dispõem desse serviço. Estudos mostram que, em comparação com a generalidade da população, os primeiros tendem a ser mais velhos, mais oriundos de zonas urbanas, e com maiores níveis de rendimento e instrução (Brick et al. 1995; Keeter 1995). Com a difusão do uso do telefone móvel e o aumento da percentagem da população que dispõe apenas deste último poder-se-ia pensar que os problemas de cobertura pelo telefone fixo teriam mudado de natureza, fazendo com a população excluída de uma base de sondagem limitada a números de telefone fixos pudesse ter-se tornado mais heterogénea. De facto, isso pode ter sucedido em relação à idade: comparando a população com telefone fixo com a que dispõe apenas de telefone móvel, a segunda é consideravelmente mais jovem sugerindo que o enviesamento das sondagens telefónicas em relação à população em geral no sentido de cobrir a população mais velha poderá ter sido anulado ou mesmo invertido. Contudo, o enviesamento no sentido de cobrir desproporcionalmente a população com os mais altos níveis de instrução não parece ter sido "cancelado" pelo fenómeno *cell-only* em Portugal: entre indivíduos com telefone fixo ou apenas com telemóvel, não há diferenças claras em termos educacionais, especialmente nos níveis mais baixos e mais elevados de qualificações académicas (Vicente e Reis 2009). Em Portugal, apesar da relação entre variáveis sócio-demográficas e o comportamento de voto não ser particularmente forte, há um padrão recorrente na relação entre a instrução e voto - uma correlação positiva entre a instrução e o voto em partidos de direita. Por outras palavras, ao incluir uma proporção maior de indivíduos com maiores qualificações académicas nas suas amostras do que as sondagens presenciais, as telefónicas tenderão a sobrestimar *ceteris paribus*, por comparação com as sondagens presenciais, as intenções de voto nos partidos da direita.

*H12: As intenções de voto nos partidos de direita tenderão a ser sobrestimados em sondagens telefónicas.*

O quadro 4 mostra os resultados de regressões lineares através das quais testamos as nossas hipóteses. A nossa variável dependente é $B_{ps}$. Por outras palavras, as nossas observações consistem nas intenções de voto captadas nas sondagens para cada um dos cinco principais partidos (mais as coligações PSD/CDS-PP) subtraídas dos resultados eleitorais obtidos por cada uma dessas listas, nos termos explicados na anterior secção 2. Assim, se incluirmos apenas as sondagens para as quais temos informação sobre o modo de inquirição (telefónico ou presencial) ficamos com 1214 observações.

**Quadro 3: Os determinantes de $B_{ps}$**

| | |
|---|---|
| Partido favorito à vitória | 2,47 (0,51)*** |
| Grande partido (>= 20% dos votos) | -5,03 (1,08)*** |
| Sondagem telefónica | -0,76 (0,30)** |
| Taxa de abstenção na eleição | -0,08 (0,01)*** |
| Partido de direita | -1,77 (0,41)*** |
| Grande partido*Abstenção | 0,14 (0,02)*** |
| Direita*Telefónica | 1,47 (0,49)*** |
| Constante | 3,25 (0,53)*** |
| N | 1214 |
| $R^2$ | 0,13 |

*p<0,10; **p<0,05; ***p<0,01 (bilateral); erros-padrão robustos entre parêntesis

A hipótese 9 é confirmada: partidos favoritos à vitória tendem a ter valores de intenções de voto mais elevados por comparação aos seus resultados eleitorais que partidos não favoritos à vitória na eleição. Ilustrando mais claramente com exemplos numéricos, o modelo prevê que, para um partido favorito à vitória, mantendo as restantes variáveis categóricas constantes nos seus valores modais e a abstenção constante no seu valor médio, as intenções de voto captadas para um partido favorito à vitória estejam 1,6 pontos acima do resultado eleitoral que acaba por obter, ao passo que, para os restantes partidos, se prevê uma subestimação de 0,9 pontos percentuais.

No que respeita às hipóteses 10 e 11, a introdução no modelo de termos de interacção significa que se prevê que qualquer sobrestimação ou subestimação de grandes ou pequenos partidos dependerá dos valores da abstenção. Os resultados confirmam essa ideia. O gráfico 2 mostra os valores previstos para $B_{ps}$ contingentes dos valores da abstenção e da dimensão dos partidos, mantendo as restantes variáveis categóricas constantes nos seus valores modais:

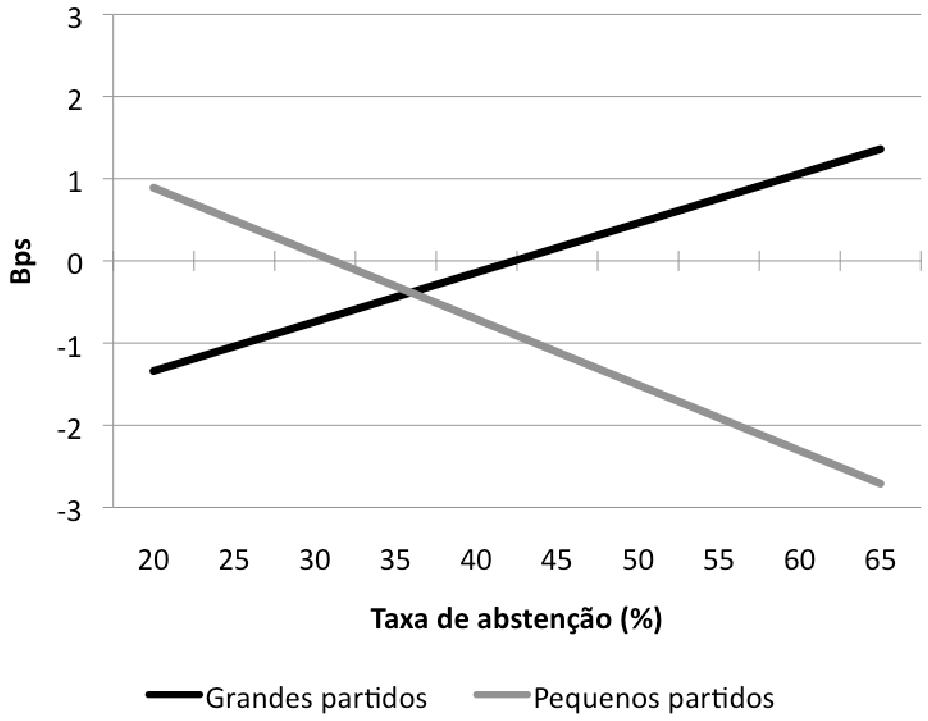



**Gráfico 2: O efeito da dimensão dos partidos condicionado pela abstenção**

O gráfico 2 revela que a tendência das sondagens para sobrestimarem ou subestimarem os grandes ou os pequenos partidos numa dada eleição depende dos valores da abstenção. Quanto mais alta a abstenção, maior a tendência para que os resultados eleitorais dos pequenos partidos acabem por ser melhores do que aquilo que as sondagens lhes atribuíam. Para os grandes partidos, a tendência é a oposta: quanto mais alta a abstenção, mais as sondagens aparecem como uma sobrestimação daquela que acaba por ser a sua votação. Finalmente, a hipótese 12 também é confirmada pela nossa análise. Os resultados do quadro 4 mostram que, em geral, os partidos de direita tendem a obter melhores resultados eleitorais do que as sondagens lhes atribuem, mas que esse efeito é quase neutralizado quando as sondagens em causa são conduzidas pelo telefone.

Em suma, este conjunto de resultados ajuda a explicar o padrão detectado no gráfico 1. Em eleições com abstenção relativamente elevada, como foram a maioria das eleições sob análise (dois terços das consideradas tiveram valores de abstenção acima dos 35%), os grandes partidos – na maior parte dos casos, o PS e o PSD – tendem a ser sobrestimados pelas sondagens. Se a isto adicionarmos o facto de o

PSD e (especialmente) o PS terem sido dados como favoritos à vitória em cerca de dois terços do total das eleições consideradas, o fenómeno da sobrestimação de ambos os partidos nas sondagens e correspondente subestimação da CDU e do CDS-PP torna-se mais compreensível. De notar, contudo, uma tendência adicional para que os partidos de direita – PSD, CDS-PP e suas coligações pré-eleitorais – sejam comparativamente subestimados pelas sondagens presenciais, que constituíram cerca de 30% do total de estudos considerados.

## 6. Conclusão

Ao contrário do que parece ser uma expectativa arreigada na opinião pública, as condições para que as sondagens obtenham distribuições das intenções de voto iguais àqueles que vêm a ser os resultados das eleições não podem ser satisfeitas, seja por limitações inerentes ao método (uso de amostras e medição de intenções e não de comportamentos), seja por características intrínsecas do objecto de estudo (instabilidade de intenções de voto), seja ainda por limitações práticas que, podendo ser combatidas, não podem ser completamente eliminadas (erros sistemáticos de amostragem e de medição). Contudo, é possível estudar os factores que fazem com que as discrepâncias entre sondagens e resultados eleitorais sejam maiores ou menores, ou que fazem com que os resultados eleitorais de determinados partidos acabem por ser maiores ou menores do que as intenções de voto que as sondagens lhes atribuem. É o que fazemos neste artigo.

Os resultados apontam claramente para quatro ideias gerais. A primeira é que o momento em que cada sondagem mede as intenções de voto para uma dada eleição tem consequências para a relação entre os seus resultados e os resultados eleitorais. O padrão que detectámos – uma relação não monotónica entre a distância em relação à eleição e o desvio absoluto médio – é, a nosso ver, particularmente interessante, sugerindo que, apesar das últimas sondagens feitas antes da eleição fornecerem as melhores aproximações àqueles que vêem a ser os seus resultados, há aspectos do comportamento eleitoral que já se encontram razoavelmente fixados e que podem ser correctamente estimados a uma distância considerável da eleição.

A segunda é que a "precisão" das sondagens depende em grande medida de características das eleições e dos partidos, que determinam padrões de sobrestimação e subestimação das intenções de voto em cada partido que são razoavelmente previsíveis. O nível da abstenção, a importância da eleição, e o favoritismo, dimensão e posicionamento ideológico dos partidos são factores relevantes a este nível.

A terceira ideia é que, apesar da discrepância inevitável entre sondagens e eleições, parte dela parece também depender de factores que não estão completamente fora do alcance de quem conduz as sondagens. Sondagens feitas por empresas com maior experiência no mercado e em contextos de maior competição entre empresas têm a tendência para gerarem resultados menos discrepantes com as eleições, sendo também visível, nalguns casos, a existência de *house effects*, ou seja, empresas cujas sondagens tendem a estar sistematicamente mais distantes dos resultados eleitorais. Contudo, alguns dos restantes factores que estão igualmente sob o controlo de quem conduz sondagens – dimensão da amostra, modo de inquirição – parecem não estar relacionados com as nossas variáveis dependentes. Na verdade, essa ausência de relação pode ocultar o facto de que as medidas de que dispomos sobre as características das sondagens são pouco finas, devido a práticas de reporte que têm sido francamente insuficientes. As recentes alterações neste domínio, que implicam a publicitação das fichas técnicas de cada sondagem no site da Entidade Reguladora para a Comunicação Social, podem trazer efeitos benéficos a este nível.

Finalmente, a quarta ideia é a existência de uma tendência secular no sentido de se verificar, ceteris paribus, uma cada vez maior discrepância entre sondagens e resultados eleitorais. A diminuição das taxas de resposta e da cobertura do telefone fixo podem estar por detrás deste fenómeno, revelando os importantes desafios que as empresas enfrentam no sentido de continuarem a fornecer informação precisa sobre as atitudes e intenções comportamentais do eleitorado.

**Referências**


Ajzen, I. (1991). The theory of planned behavior. *Organizational Behaviour and Human Decision Processes*, Vol. 50, 179-211.

Ajzen, I. (2005). *Attitudes, personality and behavior.* McGraw-Hill, New York.

Armitage, C. J. & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology*, Vol. 40, 471-499.

Asher, H. (1992). Polling and the public: What every citizen should know. Congressional Quarterly Press, Washington.

Brick, M., Waksberg, J., Kulp, D. & Starer, A. (1995). Bias in list-assisted Telephone Samples. *Public Opinion Quarterly*, Vol. 99:2, 218-235.

Crespi, I. (1988). Pre-election polling: Sources of accuracy and error. Russell Sage Foundation, New York.

Crewe,I. (1993). A nation of liars? Opinion polls and the 1992 election. *Parliamentary Affairs*, Vol. 45, 475-495.

DeSart, J. A. & Holbrook, T. M. (2003). Statewide Trial-Heat Polls and the 2000 Presidential Election: A Forecast Model. *Social Science Quarterly*, Vol. 84:3, 561–573.

Fishbein, M. & Ajzen, I. (1981). On construct of validity: A critique of Minard and Cohen's Paper. *Journal of Experimental Social Psychology*, Vol. 17, 340-350.

Gelman, A. & King, G. (1993). Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?. *British Journal of Political Science*, Vol. 23, 409-451.

Gilljam, M. & Granberg, D. (1993). Should we take don't know for an answer?. *Public Opinion Quarterly*, Vol. 57:3, 348-357.

Goodin, R. & Rice, J. M. (2009). Waking Up in the Poll Booth. *Perspectives on Politics*, Vol. 7:4, 901-910.

Holbrook, A., Green, M. & Krosnick, J. A. (2003). Telephone vs. Face-to-face interviewing of National Probability Samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, Vol. 67:1, 79-125.

Keeter, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, Vol. 59:2, 196-217.

Leve, J. & Shipman, J. (2009). *A new "interval" measure of election poll accuracy*. Artigo apresentado no encontro anual da American Association for Public Opinion Research, Miami Beach, Florida.

Magalhães, P. (2005). Pre-Election Polls in Portugal: Accuracy, Bias, and Sources of Error, 1991-2004. *International Journal of Public Opinion Research*, 17:4, 399-421.

Magalhães, P. & Moreira, D. (2007). As sondagens pré-eleitorais nas autárquicas de 2005. *Comunicação & Cultura*, Vol. 3, 157-173.

Magalhães, P. (2008). Redes sociais e participação eleitoral em Portugal. *Análise Social*, Vol. 43, 473-504.

Martin, E. A., Traugott, M. & Kennedy, Courtney (2005) A review an proposal for a new measure of poll accuracy. *Public Opinion Quarterly*, Vol. 69:3, 342-369.

Mitofsky, W. J. (1998). Was 1996 a worse year for polls than 1948?. *Public Opinion Quarterly*, Vol. 62, 230-249.

Miller, P. V (2002). *The authority and limitations of polls.* In Manza, J., Cook, F. L. & Page, B. I., eds., Navigating Public Opinion: Polls, policy and the future of American Democracy (221-231), Oxford University Press, New York.

Moon, N. (1999). *Opinion polls: History, theory, and practice.* Manchester University Press, Manchester.

Mosteller, F., Hyman, H., McCarthy, P. J., Marks, E. S. & Truman, D. B. (1949). *The pre-election polls of 1948.* Social Science Research Council, New York.

Noelle-Neumann, E. (1993). *The Spiral of Silence: Public Opinion – Our Social Skin*. Chicago University Press, Chicago.

Pereira, M. M. (2011). 25 anos de sondagens para as eleições presidenciais portuguesas: estudo de desempenho. Working-paper, ICS.

Vicente, P. & Reis, E. (2009). The mobile-only population in Portugal and its impact in a dual frame telephone survey. *Survey Research Methods*, Vol. 3:2, 105-111.

Visser, P. S., Krosnick, J. A., Marquette, J. & Curtin, M. (2000). *Improving Election Forecasting: Allocation of Undecided Respondents, Identification of Likely Voters, and Response Order Effects.* In Lavrakas, P. & Traugott, M., eds., Elecion Polls, the News Media and Democracy. Chatham House, New York.

# Erros Não Amostrais — Uma Floresta de Enganos

Sandra Aleixo [1], *sandra.aleixo@dec.isel.ipl.pt*
Maria de Fátima Brilhante [2], *fbrilhante@uac.pt*
Maria Fernanda Diamantino [3], *mfdiamantino@fc.ul.pt*
Sandra Mendonça [4], *smendonca@uma.pt*
Dinis Pestana [3], *dinis.pestana@fc.ul.pt*

*CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa*
[1] *Instituto Politécnico de Lisboa, ISEL, ADM*
[2] *Universidade dos Açores, Departamento de Matemática*
[3] *Universidade de Lisboa, Faculdade de Ciências, DEIO*
[4] *Universidade da Madeira, Centro de Ciências Exactas e da Engenharia*

"*Public agencies are very keen on amassing statistics — they collect them, add them, raise them to the n-th power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the village watchman, who just puts down what he damn well pleases.*"

Sir Josiah Stamp

"*Probably the most serious of all nonobservational errors, however, is* nonresponse."

Scheaffer *et al.* (1996, p. 52)

## 1. Introdução

A Amostragem é actualmente uma teoria coesa e bem desenvolvida. A tal ponto que há quem defenda que se deve proceder ao ajustamento dos censos usando amostragem — veja-se por exemplo na *Statistical Science* **9**, a controvérsia entre os defensores e os opositores do recurso a ajustamento com amostragem na contagem dos cidadãos dos estados dos EUA.

A Estatística desempenha um papel de relevo na transformação de informação em conhecimento; e a qualidade dos resultados depende inevitavelmente da qualidade da informação recolhida. O papel da Amostragem, nesse contexto, é relevante, nomeadamente controlando o erro amostral. Mais precisamente, nas diversas estratégias de obtenção de amostras aleatórias é conhecido o tamanho $n$ que a amostra deve ter para garantir a precisão pretendida, i.e. o número de unidades amostrais necessárias para a diferença entre o verdadeiro valor do parâmetro de interesse as correspondentes estimativas — que naturalmente variam de amostra para amostra — não se afastarem mais do que uma quantidade $B$, com um grau de confiança pretendido. Há no entanto erros que de um modo geral só podem ser parcialmente controlados, e que nada têm que ver com a variabilidade inerente dos estimadores baseados em amostras aleatórias.

Em geral a expressão "erro amostral" designa o erro decorrente da variabilidade das estimativas de amostra para amostra, e a expressão "erros alheios à amostragem" todos os outros erros que

ocorrem na recolha de informação por amostragem: erros de especificação, erros de cobertura, não-resposta, erros nas respostas, e erros de processamento.

Esta classificação dos erros alheios à amostragem é a adoptada por Tanur (2011), e simplifica a classificação de Deming (1944), que arrolou "13 factores que afectam a utilidade das sondagens". Este trabalho de Deming corresponde a um progresso notável, uma vez que a perspectiva mais matematizada da Estatística tende a esquecer que a implementação não se processa num mundo perfeito que se vergue à tirania das hipóteses dos teoremas — Neyman (1934), por exemplo, assume ingenuamente que os dados de amostragem estão isentos de erros não amostrais. A partir do trabalho de Deming (1944) deixou de ser possível assumir que as sondagens usavam, por definição, dados sem erros, e o erro total (Biemer, 2010, é uma panorâmica interessante) passou a ser uma área de investigação importante. Mosteller (1978), com a sua habitual argúcia e profundidade de visão, é incontornável.

Os **erros de especificação** ocorrem na fase de planeamento, e correspondem na generalidade a adoptar metodologias de recolha de informação que não são os adequados para a obtenção dos dados necessários para atingir os objectivos propostos. Por exemplo, não colocar algumas questões pertinentes — por exemplo, num estudo sobre sucesso escolar, não inquirir sobre a qualidade da alimentação ou sobre as horas de sono. Formular perguntas ambíguas (num inquérito encomendado pelo Mi(ni)stério da Educação pedia-se que se classificasse na famosa escala de 1 a 10 qual era a importância de Deus na vida do inquirido) é mais propício a confundir o nosso entendimento da realidade do que a esclarecê-lo. Colocar directamente questões embaraçosas, em vez de usar aleatorização das perguntas ("*random response*") é um convite à mentira ou recusa a responder. Perguntas tendenciosas, que influenciam as respostas dos inquiridos, são um erro de especificação comum, uma vez que as camadas sociais mais fragilizadas tentam responder por forma a causar boa impressão. Num estudo recente de Oliveira (2011) sobre hábitos alimentares de estudantes prepararam-se dois inquéritos, num deles as perguntas tinham sido viciadas com uma introdução sobre o que é uma refeição equilibrada e quais os nutrientes necessários para o cérebro ter um bom desempenho — como se esperava, os resultados dos dois inquéritos levam a concluir que há heterogeneidade de duas populações que na realidade são mesma (os formulários tinham sido intercalados por forma a em cada turma se entregarem alternadamente uma e outra redacção).

Os **erros de cobertura** mais comuns são por defeito: parte das unidades da população alvo estão ausentes da base de amostragem que é usada. Um exemplo simples: um inquérito telefónico que só use telefones fixos deixaria actualmente muitos habitantes de Portugal fora da base de amostragem. Por outro lado, podem ocorrer também erros de cobertura por excesso, por exemplo quando os inquiridos não deveriam pertencer, de facto, à base de amostragem — por exemplo, quando o entrevistador aproveita, para ser pago, as respostas de um menor numa sondagem sobre intenções de voto. Note-se que a escolha da base de amostragem pode ter uma influência grande em erros de interpretação sobre a estratégia amostral que está a ser usada. Por exemplo se se usar amostragem aleatória simples sem reposição na selecção de números de telemóvel, os inquiridos estão a ser seleccionados afinal de acordo com um plano amostral em que a probabilidade de selecção varia com o número de telemóveis de cada potencial inquirido, e deve-se na estimação do total, por exemplo, ter o cuidado de usar a teoria de Hansen and Hurwitz (1943).

A **não-resposta** é um problema incontornável em amostragem — Scheaffer *et al.* (1996), que não resistimos a citar textualmente, não hesitam em atribuir-lhe protagonismo entre os erros alheios à amostragem.

Há sempre quem não esteja acessível, não queira responder a algumas das perguntas ou à totalidade de um questionário, e se a resposta ao questionário é feita de forma não presencial, pedindo por exemplo que se devolva pelo correio (usando um sobrescrito fornecido, com selo pago), uma larga fracção dos inquiridos não responde. Apesar de tentar incluir-se incentivos diversos, a

taxa de resposta muitas vezes não atinge os 10% (Bourque and Fielder, 2003, p. 16). Diversos livros justamente célebres em amostragem (Scheaffer *et al.*, 1996); Barnett, 2002; Singh, 2003, para citar apenas alguns) dedicam algumas páginas ou mesmo capítulos inteiros a não-resposta e como acomodá-la, e é interessante observar as estimativas cada vez mais pessimistas sobre a taxa de não-resposta com a banalização das sondagens. Bethlehem (2011) considera que em geral excede 50%, e muitos estudos de casos mostram que pode rondar os 90%, e exceder mesmo esse valor no caso de não se estimular o retorno de respostas, ou o inquérito ser de alguma forma incómodo, ou simplesmente maçador.

Barnett (2002) dá indicações úteis sobre como contornar o viés devido a não-resposta usando pós-estratificação. Mas esse método admite que se sabe quem não respondeu. Ora nos muito difundidos inquéritos usando as facilidades da Web nem a taxa de resposta é conhecida! O interessante verbete de Manfreda *et al.* (2011) levanta questões pertinentes, e possivelmente a comunidade estatística, e nomeadamente a *IASS — International Association of Survey Statisticians*, deveriam não só investir mais na investigação da qualidade das sondagens usando a *Web* e alertar os utilizadores deste tipo de sondagens do viés que elas podem ocasionar. Interrogamo-nos, por exemplo, sobre o valor das conclusões do *Estudo da Satisfação e Motivação dos Académicos no Ensino Superior Português (ESMAESP) - PTDC/ESC/67784/2006* a que respondemos recentemente, pois é tipicamente uma questão em que os grupos de respondentes e de não-respondentes podem diferir muito susbtancialmente entre si.

Singh (2003) tem um tratamento em profundidade do problema da não-resposta, e do tratamento de dados omissos, que é também uma consequência da não-resposta. É a nossa recomendação para um primeiro contacto com o importante tema de imputação de dados omissos (inclusive imputação múltipla), um tema de investigação em que muito se tem evoluído mas que continua a ser um desafio a nível de estatísticas oficiais, apesar de a liberalização progressiva do recurso a dados administrativos ter sido nos últimos anos um paliativo no tratamento dessa questão tão problemática.

O essencial do presente trabalho centra-se na questão da não-resposta, pelo que neste ponto não avançamos mais a discussão.

Os **erros nas respostas** devem-se a múltiplos factores, nomeadamente a serem colocadas perguntas ambíguas, ou embaraçosas, ou a que o inquirido não sabe responder, ou que acumulam estes problemas. Por exemplo, no Censo 2001 o INE foi forçado a incluir uma pergunta sobre o número de deficientes que viviam na habitação — deficiência é um conceito vago para a generalidade dos falantes, com interpretações muito diversas. Houve pressão para se incluir uma questão sobre quantos homossexuais viviam na habitação, que não resultou, mas se a questão tivesse sido incluída, o facto de o chefe de família considerar a questão incómoda — ou simplesmente ignorar as determinações dos membros do agregado familiar, iria provavelmente resultar numa subestimativa da proporção verdadeira. Os inquéritos muito longos, e/ou de resposta obrigatória, levam muitas vezes os inquiridos a dar respostas fáceis falsas. No estudo de prevalência de doenças, os inquiridos podem ignorar que são portadores, ou terem vergonha de admitir que têm a doença em causa, se por exemplo puder ser adquirida por transmissão sexual, ou falta de higiene.

Os **erros de processamento** podem ocorrer em diversas fases do estudo. Inclusivamente, podem começar logo na fase de notação da informação — conhecemos directamente o caso de um agente recenseador que decidiu responder por todos os vizinhos sem os interrogar, por "saber o que eles iam responder" (a tradução adequada é: por gastar menos tempo e ter menos trabalho). A frase de Sir Josiah Stamp que citamos na abertura é uma versão aligeirada de outra que não conseguimos localizar, em que se chama a atenção para a possibilidade de todos os inquéritos terem sido respondidos pelo barbeiro da vila, que é também presidente da câmara, e se considera um conhecedor exímio dos seus eleitores.

## 2. Não-Resposta e Rarefação

A determinação do tamanho $n$ que uma amostra deve ter para produzir estimativas do parâmetro de interesse com a precisão desejada é, porventura, o mais relevante contributo da Amostragem para o avanço da Ciênca. Apenas para recordar os resultados nas estratégias amostrais mais simples:

A determinação da dimensão da amostra necessária para obter a precisão que desejamos quando estimamos parâmetros populacionais é um tema chave na teoria da amostragem e suas aplicações como ferramenta metodológica nas ciências experimentais. Por exemplo, se o nosso objectivo é estimar a média populacional a partir da média amostral, de modo a que a amplitude do intervalo de confiança $(1-\alpha)\times 100\%$ seja limitado por $B$, a dimensão da amostra $n = n_{_G}$ é o menor inteiro maior do que

$$\frac{4z^{2}_{1-\frac{\alpha}{2}}\sigma^{2}}{B^{2}}$$

no caso da amostragem com reposição (independência), e

$$\frac{\nu}{1+\frac{(\nu-1)B^{2}}{4z^{2}_{1-\frac{\alpha}{2}}\sigma^{2}}}$$

na amostragem aleatória simples sem reposição (isto é, permutabilidade em vez de independência) a partir de uma população finita de dimensão $\nu$.

Anotamos, de passagem, que o uso de quantis gaussianos na segunda daquelas expressões, é justificada por um importante desenvolvimento do Teorema Limite Central, para o caso de permutabilidade, que se deve a Erdös and Rényi (1959); veja-se a demonstração em fontes documentais mais acessíveis, por exemplo Rényi (2007) ou Pestana e Velosa (2010).

Na prática, o $\sigma^{2}$ desconhecido é substituído por uma estimativa $s^{2}$, e o uso de quantis gaussianos $z_{1-\frac{\alpha}{2}}$ é justificado pelo teorema limite central clássico no caso de independência, e sua extensão para parcelas permutáveis (Erdös and Rényi, 1959), quando a amostragem é feita sem reposição a partir de populações finitas. No entanto, em muitas situações de amostragem, algumas das unidades seleccionadas para a amostra aleatória acabam por originar não-resposta.

No entanto, aquele notável conjunto de resultados, de que os arrolados não passam de exemplos correspondentes às estratégias amostrais mais elementares, parte do pressuposto de que "tudo vai correr bem", isto é que fixado esse $n$ se podem seleccionar $n$ unidades amostrais ao acaso no rol da população. Ora na prática é muito comum uma parte das unidades amostrais escolhidas "escaparem" da amostra, por exemplo recusando-se a serem observadas ou escusando-se a responder a questões que são o objecto da sondagem. A *não resposta* é um obstáculo importante ao uso efectivo da metodologia de amostragem; como todos os obstáculos, leva a novos desenvolvimentos, havendo por exemplo um forte investimento na teoria da imputação múltipla (Singh, 2003, Cap. 12, e em particular pp. 1021–1025).

Há assim uma certa dose de lirismo, quase delírio, na teoria postulando a dimensão que a amostra deve ter para as estimativas terem, com um elevado grau de confiança, a acríbia desejada.

A forma simplificada, porventura em excesso, de contornar o problema de uma taxa de não-resposta elevada, que quase fatalmente vai acarretar viés na estimação dos parâmetros que se pretende conhecer com um grau de precisão pré-determinado, é multiplicar $n$ pelo inverso da estimativa taxa de resposta. De facto, a fracção de não-respostas pode ser bastante elevada, e em inquéritos com resposta a devolver por correio, por exemplo, o questionário é enviado a um

vasto número de indivíduos — uma regra prática *ad hoc* é $\frac{n_{_G}}{\tilde{p}}$, onde $\tilde{p}$ é a percentagem prevista de formulários devolvidos, usualmente uma estimativa grosseira baseada em estudos similares e populações alvo —, uma vez que a experiência acumulada mostra que apenas uma pequena percentagem $p$ deles devolverá os formulários.

Outra possibilidade é recalcular a dimensão da amostra tendo em linha de conta a estimativa da taxa de resposta. Inspirando-nos em Aleixo *et al.* (2007) e nos desenvolvimentos em Diamantino (2008), vamos basear-nos no facto de a rarefação de Rényi (1956) e a filtragem geométrica de Kovalenko (1965) e Kozubowsky (1994) serem assintoticamente equivalentes (no sentido em que levam à mesma lei limite, mas não deixe de se observar que uma limitação severa a esta abordagem é não estar provado que as velocidades de convergência para essa lei limite são iguais nos dois procedimentos) para propor metodologias mais consistentes. Uma parte substancial dos resultados decorre do uso do método delta (Pestana e Velosa, 2010, pp. 1046–1048), exaustivamente tratado em Chandra (1999) ou DasGupta (2008), de que expomos apenas os rudimentos. No que se refere ao uso de rarefação, anote-se de passagem que para além de taxas de não-resposta muito elevadas, há muito interesse no conhecimento de populações que elas mesmas são raras (Kalton and Anderson, 1986).

Mais precisamente, vamos considerar o caso da rarefação (filtragem aleatória) em que, cada unidade incluída na amostra na etapa de planeamento permanece nela com probabilidade $p$, ou sai dela com probabilidade $1-p$, independentemente de qualquer outra.

Adiante investigamos os resultados obtidos ao usar uma amostra com dimensão aleatória $N \frown BinomialNegativa(n_{_G}, p)$, em vez da regra *ad hoc* de dimensão $\frac{n_{_G}}{\tilde{p}}$. Para valores muito pequenos de $p$, o limite desse processo de filtragem é um processo de rarefação de Rényi (1956) da amostra inicial, e no ponto de vista dos resultados de Kovalenko (1965) e de Kozubovsky (1994), o processo de rarefação de Rényi é equivalente a parar aleatoriamente a soma de variáveis aleatórias i.i.d., com subordinador independente $V \frown$ Geométrica$(p)$.

A distribuição assintótica de $T = \sum_{k=1}^{V} X_{_k}$, assumindo a existência da variância da distribuição parente, é Laplace generalizada e, em particular, Exponencial quando as parcelas são positivas. Este pode ser um resultado muito útil quando se faz amostragem de acontecimentos raros.

Não deixamos de referir que os resultados propostos levam a maior variabilidade dos estimadores, pois quando se adopta um ponto de vista hierárquico no cálculo da variância passa a haver duas parcelas a considerar, o valor médio da variância condicional e a variância do valor médio condicional (Pestana e Velosa, 2010, p. 951–952). No entanto, defendemos que a aleatorização é o caminho natural para se obter resultados cada vez mais realistas, e isso leva naturalmente a uma observação em que, devido à escala, a variabilidade é mais "visível", tal como numa laranja à distância do nosso braço são visíveis rugosidades que a uma distância de alguns metros parecem não existir. A maior variabilidade, neste contexto, deve ser encarada como um progresso.

### O método delta

O método delta é, em traços muito gerais, a truncatura da expansão em série de Taylor de uma função com vista à obtenção de aproximações, nomeadamente, para os momentos de uma estatística de interesse. A prática habitual consiste em truncar a expansão após o primeira derivada. No caso de esta se anular no ponto em que é efectuada a expansão recorre-se ao termo correspondente à segunda derivada.

Suponhamos que em vez de um parâmetro $\theta$ desejamos estimar uma função desse parâmetro — por exemplo, queremos estimar $\frac{1}{\mu}$ em vez de estimar o valor médio populacional $\mu$. Mas o inverso de um variável aleatória com valor médio finito até pode não ter valor médio. Porém em situações regulares o recurso à expansão da função de interesse em série de Taylor pode proporcionar aproximações interessantes, truncando no termo linear ou no termo de segunda ordem. Vejamos:

Sejam $\boldsymbol{T} = (T_1, \ldots, T_n)$ e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, onde $\theta_k = \mathbb{E}(T_k)$, $k = 1, \ldots, n$. Seja $g(\boldsymbol{T})$ um estimador de um parâmetro que nos interessa, sendo $g$ uma função diferenciável, e denote-se

$$g'_k(\boldsymbol{\theta}) = \frac{\partial}{\partial t_k} g(\boldsymbol{t}) \mid_{\substack{t_1=\theta_1 \\ \cdots \\ t_n=\theta_n}} .$$

A expansão em série de Taylor de $g$ em torno de $\boldsymbol{\theta}$, de primeira ordem, é

$$g(\boldsymbol{t}) = g(\boldsymbol{\theta}) + \sum_{k=1}^{n} g'_k(\boldsymbol{\theta})\,(t_k - \theta_k) + R_1.$$

Tomando valores médios,

$$\mathbb{E}\,[g(\boldsymbol{T})] \approx g(\boldsymbol{\theta}),$$

pois $\sum_{k=1}^{n} g'_k(\boldsymbol{\theta})\,\mathbb{E}\,(T_k - \theta_k) = 0$. Por outro lado, no que refere a variância do estimador,

$$\operatorname{var}[g(\boldsymbol{T})] \approx \mathbb{E}\left[\left(g(\boldsymbol{T}) - g(\boldsymbol{\theta})\right)^2\right] \approx \mathbb{E}\left[\left(\sum_{k=1}^{n} g'_k(\boldsymbol{\theta})\,(T_k - \theta_k)\right)^2\right]$$

obtendo-se a aproximação

$$\operatorname{var}[g(\boldsymbol{T})] \approx \sum_{k=1}^{n} \left[g'_k(\boldsymbol{\theta})\right]^2 \operatorname{var}(T_k) + 2\sum_{i>j} g'_i(\boldsymbol{\theta})\, g'_j(\boldsymbol{\theta}) \operatorname{cov}\left(T_i, T_j\right).$$

Pestana e Velosa (2010, pp. 1046–1048) enunciam os resultados fundamentais; Chandra (1999) e DasGupta (2008) contêm uma exposição mais circunstanciada.

**Dimensão da amostra para estimação da média com o grau de precisão desejado**

Considerem-se as duas situações seguintes:

1. $(X_1, \ldots, X_n)$ é uma amostra aleatória de dimensão $n$, onde os $X_i$ independentes são tais que $X_i \overset{\mathrm{d}}{=} X$, com $\mathbb{E}(X) = \mu$ e $\operatorname{var}(X) = \sigma^2$. Nesta situação, podemos usar $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ para estimar $\mu$ com um limite de erro ("*error bound*") $B$, com confiança $(1-\alpha)\times 100\%$, usando uma amostra de dimensão $n_G$, o menor inteiro maior do que $\dfrac{4z^2_{1-\frac{\alpha}{2}}\sigma^2}{B^2}$.

2. $(X_1, \ldots, X_\nu)$ é uma população finita com média $\mu = \frac{1}{\nu}\sum_{k=1}^{\nu} X_k$ e variância $\sigma^2 = \frac{1}{\nu-1}\sum_{k=1}^{\nu}(X_k - \mu)^2$. A amostragem aleatória simples sem reposição garante que todas as

$\binom{\nu}{n}$ amostras de dimensão $n$ são equiprováveis. Note que, nesta situação, os $X_k$ já não são independentes, mas a sua dependência mútua é fraca, e o teorema limite central para variáveis aleatórias permutáveis pode ser usado para determinar a dimensão da amostra necessária para obter a precisão que desejamos: $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ estima $\mu$ com um limite de erro padrão $B$, com confiança $(1-\alpha)\times 100\%$, usando uma amostra de dimensão $n_G$, o menor inteiro maior do que $\dfrac{\nu}{1+\frac{(\nu-1)B^2}{4z^2_{1-\frac{\alpha}{2}}\sigma^2}}$. (Como não pode surgir confusão, usamos o mesmo símbolo $n_G$ quer no caso independente, quer no caso permutável).

Assuma-se, no entanto, que sabemos que a amostra será sujeita a uma filtragem-$p$, isto é, cada $X_k$ será efectivamente observado com probabilidade $p$, independentemente de cada um dos outros. Precisamos, portanto, de uma amostra maior de dimensão $N$, de modo a que a amostra filtrada tenha aproximadamente dimensão $n_G$. No que se segue iremos comparar os resultados usando uma amostra aleatória de dimensão $N \frown BinomialNegativa(n_G, p)$ com os resultados obtidos usando dimensão determinística $\dfrac{n_G}{p}$.

Observe que se $(Y_1,\ldots,Y_N)$ é uma amostra de $Y_k$ independentes tais que $Y_k \overset{\mathrm{d}}{=} Y \frown Bernoulli(p)$, independente de $(X_1,\ldots,X_N)$, a amostra $(Z_1,\ldots,Z_{n^*})$ onde os $Z_k$ são os $X_kY_k$ não nulos, é uma amostra filtrada-$p$, e $T = \sum_{i=1}^{n^*} Z_i = \sum_{k=1}^{N} X_kY_k$. Observe que $\mathbb{E}(X_kY_k) = p\mu$ e que $\mathrm{var}(X_kY_k) = p(\sigma^2 + (1-p)\mu^2)$.

**Dimensão da amostra determinística usando a regra *ad hoc* — estimador $\tilde{\mu}_1 = \dfrac{T}{n_G}$**

Se usarmos $N = \frac{n_G}{p}$, o valor esperado e a variância do estimador $\tilde{\mu}_1 = \frac{T}{n_G}$ são:

1. No caso da amostragem independente,

$$\mathbb{E}(\tilde{\mu}_1) = \mu \qquad \text{e} \qquad \mathrm{var}(\tilde{\mu}_1) = \frac{\sigma^2 + (1-p)\mu^2}{n_G}.$$

Portanto, a precisão da estimativa será muito pior do que a esperada sempre que $\mu \gg 0$.

2. Na amostragem aleatória simples sem repetição a partir de uma população com dimensão $\nu$, com a correcção da dimensão da amostra finita para a variância, obtemos resultados semelhantes:

$$\mathbb{E}(\tilde{\mu}_1) = \mu \qquad \text{e} \qquad \mathrm{var}(\tilde{\mu}_1) = \frac{\nu - N}{\nu - 1}\,\frac{\sigma^2 + (1-p)\mu^2}{n_G},$$

e, portanto, há um valor $\dfrac{\nu - N}{\nu - 1}\,\dfrac{(1-p)\mu^2}{n_G}$ a mais na variância do estimador, quando comparado com a situação de não filtragem.

**Dimensão da amostra aleatória Binomial Negativa — estimador $\tilde{\mu}_2 = \dfrac{T}{pN}$**

Seja $N \frown BinomialNegativa(n_G, p)$, $\mathbb{E}(N) = \dfrac{n_G}{p}$, $\text{var}(N) = \frac{n_G(1-p)}{p^2}$, e $\mathbb{E}(N^2) = \dfrac{n_G(n_G+1-p)}{p^2}$, e consideremos o estimador $\tilde{\mu}_2 = \dfrac{T}{pN}$.

Usando a expansão de Taylor linearmente truncada,

$$\tilde{\mu}_2 = \frac{T}{pN} \approx \frac{\mathbb{E}(T)}{p\mathbb{E}(N)} + \frac{1}{p\mathbb{E}(N)}\left[T - \mathbb{E}(T)\right] - \frac{\mathbb{E}(T)}{p[\mathbb{E}(N)]^2}\left[N - \mathbb{E}(N)\right].$$

$\mathbb{E}(T) = \mathbb{E}\left[\mathbb{E}(T|N)\right] = \mathbb{E}(Np\mu) = \frac{n_G}{p}\, p\mu = n_G\mu$, e assim $\mathbb{E}(\tilde{\mu}_2) \approx \dfrac{n_G\mu}{p\frac{n_G}{p}} = \mu$. No que respeita à variância do estimador $\tilde{\mu}_2$:

1. No caso da amostragem independente, $\text{var}(T|N) = Np(\sigma^2 + (1-p)\mu^2)$, e assim

$$\text{var}(T) = \mathbb{E}\left[\text{var}(T|N)\right] + \text{var}\left[\mathbb{E}(T|N)\right] = n_G(\sigma^2 + 2(1-p)\mu^2).$$

Portanto,

$$\text{var}(\tilde{\mu}_2) \approx \frac{1}{n_G^2} n_G(\sigma^2 + 2(1-p)\mu^2) + \left(\frac{n_G\mu}{p\left(\frac{n_G}{p}\right)^2}\right)^2 \frac{n_G(1-p)}{p^2} = \frac{\sigma^2 + 3(1-p)\mu^2}{n_G}.$$

2. Na população finita, com dimensão $\nu$, temos o factor de correcção para populações finitas, $\text{var}(T|N) = \frac{\nu-N}{\nu-1}\, Np(\sigma^2 + (1-p)\mu^2)$, e assim

$$\text{var}(T) = n_G\left[\frac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right]\sigma^2 + n_G\left[1 + \frac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right](1-p)\mu^2.$$

Portanto

$$\text{var}(\tilde{\mu}_2) \approx \frac{\left[\frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right]\sigma^2 + \left[2 + \frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right](1-p)\mu^2}{n_G}.$$

Em consequência, $\tilde{\mu}_2$ é menos eficiente do que $\tilde{\mu}_1$. Observe que o valor esperado do denominador de $\tilde{\mu}_2$ é $n_G$ e, portanto, estamos a dividir uma soma mais variável $T$ (com um número aleatório de parcelas) por um valor próximo do mesmo $n_G$ que usámos no denominador de $\tilde{\mu}_1$, e assim este resultado faz sentido.

**Dimensão da amostra aleatória Binomial Negativa — estimador $\tilde{\mu}_3 = \dfrac{T}{W}$**

Uma abordagem mais sofisticada seria contar o número de parcelas não nulas, $W = \sum_{k=1}^{N} Y_k$. Sendo $W|N \frown Binomial(N, p)$, para dividir a soma $T = \sum_{k=1}^{N} X_k Y_k$, isto é, para usar o estimador $\tilde{\mu}_3 = \dfrac{T}{W}$.

No entanto, neste cenário $\mathbb{E}(W) = n_{_G}$, $\mathrm{var}(W) = 2n_{_G}(1-p)$ e a variância de $\dfrac{T}{W} \approx \mu + \dfrac{1}{n_{_G}}(T - n_{_G}\mu) - \dfrac{\mu}{n_{_G}}(W - n_{_G})$ é

1. No esquema de amostragem independente

$$\mathrm{var}(\tilde{\mu}_{_3}) \approx \frac{1}{n_{_G}^{^2}} n_{_G}(\sigma^{^2} + 2(1-p)\mu^{^2}) + \left(\frac{\mu}{n_{_G}}\right)^{^2} 2n_{_G}(1-p) = \frac{\sigma^{^2} + 4(1-p)\mu^{^2}}{n_{_G}}.$$

2. Na população finita de dimensão $\nu$, a variância de $\tilde{\mu}_{_3}$ quando a amostragem é sem reposição é

$$\mathrm{var}(\tilde{\mu}_{_3}) \approx \frac{\left[\dfrac{(\nu+1)p - (n_{_G}+1)}{p(\nu-1)}\right]\sigma^{^2} + \left[3 + \dfrac{(\nu+1)p - (n_{_G}+1)}{p(\nu-1)}\right](1-p)\mu^{^2}}{n_{_G}}.$$

Portanto, quanto mais variabilidade introduzimos, menor eficiência obtemos do estimador. Parece que a única forma de alcançar os nossos objectivos seria usar um esquema de amostragem inversa desajeitado, continuando a amostragem até a dimensão da amostra atingir $n_{_G}$.

Na secção que se segue tentaremos uma abordagem alternativa, quando o parâmetro de filtragem $p$ está próximo de zero. Esta filtragem radical foi designada "rarefação" por Rényi (1956).

**Filtragem Geométrica**

Seja $Y = \sum_{k=1}^{V} X_{_k}$, onde os $X_{_k}$ variáveis aleatórias independentes tais que as parcelas $X_{_k} \stackrel{\mathrm{d}}{=} X \geq 0$, $k = 1, 2, \ldots$, são independentes da variável subordinadora $V \frown$ Geométrica$(p_{_n})$, e $\mathbb{E}(X) = \mu$. Consequentemente a média da soma aleatoriamente interrompida geométrica $Y$ é $\delta = \dfrac{\mu}{p_{_n}}$. A função característica de $Y$ é $\varphi_{_Y}(t) = \mathcal{G}_{_V}(\varphi_{_X}(t))$, onde $\mathcal{G}_{_V}$ é a função geradora de probabilidade de $V$. Então

$$\varphi_{_Y}(p_{_n}t) = \frac{1}{1 + \dfrac{1 - \varphi_{_X}(p_{_n}t)}{p_{_n}\varphi_{_X}(p_{_n}t)}} = \frac{1}{1 + \dfrac{1 - \varphi_{_X}(p_{_n}t)}{p_{_n}t}\dfrac{t}{\varphi_{_X}(p_{_n}t)}}.$$

Como $\dfrac{1 - \varphi_{_X}(p_{_n}t)}{p_{_n}t} \xrightarrow[p_{_n}\to 0]{} -\varphi'_{_X}(0) = -\mathrm{i}\mu$ e $\varphi_{_X}(p_{_n}t) \xrightarrow[p_{_n}\to 0]{} \varphi_{_X}(0) = 1$, o limite do lado direito em cima é $\frac{1}{1-\mathrm{i}\mu t}$ e consequentemente $\varphi_{_Y}(t) = \dfrac{1}{1 - \mathrm{i}\delta t}$, que é a função característica de uma variável aleatória exponencial com média $\delta$.

Assim, o limite de um processo geometricamente rarefeito com parcelas positivas com valor esperado finito é exponencial. Este resultado assintótico para o processo de rarefação foi inicialmente descoberto por Rényi (1956); Kovalenko (1965) estabeleceu que as transformadas de Laplace de variáveis aleatórias positivas que são estáveis com respeito à rarefação elementar são da forma $L(s) = \dfrac{1}{1 + cs^{\delta}}$, $c > 0$, $\delta \in (0, 1]$, o caso $\delta = 1$ — isto é, limite exponencial — correspondendo a variância finita. Os resultados de Kovalenko mostram que isto coincide com a classe

de somas aleatoriamente interrompidas $\sum_{k=0}^{V} X_k$ com parcelas positivas i.i.d., independentes da variável subordinadora $V \frown$ Geométrica$(p)$. A teoria geral das somas aleatoriamente interrompidas geométricas (Kozubowski, 1994) conduz a resultados semelhantes para a função característica de parcelas cujo suporte não é necessariamente positivo, e em particular a lei limite simétrica para somas geométricas de parcelas independentes de variância finita é a distribuição Laplace.

Não conhecemos qualquer investigação sistemática sobre até onde a rarefação deve ir de modo a que o resultado assintótico possa ser tomado como uma boa aproximação. A exponencial é estável relativamente à filtragem-$p$, ou seja, a exponencial filtrada-$p$ é ainda exponencial, qualquer que seja o valor de $p \in (0, 1]$, no entanto, isto é uma situação excepcional.

Assuma-se agora que $Y = \sum_{k=1}^{V} X_k \overset{\circ}{\frown} Exponencial(\delta)$; de

$$W = \frac{Y}{V} \approx \mu + p(Y - \delta) - p^2\delta\left(V - \frac{1}{p}\right)$$

obtemos que $\mathbb{E}(W) \approx \mu$ e $\text{var}(W) \approx (2 - p)\mu^2$.

Sejam $(W_1, \ldots, W_n)$ réplicas independentes de $W$, $\tilde{\mu}_4 = \overline{W} = \frac{1}{n}\sum_{k=1}^{n} W_k$, para as quais o teorema do limite central se verifica. Como a variância de $\tilde{\mu}_4$ é $\dfrac{(2-p)\mu^2}{n}$, se pretendermos que o intervalo de confiança a $(1-\alpha)\times 100\%$ seja limitado por $B$, devemos tomar uma amostra de dimensão $n_E$, o menor inteiro maior do que $\dfrac{4z^2_{1-\frac{\alpha}{2}}(2-p)\mu^2}{B^2}$.

## 3. Breve estudo computacional, apoiado em populações GLE

Como para $\beta > -1$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}|x|^{\frac{2}{1+\beta}}\right\} \mathrm{d}x = 2^{\frac{\beta+1}{2}}(\beta+1)\Gamma\left(\frac{\beta+1}{2}\right) = 2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right),$$

a função

$$f(x \mid \beta,\, \lambda,\, \delta) = \frac{1}{2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right)\delta} \exp\left\{-\frac{1}{2}\left|\frac{x-\lambda}{\delta}\right|^{\frac{2}{1+\beta}}\right\} \mathrm{I}_{\mathbb{R}}(x)$$

é a função densidade de probabilidade de uma variável aleatória $X_{\beta,\,\lambda,\,\delta}$ para qualquer $\beta > -1$, $\lambda \in \mathbb{R}$ e $\delta > 0$. Temos assim uma família parametrizada a qual contém em particular as variáveis aleatórias Gaussiana$(\beta = 0)$ e Laplace $(\beta = 1)$, pelo que adoptámos a sigla *GLE*, de *Gaussian–Laplace Extended family.* Denotamos $X_{\beta,\,0,\,1} = X_\beta$.

Como

$$\mathbb{E}\left(X_\beta^{2k}\right) = 2^{k(\beta+1)}\,\frac{\Gamma\left(k(\beta+1)+\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)},$$

o valor médio $\mu$ de $X_{\beta,\,\lambda,\,\delta}$ é $\mu = \lambda$, a variância $\sigma^2$ é

$$\sigma^2 = \frac{2^{\beta+1}\Gamma\left(3\,\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)}\,\delta^2,$$

e a curtose $\gamma_2$ é

$$\gamma_2 = \frac{\Gamma\left(\frac{\beta+1}{2}\right)\,\Gamma\left(5\,\frac{\beta+1}{2}\right)}{\left[\Gamma\left(3\,\frac{\beta+1}{2}\right)\right]^2} - 3\,,$$

a qual aumenta para $\infty$ com $\beta$, cf. Figura 1.

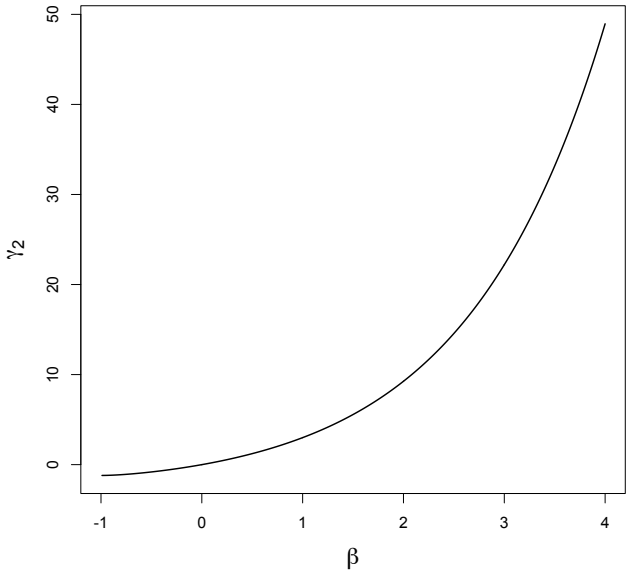



Figura 1: $\gamma_2$ em função de $\beta$

Para $\beta \approx -1$ variável aleatória $X_{\beta,\,\lambda,\,\delta}$ tem uma curtose muito baixa (por exemplo, para $\beta = -0.999$ a curtose é $-1.1999$, não muito distante do limite inferior $-2$ que a curtose pode atingir).

Assim esta família parametrizada, da qual as variáveis aleatórias Gaussiana($\beta = 0$) e Laplace ($\beta = 1$) são casos especiais, abarca uma vasta variedade de caudas leves e pesadas; a curtose desempenha um papel importante em estudos sobre a velocidade de convergência através do limite central, ver Barndorff-Nielsen and Cox (1989). Esta é a principal razão que nos conduz ao uso desta família numa avaliação preliminar dos benefícios de determinar a dimensão da amostra considerando que, sob filtragem, a lei limite é Laplace.

Para o caso de populações com suporte positivo, usamos a família de funções densidade de probabilidade de $W_{\beta,\lambda,\delta} = \lambda + \delta W_\beta$, $\beta > 0$, com

$$f_{W_\beta}(x) = \frac{\exp\left(-x^\beta\right)}{\Gamma\left(1+\frac{1}{\beta}\right)}\;\mathrm{I}_{(0,\infty)}(x).$$

Observe que $W_1 \frown \mathit{Exponencial}(1)$ e $\sqrt{2}W_2$ é a variável aleatória gaussiana dobrada ("*folded Gaussian*"), que o valor médio de $W_{\beta,\lambda,\delta}$ é

$$\mu = \lambda + \delta\,\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} = \lambda + \frac{2^{\frac{2}{\beta}-1}\delta}{\sqrt{\pi}}\,\Gamma\left(\frac{2+\beta}{2\beta}\right),$$

e a variância é $\sigma^2 = \left[\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} - \left(\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right)^2\right]\delta^2$.

Seja $Y_\alpha \frown Gama(\alpha, 1)$; então a função densidade de probabilidade de $V_\alpha = Y_\alpha^\alpha$ é $f_{V_\alpha}(x) = \frac{e^{-x^{\frac{1}{\alpha}}}}{\Gamma(\alpha+1)} I_{(0,\infty)}(x)$, e portanto a variável aleatória $W_\beta$ é

$$W_\beta = V_{\frac{1}{\beta}} = Y_{\frac{1}{\beta}}^{\frac{1}{\beta}}.$$

Por outro lado, com $B$ uma variável aleatória Bernoulli com suporte $\{-1, 1\}$, independente de $V_\alpha$, a função densidade de probabilidade de $T_\alpha = 2^\alpha B V_\alpha$ é $f_{T_\alpha}(x) = \frac{e^{-\frac{1}{2}|x|^{\frac{1}{\alpha}}}}{2^{\alpha+1}\Gamma(\alpha+1)} I_{\mathbb{R}}(x)$, e consequentemente a variável aleatória $X_{\beta,\lambda,\delta}$ é

$$X_{\beta,\lambda,\delta} = \lambda + 2^{\frac{1+\beta}{2}} \delta B Y_{\frac{1+\beta}{2}}^{\frac{1+\beta}{2}}.$$

Assim a geração de números pseudo-aleatórios das populações que escolhemos investigar — e que será mais detalhada nos Apêndices — pode usar os métodos eficientes de geração de números aleatórios gama. A qualidade das populações geradas foi verificada comparando os momentos empíricos e populacionais de menor ordem, uma estimativa de máxima verosimilhança $\hat{\beta}$ com o verdadeiro $\beta$, e o teste de ajustamento de Kolmogorov-Smirnov, com resultados altamente satisfatórios em todos os casos.

Para esta fase da investigação gerámos populações finitas de dimensão $\nu = 5000(5000)20000$ elementos $x_k$ de $X_\beta$, para $\beta = -0.75(0.25)1.5(0.5)3$, e de $W_\beta$, para $\beta = 0.25(0.25)1.5(0.5)6$; em cada caso calculámos a verdadeira média populacional gerada $\mu^* = \frac{1}{\nu}\sum_{k=1}^{\nu} x_k$ e a variância $\sigma^{*2} = \frac{1}{\nu-1}\sum_{k=1}^{\nu}(x_k - \mu^*)^2$.

## Métodos, critérios e conclusões

Com o objectivo de investigar a proximidade das somas filtradas$-p$, atrás definidas, com o limite Exponencial , comparámos a média e o desvio padrão empíricos de $\sum_{k=1}^{V} W_{\beta,k}$, onde os independentes $W_{\beta,k} \stackrel{\mathrm{d}}{=} W_\beta$, tal como foram definidos na Secção 3, são independentes de $V \frown$Geométrica$(p)$. Esta é uma avaliação muito preliminar, apenas uma indicação grosseira de quanto longe estamos da situação ideal.

Como esperado, a aproximação deteriora-se com o afastamento de $\beta$ a partir de 1, no intervalo $\beta \in (0.25, 0.95) \cup (1.1, 4)$ precisamos de $p \approx 0.01$ para obter resultados razoáveis.

Como a rarefação de Rényi e a geo-estabilidade de Kovalenko são dois modos de obter as mesmas leis assintóticas, os resultados foram verificados com investigação semelhante de amostras aleatórias provenientes de $\sum W_{\beta,k} Y_k$, onde $Y_k \frown Bernoulli(p)$.

Resultados de simulação para comparação da eficiência de $\tilde{\mu}_4$ com $\tilde{\mu}_1$ estão ainda muito incompletos. Em termos teóricos, pensamos que $\tilde{\mu}_4$ é mais fidedigno, tendo em conta que estamos

a lidar com acontecimentos raros, uma vez que a variância de $\tilde{\mu}_1$ é maior do que a que tem sido usada no cálculo da dimensão da amostra necessária para ter um limite de erro $B$. Mas até agora, não temos evidência de simulação sistemática para apoiar isto.

## Apêndices

### 1. Propriedades estruturais da família GLE com suporte positivo

Para o caso de populações com suporte positivo, usamos a família de funções densidade de probabilidade de

$$W_{\beta,\lambda,\delta} = \lambda + \delta W_{\beta}, \qquad \beta > 0,$$

com

$$f_{W_\beta}(x) = \frac{\exp\left(-x^{\beta}\right)}{\Gamma\left(1+\frac{1}{\beta}\right)}\ \mathrm{I}_{(0,\infty)}(x).$$

Observe que $W_1$ é a variável aleatória Exponencial(1) (ou Laplace dobrada) e $\sqrt{2}W_2$ é a variável aleatória gaussiana dobrada.

O valor médio de $W_{\beta,\lambda,\delta}$ é

$$\mu = \lambda + \delta\,\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} = \lambda + \frac{2^{\frac{2}{\beta}-1}\delta}{\sqrt{\pi}}\,\Gamma\left(\frac{2+\beta}{2\beta}\right),$$

e a variância de $W_{\beta,\lambda,\delta}$ é

$$\sigma^{2} = \left[\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} - \left(\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right)^{2}\right]\delta^{2}.$$

Seja $Y_\alpha \frown Gama(\alpha,1)$; então a função densidade de probabilidade de $V_\alpha = Y_\alpha^{\alpha}$ é $f_{V_\alpha}(x) = \frac{\mathrm{e}^{-x^{\frac{1}{\alpha}}}}{\Gamma(\alpha+1)}\ \mathrm{I}_{(0,\infty)}(x)$, e portanto a variável aleatória $W_\beta$ é

$$W_{\beta} = V_{\frac{1}{\beta}} = Y_{\frac{1}{\beta}}^{\frac{1}{\beta}}.$$

### 2. Propriedades estruturais da família GLE simétrica com suporte real

Como, para $\beta > -1$, se tem

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\,|x|^{\frac{2}{1+\beta}}\right\} dx = 2^{\frac{\beta+1}{2}}(\beta+1)\Gamma\left(\frac{\beta+1}{2}\right) = 2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right)$$

então a função

$$f(x \mid \beta,\, \lambda,\, \delta) = \frac{1}{2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right)\delta}\exp\left\{-\frac{1}{2}\left|\frac{x-\lambda}{\delta}\right|^{\frac{2}{\beta+1}}\right\}\ \mathrm{I}_{\mathbb{R}}(x)$$

é a função densidade de probabilidade de uma variável aleatória $X_{\beta,\,\lambda,\,\delta}$ para qualquer $\beta > -1$, $\lambda \in \mathbb{R}$ e $\delta > 0$. Temos assim uma família parametrizada a qual contém em particular as variáveis aleatórias Gaussiana($\beta = 0$) e Laplace ($\beta = 1$). Denotamos $X_{\beta,\,0,\,1} = X_{\beta}$.

Como

$$\mathbb{E}\left(X_{\beta}^{2k}\right) = 2^{k(\beta+1)}\,\frac{\Gamma\left(k(\beta+1)+\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)},$$

o valor médio $\mu$ de $X_{\beta,\,\lambda,\,\delta}$ é

$$\mu = \lambda,$$

a variância $\sigma^2$ de $X_{\beta,\,\lambda,\,\delta}$ é

$$\sigma^2 = \frac{2^{\beta+1}\Gamma\left(3\,\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)}\,\delta^2,$$

e a curtose $\gamma_2$ de $X_{\beta,\,\lambda,\,\delta}$ é

$$\gamma_2 = \frac{\Gamma\left(\frac{\beta+1}{2}\right)\,\Gamma\left(5\,\frac{\beta+1}{2}\right)}{\left[\Gamma\left(3\,\frac{\beta+1}{2}\right)\right]^2} - 3,$$

a qual aumenta para $\infty$ com $\beta$.

Para $\beta \approx -1$ variável aleatória $X_{\beta,\,\lambda,\,\delta}$ tem uma curtose muito baixa (por exemplo, para $\beta = -0.999$ a curtose é $-1.1999$, não muito distante do limite inferior $-2$ que a curtose pode atingir).

## 3. Geração de números pseudo-aleatórios de populações GLE

Aleixo *et al.* (2010), ao abordarem directamente a geração dos números pseudo-aleatórios das populações com distribuição pertencente à família GLE , experimentaram vários métodos sugeridos em Devroye (1986), Kundu e Gupta (2007), Law e Kelton (1982) e Ross (1997), concluindo que o método mais adequado em cada caso dependia do valor do parâmetro $\beta$ considerado.

- Para $0 < \beta < 1$, método da rejeição utilizando a distribuição $Pareto(\alpha)$, com $\alpha$ óptimo.
- Para $\beta = 1$, método da inversão para gerar exponenciais unitárias.
- Para $1 < \beta < 6$, método da rejeição utilizando a distribuição $Exponencial(1)$.
- Para $\beta \geq 6$, método da rejeição utilizando a distribuição $Weibull(0, \alpha, \gamma)$ (a exponencial unitária deixa de funcionar).

Por exemplo, para $\alpha = 0.7$ e $\gamma = 1.3$, obtêm-se bons resultados para os números pseudo-aleatórios gerados usando o método da rejeição com a Weibull, para $6 \leq \beta \leq 15$. Já para $16 \leq \beta \leq 20$, pode usar-se o método da rejeição com a $Weibull(0, 0.75, 1.7)$. No entanto, não é possível obter uma expressão geral para os parâmetros de escala e forma da Weibull para cada valor de $\beta$. Os valores adequados destes parâmetros dependem do parâmetro $\beta$ considerado.

Mas, tendo em conta as relações existentes entre as variáveis aleatórias com distribuição pertencente à família GLE e variáveis aleatórias com distribuição Gama, a geração de números aleatórios das populações que pretendemos obter, com $\beta \neq 1$, pode ser feita indirectamente, usando os métodos eficientes de geração de números aleatórios gama já bem conhecidos e testados. Assim vamos optar por este processo de geração, para $\beta \neq 1$.

Uma vez que as relações com produtos de potências de gamas independentes são usadas para a geração no caso de suporte positivo, as correspondentes populações simétricas são obtidas multiplicando por uma Bernoulli com $p = 0.5$ e com suporte $\{-1, 1\}$, independente da anteriormente gerada para suporte positivo. No entanto, para $-1 < \beta \leq 0$, estas populações simétricas têm que ser geradas com base na relação existente entre as variáveis aleatórias com distribuição pertencente à família GLE de suporte $\mathbb{R}$ e variáveis aleatórias com distribuição Gama.

Consulte-se Aleixo *et al.* (2010) e as referências aí indicadas para mais detalhes, inclusive estudos de validade das tabelas de pseudo-aleatórios assim gerados, que podem ser obtidas nas páginas de S. Aleixo ou de M. F. Diamantino em `www.ceaul.fc.ul.pt`.

## Bibliografia


Aleixo, S., Brilhante, M. F., Diamantino, F, Mendonça, S., and Pestana, D. (2007). Non-Response and Sample Size, *Bulletin of the International Statistical Institute* **LXII**, 4804–4807.

Aleixo, S., Diamantino, M. F., and Pestana, D. D. (2010). The GLE Distributions Family. Notas e Comunicações do CEAUL, 8/2010.

Barndorff-Nielsen, O., and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, London.

Barnett, V. (2002) *Sample Surveys: Principles and Methods*, 3rd ed., Arnold, London.

Bethlehem, J. (2011). Nonresponse in surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, 982 –983.

Biemer, P. P. (2010) Overview of design issues: total survey error. In Marsden, P., and Wright, J. (eds.) *Handbook of Survey Research*, 2nd ed., Cap. 2, Bingley, United Kingdom.

Bourque, L. B., and Fielder, E. P. (2003). *How to Conduct Self-Administered and Mail Surveys*, Sage Publ., Thousand Oaks.

Chandra, T. K. (1999). *A First Course in Asymptotic Theory of Statistics*, Narosa, New Delhi.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.

Deming, W. E. (1944). On errors in surveys. *Am. Sociol. Rev.* **9**, 359–369.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

Diamantino, M. F. (2008). *Contribuição ao Estudo de Dados em Falta*, dissertação de doutoramento, Universidade de Lisboa.

Erdös, P., and Rényi, A. (1959). On a central limit theorem for samples from a finite population, *Publ. Math. Instit. Hungar. Acad. Sci.* **4**, 49–61.

Hansen, M. M., and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14**, 333–362.

Hansen, M. M., and Hurwitz, W. N. (1946). The problem of nonresponse in sample survys, *J. Am. Statist. Assoc.* **41**, 517–529.

Kalton, G., and Anderson, D. (1986). Sampling rare populations, *J. Royal Statist. Soc.* **A 149**, 65–82.

Kovalenko, I. N. (1965). On a class of limit distributions for rarefied flows of homogeneous events, *Lit. Mat. Sbornik* **5**, 569–573. (*Selected Transl. Math. Statist. and Prob.* **9**, Providence, Rhode Island, 1971, 75– 81.)

Kozubowski, T. J. (1994). Representation and properties of geometric stable laws, *Approximation, Probability, and Related Fields*, Plenum, New York. 321–337.

Kundu, D., and Gupta, R. D. (2007). A convenient way of generating gamma random variables

using generalized exponential distribution, *Comp. Stat. & Data Analysis*, **51**, 2796–2802.

Law, A. M., and Kelton, W. D. (1982). *Simulation, Modeling and Analysis*, Academic Press, New York.

Manfreda, K. L., Berzelak, N., and Vehovar, V. (2011). Nonresponse in Web surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, 984–987.

Mosteller, F. (1978). Errors I: nonsampling errors. In: Kruskal, W. H., and Tanur, J. M. (eds.) *International Encyclopedia of Statistics*. Free Press, New York, 208–229.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Roy. Stat. Soc.* **97**, 558–606.

Oliveira, A. A. (2011). *Problemas Mal Resolvidos em Amostragem — Não–Resposta e Formulação Tendenciosa de Questões*, DEIO, FCUL.

Pestana, D. e Velosa, S. (2010). *Introdução à Probabilidade e à Estatística*, 4ª edição Fundação Gulbenkian, Lisboa.

Rényi, A. (1956). A characterization of the Poisson process, *MTA Mat. Kut. Int. Közl.* **1**, 519–527. (Reeditado em *Selected Papers of Alfréd Rényi*, P. Turán, ed, Akadémiai Kiadó, Budapest, 1976, vol. I, p. 622–628.)

Rényi, A. (2007). *Probability Theory*, Dover, New York.

Ross, S. M. (1997). *Simulation. Statistical Modeling and Decision Science*, 2nd ed., McGraw-Hill, San Diego.

Scheaffer, R. L., Mendenhall III, W., and Ott, R. L (1996) *Elementary Survey Sampling*, 5th ed., Duxbury Press, Belmont.

Singh, S. (2003). *Advanced Sampling Theory with Applications, How Michael 'Selected' Amy*, Kluwer, Dordrecht.

Tanur, J. M. (2011). Nonsampling errors in surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, 988–991.

*Investigação parcialmente financiada por FCT/OE.*

# Sobrevivência no R

Valeska Andreozzi, *valeska.andreozzi@fc.ul.pt*

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

O número de bibliotecas que disponibilizam funções para a análise de dados de sobrevivência no R tem crescido de forma exponencial nos últimos cinco anos. De acordo Allignol e Latouche [1] o R dispõe atualmente de 124 bibliotecas, com as mais variadas funcionalidades sobre análise de sobrevivência. Dos métodos clássicos de análise de sobrevivência, como estimador Kaplan-Meier, modelos paramétricos, modelos de Cox, modelos de vida acelerado, passando pelos modelos de fragilidade, modelos multi-estados (modelos para eventos recorrentes e para risco competitivo) até chegar aos modelos baseados em métodos bayesianos, estão disponíveis (ver Quadro Resumo). Além da vasta oferta na área da modelação, também é de ressaltar o avanço na disponibilização de técnicas para tratar características da análise de sobrevivência, como censura à esquerda, censura intervalar, censura dupla e truncamento.

Quadro Resumo: Algumas funcionalidades do R e suas principais bibliotecas.

| Funcionalidades | Bibliotecas |
|---|---|
| Estimador Kaplan-Meier | `survival; rms; prodlim` |
| Modelo de Cox | `survival; rms; eha; coxphf; coxphw; coxrobust; timereg` |
| Modelo de vida acelerado | `survival; rms; eha; lss` |
| Modelo aditivo | `survival; timereg` |
| Modelos multi-estados | `survival; mvna; etm; mstate;` |
| Modelos de risco competitivo | `Surv2sample; cmprsk; timereg; CompetingRiskFrailty;` |
| Modelos para eventos recorrentes | `survival; rms; frailtypack; survrec` |
| Modelos de fragilidade | `survival; kinship; frailtypack; lmec` |
| Modelo com abordagem bayesiana | `survBayes; bayesSurv; BMA; DPpackage; LDDPsurvival; BayHaz; splinesurv;` |
| *Inverse probability weights* | `ipw` |

Fonte: Allignol e Latouche [1].

Como pode ser visto no Quadro Resumo, a biblioteca `survival` e uma das mais completas bibliotecas do R sobre análise de sobrevivência. A seguir serão ilustradas suas principais funcionalidades através de exemplos disponíveis na página da internet do livro: Análise de Sobrevida: Teoria e Aplicações em Saúde[1] [2].

[1] http://sobrevida.fiocruz.br/

**Exemplo: Transplante de medula óssea**

Estes dados provêm de uma coorte de 96 pacientes submetidos a transplante de medula óssea (TMO) para tratamento de leucemia mielóide crônica, no período de Junho de 1986 a Junho de 1998, no Centro de Transplante de Medula Óssea do Instituto Nacional do Câncer do Rio de Janeiro. O objetivo do estudo foi investigar o efeito de fatores prognósticos para ocorrência de doença do enxerto contra hospedeiro aguda e crónica, da sobrevivência livre de doença e da sobrevivência global.
As covariáveis registadas para cada paciente estão descritas na página da internet do livro[2]. Os dados estão organizados em dois ficheiros. O *tmoclas.dat* tem o formato clássico, com uma linha para cada paciente. O ficheiro *tmopc.csv* está preparado para uma análise de sobrevivência com covariável dependente do tempo havendo repetição de linhas para cada paciente de acordo com o número de mudanças nas covariáveis ao longo do estudo. Os dados podem ser lidos diretamente da internet fornecendo o caminho da página à função `read.table()`.

```
> tmo <- read.table("http://sobrevida.fiocruz.br/dados/tmoclas.dat", header=T, sep=",")
```

Nesse tutorial será utilizado as seguintes variáveis:

- `idade` : idade do paciente na data do transplante
- `status`: 0 = censura, 1 = óbito
- `os`: tempo de sobrevivência decorrido desde o transplante até o óbito
- `deag`: doença enxerto aguda: 0 = não, 1 = sim
- `inicio`: data de entrada do paciente no estudo ou do início da mudança de covariável
- `fim`: data de mudança de covariável ou do fim do estudo

Para que o R assuma que os valores numéricos da variável discreta `deag` sejam categorias, `deag` foi declarada como *factor* e as primeiras 6 linhas da base de dados é listada através da função `head()`.

```
> tmo$deag <- factor(tmo$deag)
> head(tmo)
  id sexo idade status   os plaq tempplaq deag tempdeag decr tempdecr fase
1  1    2    30      0 1000    1        9    0     3527    0     3527    1
2  2    2    38      1   39    0       39    1       28    0       39    1
3  3    1    23      1  434    1       27    1       36    1      268    1
4  4    2     5      1   69    0       69    1       24    0       69    1
5  5    2    15      1  672    1       83    1       22    1      446    1
6  6    1    23      1   98    1        0    1       22    0       98    1
```

Para tornar disponível as funções da biblioteca `survival` é necessário carregá-la na sessão de trabalho:

```
> library(survival)
```

**Estimador Kaplan-Meier**

O tempo até um determinado evento pode ser estudado através da função de sobrevivência descrita em cada tempo, $t$, por $S(t) = Pr(T > t)$. O estimador de Kaplan-Meier estima a função de sobrevivência de forma não paramétrica através do produto das probabilidades de sobrevivência até o tempo $t$.
Na biblioteca `survival` o Kaplan-Meier é estimado através da função `survfit()`. A função `summary()` retorna em detalhe o resultado do estimador.

```
> km.fit <- survfit(Surv(os,status)~1, data=tmo)
> km.fit
Call: survfit.formula(formula = Surv(os, status) ~ 1, data = tmo)
records   n.max n.start  events  median 0.95LCL 0.95UCL
     96      96      96      48     475     370      NA
```

[2] http://sobrevida.fiocruz.br/dados/transplante.html

```
> summary(km.fit)
Call: survfit.formula(formula = Surv(os, status) ~ 1, data = tmo)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   31     96       1    0.990  0.0104        0.969        1.000
   32     95       1    0.979  0.0146        0.951        1.000
   39     93       1    0.969  0.0178        0.934        1.000
…
```

O gráfico de sobrevivência do Kaplan-Meier é feito utilizando a função `plot()`.

```
> plot(km.fit, conf.int=F, xlab="Dias", ylab="S(t)")   #Figura 1a
```

Nos gráficos da Figura 1, as censuras são marcadas ao longo da curva de sobrevivência por um traço vertical.

Para estratificar a sobrevivência por uma variável basta acrescentar esta variável, no exemplo `deag`, na fórmula da função `survfit()`.

```
> km.deag <- survfit(Surv(os,status)~deag, data=tmo)
> plot(km.deag, conf.int=F,xlab="Dias", ylab="S(t)", lty=c(1,2)) #Figura 1b
> legend(x="bottomleft", legend=c("Não","Sim"), lty=c(1,2), title="Deag", bty="n")
> title("Kaplan-Meier",cex=.3)
```

a) b)

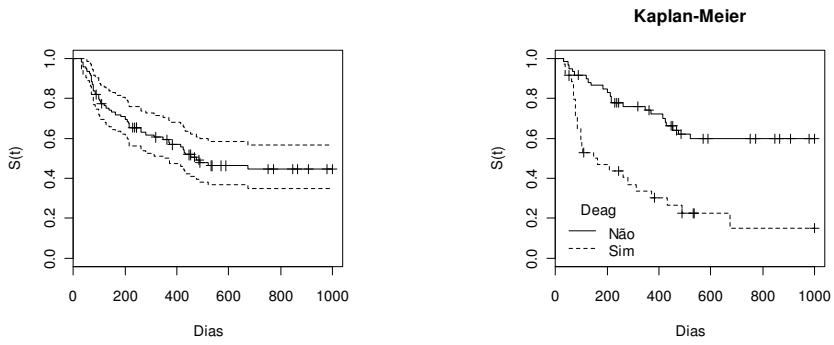



Figura 1: Gráfico da sobrevivência estimada por Kaplan-Meier

O teste de log rank pode ser calculado através da função `survdiff()` para testar se há diferenças entre as curvas de sobrevivência do gráfico da Figura 1b). Para calcular o teste de Peto basta acrescentar o argumento rho=1 na mesma função.

```
> survdiff(Surv(os,status)~deag, data=tmo)   #Teste de Log rank
> survdiff(Surv(os,status)~deag, data=tmo,rho=1)   #Teste de Peto
```

**Modelo Paramétrico**

Os modelos paramétricos assumem que a variável aleatória tempo T segue uma distribuição de probabilidade conhecida. Em geral, as distribuições mais utilizadas são: exponencial, lognormal e Weibull.

A função `survreg()` faz estimação paramétrica da curva de sobrevivência e possui a seguinte sintaxe:

```
 fit <- survreg(formula, data=dados, dist="distribuição")
```

sendo o argumento `dist` a distribuição a ser ajustada: `dist="exp"` ajusta a distribuição exponencial; `dist="weibull"` ajusta distribuição Weibull. O argumento `formula` possui o seguinte formato: `Surv(tempo,status)`. O objeto criado, `fit`, contém o resultado do ajuste da distribuição e uma série de estatísticas de interesse que podem ser vistos com a função `summary()`.

Assumindo que o tempo de sobrevivência (variável `os`) dos dados TMO segue uma distribuição exponencial, tem -se no R:

```
> exp.fit <- survreg(Surv(os,status)~1,data=tmo,dist="exponential")
> summary(exp.fit)
Call:
survreg(formula = Surv(os, status) ~ 1, data = tmo, dist = "exponential")
            Value Std. Error    z p
(Intercept)   6.8      0.144 47.1 0

Scale fixed at 1

Exponential distribution
Loglik(model)= -374.2   Loglik(intercept only)= -374.2
Number of Newton-Raphson Iterations: 5
n= 96
```

Para incluir uma variável no modelo de regressão paramétrico basta acrescentar na fórmula o nome da variável ou variáveis.

```
> exp.fit <- survreg(Surv(os,status)~deag,data=tmo,dist="exponential")
```

A comparação de modelos paramétricos pode ser feita graficamente com a sobreposição das curvas de sobrevivência estimada parametricamente e pelo Kaplan-Meier como mostra a Figura 2.

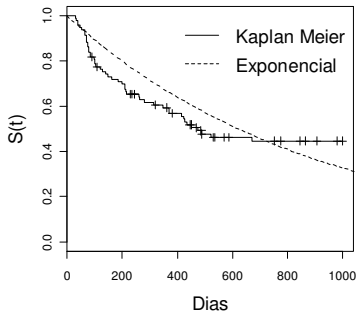



Figura 2: Gráfico da sobrevivência estimada pelo Kaplan-Meier e parametricamente (distribuição exponencial)

```
> plot(km.fit, ylab="S(t)", xlab="Dias",
+ lty=1, lwd=1.5,cex.lab=1.5,
+ conf.int = F)

> curve(exp(-exp(-exp.fit$coef)*x),
+ from=0, to=1200,add=T, lty=2, lwd=1.5)

> legend("topright", legend=c("Kaplan
+ Meier", "Exponencial"), lty=c(1,2),
+ bty="n", lwd=1.5, cex=1.5)
```

**Modelo de Cox**

Modelo de riscos proporcionais de Cox e suas extensões são os modelos mais amplamente utilizados em análise de sobrevivência na área da Saúde. A sintaxe da função para estimar o modelo de Cox no R segue a forma:

```
> modelo <- coxph(formula, data=dados)
```

em que o argumento formula é igual a `Surv(tempo,status)~x1+x2`. Para os dados do TMO, tem-se:

```
> cox.fit <- coxph(Surv(os,status)~ idade + deag, data=tmo)
> summary(cox.fit)
Call:
coxph(formula = Surv(os, status) ~ idade + deag, data = tmo)
  n= 96, number of events= 48

           coef exp(coef)  se(coef)      z Pr(>|z|)
idade -0.004738  0.995274  0.013031 -0.364    0.716
deag1  1.213254  3.364414  0.298652  4.062 4.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
idade    0.9953     1.0047    0.9702     1.021
deag1    3.3644     0.2972    1.8737     6.041
```

```
Rsquare= 0.164   (max possible= 0.984 )
Likelihood ratio test= 17.25  on 2 df,   p=0.0001797
Wald test            = 17.66  on 2 df,   p=0.0001461
Score (logrank) test = 19.81  on 2 df,   p=5.003e-05
```

Novamente, a função `summary()` é utilizada para listar o resultado do ajuste do modelo de Cox. São apresentados os coeficientes, seus erros-padrão e o teste de Wald associado a cada covariável, as razões de risco e os respectivos intervalos de confiança, além de alguns testes estatísticos.
A seleção de covariáveis do modelo de Cox é feita através do teste da razão de verossimilhanças. No R a função `anova(modelo1,modelo2)` é utilizada para tal efeito. Veja o exemplo da comparação do modelo de Cox estimado anteriormente com um modelo que inclui somente a idade dos pacientes.

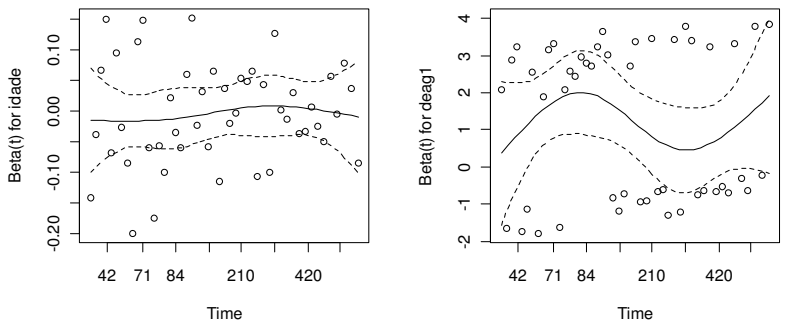
```
> cox.fit2 <- coxph(Surv(os,status) ~ idade, data=tmo)
> anova(cox.fit,cox.fit2)
Analysis of Deviance Table
 Cox model: response is  Surv(os, status)
 Model 1: ~ idade + deag
 Model 2: ~ idade
   loglik  Chisq Df P(>|Chi|)
1 -190.42
2 -198.50 16.168  1 5.797e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Análise de Resíduos**
Os elementos para elaborar uma análise de resíduos dos modelos de Cox estão listados na Tabela 2.

Tabela 2. Elementos para análise de resíduos

| | |
|---|---|
| Avaliar a proporcionalidade de cada variável | Gráficos dos resíduos de Schoenfeld contra o tempo |
| Identificar pontos atípicos | Gráficos dos resíduos martingale ou deviance |
| Estudar a forma funcional da variável | Gráficos dos resíduos martingale do modelo nulo contra covariável |
| Identificar pontos influentes | Gráficos dos resíduos escore contra covariável |

Os gráficos dos resíduos do modelo de Cox ajustado anteriormente (objeto `cox.fit`) encontram-se nas Figuras 3-5 com as respectivas funções do R.

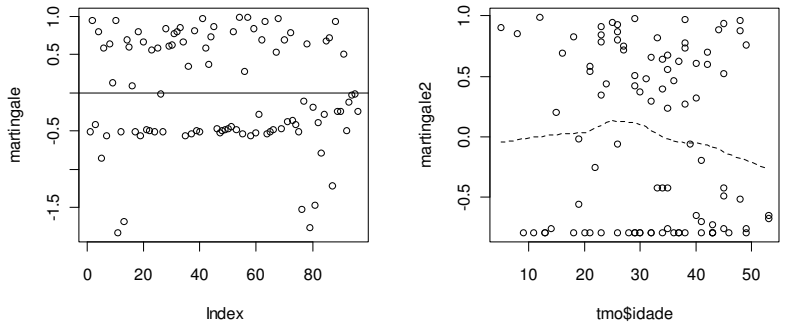



```
> shoenfeld <-
+ cox.zph(cox.fit)

> par(mfrow=c(1,2))

> plot(shoenfeld)
```

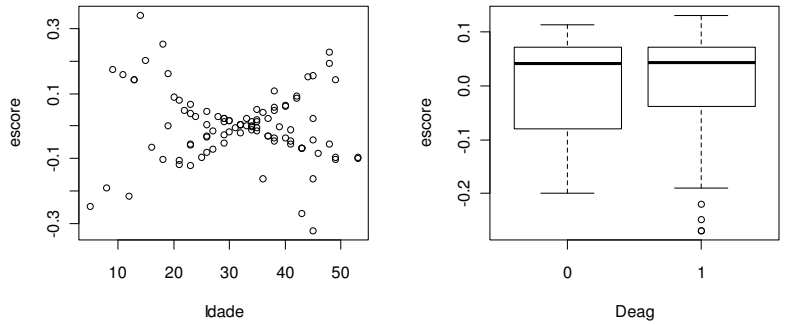
Figura 3. Resíduo de Schoenfeld




```
> resm <- resid(cox.fit,
+ type="martingale")
> plot(resm)
> abline(h=0)
> cox.nulo<-
+ coxph(Surv(os,status)~1,
+ data=tmo)
```

```
> resm2 <- resid(cox.nulo,
+ type="martingale")
> plot(tmo$idade,resm2)
> lines(lowess(tmo$idade,
+ resm2,iter=0),lty=2)
```

Figura 4. Resíduo Martingale




```
> escore <- resid(cox.fit,
+ type="dfbetas")
> par(mfrow=c(1,2))
> plot(tmo$idade,escore[,1],
+xlab="Idade",ylab="escore")
> plot(tmo$deag,
+ escore[,2],xlab="Deag",
+ ylab="escore")
```

Figura 5. Resíduo de Escore

**Modelo de Cox estendido**

Para lidar com covariáveis que mudam no tempo ou com efeitos de covariáveis que não são lineares, é necessário estender os modelos de risco proporcionais de Cox para permitir a inclusão de tais características. No R, o modelo de Cox estendido é estimado utilizando-se a mesma função `coxph()`, sendo o objeto sobrevivência gerado pela função `Surv()` agora formulado através do processo de contagem, ou seja, tem que ser indicado o início e fim de cada período de acompanhamento.

```
> modelo <- coxph(Surv(inicio,fim,status) ~ x1+x2,data=dados)
```

O ficheiro *tmopc.csv* contém informação detalhada da variável `deag`, agora dependente do tempo. A seguir encontra-se o resultado da estimação do modelo Cox considerando uma covariável que muda ao longo do tempo:

```
> tmo2 <- read.table("http://sobrevida.fiocruz.br/dados/tmopc.csv",header=T,sep=";")

> head(tmo2)
  id sexo idade status inicio  fim deag decr recplaq fasegr
1  1    2    31      0      0    9    0    0       0    CP1
2  1    2    31      0      9 3527    0    0       1    CP1
3  2    2    38      0      0   28    0    0       0    CP1
4  2    2    38      1     28   39    1    0       0    CP1
5  3    1    23      0      0   27    0    0       0    CP1
6  3    1    23      0     27   36    0    0       1    CP1
```

```
> cox.fit3 <- coxph(Surv(inicio,fim,status)~idade+deag,data=tmo2)

> summary(cox.fit3)
…
           coef exp(coef)  se(coef)      z Pr(>|z|)
idade -0.008896  0.991143  0.011980 -0.743    0.458
deag   1.124783  3.079547  0.280019  4.017  5.9e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
idade    0.9911     1.0089    0.9681     1.015
deag     3.0795     0.3247    1.7788     5.331
…
```

O resultado do modelo ajustado é listado através da função `summary()`.
Fica para um próximo tutorial os modelos mais complexos de análise de sobrevivência. Enquanto isso, o leitor pode utilizar aa *vignettes* do R que são tutoriais disponíveis pelas próprias bibliotecas. Utilize a função `vignette()` para listar as *vignettes* disponíveis do seu R. A sintaxe para aceder a uma *vignette* segue o seguinte formato:

```
> vignette("nome da vignette", package="nome da biblioteca")
```

Alguns exemplos em sobrevivência são:

```
> vignette("Tutorial", package="mstate") #modelos multi-estados

> vignette("parametric", package="eha") # modelos paramétricos

> vignette("intervals_overview",package="intervals") # Overview da biblioteca intervals
```

**Referências**


[1] Allignol, A and Latouche, A. CRAN Task View: Survival Analysis. Version: 2010-11-30. http://cran.r-project.org/web/views/Survival.html. (visitada em 17 de Fevereiro de 2011)

[2] Carvalho, MS; Andreozzi, VL; Codeço, CT; Barbosa, MTS; Shimakura, SE. 2005. *Análise de Sobrevida*: *Teoria e Aplicações em Saúde*. Editora Fiocruz. Rio de Janeiro.

# Pós - Doc

## Tese de Doutoramento: Estimação Não Paramétrica em Modelos de Regressão de Dados de Contagem com Excesso de Zeros

(Boletim SPE Outono de 2007, p. 51)

José António Santos, *josant.santos@gmail.com*

*ISEGI - Universidade Nova de Lisboa*

O Doutoramento em Matemática e Estatística, pelo Departamento de Matemática do ISA-UTL, que concluí em Abril de 2007, foi um trabalho muito gratificante que se estendeu por cerca de quatro anos. Houve fases de dificuldades, sobretudo no final, que não foram sentidas como dolorosas. Foram sempre tidas como desafios e oportunidades de aprendizagem e de desenvolvimento pessoal e científico. A Professora Manuela Neves estava ali.

Pouco senti com a obtenção do grau de "Doutor". Tenho para comigo que "D(d)outor" não é nome de gente. Lembro-me bem, aquando de uma visita à Faculdade de Estatística da Universidade de Munique, em 2003, o Professor Gerhard Tutz, no nosso primeiro contacto, quando o tratei por "Professor", me dizer: "*Do not Professor me!*".

Importante foi o trabalho realizado no Doutoramento e, sobretudo, o caminho aberto e o trabalho futuro que este prometia. Digamos, o que não foi feito, ou ficou por fazer…

Devo dizer que foram particularmente estimulantes e decisivos os comentários que recebi dos membros do Júri. Foram críticas muito úteis, incentivos à prossecução dos trabalhos e o elogio à apresentação.

Sem o incentivo do Júri do Doutoramento tudo seria mais difícil. O reconhecimento do trabalho catalisou a minha vontade e força de continuar.

Devo dizer que há dois nomes importantes a destacar durante esses tempos e agora, sem descortesia para muitos outros!

A Professora Manuela Neves, que me orientou e com quem tenho a mercê de continuar a trabalhar. Daí retiro grande ganho da sua grande cultura científica, bem como da sua honestidade, rectidão e enorme simpatia. A Professora Manuela Neves tem lutado toda uma vida. Bertolt Brecht dela diria, estou certo, que foi, é e será "imprescindível"!

O Professor António St. Aubyn foi o meu Mestre na Matemática. Com o máximo rigor, dedicação e exigência, infelizmente mal compreendido por alguns alunos que, muitos, mais tarde viriam a reconhecer a bondade destes desideratos, o Professor António St. Aubyn abriu-me as portas da paixão pela matemática e pelos seus mistérios! Quanto lhe estou em dívida!

Feitas as minhas singelas, justas e humildes homenagens (de aluno!) a figuras tão grandes, falarei agora do que ando por aí a fazer.

Diz-se "*no hay camino, se hace camino al andar*". Assim tem sido o meu percurso desde 2007 até agora, em que as coisas vão acontecendo, não raro de forma casuística. Tenho continuado a batalhar na área, de natureza mais teórica, dos modelos de regressão não paramétrica ou semi-paramétrica. Num âmbito mais geral, tenho trabalhado na modelação semi- ou não paramétrica e, em particular, na comparação do desempenho de modelos com diferentes "graus" semi-paramétricos.

Além disso, têm surgido propostas de modelação de dados de natureza espacial e/ou temporal, com casos na área do ambiente e da ecologia.

Outras matérias que têm merecido atenção são a análise de variância a dois factores de natureza forçosamente não paramétrica (quem quer ajudar?, fica o convite!), os métodos de análise de regressão de dados discretos e de dados com censura ou com truncatura, e os modelos de mistura.

Entretanto, veio a decisão recente de sair do ISEGI, para estar disponível para novos desafios. Um novo desafio que apareceu e cresceu, tomando uma dimensão bastante grande e consistente, tem sido a investigação em colaboração com o Departamento de Biologia e a unidade de investigação CESAM da Universidade de Aveiro, desde 2007. Aqui o meu interlocutor tem sido o Professor Carlos Miguez Barroso (é verdade, é excelente trabalhar com biólogos, uma asserção que coloco em termos inferenciais…), de grandes qualidades pessoais, científicas e profissionais e a quem me ligam laços de amizade e já não somente profissionais. O trabalho em bioestatística que se tem desenvolvido tem tido aplicações na biologia marinha, ecotoxicologia e poluição.

O meu desvio para a área da bioestatística tem-se acentuado, agora com uma nova colaboração que surgiu, cresce e se desenvolve com o Serviço de M. F. Reabilitação do Hospital de Curry Cabral.

E mais coisas virão. Como disse Friedrich Nietzsche, "é necessário ter o caos cá dentro para gerar uma estrela".

Com o convite a colaborações frutuosas, os meus melhores cumprimentos a todos,

José António Santos

# Tese de Doutoramento: Distance-Based Methods for Classification and Clustering of Time Series

(Boletim SPE Outono de 2007, p. 52)

Jorge Caiado, *jcaiado@iseg.utl.pt*

*ISEG – Universidade Técnica de Lisboa*

Prezados colegas,

Em resposta ao apelo do Prof. Fernando Rosado, venho partilhar com os sócios individuais e colectivos da SPE e leitores das suas publicações periódicas a minha experiência profissional, após a conclusão do doutoramento, no que se refere às actividades de investigação.

Em Dezembro de 2006 e, decorridos quatro anos de trabalho árduo de investigação e muita transpiração, conclui, no Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa, o doutoramento em Matemática Aplicada à Economia e à Gestão, sob a orientação dos professores Nuno Crato (ISEG) e Daniel Peña (Universidad Carlos III de Madrid). O tema nele desenvolvido centrou-se na classificação e agrupamento de séries temporais (ver Boletim SPE Outono de 2007). Em particular, foram propostas medidas de distância entre séries temporais baseadas na sua estrutura de autocorrelações, densidade espectral e volatilidades. Como aplicações económicas, estes métodos foram utilizados para identificar semelhanças e dissemelhanças entre índices de produção industrial nos Estados Unidos e para classificar e agrupar os principais países desenvolvidos e as acções do índice Dow Jones. A tese encontra-se agora publicada em livro internacional [Caiado, J. (2010), *Classification and Clustering of Time Series*, Lambert Academic Publishing, Saarbrücken (Germany), ISBN 978-3-8383-4181-1], e está disponível para aquisição no portal Amazon.com (http://www.amazon.co.uk).

Uns meses após ter terminado a etapa do doutoramento, surgiu uma oportunidade de dar continuidade ao trabalho de investigação nele realizado. No âmbito do Programa Ciência 2007 da Fundação para a Ciência e Tecnologia (FCT), concorri a uma posição de investigador auxiliar no Centro de Matemática Aplicada à Previsão e Decisão Económica (CEMAPRE) do ISEG. Esta posição consiste num contrato anual em regime de tempo integral e exclusividade, renovável por 5 anos, dependendo de avaliação independente e da aprovação da FCT. O plano de trabalho submetido ao CEMAPRE, assente na investigação teórica e aplicada da temática de classificação de séries temporais económicas e financeiras, foi aceite pelo júri do concurso. Em Outubro de 2008, iniciei assim uma nova etapa de investigação que me levou a suspender a actividade docente no ensino superior politécnico que exerci quase ininterruptamente desde 1995 até 2008.

Neste projecto de investigação, em colaboração com um colega do Programa Ciência 2007, propus-me estudar a estrutura dinâmica e a predictabilidade dos mercados financeiros internacionais. Para o efeito, recolhemos dados dos índices diários bolsistas de 23 mercados desenvolvidos e 23 mercados emergentes no período de 1995 a 2009, construídos pela Morgan Stanley Capital International (MSCI).

No que refere à estrutura dinâmica, começamos por estudar o comportamento dinâmico dos retornos dos índices dos mercados desenvolvidos e emergentes em termos das principais estatísticas univariadas, da dependência a curto prazo, da memória, da heteroscedasticidade condicionada, e dos efeitos assimétricos. Em seguida, para investigar a estrutura subjacente entre aquelas características,

utilizámos uma técnica estatística de análise factorial separadamente para os grupos de mercados desenvolvidos e mercados emergentes. Por último, utilizando a análise factorial para gerar os scores dos factores, procurámos identificar clusters de países e outliers multivariados. Os resultados obtidos sugerem que existem diferenças significativas entre a estrutura de dependência dos retornos dos mercados desenvolvidos e emergentes. Contudo, a estrutura não é constante ao longo dos períodos em análise.

No que concerne à predictabilidade, recorremos a testes de hipóteses de passeio aleatório, baseados em rácios de variâncias, para investigar a interdependência dos mercados accionistas globais em termos da previsibilidade dos retornos de índices, e averiguámos se o padrão de agrupamento dos mercados evoluiu nos anos recentes. Primeiro, examinámos a validade da hipótese de passeio aleatório nos mercados individuais usando testes de rácios de variâncias convencionais, de sinal e de posição. Posteriormente, empregámos técnicas de escalonamento multidimensional e de agrupamento para examinar a interdependência dos rácios de variância entre os mercados accionistas.

Mais recentemente, estamos a investigar a presença de dependências determinísticas não-lineares entre os mercados accionistas internacionais através da análise e quantificação de gráficos de recorrências. Os gráficos de recorrências permitem detectar visualmente padrões determinísticos em sistemas dinâmicos, sendo, no nosso caso, particularmente úteis para averiguar a existência de padrões comuns que reflictam o grau de integração dos mercados financeiros. Por sua vez, a análise de quantificação de recorrências dispõe dos instrumentos para a quantificação das estruturas dos mercados internacionais e a detecção de transições críticas no sistema. Os primeiros resultados obtidos apontam para uma integração mais forte a nível regional do que a nível global dos mercados financeiros, em termos de padrões determinísticos. Os mercados desenvolvidos (como os Estados Unidos e o Japão) apresentam menores níveis de determinismo não-linear do que os mercados emergentes. Este resultado é consistente com o facto dos mercados mais desenvolvidos exibirem, em geral, menores níveis de predictabilidade a curto-prazo.

Esperemos que a investigação continue a produzir resultados interessantes que possam contribuir não só para a produção e divulgação científica mas também para a resolução e interpretação de problemas reais, quer na ciência quer na vida quotidiana.

Saudações académicas e científicas,

Jorge Caiado
Investigador Auxiliar do CEMAPRE/ISEG

# Tese de Doutoramento: Desenvolvimento de distribuições quase-exactas para vários cenários de utilização da estatística Lambda de Wilks

(Boletim SPE Outono de 2007, p. 48)

Luís Miguel Grilo, *lgrilo@ipt.pt*

*Área de Matemática da Escola Superior de Tecnologia*
*do Instituto Politécnico de Tomar*

Caros colegas e amigos,

As provas públicas do meu Doutoramento em Matemática e Estatística decorreram no Instituto Superior de Agronomia da Universidade Técnica de Lisboa, no dia 17 de Julho de 2006. Foi um momento que me envolveu num turbilhão de emoções e sentimentos, dado que correspondeu à concretização de um desafio considerável, sendo que nem sempre foi fácil investigar, leccionar e residir em locais separados por grandes distâncias.

Naquele 2.º Semestre de 2005/2006 leccionava a unidade curricular de Probabilidades e Estatística aos cursos de Engenharia no Instituto Politécnico de Tomar (IPT) e colaborava com o Departamento de Matemática da Universidade de Évora, na leccionação da unidade curricular de Análise de Dados Multivariados às licenciaturas de Geografia e Psicologia. Simultaneamente, era o responsável local pela organização da Conferência Internacional SCRA2006|FIM XIII (Statistics, Combinatorics and Related Areas and XIII Conference of the Forum for Interdisciplinary Mathematics), que decorreu no *Campus* do IPT em Setembro de 2006 e que, para além das interessantes comunicações apresentadas, teve como ponto alto o Doutoramento *Honoris Causa* do Professor Doutor C. R. Rao, atribuído pela Universidade Nova de Lisboa. Assim, foi numa altura relativamente atribulada que surgiu a data e a discussão da minha dissertação e nem o facto de ter dois artigos aceites para publicação em revistas internacionais e de já ter realizado várias comunicações em eventos científicos nacionais e internacionais de Estatística (uma delas nos EUA, onde recebi o prémio - Best Presentation Award) me deixou psicologicamente mais à vontade. As provas públicas de um Doutoramento são, pela sua natureza e formato, um momento diferente em que somos confrontados com interessantes questões e comentários. Recordo, ainda hoje, muitas das palavras proferidas pelos sete membros que constituíram o Júri, em particular, as dos arguentes: Professores Doutores Dinis Duarte Pestana (FC-UL) e João Tiago Mexia (FCT-UNL).

Na minha dissertação foram apresentadas distribuições quase-exactas para a estatística Lambda de Wilks generalizada. Estas distribuições situam-se muito mais próximas (em termos de função característica, função densidade de probabilidade, função distribuição cumulativa, momentos e quantis) da distribuição exacta do que as assimptóticas, nomeadamente, para amostras de dimensão reduzida. Considero que a investigação nesta área foi muito enriquecedora dado que envolveu o estudo de distribuições de alguma complexidade analítica, bem como a sua implementação informática no *software* Mathematica, o que permitiu, não só o cálculo de momentos e quantis quase-exactos, como também a comparação entre distribuições exactas, quase-exactas e assimptóticas, através de duas medidas desenvolvidas com base nos limites de Berry-Esseen. Na sequência da dissertação, tenho continuado a desenvolver investigação e a publicar em co-autoria com o meu, agora, ex-orientador de

Doutoramento (Professor Doutor Carlos Agra Coelho, FCT-UNL). Mais uma vez, com o seu incentivo, amizade e excelente colaboração temos produzido trabalhos na área da Teoria das Distribuições, em Estatística Multivariada, e no âmbito do Centro de Matemática e Aplicações da Universidade Nova de Lisboa, ao qual pertencemos.

Desde 1999 que desenvolvo a actividade de docente na área de Matemática do IPT, onde o facto de ter obtido mais um grau académico, não implicou alteração na minha categoria profissional, dado que permaneço na categoria de Professor Adjunto. Todavia, em Fevereiro de 2007, fui nomeado Director do Centro de Sondagens e Estudos Estatísticos (CSEE) pelo então Presidente do IPT e, enquanto director deste Centro, tenho colaborado e desenvolvido projectos de interesse para o IPT, nomeadamente participo na avaliação interna semestral da Instituição, realizada com base em inquéritos de leitura óptica efectuados aos alunos e docentes; por outro lado, tenho colaborado em actividades de investigação efectuadas pelos docentes do IPT, no contexto da sua formação pós-graduada (Mestrados e Doutoramentos), e em trabalhos desenvolvidos pelos alunos do IPT, principalmente, a nível de projectos finais de 1.º e 2.º ciclos (em 2010 conclui a minha primeira orientação de uma dissertação de Mestrado em Engenharia Química); tenho, ainda, procurado dinamizar e apoiar a realização de eventos (cursos breves, seminários, *workshops*, conferências, etc.) nas diferentes áreas de aplicação da Estatística, em particular, coordeno e lecciono um curso breve em "Análise de Dados com o SPSS", onde têm participado especialistas de diversas áreas do conhecimento que utilizam a Estatística no decurso da sua actividade profissional, bem como alunos e docentes do IPT e de outras Instituições.

Paralelamente à minha actividade no IPT, é um privilégio continuar a participar em reuniões científicas de Estatística e ser membro de Comissões Organizadoras e de Conselhos Científicos de alguns desses eventos. É, ainda, muito interessante desempenhar, pontualmente, a tarefa de *referee* para uma revista internacional de Estatística e ser, frequentemente, contactado para colaborar em estudos que envolvem a análise e o tratamento de dados (actualmente, são muitos os investigadores das ciências sociais e humanas e das ciências médicas que solicitam apoio).

Em 1992, quando comecei a desenvolver aplicações estatísticas em gestão de stocks, no controlo de qualidade, ou a modelar as vendas de papel higiénico e dos lenços de bolso na Renova (Fábrica de Papel do Almonda, S.A.) ou, em 1993, quando, simultaneamente, comecei a leccionar as disciplinas de Estatística e de Métodos Quantitativos no ISLA – Santarém (onde, actualmente, sou membro convidado e secretário eleito do Conselho Científico), estava, obviamente, longe de imaginar os caminhos que a Estatística me permitiria percorrer…

Como outros colegas já escreveram neste espaço, após a conclusão do Doutoramento, as responsabilidades e as solicitações profissionais crescem consideravelmente e se, como escreveu Pitágoras, "com ordem e com tempo encontra-se o segredo de fazer tudo e tudo fazer bem", importa não esquecer que a não verificação de, pelo menos, um daqueles pressupostos pode comprometer seriamente os resultados finais. Efectivamente, por vezes, parece ser necessária alguma arte para compatibilizar o ensino com outras actividades, incluindo a investigação e, obviamente, sem esquecer a vida familiar. Considero ainda que, para investigar nesta ciência transversal, de utilidade inquestionável, que é a Estatística, é conveniente para além de ordem e tempo, alguma persistência e resistência (mental e, até, física!), à semelhança do que acontece nas longas partidas de Xadrez que disputei como jogador federado durante alguns anos ou nas maratonas de BTT que pratico, com a regularidade possível. Não obstante, o mais fácil acesso à informação e o apoio e colaboração de especialistas de grande valia que, actualmente, se encontram afectos a diversos centros de investigação por todo o país, tem permitido a jovens investigadores o desenvolvimento de trabalhos de muita qualidade, o que deixa antever um futuro promissor para a Estatística em Portugal. No meu caso, e se me é permitido, tenho procurado estar disponível para trabalhar e aprender, bem como procuro nortear a minha vida com base no princípio "esteja eu onde estiver, estou apenas a começar" e penso que os resultados alcançados são relativamente positivos e encorajadores.

Para todos um forte abraço e votos de sucessos académicos e profissionais,

Luís Grilo

# Tese de Doutoramento: A Estatística Bayesiana na Identificação Forense Análise e avaliação de vestígios de DNA com redes bayesianas

(Boletim SPE Outono de 2007, p. 50)

Marina Andrade, *Marina.Andrade@iscte.pt*

*ISCTE - Instituto Universitário de Lisboa*

Caros colegas e amigos,

As minhas provas públicas de Doutoramento tiveram lugar no ISCTE – IUL, no dia 16 de Abril de 2007. Foi naturalmente um dia importante, por ser a conclusão de uma fase relativa à minha formação, por ser a conclusão de um objectivo anteriormente proposto, e por ser o início de uma fase de novos desafios e propostas.

Em Agosto do mesmo realizaram-se em Lisboa o XV Congresso Anual da Sociedade Portuguesa de Estatística, no ISCTE – IUL, e a 56th Session of the ISI, e em Copenhaga o 22nd Congress of the ISFG - International Society for Forensic Genetics. A oportunidade de participar nestes eventos, quase imediatamente a seguir à obtenção do grau de Doutor, foi um estímulo determinante para o início desta fase da minha carreira de investigação.

Tenho prosseguido essa carreira, integrada no centro de investigação UNIDE – IUL, trabalhando basicamente nos temas:

- Identificação civil e criminal e estudo de cenários criminais através da análise de perfis de DNA com recurso a redes bayesianas,
- Aplicação de modelos de Filas de Espera ao estudo de problemas financeiros, económicos e sanitários (epidemias),
- Antropologia forense,
- Aplicação da Teoria do Caos à Gestão de Recursos,

em conjunto com a actividade docente no ISCTE – IUL em que tenho leccionado disciplinas no âmbito da Matemática e da Estatística nos primeiro, segundo e terceiro ciclos.

Apresentei dois seminários sobre o primeiro dos temas enumerados acima no DEIO da Faculdade de Ciências da Universidade de Lisboa e no Departamento de Matemática da Universidade do Minho. Ainda sobre o mesmo tema tive a oportunidade de leccionar um curso, como professor visitante, na Universidade de Economia e Informática de Bratislava, República Eslovaca.

Com a divulgação do trabalho científico e a participação, relativamente frequente em reuniões cientificas, proporcionou-se a possibilidade de colaborar na edição de revistas cientificas. Deste modo:

- Sou editora do Portuguese Journal of Quantitative Methods (revista editada pelo Departamento de Métodos Quantitativos do ISCTE - IUL),
- Pertenço ao Advisory Committee da Statistical Review (Journal of the Greek Statistical Association),
- Integro o Advisory and Review Board do International Journal of Academic Research,
- Integro o International Review Board do Journal of Mathematics and Technology,
- Integrei o Scientific Committee da 6th National and International Conference of the Hellenic Society for Systemic Studies (HSSS),
- Fiz revisão cientifica de artigos para as revistas Aplimat – Journal of Applied Mathematics, International Journal of Academic Research and Journal of Mathematics and Technology,

- Fiz revisão cientifica de artigos para as Actas das Conferências Aplimat, HSSS e XV Congresso Anual da SPE.

Participei como arguente em dois júris de Doutoramento no ISCTE – IUL. Participei como Presidente num júri de Mestrado no ISCTE – IUL e como arguente em três júris de Mestrado no ISCEM.
Orientei duas teses de Mestrado no ISCTE – IUL, já concluídas.
Termino apresentando uma lista dos artigos publicados em revistas científicas internacionais:


Ferreira, M. A. M. e Andrade, M. (2011). "Fundaments of theory of Queues", International Journal of Academic Research, 3(1), Forthcoming.

Andrade, M. e Ferreira, M. A. M. (2011). "Some considerations about forensic DNA evidences ", International Journal of Academic Research, 3(1), Forthcoming.

Ferreira, M. A. M. e Andrade, M. (2011). "The M/G/oo Queue Busy Period Distribution Exponentiality", Aplimat- Journal of Applied Mathematics, Forthcoming.

Andrade, M. e Ferreira, M. A. M. (2011). "Paternities search in a very uncommon situation through object-oriented Bayesian networks", Journal of Mathematics and Technology, 2(1), Forthcoming.

Ferreira, M. A. M. e Andrade, M. (2011). "M/G/oo Queue System Transient Behaviour with Time Origin at an Operation Beginning Instant and Occupation", Journal of Mathematics and Technology, 2(1), Forthcoming.

Ferreira, M. A. M. e Andrade, M. (2011). "Management Optimization Problems", International Journal of Academic Research, 3(2). Forthcoming.

Ferreira, M. A. M. e Andrade, M. (2011). "Stochastic Processes in Networks of Queues with Losses: A Review", International Journal of Academic Research, 3(2). Forthcoming.

Andrade, M. e Ferreira, M. A. M. (2010). "Janus Probability two faces in court", Statistical Review-Journal of the Greek Statistical Association, 6(1) Forthcoming.

Ferreira, M. A. M., Andrade, M. e Matos, M. C. (2010). "Separation theorems in Hilbert spaces convex programming", Journal of Mathematics and Technology, 1(5), pp. 20-27.

Filipe, J. A., Ferreira, M. A. M., Coelho, M., Pedro, M. I. e Andrade, M. (2010). "Analysing Fisheries Management through Complexity and Chaos Theories Framework", Journal of Mathematics and Technology, 1(2), pp. 5-12 .

Andrade, M. e Ferreira, M. A. M. (2010). "Solving civil identification cases with DNA profiles databases using Bayesian networks", Journal of Mathematics and Technology, 1(2), pp. 37-40.

Andrade, M., (2010). “A Note on Foundations of Probability”. Journal of Mathematics and Technology (JMT),Volume 1 (1), pp. 96-98.

Ferreira, M. A. M. e Andrade, M. (2010). "Looking to a M/G/00 system occupation through a Ricatti equation", Journal of Mathematics and Technology, 1(2), pp. 58-62.

Ferreira, M. A. M. e Andrade, M. (2010). "M/M/m/m Queue System Transient Behavior", Journal of Mathematics and Technology, 1(1), pp. 49-65.

Andrade, M. e Ferreira, M. A. M. (2010). "Civil Identification Problems with Bayesian Networks using Official DNA Databases", Aplimat- Journal of Applied Mathematics, 3(3), pp. 155-162.

Ferreira, M. A. M. e Andrade, M. (2010). "Algorithm for the calculation of the Laplace-Stieltjes transform of the sojourn time of a costumer in an open network of queues, with a product form equilibrium distribution...", Journal of Mathematics and Technology, 1(4), pp. 31-36.

Andrade, M. e Ferreira, M. A. M. (2010). "Evaluation of Paternities with less usual Data using Bayesian Networks", IEEE Xplore (BMEI 2010 IEEE Catalog Number CFP1093D-PRT), pp. 2475-2477.

Ferreira, M. A. M. e Andrade, M. (2010). "M/G/oo Queue Busy Period Tail", Journal of Mathematics and Technology, 1(3), pp. 11-16.

Ferreira, M. A. M. e Andrade, M. (2010). "M/G/oo System Transient Behavior with Time Origin at the Beginning of a Busy Period Mean and Variance", Aplimat- Journal of Applied Mathematics, 3(3), pp. 213-221.

Andrade, M., Ferreira, M. A. M., Abrantes, D., Pontes, M. L. e Pinheiro, M. F. (2010). "Object-oriented Bayesian Networks in the evaluation of paternities in less usual environments", Journal of Mathematics and Technology, 1(1), pp. 161-164.

Matos, M. C., Ferreira, M. A. M. e Andrade, M. (2010). "Code Form Game", International Journal of Academic Research, 2(1), pp. 135-141.

Andrade, M. e Ferreira, M. A. M. (2009). "Bayesian Networks in Forensic Identification Problems", Aplimat- Journal of Applied Mathematics, 2(3), pp.13-30.

Ferreira, M. A. M., Andrade, M. e Filipe, J. A. (2009). "Networks of Queues with Infinite Servers in Each Node Applied to the Management of a Two Echelons Repair System", China-USA Business Review, 8 (8), pp. 39-45 and 62

Ferreira, M. A. M. e Andrade, M. (2009). "M/G/oo Queue System Parameters for a Particular Collection of Service Time Distributions", AJMCSR-African Journal of Mathematics and Computer Science Research, 2(7), pp. 138-141.

Ferreira, M. A. M. e Andrade, M. (2009). "The Ties Between the M/G/inf Queue System Transient Behaviour and the Busy Period", International Journal of Academic Research, 1(1), pp. 84-92.

Ferreira, M. A. M. e Andrade, M. (2009). "A note on Dawnie Wolfe Steadman, Bradley J. Adams, and Lyle W. Konigsberg, ...", International Journal of Academic Research, 1(2), pp.23-36.

Ferreira, M. A. M., Andrade, M., Filipe, J. A. and Selvarasu, A. (2009). "The Management of a Two Echelons Repair System Using Queueing Networks with Infinite Servers Queues", Annamalai International Journal of Business Studies and Research - AIJBSR, Vol:1(1), pp. 32-37.

Andrade, M. e Ferreira, M. A. M. (2009). "Criminal and Civil Identification with DNA Databases Using Bayesian Networks", International Journal of Security-CSC Journals, 3 (4), pp. 65-74.

Andrade, M., Ferreira, M. A. M., Filipe, J. A. e Coelho, M. (2008). "Paternity Dispute: is it important to be conservative?", Aplimat- Journal of Applied Mathematics, 1(2).

Ferreira, M. A. M., Andrade, M. e Filipe, J. A. (2008). "The Ricatti Equation in the M/G/oo Busy Cycle Study", Journal of Mathematics Statistics and Allied Fields, 2(1).

Andrade, M., Ferreira, M. A. M. e Filipe, J. A. (2008). "Evidence Evaluation in DNA Mixture Traces", Journal of Mathematics Statistics and Allied Fields, 2(2).

Abrantes, D., Pontes, M. L., Pinheiro, M. F., Andrade, M. e Ferreira, M. A. M. (2008). "Towards systematic probabilistic evaluation of parentage casework in forensic genetics: A modest attempt to define a general standardized approach to simple and complex cases", Forensic Science International : Genetics Supplement Series. Elsevier, Vol. 1 (1), pp. 635-637.

Filipe, J. A., Ferreira, M. A. M., Coelho, M. e Andrade, M. (2008). "Anticommons Destroy Value. Portugal`s Aquaculture Case", Aplimat- Journal of Applied Mathematics, 1(2).

# Tese de Doutoramento: Estimation and Testing for Distributions with Light, Heavy and Super-heavy Tails

(Boletim SPE Outono de 2007, p. 49)

Claudia Neves, *claudia.neves@ua.pt*

*CEAUL e Universidade de Aveiro*

Na sucessão de acontecimentos aleatórios, coincidências e acasos com sentido que constituem as nossas vidas, esta breve crónica pós-doc surge em altura de balanço pessoal. Digo isto, quer no sentido mais comum da contabilidade de fracassos e sucessos acessória àqueles momentos de introspecção que às vezes surgem do nada e por pouco menos de nada, quer na acepção da palavra "balanço" como fase preparatória de movimento noutra direcção.

Passo a explicar melhor: em Dezembro de 2006 a minha tese de doutoramento veio ocupar a posição 72 na lista dos doutoramentos realizados no Departamento de Estatística e Investigação Operacional (DEIO) da Faculdade de Ciências da Universidade de Lisboa - disponível em http://www.deio.fc.ul.pt/teses_doutoramento.php. O total é um número impressivo de 99 doutoramentos no DEIO desde 1981. Desde então, e tal como acontece aos docentes da carreira universitária, passei a exercer funções como Professora Auxiliar. No meu caso particular, no Departamento de Matemática da Universidade de Aveiro em regime experimental por cinco anos, segundo as mais recentes alterações ao Estatuto da Carreira Docente Universitária. Naturalmente, num regime experimental, há que experimentar: até ao momento leccionei sete disciplinas distintas. O regime dura até Dezembro próximo e como a vindima só acaba com o lavar dos cestos, há que continuar nesta senda e experimentar! Posso asseverar que, algumas das disciplinas, até experimentei mais do que uma vez, não vá o leitor ficar com a impressão de que estou a falar de breves ensaios pedagógicos em turmas com menos de 20 alunos, como que de ensaios sem grupo de controlo se tratasse. Ou como quem diz deixou de fumar sem nunca ter realmente inalado. Nada disso. São cinco anos a experimentar intensivamente, em obediência à Lei dos Grandes Números e em campo aberto à aplicação do Teorema Limite Central. O balanço final espectável está próximo: muitos fracassos para poucos sucessos. Nada mal para quem estuda acontecimentos raros.

A actividade de investigação decorre, naturalmente, na intermitência do anterior. Em tom de balanço introspectivo, o trabalho de investigação que tenho desenvolvido inscreve-se no domínio da Teoria de Valores Extremos e Aplicações, com especial destaque para métodos assintóticos de inferência estatística envolvendo o índice de valores extremos. Neste contexto as distribuições de cauda super-pesada têm merecido especial destaque. Mais recente é o estudo de métodos em Estatística de Extremos que visam a detecção e concomitante estimação de uma eventual tendência temporal ou espacial. Este estudo, no âmbito do projecto de investigação EXES - Extremos Espaciais, (financiado pela Fundação para a Ciência e Tecnologia), foi orientado para a aplicação mais imediata a níveis elevados de precipitação no Norte da Europa.

*Many people dream of success. To me success can only be achieved through repeated failure and introspection. In fact, success represents 1 per cent of your work, which results from the 99 percent that is called failure.* Soichiro Honda (1906-1991).

O segundo tipo de balanço, em jeito da mudança anunciada, resulta no deslocar do foco de investigação das caudas super-pesadas para centrar a atenção em distribuições de cauda leve com limite superior do suporte finito, quase diametralmente opostas às primeiras. O trabalho actualmente em curso, no âmbito do projecto EXTREMA - Estatística de Extremos no Mundo Actual (financiado pela Fundação para a Ciência e a Tecnologia, FCT), incide mais precisamente na caracterização de uma classe particular de distribuições de probabilidade, de cauda leve, e subsequente estimação do limite superior do suporte. Um exemplo de aplicação interessante é o da análise das melhores marcas alcançadas no atletismo.

Em Fevereiro de 2011 iniciarei a colaboração num novo projecto de investigação da FCT: o projecto ENES - Extremos no Espaço. Afinal, na linha temporal do período de regime experimental tem cabimento qualquer coisa como um projecto e meio de investigação da FCT. Ficará sempre o consolo de que o todo é certamente mais do que a soma das nossas experiências e um importante denominador comum: o trabalho desenvolvido com os Professores Ivette Gomes, Isabel Fraga Alves e Laurens de Haan, legítimos precursores das minhas actividades pós-doc.

# Tese de Doutoramento: Estimação de parâmetros de acontecimentos raros.

(Boletim SPE Outono de 2007, p. 50)

Frederico Caeiro, *fac@fct.unl.pt*

*Faculdade de Ciências e Tecnologia e Centro de Matemática e Aplicações*
*Universidade Nova de Lisboa*

Na minha dissertação de doutoramento abordei a estimação de diversos parâmetros relevantes para a área de valores extremos. Optei por fazer aqui uma breve referência a um desses parâmetros, o índice de valores extremos, e a apresentar alguns resultados acerca da estimação semi-paramétrica MVRB (do inglês "Minimum-Variance Reduced Bias") deste parâmetro (Caeiro *et al*., 2005).

Em Estatística de Extremos, o índice de valores extremos tem um papel relevante na estimação de parâmetros relacionados com acontecimentos extremos, como por exemplo um quantil elevado ou um período de retorno associado a um nível elevado. O índice de valores extremos é um parâmetro real, e quanto maior o seu valor, mais pesada é a cauda do modelo $F$ subjacente aos dados. Vamos assumir que $F$ pertence ao domínio de atracção para máximos da distribuição de valores extremos, com função de distribuição $G_\gamma(x) = \exp\{-(1+\gamma x)^{-1/\gamma}\}$, $1+\gamma x > 0$. Neste modelo, $\gamma$ é o índice de valores extremos, o parâmetro que interessa estimar. Assumimos também que $F$ é um modelo de cauda pesada, isto é, um modelo pertencente ao domínio de atracção para máximos do modelo de valores extremos com $\gamma > 0$.

A estimação deste parâmetro começou por ser de índole paramétrica. Os primeiros estimadores semi-paramétricos deste parâmetro, baseados nas k+1 maiores observações, $X_{n-k:n} \le X_{n-k+1:n} \le \ldots \le X_{n:n}$, da amostra aleatória, $X_1, X_2, \ldots, X_n$, são o estimador de Hill (1975),

$$H(k) = \frac{1}{k}\sum_{i=1}^{k} \ln X_{n-i+1:n} - \ln X_{n-k:n} \,,$$

válido apenas para $\gamma > 0$, e o estimador de Pickands (1975),

$$P(k) = \frac{1}{\ln 2} \ln \frac{X_{n-\lfloor k/4 \rfloor:n} - X_{n-\lfloor k/2 \rfloor:n}}{X_{n-\lfloor k/2 \rfloor:n} - X_{n-k:n}},$$

válido para qualquer $\gamma$ real. Estes estimadores são muito sensíveis à escolha do nível $k$. Por um lado, este estimadores têm variância elevada para $k$ pequeno, e por outro lado, têm viés elevado para $k$ elevado. Consequentemente o erro quadrático médio (EQM), enquanto função de $k$, costuma apresentar um padrão em forma de "V", o que dificulta a escolha do nivel óptimo, isto é, o nível que minimiza o EQM.

O estimador de Hill tem sido estudado por vários autores (Hall 1982, Mason 1982, de Haan and Peng 1998, entre outros). Sabe-se que é consistente desde que $k$ seja uma sequência intermédia de valores inteiros, isto é, uma sequência de valores inteiros verificando $k \to \infty$ e $k/n \to 0$, quando $n \to \infty$. Impondo algumas condições adicionais acerca do comportamento de 2ª ordem da cauda do modelo $F$, garantimos a normalidade assintotica deste estimador. Podemos escrever $H(k) \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}} Z_k + \frac{\gamma\beta}{1-\rho}\left(\frac{n}{k}\right)^{\rho}(1+o_p(1))$, sendo $Z_k$ uma variável aleatória com distribuição normal padrão, $\rho \leq 0$ e $\beta$ parâmetros de segunda ordem do modelo $F$. Consequentemente se escolhermos $k$ de modo que $\sqrt{k}\gamma\beta\left(\frac{n}{k}\right)^{\rho} \to \lambda$, finito, então $\sqrt{k}\left(H(k)-\gamma\right) \xrightarrow{d} N\left(\frac{\lambda}{1-\rho}, \gamma^2\right)$.

O termo dominante de viés deste estimador, $\gamma\beta(n/k)^{\rho}/(1-\rho)$, influencia o seu comportamento, especialmente quando $\rho$ está próximo de 0. Devido a esta propriedade, muitos investigadores, entre os quais mencionamos Peng (1998), Beirlant *et al.* (1999), Feuerverger and Hall (1999), Gomes *et al.* (2000), Beirlant *et al.* (2002), Caeiro and Gomes (2002), removeram o termo dominante de viés de modo adequado. Em todos estes trabalhos, os autores estudaram novos estimadores de viés reduzido, com variância assintótica maior ou igual a $(\gamma(1-\rho)/\rho)^2$. Esta variância é sempre superior a $\gamma^2$, a variância do estimador de Hill, que corresponde à mais pequena variância que qualquer estimador do índice de valores extremos pode ter (por ser um estimador obtido por máxima verosimilhança). Todos os estimadores de viés reduzido, dos trabalhos acima referidos, são geralmente mais eficientes que o estimador de Hill, mas apenas em níveis próximos do nível óptimo.

Atendendo à expressão do termo dominante de viés de $H(k)$, mais recentemente Caeiro *et al.* (2005) procederam à estimação directa e remoção desse termo e estudaram os seguinte estimadores de Hill corrigidos,

$$CH(k) = H(k)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right)$$

e

$$\overline{CH}(k) = H(k)\exp\left(-\frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right)$$

onde $(\hat{\rho}, \hat{\beta})$ são estimadores consistentes dos parâmetros de segunda ordem $(\rho, \beta)$. Se a estimação de $\rho$ e $\beta$ for feita num nível $k_1$, de ordem superior ao nível $k$ usado na estimação do índice de valores extremos, e se $(\hat{\rho}-\rho)\ln(n/k) = o_p(1)$, conseguimos remover o termo dominante de viés sem aumentar a variância assintótica $\gamma^2$. Estes dois estimadores do índice de valores extremos são estimadores MVRB porque, relativamente ao estimador de Hill, mantêm a mesma variância assintótica, $\gamma^2$, e têm viés assintótico de ordem inferior. Estas propriedades permitem-nos assegurar que o EQM dos estimadores MVRB é menor ou igual que o EQM do estimador de Hill, para todos os níveis $k$.

Para mais detalhes acerca dos estimadores MVRB sugerimos Reiss and Thomas (2007), Capítulo 6 e Caeiro *et al.* (2009).

**Referências**


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

Beirlant, J., Dierckx, G., Guillou, A., Starica, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes,* 5(2):157–180.

Caeiro, F., Gomes, M.I. (2002). A class of asymptotically unbiased semi-parametric estimators of the tail index. *Test*, 11(2), 345–364.

Caeiro, F., Gomes, M.I. and Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator. *Revstat,* 3(2), 113-136.
Caeiro, F., Gomes, M.I. and Henriques-Rodrigues, L. (2009). Reduced-bias tail index estimators under a third order framework. *Communications in Statistics - Theory and Methods*, 38(7), 1019-1040.
Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution. *Ann. Statist.*, 27, 760–781.
Gomes, M.I., Martins, M.J. and Neves, M. (2000). Alternatives to a semiparametric estimator of parameters of rare events - the Jackknife methodology. *Extremes*, 3(3), 207-229.
Haan, L. de, Peng, L. (1998). Comparison of tail index estimators. Stat. Neerl. 52, 60–70.
Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. R. Stat. Soc. B*, 44, 37–42.
Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3, 1163-1174.
Mason, D.M. (1982). Laws of large numbers for sums of extreme values. *Ann. Probab.*, 10, 754–764.
Peng, L. (1998). Asymptotically unbiased estimator for the extreme-value index, *Statistics and Probability Letters*, 38(2), 107–115.
Reiss, R.-D., and Thomas, M.(2007). Statistical Analysis of Extreme Values, with Application to Insurance, Finance, Hydrology and Other Fields, 3rd edition, Birkhäuser Verlag.

# Tese de Doutoramento: Análise Preditiva em Populações Finitas

(Boletim SPE Outono de 2007, p. 53)

Susana Rosado-Ganhão, *srosado@fa.utl.pt*

*Faculdade de Arquitectura da Universidade Técnica de Lisboa*

No dia 29 de Janeiro de 2007 fiz as minhas provas de doutoramento gravidíssima - a Patrícia tinha 7 meses de gestação!

Após saber o resultado das provas senti-me ao mesmo tempo aliviada por ter sido bem sucedida e apreensiva, pois este é um passo muito importante para alguém, como eu, que está na carreira universitária.

É claro que os 6 meses que se seguiram foram de profunda e pura investigação humana – uma bebé recém-nascida para alimentar, tratar e principalmente mimar; para além das outras duas princesas que já faziam parte do agregado familiar!
Não houve mãos a medir, como se pode imaginar…

Com a disponibilidade fundamental e imprescindível dos avós, comecei então a sentir os efeitos, mais imediatos, do grau de doutor obtido: a partir desta altura passamos a ter lugar no Conselho Científico, uma realidade que agora se alterou com os novos Estatutos das Faculdades, pois este órgão passou a ser eleito. Ao nível do Departamento, tem-se um papel mais participativo uma vez que passamos a integrar o Conselho de Departamento. Para além disso somos, também, convidados a fazer parte de júris de provas académicas.

Nesta nova etapa as solicitações são muitas, e comecei a participar em grupos de trabalho com temas de investigação diversos, dada a especificidade da Faculdade onde lecciono e onde construo a minha carreira: Faculdade de Arquitectura da Universidade Técnica de Lisboa (FAUTL).

Um dos projectos, na área da Arquitectura, cuja equipa de investigação integro é o Naadir: uma abordagem ao desenho arquitectónico integrando descrições computacionais.
Este consiste em ampliar e potenciar o papel da perspectiva na representação gráfica do espaço. A área principal de interesse é a do desenho arquitectónico e urbanístico, o seu ensino e as suas práticas, para as quais se pretende contribuir através de uma nova abordagem didáctica e de uma ferramenta computacional de trabalho melhorado. A preocupação central da equipa é o modo como os processos de desenho influenciam a racionalização do espaço percepcionado e o decurso da concepção. O principal objectivo deste projecto é conseguir a implementação computacional de um sistema de perspectiva expandido, desenvolvido pela equipa, que permitirá cumprir as tarefas específicas compreendidas por este sistema e a sua repercussão na prática e didáctica do desenho conceptual.
Esta equipa reúne estrategicamente um grupo diversificado de investigadores, cujas competências e conhecimentos abarcam áreas do Desenho, da Geometria, da Matemática, da Computação e da Arquitectura. Fui convidada a pertencer a esta equipa, pelos meus conhecimentos ao nível da Matemática e da programação computacional, dos quais a minha tese de doutoramento é espelho, uma vez que desenvolvi programas para realizar simulações que me permitissem testar a metodologia desenvolvida, na tese, ao nível da análise preditiva bayesiana.
Este projecto está a decorrer e tem sido extremamente interessante trabalhar com uma equipa cientificamente tão diversificada.

Uma outra porta que se abre a partir do momento em que somos Professores Doutores, é a da orientação de dissertações de alunos.
Co-orientei dois alunos do Mestrado Integrado em Arquitectura da FAUTL, cujas teses incidiram sobre a avaliação da sustentabilidade de edifícios.
É um tema extremamente interessante e actual, uma vez que as preocupações ambientais são uma constante nas nossas vidas, nos tempos que correm.
Essas teses tiveram como objectivo comum quantificar a sustentabilidade ambiental dos edifícios, fazendo um estudo repartido em 3 grandes fases fundamentais – construção, utilização e reabilitação, e incidiram sobre um conjunto de edifícios de determinadas épocas de construção: uma é sobre o Parque das Nações[1] e a outra sobre os Edifícios Gaioleiros[2].

Inicialmente, a partir das avaliações dos edifícios estudados, elaboraram-se fichas de levantamento da informação, que foram analisadas criticamente, para depois poderem constituir o conjunto de indivíduos e variáveis utilizados no cálculo do índice quantitativo. Posteriormente atribuiu-se a importância percentual a cada variável, no número e no limite das classes em que as variáveis são definidas, com base no método EPM - Environmental Preference Method.
Aplicou-se, então, a Discriminação Baricêntrica – adaptação da Análise Factorial das Correspondências à discriminação em torno de pólos que são o centro de gravidade dos atributos de partida (Ribeiro, 1999)[3] - aos elementos definidores dos extremos da escala de qualidade (eixo discriminante), constituindo a coordenada da projecção dos indivíduos nessa escala a representação quantitativa da qualidade.
Os resultados foram comparados com mais informação, obtida posteriormente a partir de indivíduos desconhecidos ou retirados aleatoriamente do conjunto de dados, de modo a modificar alguns parâmetros, repetindo-se todo o processo até que o índice esteja calibrado, validando a metodologia e a consequente utilização do índice como instrumento de previsão.
Chegou-se a conclusões muito interessantes ao nível da sustentabilidade dos edifícios, sempre muito bem justificadas com os inúmeros estudos e tratados escritos sobre este assunto.
A literatura sobre a sustentabilidade é vastíssima e não se restringe aos edifícios, embora este seja um dos temas intimamente relacionados com a Arquitectura e o Urbanismo.

Um outro projecto em que estou envolvida, e que me dá bastante satisfação, diz respeito à divulgação da Matemática e sua ligação/aplicação à "realidade", para os alunos do ensino secundário, e também para os alunos da FAUTL.
A iniciativa, da Reitoria da UTL coordenada pelo Prof. Nuno Crato, chama-se "Rotas Matemáticas da UTL" e consiste num encontro anual com alunos do ensino secundário que se deslocam às escolas da UTL e assistem a palestras e actividades que revelam a ligação da Matemática com a arte, a economia, a estatística, a sociologia, a medicina, a medicina veterinária, o desporto, etc.

As palestras que fazemos têm também sido parte integrante de um seminário que eu e o meu colega Prof. Jorge Ribeiro organizamos na FAUTL, já com duas edições. Intitulado "Momentos Matemáticos" tem como objectivo despertar os alunos da FAUTL para as maravilhas da Matemática e o seu papel fundamental em todos os ramos do conhecimento.

Em termos de leccionação, passei a fazer parte do corpo docente do Curso de Doutoramento em Arquitectura, a convite dos meus colegas que leccionam neste curso. Esta transmissão de conhecimento a um nível mais avançado e com uma plateia muito mais estimulante que está noutra fase da sua vida e exige outro tipo de abordagem dos assuntos, da nossa parte, é uma experiência muito interessante e enriquecedora.

---

[1] Barreto, A.L.C. (2010). *Avaliação da Sustentabilidade de Edifícios: Parque das Nações.* Dissertação de Mestrado em Arquitectura, FAUTL, Lisboa.
[2] Heleno, A.R. (2010) *Avaliação da Sustentabilidade de Edifícios: os edifícios "gaioleiros"*. Dissertação de Mestrado em Arquitectura, FAUTL, Lisboa.
[3] Ribeiro, J. (1999) Formulação de índices quantitativos com base na Discriminação Baricêntica. Tese de Doutouramento, UTL/IST, Lisboa.

Não me querendo alongar, julgo ter dado uma perspectiva geral do meu "pos-doc" e espero daqui a 3 anos ter muitas outras experiências de trabalho e projectos interessantes para partilhar.

Devo acrescentar que foi muito interessante escrever este artigo e "rever" estes últimos anos de trabalho. Obrigada pelo desafio e pela ideia tão inovadora.

Saudações académicas
Susana Rosado-Ganhão

# Tese de Doutoramento: Distribuições Conjugadas e Aproximações

(Boletim SPE Outono de 2007, p. 52)

Madalena Malva, *malva@estv.ipv.pt*

*Instituto Politécnico de Viseu*

Quando eu nasci,
ficou tudo como estava,
Nem homens cortaram veias,
nem o Sol escureceu,
nem houve Estrelas a mais...
Somente,
esquecida das dores,
a minha Mãe sorriu e agradeceu.

Quando eu nasci,
não houve nada de novo
senão eu.

As nuvens não se espantaram,
não enlouqueceu ninguém...

P'ra que o dia fosse enorme,
Bastava
toda a ternura que olhava
nos olhos de minha Mãe...

José Régio

Quando o Prof. Fernando Rosado me convidou para participar nesta secção do boletim da SPE foi do poema de José Régio que me lembrei. Porquê? Porque apesar das "dores de parto" quando o júri me comunicou a sua decisão "esquecida das dores apenas **eu** sorri e agradeci.". Estou a ser mal agradecida, claro que a família e os amigos jubilaram com o meu sucesso mas "As nuvens não se espantaram, não enlouqueceu ninguém..."

Passados quatro anos, a nível profissional continua quase tudo na mesma, passei de equiparada a professora adjunta a professora adjunta com contrato por tempo indeterminado graças à nova Lei n.º 7/2010, de 13 de Maio, sem esta lei provavelmente continuaria na mesma situação. Além disso, o meu curriculum começou a ser mais solicitado porque a certificação dos cursos exige um determinado número de doutores por área! Eu também passei a ser mais requisitada pelos colegas das áreas da engenharia e ciências sociais que no desenvolvimento dos seus projectos de investigação necessitam de fazer tratamento estatístico de dados. Pena é que julguem que uma pessoa doutorada em Estatística sabe tudo sobre Estatística, e se não sabe "facilmente" percebe do assunto se ler umas linhas! Pena é que também a maioria destes trabalhos só utilizem a chamada "estatística do *p*-value", e na maior parte dos casos ser apenas isso que interessa às pessoas!

Quanto ao trabalho cientifico continuo interessada e a trabalhar na área do meu doutoramento, ou seja, nas distribuições conjugadas e suas aplicações.

Resultados assintóticos são inspiração e justificação de importantes aproximações usadas em modelação estatística, mesmo quando o intuito é analisar uma mão-cheia de dados. Há, de facto, situações em que a convergência é rápida, mas muitas outras são conhecidas em que a convergência é deveras lenta. Os estudos sobre velocidades de convergência são naturalmente uma parte essencial do *corpus* das convergências estocásticas.

Ao Teorema Limite Central, se o cerne da investigação não se situar nas questões de velocidade de convergência, basta existência de segundo momento (e mesmo esta exigência pode ser relaxada, usando o conceito de variação lenta de Karamata). Terceiro e quarto momentos — e portanto,

indirectamente, assimetria e achatamento da população parente —, por outro lado, são os instrumentos adequados para a abordagem inicial das questões de velocidade de convergência.
Reexpressando assimetria e achatamento em termos dos cumulantes de Thiele--Fisher, e retomando ideias implícitas nas expansões de Gram--Charlier e de Edgeworth, por um lado, e de Cornish—Fisher por outro, reencontramos outros instrumentos preciosos, tais como as transformadas de Esscher e distribuições conjugadas de Cramér--Khintchine, e as aproximações usando ponto de sela (que são a base de toda a área de *small sample assymptotics*); as expansões de Edgeworth diferidas (*tilted Edgeworth expansions*, na linguagem de Barndorff-Neilsen e Cox), são, de facto, o traço de união de todos estes resultados.
Por outro lado, aproximações excelentes são válidas em situações inesperadas, como no exemplo da estável de Lévy, que apesar de não ter sequer primeiro momento pode ser aproximada de forma muito adequada com aqueles instrumentos, que pareciam ter sido talhados para circunstâncias bem diversas.
A investigação de expansões de Edgeworth diferidas levou-me naturalmente ao estudo de famílias exponenciais (e polinómios ortogonais associados), e particularmente às famílias naturais de Morris, com variância que é função quando muito quadrática do valor médio.
Por outro lado, o exemplo da Lévy levou-me a concentrar algum esforço na aproximação por leis estáveis para somas, obtendo resultados parciais mais interessantes no caso de a distribuição parente estar no domínio de atracção *standard*. Por outro lado, o recurso a *tilting* permite bons resultados na chamada zona de grandes desvios, e abre a perspectiva de a velocidade de convergência no Teorema Limite Central melhorar de O( $1/\sqrt{n}$) para O(1/n), desde que se use devidamente a teoria de que Daniels foi pioneiro. Uma incursão pela teoria da informação, perspectivando resultados assintóticos como os que correspondem a entropia máxima em situações bem tipificadas, ajuda a compreender este progresso notável.
Assim, a minha vida profissional segue entre as "duas estatísticas" anteriores! Uma por gosto outra porque tem de ser…


- Cramér, H. (1963). On asymptotic expansions for sums of independent random variables with limiting stable distributions, *Sankhya Ser.,* **25**, 12-24.
- Daniels, H. E. (1954). Saddlepoint approximations in Statistics, *Ann. Math. Statist.,* **25**, 631-650.
- Edgeworth, F.\ Y. (1905). The Law of Error, *Cambridge Philos. Soc*., **20**, 36-66.
- Fisher, R. A. (1934). Two new properties of mathematical likelihood, *Proc. Royal Soc*., **144**, 295-307.
- Morris, C. N. (1982). Natural exponencial families with quadratic variance functions, *Ann. Stat*., **10**, 65-80.
- Morris, C. N. (1983). Natural exponencial families with quadratic variance functions: statistical theory, *Ann. Stat.*, **2**, 515-529.
- Thiele, T. N. (1903). Theory of Observations, Layton, London (republicado em *Ann. Math. Stat.*, **2**, 165-308, 1931).

# Ciência Estatística

## • Artigos Científicos Publicados


Caeiro, F. and Gomes, M.I. (2010). An asymptotically unbiased moment estimator of a negative extreme value index. *Discussiones Mathematicae Probability and Statistics* **30**:1, 5-19.

Caiado, J. and N. Crato (2010). **"**Identifying common dynamic features in stock returns", *Quantitative Finance*, 10, 797-807.

Ferreira, L.N.; Ferreira, P.L.; Pereira, L.N.; Brazier, J. e Rowen, D. (2010). A Portuguese Value Set for the SF-6D. *Value in Health,* 13:5, 624-630.

Ferreira, M. (2010). Estimation of the parameter of a pARMAX model. *REVSTAT* 8(2), 139-149.

Ferreira, M. e Canto e Castro, L. (2010). Asymptotic and Pre-Asymptotic Tail Behavior of a Power Max-Autoregressive Model. *ProbStat Forum* 3(8), 91-107.

Gil, Pedro M., Figueiredo, F. and Afonso, O. (2010). Equilibrium Price Distribution with Directed Technical Change. *Economics Letters*, 108, 130-133.

Gomes, M.I. and Henriques-Rodrigues, L. (2010). Comparison at optimal levels of classical tail index estimators: a challenge for reduced-bias estimation? *Discussiones Mathematica: Probability and Statistics* 30:1, 35-51.

Grilo, L. M. e Coelho, C. A. (2010). Near-exact distributions for the generalized Wilks Lambda statistic. *Discussiones Mathematicae Probability and Statistics*, 30 (1) 53-86.

Mendes, S., Mª José Fernández Gómez, Mário Jorge Pereira, Ulisses Miranda Azeiteiro e Mª Purificación Galindo-Villardón (2010). The efficiency of the Partial Triadic Analysis method: an ecological application. *Biometrical Letters*. Vol. 47 (2010), No. 2 , 83-106.

Neves, M.M. and Cordeiro, C. (2010). Exponential smoothing and resampling techniques in time series prediction. *Discussiones Mathematicae. Probability and Statistics* 30:1, 87-101.

Prata Gomes, D. and Neves, M.M. (2010). Extremal behaviour of stationary processes: the calibration technique in the extremal index estimation. *Discussiones Mathematicae. Probability and Statistics* 30, 21-33.

Ramos, S., Antónia Amaral Turkman, Marília Antunes. (2010). Bayesian classification for bivariate normal gene expression. *Computational Statistics and Data Analysis* 54, 2012-2020.


## • Revistas

**Título:** *Portuguese Journal of Quantitative Methods*
**Editores:** José António Filipe e Marina Andrade
Ano: 2010 (1ª Edição). Editora: Edições Sílabo. ISSN: 1647-7987

## • 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. **http://www.fundacion.uc3m.es/IWAP2010/Extended_Abstracts.html** Abstract available : In Arribas, A., Glaz, J., Jiménez, R. and Romo, J. (eds.). IWAP 2010, Universidad Carlos III de Madrid editions, 56.

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.
**http://www.springer.com/statistics/computanional+statistics/book/978-3-540-32825-4**

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. **http://www.springer.com/mathematics/probability/book/978-3-0348-0008-2**

## • Livros

**Título:** *Análise de Equações Estruturais - Fundamentos teóricos, Software & Aplicações.*
**Autor:** João Marôco
Ano: 2010. Editora: ReportNumber, Lda. ISBN: 978-989-96763-1-2

**Título:** *Estatística e Probabilidades. Aplicações e Soluções em SPSS.*
**Autoras:** Anabela Afonso e Carla Nunes
Ano: 2011. Editora: Escolar Editora. ISBN:978-972-592-299-6

## • Teses de Mestrado

**Título:** *Interval Time Series Approach to High-Frequency Data: An Application to PSI20*
**Autora:** Nazarii Salish, *nazarii.salish@gmail.com*
**Orientador:** Paulo M. M. Rodrigues

**Título:** *Análise da evolução das actividades económicas em Portugal através da metodologia Statis*
**Autora:** Glória Silvana Gonçalves, *080414001@fep.up.pt*
**Orientadoras:** Adelaide Figueiredo e Fernanda Otília Figueiredo

## • Teses de Doutoramento

**Título:** *Estimação robusta em modelos lineares de equações simultâneas*
**Autora:** Anabela Virgínia dos Santos Flores da Rocha, *anabela.rocha@ua.pt*
**Orientadores:** Maria Manuela Souto Miranda e João António Branco

Na minha tese apresento um estudo sobre Modelos de Equações Simultâneas (*SEM*), que são modelos estatísticos com muita tradição em estudos de Econometria, uma vez que permitem representar e estudar uma vasta gama de processos económicos. São apresentados os estimadores mais usados em *SEM,* resultantes da aplicação do Método dos Mínimos Quadrados ou do Método dos Momentos Generalizado (*GMM*), assim como as respectivas propriedades, focando, em particular, a falta de robustez estatística. A robustez de um estimador é uma propriedade importante, uma vez que um método não robusto pode conduzir a resultados enganadores quando são violadas as hipóteses subjacentes ao modelo assumido. A questão assume ainda maior relevância quando os modelos em estudo são complexos, como é o caso dos *SEM*.
O principal objectivo da investigação foi o de procurar métodos robustos para estimar os parâmetros dos SEM, tendo-se desenvolvido um estimador robusto a que se deu o nome de *GMMOGK,* pelo facto de constituir uma versão robusta do estimador *GMM,* com recurso ao estimador Ortogonalizado de Gnanadesikan-Kettenring. Para avaliar o desempenho do novo estimador foi efectuado um estudo de simulação e foi também estudada a sua aplicação a um conjunto de dados reais. O estimador robusto tem um bom desempenho nos modelos heterocedásticos considerados e, nessas condições, comporta-se melhor do que os estimadores não robustos usados no estudo. Contudo, e tal como acontece com os estimadores tradicionais nos SEM, verificou-se que quando a análise é feita em cada equação separadamente, a especificidade da forma de cada equação e da estrutura de dependência do sistema, são dois aspectos que influenciam o desempenho do estimador. Para enquadrar a investigação, o texto inclui uma revisão de aspectos essenciais dos *SEM*, o seu papel em Econometria, os principais métodos de estimação, com particular ênfase no *GMM*, e uma curta introdução à estimação robusta.

Anabela Rocha

**Título:** *Métodos estatísticos de screening em classificação supervisionada*
**Autora:** Sandra Cristina de Faria Ramos, *sfr@isep.ipp.pt*
**Orientadoras:** Maria Antónia Amaral Turkman e Marília Antunes

Na minha tese apresentam-se as contribuições resultantes de um trabalho de investigação sobre métodos bayesianos de *screening* em classificação supervisionada num cenário bivariado, ou seja, métodos que permitem atribuir a um novo indivíduo uma categoria de entre um conjunto de categorias mutuamente exclusivas, com base na observação de vectores de características bidimensionais nesse indivíduo.

Considerando a formulação do problema de *screening* do ponto de vista preditivo bayesiano mostra-se como se pode construir uma região de classificação óptima quando se admite um modelo gaussiano bivariado para o vector de características. Seguidamente introduzem-se alterações no modelo inicial de forma a remover restrições no que respeita a pressupostos distribucionais. Nesta generalização consideram-se duas abordagens não paramétricas. A primeira usa métodos do núcleo multivariados para estimar a distribuição preditiva de uma observação futura condicional às várias categorias da variável resposta. A segunda usa os actuais métodos bayesianos não paramétricos para estimar essa distribuição preditiva.

É proposta uma regra de classificação baseada em múltiplos pares de variáveis, que resulta da combinação da classificação e de quantidades preditivas *a posteriori* resultantes da aplicação do método a cada par de variáveis.

Para ultrapassar os problemas de cálculo associados com a obtenção da região de classificação e das probabilidades preditivas desenvolveu-se um conjunto de algoritmos assentes em métodos de integração numérica e de simulação estocástica. Todos os programas computacionais foram implementados em ambiente R e permitem obter a região de especificação de forma automática.

São apresentados e discutidos resultados da ilustração da metodologia proposta quando aplicada a conjuntos de dados reais correspondentes a níveis de expressão genética e a conjuntos de dados simulados. O estudo de simulação mostrou que as abordagens não paramétricas, com destaque para a bayesiana, são as mais flexíveis, visto apresentarem bons resultados mesmo na presença de classes não separáveis por funções paramétricas simples e/ou na presença de pequenas amostras provenientes de populações não normais.

O classificador bayesiano que se apresenta generaliza os métodos de classificação clássicos, pois permite a obtenção de fronteiras paramétricas flexíveis, sem necessidade de fixar previamente a sua forma e possibilita o cálculo de um conjunto de quantidades preditivas de interesse.

Sandra Ramos

## • Edições SPE - Mini Cursos

**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

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

# PRÉMIOS "ESTATÍSTICO JÚNIOR 2011"

Está aberto, até 27 de Maio de 2011, o concurso para atribuição dos prémios **"Estatístico Júnior 2011"**, de acordo com o seguinte regulamento:

**1.** A atribuição dos prémios **"Estatístico Júnior 2011"** 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 dos ensinos Básico e Secundário pelas áreas da Probabilidade e da Estatística.

**2.** Os candidatos aos prémios **"Estatístico Júnior 2011"** 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 2010-2011.

**3.** As candidaturas podem ser individuais ou em **grupo com um máximo de 3 alunos**. Do grupo pode ainda fazer parte um professor do Ensino Básico ou Secundário a quem 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 póster formato A2 que resuma os principais aspectos do trabalho. O trabalho (póster e texto escrito) deverá ser **enviado impresso em papel para efeitos de avaliação**.

**6.** Poderão ser atribuídos prémios **"Estatístico Júnior 2011"** 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 Esino Básico e Secundário, e 600 euros para a categoria Cursos CEF-EFA.

**7.** Ao professor orientador do trabalho classificado em 1.° lugar, em cada categoria, é ainda atribuída uma anuidade grátis como sócio da SPE, ajudas de custo para participação no XVII 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 1.° lugar será também oferecida uma ampliação do correspondente póster que será colocado na Sessão de Pósteres 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 dos Correios 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/static/docs/BoletimCandidaturaPEJ11.pdf
http://www.spestatistica.pt/static/docs/RegulamentoPEJ11.pdf

**10.** A admissibilidade e apreciação dos trabalhos submetidos a concurso são da competência de um júri, cujas 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 2011"** 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 2011"** poderão não ser atribuídos.

Apoio Porto Editora

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

# PRÉMIO SPE 2011

Está aberto, até **15** de **Junho** de **2011**, o concurso para atribuição do **Prémio SPE 2011**, de acordo com o seguinte regulamento:

1. Pretendendo dar destaque ao XIX Congresso Anual da **SPE,** a principal reunião científica organizada pela Sociedade Portuguesa de Estatística, é instituído o **Prémio SPE 2011**.
2. Este prémio destina-se a estimular a actividade de estudo e investigação científica em Probabilidade e Estatística entre os jovens que trabalham nestas áreas.
3. O **Prémio SPE 2011** é constituído por uma quantia de 1000 euros.
4. Ao **Prémio SPE 2011** podem concorrer trabalhos originais sobre temas de Probabilidade e Estatística, desde que não tenham sido objecto de qualquer prémio atribuído por outra instituição.
5. Os autores dos trabalhos candidatos ao **Prémio SPE 2011** devem ser estudantes ou investigadores em alguma instituição portuguesa ou bolseiros portugueses, devem ser sócios da **SPE** e não devem ter completado os 35 anos de idade até 15 de Junho de 2011. Os autores não devem ter recebido o Prémio SPE nas quatro edições anteriores.
6. O trabalho deve ser escrito em português ou inglês e não poderá exceder 25 páginas A4.
7. As candidaturas deverão vir acompanhadas do trabalho concorrente e do *curriculum vitae* dos autores e ser dirigidas ao Presidente da **SPE**, em carta registada, para a morada abaixo indicada. O carimbo do correio validará a data de entrega.
8. A admissibilidade e a apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição será da responsabilidade da Direcção da **SPE**.
9. O júri é soberano nas suas decisões, não havendo lugar a recurso.
10. O trabalho galardoado com o **Prémio SPE 2011** será apresentado em sessão plenária pelo seu autor ou autores no XIX Congresso Anual da **SPE** e será publicado nas respectivas Actas.
11. A atribuição do **Prémio SPE 2011** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será feita no XIX Congresso Anual da **SPE** na sessão plenária da sua apresentação.
12. O **Prémio SPE 2011** poderá não ser atribuído.

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

# Índice