CONTRIBUIÇÃO À ANÁLISE DE DADOS CATEGORIZADOS

Rita Maria César e Sá Fernandes de Vasconcelos

UNIVERSIDADE DA MADEIRA
JULHO 1994

CONTRIBUIÇÃO À ANÁLISE DE DADOS CATEGORIZADOS

Rita Maria César e Sá Fernandes de Vasconcelos

Dissertação submetida com vista à obtenção do doutoramento em Estatística Matemática, na Universidade da Madeira.

Julho 1994

AGRADECIMENTOS

Ao Prof. Doutor Dinis Pestana agradeço a dedicação e o empenho com que me orientou, a riqueza de ideias e conhecimentos sempre actualizados que frutificaram nesta tese, a infindável paciência para ler os inúmeros *faxes* que lhe enviei ao longo destes três anos, a voz entusiasmante que ouvi ao telefone nos momentos mais difíceis...a compreensão.

Não posso deixar de agradecer ao Prof. Doutor António St.'Aubyn, com quem iniciei esta carreira e que me lançou nesta área, a confiança que sempre demonstrou ter no meu trabalho.

À Universidade da Madeira agradeço todo o apoio que me concedeu, sem o qual este trabalho não se teria concretizado.

Ao Serviço de Cardiologia do Centro Hospitalar do Funchal, agradeço a cedência dos dados e o interesse demonstrado na realização deste trabalho. Ao Dr. António de Almada Cardoso agradeço a confiança que depositou neste projecto. Ao Dr. António Drumond agradeço a disponibilidade com que sempre me recebeu e as frutíferas trocas de impressões sobre as implicações dos resultados a que fui chegando.

Ao Centro de Estatística e Aplicações da Universidade de Lisboa agradeço o apoio bibliográfico e o *software* que disponibilizou para utilização neste trabalho.

Aos meus colegas do Departamento agradeço todo o apoio que me deram. Um agradecimento especial aos meus colegas de gabinete, José Carmo, José Castanheira, António Casimiro Silva e Miguel Nunes, pela amizade, pelo excelente ambiente de trabalho e

pelo encorajamento nos momentos desanimadores. Registo também o interesse do Prof. Doutor Streit, e a influência que tem tido no Departamento, valorizando a investigação científica. Ao António Pires quero agradecer o apoio, sempre pronto, prestado na área da informática.

À Luisa Canto e Castro Duarte agradeço as discussões que foram tão importantes para a consolidação dos resultados que obtive.

Ao Rui Bettencourt agradeço a indispensável ajuda no "transporte" do enorme ficheiro do Serviço de Cardiologia para o computador que utilizei para desenvolver este trabalho.

Ao meu sogro, que não chegou a ver a conclusão desta tese, agradeço as palavras de confiança, esperança e alegria no desenrolar deste trabalho e que não me abandonaram nos piores momentos. À minha família e à família do meu marido agradeço a amizade com que acompanharam este meu empenhamento no trabalho. Aos meus pais quero agradecer o apoio que deram aos meus filhos, o qual tornou possível a minha dedicação plena a este projecto. À Ângela agradeço a amizade, a disponibilidade e a dedicação aos meus filhos durante as minhas ausências.

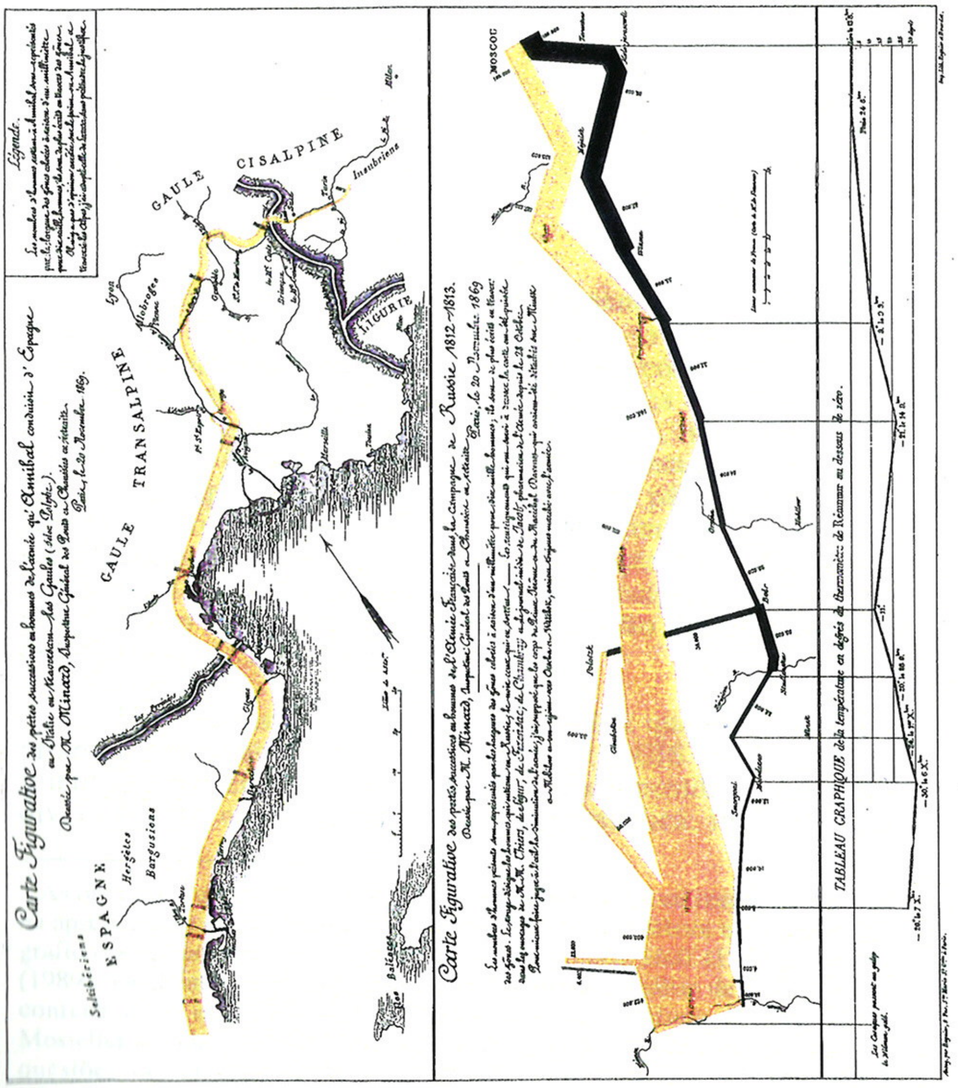
Às minhas Amigas agradeço o entusiasmo que souberam transmitir-me, o encorajamento nos momentos certos e o apoio que deram aos meus filhos.

Ao Filipe quero agradecer a infinita compreensão e o carinho com que me acompanhou nesta "travessia do deserto". Aos meus filhos, Francisco e Catarina, quero agradecer a confiança que tiveram em mim e o entusiasmo, algumas vezes superior ao meu, com que enfrentaram a minha dedicação a este projecto.

Índice

III Análise Exploratória de uma base de dados sobre Cardiologia. Atribuição de uma dimensão espacial à doença coronária.

IV Aplicação do teste de ajustamento do Qui-Quadrado à análise da incidência da doença coronária em diferentes regiões.


Carte Figurative
GAULE CISALPINE
GAULE TRANSALPINE
ESPAGNE
LIGURIE
Carte Figurative
Russie 1812-1813.
MOSCOU
TABLEAU GRAPHIQUE

Introdução

1. A utilização de gráficos na apresentação de trabalhos estatísticos é uma arte, e começa a tornar-se uma ciência[1]; é bastante vulgar encontrar afirmações de senso comum, no sentido de um bom gráfico ser melhor do que mil palavras — e decerto não há gráficos que se possam comparar a mapas, no que respeita transmissão de informação. Tufte (1983) apresenta um "mapa" das campanhas napoleónicas, que reproduzimos na página anterior, da autoria de Minard, como o melhor gráfico jamais feito: contém simultâneamente informações sobre a deslocação dos exércitos (no espaço e no tempo), evolução do inimigo principal — as terríveis temperaturas do inverno russo — e a redução dos efectivos do exército, qual rio que se vai esgotando.

Não é consequentemente de estranhar que uma das mais promissoras novas áreas de investigação seja GIS — Geographic Intelligent Systems —, que combina conhecimentos de áreas tão diversificadas quanto Estatística e Geoestatística, Informática e

[1] As obras de Tufte (1983, 1990) deram nova dimensão à utilização de gráficos na apresentação de trabalhos estatísticos — mostrando embora a que ponto um gráfico pode ser enganoso; a este respeito ver também Huff (1954), e Leland (1989). Também as novas abordagens exploratórias à análise de dados vieram conferir novo estatuto a técnicas gráficas e semi-gráficas (Tukey, 1977; Mosteller e Tukey, 1977). E a abordagem geométrica tem sido frutuosa em questões tão diversas como estudentização em situações não gaussianas (Hotteling, 1961; Efron, 1969), análise da variância (Saville, 1980), ou mesmo teoria da medida.

Gestão de Informação, Geografia, Geociências. No campo específico da Estatística, depois dos trabalhos seminais de Matheron (1965, 1971) e da súmula de Mardia (1972) houve um desenvolvimento explosivo de investigação na área de estatística e processos estocásticos espaciais, de que recentemente a monografia de Cressie (1991) faz o ponto da situação. A bibliografia regista outras obras relevantes neste campo por nós consultadas.

A relevância da dimensão espacial na interpretação dos dados parece ter tido a sua primeira utilização intencional — e coroada de sucesso — no trabalho de John Snow (1813-1858), um anestesista de Londres que era o obstetra da Rainha Victória. Snow tinha a ideia que a cólera — a doença epidemiológica clássica do século XIX — podia ser propagada pelo abastecimento de água contaminada. Através do uso de mapas indicando a distribuição geográfica das mortes provocadas por cólera, demonstrou que associações entre mortes por cólera e abastecimento de água contaminada resultaram numa distribuição geográfica coincidente.

A história colectiva é hoje descrita em mapas[2], e mesmo a história individual pode ser elipticamente referida por uma selecção de lugares, como no notável poema "Passagem de Rimbaud, II", de Mário Cesariny de Vasconcelos:

> Mazan Charleville Bruxelas Charleville Paris Charleville Paris Charleville Bruxelas Londres Charleville Londres Roche Bouillon Londres Bruxelas,
>
> Roche Charleville Paris Londres Alemanha Suíça Itália Marselha Charleville Holanda Batávia Bordeus Charleville Viena Charleville,
>
> Holanda Hamburgo Suécia Dinamarca Marselha Alexandria Roma Charleville Hamburgo Charle-

[2] Consultar, por exemplo, o Grande Atlas dos Descobrimentos (1992), ou a excelente colecção de atlas históricos publicados pela Phaidon (cf. Bibliografia).

ville Suíça S. Gotardo Lugano Génova Alexandria Chipre Charleville Egito Aden,

Djeddah Suakin Hodeidah Massava Aden Zeylah Harar Bubassa Harar Aden Zeylah Harar Huebbe Harar Aden Tadjura Ankober Amtoto Harar Aden Cairo Aden Harar Aden Harar Zeylah Aden Marselha Roche Paris Lião Marselha.

(in *Planisfério e outros poemas*, p. 55)

Hoje em dia grandes empresas públicas ou privadas vivem exclusivamente da produção de ficheiros de "fronteiras" de mapas digitais, e da produção de *software* de utilização, quer se trate de mapas no sentido corrente do termo, quer as fronteiras tenham um sentido difuso (por exemplo, quando o que se pretende "cartografar" são zonas tróficas em regiões marinhas, ou rotas de migrações). Mapas digitais são considerados instrumentos preciosos por utilizadores que vão desde gestores de redes de supermercados a executivos de administração central e local, passando por responsáveis por sectores de desenvolvimento ou de produção de empresas públicas ou privadas de âmbito nacional, como telecomunicações. Em Portugal, ao tempo da investigação conducente a esta investigação, há grandes projectos com financiamento público, mas os produtos finais ainda não estão disponíveis ou não se adequam ao nosso objectivo, o que nos obrigou a produzir um mapa digital da RAM, como adiante detalharemos.

2. O nosso objectivo inicial foi a exploração de um banco de dados de doentes cardíacos da RAM que tem vindo a ser coligido pelos serviços de Cardiologia do Centro Hospitalar do Funchal.

No **Capítulo I** procedemos a uma descrição detalhada do

banco de dados, e em Anexo se apresenta uma *diskette* com o estado do referido banco de dados em 91.12.31, data em que procedemos às derradeiras actualizações de resultados. No referido Capítulo I apresentamos também uma análise sucinta dos dados, usando técnicas bem conhecidas (análise de sobrevivência, testes de independência ou de homogeneidade). Não reclamamos qualquer originalidade nos métodos estatísticos aí utilizados, obviamente. Note-se no entanto que as palavras iniciais sobre o valor explanatório dos gráficos são também válidas no que refere exemplos e "*case studies*", e que as dificuldades e limites de utilização de instrumentos teóricos são frequentemente o fermento de novos desenvolvimentos. Em particular, no caso que nos ocupa, defrontámos o problema de haver o cruzamento de muitas variáveis, de que decorre quase inevitavelmente a observação de zeros, amostrais ou essenciais, e de valores esperados muito baixos, que deixam larga dúvida sobre a utilização de resultados assintóticos.

Como por outro lado os novos desenvolvimentos de algoritmos exactos (*XactStat*, v.g.) por si só não são solução (quando a dimensão da amostra cresce deixam de ser implementáveis nos computadores para que o referido *software* existe, sem que necessariamente do crescimento da amostra decorra que estamos em condições de recorrer aos resultados assintóticos) no caso de bancos de dados tão extensos como aqueles de que dispomos, fomos inevitavelmente levados a investigar questões de eficiência e de robustez, e mais precisamente de "liberalidade" das técnicas estatísticas vulgarmente utilizadas, como adiante se verá.

No **Capítulo II** descrevemos algum do trabalho que tivémos de enfrentar na construção de um mapa digital da RAM, ainda que devamos ao CEAUL o ter disponibilizado para nossa utilização *software* excelente, sem o qual a tarefa teria excedido as nossas possibilidades. Ainda que não sobrecarreguemos o texto com informação excessiva — que nos pareceu irrelevante incluir —, um leitor atento decerto se apercebe que este empreendimento

é tão moroso e ingrato como o trabalho de simulação, e que muitas vezes se resume em poucas linhas muito trabalho efectivo. A este respeito, permita-se-nos uma breve anedota:

> Um dia o pobre Rato, sentindo-se incapaz de continuar a enfrentar as constantes perseguições do Gato, foi consultar o senhor Mocho, que todos sabiam ser um estimável sábio. Este, depois de ouvir o problema, descobriu imediatamente a solução, e aconselhou: "Transformas-te em Cão, e passas tu a perseguir o Gato!"
>
> O Rato ia retirar-se muito agradecido, quando teve uma pequena hesitação, e perguntou: "Senhor Mocho, como é que me transformo em Cão?"
>
> Ao que o Mocho respondeu, sobranceiro: "Nunca me ocupo com detalhes de implementação."

Como seria ousadia considerar-me sábia, o Capítulo II corre o risco de ser considerado maçador por nele ter tido que relatar os detalhes de implementação que tão longamente me ocuparam.

No **Capítulo III** retomamos a análise dos dados, agora de um ponto de vista mais exploratório, e justificando a necessidade de considerar uma dimensão espacial, que melhor os enquadra e explica. De facto, a evolução demográfica de zonas de povoamento recente, mormente de zonas "remotas" como a RAM que só a recente evolução dos transportes colocou de facto na "aldeia global", leva a que se mantenham ainda muitas características marcadamente locais, nomeadamente uma certa estabilidade do *stock* genético, e permanência de hábitos alimentares. Assim, como discutiremos em maior detalhe no referido capítulo, dispomos de um banco de dados sobre um "laboratório" natural que alia as vantagens de estrutura simples (as delimitações administrativas estão aliadas a delimitações geográficas e etnográficas) e diversidade reduzida, ainda que marcada.

Introdução

Muito das técnicas utilizadas na análise de tabelas de contingência gravitam em torno da distribuição amostral de estatísticas de teste que operam apenas sobre valores observados e valores esperados — isto é, na visão abrangente de Berkson (1980), estatísticas de qui-quadrado. É conhecimento comum que sob hipóteses muito "permissivas" (em termos vagos — amostragem poissoniana) a distribuição condicional do número de observações por grupo, condicional na dimensão da amostra e totais marginais, tem distribuição multinomial, e estatísticas tais como

(I.1.a) $$\chi^2_P = \sum_i ((o_i - e_i)^2 / e_i)$$ (Pearson)

(I.1.b) $$\chi^2_I = \sum_i ((o_i - e_i)^2 / o_i)$$ (Neyman)

(I.1.c) $$\chi^2_\lambda = 2 \sum_i o_i \ln(o_i / e_i)$$ (verosimilhança)

(I.1.d) $$\chi^2_I = 2 \sum_i e_i \ln(e_i / o_i)$$ (Kullback)

(I.1.e) $$\chi^2_I = \sum_i np_i q_i (\mathrm{lgt}\ p_i - \mathrm{lgt}\ P_i)^2$$ (logit)

têm distribuição assintótica de qui-quadrado, com um número de graus de liberdade igual ao número de parcelas menos um, menos o número de valores ajustados a partir dos dados amostrais. É porventura desejável referir desde já que aquelas estatísticas podem ser imersas na família $\{ I^\lambda ; \lambda \in \mathbb{R} \}$, definido-se

$$(I.2) \qquad 2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_i X_i \left\{ \left(\frac{X_i}{E_i} \right)^{\lambda} - 1 \right\}$$

sendo as quatro primeiras os casos particulares $\lambda = 1$, $\lambda = -2$, $\lambda = 0$, $\lambda = -1$, respectivamente.

No **Capítulo IV** abordamos eventuais vantagens de utilizar como estatística de teste, na análise de tabelas de contingência, estatísticas desta família com um valor do parâmetro λ diferente dos atrás referidos. Começamos por observar que do ponto de vista assintótico a escolha de λ é irrelevante, visto que a distribuição assintótica é sempre a distribuição de qui-quadrado com o número de graus de liberdade atrás referido.

Porém, do ponto de vista prático não temos exactamente a mesma situação, e na literatura estatística recente abundam exemplos de circunstâncias em que "transformações-potência" dos dados (Hoaglin et al, 1993; Tukey, 1957) são recomendáveis. De facto, ainda que a distribuição limite seja a mesma, a velocidade de convergência para a distribuição limite pode ser bem diversa. É bem conhecido, no âmbito de convergência de extremos normados, a existência de fenómenos de "penultimate approximation" (Fisher e Tippett, 1928; Gomes, 1978; Gomes e Pestana, 1984), e o exemplo de Haldane e Jayakar (1963) de que uma sucessão de máximos de quis-quadrados convenientemente normados converge mais rapidamente para a distribuição limite Gumbel do que a sucessão de máximos das normais de que são o quadrado. No Capítulo IV abordamos esta problemática — para que infelizmente não conseguimos uma solução completa.

O nosso intuito é proceder ao desenvolvimento da função característica da variável aleatória (I.2), considerando que a distribuição do vector $\underset{\sim}{O}$ (o número de observações por grupo condicional na dimensão da amostra e totais marginais) é multinomial.

Uma vez que a expressão da função característica de uma função não linear de uma variável aleatória não tem expressão simples (e se pode mesmo revelar um problema intransponível — não mais há que recordar que ainda hoje não é conhecida a expressão analítica da função característica da lognormal, e que as demonstrações independentes de Thorin (1977) e de Bondesson (1979) da sua divisibilidade infinita depende de resultados sofisticados sobre o comportamento de funções analíticas na vizinhança da origem, associados aos teoremas Tauberianos, seleccionámos dos muitos resultados parcelares que trabalhámos o de exposição mais simples, esperando que algum continuador consiga resultados mais gerais.

Recorrendo a técnicas de cálculo diferintegral (isto é, a integrais e derivadas de ordem não inteira), e trabalhando sobre a definição que do nosso ponto de vista mais se adequava ao resultado em vista, procurámos determinar os momentos fraccionários de uma binomial genérica (mais geralmente, damos a expressão genérica do momento de ordem α em termos de derivadas fraccionárias da função característica no ponto 0). Com base nestes resultados, e extensões diferintegrais da expansão em série de Taylor, obtivémos uma expansão em série da função característica de uma potência, não necessariamente inteira, de uma variável aleatória Binomial. Os resultados, ainda que parcelares, parecem-nos de qualquer forma dignos de registo. Em muitas situações, dadas as relações simples que se podem estabelecer entre "auto-normalização" (no sentido de Logan, Mallows, Rice e Shepp, 1973), desenvolvemos o nosso raciocínio, para maior simplicidade de exposição, em termos de transformações-potência sobre multinomiais, em lugar de trabalharmos com a expressão (I.2), que se torna excessivamente pesada.

É nossa intenção investigar se a convergência da expressão (I.2) para $(1-2it)^{-\nu/2}$, a função característica do qui-quadrado limite, é mais rápida no caso de tomarmos potências dos valores observados.

Finalmente, e ainda no intuito de investigar se porventura

existe(m) valor(es) λ fraccionários tais que a expressão (I.2) apresente vantagens em relação a alguma das formas mais tradicionais da estatística de qui-quadrado, abordaremos a problemática do comportamento das estatísticas de teste face a violações das hipóteses teóricas, nomeadamente sobre a distribuição da distribuição parente. Adoptamos a terminologia "teste conservador" e "teste liberal" introduzida por Benjamini (1983), e investigaremos as referidas características quando optamos por uma transformação-potência como alternativa à estatística clássica do qui-quadrado. A desiribilidade de usar a expressão (I.2) com valores fraccionários de λ é eventualmente justificada em termos dos limites de Chernoff (Chernoff, 1952; Kraft, 1955) para o erro total, para cujo cálculo recorremos aos produto interno de Kakutani e ao integral de Hellinger.

Referências:

Agrestti, A. (1984) *Analysis of Ordinal Categorical Data.* Wiley, New York.

Atlas Pro—Geographic Data Analysis and Presentation. Strategic Mapping, Inc.

Benjamini, Y. (1983) *Is the t Test Really Conservative When the Parent Distribution is Long-Tailed?* , JASA, 78, pp. 645-654.

Berkson, J. (1980) *Minimum Chi-Square, not Maximum Likelihood!* , The Annals of Statistics, 8, pp. 457-487.

Bishop, Y., Fienberg, S. e Holland, P. (1975) *Discrete Multivariate Analysis*. The MIT Press, London.

Bondesson, L. (1979) *A General Result on Infinite Divisibility.* Ann. Probability, 7, pp. 965-979.

Canto e Castro, L. (1992) *Sobre a Teoria Assintótica de Extremos*. Tese de Doutoramento, F.C.U.L. .

Cesariny de Vasconcelos, *Planisfério e outros poemas*, Guimarães Editores

Cressie, N. (1991) *Statistics of Spatial Data*. Wiley, New York.

Chernoff, H. (1952) *A measure of asymptotic efficiency for tests of a hipothesis based on the sum of observations*. Annals of Mathematical Statistics, 23, pp. 493-507.

Efron, B. (1969) *Student's t Test under Symmetry Conditions*, JASA, 63, pp. 1278-1302.

Everitt, B. (1977) *The Analysis of Contingency Tables*. Chapman and Hall, London.

Fisher, R. e Tippett, L. (1928) *Limiting forms of the frequency distribution of the largest or smallest member of a sample*. Proc. Camb. Phil. Soc. 24, 180.

Gomes, I. (1978) *Some Probabilistic and Statistical Problems in Extreme Value Theory*. Tese de Doutoramento, Universidade de Sheffield.

Gomes, I. e Pestana, D. (1984) *Domains of Attraction and Penultimate Behaviour*. Abs. 16th European Meeting Statisticians, Marburg.

Goodman, L. (1968) *The Analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries*. J.A.S.A., 63, pp. 1091-1131.

Goodman, L. (1970) The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications. J.A.S.A., 65, pp. 226-256.

Goodman, L. (1985) *The 1983 The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries*. The Annals of Statistic, 13, pp. 10-69.

Grant, N. (1992) *Grande Atlas dos Descobrimentos*. Editora Civilização.

Haldane, J. e Jayakar, S. (1963) *The distribution of extremal values in samples from a normal distribution*. Biometrika, 50, pp. 89-94.

Hoaglin, D., Mosteller, F. e Tuckey, J. (1993) *Análise Exploratória de Dados. Técnicas Robustas*. Coleccção Novas Tecnologias, Estatística, Vol.1, Edições Salamandra.

Hotteling, H. (1961) *The behavior of some standard statistical tests under nonstandard conditions*. Proc. 4th Berkeley Symposium, I, pp. 319-359.

Huff, D. (1954) *How to Lie with Statistics*. Pelican Books, London.

Kraft, C. (1955) *Some conditions for consistency of statistical procedures*. University of California Publications in Statistics, University of California Press, California.

Leland, W. (1989) *The System for Statistics*. Evanston, IL: SYSTAT, Inc.

Logan, B., Mallows, C., Rice, S. e Shepp, L. (1973) *Limit Distributions of Self-Normalized Sums*. The Annals of Probability, 1, pp. 788-809.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press, London

Matheron, G. (1965) *Les Variables Regionalisés et leur Estimation*. Masson et Cie., Paris.

Matheron, G. (1971) *The Theory of Regionalized Variables and its Applications*. Ecole Nationale Superieure des Mines de Paris.

Mehta, C., Patel, N. (1989-91) *StatXact User Manual*. Cytel Software Corporation, Cambridge, Massachussets.

Mosteller, F. e Tukey, J. (1977) *Data Analysis and Regression*. Reading MA: Addison-Wesley.

Saville, D. and Wood, G. (1980) *Statistical Methods: The Geometric Approach*

Springer Verlag.

Thorin, O. (1977) *On the Infinite Divisibility of the Lognormal Distribution*. Scand. Actuarial J., pp. 121-148.

Tufte, E. R. (1983) *The Visual Display of Quantitative Information* . Graphics Press, U.S.A.

Tufte, E. R. (1990) *Envisioning Information*. Graphics Press, Cheshire, Connecticut.

Tukey, J. (1957) *On the comparative anatomy of transformations*. Annals of Mathematical Statistics, 28, pp. 602-632.

Tukey, J. (1977) *Exploratory Data Analysis*. Reading MA: Addison-Wesley.

Upton, G. (1978) *The Analysis of Cross-Tabulated Data*. Wiley, New York.

Referências sobre Atlas Históricos

Baines, J., Málek, J. (1984) *Atlas of Ancient Egypt*. Phaidon, Oxford.

Blunden, C., Elvin, M. (1983) *Cultural Atlas of China*. Phaidon, Oxford.

Chadwick, H., Evans, G. R. (1987) *Atlas of the Christian Church*. Phaidon, Oxford.

Collcut, M., Jansen, M., Kamakura, I. (1988) *Cultural Atlas of Japan*. Phaidon, Oxford.

Cornell, T., Mathews, J. (1982) *Atlas of the RomanWorld*. Phaidon, Oxford.

de Lange, N. (1984) *Atlas of the Jewish World*. Phaidon, Oxford.

Levy, P. (1980) *Atlas of the Greek World*. Phaidon, Oxford.

Mathews, D. (1983) *Atlas of Medieval Europe.* Phaidon, Oxford.

Milner - Gulland, R., Dejevstay, N. (1989) *Atlas of Russia*. Phaidon, Oxford.

Murray, J. (1981) *Cultural Atlas of Africa*. Phaidon, Oxford.

Ribeiro, O., Lautensach, H. e Daveau, S. (1989) *Geografia de Portugal*, Vol. III, Edições João Sá da Costa, Lisboa.

Robinson, F. (1982) *Atlas of the Islamic World Since 1500.* Phaidon, Oxford.

Capítulo I

Análise de Sobrevivência de Doentes Cardíacos. — Exploração do Banco de Dados do Centro Hospitalar do Funchal

1 — Introdução

O Serviço de Cardiologia do Centro Hospitalar do Funchal criou um ficheiro de dados relativos a todos os doentes que recorreram a este Serviço a partir do ano de 1983, onde constam as informações que os médicos consideraram relevantes a uma análise estatística da doença cardíaca na Região Autónoma da Madeira.

Esta base de dados é constituída por observações de 94 variáveis respeitantes a 1835 doentes. As variáveis contêm informações pessoais do doente (Idade, Sexo, Profissão, etc.), a história clínica do doente anterior à crise cardíaca (Hipertensão, Diabetes, Tabaco, Angina Prévia, etc.), informações sobre o estado actual do doente (Diagnóstico, Killip, Morte, Sobrevida, etc.), informações sobre os tratamentos ministrados ao doente (DNI, Aspirina, Betablock, etc.), complicações da crise (Complicações Cardíacas Tardias, Complicações Não Cardíacas, etc.), e informações semelhantes referentes a crises cardíacas anteriores. Em apêndice (Apêndice I) apresentamos uma listagem e descrição das mesmas.

O último ano de observações que considerámos é o de 1991, data em que iniciámos este estudo.

Numa primeira fase iniciámos uma análise estatística considerando os oito diagnósticos constantes da variável DIAGNOST — angina instável, enfarto agudo do miocárdio, arritmia, BAV, EAP, TAMPON, embolia pulmonar, OUTRO. Mas, posteriormente, foi-nos sugerido pelos especialistas de cardiologia do Centro Hospitalar do Funchal que o estudo incidisse apenas sobre os dois primeiros diagnósticos por serem os de maior gravidade clínica, e por forma a que não houvesse dispersão de informação e a que as respectivas conclusões fossem evidenciadas.

Assim, a base de dados de que dispomos neste estudo é constituída por observações de 94 variáveis referentes a 1070 doentes que tiveram uma crise de Enfarto Agudo do Miocárdio (EAM) ou de Angina Instável (ANG. INST.) entre os anos de 1983 e 1991.

A primeira abordagem estatística que fizémos a esta base de dados e que constitui o presente capítulo foi a Análise da Sobrevivência. O uso de modelos probabilísticos para dados de sobrevivência é de grande importância nas ciências biomédicas, não só por descrever a forma do tempo de sobrevivência, como por frequentemente levantar novas questões sobre as doenças em estudo.

A variável Sobrevida, constante do ficheiro de dados, mede o tempo que decorre desde o diagnóstico de uma crise de enfarto ou de angina até à morte. Estes tempos estão sujeitos a variações aleatórias e, como qualquer variável aleatória podemos ajustar-lhe uma distribuição. A distribuição da Sobrevida é descrita por três funções que são matematicamente equivalentes, mas que, na prática, podem ser utilizadas para ilustrar diferentes aspectos dos dados: a função de sobrevivência, a função densidade e a função hazard. Estimamos estas funções e evidenciamos a necessidade de fazer análises estatísticas separadas para os dois

diagnósticos referidos anteriormente.

É de salientar que optámos por não utilizar toda a informação disponível, isto é, os dados censurados correspondentes aos doentes que ainda sobreviviam na data em que iniciámos o estudo não foram considerados, porque os especialistas de cardiologia consideravam que era provável que houvesse casos de doentes que já tivessem falecido sem que esta informação tivesse sido actualizada no ficheiro, o que poderia destorcer os resultados da análise dadas as características da variável Sobrevida.

Nos últimos anos, a identificação de factores de risco e de variáveis prognóstico relacionados com a sobrevida tem-se tornado importante. Incluímos também neste capítulo um estudo sobre estas variáveis. Como as variáveis prognóstico não foram, *a priori*, completamente especificadas, decidimos agrupar as variáveis constantes do ficheiro de dados em dois grupos — o grupo das variáveis que dizem respeito ao tratamento ministrado ao doente e o grupo formado pelas variáveis que incluem informações pessoais e a história clínica do doente.

O primeiro grupo não foi, no entanto, analisado por ter apenas interesse, e até sentido, ser realizado com a orientação específica dos médicos cardiologistas e tendo em consideração que não são variáveis prognóstico gerais.

Os resultados desta análise levantam a questão que deu origem a esta Tese— é relevante investigar a atribuição de uma dimensão espacial à doença coronária.

2 — Estimação da função de Sobrevivência, da função Densidade e da função Hazard

Estimativas obtidas da utilização de métodos não-paramétricos e gráficos podem ser úteis na escolha de uma distribuição teórica adequada aos dados de sobrevivência. Assim, optámos por utilizar métodos não-paramétricos na análise dos dados que possuímos, antes de tentar o ajustamento a uma distribuição teórica.

Os métodos não-paramétricos são menos potentes do que os métodos paramétricos quando os tempos de sobrevivência seguem uma distribuição teórica, mas mais robustos quando não conhecemos uma distribuição teórica adequada.

No que se segue representaremos por T o tempo de sobrevivência.

2.1 — *Função de Sobrevivência*

A função de sobrevivência, representada por S(t), é definida como sendo a probabilidade de que um doente sobreviva para além de t, isto é, a cauda direita da distribuição do tempo de vida

$$S(t) = P(T>t) .$$

Tendo em consideração a definição de função de distribuição F(t) de T, poderemos escrever

$$S(t) = 1-F(t) .$$

Chamamos curva de sobrevivência ao gráfico de S(t), que é uma forma extremamente útil de representarmos o comportamento da sobrevivência.

Estimámos a função de sobrevivência como sendo a proporção de doentes que sobrevivem para além de t.

TABELA 2.1

Sobrevida t (dias)	Nº doentes com Sobrevida > t	$\hat{S}(t)$
[0 , 100 [	275	1
[100, 200 [	121	0.44
[200 , 300 [	103	0.3745
[300 , 400 [	97	0.3527
[400 , 500 [	84	0.3055
[500 , 600 [	73	0.2655
[600 , 700 [	66	0.24
[700 , 800 [	58	0.2109
[800 , 900 [	51	0.1855
[900 , 1000 [	45	0.1636
[1000 , 1100 [	41	0.1491
[1100, 1200 [	37	0.1345
[1200 , 1300 [	33	0.12
[1300 , 1400 [	28	0.1018
[1400, 1500 [	22	0.08
[1500 , 1600 [	18	0.0655
[1600 , 1700 [	17	0.0618
[1700 , 1800 [	13	0.0473
[1800 , 1900 [	11	0.04
[1900 , 2000 [	10	0.0364
[2000 , 2100 [	8	0.0291
[2100 , 2200 [	4	0.0145
≥2200	3	0.0109

As duas primeiras colunas da Tabela 2.1 referem-se aos

valores observados da variável Sobrevida nos 275 doentes que sofreram uma crise cardíaca entre os anos de 1983 e 1991. Os tempos de sobrevivência foram agrupados em intervalos de 100 dias e a função de sobrevivência estimada, que apresentamos na mesma tabela, foi calculada no início de cada intervalo.

Observando a Fig.2.1, onde está representada a curva de sobrevivência estimada, podemos verificar que, pela forma abrupta como decresce, o tempo de sobrevivência a uma crise destas doenças é, frequentemente, curto.

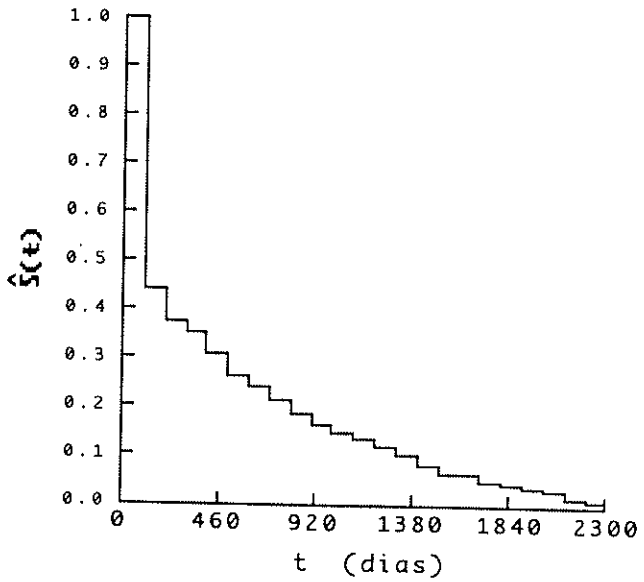


Fig.2.1 - Curva de sobrevivência estimada para as doenças em estudo

2.2 — *Função Densidade*

Parece razoável admitir que a Sobrevida seja uma variável aleatória absolutamente contínua. A sobrevida T tem uma função densidade de probabilidade definida por

$$f(t) = \lim_{\Delta t \to 0} P\{\text{de um doente morrer no intervalo } (t\ ,\ t + \Delta t)\} / \Delta t\ .$$

A função densidade pode ser grosseiramente estimada como sendo a proporção de doentes com sobrevida pertencente a um intervalo de tempo por unidade de tempo do intervalo. Está muito para além do escopo desta dissertação sequer abordar os métodos sofisticados de estimação de densidades usando teoria de kernels ou outra.

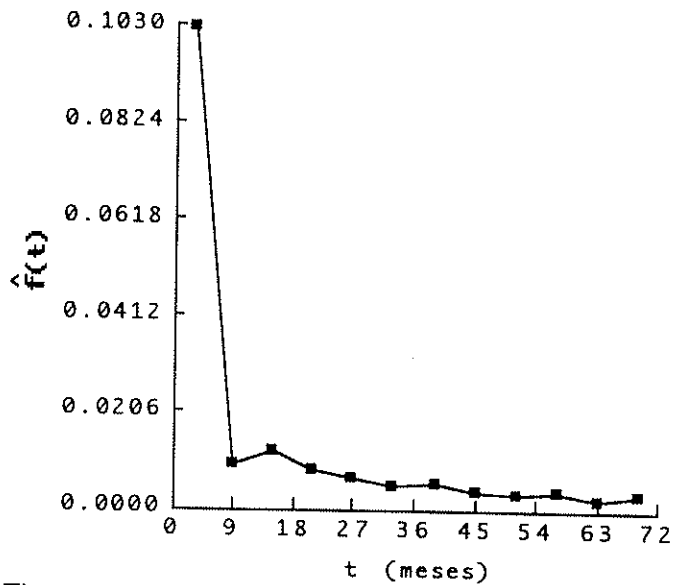


Fig. 2.2 - Função densidade estimada para as doenças em estudo

Na Fig.2.2 está representada a função densidade estimada. A proporção de mortes é grande nos primeiros seis meses após a crise e, a partir daí decresce apresentando pequenas flutuações.

As classes consideradas no agrupamento dos dados têm amplitude de aproximadamente 6 meses (180 dias) apesar da Sobrevida ter sido originalmente registada em dias.

TABELA 2.2

Sobrevida t (meses)	Nº de mortes no intervalo	Nº de doentes vivos no ponto médio do intervalo	$\hat{f}(t)$	$\hat{h}(t)$
[0 , 6 [	169	122	0.1024	0.2309
[6 , 12 [	16	99	0.0097	0.0269
[12 , 18 [	20	78	0.0121	0.0427
[18 , 24 [	14	63	0.0085	0.037
[24 , 30 [	11	50	0.0067	0.0367
[30 , 36 [	8	41	0.0049	0.0325
[36 , 42 [	9	35	0.0055	0.0429
[42 , 48 [	6	25	0.0036	0.04
[48 , 54 [	5	17	0.003	0.049
[54 , 60 [	6	12	0.0036	0.0833
[60 , 66 [	3	10	0.0018	0.05
[66 , 72 [	5	6	0.003	0.1389
≥ 72	3	—	—	—

Foi uma escolha de algum modo arbitrária que teve em conta a obtenção de valores estimados para a função densidade

numa escala susceptível de originar uma representação gráfica aceitável e, simultaneamente, adequada aos dados que possuímos. Devido ao facto de existirem muitos doentes com Sobrevidas baixas deparámos com o problema de o valor estimado da função densidade para a primeira classe ser muito superior aos valores estimados para as restantes classes, quando a amplitude das classes era considerada em dias, originando representações gráficas inúteis.

As três primeiras colunas da Tabela 2.2 apresentam os dados dos tempos de sobrevivência referentes ao agrupamento considerado. A quarta coluna contem os valores estimados para a função densidade.

2.3 — *Função Hazard*

A função hazard do tempo de sobrevivência T, que desempenha um importante papel na análise de dados de sobrevivência, e que representaremos por h(t), é definida como sendo a probabilidade de um doente morrer durante um pequeno intervalo de tempo dado que sobreviveu até ao início do intervalo, ou seja

$$h(t) = \lim_{\Delta t \to 0} P\{\text{um doente que sobreviveu até } t \text{ morrer no intervalo } (t+\Delta t)\} / \Delta t \ .$$

Daí decorre que a função hazard pode ser expressa em termos da função de distribuição F(t) e da função densidade de probabilidade f(t):

$$h(t) = f(t) / \{1 - F(t)\} \ .$$

Utilizámos a estimativa mais conservativa da função

hazard, na qual o número de doentes que morre por unidade de tempo no intervalo é dividido pelo número de sobreviventes no ponto médio do intervalo. A última coluna da Tabela 2.2 apresenta as estimativas obtidas, $\hat{h}(t)$, para o agrupamento considerado.

Na Fig. 2.3 podemos observar a representação gráfica da função hazard estimada. Durante um período inicial o risco é grande. Passado este período, $\hat{h}(t)$ apresenta pequenas flutuações que se tornam acentuadas a partir dos 52 meses, aproximadamente, apresentando-se relativamente elevada para os doentes que sobreviveram até aos 66 meses. Assim, o prognóstico de um doente que tenha sobrevivido os primeiros 9 meses é melhor do que o de um doente que tenha acabado de ter uma crise cardíaca, se não tivermos em conta os factores prognóstico.

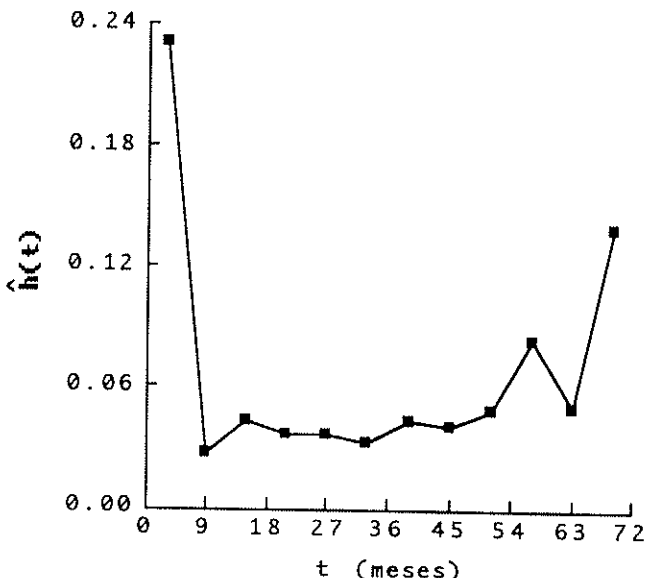


Fig. 2.3 - Função hazard estimada para as doenças em estudo

3 — Identificação de potenciais factores prognóstico

A análise de dados de sobrevivência inclui, também, o exame da relação individual de cada variável com a Sobrevida, para verificarmos a possibilidade de os dados provirem de populações dissemelhantes e para termos uma ideia preliminar sobre quais as variáveis que podem vir a revelar-se importantes no prognóstico.

Apresentamos com maior detalhe o estudo da variável Diagnóstico pela importância que tem a comparação das curvas de sobrevivência para as duas doenças consideradas.

3.1 — *Relação entre a variável Diagnóstico e a variável Sobrevida*

Com a intenção de comparar as curvas de sobrevivência para os doentes de enfarto e de angina e de verificar se estes dois diagnósticos dividiam os dados em dois grupos heterogéneos justicando análises estatísticas separadas, representámos no mesmo gráfico as correspondentes curvas de sobrevivência. É apenas uma primeira exploração impressionista, adiante complementada com uma análise objectiva.

Na Fig.3.1 apresentamos as curvas de sobrevivência estimadas para os dois diagnósticos e, na Fig.3.2 as respectivas funções de distribuição empíricas. As diferenças acentuam-se nos tempos de sobrevivência mais baixos. O doente que sofreu uma crise de angina apresenta mais frequentemente uma sobrevida superior a 100 dias do que o doente que sofreu uma crise de enfarto.

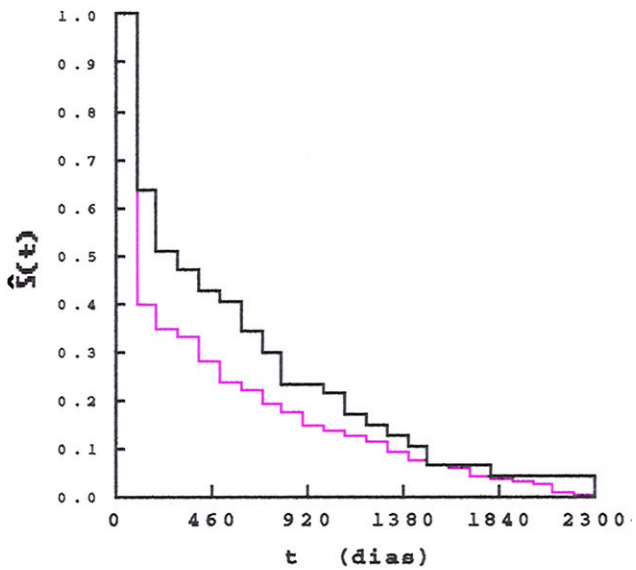


Fig.3.1 - Curvas de sobrevivência estimadas para os dois diagnósticos

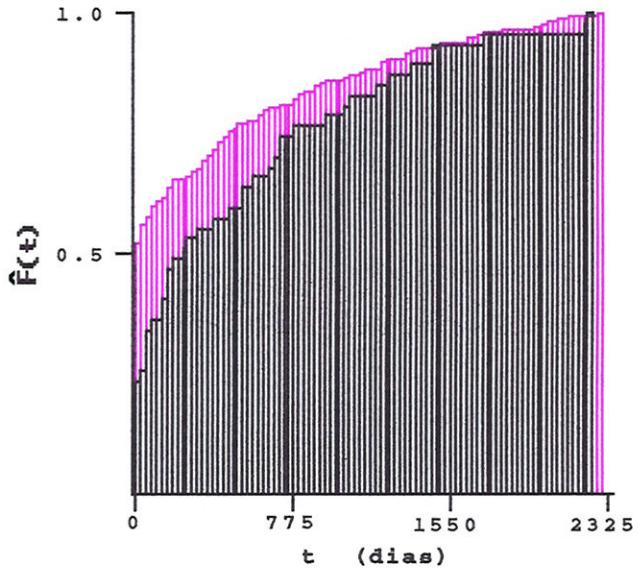


Fig.3.2 - Funções de distribuição empíricas da Sobrevida relativamente aos dois diagnósticos

A curva de sobrevivência estimada para os doentes de enfarto é mais abrupta do que a obtida para os doentes de angina o que significa que os primeiros apresentam um tempo de sobrevivência a uma primeira crise inferior ao dos doentes de angina até cerca dos 1500 dias, a partir dos quais as sobrevidas a uma primeira crise tendem a aproximar-se em ambos os diagnósticos.

Na Fig. 3.3 estão representadas as funções densidade estimadas referentes aos tempos de sobrevivência em cada um dos diagnósticos.

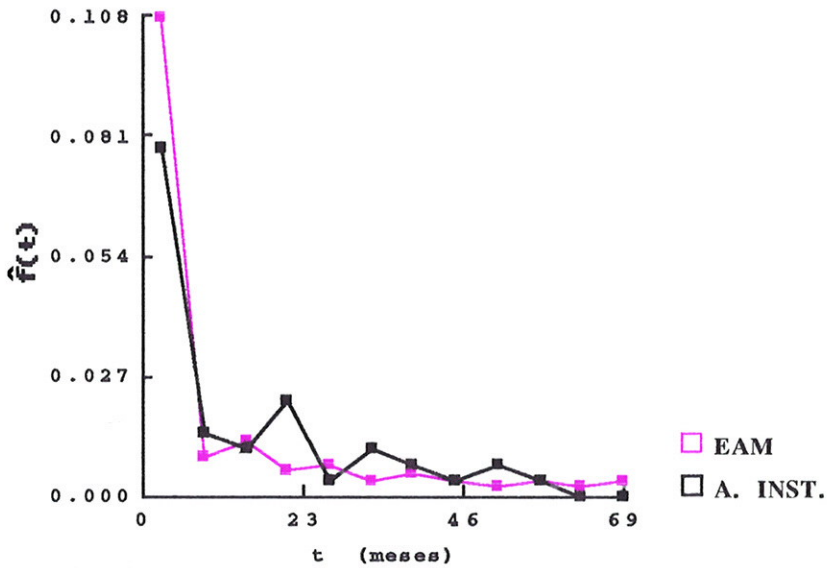


Fig. 3.3 - Funções densidade estimadas para os dois diagnósticos

A forma da função densidade estimada para a sobrevida dos doentes de enfarto é muito semelhante à estimada para a

globalidade dos doentes. Este facto não é surpreendente por o número de mortes por enfarto ser de 228 enquanto que o número de mortes por angina é de 47, levando assim a que os dados relativos ao enfarto dominem as estimativas obtidas para as funções apresentadas por não existir uma diferença acentuada no forma das funções densidade estimadas para os dois diagnósticos.

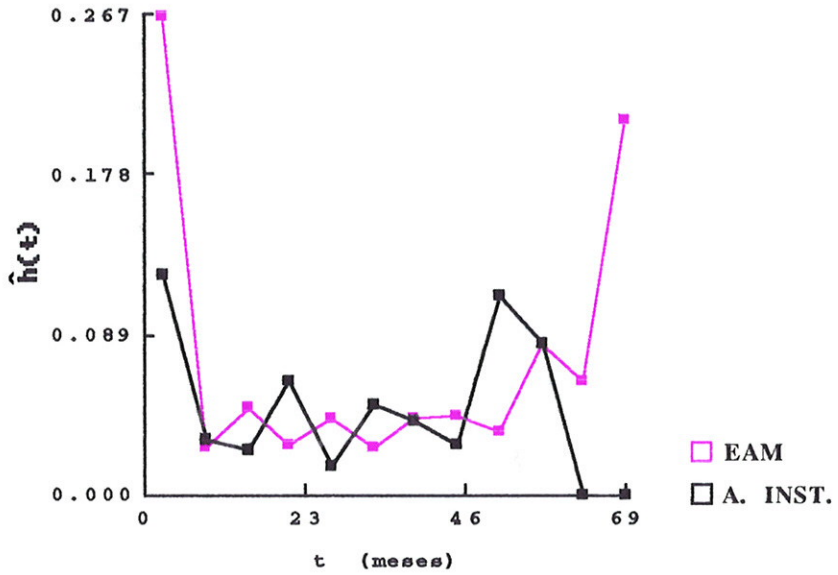


Fig. 3.4 - Funções hazard estimadas para os dois diagnósticos

As funções hazard estimadas para cada um dos diagnósticos têm formas diferentes como se pode verificar pelos gráficos apresentados na Fig. 3.4. A função hazard estimada para a Sobrevida dos doentes de enfarto é semelhante à estimada para a globalidade dos doentes. No entanto, a função hazard estimada para os doentes de angina tem uma forma diferente, apresentando, para a partir dos 9 meses, flutuações em sentido oposto às

flutuações apresentadas pelas funções hazard estimadas acima referidas. Podemos, também, observar que tanto a função densidade estimada como a função hazard estimada para os tempos de sobrevivência dos doentes de angina apresentam flutuações mais acentuadas nos valores centrais, do que a função hazard estimada para os doentes de enfarto.

As sobrevivências correspondentes aos dois diagnósticos foram comparadas usando o teste de Mann-Whitney. O valor obtido para a estatística de teste levou à rejeição da hipótese de igualdade das funções de distribuição dos tempos de sobrevivência para os dois diagnósticos, conforme podemos observar na Tabela 3.1. Note-se que em rigor a aplicação do teste apenas dá indicação precisa sobre a localização das distribuições, mas que os estudos complementares a que procedemos revelam também inequívocas dissemelhanças na forma das distribuições.

As medianas correspondentes aos dois diagnósticos estão apresentadas na Tabela 3.1, assim como os valores observados para os 1° e 3° quartis. Como se pode observar a sobrevida de um doente de angina apresenta valores mais elevados que a sobrevida de um doente de enfarto.

TABELA 3.1

		N° de doentes	N° de mortes	Me (dias)	Q1	Q3	R	Est. de Teste	Valor p
Diagnóstico	Ang. Inst.	304	47	230	27	775	7923	6795	0.004
	E.A.M.	766	228	18.5	2	451	30027		

3.2 — *Estudo da relação das restantes variáveis com a variável Sobrevida*

De acordo com o que concluímos no parágrafo anterior,

decidimos analisar separadamente para cada diagnóstico as relações entre as restantes variáveis e a variável Sobrevida uma vez que as potenciais variáveis prognóstico podem não ser as mesmas para os doentes de angina e para os doentes de enfarto. Neste caso, não teria mesmo sentido pensarmos numa "estratificação" da função de sobrevivência por níveis da variável Diagnóstico para a análise dos efeitos simultâneos das potenciais variáveis prognóstico nos tempos de sobrevivência e avaliação da importância relativa das mesmas.

A Tabela 3.2 refere-se aos resultados obtidos para os tempos de sobrevivência dos doentes de enfarto. Apresentamos, para cada nível de cada variável, o número de doentes observados, o número de mortes observadas, a mediana e os 1° e 3° quartis dos tempos de sobrevivência observados, e o resultado do teste de comparação das funções de sobrevivência através do valor aproximado do nível de significância a que a hipótese da igualdade das localizações das distribuições seria rejeitada. Utilizámos o teste de Mann-Whitney nas variáveis com dois níveis com m e n observações, cuja estatística de teste é

$$U = T - n(n+1)/2$$

em que $T = \sum_{j=1}^{n} R_j$, onde R_j denota o *rank* ou ordem da j-ésima observação na amostra combinada, e, representando por N o número total de observações nos dois grupos, $\mu_T = n(N+1)/2$ e $\sigma_T^2 = \frac{mn(N+1)}{12}$ no caso de não existirem observações empatadas. Se existirem empates a variância corrigida é

$$\sigma_T^2 = \frac{mn(N+1)}{12} - \left(\frac{mn}{12N(N-1)} \sum_{i=1}^{g} K_i \right),$$

em que g representa o número de grupos de empates, $K_i = (\tau_i - 1)\tau_i(\tau_i + 1)$, onde τ_i é o número de observações empatadas no i-ésimo grupo de empates.

Utilizámos o teste de Kruskal-Wallis nas variáveis com mais de dois níveis; neste caso a expressão da estatística de teste é

$$T = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

quando não existem observações repetidas, sendo k o número de grupos considerados em H_0, n_j o número de observações no grupo j, R_j a soma das ordens atribuídas na amostra combinada às observações do grupo j e N o número total de observações. Se existirem observações repetidas a expressão da estatística de teste corrigida é

$$T = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum_{j=1}^{g} K_j}{N^3 - N}}$$

em que g é o número de grupos de empates e $K_j = \tau_j^3 - \tau_j = (\tau_j - 1)\tau_j(\tau_j + 1)$, sendo τ_j o número de observações no j-ésimo grupo de empates (o teste de Kruskal-Wallis é, obviamente, uma extensão a k amostras do teste de Mann-Whitney para duas amostras). A escolha destes testes de entre os vários possíveis para dados de sobrevivência teve em linha de conta a sua acessibilidade no *software* de que dispunhamos.

TABELA 3.2

Variável		N° de doentes	N° de mortes	Me (dias)	Q1	Q3	R	Est. de Teste	Valor p
Idade	C0	25	0	—	—	—	—		
	C1	328	72	161.5	3	885	9333		*
	C2	228	75	29	4	433	8972	10.436	0.015
	C3	125	55	5	1	293	5306.5		
	C4	60	26	8	2	208	2494.5		
Sexo	M	503	120	63	3	634	15012		*
	F	263	108	8	1	292	11094	7752	0.01
Ang.Prev.	Não	464	126	13	2	433	14148.5	6147.5	0.573
	Sim	301	102	44	2	512	11957.5		
N°EAMPrev.	Não	696	196	18	2	451	22537.5		
	≥1	67	32	18	1	361	3568.5	3231.5	0.782
Hipert.Prev.	Não	395	110	19.5	2	419	12429		
	Sim	367	116	15	2	434	13222	6324	0.909
Diabetes	Não	567	155	20	2	490	17916.5		
	Sim	197	73	14	2	419	8189.5	5826.5	0.716
Tabaco	Não	143	40	10	1	162	1894		
	Sim	198	53	361	6	1028	3530		*
	<20							8.623	0.013
	Sim	159	27	179	8	918	1836		
	>20								
Arrit.Isq.	Não	534	132	72	2	566	16125		*
	Sim	228	95	8	2	367	9753	7347	0.027
Loc.EAMAct.	Ant.	267	90	23.5	2	488	7774.5		
	Inf.	321	83	35	3	787	7276.5	3679.5	0.866
Killip	I	475	77	85	6	918	7286		
	II	162	72	163	9	634	8383.5		*
	III	74	35	10	1	208	5038	14.486	0.002
	IV	52	43	1	0	2	5170.5		

Eliminámos do estudo de cada variável os casos em que faltavam informações. Também não apresentamos o estudo das variáveis com poucas observações e em que não fosse aceitável agrupar os respectivos níveis.

As variáveis que aparecem como podendo ter importância significativa para o prognóstico da sobrevivência são: Idade, Sexo, Tabaco, Arritmia Isquémica e Killip.

Relativamente aos 228 doentes com Sobrevida inferior a 2286 dias (aproximadamente 6 anos e 3 meses), existem diferenças significativas entre as localizações das distribuições da Sobrevivência para as diferentes classes etárias consideradas — <43, $[43, 66[$, $[66, 75[$, $[75, 80[$, ≥ 80. Não foram registadas mortes no grupo etário com menos de 43 anos. A sobrevida mediana diminui conforme a idade avança, excepto no caso correspondente aos doentes com idade superior ou igual a 80 anos. Neste caso, a sobrevida mediana aumenta ligeiramente em relação à classe etária anterior, sendo, no entanto, o 3° quartil inferior ao da classe anterior.

O enfarto agudo do miocárdio é uma doença que afecta mais os homens do que as mulheres (embora seja do âmbito da análise exploratória, podemos adiantar que a proporção de mulheres na população afectadas pela doença é inferior à correspondente proporção de homens, mas a proporção vai aumentando conforme a idade avança, atingindo um valor próximo ao dos homens na classe etária mais avançada). No entanto, como podemos observar na Tabela 3.2, a Sobrevida mediana das mulheres é consideràvelmente inferior à dos homens. Atendendo ao que referimos sobre a incidência da doença nestes dois grupos, impõe-se verificarmos se a variável idade será, neste caso, uma variável *confounding*, ou seja, se as sobrevidas dos homens são idênticas às das mulheres em cada classe etária considerada.

Na Tabela 3.3 apresentamos os resultados obtidos, que fornecem evidência de que não existem diferenças significativas

entre as localizações das distribuições de sobrevivência dos homens e das mulheres em cada classe etária considerada. Eliminamos, pois a variável Sexo do grupo de variáveis prognóstico.

TABELA 3.3

Classe Etária		**Nº de mortes**	**R**	**Est. de Teste**	**Valor p**
C1	M	49	1900.5		
	F	23	727.5	675.5	0.176
C2	M	36	1423.5		
	F	39	1426.5	757.5	0.556
C3	M	28	859.5		
	F	27	680.5	453.5	0.201
C4	M	7	109.5		
	F	19	241.5	81.5	0.383

Os resultados obtidos para a variável Tabaco, aparentemente duvidosos, são, em parte, explicados pela interacção entre esta variável e a variável Idade, como podemos confirmar pelo resultado do teste de independência do Qui-Quadrado aplicado aos dados da Tabela 3.4.

TABELA 3.4

		IDADE				
		C1	C2	C3	C4	TOTAL
TABACO	Não	33	48	40	22	143
	Sim, <20	24	20	12	2	58
	Sim, >20	15	7	3	2	27
	TOTAL	72	75	55	26	228

O valor obtido para a estatística de teste, x^2=19.056, leva-nos à rejeição, ao nível de significância de 0.05, da hipótese de o facto de um doente ser fumador, por exemplo, ser independente da respectiva idade.

Como podemos observar os doentes que apresentam uma sobrevida inferior a 2286 dias e que fumam distribuem-se pelas classes de idade mais baixas que são aquelas cujos doentes apresentam uma sobrevida maior. Apesar de não termos disponíveis estatísticas sobre o consumo de tabaco na R.A.M., pensamos que este não se afastará muito do observado na maioria dos países, indicando que a proporção de não-fumadores na população é maior do que a proporção de fumadores. Quanto à incidência da doença, observaram-se 392 doentes não-fumadores e 371 fumadores, o que nos leva a arriscarmos a afirmação de que a doença atinge mais os fumadores do que os não-fumadores. Além disso, tendo em consideração os dados da Tabela 3.4, os fumadores são atingidos numa idade mais jovem que os não-fumadores.

Consideramos importante realçar que a sobrevida mediana dos doentes que fumam mais de 20 cigarros por dia é cerca de metade da sobrevida mediana dos doentes que fumam menos de 20 cigarros por dia, evidenciando o efeito do consumo excessivo de tabaco.

Não nos foi possível desenvolver a análise da Sobrevivência por classe etária uma vez que os números de mortes nos três níveis da variável Tabaco são muito diferentes chegando a atingir valores muito baixos nas classes etárias mais avançadas. No entanto, para a classe de idade C1, a única a que aplicámos o teste de Kruskal-Wallis, obtivémos um resultado idêntico ao apresentado na Tabela 3.2. Uma análise apenas exploratória indica que para as restantes classes as diferenças nas localizações das distribuições da Sobrevivênvia deixam de ter importância.

Relativamente às 228 mortes observadas , as sobrevidas dos doentes que sofriam de arritmia isquémica é significativamente

inferior à sobrevida dos doentes que não sofriam desta perturbação. No caso desta variável, não podemos concluir que a variável Idade seja uma variável *confounding*, uma vez que as sobrevidas relativas a estes dois grupos de doentes não são idênticas em cada classe de Idade considerada. Para verificar se a incidência de doentes com arritmia isquémica é superior nos doentes não-fumadores do que nos fumadores, contribuindo assim para explicar as baixas sobrevidas observadas no doentes não-fumadores, aplicámos o teste de Mantel-Haenszel que tem sido usado em muitos estudos clínicos e epidemiológicos como um método de controlar os efeitos de variáveis *confounding*. Os dados foram estratificados pelos níveis da variável Idade e obteve-se uma sequência de tabelas 2x2, uma para cada estrato. A hipótese nula a ser testada é

$$H_0: \begin{array}{l} p_{11}=p_{12} \\ p_{21}=p_{22} \\ \vdots \\ p_{41}=p_{42} \end{array}$$

em que p_{ij} = P(sofrer arrit.isq. | nível j da v. Tabaco, estrato i) . O resultado obtido foi :

TABELA DE TABACO (LINHAS) POR ARRITISQ (COLUNAS)
PARA OS SEGUINTES VALORES:
CLASSE$ = C1

FREQUENCIAS

	1.000	2.000	TOTAL
1.000	20	12	32
2.000	26	13	39
TOTAL	46	25	71

inferior à sobrevida dos doentes que não sofriam desta perturbação. No caso desta variável, não podemos concluir que a variável Idade seja uma variável *confounding*, uma vez que as sobrevidas relativas a estes dois grupos de doentes não são idênticas em cada classe de Idade considerada. Para verificar se a incidência de doentes com arritmia isquémica é superior nos doentes não-fumadores do que nos fumadores, contribuindo assim para explicar as baixas sobrevidas observadas no doentes não-fumadores, aplicámos o teste de Mantel-Haenszel que tem sido usado em muitos estudos clínicos e epidemiológicos como um método de controlar os efeitos de variáveis *confounding*. Os dados foram estratificados pelos níveis da variável Idade e obteve-se uma sequência de tabelas 2x2, uma para cada estrato. A hipótese nula a ser testada é

$$H_0: \; p_{11}=p_{12}$$
$$p_{21}=p_{22}$$
$$\vdots$$
$$p_{41}=p_{42}$$

em que p_{ij} = P(sofrer arrit.isq. | nível j da v. Tabaco, estrato i) . O resultado obtido foi :

TABELA DE TABACO (LINHAS) POR ARRITISQ (COLUNAS)
PARA OS SEGUINTES VALORES:
CLASSE$ = C1

FREQUENCIAS

	1.000	2.000	TOTAL
1.000	20	12	32
2.000	26	13	39
TOTAL	46	25	71

TABELA DE TABACO (LINHAS) POR ARRITISQ (COLUNAS)
PARA OS SEGUINTES VALORES:
CLASSE$ = C2
FREQUENCIAS

	1.000	2.000	TOTAL
1.000	28	20	48
2.000	17	10	27
TOTAL	45	30	75

TABELA DE TABACO (LINHAS) POR ARRITISQ (COLUNAS)
PARA OS SEGUINTES VALORES:
CLASSE$ = C3
FREQUENCIAS

	1.000	2.000	TOTAL
1.000	20	20	40
2.000	9	6	15
TOTAL	29	26	55

TABLE OF TABACO (LINHAS) POR ARRITISQ (COLUNAS)
PARA OS SEGUINTES VALORES:
CLASSE$ = C4
FREQUENCIAS

	1.000	2.000	TOTAL
1.000	11	11	22
2.000	1	3	4
TOTAL	12	14	26

ESTATÍSTICA DO QUI-QUADRADO DE MANTEL-HAENSZEL = 0.265
PROBABILIDADE = 0.607

Conclui-se que não existe evidência suficiente para rejeitar a hipótese de que a incidência de doentes com arritmia isquémica é idêntica no grupo de não-fumadores e no grupo de fumadores. E, como seria então de esperar, se observarmos o que se passa com os doentes da classe de idade C1 que não sofrem de arritmia isquémica, a sobrevida dos doentes fumadores continua a ser superior à dos doentes não-fumadores. Quanto aos doentes da mesma classe etária que sofrem de arritmia isquémica, a situação altera-se, passando a não existir diferença significativa (ao nível de significância de 0.05) entre as localizações das distribuições da sobrevivência dos fumadores e dos não-fumadores. Na Tabela 3.5 apresentamos os resultados obtidos.

TABELA 3.5

IDADE C1					
	Tabaco	Nº de mortes	R	Est. de Teste	Valor p
Não	Não	20	365.5		
	Sim	26	715.5	155.5	0.02
Arrit.Isq.					
Sim	Não	12	125		
	Sim	13	200	47	0.09

A classe de enfarto (Killip), influencia a Sobrevida de uma forma pouco clara uma vez que as classes de enfarto menos graves apresentam sobrevidas mais baixas, conforme consta na Tabela 3.2. Na Tabela 3.5 podemos observar que as proporções de doentes nas classes etárias C3 e C4 com enfartos de classe III e IV são consideràvelmente superiores às mesmas proporções nos casos de

enfartos de classes I e II. O valor obtido para estatística de teste do Qui-Quadrado, x^2=23.892 leva à rejeição, ao nível de significância de 0.05, da hipótese de que a classe de enfarto é independente da classe etária a que o doente pertence.

TABELA 3.6

		KILLIP				
		I	II	III	IV	TOTAL
IDADE	C1	34	23	3	12	72
	C2	25	28	10	12	75
	C3	12	15	15	12	54
	C4	6	6	7	7	26
	TOTAL	77	72	35	43	227

A variável Idade não é, no entanto, uma variável *confounding* relativamente à classe de enfarto, uma vez que as localizações das distribuições das sobrevivências para as classes de enfarto continuam a apresentar diferenças significativas nas três primeiras classes etárias consideradas. Na última classe etária não existem diferenças significativas entre as localizações das distribuições da sobrevivência relativas às quatro classes de enfarto.

No que diz respeito às variáveis Tabaco e Arritmia Isquémica chegámos às mesmas conclusões que para a variável Idade. Tentámos estudar os efeitos das interacções entre estas variáveis nas localizações das distribuições das sobrevivências relativamente às classes de enfarto mas o número de observações torna-se extremamente reduzido. Pensamos, no entanto, que os referidos efeitos possam contribuir para a explicação dos resultados que obtivémos, dado que, considerando as poucas situações com números de observações que permitiam uma análise estatística, pudémos observar que, por exemplo, para os doentes

com idade pertencente à classe C1, fumadores e que sofreram arritmia isquémica não existiam diferenças significativas entre as localizações das distribuições das sobrevivências relativas às classes de enfarto consideradas.

Analisando apenas os doentes na classe de idade C1, pudémos verificar que de entre os não-fumadores a maior proporção de doentes apresentam situações em relação a outras potenciais variáveis prognóstico constantes do ficheiro e cujo estudo não apresentámos por terem um número muito reduzido de observações, que explicam as curtas sobrevidas observadas. De entre os doentes fumadores a mesma proporção é inferior, ficando, assim, uma maior proporção de casos por explicar. Na busca de outras possíveis potenciais variáveis prognóstico que pudessem contribuir para clarificar estas situações, deparámos com a variável Morada e criámos uma nova variável que designámos por Concelho. Ao longo deste trabalho abordaremos várias vezes o que esta variável representa. Como referimos na Introdução deste trabalho, o facto deste estudo incidir em dados referentes à população de uma ilha e tendo em conta o modo como esta foi colonizada, dão grande importância à consideração da variável Concelho no estudo que estamos a desenvolver.

Ao analisarmos a possibilidade da variável Concelho ser uma potencial variável prognóstico, deparámos com o problema de existir um concelho (Funchal) com um número de mortes muito superior aos restantes e, de entre estes, o número de mortes apresentar valores consideravelmente diferentes. No entanto, escolhendo dois concelhos com números de mortes não muito diferentes, Câmara de Lobos e Machico, por exemplo, os resultados da análise revelam que as localizações das distribuições das sobrevivências são significativamente diferentes (o valor p é 0.038).

Não consideramos correcto agrupar concelhos por forma a tornar possível a continuação da análise que temos vindo a desenvolver, sem verificar, primeiro, se é plausível proceder a tal

agrupamento e quais os concelhos a agrupar.

Torna-se, pois, imprescindível uma análise exploratória desta base de dados que tenha em conta a variável Concelho. Os dados serão analisados globalmente e estratificados pelos concelhos, pois consideramos que esta estratificação pode ser, de entre as correspondentes às variáveis prognóstico que estudámos, a mais "rica" no sentido de nos permitir identificar variáveis *confounding* e potenciais variáveis prognóstico que não constam desta base de dados, mas que poderemos indicar à equipa médica como sendo importante a inclusão das mesmas na observação de casos futuros.

A análise dos dados da sobrevivência dos doentes de angina é desenvolvida de modo semelhante à análise que apresentámos relativamente aos dados da sobrevivência dos doentes de enfarto pelo que seria repetitivo apresentá-la neste trabalho, embora as potenciais variáveis prognóstico não sejam exactamente as mesmas — a variável Killip que representa o grau de enfarto, por exemplo, não é considerada como uma potencial variável prognóstico relativamente à sobrevivência de um doente de angina.

Também para os dados da sobrevivência dos doentes de angina a variável Concelho se apresenta como uma potencial variável prognóstico.

Referências

Lee, E. (1992) *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, Inc.

Mosteller, F. e Rourke (1973) *Sturdy Statistics*. Addison-Wesley.

Capítulo II

Construção dos mapas da Ilha da Madeira e da Ilha do Porto Santo

1 — Introdução

Tal como em algumas circunstâncias não faz sentido ordenar os dados da amostra, porque se reconhece que a ordem cronológica das observações é relevante, e há que usar as técnicas de sucessões cronológicas, inúmeros exemplos mostram que a localização dos dados da amostra no espaço é da maior relevância em toda a análise estatística posterior — e assim tem-se assistido, nos últimos anos, a um desenvolvimento explosivo de estatística espacial, com inúmeras aplicações em geografia, geoestatística, prospecção mineira, estudos faunísticos (nomeadamente estudo de bancos de pescado), epidemiologia, saúde pública, etc.

A doença coronária é afectada por uma variedade de factores, entre os quais os que dizem respeito ao tipo de vida, ao meio ambiente e ao local onde vivem as pessoas doentes. As características destes locais (incluindo a situação sócio-demográfica e ambiental) oferecem uma fonte de informação importante para estudos de investigação epidemiológica. Nesta perspectiva, a doença também apresenta uma dimensão espacial.

Já há mais de um século, os epidemiologistas e outros cientistas da área da medicina, começaram a explorar o poder dos mapas na ajuda à assimilação e compreensão da dinâmica espacial da doença.

Hoje em dia, a doença coronária está a ser um alvo importante de investigação e o aspecto geográfico está a ser considerado. A modelação da distribuição geográfica da doença coronária é extremamente útil para a modelação estatística da doença e para posteriores planeamentos de prevenção primária. Através da modelação geográfica podem detectar-se formas de distribuições, semelhantes ou opostas, interessantes.

No presente capítulo abordamos apenas a problemática da construção de mapas adequados a tratamento estatístico. Assim, o objectivo primordial não é o de apresentação de cartas geográficas de grande qualidade gráfica, antes a construção de mapas (que obviamente interessa que sejam detalhados e correctos) utilizáveis conjuntamente com um programa estatístico acessível, uma base de dados e uma folha de cálculo.

2 — Dois métodos de discretização de mapas

Para que o Systat, que é o programa de estatística (usámos a versão para Macintosh devido ao excelente grafismo que este tipo de computador oferece há vários anos) que vamos usar na análise estatística de dados, desenhe um mapa é necessário criar um ficheiro de texto, num programa de texto, com as coordenadas (cartesianas ou geográficas) de alguns dos pontos que formam as linhas que delimitam o mapa — é a este processo que chamamos a discretização de um mapa. Na primeira linha deste ficheiro tem de ser indicado um número de identificação do mapa e o número de pontos que vamos introduzir para o desenho do mesmo. Segue-se, então, a lista das coordenadas dos pontos. Para que o mapa forme um polígono basta que o primeiro ponto da lista seja o mesmo que aparece no fim da lista. Este ficheiro tem de ser guardado em

forma de texto. Usámos o programa Microsoft Word para criar este ficheiro.

Temos, também, que criar no Systat um ficheiro de dados, que terá de incluir o mesmo número de identificação do mapa que foi indicado no ficheiro de texto, as abcissas máxima e mínima e as ordenadas máxima e mínima dos pontos que se encontram no ficheiro de texto. Note-se que no caso de estarmos a trabalhar com coordenadas geográficas, a longitude corresponderá à abcissa e a latitude à ordenada. Este ficheiro tem, também, que incluir a abcissa e a ordenada de um ponto do interior do mapa, que servirá de referência para qualquer representação (nome, gráficos, etc.) que se queira fazer sobre o mapa. O nome a atribuir a este ficheiro pode ser um nome qualquer.

O ficheiro de texto é importado pelo Systat na opção de mapa e guardado no Systat com o mesmo nome atribuído ao ficheiro de texto, seguido da extensão " .Map " .

Utilizámos dois métodos para a obtenção no Systat dos mapas das Ilhas da Madeira e do Porto Santo. Designámo-los por Método 1 e Método 2 e serão descritos nos parágrafos seguintes.

Estes dois métodos diferem no processo utilizado para a obtenção das coordenadas dos pontos que formam as linhas que delimitam as figuras. Além disso, o Método 1 utiliza coordenadas cartesianas, enquanto que o Método 2 utiliza coordenadas geográficas. É de salientar, também, que com o Método 2 é muito mais simples obter as coordenadas dos pontos a introduzir no ficheiro de texto, pelo que permite a utilização de um número muito maior de pontos, o que aumenta a precisão do mapa.

Realce-se que pretendemos obter os mapas para apresentações com fins estatísticos e, como tal, não se justifica a utilização de Sistemas Geográficos minuciosamente precisos. Além disso, o que

nos propomos é obter os mapas num programa que os utilize interactivamente com técnicas estatísticas. Isto não invalidou, no entanto, que fossemos o mais rigorosos possível nas técnicas utilizadas na sua construção. Procurámos, também, obter um mapa com qualidade gráfica.

Todos os mapas referidos no texto podem ser consultados na Galeria de Mapas que se encontra no fim deste capítulo.

2.1 — *Construção dos mapas através da manipulação das coordenadas*

A esta técnica utilizada na construção dos mapas chamámos **Método 1**. Para a obtenção dos mapas, introduzimos no *scanner* os mapas da Ilha da Madeira e da Ilha do Porto Santo à escala 1:250000 e guardámos a imagem obtida em formato PICT. Usámos o programa SuperPaint 2.0 para abrir este ficheiro PICT e, através da opção *view - floating windows - show coordinates*, obtivémos as coordenadas (aproximadas à segunda casa decimal) dos pontos que delimitam os concelhos que constituem a Ilha da Madeira e a Ilha do Porto Santo. Tentámos obter as coordenadas do maior número de pontos possível. Esta lista de pontos foi escrita, por concelhos, num ficheiro de texto que guardámos com a opção tipo texto apenas. Ao iniciar a lista de pontos de cada concelho tivémos que escrever um número de identificação do concelho, seguido do número de pontos que constituíam a linha que delimita o concelho. O separador utilizado foi o tabulador. Cada concelho corresponde a uma linha poligonal que, para ser fechada basta que o primeiro ponto da lista de pontos que constituem o concelho, coincida com o último ponto da mesma lista. Este ficheiro de texto é importado pelo Systat na opção de mapa e guardado com um nome qualquer seguido dos da extensão .Map . A origem do eixos coordenados no Systat, situa-se no canto superior esquerdo do écran, o que nos obrigou a escrever, no

ficheiro de texto, as abcissas de todos os pontos positivas e as ordenadas negativas.

O Método 1 utilizou, assim, 1925 pontos para a obtenção do mapa da Ilha da Madeira e 311pontos para a obtenção do mapa da Ilha do Porto Santo. Note-se que os pontos comuns que dizem respeito às fronteiras dos concelhos adjacentes não foram repetidos, por razões que explicaremos à frente.

No Systat foi criado um ficheiro que contém dados sobre os 10 concelhos da Ilha da Madeira e outro que contém dados sobre a Ilha do Porto Santo. Entre estes dados encontram-se as abcissas máxima e mínima de cada concelho, assim como as ordenadas máxima e mínima, a abcissa e a ordenada do ponto do interior do concelho que servirá de referência para qualquer representação (nome do concelho, gráficos, etc.) que se queira fazer sobre o mapa do concelho e, ainda, o número de identificação do mesmo, que tem de coincidir com o que lhe foi atribuído no ficheiro de texto. O nome deste ficheiro criado no Systat tem de ser o mesmo que o que foi atribuído ao ficheiro de texto importado, mas sem os quatro caracteres finais. No ApêndiceII encontram-se editados alguns dos ficheiros de dados criados no Systat.

Deparámos com um problema quando obtivémos o mapa no programa Systat. Em diversos pontos, o mapa não fechava. Os pontos de união de alguns concelhos adjacentes, apesar de terem as mesmas coordenadas, não coincidiam no mapa. Criámos sucessivos ficheiros de texto e de dados no Systat, de forma a acrescentarmos um concelho de cada vez no mapa, iniciando com o concelho da Calheta e continuando pela costa Sul. Este processo levou-nosa verificar que, ao introduzir o concelho de Machico, o mapa, que até aí estava fechado, apresentava pontos de união, pertencentes a concelhos anteriormente introduzidos, não coincidentes. Se, seleccionássemos no ficheiro de dados do Systat todos os concelhos até ao momento introduzidos, com excepção do

concelho de Machico, o mapa referente voltava a apresentar-se fechado. Conclui-se que a escala usada na representação de um mapa depende da área que este cobre. O que parece acontecer é que o programa está feito de modo a não repetir o desenho dos pontos que formam a fronteira entre dois concelhos adjacentes. Como tal, os pontos de união, que são pontos comuns aos concelhos, só são traçados uma vez. Como existem muitos pontos próximos destes, no desenho do mapa a uma escala normal pode não se notar qualquer anomalia. No entanto, quando ampliamos este mesmo mapa, estas discrepâncias são evidenciadas. Quando ampliamos o mapa, o que pensamos verificar-se é que existe uma maior densidade de coordenadas cartesianas identificadas com as coordenadas de écran e, portanto, mais pontos, originalmente distintos, são identificados com os tais pontos comuns que o programa só desenha um vez. Assim, passam a existir mais pontos da lista que constitui o concelho que não são desenhados, aparecendo espaços por preencher, na linha que delimita o concelho, na zona de união com outros concelhos. Este raciocínio repete-se para o caso de introduzirmos novos concelhos e, portanto, mais pontos em coordenadas cartesianas, uma vez que o número de pontos de écran é fixo. Conseguimos fechar o mapa acrescentando, por tentativas, pontos que, efectivamente, não pertenciam nem a um nem a outro dos concelhos em questão. É um processo que pode ser extremamente moroso.

Devido ao facto de termos tido de acrescentar pontos para fechar o mapa, a representação geográfica de alguns concelhos isoladamente, obtida pela respectiva selecção no ficheiro de dados do Systat, apresenta ligeiras alterações. Foi este o motivo que nos levou a optarmos por não repetir no ficheiro de texto as coordenadas dos pontos que formam as fronteiras entre concelhos. Como na futura utilização dos mapas na análise estatística de dados vamos precisar da representação do mapa de cada concelho isoladamente, criámos paralelamente um ficheiro de texto e um ficheiro de dados no Systat para cada concelho.

Na Fig.1 está representado o mapa das ilhas da Madeira e do Porto Santo obtido no Systat, tendo sido utilizado o Método 1. O mapa da Ilha do Porto Santo não está na posição correcta relativamente à Ilha da Madeira, porque o mapa conjunto das ilhas da Madeira e do Porto Santo seria desenhado no Systat a uma escala demasiado pequena. Isto levaria a que não conseguíssemos obter a representação dos mapas das duas ilhas simultaneamente na janela de trabalho, a partir de uma escala ainda muito reduzida. Optámos, então, por colarmos uma fotocópia do mapa da Ilha do Porto Santo numa posição mais próxima da Ilha da Madeira. A figura obtida apresentava, ainda, a vantagem de poder ser introduzida no scanner numa só vez.

Ao observarmos o mapa da Ilha da Madeira obtido no Systat, Fig.2, verificámos que, ao ampliá-lo, Fig.3 e Fig.4, perdia qualidade e pior acontecia quando o programa traçava o mapa de um concelho isoladamente. Assim, decidimos desenhar um novo mapa com um número muito maior de pontos.

2.2 — *Construção dos mapas através de um programa de execução e manipulação de mapas*

Como pretendíamos obter no Systat um mapa mais preciso, utilizámos um mapa do Arquipélago da Madeira à escala 1:50 000, fornecido pelo Instituto Geográfico e Cadastral, para introduzir no scanner. Este mapa consta de três cartas geográficas pelo que teve de ser introduzido no scanner em partes, cada uma contendo uma área relativamente pequena do mapa, como se pode verificar pela Fig.5.

Utilizámos o programa Atlas Pro, que é um programa de execução e manipulação de mapas (versão para Macintosh), para importar as imagens obtidas através do scanner, guardadas em formato PICT e criar um ficheiro de texto com as longitudes e

latitudes dos diversos pontos que formam as linhas que delimitam os concelhos. A esta técnica utilizada na construção dos mapas chamámos **Método 2**.

Numa folha de papel vegetal copiámos, com um traço extremamente fino por forma a que a cópia fosse bastante fiel, uma pequena parte do mapa acrescentando as linhas necessárias por forma a obter uma linha poligonal fechada para que no programa Atlas Pro seja possível a criação de um ficheiro de texto com as coordenadas geográficas dos pontos. É, também, necessário acrescentar dois pontos de referência, um no extremo superior esquerdo e outro no extremo inferior direito, cujas longitudes e latitudes têm de ser indicadas quando o Atlas Pro calibrar aquela área do mapa. A Fig.5 é a imagem obtida no scanner duma parte do concelho do Porto Moniz e dos respectivos pontos de referência, guardada em formato PICT.

O programa Atlas Pro importa este ficheiro PICT como um simples objecto e, quando seleccionamos o comando *Geography* no menu *Edit* deste programa, então no menu *Tools* ao escolhermos a opção *Auto-Trace* o programa automaticamente determina os limites interiores da área. Mas para que este processo dê os resultados pretendidos, é necessário que a área em questão esteja completamente visível na janela de trabalho, assim como os pontos de referência anteriormente referidos. Este é mais um motivo para introduzirmos o mapa das ilhas da Madeira e do Porto Santo em partes relativamente pequenas. Também é necessário que não exista nenhuma interrupção nas linhas que delimitam a área, para que o programa possa correctamente determinar os limites da mesma. É por esta razão que, ao copiarmos para o papel vegetal uma porção do contorno do mapa, tivémos de acrescentar linhas de forma a obtermos um polígono. Seguidamente calibrámos aquela parte do mapa, seleccinando no menu *Geo* o comando *Lat/Lon Options* e escolhendo a opção *Calibration*. Foi, então, que tivémos de indicar com o cursor os pontos de referência e as respectivas

longitudes e latitudes. Note-se que, ao indicarmos estes pontos, é extremamente difícil colocarmos o cursor precisamente em cima de cada um deles, podendo este facto originar alguma imprecisão no processo de calibração.

Obtivémos o ficheiro de texto com as coordenadas geográficas, (aproximadas à quarta casa decimal), dos pontos que formam as linhas que delimitam aquela área do mapa, seleccionando no menu *File*, o comando *Export* e a opção *Geography* que converte coordenadas de écran em texto na forma de coordenadas latitude/longitude. O ficheiro de texto obtido pode ser aberto em qualquer programa de texto, e alterado se o desejarmos. Se o importarmos para o Atlas Pro, através do comando *Geographic Utilities* no menu *File* e da opção *Conversions - Text to Atlas Pro Geography* , convertemos o ficheiro de coordenadas geográficas em formato de texto, num ficheiro geográfico no formato usado pelo Atlas Pro. Na Fig.6 apresentamos o mapa obtido a partir do ficheiro geográfico correspondente à parte do concelho do Porto Moniz introduzida no scanner. O mapa está desenhado com grande precisão, o que nos indica que o número de pontos e as respectivas coordenadas reproduzem correctamente aquela área do mapa do arquipélago da Madeira.

Houve situações em que, simultaneamente, introduzimos duas áreas correspondentes a concelhos diferentes sendo, portanto, separadas por uma parte da fronteira comum. A única alteração em relação ao processo utilizado anteriormente é no momento em que o programa determina os limites interiores da área. Utilizámos a opção Auto-Trace duas vezes, uma para cada área, e atribuímos-lhes nomes diferentes.

Como pode verificar-se no ficheiro de texto que editámos e que aparece no ApêndiceIII, contendo as coordenadas geográficas da parte do concelho do Porto Moniz, aparece na primeira linha o

nome que escolhemos para identificar a área, o nome que escolhemos para legendar a área, assim como o número de pontos que o programa utilizou para definir as linhas que a delimitam. Seguem-se as coordenadas geográficas dos pontos. No caso de termos introduzido simultaneamente mais do que uma área, no ficheiro de texto correspondente aparece na primeira linha o nome escolhido para identificação da primeira área que o programa determinou, o nome escolhido para legendar a mesma, assim como o número de pontos que o programa utilizou para a determinar. Segue-se a lista das coordenadas geográficas daqueles pontos. No fim desta lista surge uma linha com o nome de identificação da segunda área que o programa determinou, o nome escolhido para legendar a mesma, assim como o número de pontos que o programa utilizou para a determinar. Nas linhas seguintes estão as coordenadas geográficas daqueles pontos. Esta representação seria idêntica para as restantes áreas.

Importámos um ficheiro de texto nas condições acabadas de referir para o Atlas Pro e verificámos que a representação do ficheiro geográfico obtido apresentava cada fronteira entre concelhos adjacentes como sendo linhas não coincidentes. O programa digitalizou a fronteira entre duas áreas adjacentes, de forma diferente para cada uma delas. Isto deve-se ao facto de o programa determinar automaticamente os limites interiores de cada área. Note-se que, como já referimos no início deste parágrafo, utilizámos um "rotring finograph ef" para copiarmos para o papel vegetal as áreas a introduzir, de modo a obtermos um traço extremamente fino. Consequentemente, tivémos de optar por uma destas linhas para fronteira entre os concelhos adjacentes.

Este ficheiro de texto, depois de feitas algumas alterações, será importado pelo Systat na opção Map. Uma das alterações que fizémos consistiu em substituir a linha de identificação de cada área, por uma linha que contém o número de identicação do concelho a que aquela área se refere e o número de pontos que

vamos introduzirnesta área. A outra alteração que fizémos consistiu em procurarmos no ficheiro de texto as coordenadas dos pontos que formam as linhas que acrescentámos ao mapa de forma a obtermos um polígono, tendo em conta que o programa digitaliza o polígono no sentido dos ponteiros do relógio e que, para facilitarmos o processo, acrecentámos uma linha vertical e uma horizontal que são fáceis de detectar através das alterações na longitude e na latitude.

Assim, começámos por copiar a primeira página de coordenadas, do ficheiro de texto original, para um novo ficheiro de texto, na primeira linha do qual corrigimos o número de pontos. O Atlas Pro importa qualquer sequência de pontos no formato idêntico ao do ficheiro de texto editado, sem exigir que a referida sequência forme um polígono. Depois de o Atlas Pro ter importado este novo ficheiro de texto, obtivémos a representação geográfica daquela sequência de pontos. Foi, então, fácil de nos apercebermos onde é que o programa começou a digitalizar a área. Voltámos ao ficheiro de texto original e depois de detectarmos as coordenadas dos pontos que formavam as linhas vertical e horizontal anteriormente referidas, eliminámo-las do ficheiro e realizámos neste os ajustes necessários, de modo a obtermos novamenteuma sequência de pontos. Obviamente este ficheiro assim obtido foi importado pelo Atlas Pro para nos assegurarmos de que a sequência de pontos correspondia exactamente à área que queríamos digitalizar. Houve situações em que foi necessário proceder a pequenas alterações. No caso em que o ponto a que diziam respeito as primeiras coordenadas geográficas da lista se situar na parte do mapa que queríamos guardar, mantivémos a sequência inicial de pontos até se atingir as coordenadas dos pontos que formavam as linhas vertical e horizontal e, depois de estas terem sido eliminadas, juntámos à sequência final de pontos a inicial.

Como exemplo do exposto atrás, atentemos no ficheiro de

texto referente à digitalização da parte do concelho do Porto Moniz que introduzimos no scanner e que está representado na Fig.5, que editámos e que se encontra no ApêndiceIII. Procedendo como indicámos atrás, verificámos que o primeiro ponto da lista pertence ao contorno do mapa que nos interessa guardar. Consultando a imagem introduzida no scanner e tendo em conta, como já referimos, que a imagem é digitalizada no sentido dos ponteiros do relógio, seguimos a sequência de pontos até encontrarmos aquele em que a latitude pára de diminuir e, simultaneamente, a longitude começa a diminuir. Rapidamente se verifica que é o primeiro ponto assinalado na página 6 daquele ficheiro. A sequência de pontos que se segue forma a linha horizontal que procurávamos. Seguidamente procurámos o ponto em que a latitude começa a aumentar e, simultaneamente, a longitude estabiliza. Encontrámos o ponto assinalado na página 7. A sequência de pontos que se segue forma a linha vertical que nos interessava identificar. Foi, depois fácil detectar o ponto onde a latitude deixa de aumentar. Assinalámos este ponto na página 8. A sequência seguinte já nos interessava guardar, pois dizia respeito à parte do mapa em que estávamos interessados. Seguidamente eliminámos do ficheiro as sequências referentes às linhas horizontal e vertical, copiámos a sequência inicial de pontos e colámo-la a seguir à sequência final e fizémos a consequente rectificação no número de pontos constantes do ficheiro.

Feitas as alterações referidas, obtivémos um ficheiro de texto em condições de ser importado pelo programa Systat na opção Map.

Quando obtivémos, por este processo, a primeira área referente ao mapa da Ilha da Madeira, criámos no Systat um ficheiro de dados contendo o número de identificação do concelho, que tem de ser o mesmo que indicámos no ficheiro de texto, as latitudes máxima e mínima do concelho, as respectivas longitudes máxima e mínima, assim como a longitude e a latitude do ponto

interior do concelho que servirá de referência para qualquer representação (nome do concelho, gráficos, etc.) que se queira fazer sobre o mapa do concelho.

Este ficheiro de dados foi acrescentado sempre que introduzimos novos concelhos e o ficheiro de texto foi sucessivamente actualizado com as novas sequências de pontos que obtínhamos. Editámos o ficheiro final, que pode ser consultado no ApêndiceII. Nos casos em que a sequência a acrescentar dizia respeito a um concelho ainda não introduzido no ficheiro de texto, criámos neste uma nova linha contendo o número de identificação do concelho e o número de pontos que constituíam a referida sequência e, seguidamente a lista das coordenadas geográficas dos mesmos. Nos casos em que a sequência a acrescentar pertencia a concelhos já introduzidos no ficheiro de texto, começámos por verificar se estes novos pontos se encontravam na mesma ordem dos pontos até aí introduzidos, ou se a ordem teria de ser invertida de forma a que, o total de pontos na listado concelho formasse uma sequência. Nos casos em que foi necessário inverter a ordem, utilizámos um programa que fizémos, uma vez que a lista de pontos era grande.

Paralelamente e, porque decidimos, pelo mesmo motivo que referimos na descrição do Método 1, introduzir os pontos que formam as fronteiras só num dos concelhos a que elas se referem, criámos um novo ficheiro de texto e um correspondente ficheiro de dados no Systat para cada concelho. Este ficheiro de texto foi sucessivamente acrescentado cada vez que obtivémos mais coordenadas de pontos pertencentes ao respectivo concelho, até obtermos as coordenadas geográficas de todos os pontos que formam o polígono que o representa e tendo em atenção que o último ponto da sequência tem de ser igual ao primeiro.

Nos casos em que foram digitalizadas, simultaneamente, duas ou mais áreas, depois de determinadas as sequências do ficheiro de

texto original que queríamos obter, copiámo-las para o ficheiro de texto da Ilha da Madeira e colámo-las nas partes correspondentes aos concelhos a que elas se referiam, depois de feitas as verificações e as alterações quando necessárias, que indicámos anteriormente.

O mapa da Ilha do Porto Santo foi obtido exactamente pelo mesmo processo. Não o incluímos no ficheiro da Ilha da Madeira porque o mapa obtido era apresentado a uma escala extremamente pequena, o que mesmo depois de ampliado, o tornava inútil para a utilização que lhe pretendemos dar por não ser possível obter a representação dos mapas das duas ilhas, simultaneamente, na janela de trabalho. Além disso, ao introduzir a Ilha do Porto Santo no ficheiro da Ilha da Madeira, o mapa desta ilha voltava a abrir em alguns pontos de união de concelhos, o que faria com que esta primeira parte do projecto se arrastasse ainda por mais tempo, pois como foi referido em 2.2 o processo utilizado para voltar a fechar o mapa pode ser extremamente moroso.

Optámos por criar um ficheiro de texto e um ficheiro de dados no Systat só para a Ilha do Porto Santo. O Systat importou na opção de Map, os ficheiros de texto finais obtidos para cada um dos concelhos da Ilha da Madeira, para a Ilha do Porto Santo, assim como o ficheiro completo da Ilha da Madeira. O Método 2 utilizou, assim, 7950 pontos para a obtenção do mapa da Ilha da Madeira e 1203 pontos para a obtenção do mapa da Ilha do Porto Santo.

Na representação gráfica dos mapas, deparámos com os mesmos problemas que surgiram quando utilizámos o Método 1, os quais foram resolvidos pelo mesmo processo.

A Fig.7 e a Fig.8 representam o mapa da Ilha da Madeira, obtido no Systat e utilizando o Método 2, com ampliações diferentes.

Na Fig.12 pode observar-se o mapa da Ilha do Porto Santo, obtido no Systat e utilizando o Método 2.

Para a obtenção dos mapas que estão representados nas Fig.9, Fig.10, Fig.11, Fig.12 e Fig.13, servimo-nos das coordenadas dos pontos do interior de cada concelho que foram introduzidos no ficheiro de dados do Systat como pontos de referência para representações que pretendessemos fazer sobre o mapa.

É possível obter o mapa conjunto das ilhas da Madeira e do Porto Santo a uma escala razoável, desenhando o mapa da Ilha da Madeira e, na mesma janela de trabalho, desenhando, também o mapa da Ilha do Porto Santo. Seguidamente podemos deslocar cada um deles como quisermos. Foi utilizando este processo que obtivémos os mapas que estão representados nas Fig.13 e Fig.14.

3 — Comparação dos mapas obtidos pelos dois métodos

Observando a Fig.15 que representa os mapas da Ilha do Porto Santo obtidos no Systat pelos dois métodos à escala normal, pode verificar-se que o Método 2 originou um mapa de melhor qualidade do que o Método 1.

Na utilização que vamos dar aos mapas que construímos, precisaremos de os obter bastante ampliados. Na Fig.16 e na Fig.17 estão representados os mesmos mapas mas, agora, desenhados numa escala maior. É evidente a diferença de qualidade dos mapas obtidos. O Método 2 usou 1203 pontos na construção do mapa, enquanto que o Método 1 usou, apenas, 311 pontos. O mesmo se pode confirmar pela Fig.18 que representa, à escala normal utilizada pelo Systat, os mapas do concelho de Santa Cruz obtidos pelos dois métodos. Note-se que os polígonos que representam o

concelho são diferentes. O que está correcto é o que foi obtido utilizando o Método 2, uma vez que o mapa original utilizado neste método foi fornecido pelo Instituto Geográfico e Cadastral. Neste caso, o Método 2 usou 943 pontos na construção do mapa, enquanto que o Método 1 usou, apenas, 243.

Na Fig.19 podemos observar os mapas da Ilha da Madeira obtidos no Systat utilizando os dois métodos. Mesmo a uma escala reduzida confirma que o Método 2 dá melhores resultados que o Método 1. Este utilizou 1925 pontos para a obtenção do mapa e o Método 2 usou 7950 pontos.

Pode, agora, entender-se o motivo que nos levou a construir um novo mapa utilizando mais pontos na sua digitalização. A precisão do mapa obtido depende, em grande parte, da escala do mapa original, uma vez que esta decide o número de pontos a serem digitalizados. Além disso, no Método 2, a digitalização do mapa é automática, o que o torna mais simples que o Método 1 e, também, mais preciso.

4 — Conclusões

A escolha do programa Systat para utilizarmos no desenvolvimento deste projecto está, pois, justificada. A capacidade deste programa para executar análises estatísticas já era por nós conhecida. Verificámos agora, que o mesmo desenha mapas com uma precisão aceitável e que os utiliza em interacção com técnicas estatísticas de uma forma que é de uma importâcia relevante para o trabalho que nos propomos realizar.

Vai-nos permitir realizar apresentações não só cientificamente interessantes, mas também precisas do ponto de

vista geográfico.

É uma técnica extraordinária para nos ajudar numa procura "cega" de interacções num grupo de variáveis.

GALERIA DE MAPAS

ILHAS DA MADEIRA E DO PORTO SANTO (obtidas pelo Método 1)

Fig. 1

ILHA DA MADEIRA (obtida pelo Método 1)

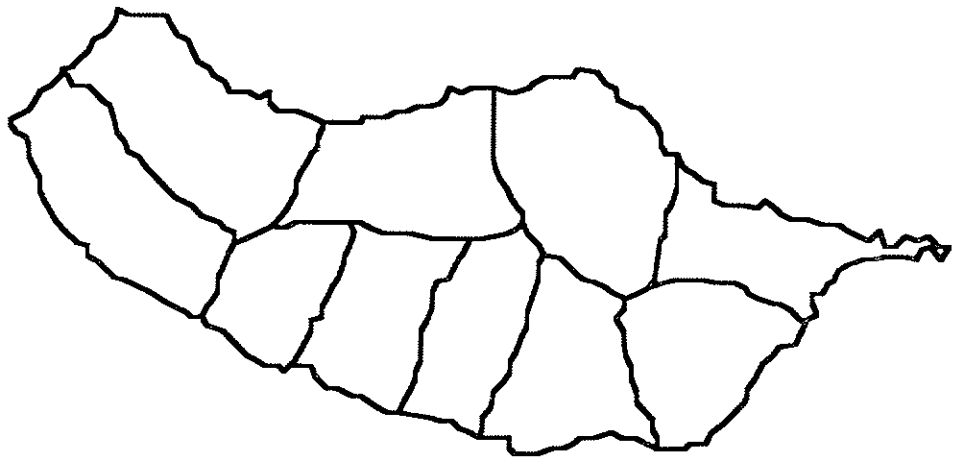


Fig. 2

ILHA DA MADEIRA (obtida pelo Método 1)

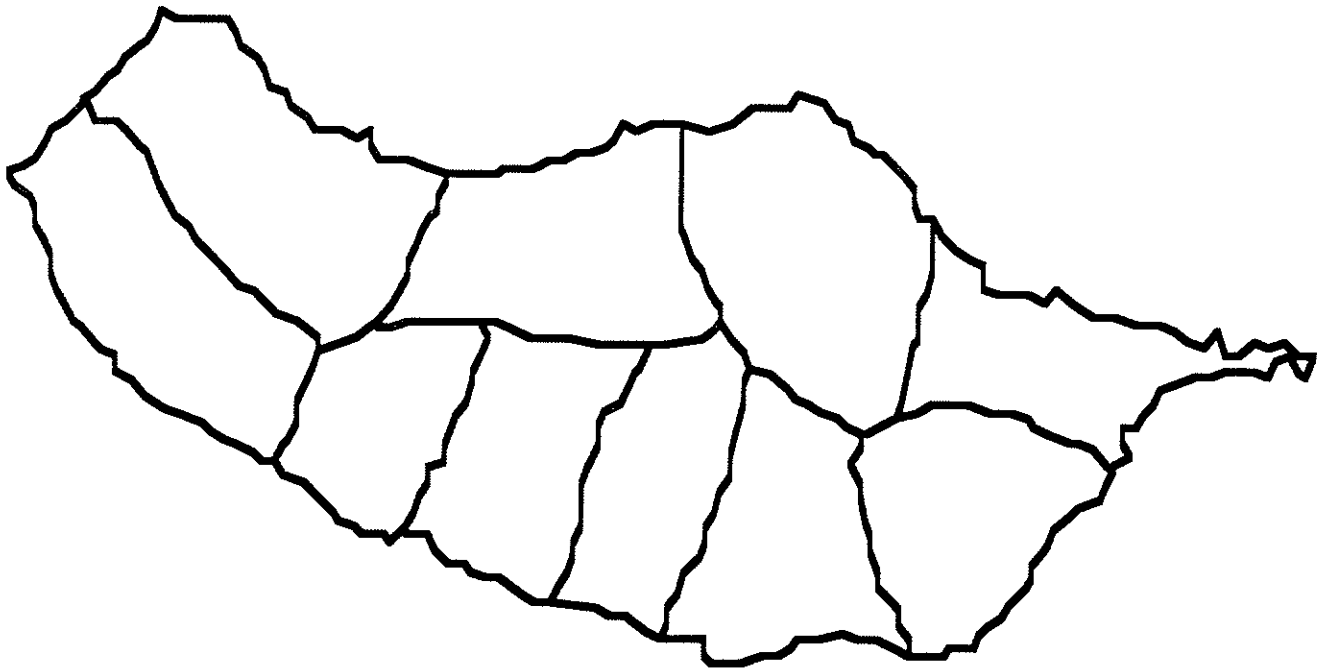


Fig. 3

ILHA DA MADEIRA (obtida pelo Método 1)

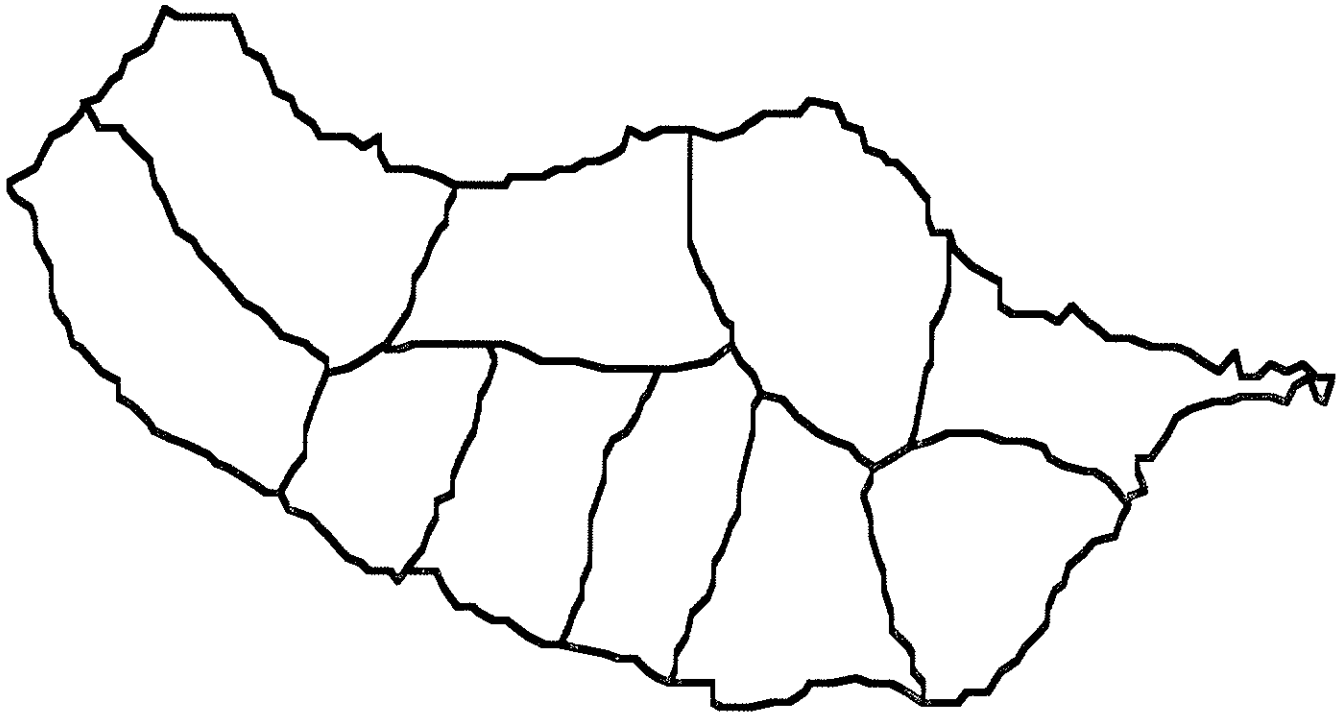


Fig. 4

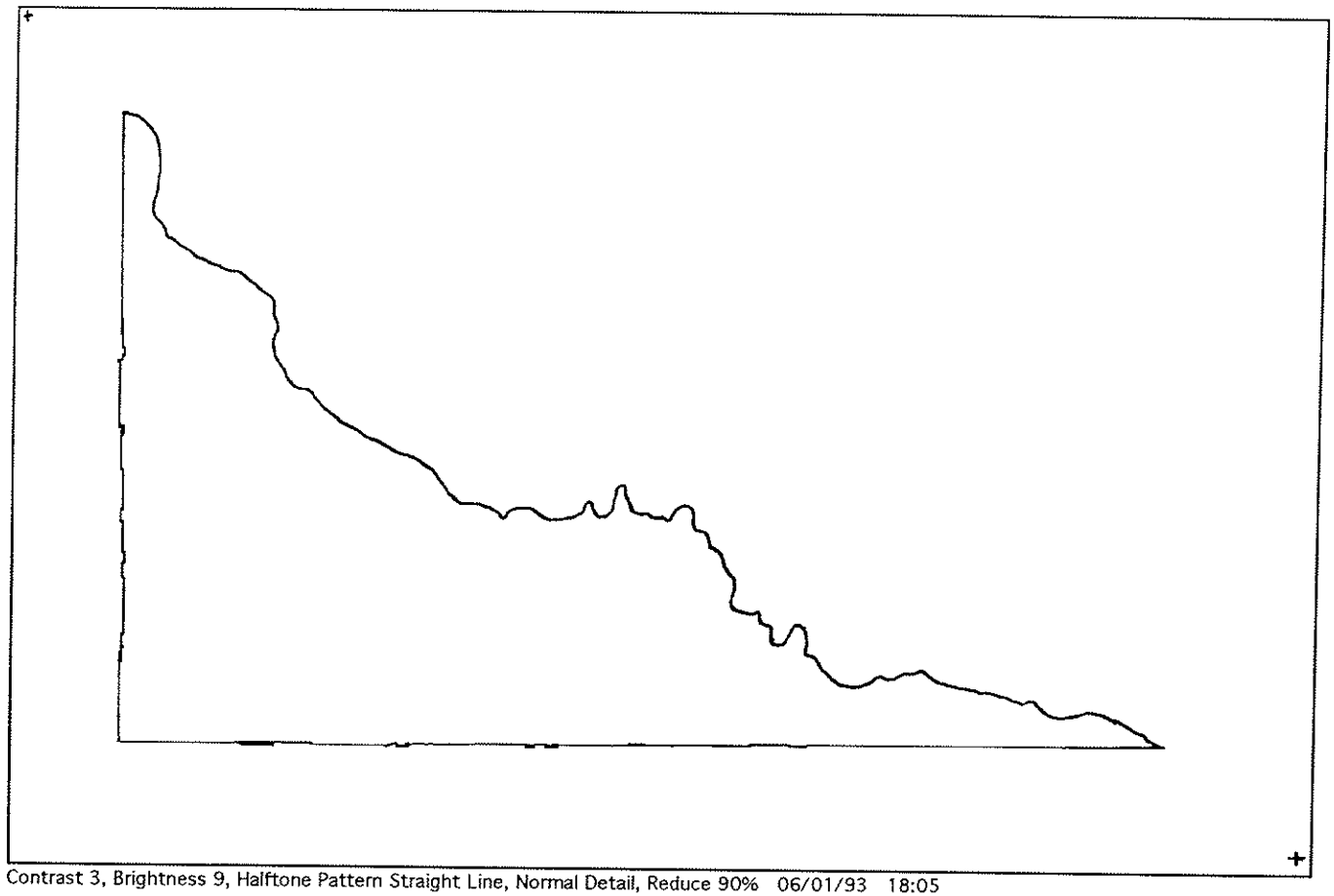


Fig. 5 - Imagem obtida no scanner de uma parte do concelho do Porto Moniz (redução de 90%)

Mapa de parte do concelho do Porto Moniz, obtido em Atlas Pro

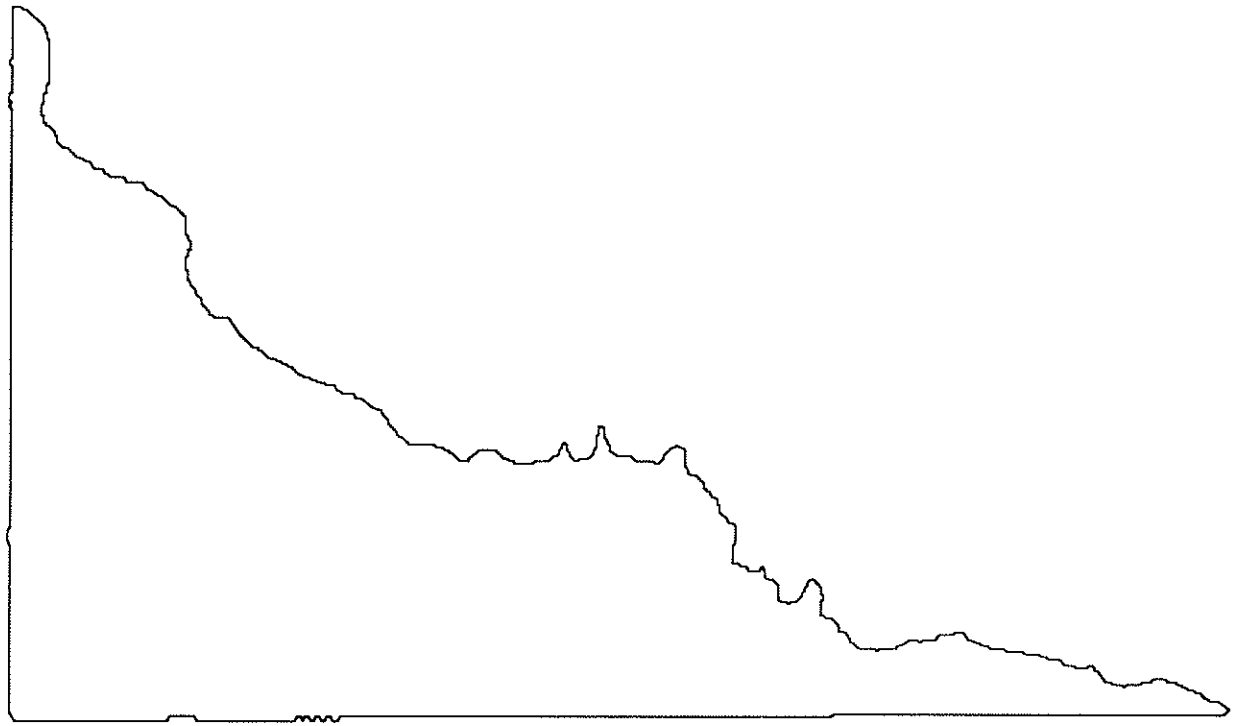


Fig. 6

ILHA DA MADEIRA (obtida pelo Método 2)

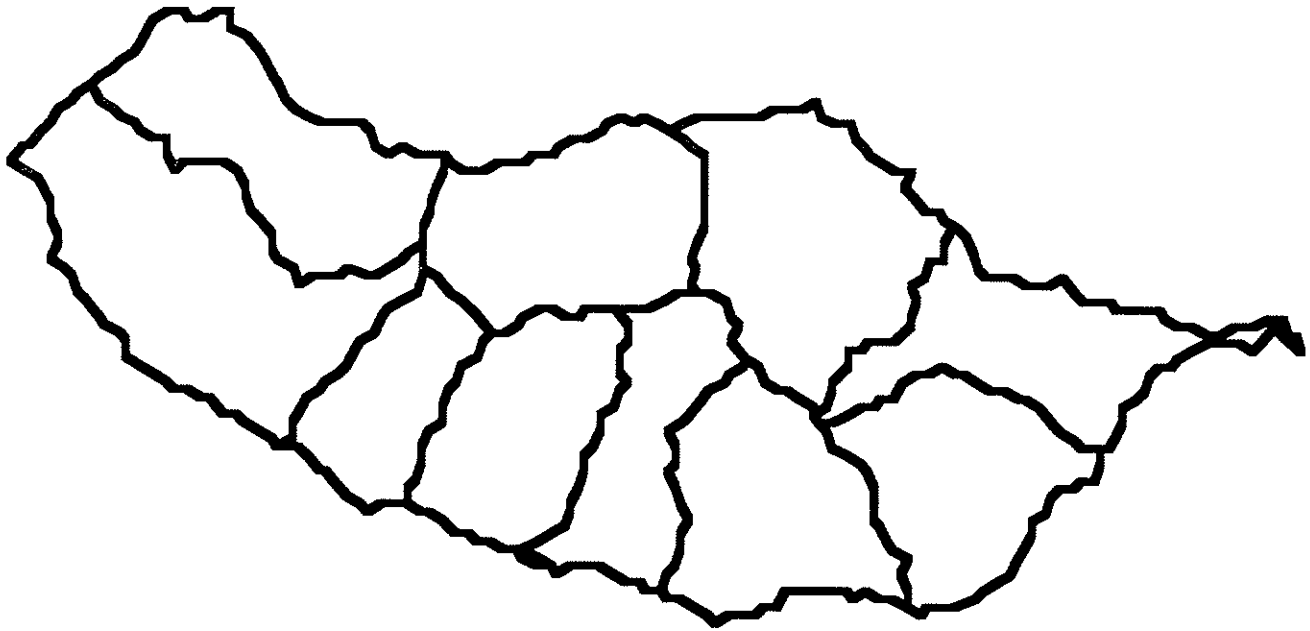


Fig. 7

ILHA DA MADEIRA (obtida pelo Método 2)

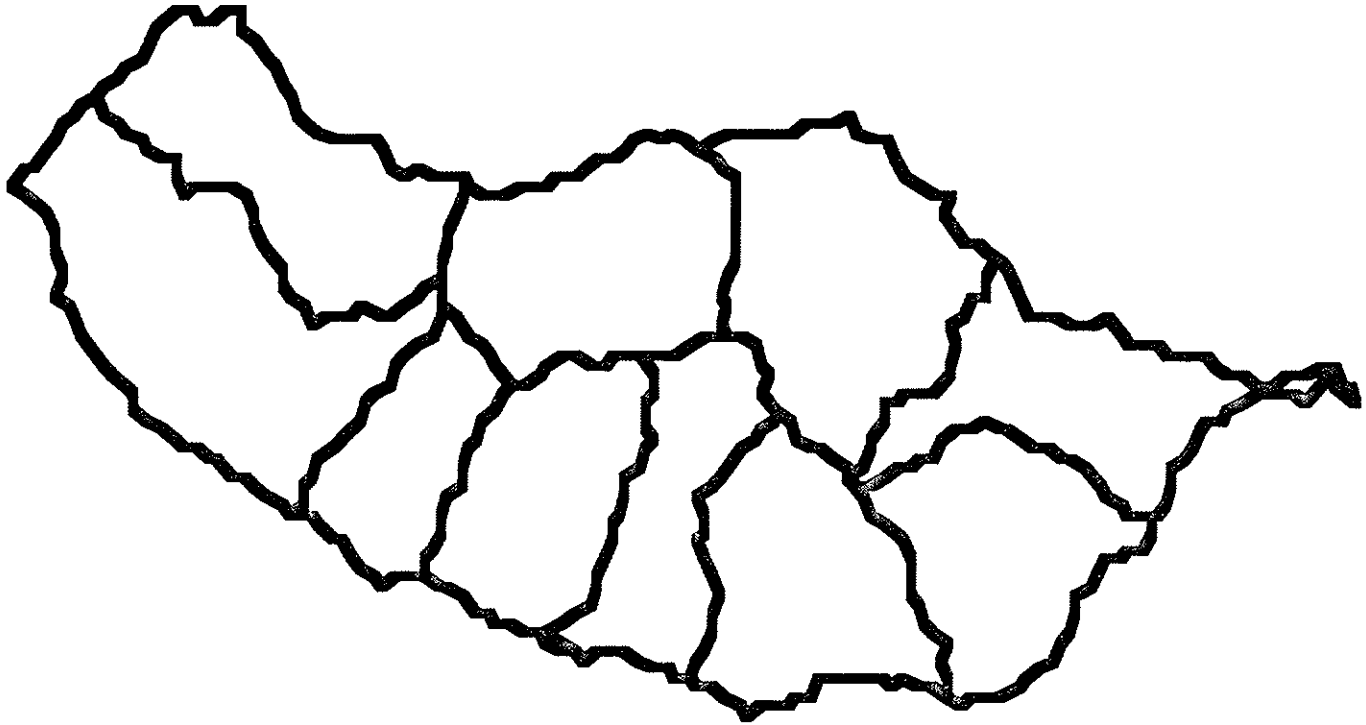


Fig. 8

ILHA DA MADEIRA

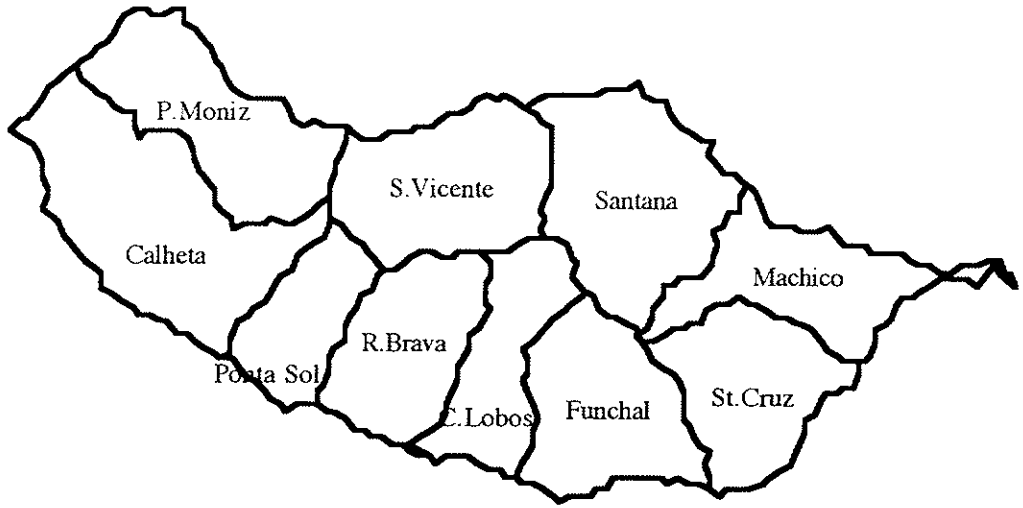


Fig. 9

ILHA DA MADEIRA

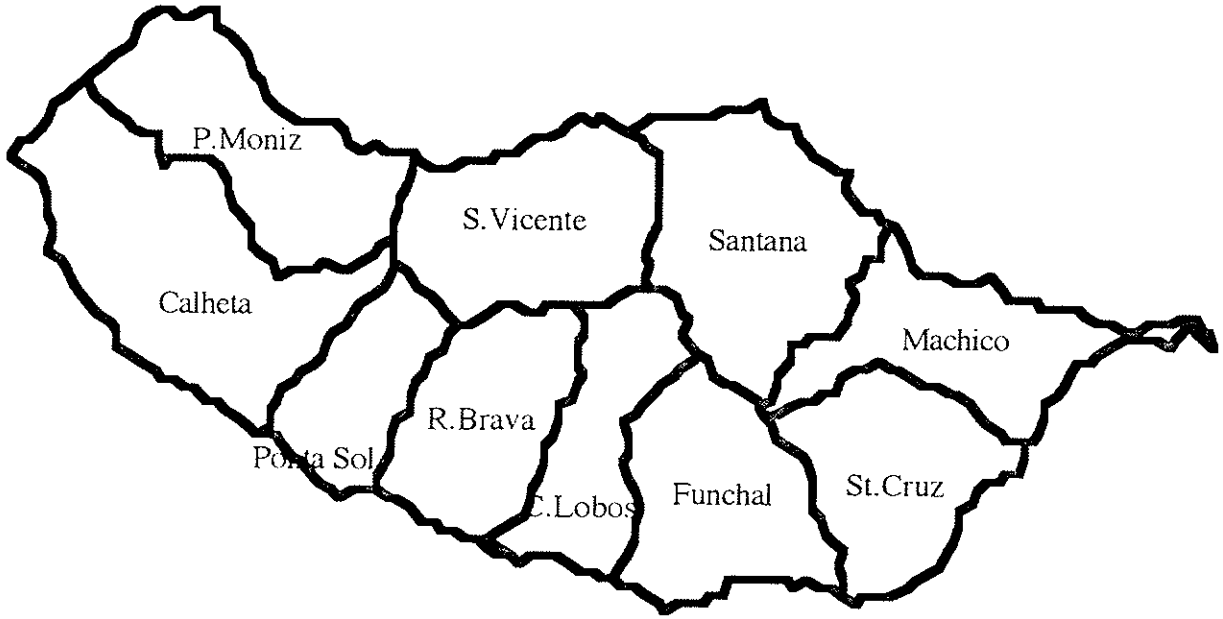


Fig. 10

ILHA DA MADEIRA

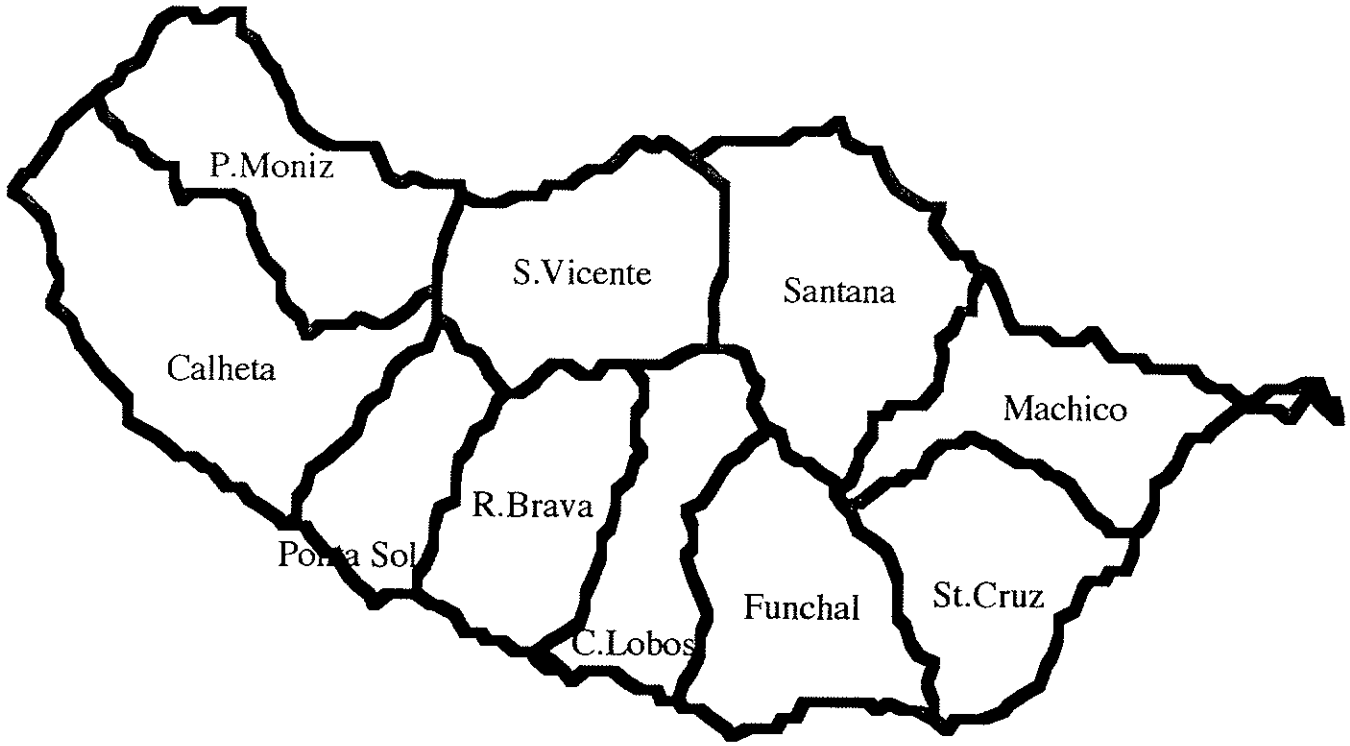


Fig. 11

Fig. 12

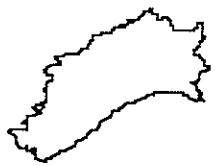
ILHAS DA MADEIRA E DO PORTO SANTO

Fig. 13

ILHAS DA MADEIRA E DO PORTO SANTO (obtidas pelo Método 2)

Fig. 14

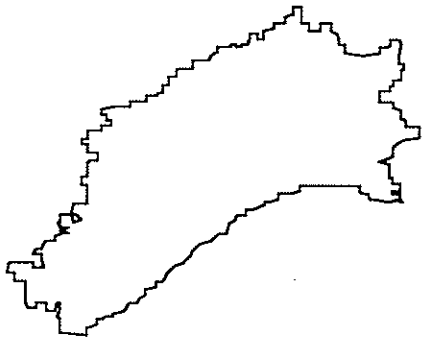
ILHA DO PORTO SANTO (obtida pelo Método 1)



ILHA DO PORTO SANTO (obtida pelo Método 2)

Fig. 15

ILHA DO PORTO SANTO (obtida pelo Método 1)



ILHA DO PORTO SANTO (obtida pelo Método 2)

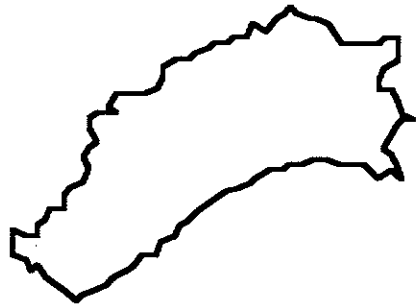
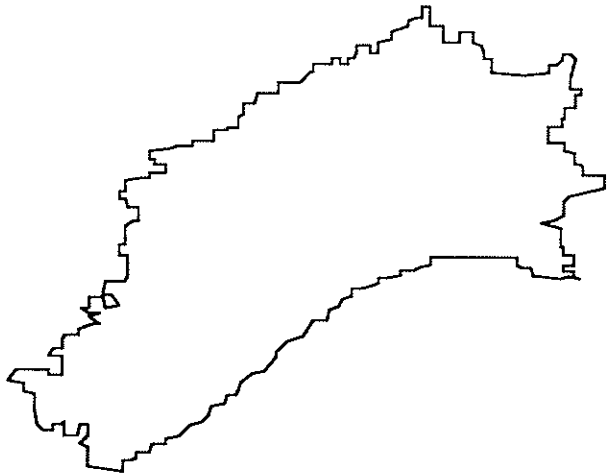


Fig. 16

ILHA DO PORTO SANTO (obtida pelo Método 1)



ILHA DO PORTO SANTO (obtida pelo Método

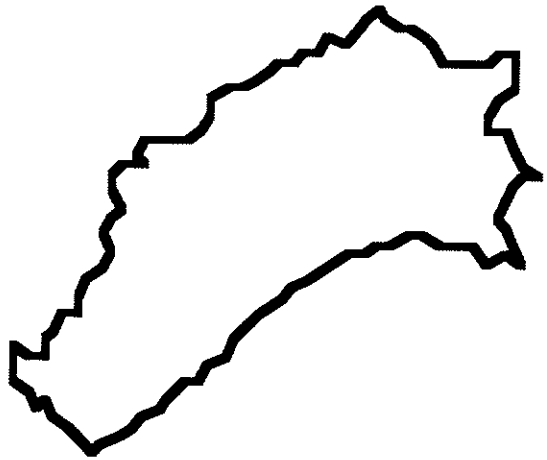


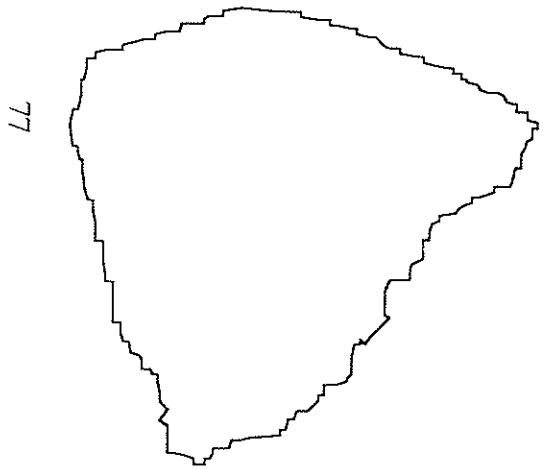
Fig. 17

vamos introduzirnesta área. A outra alteração que fizémos consistiu em procurarmos no ficheiro de texto as coordenadas dos pontos que formam as linhas que acrescentámos ao mapa de forma a obtermos um polígono, tendo em conta que o programa digitaliza o polígono no sentido dos ponteiros do relógio e que, para facilitarmos o processo, acrecentámos uma linha vertical e uma horizontal que são fáceis de detectar através das alterações na longitude e na latitude.

Assim, começámos por copiar a primeira página de coordenadas, do ficheiro de texto original, para um novo ficheiro de texto, na primeira linha do qual corrigimos o número de pontos. O Atlas Pro importa qualquer sequência de pontos no formato idêntico ao do ficheiro de texto editado, sem exigir que a referida sequência forme um polígono. Depois de o Atlas Pro ter importado este novo ficheiro de texto, obtivémos a representação geográfica daquela sequência de pontos. Foi, então, fácil de nos apercebermos onde é que o programa começou a digitalizar a área. Voltámos ao ficheiro de texto original e depois de detectarmos as coordenadas dos pontos que formavam as linhas vertical e horizontal anteriormente referidas, eliminámo-las do ficheiro e realizámos neste os ajustes necessários, de modo a obtermos novamenteuma sequência de pontos. Obviamente este ficheiro assim obtido foi importado pelo Atlas Pro para nos assegurarmos de que a sequência de pontos correspondia exactamente à área que queríamos digitalizar. Houve situações em que foi necessário proceder a pequenas alterações. No caso em que o ponto a que diziam respeito as primeiras coordenadas geográficas da lista se situar na parte do mapa que queríamos guardar, mantivémos a sequência inicial de pontos até se atingir as coordenadas dos pontos que formavam as linhas vertical e horizontal e, depois de estas terem sido eliminadas, juntámos à sequência final de pontos a inicial.

Como exemplo do exposto atrás, atentemos no ficheiro de

SANTA CRUZ (Método 1)



SANTA CRUZ (Método 2)

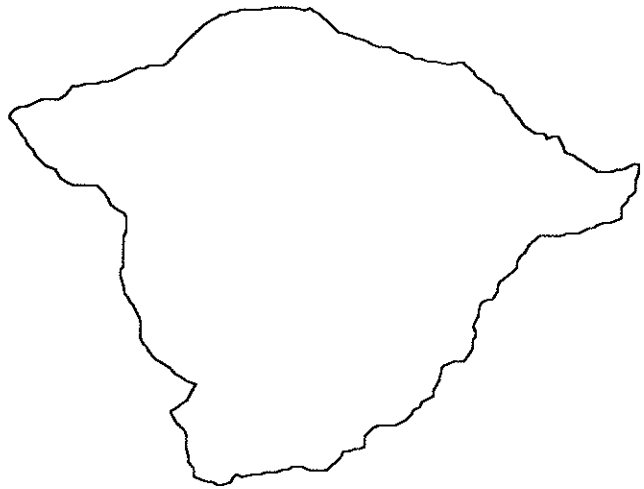
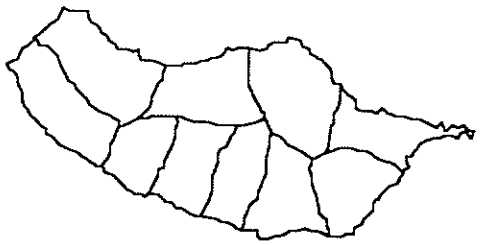


Fig. 18

ILHA DA MADEIRA (obtida pelo Método 1)

ILHA DA MADEIRA (obtida pelo Método 2)



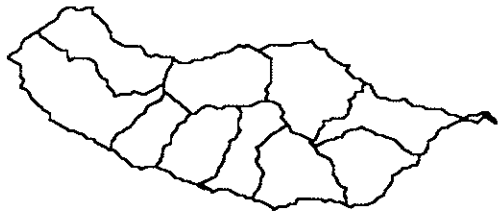


Fig. 19

Referências

Atlas Pro—Geographic Data Analysis and Presentation. Strategic Mapping, Inc.

Cogels, O. (1992) *Geomanagement system - a new concept for communication, integration, and analysis of georeferenced information*. New Techniques and Technologies for Statistics Conference, Bonn.

Cuffand, D. e Mattson, M. (1983) *Thematic maps. Their design and production*.

Leland, W. (1989) *The System for Statistics*. Evanston, IL: SYSTAT, Inc.

Lewis, P. (1977) *Maps and Statistics*. Methuen & Co.

Mosteller, F. e Tukey, J. (1977) *Data Analysis and Regression*. Addison-Wesley Publishing Company.

Scholter, H. e Lepper, M. (1992) *The benefits of the application of geographical information systems in public and environmental health*. New Techniques and Technologies for Statistics Conference, Bonn.

Smith, T., Menon, S., Star, J. e Estes, J. (1987) *Requirements and principles for the implementation and construction of large-scale geographic information systems*. Int. J. Geographical Informatios Systems, pp. 13-31.

Tukey, J. (1979) *Methodology, and the statistician's responsability for both accuracy and relevance*. JASA **74**, pp. 786-793.

Vasconcelos, R. (1992) *Doenças cardio-vasculares na Região Autónoma da Madeira. I Preparação de cartas geográficas para utilização estatística*. Notas e Cominicações do Centro de Estatística e Aplicações da U.L.-INIC **Nota nº 7**.

Capítulo III

Análise Exploratória de uma base de dados sobre Cardiologia. Atribuição de uma dimensão espacial à doença coronária.

1 — Introdução

A análise da base de dados do Serviço de Cardiologia do Centro Hospitalar do Funchal revela-se extremamente atraente essencialmente pelas seguintes razões:

— O Hospital do Funchal é o único a que podem recorrer os habitantes da Ilha da Madeira que sofram uma crise cardíaca, o que permite alargar as conclusões da análise à Região Autónoma da Madeira.

— Grande quantidade de informação sobre cada doente está informatizada como foi referido no Capítulo I.

— Foi possível acrescentar ao ficheiro de dados variáveis de informação geográfica, de modo a analisarmos a atribuição de uma dimensão espacial à doença coronária. Esta análise é extremamente relevante uma vez que as divisões administrativas têm profundas raízes históricas, e reflectem consequentemente aspectos demográficos, que no caso da Região Autónoma da Madeira serão em última análise uma tradução, ainda que imperfeita, dos diferentes *stocks* genéticos resultantes de vagas

colonizadoras diversificadas. Este facto justifica a opção de considerarmos como zonas geográficas de estudo os concelhos ou grupos de concelhos o que não será porventura a abordagem mais científica, e deveríamos deixar que os próprios dados indicassem os limites regionais. Mas, numa primeira abordagem, é decerto o mais conveniente.

2 — Sobre a modelação da distribuição geográfica das doenças

Ao desenvolvermos a análise exploratória desta base de dados, começámos pelo estudo global de cada uma das variáveis básicas do ficheiro de dados escolhidas sob a orientação dos especialistas de cardiologia. Este estudo rapidamente se revelou insuficiente para uma interpretação consistente das conclusões da análise sobre a incidência das doenças.

O estudo de cada variável por concelhos mostrou-se relevante na análise exploratória de uma base de dados deste tipo.

Na Análise de Sobrevivência desenvolvida no Capítulo I, já tínhamos concluído sobre a importância de uma análise exploratória por zonas geográficas.

Não nos parece, pois, despropositado, numa primeira abordagem, considerar a variável qualitativa "Concelho", e investigar da sua relevância na ocorrência e evolução das doenças coronárias em estudo.

O objectivo da análise exploratória dos dados do ficheiro do Serviço de Cardiologia do Centro Hospitalar do Funchal é detectar quais as variáveis e quais as interacções entre variáveis, que devem ser introduzidas no estudo estatístico que nos propomos realizar para representar o comportamento de duas doenças, o enfarto

agudo do miocárdio e a angina instável, na Região Autónoma da Madeira.

Uma análise exploratória das variáveis envolvidas, global e por concelhos, através da utilização de técnicas robustas e da interacção de técnicas de análise multivariada discreta aplicadas a dados demográficos da Região Autónoma da Madeira, sugere a inclusão de uma variável "Concelho" no modelo, atribuindo assim uma dimensão espacial às doenças. Tem a vantagem de ao fazermos o ajustamento de um modelo estatístico aos dados tornar mais clara a interpretação dos resíduos e, consequentemente, permitir formular um novo modelo que represente de uma forma mais próxima da realidade o comportamento daquelas doenças na Região.

A modelação da distribuição geográfica das doenças, consiste numa análise exploratória dos dados do ficheiro por zonas geográficas. Como referimos anteriormente, estas zonas geográficas podem ser formadas por concelhos ou grupos de concelhos.

A modelação da distribuição geográfica da doença tem sentido desde que se confirme que o seu comportamento não é idêntico em todos os concelhos, e permite-nos detectar formas de distribuições, semelhantes ou opostas, interessantes.

Para a modelação da distribuição geográfica das doenças será, por vezes, necessário recorrer a dados demográficos relativos à população residente na Região Autónoma da Madeira para consolidação da informação que os dados nos fornecem. Como o ficheiro de dados se refere aos anos de 1983 a 1991, poderia questionar-se se deveríamos usar os resultados dos Censos de 1981 ou os dos Censos de 1991 (cf. Apêndice IV). Consideramos que os resultados dos Censos de 1991 reflectem mais fielmente a situação demográfica da Região para os anos a que o ficheiro de dados se refere. Assim, optámos pelos resultados dos Censos de 91 sempre

que a respectiva utilização fosse sugerida pelo evoluir da análise exploratória, embora o facto de os resultados não serem definitivos suscite algumas reservas no ajustamento de modelos a esta colecção de dados.

Para a modelação geográfica da doença utilizaremos os mapas da Ilha da Madeira e da Ilha do Porto Santo que construímos com o objectivo de os utilizarmos em interacção com técnicas estatísticas na análise exploratória por concelhos, no programa Systat, recorrendo a comandos para maleabilidade suplementar, sempre que uma técnica exploratória que nos interesse não esteja disponível no *menu standard*. Através dos mapas usamos melhor a representação gráfica das variáveis em estudo. Trata-se de uma maneira de exprimir dados espaciais.

Como referimos, é uma técnica extraordinária para a ajuda numa procura "cega" de interacções num grupo de variáveis. Além disso, evitamos a situação de que é suficiente fazer o que é simples e obviamente exacto, dando ênfase à ideia de que temos de pensar, frequentemente, em termos de "dados = ajustamento + resíduos" e tentar sempre extrair mais dos resíduos (Tukey, 1977).

3 — A exposição de resultados da análise exploratória

Uma análise de dados é especialmente difícil de expor porque, pela sua estrutura, os diversos pontos de interesse a analisar não encaixam uns nos outros, nem parecem susceptíveis de um tratamento ordenado. Consequentemente, nós repetidamente ao trabalharmos um ponto de interesse novo, recorremos a resultados já anteriormente obtidos. Temos, frequentemente, vantagem em combinar, para a análise de uma mesma colecção de dados, resultados obtidos a partir de programas estatísticos

diferentes.

Na arte da análise de dados repetidamente precisamos de dar ênfase a resultados mais profundos para o que recorremos, frequentemente, a processos gráficos. Os gráficos são de grande valor essencialmente pelo papel que desempenham como indicadores. Muitas vezes, para conseguirmos que os dados forneçam uma indicação especial, pode ser necessária uma análise sofisticada e, até, complexa. As técnicas de indicação muitas vezes ultrapassam a estatística descritiva elementar.

Para um analista de dados, uma primeira avaliação dos dados consiste no que eles indicam, no que revelam, especialmente em estudos demográficos. Neste caso, geralmente, através dos mapas usamos melhor a representação gráfica e tornamos o observador mais eficiente em gerar métodos para extrair evidência. Ajuda a compreensão e interpretação do mundo que nos rodeia. Os recentes e extensos progressos da estatística espacial e da geoestatística mostram a crescente importância desta área pelo que não consideramos tempo perdido o dedicado à construção dos mapas.

Apresentaremos nos parágrafos seguintes uma sucessão de gráficos e de mapas que pensamos ser a forma mais correcta de apresentação dos resultados obtidos.

Para uma melhor visualização das referências que surgem no texto, incluiremos o mapa das Ilhas da Madeira e do Porto Santo com os concelhos devidamente identificados, como o que aparece na Fig.4.1 da secção seguinte, sempre que considerarmos conveniente.

ILHAS DA MADEIRA E DO PORTO SANTO

Fig.4.1

4 — Estudo do risco individual de contracção e da incidência das doenças

4.1 — *Um único grupo etário*

A Fig.4.2 apresenta as proporções de Habitantes, Doentes (H) (relativamente ao número de habitantes do Concelho na facha etária da incidência das doenças), Doentes (relativamente ao número total de doentes no ficheiro de dados) e de Mortes, para

cada concelho. Recorremos à standardização das proporções, para facilitar as comparações entre concelhos.

PROPORÇÕES DE : HABIT, DOENT(H), DOENT(D), MORT(D)
(STAND)

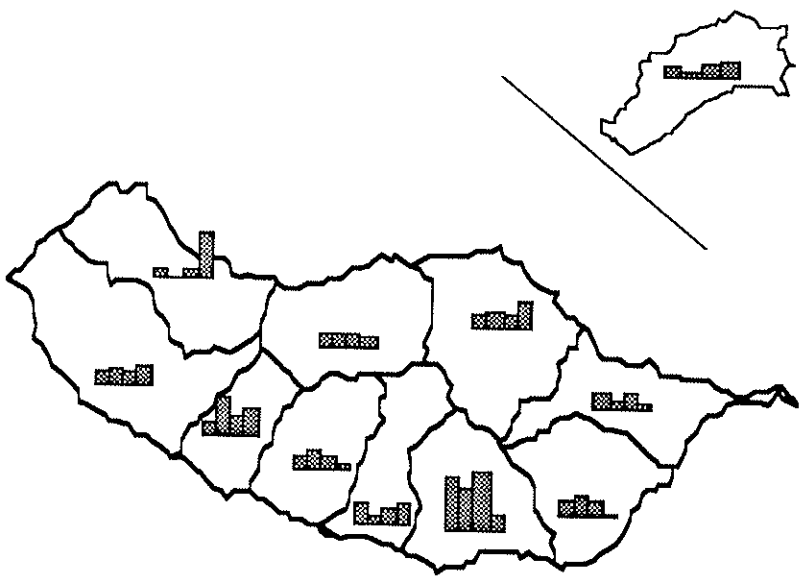


Fig.4.2

Comecemos por observar a proporção de Doentes(H) nos diferentes concelhos. O concelho com maior proporção de doentes é o do Funchal, seguido do concelho da Ponta do Sol. O concelho com menor proporção de doentes é o do Porto Moniz, seguido do Porto Santo e de Câmara de Lobos. Estas proporções foram calculadas com base no número de habitantes no concelho com 25 ou mais anos. Por indicação dos especialistas de cardiologia, a Idade é uma das variáveis que deve *a priori* ser considerada no

modelo estatístico que nos propomos construir e a incidência das doenças tem de ser considerada relativamente à população do concelho na facha etária atingida pelas doenças.

Na Fig.4.3 apresentamos o gráfico de caule-e-folhas das idades dos doentes.

Verificamos que os doentes têm idades compreendidas entre os 26 e os 91 anos —informação por demais vaga, imediatamente precisada pela informação que os 50% centrais estão entre os 61 e os 77 anos. Voltaremos à interpretação deste gráfico mais tarde.

GRAFICO DE CAULE-E-FOLHAS (IDADE DOS DOENTES)

```
2  66
3  2366
   ***OUTSIDE VALUES***
3  77
3  889999
4  0001111
4  222223333
4  444444455555
4  666667777777777
4  888888899999999
5  000000000111111111111111111111
5  22222222222223333333333333333
5  444444444444444444455555555555555555555
5  666666666667777777777777777777
5  888888888888888888888999999999999999999999999999
6H 000000000000000000000000000000001111111111111111111111111111111111111
6  22222222222222222222222222223333333333333333333333333333
6  44444444444444444444444444444444444444555555555555555555555555555555555555555555555
6  66666666666666666666666666666666666666677777777777777777777777777777777777777
6  88888888888888888888888888888888888888888999999999999999999999999999999999999999
7N 00000000000000000000000000000000000000000000000011111111111111111111111111111111111111111
7  22222222222222222222222222223333333333333333333333333333333333333
7  44444444444444444444444444444444444444455555555555555555555555555555555555555555555
7H 66666666666666666666666666666666666666666777777777777777777777777777777777777
7  88888888888888888888888888888888888999999999999999999999999999999999999999
8  00000000000000000000000000000001111111111111111111111
8  22222222222222222222222222222222222333333333333333333333333
8  44444444444444444455555555555555555555555555
8  66666666667777777777777
8  88888888899999999
9  0000000000000001
```

Fig.4.3

Ao consultarmos os resultados pré-definitivos dos Censos/91, verificamos que as classes de idades consideradas para os habitantes dos concelhos, são : 0 a 14 anos, 15 a 24 anos, 25 a 64 anos e 65 e mais anos. Não temos, pois, muita maleabilidade na escolha das classes de idades a considerar como sendo as classes típicas de incidência da doença, quando o nosso propósito, neste momento, é comparar dados populacionais dos concelhos, com dados correspondentes do ficheiro do Serviço de Cardiologia. Perdemos grande informação se desprezarmos a classe dos 25 a 64 anos, pois, como se pode verificar pelo gráfico de caule-e-folhas esta classe abrange uma proporção considerável de doentes. Assim, começámos por considerar, em cada concelho, a população com 25 e mais anos.

Estamos interessados em verificar se os dados fornecem evidência de que a incidência das doenças não é a mesma em todos os concelhos, ou seja, se algumas das diferenças indicadas na Fig.4.2 têm significado estatístico.

TABELA 4.1

CONCELHO	Nº DOENTES	Nº HABITANTES
CALHETA	38	7933
P.SOL	44	5117
P.MONIZ	4	2092
S.VICENTE	19	4562
R.BRAVA	44	7543
C.LOBOS	57	14627
FUNCHAL	703	68566
ST.CRUZ	76	13588
MACHICO	47	11682
SANTANA	30	6253
P.SANTO	8	2662
TOTAL	1070	144625

A Tabela 4.1 representa o número de doentes em cada concelho, entre os anos de 1983 e 1991. Acrescentámos à tabela o número de habitantes de cada concelho com 25 e mais anos.

Gostaríamos de poder analisar se o risco individual de contrair as doenças é o mesmo para todos os concelhos, no grupo etário considerado.

Para respondermos a esta questão considerámos o seguinte modelo para os dados da Tabela 4.1, considerando que o número de doentes num concelho entre os anos de 1983 e 1991 é uma variável aleatória de Poisson:

$$(4.1)\quad \begin{cases} X_i \cap P(\lambda_i) \\ \lambda_i \,/\, \lambda_\cdot = \psi_i N_i \,/\, N_\cdot \end{cases} \qquad i = 1,\ldots,I \qquad (4.2)$$

em que X_i representa o número de doentes entre 1983 e 1991, as variáveis X_i são independentes e N_i representa o número de habitantes, com 25 e mais anos, no concelho i, com i = 1,...,I. ψ_i representa o risco de contracção das doenças no concelho i.

A hipótese de que o risco de contracção das doenças é o mesmo em todos os concelhos, pode ser formulada da seguinte maneira:

$$H_0 : \psi_i = \psi \,, \quad \forall i \,.$$

em que ψ, que representa o risco comum de contracção das doenças, é a média ponderada dos riscos nos diversos concelhos, isto é,

$$\psi = \sum_i \psi_i N_i \,/\, N_\cdot = 1$$

De (4.2) sai, então, que a hipótese H_0 é equivalente a

$$\lambda_i / \lambda_. = N_i / N_. .$$

Para testar H_0 utilizaremos a estatística de teste de Pearson

$$X^2 = \sum_i (X_i - nN_i / N_.)^2 / (nN_i / N_.)$$

que, sob H_0, tem aproximadamente uma distribuição Qui-Quadrado com I-1 graus de liberdade.

Vejamos o que aconteceria se fôssemos aplicar o modelo considerado aos onze concelhos e testar H_0 .

O valor obtido para a estatística de teste, $x^2 = 161.9944$, é demasiado elevado e como podemos observar na Tabela 4.2 e na Fig.4.4, os resíduos fornecem indicação de que o modelo não é adequado aos dados referentes aos onze concelhos.

TABELA 4.2

CONCELHO	R.PEARSON
CALHETA	-2.7
P.SOL	1.0
P.MONIZ	-2.92
S.VICENTE	-2.54
R.BRAVA	-1.58
C.LOBOS	-4.92
FUNCHAL	8.69
ST.CRUZ	-2.45
MACHICO	-4.24
SANTANA	-2.39
P.SANTO	-2.64

Não é surpreendente que os resíduos sejam quase todos negativos, uma vez que estamos a comparar concelhos com populações, no grupo etário considerado, de dimensões muito diferentes. O facto de o concelho do Funchal ter um número de habitantes e, consequentemente, um número de doentes, muito superiores aos dos outros concelhos, dá origem a que exista uma forte tendência para as frequências esperadas das células serem sistematicamente maiores do que as respectivas frequências observadas, originando, assim, resíduos negativos.

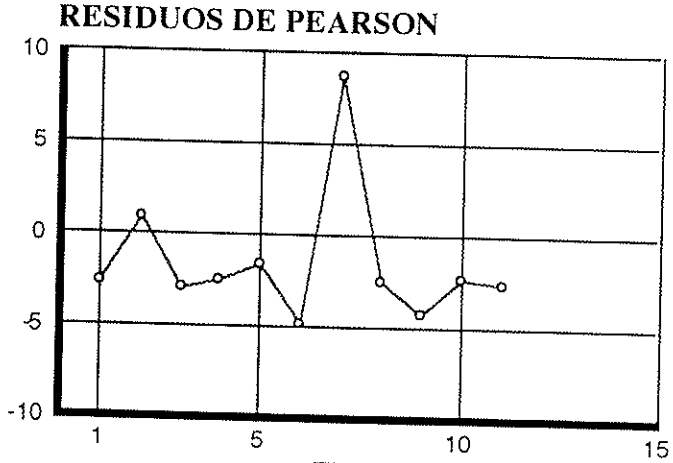


Fig.4.4

O modelo seria adequado e poderíamos testar a hipótese H_0 por forma a obtermos uma resposta à questão levantada, se os onze concelhos tivessem populações naquele grupo etário com dimensões não muito diferentes.

No Capítulo IV apresentamos um estudo detalhado sobre a escolha da estatística de teste de Pearson e sobre as questões levantadas sobre a aplicabilidade deste teste.

Na impossibilidade de responder à questão formulada relativamente aos onze concelhos, podemos, contudo, responder à mesma questão mas apenas relativamente aos grupos de concelhos com dimensões não muito diferentes.

Formámos quatro grupos de concelhos. O Grupo 1 inclui os concelhos de São Vicente, da Ponta do Sol e de Santana. O Grupo 2 é formado pelos concelhos de Câmara de Lobos, Santa Cruz e Machico. O Grupo 3 inclui os concelhos da Calheta e da Ribeira Brava. O Grupo 4 é formado pelos concelhos do Porto Moniz e do Porto Santo. Ao aplicarmos o mesmo modelo a cada um destes grupos, verificámos que no Grupo 1 H_0 é rejeitada, enquanto que nos restantes grupos os dados não fornecem evidência para rejeitarmos H_0 . Na Tabela 4.3 apresentamos os resultados obtidos.

TABELA 4.3

GRUPOS	X^2	GR. DE LIB.	VALOR CRITICO
GRUPO1	10.02864	2	5.991
GRUPO2	5.36557	2	5.991
GRUPO3	0.6726	1	3.841
GRUPO4	0.55462	1	3.841

Os riscos considerados no modelo são riscos relativos. Podemos apenas referir-nos ao risco num determinado concelho quando comparado com os riscos nos restantes concelhos incluídos no modelo.

Rejeitar H_0, quando aplicámos o modelo ao Grupo 1, significa, pois, que se considerarmos os concelhos de São Vicente, Ponta do Sol e Santana, os dados não fornecem evidência de que o risco de contrair as doenças seja idêntico nos três concelhos. O gráfico apresentado na Fig.4.2 já nos tinha fornecido indicação de que, relativamente aos habitantes na faxa etária considerada, Ponta

do Sol era o segundo concelho com maior proporção de doentes. Pela análise dos resíduos de Pearson obtidos, verificamos que o maior corresponde a este concelho. Temos pois indicação de que o concelho da Ponta do Sol tem um maior risco individual de contracção das doenças relativamente aos concelhos de São Vicente e de Santana.

Uma nota tem de ser acrescentada: estamos a trabalhar com populações de dimensões grandes. É difícil estabelecer a partir de que valor é que a diferença entre as dimensões das populações vai afectar a análise.

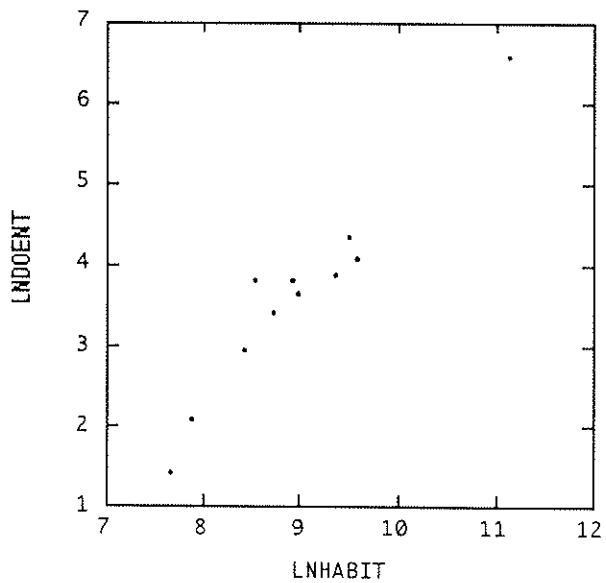


Fig.4.5 - Relação entre ln(Nº Habitantes) e ln(Nº Doentes)

Devido à impossibilidade de comparar simultaneamente um

maior número de concelhos quanto ao risco individual de contracção das doenças, decidimos estudar as proporções de doentes em cada concelho de uma forma mais rigorosa. Ajustámos aos dados transformados uma recta resistente de três grupos. A escolha da transformação utilizada é justificada por os dados provirem de contagens e pelo gráfico apresentado na Fig.4.5 .

Obtivémos a seguinte recta ao fim da segunda iteração :

$$\hat{y} = -7.447 + 1.22802\ x\ .$$

Apresentamos os resíduos obtidos, ordenados, na Tabela 4.4 e, na Fig. 4.6 representamo-los relativamente aos correspondentes valores estimados dos logaritmos dos números de doentes.

TABELA 4.4

CONCELHO	RESIDUO	RESID. CENTR. E REDUZ.
P. MONIZ	-0.556	-1.6514
C. LOBOS	-0.2873	-0.9293
MACHICO	-0.2043	-0.7063
P. SANTO	-0.1587	-0.5837
S. VICENTE	0.0447	-0.0371
CALHETA	0.0585	0
ST. CRUZ	0.0908	0.0868
SANTANA	0.1143	0.15
R. BRAVA	0.2669	0.5601
FUNCHAL	0.3277	0.7235
P. SOL	0.7435	1.841

Uma análise dos resíduos, centrados pela mediana e reduzidos pela dispersão-quartal, indica-nos que, efectivamente, o

concelho da Ponta do Sol, (o maior resíduo em valor absoluto), tem um número de doentes substancialmente maior do que o esperado se admitirmos uma relação linear entre o logaritmo do número de habitantes com 25 anos e mais, (x), e o logaritmo número de doentes, (y). O mesmo se pode dizer do concelho do Funchal. Admitindo a mesma relação, os concelhos do Porto Moniz, de Câmara de Lobos e de Machico têm um pequeno número de doentes relativamente ao número de habitantes. No entanto, nenhum concelho constitui um "outlier". A colecção de resíduos centrados e reduzidos tem barreiras de "outliers" compreendidas entre os valores -2 e +2. Nenhum dos resíduos apresentados na Tabela 4.4 está fora destas barreiras.

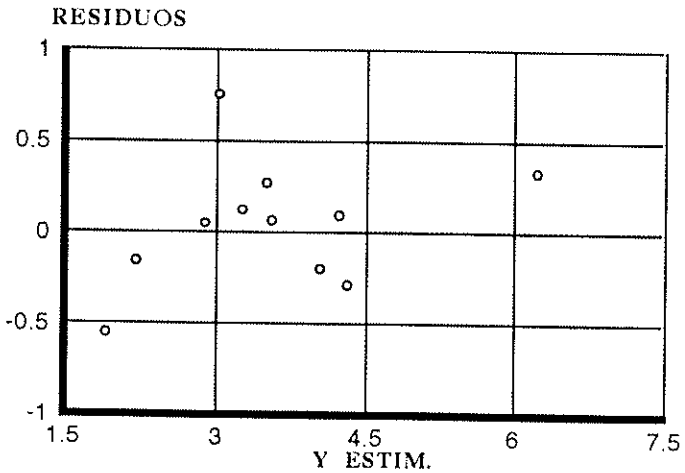


Fig.4.6 - Representação gráfica dos resíduos relativamente aos valores de y estimados.

4.2 — *Dois grupos etários*

As incidências que acabámos de estudar são, ainda, incidências grosseiras uma vez que ignoram a informação adicional fornecida pela classificação das idades. Cerca de 33% dos doentes têm idades compreendidas entre os 25 e os 64 anos, e cerca de 67% têm idades superiores ou iguais a 65 anos. Esta informação aliada ao facto de na população aquele primeiro grupo etário ser muito mais representativo do que o segundo, dá-nos indicação que as doenças em causa atingem mais um grupo etário do que outro. Assim uma forma menos grosseira de estudar a incidência das doenças nos diferentes concelhos, e o risco de contracção das doenças nos diferentes concelhos, deveria passar por incluir na análise a variável Idade com dois níveis—um nível representaria o grupo etário dos 25 aos 64 anos, o outro nível representaria o grupo com 65 e mais anos. Consideramos que existe um risco individual de contracção das doenças para cada concelho e para cada grupo etário.

Para estudarmos o risco individual de contracção das doenças nos diversos concelhos, temos de seleccionar os concelhos com populações de dimensões não muito diferentes e, além disso, em que a distribuição da Idade seja idêntica.

Assim, para os grupos de concelhos considerados no caso anterior, ajustámos um modelo log-linear sem interacção

$$\ln E[Y_{ij}] = u + u_{1(i)} + u_{2(j)}$$

sujeito às restrições

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0$$

em que $Y_{ij} \cap P(\lambda_{ij})$, representa o número de habitantes do Concelho i no grupo etário j .

O facto de este modelo ser adequado aos dados de um grupo de concelhos implica que, para uma tabela de contingência IxJ, a distribuição da Idade é a mesma nos concelhos que constituem esse grupo. Basta ter em conta que, se designarmos por m_{ij} a frequência esperada na célula (i,j) da tabela, sob a hipótese da interacção ser nula, tem-se

$$\frac{m_{ij}}{m_{rj}} = e^{u1(i) - u1(r)} = \frac{m_{i.}}{m_{r.}}, \quad \forall j = 1,\ldots,J$$

e

$$\frac{m_{ik}}{m_{it}} = e^{u2(k) - u2(t)} = \frac{m_{.k}}{m.t}, \quad \forall i = 1,\ldots,I.$$

Logo,

$$\frac{m_{ij}}{m_{i.}} = \frac{\frac{m_{i.} \times m_{rj}}{m_{r.}}}{m_{i.}} = \frac{m_{rj}}{m_{r.}}, \quad \forall j = 1,\ldots,J .$$

O modelo foi rejeitado em todos os grupos de concelhos. No entanto, no Grupo1, a análise dos resíduos forneceu indicação de que as células que mais contribuíam para a falta de ajustamento eram as correspondentes ao concelho da Ponta do Sol. Ajustámos, então, o mesmo modelo apenas aos dados dos concelhos de São Vicente e Santana. O modelo não foi rejeitado, pelo que podemos

considerar que estes concelhos são "comparáveis", isto é, têm populações de dimensões não muito diferentes e a distribuição da Idade é idêntica nos dois concelhos.

A Tabela 4.5 é a tabela de contingência correspondente ao número de doentes nestes dois concelhos, à qual acrescentámos o número de habitantes em cada grupo etário e em cada concelho.

TABELA 4.5

IDADE / CONCELHO	25 A 64	65 E MAIS	TOTAL	Nº DE HABIT
S.VICENTE	6	13	19	4562
SANTANA	8	22	30	6253
TOTAL	14	35	49	
Nº DE HABIT	8290	2525		10815

Considerámos o seguinte modelo para os dados :

$$\begin{cases} X_{ij} \cap P\left(\lambda_{ij}\right) \\ \lambda_{ij} / \lambda_{..} = \theta_{ij} N_{ij} / N_{..} \end{cases}$$

em que X_{ij} representa o número de doentes do concelho i no grupo etário j e as variáveis X_{ij}, i,j=1,2, são independentes. N_{ij} representa o número de habitantes no concelho i com idade no grupo etário j. θ_{ij} representa o risco individual de contracção das doenças no concelho i e na classe etária j.

Como admitimos que não existe interacção entre a Idade e o

Concelho, nas populações dos dois concelhos, podemos escrever:

$$\frac{N_{ij}}{N_{i.}} = \frac{N_{.j}}{N_{..}}$$

e, portanto,

$$\frac{N_{ij}}{N_{..}} = \frac{N_{i.}}{N_{..}} \times \frac{N_{.j}}{N_{..}} .$$

Se considerarmos que φ_i representa o risco no nível i da variável Concelho e ψ_j representa o risco no nível j da variável Idade, a hipótese de não-interacção entre as variáveis Concelho e Idade no que diz respeito ao risco de contrair as doenças pode, então, ser expressa como

$$\theta_{ij} = \varphi_i \times \psi_j$$

que corresponde a

$$\lambda_{ij} = \lambda_{..}\varphi_i\psi_j N_{i.}N_{.j} / N_{..}^2 \qquad (4.3)$$

ou ao modelo log-linear sob a hipótese de não-interacção

$$\ln \lambda_{ij} = u + u_{1(i)} + u_{2(j)}$$

sujeito às restrições

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0,$$

$$\text{com} \quad \lambda_{..} = e^{u}, \quad \varphi_i \frac{N_{i.}}{N_{..}} = e^{u_{1(i)}} \quad \text{e} \quad \psi_j \frac{N_{.j}}{N_{..}} = e^{u_{2(j)}} .$$

Fomos, então, testar o ajustamento deste modelo aos dados da Tabela 4.5. Se o modelo não for rejeitado, o que significa que os dados não fornecem evidência da existência de interacção entre as variáveis Concelho e Idade, estamos em condições de testar hipóteses marginais que dizem respeito ao risco de contracção das doenças nos concelhos.

No caso de rejeitarmos o ajustamento do modelo podemos, apenas, concluir que o risco de um indivíduo de São Vicente, por exemplo, contrair as doenças depende do grupo etário a que pertence.

O valor observado da estatística de teste foi $g^2 = 0.14$, o que levou à não rejeição do modelo.

Estamos, pois, em condições de testar a hipótese de o risco de contracção das doenças ser o mesmo nos dois concelhos, que é formulada da seguinte maneira:

$$H_0' : \varphi_i = \varphi$$

em que $\varphi = \sum_i \varphi_i \frac{N_{i.}}{N_{..}}$.

Note-se que $\varphi\psi = 1$, pois de (4.3) vem que

$$\sum_i \sum_j \lambda_{ij} = \lambda_{..} \varphi\psi \Leftrightarrow \varphi\psi = \frac{\lambda_{\cdots}}{\lambda_{..}} = 1$$

De (4.3) sai que H_0' é equivalente a

$$\lambda_{i.} / \lambda_{..} = N_{i.} / N_{..} .$$

Como se pode verificar, de (4.3), podemos escrever

$$\sum_j \lambda_{ij} = \lambda_{..} \varphi_i \frac{N_{i.}}{N_{..}} \sum_j \psi_j \frac{N_{.j}}{N_{..}} \qquad (4.4)$$

que, admitindo $\varphi_i = \varphi$, pode, ainda, ser escrito

$$\frac{\lambda_{i.}}{\lambda_{..}} = \frac{N_{i.}}{N_{..}} \varphi\psi .$$

Tendo em conta que $\varphi\psi = 1$, tem-se, então

$$\frac{\lambda_{i.}}{\lambda_{..}} = \frac{N_{i.}}{N_{..}} .$$

Por outro lado, admitindo que

$$\lambda_{i.} / \lambda_{..} = N_{i.} / N_{..}$$

de (4.4), podemos escrever

$$\varphi_i \psi = 1$$

o que é equivalente a

$$\varphi_i = \frac{1}{\psi} .$$

Como $\varphi\psi = 1$, vem que

$$\varphi_i = \varphi .$$

Para testarmos H'_0 a estatística de teste da razão de verosimilhanças é

$$G^2 = 2 \sum_{i=1}^{2} X_{i.} \left(\ln X_{i.} - \ln \left(n \frac{N_{i.}}{N_{..}} \right) \right)$$

que segue aproximadamente uma distribuição Qui-Quadrado com 1 grau de liberdade.

Ou, se usarmos a estatística de teste de Pearson,

$$X^2 = \sum_{i=1}^{2} \frac{\left(X_{i.} - n \frac{N_{i.}}{N_{..}} \right)^2}{n \frac{N_{i.}}{N_{..}}}$$

que segue aproximadamente a mesma distribuição.

O valor que obtivémos para a estatística de teste da razão de verosimilhanças foi g^2=1.33653. Logo, a hipótese de o risco individual de contracção das doenças ser o mesmo nos dois concelhos, não pode ser rejeitada.

Deparámos, novamente, com a situação de existirem muito poucos concelhos "comparáveis", o que não nos permitiu avançar

muito mais nas conclusões sobre o risco individual de contracção das doenças nos concelhos.

Mas, o facto de a distribuição da Idade não ser idêntica nos concelhos do Grupo1, havendo indicação através de uma análise dos resíduos, de que o concelho da Ponta do Sol tem um número elevado de habitantes no grupo etário dos 65 e mais anos, levou-nos a estudar o número de doentes em cada concelho relativamente ao número de habitantes do concelho no grupo etário de 65 e mais anos, através do ajustamento aos dados transformados de uma recta resistente de três grupos. Optámos por trabalhar com os dados transformados (transformação logarítmica das duas variáveis), o que é justificado pelo gráfico apresentado na Fig.4.7.

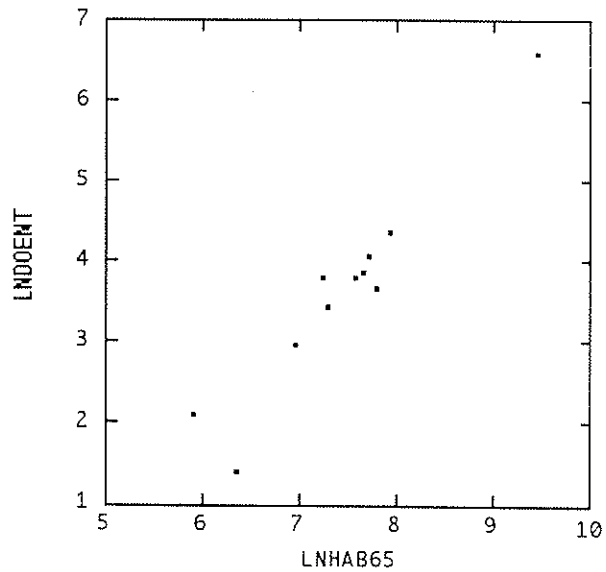


Fig.4.7 - Relação entre ln(Nº Habitantes de 65 e mais anos) e ln(Nº Doentes)

A recta resistente ajustada ao fim da terceira iteração foi

$$\hat{y} = -5.2476 + 1.19866x \ ,$$

em que y =ln(nº de doentes) e x=ln(nº de habitantes com 65 e mais anos).

Na Tabela 4.6 apresentamos os dados, os resíduos e os resíduos centrados pela mediana e reduzidos pela dispersão-quartal.

TABELA 4.6

CONCELHO	LN X	LN Y	RESIDUOS	RESID. CENTR. E REDUZ.
P.MONIZ	6.3509	1.3863	-0.9787	-3.6097
C.LOBOS	7.7253	4.0431	0.0307	0.2644
MACHICO	7.6549	3.8501	-0.0779	-0.1524
P.SANTO	5.9216	2.0794	0.229	1.0255
S.VICENTE	6.9594	2.9444	-0.1504	-0.4306
CALHETA	7.8124	3.6376	-0.4792	-1.6926
ST.CRUZ	7.9516	4.3307	0.047	0.327
SANTANA	7.2944	3.4012	-0.0947	-0.2168
R.BRAVA	7.5668	3.7842	-0.0382	0
FUNCHAL	9.4664	6.5554	0.456	1.8968
P.SOL	7.2513	3.7842	0.34	1.4515

4.3 — *Conclusão*

O concelho do Porto Moniz constitui um "outlier", com um número de doentes demasiado pequeno relativamente ao número

de habitantes do concelho com 65 e mais anos. Na análise que realizámos anteriormente tínhamos referido que este concelho tinha poucos doentes em relação ao número de habitantes com 25 e mais anos.

Câmara de Lobos tem um número de doentes baixo relativamente ao número de habitantes com 25 e mais anos, mas o mesmo não se pode concluir relativamente aos habitantes do concelho com 65 e mais anos. Neste grupo etário, o número de doentes é coerente com o ajustamento linear aos dados transformados.

O concelho de Machico tem, também, tendência para ter poucos doentes relativamente ao número de habitantes com 25 e mais anos, o mesmo se podendo dizer relativamente ao grupo etário de 65 e mais anos.

Quanto ao concelho do Porto Santo, os resíduos fornecem indicação de que o número de doentes é inferior ao que seria coerente com o número de habitantes com 25 e mais anos mas, relativamente ao número de habitantes no grupo etário de 65 e mais anos, o número de doentes é elevado.

São Vicente tem um número de doentes coerente com o número de habitantes do concelho com 25 e mais anos, o mesmo se podendo concluir relativamente ao grupo etário de 65 e mais anos.

No concelho da Calheta, o número de doentes relativamente à população do concelho com 25 e mais anos é coerente com o ajustamento linear aos dados transformados. No que diz respeito ao número de doentes relativamente ao número de habitantes no grupo etário 65 e mais anos, a situação é bem diferente. Neste caso o resíduo evidencia uma tendência acentuada para o número de doentes ser baixo.

O concelho de Santa Cruz não é um concelho em que o número de doentes se afaste dos padrões esperados sob um ajustamento linear aos dados transformados, para o grupo etário de 25 e mais anos. O mesmo se pode dizer relativamente à classe de 65 e mais anos.

O número de doentes do concelho de Santana, é coerente com o número de habitantes do concelho com 25 e mais anos. O mesmo se pode dizer em relação ao grupo etário de 65 e mais anos.

A análise dos resíduos indica que o concelho da Ribeira Brava tem um número de doentes coerente com o número de habitantes, quer relativamente ao grupo etário de 25 e mais anos, quer relativamente ao grupo de 65 e mais anos.

No concelho do Funchal, o número de doentes tem tendência a ser elevado relativamente ao número de habitantes com 25 e mais anos, e esta tendência acentua-se no grupo etário de 65 e mais anos. No entanto, o senso comum alerta-nos para a proximidade do Hospital , facto que tem de ser tido em conta pela maior facilidade com que um indivíduo do grupo etário 65 e mais anos pode recorrer ao atendimento do Serviço de Cardiologia.

O concelho da Ponta do Sol é o que apresenta maior número de doentes relativamente ao número de habitantes do concelho no grupo etário de 25 e mais anos. É de realçar que este número de doentes não é coerente com o número de habitantes com 65 e mais anos. O número de doentes relativamente ao número de habitantes neste grupo etário é, também, bastante elevado.

De tudo o que foi analisado, podemos concluir que as doenças afectam de forma diferente as populações de diversos concelhos, quer no que diz respeito ao grupo etário mais atingido pelas doenças, quer no que se refere ao número de doentes , existindo evidência de que o concelho da Ponta do Sol é o de maior incidência da doença e que os concelhos de Porto Moniz, de Machico e de São Vicente são os de menor incidência das doenças.

Quanto ao risco individual de contracção das doenças, dados os grupos etários considerados, pudémos apenas concluir que os dados fornecem evidência estatística de que relativamente ao grupo

de concelhos da Ponta do Sol, de São Vicente e de Santana, o primeiro apresenta um risco individual de contracção das doenças maior do que os restantes, dado o grupo etário de 25 e mais anos. Não existe evidência estatística de que, dado o mesmo grupo etário, os concelhos de Câmara de Lobos, Santa Cruz e Machico apresentem riscos diferentes de contracção das doenças. O mesmo se pode dizer em relação aos concelhos da Calheta e da Ribeira Brava. Os dados referentes aos concelhos do Porto Moniz e do Porto Santo, também não fornecem evidência estatística de que o risco de contracção das doenças seja maior num destes concelhos relativamente ao outro. Se considerarmos os grupos etários de 25 a 64 anos e de 65 e mais anos, apenas tivémos condições de comparar o risco individual de contracção das doenças nos concelhos de São Vicente e de Santana. Os dados não forneceram evidência estatística de que existam diferenças nestes dois concelhos quanto ao risco individual de contracção das doenças.

A análise do risco individual de contracção das doenças ficou, assim, limitada ao respectivo estudo comparativo em pequenos grupos de concelhos, uma vez que os concelhos apresentam características demográficas suficientemente diferentes para não nos permitirem uma comparação global.

Tendo em conta a existência de evidência estatística de que o risco individual de contracção das doenças de entre os poucos casos susceptíveis de comparação não é o mesmo em todos os concelhos, e os resultados da análise da incidência das doenças, estamos em condições de concluir que uma modelação da distribuição geográfica das doenças se torna, pois, relevante na análise exploratória do ficheiro de dados do Serviço de Cardiologia do Centro Hospitalar do Funchal.

5 — Estudo da variável Morte

5.1 — *Estudo do Risco Individual e da Incidência da Mortalidade*

Se observarmos, no gráfico apresentado na Fig.4.2, a proporção de mortes por concelho, vemos que existe indicação de o concelho do Porto Moniz apresentar a maior proporção de mortes, seguido dos concelhos de Santana e da Ponta do Sol. Os concelhos de Santa Cruz, Machico e da Ribeira Brava são os que apresentam menor proporção de mortes. Estas proporções foram calculadas relativamente ao número de doentes de enfarto e de angina do concelho.

Aplicámos aos dados dos concelhos com números de doentes não muito diferentes o modelo (4.1), para testar a hipótese de o risco individual de mortalidade ser idêntico neste grupo de concelhos. Este grupo era constituído pelos concelhos da Calheta, da Ponta do Sol, da Ribeira Brava e de Machico. A Tabela 5.1 representa os dados referentes ao número de mortes e ao número de doentes de cada concelho.

TABELA 5.1

CONCELHO	Nº MORTOS	Nº DOENTES
CALHETA	11	38
P.SOL	14	44
R.BRAVA	8	44
MACHICO	8	47
TOTAL	41	173

Neste caso X_i representa o número de mortes no concelho i e

N_i representa o número de doentes no concelho i, com i=1,...,4 . O valor obtido para a estatística de teste de Pearson foi $x^2 = 3.11505$, o que significa que os dados não fornecem evidência para rejeitarmos a hipótese de igualdade do risco individual de mortalidade nos quatro concelhos.

Estudámos, então, as proporções de mortes nos diferentes concelhos de uma forma mais rigorosa, à semelhança do que fizémos no parágrafo anterior.

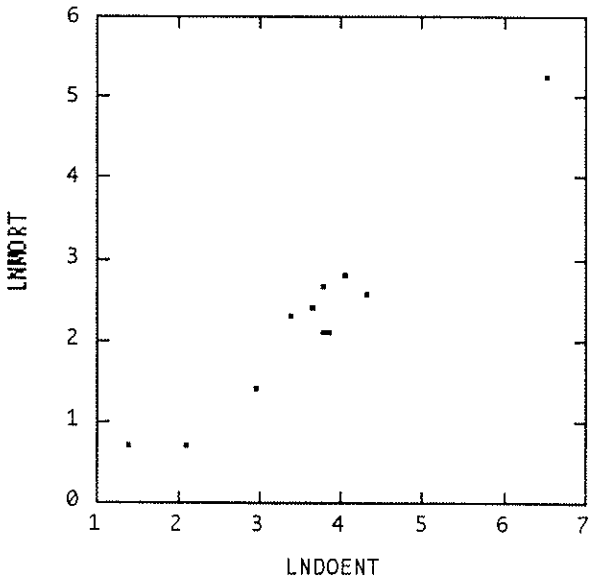


Fig.5.1 - Relação entre ln(Nº Mortos) e ln(NºDoentes)

Ajustámos aos dados transformados uma recta resistente de três grupos. Na Fig.5.1 representamos os pontos correspondentes aos dados transformados, que sugerem o ajustamento linear.

Obtivémos a seguinte recta

$$\hat{y} = -0.6639 + 0.81324x \ .$$

A Tabela 5.2 apresenta os resíduos e os resíduos centrados pela mediana e reduzidos pela dispersão-quartal. A análise dos resíduos fornece indicação de que os concelhos de Machico, de São Vicente, da Ribeira Brava, do Porto Santo e de Santa Cruz têm um número de mortes baixo relativamente ao número de doentes do concelho. Os concelhos do Funchal, do Porto Moniz e da Ponta do Sol têm um número de mortes elevado em relação ao número de doentes do concelho.

TABELA 5.2

CONCELHO	LN X	LN Y	RESIDUOS	RESID. CENTR. E REDUZ.
P.MONIZ	1.3863	0.6931	0.2296	0.2303
C.LOBOS	4.0431	2.7726	O.1485	0.0821
MACHICO	3.8501	2.0794	-0.3878	-0.8981
P.SANTO	2.0794	0.6931	-0.3341	-0.8
S.VICENTE	2.9444	1.3863	-0.3443	-0.8186
CALHETA	3.6376	2.3979	0.1036	0
ST.CRUZ	4.3307	2.565	-0.293	-0.7248
SANTANA	3.4012	2.3026	0.2005	0.1771
R.BRAVA	3.7842	2.0794	-0.3342	-0.8001
FUNCHAL	6.5554	5.2575	0.5903	0.8895
P.SOL	3.7842	2.6391	0.2255	0.2228

Segundo os especialistas de cardiologia, bastaria que as populações dos concelhos que referimos como tendo um número

baixo de mortes relativamente ao número de doentes tivessem comportamentos idênticos, no que diz respeito à Idade e ao Sexo, para explicar a indicação obtida do ajustamento da recta resistente.

Recorrendo, mais uma vez, aos resultados pré-definitivos dos Censos/91 e, tendo em conta que os modelos log-lineares constituem uma útil descrição de uma estrutura de dados, tentámos conhecer melhor o comportamento das populações que constituem os cinco concelhos acima referidos. No plano subjacente à obtenção desta colecção de dados não houve, a priori, qualquer restrição na dimensão da mesma. Assim, X_{ijk}, que representa o número de indivíduos que foram colocados no nível i da variável Sexo, no nível j da variável Idade e no Concelho k, tem uma distribuição de Poisson de parâmetro λ_{ijk}. Para a nossa tabela a distribuição de probabilidade é

$$f\left(\left\{x_{ijk}\right\}\right) = \prod_{i,j,k} \frac{\lambda_{ijk}^{x_{ijk}} e^{-\lambda_{ijk}}}{x_{ijk}!} .$$

O modelo log-linear saturado correspondente é

$$\ln E[X_{ijk}] = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

sujeito às habituais restrições

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = 0 ,$$

$$\sum_i u_{12(ij)} = \sum_i u_{13(ik)} = \sum_j u_{23(jk)} = \sum_j u_{12(ij)} = \sum_k u_{13(ik)}$$

$$= \sum_k u_{23(jk)} = 0,$$

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0 .$$

Testámos a hipótese

$$H_0: \; u_{123(ijk)} = 0, \quad \forall i,j,k$$

para verificar se o modelo

$$\ln E\left[X_{ijk}\right] = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

sujeito às restrições

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = 0$$

$$\sum_i u_{12(ij)} = \sum_i u_{13(ik)} = \sum_j u_{23(jk)} = \sum_j u_{12(ij)} = \sum_k u_{13(ik)}$$

$$= \sum_k u_{23(jk)} = 0,$$

constitui uma adequada descrição da estrutura da população

daqueles cinco concelhos.

O valor obtido para a estatística do teste da razão de verosimilhanças foi g^2=25.87, o que significa que os dados fornecem evidência para rejeitarmos H_0. Os dados evidenciam a existência de uma interacção de segunda ordem entre as variáveis Sexo, Idade e Concelho. Concluímos, assim, que a interacção entre as variáveis Sexo e Idade não é a mesma nos diferentes concelhos. No entanto, a análise dos resíduos de Pearson, revelou que o concelho da Ribeira Brava é o que apresenta maiores resíduos. Assim, ajustámos o mesmo modelo aos dados dos concelhos de Machico, Porto Santo, São Vicente e Santa Cruz. O valor obtido para a estatística de teste foi $g^2 = 9.22$, o que levou à rejeição do modelo. Desta vez, o concelho de São Vicente apresentava o resíduo que mais contribuía para o desajustamento do modelo. Como tal, ajustámos o mesmo modelo mas, apenas aos concelhos de Machico, do Porto Santo e de Santa Cruz. O valor observado para a estatística de teste foi $g^2 = 2.52$, o que levou à não rejeição da hipótese de estes três concelhos terem comportamentos idênticos no que se refere à interacção entre as variáveis Sexo e Idade.

Fica, então, por explicar o baixo número de mortes em relação ao número de doentes dos concelhos da Ribeira Brava e de São Vicente.

O evoluir da análise exploratória, nomeadamente na análise da variável Diagnóstico, poderá esclarecer (ou não) o que ficou por explicar, realçando, uma vez mais a relevância de uma análise exploratória por concelhos.

5.2 — *Conclusão*

Neste momento, em relação a estas doenças, poderíamos ironicamente, (uma vez que consideramos que a variável Concelho representa muito mais do que o simples facto de um indivíduo ser habitante de um dado concelho), dizer que o concelho de Machico é o local ideal para vivermos, quer o nosso grupo etário seja o de 25 a 64 anos, quer seja o de 65 e mais anos. A incidência das doenças é baixa em relação ao número de habitantes e o número de mortes também é baixo em relação ao número de doentes.

O concelho do Porto Santo também pode ser uma boa escolha para vivermos, se o nosso grupo etário for o de 25 a 64 anos. O número de doentes é baixo em relação ao número de habitantes neste grupo etário, e o número de mortes também é baixo em relação ao número de doentes.

O concelho de Câmara de Lobos também tem baixa incidência da doença em relação aos habitantes no grupo etário de 25 a 64 anos e o número de mortes é coerente com o número de doentes.

Quanto aos concelhos menos atraentes, sob o ponto de vista destas doenças, temos o concelho da Ponta do Sol como sendo o pior. É o de maior incidência das doenças e tem, também, um elevado número de mortes em relação ao número de doentes. O concelho do Funchal com tendência para um elevado número de doentes, francamente acentuada no grupo etário de 65 e mais anos, e com um elevado número de mortes em relação ao número de doentes, também não se apresenta atraente.

6 — Estudo da variável Sobrevida

6.1 — *Estudo global*

A Fig.6.1 representa o gráfico de caule-e-folhas para a variável Sobrevida.

```
 0  H000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
 1  0113344556666667889
 2  0036699
 3  0224445666689999
 4  11223355689
 5  013344689
 6  H223456799
 7  1127899
 8  01234888
 9  1236
10  24667
11  1389
12  012358
13  2335569
14  5777
15  0
    ***OUTSIDE VALUES***
16  24999
17  01
18  1
19  25
20  1478
21  4
22  148
24  16
```

Fig.6.1 - Gráfico de Caule-e-Folhas da Sobrevida

Para a execução deste gráfico, começámos por detectar os

valores periféricos da Sobrevida, com base na respectiva posição em relação às barreiras exteriores. A mediana desta colecção de dados é 74 dias, o quarto inferior é 3 dias e o quarto superior é 640 dias, o que, considerando como barreiras exteriores os pontos obtidos de

$$\text{Quarto Inferior} - 1.5 \times \text{Dispersão} - \text{Quartal}$$

e

$$\text{Quarto Superior} + 1.5 \times \text{Dispersão} - \text{Quartal}$$

nos levou a basear a escolha da escala de representação nos dados correspondentes a sobrevidas inferiores a 1600 dias. O número máximo de linhas obtido, aplicando a expressão

$$\text{Nº máximo de linhas} = [10 \times \log_{10} 260]$$

foi 24. Dividindo a amplitude desta colecção de dados pelo número máximo de linhas, obtivémos o valor 62.5 que arredondando à potência de 10 mais próxima, nos levou a considerar o valor 100 para comprimento do intervalo. Note-se que a primeira linha tem um aspecto muito pesado por ter demasiadas folhas mas, se optássemos pela representação com duas linhas por caule, o que equivalia a arredondar 62.5 para 50, íamos obter um gráfico ilegível no formato em que este texto é apresentado.

A análise deste gráfico fornece-nos indicação de que existe um número considerável de doentes com sobrevida nula. Mais de 50% dos doentes tiveram sobrevida inferior a 100 dias. 25% dos doentes tiveram uma sobrevida superior a 640 dias e o número de doentes cuja sobrevida ultrapassou os 1600 dias não é desprezável, sendo a sobrevida máxima observada de 2466 dias.

Esta análise sugere imediatamente, a análise da dispersão da

variável Sobrevida.

A Fig.6.2 representa a caixa-de-bigodes para esta colecção de dados.

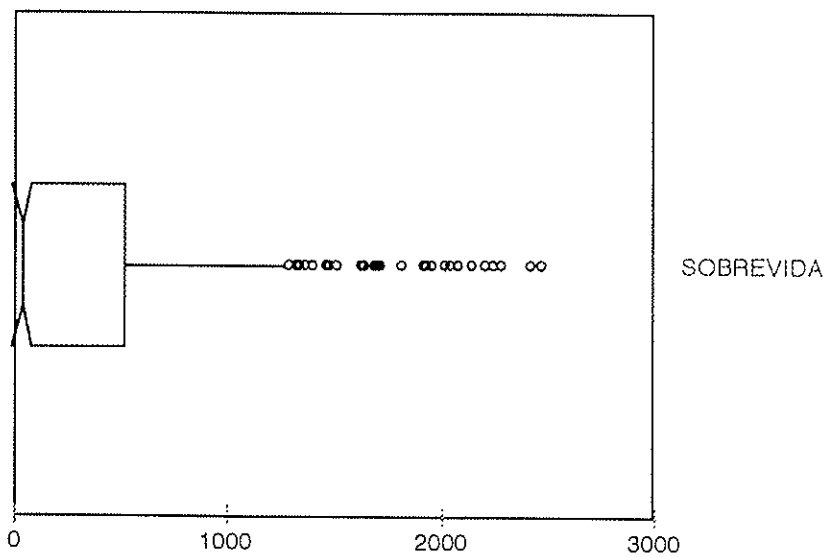


Fig.6.2 - Caixa-de-bigodes dos valores da Sobrevida

Como podemos observar a dispersão dos 50% de dados mais baixos da variável Sobrevida é muito reduzida relativamente à dispersão dos 50% mais altos, realçando, pois uma assimetria na direcção das maiores sobrevidas. Assim, temos indicação de que, realmente 50% dos doentes tiveram uma sobrevida extremamente reduzida. Os restantes doentes apresentam sobrevidas com uma dispersão relativamente grande. Chamamos a atenção para as sobrevidas invulgarmente altas, o que nos levou a procurar, dentro da informação disponível neste momento, uma explicação para o

seu aparecimento.

A primeira ideia que nos surgiu foi que estas sobrevidas invulgarmente altas estivessem associadas a doentes que tiveram a primeira crise com uma idade inferior às idades mais frequentes, isto é, observando a Fig.4.3, com idade inferior a 61 anos. Observando a Fig.6.3, temos indicação de que a explicação para os valores invulgarmente altos da Sobrevida não reside apenas na Idade em que o doente teve a primeira crise.

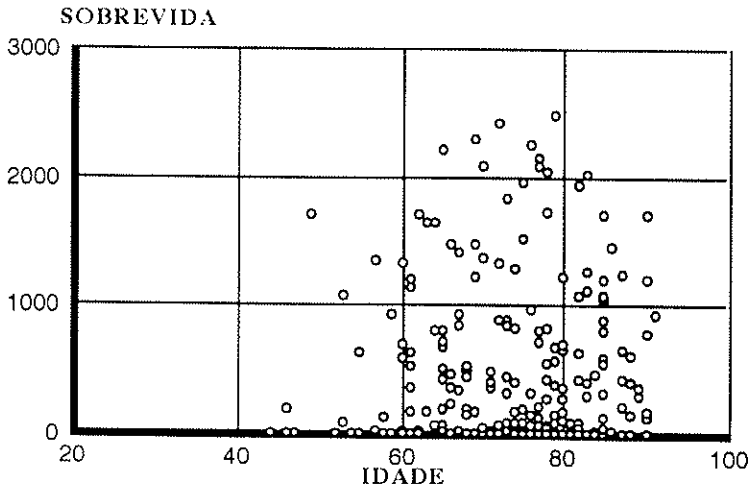


Fig.6.3 - Representação gráfica da Sobrevida relativamente à Idade da 1ª crise diagnosticada

Este gráfico também fornece indicação de que em todas as idades existem sobrevidas nulas.

6.2 — *Estudo da Sobrevida por concelhos*

Uma outra tentativa de explicar os valores invulgarmente altos da Sobrevida consistiu em observar a dispersão desta colecção de dados por concelho. A Fig.6.4 representa as caixas-de-bigodes da sobrevida dos doentes para cada concelho.

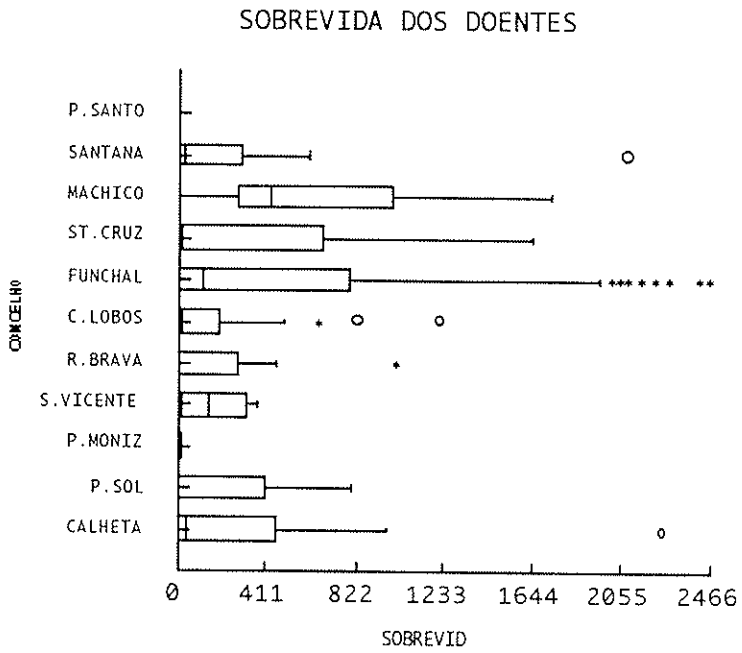


Fig.6.4

Este gráfico fornece indicação de que todos os concelhos são francamente assimétricos na direcção dos doentes com maiores sobrevidas. Em cinco concelhos existem dados com valores extremamente altos. As dispersões apresentam-se heterogéneas.

Não existe uma tendência nítida para o aumento da dispersão *versus* nível. No entanto, fizémos o gráfico de dispersão-*versus*--nível que apresentamos na Fig.6.5.

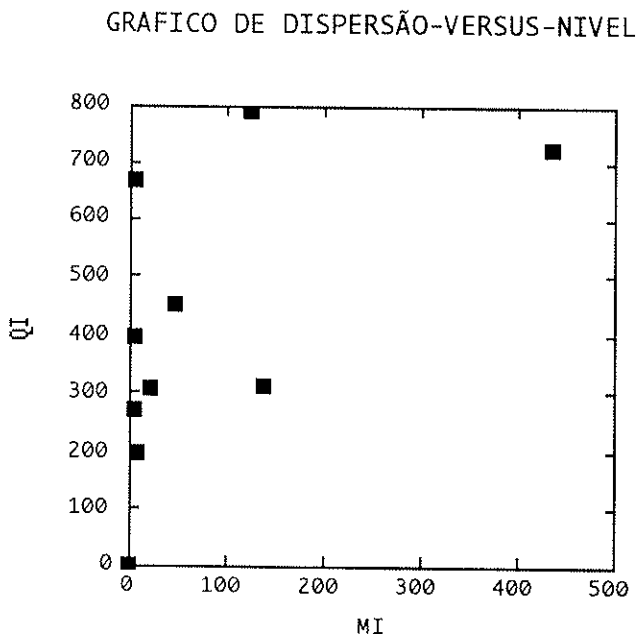


Fig.6.5

O gráfico de dispersão-*versus*-nível sugere como transformação, a raiz-quadrada da Sobrevida. A Fig 6.6 representa as caixas-de-bigodes para os dados transformados.

Esta transformação permitiu uma maior aproximação das dispersões dos onze concelhos e aproximou relativamente à zona central os outliers. As novas caixas-de-bigodes são mais fáceis de comparar.

Como podemos observar os concelhos com maiores sobrevidas medianas são Funchal e Machico. Em relação a este último concelho, esta indicação e a indicação fornecida pelo mesmo gráfico no que se refere às sobrevidas neste concelho serem quase todas muito elevadas, é coerente com o facto de ser um concelho com o mais baixo número de mortes relativamente ao número de doentes.

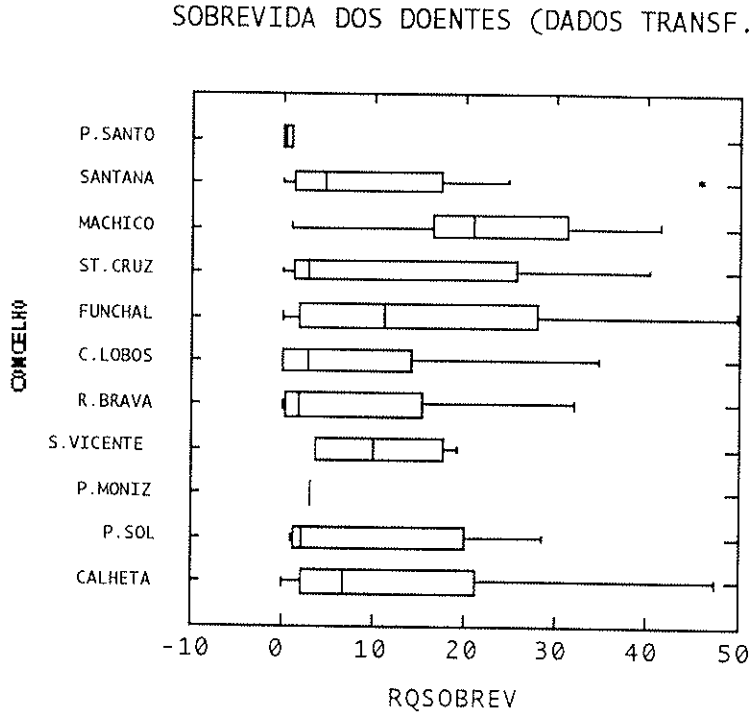


Fig.6.6

O concelho do Funchal apresenta uma cauda direita grande indicando que as 50% maiores sobrevidas se encontram muito

dispersas, podendo mesmo atingir valores muito altos.

Quanto a este concelho e uma vez que já vimos que a Idade em que o doente teve a primeira crise não explica, por si só, a sobrevida que o mesmo apresenta, esperamos que o evoluir da análise exploratória nos elucide esta situação, uma vez que é o segundo concelho com maior número de mortes relativamente ao número de doentes, nunca afastando da análise o facto de a proximidade do hospital explicar uma sobrevida maior do que a que seria de esperar para os doentes deste concelho.

O concelho da Ponta do Sol tem a sobrevida mediana mais baixa, o que é coerente com o facto de ser o concelho com maior número de mortes relativamente ao número de doentes do concelho. No entanto, o comprimento da cauda direita é muito superior ao da cauda esquerda, isto é, as 50% menores sobrevidas são extremamente baixas e todas extremamente próximas, enquanto que as 50% maiores apresentam uma dispersão grande, nunca chegando a atingir valores tão altos como em muitos dos outros concelhos.

O concelho de São Vicente apresenta uma sobrevida mediana alta e uma dispersão realtivamente pequena, o que é coerente com o baixo número de mortes relativamente ao número de doentes do concelho. No entanto, a Sobrevida nunca chega a atingir valores muito altos neste concelho.

O concelho de Santa Cruz apresenta uma sobrevida mediana muito baixa e as 50% menores sobrevidas estão muito concentradas em valores muito baixos. No entanto, a cauda direita é bastante longa, indicando que as 50% maiores dispersões atingem valores altos, o que parece estar de acordo com o facto de ser um concelho com um baixo número de mortes relativamente ao número de doentes do concelho.

7 — Estudo da variável Diagnóstico

7.1 — *Análise Global*

A variável Diagnóstico tem dois níveis : Angina Instável (Ang. Inst.), e Enfarto Agudo do Miocárdio (EAM).

A Fig.7.1 representa a proporção de doentes para cada um dos diagnósticos. Como podemos observar, existe um número muito maior de casos de enfarto do que de angina. Relativamente ao número de doentes de enfarto, a proporção de mortes é, também, mais elevada do que a proporção de mortes de angina instável relativamente ao número de diagnósticos de angina.

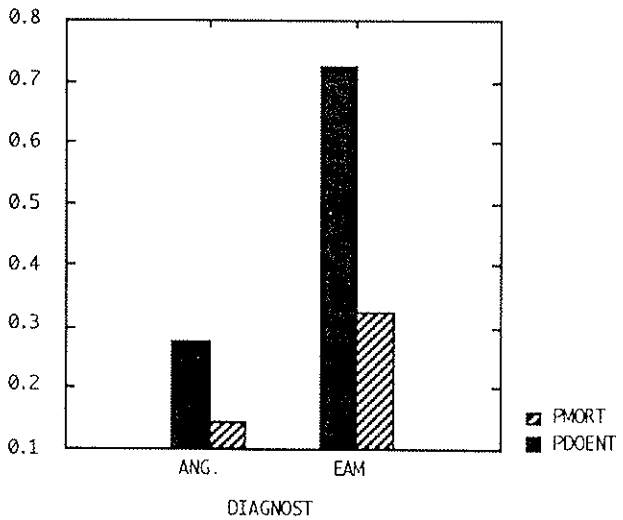


Fig.7.1 - Proporções de doentes e de mortes para cada diagnóstico.

Se observarmos a Fig.7.2 verificamos que, relativamente ao número total de mortes constantes do ficheiro de cardiologia, a proporção de doentes a quem tinha sido diagnosticado enfarto é muito superior à proporção de doentes a quem tinha sido diagnosticado angina instável.

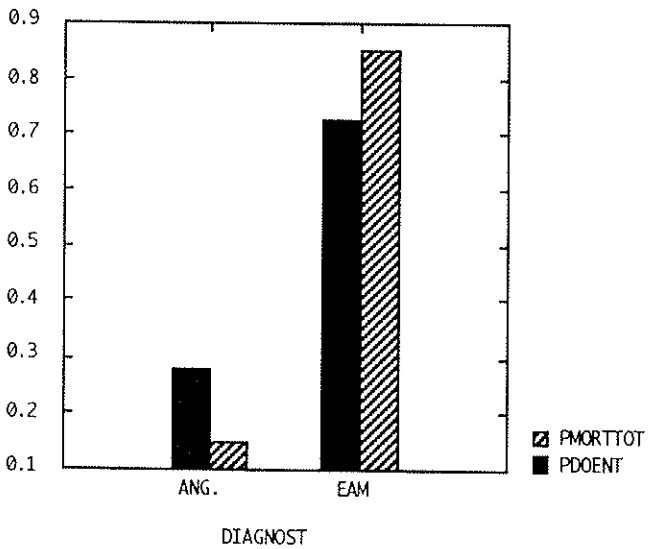


Fig.7.2 - Proporções de doentes e de mortes relativamente ao número total de doentes

Estas observações levaram-nos a estudar a Sobrevida dos doentes para cada uma destas doenças separadamente, pois é de esperar que a sobrevida dos doentes de angina seja maior do que a dos doentes de enfarto. A Fig.7.3 representa as caixas-de-bigodes para estas duas doenças.

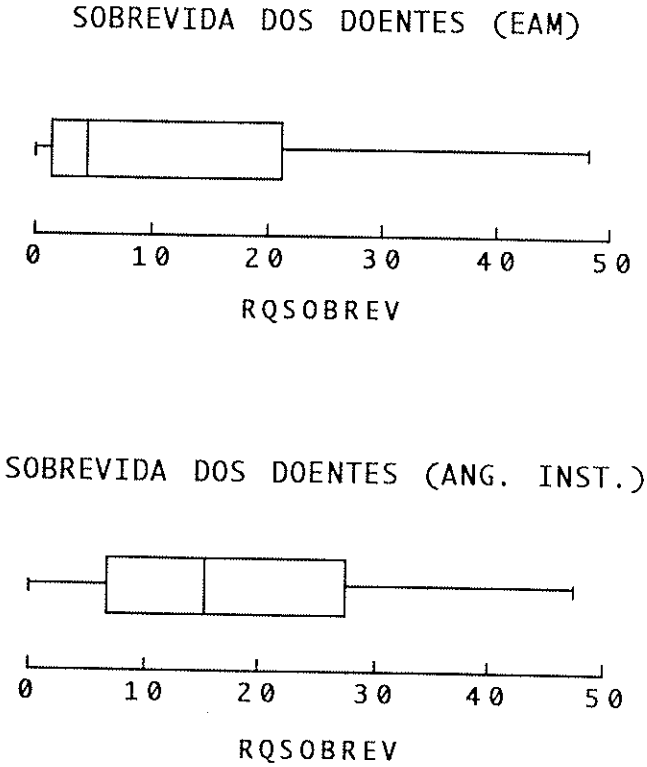


Fig.7.3

A sobrevida mediana para os doentes de angina é de 230 dias, enquanto que para os doentes de enfarto esta é de 18.5 dias. 25% dos doentes de angina tiveram uma sobrevida até 27 dias, enquanto que em relação aos doentes de enfarto as 25% menores sobrevidas não ultrapassaram 2 dias e as 50% menores sobrevidas apresentam uma dispersão muito reduzida. A dispersão da Sobrevida nas 50% maiores sobrevidas é consideravelmente maior para os doentes de enfarto do que para os doentes de angina. Note-se, no entanto, que

existem muitos valores invulgarmente elevados na Sobrevida dos doentes a quem tinha sido diagnosticado enfarto, pelo que transformámos os dados por forma a facilitar a comparação das sobrevidas relativas aos dois diagnósticos.

Os resultados obtidos em 5.2, sugeriram-nos que estes valores invulgarmente altos da Sobrevida para os doentes de enfarto sejam os mesmos que apareciam no concelho do Funchal e que, neste momento, pensamos que possam ser atribuídos ao facto dos doentes correspondentes habitarem muito próximo do Hospital e, consequentemente mais facilmente poderem recorrer ao atendimento do Serviço de Cardiologia.

Mas, o que acabámos de referir sugeriu-nos uma análise da Sobrevida dos doentes de angina instável e de enfarto, separadamente, por concelhos, o que será tratado na secção seguinte.

7.2 — *Estudo por concelhos*

A Fig.7.4 representa os perfis das proporções de doentes de angina instável e de enfarto em cada concelho. Ao observá-la questionamo-nos se existirá um padrão para a relação entre o número de doentes de angina instável e o número de doentes de enfarto, válido em todos os concelhos, ou se a indicação fornecida pelo gráfico de que os concelhos de São Vicente e de Machico têm proporções de doentes de angina instável superiores aos restantes concelhos é suficientemente forte para nos atrevermos a dizer que o referido padrão não é válido em todos os concelhos.

Estudámos, então, estas proporções de uma forma mais detalhada.

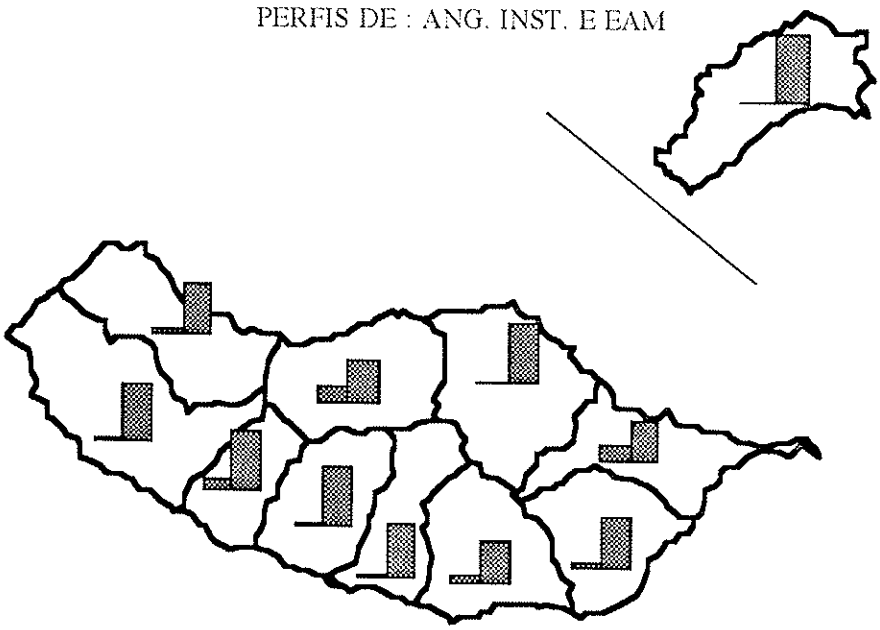


Fig.7.4

Optámos pela transformação logarítmica dos dados, uma vez que os valores da variável resposta são contagens, obtendo a tabela de dados que representámos por Tabela 7.1.

Aos dados da Tabela 7.1 ajustámos, então, pela técnica do refinamento pela mediana, o modelo aditivo

$$y_{ij} = m + a_i + b_j + e_{ij}$$

em que m é um valor típico global para toda a tabela, a_i é a contribuição incremental do concelho i, b_j é a contribuição incremental do diagnóstico j e e_{ij} representa o resíduo da célula correspondente da tabela de dados.

TABELA 7.1

CONCELHO	Log(ANG. INST.)	Log(EAM)
CALHETA	2.079	3.401
P.SOL	2.398	3.497
P.MONIZ	0	1.099
S.VICENTE	1.946	2.485
R.BRAVA	2.079	3.548
C.LOBOS	2.398	3.829
FUNCHAL	5.375	6.188
ST.CRUZ	2.944	4.043
MACHICO	2.833	3.401
SANTANA	1.609	3.219
P.SANTO	0	1.946

TABELA 7.2

CONCELHO	ANG. INST.	EAM	a_i
CALHETA	-0.111	0.111	-0.0915
P.SOL	0	0	0.1165
P.MONIZ	0.0005	-0.0005	-2.282
S.VICENTE	0.285	-0.285	-0.616
R.BRAVA	-0.2025	0.2025	0
C.LOBOS	-0.1655	0.1655	0.282
FUNCHAL	0.1435	-0.1435	2.95
ST.CRUZ	0.0005	-0.0005	0.662
MACHICO	0.266	-0.266	0.2855
SANTANA	-0.255	0.255	-0.4175
P.SANTO	-0.423	0.423	-1.8585
b_j	-0.55	0.55	2.8315

Na Tabela 7.2 apresentamos os resíduos resultantes do ajustamento do modelo aditivo, bem como as contribuições de cada concelho e de cada diagnóstico, assim como a valor típico global.

Na Fig.7.5 apresentamos o gráfico de caule-e-folhas que obtivémos para os resíduos.

GRAFICO DE CAULE-E-FOLHAS PARA OS RESIDUOS

```
-4   2
-3
-2   8650
-1 H 641
-0 M 00
 0   0000
 1 H 146
 2   0568
 3
 4   2
```

Fig.7.5

Os resíduos não apresentam nenhuma tendência evidente, que justifique um gráfico de diagnóstico com o objectivo de acrescentar ao modelo inicial mais um termo correspondente à contribuição conjunta das variáveis Concelho e Diagnóstico. O modelo aditivo ajustar-se aos dados na escala logarítmica significa que as contribuições da variável Diagnóstico são aproximadamente as mesmas nos diferentes concelhos. Existe, pois, indicação de um aumento do número de doentes de angina para o enfarto semelhante para cada concelho. Isto significa que temos indicação da existência de uma relação padrão entre o número de doentes

com angina instável e o número de doentes com enfarto, válido na generalidade dos concelhos.

Por analogia com os modelos log-lineares na descrição de frequências esperadas, podemos dizer que esta análise fornece indicação da não-existência de uma interacção entre as variáveis Concelho e Diagnóstico.

A análise das contribuições dos diversos concelhos não tem qualquer interesse ser reanalizada aqui, uma vez que já foi convenientemente abordada na secção 4. O mesmo podemos dizer em relação à contribuição dos diagnósticos, uma vez que a Fig.7.1 já tinha fornecido uma forte indicação de que o número de doentes com angina instável é consideravelmente inferior ao número de doentes com enfarto.

No entanto, estes resultados confirmam as indicações da Fig.7.4. Se analisarmos os resíduos da Tabela 7.2, poderemos ver que os concelhos que mais se afastam do comportamento padrão no que diz respeito aos dois diagnósticos são os concelhos de Santana e de Porto Santo, com um reduzido número de doentes com angina instável, e os concelhos de São Vicente e de Machico com um número de doentes de angina elevado. Na Fig.7.4 já tínhamos chamado a atenção para o facto de os perfis dos concelhos do Porto Santo e de Santana serem semelhantes, o mesmo se podendo dizer relativamente aos perfis de São Vicente e de Machico.

Note-se que os concelhos de Santana e do Porto Santo, apesar de terem perfis idênticos, têm números de mortes relativamente ao número de doentes do concelho muito diferentes, tendo o primeiro um número mais elevado.

Os concelhos da Ponta do Sol, do Porto Moniz e de Santa Cruz são os que têm menores resíduos e, como tal, são os que mais

se aproximam do referido comportamento padrão. Uma interpretação possível do valor nulo do resíduo correspondente ao concelho da Ponta do Sol, é que a referida relação padrão será semelhante à apresentada neste concelho.

O facto de o Concelho de Machico ter um número excessivo de anginas relativamente aos doentes do concelho, sugere que, e uma vez que já vimos que a Sobrevida dos doentes de angina é superior à dos doentes de enfarto, a variável Diagnóstico pode contribuir para a explicação da baixa proporção de mortes observada neste concelho. O mesmo se pode dizer relativamente ao concelho de São Vicente.

O concelho do Funchal apresenta tendência para um número elevado de anginas, o que pode contribuir para explicar o facto de a Sobrevida apresentar valores elevados neste concelho.

Este facto, aliado ao facto do concelho da Ponta do Sol não ter um número excessivo de doentes de enfarto relativamente ao número de doentes do concelho, e, apesar disso, apresentar um elevado número de mortes em relação ao número de doentes do concelho, leva-nos a analisar, separadamente a Sobrevida dos doentes para ambos os diagnósticos, mas agora, por concelhos.

A Fig.7.6 representa a caixa-de-bigodes para a Sobrevida dos doentes de enfarto, nos diversos concelhos. A Fig.7.7 representa a caixa-de-bigodes para a Sobrevida dos doentes de angina instável, por concelhos.

Estes gráficos fornecem indicação de que, os doentes do concelho de Machico apresentam uma Sobrevida elevada, quer relativamente à angina instável, quer em relação ao enfarto.

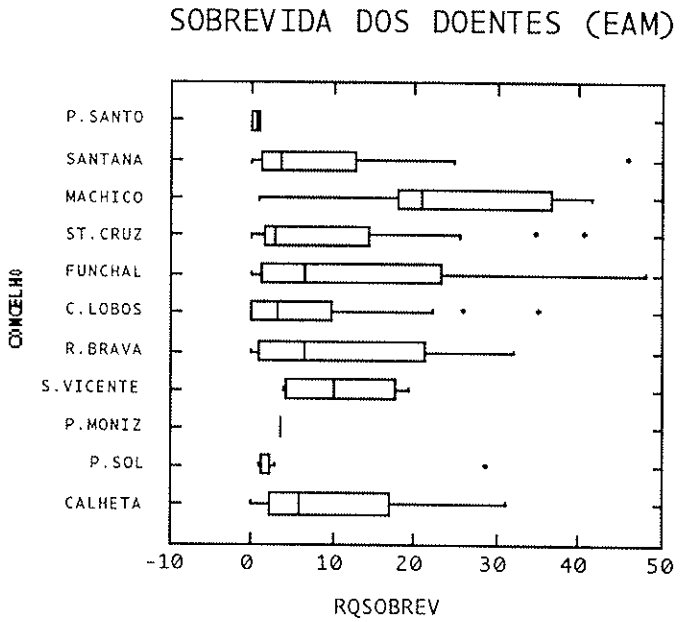


Fig. 7.6

Os doentes de angina instável do Funchal apresentam a Sobrevida elevada relativamente aos doentes dos restantes concelhos.

Quanto ao concelho da Ponta do Sol, a Sobrevida dos doentes de enfarto é extremamente reduzida quando comparada com a dos doentes dos restantes concelhos e, ficamos esclarecidos dos valores relativamente elevados que as 50% maiores sobrevidas podiam apresentar neste concelho quando não fizémos a distinção entre os dois diagnósticos. A Sobrevida mediana dos doentes de angina deste concelho é elevada. No entanto, as 50% maiores sobrevidas são muito próximas, não atingindo valores tão elevados como no Funchal ou em Santa Cruz.

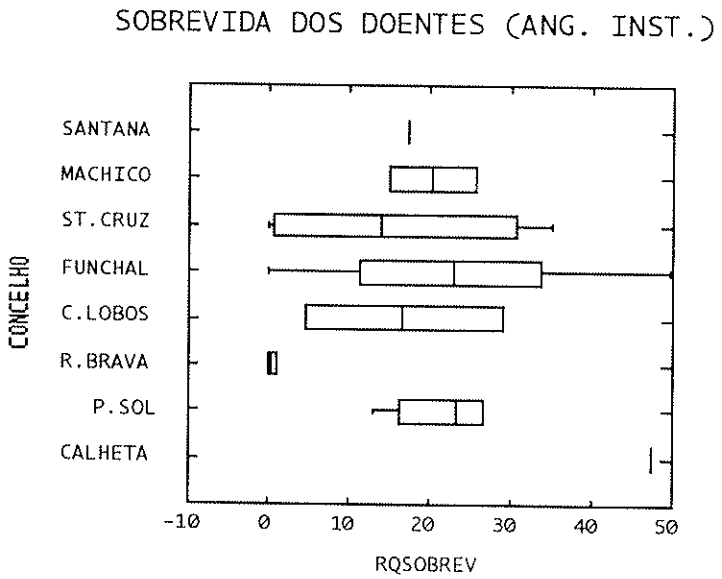


Fig.7.7

A Fig. 7.8 é uma representação gráfica extremamente útil para a restante fase da análise exploratória do ficheiro de cardiologia. Repare-se que está representada gráficamente a informação contida nos resíduos da Tabela 7.2.

Consultando esta figura rapidamente temos informação se, num dado concelho, o número de doentes de angina instável (enfarto) é elevado relativamente ao mesmo número nos restantes concelhos ou se aquele concelho tem uma baixa incidência de doentes de angina.

PERFIS DE : ANG. INST. E EAM (STAND)

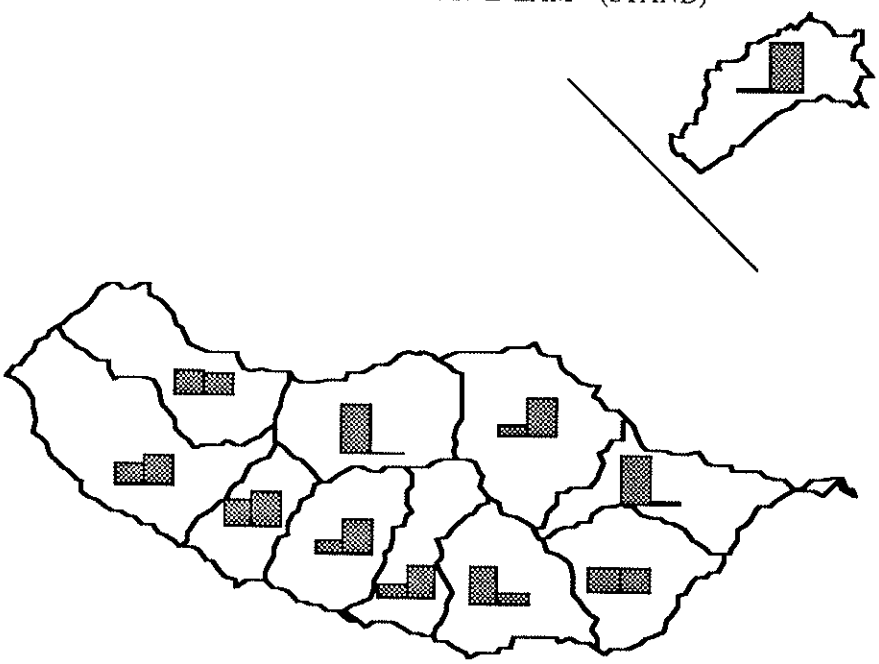


Fig.7.8

Note-se que, neste gráfico, as duas colunas de cada concelho quando comparadas entre si, apenas fornecem indicação da prevalência da incidência de um dos diagnósticos quando comparada com a incidência do mesmo diagnóstico nos restantes concelhos.

Por exemplo, relativamente ao concelho do Porto Moniz, podemos dizer que há indicação de que em relação o número de doentes de angina instável, este concelho não é dos que apresenta menores incidências, embora esta incidência seja relativamente

maior do que a dos doentes de enfarto do mesmo concelho relativamente às dos restantes concelhos. Digamos que mentalmente atribuímos uma ordem a cada concelho para cada um dos diagnósticos. Assim, voltando ao concelho do Porto Moniz, verificamos que a ordem que lhe foi atribuída relativamente à incidência de doentes de angina é superior à ordem que ao mesmo foi atrubuída relativamente à incidência do número de doentes de enfarto.

Não é nossa intenção apresentar exaustivamente esta análise exploratória para todas as variáveis do ficheiro de dados. O nosso objectivo era mostrar como pensamos ser a melhor forma de desenvolver tal análise, como esta abordagem enriquece o conhecimento do comportamento da doença coronária numa região e como levanta questões pertinentes que de outra forma nos teriam passado desapercebidas. Incluímos no parágrafo que se segue mais um exemplo que confirma o que acabámos de referir.

8 — Detecção de variáveis "escondidas"

ILHAS DA MADEIRA E DO PORTO SANTO

Neste parágrafo mostraremos, apenas com um exemplo, a importância da atribuição de uma dimensão espacial às doenças sob estudo na interpretação de uma base de dados clínicos.

Permite confirmar a relevância de desenvolver uma análise exploratória com a ajuda de informações geográficas.

Na Fig.8.1 podemos observar os perfis, para cada concelho,

das proporções de doentes relativamente ao número de habitantes, de mortes relativamente ao número de doentes, e de doentes cujo diagnóstico foi angina e de doentes cujo diagnóstico foi enfarto agudo do miocárdio ambas relativamente ao número de doentes. Os dados foram standardizados de modo a que cada coluna possa ser comparada com a coluna correspondente em qualquer outro concelho.

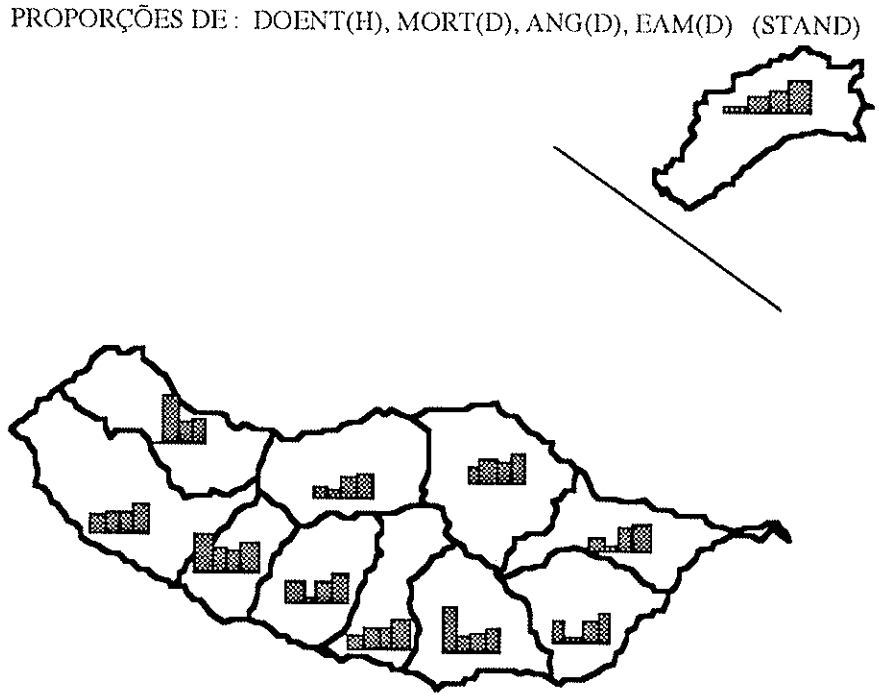


Fig.8.1

Como exemplo, concentremos a nossa atenção no concelho de Câmara de Lobos. Em Câmara de Lobos a incidência destas

doenças é baixa, mas a proporção de mortes é relativamente elevada. À maioria dos doentes foi diagnosticado enfarto agudo do miocárdio.

Podemos observar na Fig.8.2 que representa a caixa-de-bigodes da Sobrevida dos doentes a quem foi diagnosticado enfarto que, neste concelho a Sobrevida destes doentes é extremamente curta quando comparada com a Sobrevida dos doentes de outros concelhos.

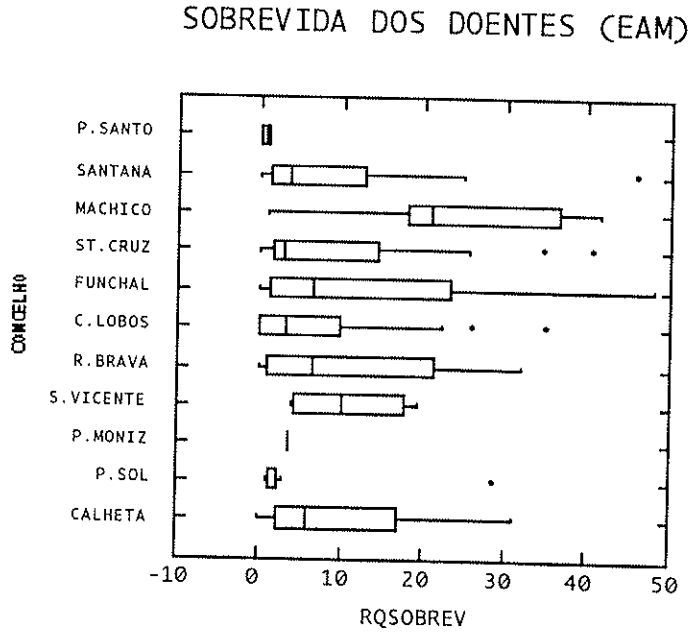


Fig. 8.2

Observando a Fig.8.3 onde estão representadas, para cada concelho, as proporções de doentes nas classes etárias 25 a 54 anos, 55 a 64 anos, 65 a 79 anos e 80 a 91 anos, podemos observar

que Câmara de Lobos é um concelho com uma elevada proporção de doentes com menos de 65 anos de idade. Isto não é surpreendente uma vez que este concelho apresenta a mais baixa proporção de habitantes no grupo etário de 65 e mais anos.

PROPORÇÕES DE DOENTES POR IDADES : 25-54, 55-64, 65-79, 80-91

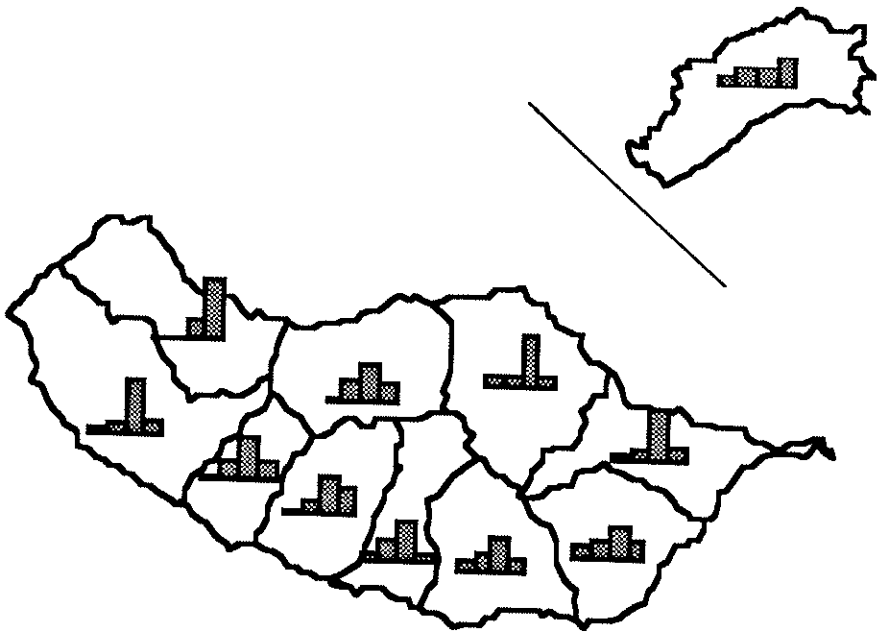


Fig. 8.3

Na tentativa de explicar a baixa Sobrevida observada nos doentes de Câmara de Lobos, fomos sob a orientação dos especialistas do Serviço de Cardiologia, analisar nos diferentes concelhos a variável Tabaco, que regista os hábitos tabágicos dos doentes. A Fig.8.4 torna visível que existe uma considerável

proporção de doentes que fumam mais de 20 cigarros por dia. No entanto, o mesmo facto pode ser observado no concelho de Machico onde os doentes de enfarto agudo do miocárdio apresentam uma Sobrevida muito elevada quando comparada com a Sobrevida dos doentes com o mesmo diagnóstico dos outros concelhos. Consequentemente os hábitos tabágicos dos doentes, pelo menos isoladamente, não explicam a baixa Sobrevida dos doentes de Câmara de Lobos.

PERFIS DE : NÃO-FUMADORES, FUM<20, FUM>20

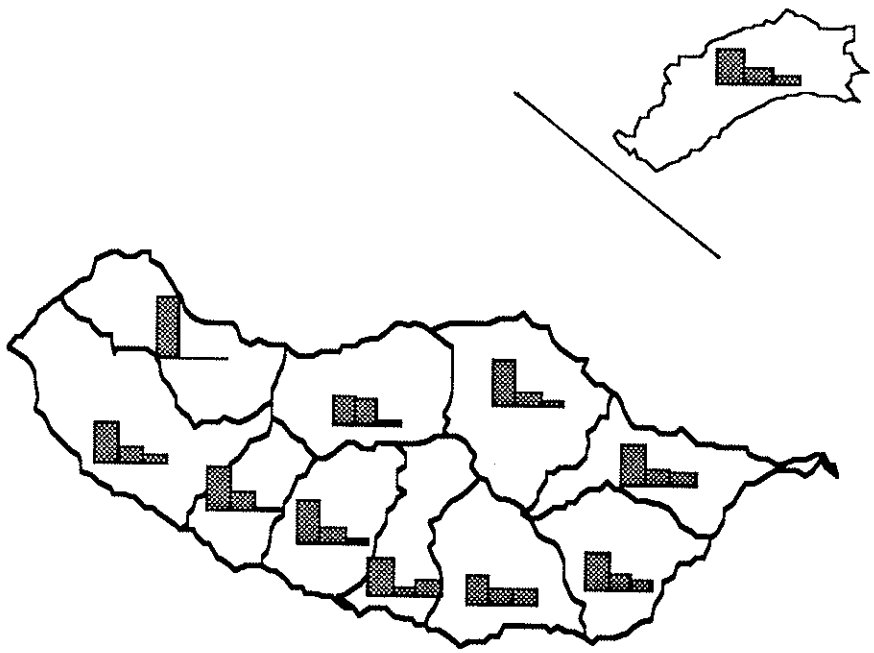


Fig. 8.4

É de realçar que o tipo de alimentação dos habitantes dos

dois concelhos é idêntico, uma vez que são concelhos onde grande parte da população se dedica à actividade piscatória. Como tal, não existe indicação de que no tipo de alimentação resida a explicação que procuramos.

Na Fig.8.5 apresentamos as proporções de doentes com e sem hipertensão prévia. O concelho de Câmara de Lobos surge com uma baixa proporção de doentes com hipertensão prévia.

PERFIS DE : PROPORÇÕES DE DOENTES NÃO-HIPERTENSOS E HIPERTENSOS

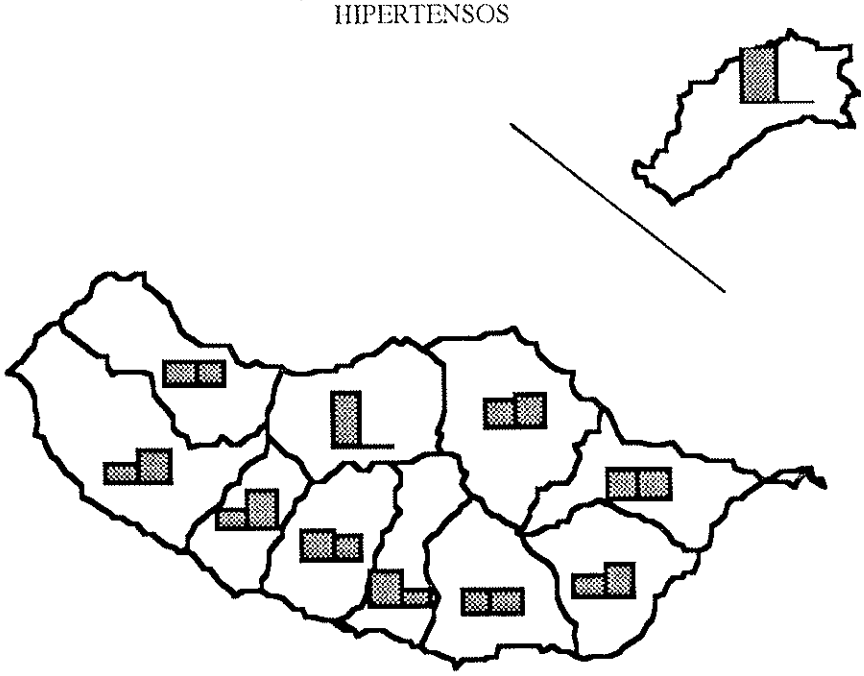


Fig. 8.5

Se considerarmos que o concelho de Câmara de Lobos é uma

zona caracterizada por um problema de alcoolismo extremamente sério desde há muito tempo (embora não existam estatísticas fiáveis sobre a assunto por ser uma região vinícola onde existem muitos pequenos produtores), podemos afirmar que existe uma variável que não consta do ficheiro de dados, que pode fornecer uma contribuição importante para a explicação da curta Sobrevida observada nos doentes de enfarto do concelho de Câmara de Lobos.

Fica, assim, evidenciada uma relação entre o consumo não moderado de álcool e a mortalidade cardíaca.

Pelo que foi dito na introdução deste capítulo, uma predisposição genética pode, também, ser considerada. A relevância da predisposição genética como factor de risco para a doença coronária tem sido abordada em estudos recentes. Contudo, o acesso à predisposição genética é impossível neste estudo.

Com este simples exemplo ficou evidenciado que a atribuição de uma dimensão espacial à doença coronária, torna possível, através da análise exploratória de dados, considerar variáveis "escondidas" e as respectivas relações com outras variáveis e contribui consideravelmente para o estudo das duas doenças a que os dados dizem respeito.

Referências:

Andersen, E. (1991) *The Statistical Analysis of Categorical Data.* Springer Verlag.

Bishop, Y., Fienberg, S. e Holland, P. (1975) *Discrete Multivariate Analysis.* The Massachussets Institute of Technology.

Cogels, O. (1992) *Geomanagement system - a new concept for communication, integration, and analysis of georeferenced information.* New Techniques and Technologies for Statistics Conference, Bonn.

Hoaglin, D., Mosteller, F. e Tukey, J. (1992) *Análise Exploratória de Dados. Técnicas Robustas*. Edições Salamandra, Lisboa

Mosteller, F. e Tukey, J. (1977) *Data Analysis and Regression.* Addison--Wesley Publishing comp.

Oberman, A. (1992) *Epidemiology and Prevention of Cardiovascular Disease.* Text Book of Internal Medicine, Vol.1, Chap.34, J.B. Lippincoh Company

Scholter, H. e Lepper, M. (1992) *The benefits of the application of geographical information systems in public and environmental health.* New Techniques and Technologies for Statistics Conference, Bonn.

Tukey, J. (1979) *Methodology, and the Statistician's Responsability for Both Accuracy and Relevance*. JASA **74**, pp. 262-269.

Vasconcelos, Rita (1992) *Doenças Cardio-Vasculares na Região Autónoma da Madeira. I. Preparação de cartas geográficas para utilização estatística.* Notas e Comunicações do Centro de Estatística e Aplicações da Universidade de Lisboa - INIC, **Nota nº7**.

Vasconcelos, Rita (1993) *The relevance of spatial statistics on the statistical model building for coronary heart disease.* 2nd European Conference on Mathematics Applied to Biology and Medicine, Lyon.

Vasconcelos, Rita (1994) *Análise Exploratória de uma base de dados sobre Cardiologia. Atribuição de uma dimensão espacial à doença coronária?*, A Estatística e o Futuro e o Futuro da Estatística, D. Pestana et al, eds. Salamandra, Lisboa, pp. 197-228.

Vasconcelos, Rita (1994) *O Mundo em que vivemos*, A Estatística e o Futuro e o Futuro da Estatística, D. Pestana et al, eds. Salamandra, Lisboa, pp. 287-294.

Capítulo IV

Aplicação do teste de ajustamento do Qui-Quadrado à análise da incidência da doença coronária em diferentes regiões.

1 — Introdução

Neste capítulo analisaremos a escolha da estatística de teste de Pearson para testar a hipótese da igualdade do risco de contracção das doenças em estudo nos diversos concelhos.

Mostraremos que, sob certas condições, o estimador do número médio de doentes num concelho obtido com base na proporção de habitantes do concelho é o que apresenta menor variabilidade.

A sensibilidade do teste do Qui-Quadrado de Pearson à detecção de discrepâncias entre frequências observadas e esperadas que clinicamente são consideradas importantes será, também, abordada.

Esta última questão levou-nos a considerar uma transformação dos dados originais, para o que apresentamos a expansão em série de Taylor da função característica de uma potência não necessariamente inteira de uma variável aleatória Binomial. Consideramos este resultado importante para o estudo da velocidade de convergência da distribuição da estatístistica de teste de Pearson para a distribuição Qui-Quadrado quando aplicamos uma transformação potência aos dados originais.

2— Sobre o teste da hipótese de igualdade do risco relativo de contracção das doenças nos diferentes concelhos

No Capítulo III utilizámos a estatística de teste de Pearson para testar a hipótese de igualdade do risco relativo de contracção da doença coronária nos diferentes concelhos da Região. Apresentamos neste parágrafo uma estudo mais detalhado sobre a aplicabilidade deste teste, uma vez que o consideramos importante em análises como a que desenvolvemos neste trabalho.

Voltamos a formular a questão posta anteriormente :

—Gostaríamos de poder analisar se o risco relativo de contrair as doenças é o mesmo para todos os concelhos no grupo etário considerado, o que é equivalente a verificarmos se podemos considerar que a variabilidade amostral justifica as diferenças entre as proporções de doentes em cada concelho relativamente ao número total de doentes na região e as correspondentes proporções de habitantes.

Para respondermos a esta questão considerámos o seguinte modelo para os dados da Tabela 2.1, supondo que o número de doentes num concelho entre os anos de 1983 e 1991 pode ser adequadamente modelado por uma variável aleatória de Poisson:

$$\begin{cases} X_i \cap P(\lambda_i) \\ \lambda_i / \lambda_. = \psi_i N_i / N_. \end{cases} \qquad i = 1,\ldots,k$$

em que X_i representa o número de doentes entre 1983 e 1991, as variáveis X_i são independentes e N_i representa o número de habitantes, com 25 e mais anos, no concelho i, com i = 1,...,k. ψ_i

representa o risco relativo de contracção das doenças no i-ésimo concelho.

TABELA 2.1

CONCELHO	Nº DOENTES	Nº HABITANTES
CALHETA	38	7933
P.SOL	44	5117
P.MONIZ	4	2092
S.VICENTE	19	4562
R.BRAVA	44	7543
C.LOBOS	57	14627
FUNCHAL	703	68566
ST.CRUZ	76	13588
MACHICO	47	11682
SANTANA	30	6253
P.SANTO	8	2662
TOTAL	1070	144625

A hipótese de que o risco relativo de contracção das doenças é o mesmo em todos os concelhos, pode ser formulada da seguinte maneira :

$$H_0 : \psi_i = \psi \ , \ \forall i \ .$$

em que ψ, que representa o risco comum de contracção das doenças, é a média ponderada dos riscos nos diversos concelhos, isto é,

$$\psi = \sum_{i=1}^{k} \psi_i N_i / N. = 1$$

A hipótese H_0 é equivalente a

$$\lambda_i / \lambda_. = N_i / N. \ , \ \forall i$$

Sendo $X_1,...,X_k$ k variáveis aleatórias de Poisson independentes, com parâmetros $\lambda_1,...,\lambda_k$ respectivamente, a distribuição condicional de $(X_1,...,X_k)$ dado $\sum_{i=1}^{k} X_i = n$, é multinomial de parâmetros n e **p**, em que $\mathbf{p} = \left(\frac{\lambda_1}{\lambda_.},...,\frac{\lambda_k}{\lambda_.}\right)$.

Para testar H_0 utilizámos, então, a estatística de teste de Pearson

$$X^2 = \sum_{i=1}^{k} \left(X_i - nN_i / N.\right)^2 / \left(nN_i / N.\right)$$

que, sob H_0, tem aproximadamente uma distribuição Qui-Quadrado com k-1 graus de liberdade.

— *Passemos a alguns comentários sobre a escolha da estatística de Pearson para testar H_0:*

As estatísticas mais frequentemente usadas em testes de ajustamento de frequências observadas, (de Pearson, logaritmo da razão das verosimilhanças, logaritmo da razão das verosimilhanças transformado, etc.), são casos especiais da família de estatísticas $\{I^{\alpha}; \alpha \in \mathbb{R}\}$, com

$$2nI^{\alpha} = \frac{2}{\alpha(\alpha+1)} \sum_{i=1}^{k} \left\{ X_i \left(\frac{X_i}{E_i} \right)^{\alpha} - 1 \right\} ; \alpha \in \mathbb{R}$$

onde $\{X_i ; i=1,...,k\}$ representam as frequências observadas e $\{E_i ; i=1,...,k\}$ as frequências esperadas e $2nI^{\alpha}$ é definida por continuidade para $\alpha=0$ e $\alpha=-1$:

$$2nI^{0} = \lim_{\alpha \to 0} 2nI^{\alpha} = 2 \sum_{i=1}^{k} X_i \ln \left(\frac{X_i}{E_i} \right)$$

e

$$2nI^{-1} = \lim_{\alpha \to -1} 2nI^{\alpha} = 2 \sum_{i=1}^{k} E_i \ln \left(\frac{E_i}{X_i} \right).$$

A escolha do α e, consequentemente, da estatística de teste a usar, baseou-se na sensibilidade da estatística a grandes quocientes de $\frac{x_i}{E_i}$ ou $\frac{E_i}{x_i}$. Da expressão de $2nI^{\alpha}$, podemos observar que um valor elevado de $\frac{x_i}{E_i}$ originará um aumento na estatística quando α aumenta, $\alpha>0$. Do mesmo modo, um valor elevado de $\frac{E_i}{x_i}$ resultará num aumento da estatística quando $|\alpha|$ aumenta, $\alpha<0$. Assim, se quisermos precaver-nos contra o efeito na estatística de uma situação semelhante às descritas, devemos escolher $|\alpha|$ pequeno. Para um estudo aprofundado dos efeitos da magnitude de $\frac{x_i}{E_i}$ e de $\frac{E_i}{x_i}$ sobre a estatística de teste consulte-se Cressie e Read (1984). Na Tabela 2.2 apresentamos as frequências observadas, as

frequências esperadas e os valores da função

$$g(x_i, E_i) = \begin{cases} x_i / E_i \ , & x_i \geq E_i \\ -E_i / x_i , & x_i < E_i \end{cases}.$$

TABELA 2.2

CONCELHO	Nº DOENTES (xi)	NºHABITANTES (Ni)	FREQ. ESP. (Ei)	g(xi,Ei)
CALHETA	38	7933	58.692	-1.545
P.SOL	44	5117	37.858	+1.162
P.MONIZ	4	2092	15.478	-3.87
S.VICENTE	19	4562	33.752	-1.776
R.BRAVA	44	7543	55.806	-1.268
C.LOBOS	57	14627	108.217	-1.899
FUNCHAL	703	68566	507.282	+1.386
ST.CRUZ	76	13588	100.53	-1.323
MACHICO	47	11682	86.429	-2.881
SANT'ANA	30	6253	46.262	-1.542
P.SANTO	8	2662	19.695	-2.462
TOTAL	1070	144625		

Os concelhos do Porto Moniz e de Machico apresentam os maiores quocientes $\frac{E_i}{x_i}$ e vão dominar a estatística para $|\alpha|$ grande, $\alpha<0$. Por outro lado, temos, apenas, dois concelhos onde a função toma valores positivos, sendo estes valores inferiores, em valor absoluto, aos primeiros. Assim, o aumento da estatística com $|\alpha|$ será mais rápido para $\alpha<0$ do que para $\alpha>0$.

Como veremos na secção seguinte, a explicação para termos apenas dois concelhos onde a função $g(x_i,E_i)$ toma valores positivos reside no facto de o concelho do Funchal ter uma

proporção de doentes observados muito superior às correspondentes proporções nos restantes concelhos. Estamos, assim, interessados em detectar um afastamento do modelo neste sentido. Deu-se, pois, preferência a uma estatística que desse ênfase a grandes valores do quociente $\frac{x_i}{E_i}$, optando pela escolha de $\alpha>0$.

A escolha de $\alpha=1$, isto é da estatística de teste de Pearson, pareceu-nos apropriada e os resíduos de Pearson da Tabela 2.3 estão de acordo com o que esperávamos.

TABELA 2.3

CONCELHO	R.PEARSON
CALHETA	-2.7
P.SOL	1.0
P.MONIZ	-2.92
S.VICENTE	-2.54
R.BRAVA	-1.58
C.LOBOS	-4.92
FUNCHAL	8.69
ST.CRUZ	-2.45
MACHICO	-4.24
SANTANA	-2.39
P.SANTO	-2.64

No entanto, para efeitos comparativos, apresentamos também, quando apropriado, resultados para $\alpha=2/3$ que diversos estudos apontam como valor "óptimo", em muitas circunstâncias.

Vejamos o que aconteceria se fossemos aplicar o modelo considerado aos onze concelhos e testar H_0 .

O valor obtido para a estatística de teste , $x^2 = 162.1338$, é demasiado elevado e como podemos observar na Tabela 2.3, os resíduos fornecem indicação de que o modelo não é adequado aos dados referentes aos onze concelhos. Note-se, também, que os resíduos são quase todos negativos, excepto os referentes aos concelhos da Ponta do Sol e do Funchal.

Uma vez que $\min\{E_i\} \geq 1$ e $n \geq 10$, calculámos o valor da estatística de teste para $\alpha = 2/3$. O valor obtido $x^{*2} = 164.62368$ é, ainda, mais elevado do que o anterior.

Decidimos, então, analisar mais detalhadamente esta situação.

— *Valor elevado da estatística de teste:*

O valor tão elevado obtido para a estatística de teste do Qui-Quadrado não significa que o modelo proposto nada tenha a ver com o verdadeiro modelo subjacente àquela tabela de dados. Em alguns conjuntos de dados, em que a dimensão da amostra é muito grande, um modelo pode originar um valor elevado do Qui-Quadrado, embora nós possamos pensar que o modelo é grosseiramente, mas não exactamente, verdadeiro. Isto deve-se ao facto de a estatística do Qui-Quadrado aumentar não só com o afastamento do modelo ajustado relativamente ao verdadeiro modelo, mas também com a dimensão da amostra sempre que o modelo proposto não seja exactamente o verdadeiro modelo.

Designemos por p_i^* a verdadeira probabilidade correspondente à i-ésima célula da tabela de dados. Se $p_i = p_i^*$, $\forall i$, o valor esperado da estatística do Qui-Quadrado de Pearson

$$X^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

é k-1. Quando $p_i \neq p_i^*$ para algum i, tem-se:

$$E(X^2) = E\left[\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}\right] = E\left[\sum_{i=1}^{k} \frac{(X_i - np_i^* + np_i^* - np_i)^2}{np_i}\right]$$

$$= \sum_{i=1}^{k} \frac{E\left[(X_i - np_i^*)^2\right] + 2E\left[(X_i - np_i^*)(np_i^* - np_i)\right] + n^2 E\left[(p_i^* - p_i)^2\right]}{np_i}$$

$$= \sum_{i=1}^{k} \frac{E\left[(X_i - np_i^*)^2\right]}{np_i} + n\sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i}$$

$$= \sum_{i=1}^{k} \frac{p_i^*(1 - p_i^*)}{p_i} + n\sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^{k} \frac{p_i^* - p_i^{*2}}{p_i} + n\sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i}$$

$$= \sum_{i=1}^{k} \frac{p_i^* - p_i^{*2} + 2p_i p_i^* - p_i^2 - 2p_i p_i^* + p_i^2}{p_i} + n\sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i}$$

$$= \sum_{i=1}^{k} \frac{p_i^* - 2p_i p_i^* + p_i^2 - (p_i^* - p_i)^2}{p_i} + n\sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i}$$

$$= (n-1)\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i} + \sum_{i=1}^{k}\frac{p_i^*}{p_i} - 2\sum_{i=1}^{k} p_i^* + \sum_{i=1}^{k} p_i$$

$$= (n-1)\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i} + \sum_{i=1}^{k}\frac{p_i^*}{p_i} - 1$$

$$= (n-1)\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i} + \sum_{i=1}^{k}\frac{p_i^* - p_i + p_i}{p_i} - 1$$

$$= (n-1)\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i} + \sum_{i=1}^{k}\frac{p_i^* - p_i}{p_i} + k - 1 \qquad (2.1)$$

Os dois primeiros termos representam a alteração observada no valor esperado resultante do ajustamento do modelo errado. Como podemos verificar, quando n é grande relativamente a k, o primeiro termo é grande e o valor esperado de X^2 aumenta com o valor de n.

Note-se que, neste caso, X^2 tem assintoticamente uma distribuição Qui-Quadrado não-central com k-1 graus de liberdade e parâmetro de não centralidade

$$\eta = n\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i},$$

indicando, também, que o valor de X^2 aumenta com o de n.

Na Tabela 2.4 representamos os valores obtidos para a estatística de teste do Qui-Quadrado de Pearson para diversos valores de n. Mantivémos as proporções de doentes observadas e

de habitantes, fazendo, apenas, variar o n.

Há assim interesse, para efeitos comparativos, em registar X^2/n (cf. Mosteller e Rourke, 1973, p.191).

TABELA 2.4

n	x^2	x^2/n
200	29.088	0.15
300	44.798	0.15
500	75.801	0.15
1000	151.428	0.15
1070	162.134	0.15

Observa-se que, mantendo fixas as diferenças $(p_i^* - p_i)$ e fazendo as correspondentes alterações nos x_i e nos N_i, o primeiro termo de (2.1) é decisivo para o valor esperado de X^2.

— *Elevado número de resíduos negativos e grandes em valor absoluto:*

Repare-se que o modelo formulado,

$$\frac{\lambda_i}{\lambda} = \frac{N_i}{N}, \; i=1,...,11 \; ,$$

é equivalente a

$$\frac{\lambda_i}{N_i} = \frac{\lambda}{N}, \; i=1,...,11,$$

ou seja, a considerarmos que a proporção média de doentes é

idêntica em todos os concelhos à proporção esperada de doentes na Região.

Basta, pois, que um determinado concelho apresente uma proporção observada de doentes suficientemente mais elevada, ou mais baixa, que os outros para alterar consideravelmente as frequências esperadas para cada concelho.

Podemos ter

$$\frac{\lambda_1}{N_1} = \frac{\lambda_2}{N_2} = \ldots = \frac{\lambda_{10}}{N_{10}}$$

e $\frac{\lambda_{11}}{N_{11}}$ muito superior (inferior), o que dará origem a um valor relativamente mais elevado (baixo) de $\frac{\lambda}{N}$, uma vez que $\lambda = \sum_{i=1}^{11} \lambda_i$ e $N = \sum_{i=1}^{11} N_i$. Como consequência teremos frequências esperadas maiores (menores) para os restantes dez concelhos e os correspondentes resíduos negativos (positivos), podendo tomar valores tanto mais elevados, em valor absoluto, quanto maior for a discrepância entre a proporção de doentes desse concelho e as proporções dos restantes concelhos.

Conclui-se, assim, que um único concelho com uma proporção de doentes muito diferente dos restantes, pode perturbar consideravelmente os resultados do teste e a interpretação dos resíduos, uma vez que o cálculo das frequências esperadas para todos os concelhos é afectado pelo valor observado para este concelho, o qual uma análise dos resíduos é suposto identificar. Este facto corresponde ao fenómeno de "derrame" tão frequente na análise clássica de tabelas de dupla entrada (Hoaglin, Mosteller e Tukey, 1983, p.180).

— *Qual o caminho a seguir?*

Na tabela de resíduos apresentada, temos dois resíduos positivos e nove negativos. Estamos, certamente, numa situação semelhante à referida no parágrafo anterior. O maior resíduo, em valor absoluto, é o referente ao concelho do Funchal. Este resíduo é positivo. Somos, pois, levados a crer que este concelho terá uma proporção de doentes muito superior aos restantes concelhos. O resíduo referente ao concelho da Ponta do Sol não é grande, em valor absoluto, mas é também positivo. Isto poderá significar que também este concelho apresenta uma proporção de doentes mais elevada que os restantes nove concelhos e que a presença de um resíduo tão elevado no concelho do Funchal dissimula a presença de outro resíduo também elevado no concelho da Ponta do Sol.

Os factos que acabámos de referir aliados a termos um elevado número observado de doentes, o que contribui fortemente para explicar o valor tão elevado obtido para a estatística de teste, forneceram-nos indicação de que é plausível testar o ajustamento do modelo inicialmente formulado aos nove concelhos que apresentavam resíduos negativos.

Na Tabela 2.5 apresentamos os resíduos de Pearson obtidos.

TABELA 2.5

CONCELHO	Nº DOENTES	Nº HABITANTES	FREQ. ESPERADA	R. PEARSON
CALHETA	38	7933	36.1191	0.313
P.MONIZ	4	2092	9.5249	-1.7902
S.VICENTE	19	4562	20.7709	-0.3886
R.BRAVA	44	7543	34.3434	1.6478
C.LOBOS	57	14627	66.597	-1.176
ST.CRUZ	76	13588	61.8664	1.7969
MACHICO	47	11682	53.1883	-0.8485
SANTANA	30	6253	28.47	0.2867
P.SANTO	8	2662	12.12	-1.1834
TOTAL	323	70942		

O valor obtido para a estatística de teste de Pearson, $x^2=12.9837$, leva-nos à não rejeição da hipótese de que o risco relativo de contracção das doenças é idêntico naqueles nove concelhos.

Chegamos à mesma conclusão se usarmos a estatística de teste para $\alpha = 2/3$, uma vez que o valor que obtivémos foi $x^{*2}=13.18572$.

3 — Estimação do número médio de doentes em cada concelho, com base no número médio de doentes da região

Admitindo que o número de doentes na Região, no período de 1983 a 1991, é descrito por uma variável aleatória de Poisson, X, de parâmetro λ, e que o número de doentes no concelho i, no mesmo período de tempo, é descrito por uma variável aleatória de Poisson, X_i, de parâmetro λ_i, tem-se que

$$X = \sum_{i=1}^{k} X_i$$

em que k é o número total de concelhos na Região e as variáveis aleatórias X_i são independentes .

O valor médio de X pode, pois, ser decomposto da seguinte maneira:

$$\lambda = \sum_{i=1}^{k} \lambda_i .$$

Para estimarmos λ_i podemos considerar dois estimadores:

—o estimador de máxima verosimilhança sem quaisquer restrições,

$$\hat{\lambda}_i^{(1)} = X_i$$

—o estimador de máxima verosimilhança baseado no modelo $\frac{\lambda_i}{\lambda} = \frac{N_i}{N}$, i=1,...,k, representando N_i o número de habitantes do concelho i e N o número total da habitantes na Região no grupo etário que temos vindo a considerar,

$$\hat{\lambda}_i^{(2)} = X\frac{N_i}{N} .$$

3.1 — *Comparação dos dois estimadores*

—o estimador $\hat{\lambda}_i^{(1)}$ é um estimador centrado de λ_i, enquanto que $\hat{\lambda}_i^{(2)}$ só é um estimador centrado de λ_i se o modelo for verdadeiro, tendo em conta que, nestas condições, podemos escrever $\lambda_i = \lambda\frac{N_i}{N}$;

—cálculo da variabilidade global das estimativas relativamente aos verdadeiros valores dos parâmetros correspondentes aos diferentes concelhos, através do erro médio quadrático esperado, isto é,

$$R_m = \sum_i E\left[\left(\hat{\lambda}_i^{(m)} - \lambda_i\right)^2\right] , m=1, 2 .$$

Note-se que

$$E\left[\left(\hat{\lambda}_i^{(m)} - \lambda_i\right)^2\right] = E\left[\left(\hat{\lambda}_i^{(m)} - E\left(\hat{\lambda}_i^{(m)}\right) + E\left(\hat{\lambda}_i^{(m)}\right) - \lambda_i\right)^2\right]$$

$$= E\left[\left(\hat{\lambda}_i^{(m)} - E\left(\hat{\lambda}_i^{(m)}\right)\right)^2\right] + \left(E\left(\hat{\lambda}_i^{(m)}\right) - \lambda_i\right)^2.$$

A contribuição de cada concelho para o erro médio quadrático esperado pode, assim, ser considerada como a soma de duas componentes : a variância e o viés,

$$\text{var}\left(\hat{\lambda}_i^{(m)}\right) = E\left[\left(\hat{\lambda}_i^{(m)} - E\left(\hat{\lambda}_i^{(m)}\right)\right)^2\right]$$

$$\text{viés}\left(\hat{\lambda}_i^{(m)}\right) = \left(E\left(\hat{\lambda}_i^{(m)}\right) - \lambda_i\right)^2$$

Para $\hat{\lambda}_i^{(1)}$ o viés é sempre nulo. Tem-se, então,

$$R_1 = \sum_{i=1}^{k} \text{var}\left(\hat{\lambda}_i^{(1)}\right) = \sum_{i=1}^{k} \text{var}\left(X_i\right) = \sum_{i=1}^{k} \lambda_i = \lambda.$$

Para $\hat{\lambda}_i^{(2)}$ temos como expressão para o viés

$$\left(E\left(\hat{\lambda}_i^{(2)}\right) - \lambda_i\right)^2 = \left(\lambda \frac{N_i}{N} - \lambda_i\right)^2$$

e

$$\mathrm{var}\left(\hat{\lambda}_i^{[2]}\right) = \mathrm{var}\left(X\frac{N_i}{N}\right) = \left(\frac{N_i}{N}\right)^2 \mathrm{var}(X) = \left(\frac{N_i}{N}\right)^2 \lambda .$$

Logo,

$$R_2 = \sum_{i=1}^{k}\left[\left(\lambda\frac{N_i}{N} - \lambda_i\right)^2 + \left(\frac{N_i}{N}\right)^2 \lambda\right] .$$

As seguintes situações devem ser consideradas :

a) Quando $\frac{\lambda_i}{\lambda} = \frac{N_i}{N}$, i=1,..., k temos que

$$R_2 = \sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2 \lambda = R_1 \sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2 .$$

Uma vez que $\sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2 < 1$, conclui-se que, se o modelo for correcto,

$$R_2 < R_1 .$$

b) Vejamos, agora, a situação em que o modelo não é exactamente correcto:

Suponhamos que $\frac{N_i}{N} = \frac{\lambda_i}{\lambda} + \varepsilon_i$, i=1,...,k, com ε_i suficientemente pequeno de forma a que $\frac{\lambda_i}{\lambda}$ esteja próximo de $\frac{N_i}{N}$.

Tem-se, então,

$$R_2 = \sum_{i=1}^{k} \left\{ \left[\lambda\left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right) - \lambda_i \right]^2 + \left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right)^2 \lambda \right\}$$

$$= \sum_{i=1}^{k} \left[\lambda^2 \left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right)^2 + \lambda_i^2 - 2\lambda_i \lambda\left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right) + \left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right)^2 \lambda \right]$$

$$= \sum_{i=1}^{k} \left[\lambda(\lambda+1)\left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right)^2 - \lambda_i^2 - 2\lambda\lambda_i\varepsilon_i \right]$$

$$= \lambda \sum_{i=1}^{k} \left[(\lambda+1)\left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right)^2 - \left(\frac{\lambda_i}{\lambda} + \varepsilon_i\right)^2 + \lambda\varepsilon_i^2 \right]$$

$$= \lambda \sum_{i=1}^{k} \left[\frac{1}{\lambda^2}\left(\lambda_i + \lambda\varepsilon_i\right)^2 + \lambda\varepsilon_i^2 \right]$$

$$= R_1 \sum_{i=1}^{k} \left[\frac{1}{\lambda^2}\left(\lambda_i + \lambda\varepsilon_i\right)^2 + \lambda\varepsilon_i^2 \right].$$

Para ε_i, i=1,...,k, suficientemente pequeno aquele somatório é inferior 1, (o que se verifica desde que $\sum_{i=1}^{k} \varepsilon_i^2 < \frac{1 - \sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2}{\lambda}$), e, portanto, continuamos a ter

$$R_2 < R_1 .$$

c) O que é que se passaria se os concelhos em consideração tivessem o mesmo número de habitantes ?

(i) Admitamos o mesmo modelo,

$$\frac{\lambda_i}{\lambda} = \frac{N_i}{N} = p \ , \ i=1,\ldots,k,$$

sendo $p = \frac{1}{k}$.

Representando por $\hat{\lambda}_i^{(2)^*}$ o estimador de máxima verosimilhança baseado neste modelo, teríamos

$$\text{viés}\left(\hat{\lambda}_i^{(2)^*}\right) = \left[E\left(\hat{\lambda}_i^{(2)^*}\right) - \lambda_i\right]^2 = (\lambda p - \lambda_i)^2$$

$$\text{var}\left(\hat{\lambda}_i^{(2)^*}\right) = \left(\frac{N_i}{N}\right)^2 \lambda = \lambda p^2.$$

O erro médio quadrático esperado seria

$$R_2^* = \sum_{i=1}^{k} \left[(\lambda p - \lambda_i)^2 + \lambda p^2\right].$$

Se o modelo for correcto, tem-se

$$R_2^* = \lambda k p^2 = \lambda p = R_1 p \ .$$

Nestas condições teremos, novamente,

$$R_2^* < R_1 \ .$$

Vejamos, agora, a relação entre R_2^* e R_2.

Consideremos, então,

$$R_2 = \lambda \sum_{i=1}^{k} \left(\frac{N_i'}{N}\right)^2 \text{ e } R_2^* = \lambda p \text{, em que } p = \frac{N_i}{N} = \frac{1}{k} .$$

Tendo em conta que

$$\sum_{i=1}^{k} \frac{N_i'}{N} = \sum_{i=1}^{k} p = 1 \text{ e que, } \sum_{i=1}^{k} \left(\frac{N_i'}{N} - p\right) = 0,$$

teremos

$$\sum_{i=1}^{k} \left(\frac{N_i'}{N}\right)^2 = \sum_{i=1}^{k} \left[\left(\frac{N_i'}{N} - p\right) + p\right]^2 = \sum_{i=1}^{k} \left[\left(\frac{N_i'}{N} - p\right)^2\right] + p > p, \quad (3.2)$$

ou seja,

$$R_2^* < R_2 .$$

(ii) Se as proporções de habitantes em cada concelho não forem exactamente iguais, mas não forem muito diferentes, isto é, se admitirmos que $\frac{N_i}{N} = p + \varepsilon_i$, com $p = \frac{1}{k}$ e ε_i pequeno, i=1,...,k, tem-se

$$R_2^* = \lambda \sum_{i=1}^{k} (p + \varepsilon_i)^2 < \lambda ,$$

ou seja, teremos, ainda,

$$R_2^* < R_1 .$$

Analisemos, agora, a relação entre R_2^* e R_2.

Note-se que, por (3.2), podemos escrever

$$\sum_{i=1}^{k}\left(\frac{N_i'}{N}\right)^2 - \sum_{i=1}^{k}(p+\varepsilon_i)^2 = \sum_{i=1}^{k}\left(\frac{N_i'}{N}-p\right)^2 + p - p - \sum_{i=1}^{k}\varepsilon_i^2$$

$$= \sum_{i=1}^{k}\left(\frac{N_i'}{N}-p\right)^2 - \sum_{i=1}^{k}\varepsilon_i^2 > 0 \qquad (3.3)$$

Uma vez que considerámos os ε_i, $i=1,\ldots,k$, pequenos, a situação que estamos a analisar só se justifica se aquela diferença for positiva. Podemos, então concluir que, mesmo nestas condições,

$$R_2^* < R_2 .$$

(iii) Se o modelo não for exactamente correcto, ou seja, perante o modelo $p = \frac{\lambda_i}{\lambda} + \varepsilon_i$, $i=1,\ldots,k$, teremos

$$R_2^* = \sum_{i=1}^{k}\left[(\lambda p - \lambda_i)^2 + \lambda p^2\right] = \lambda\sum_{i=1}^{k}\left[\lambda\varepsilon_i^2 + \left(\frac{\lambda_i}{\lambda}+\varepsilon_i\right)^2\right].$$

O último somatório será inferior a 1 para ε_i, $i=1,\ldots,k$, suficientemente pequeno (basta que se tenha $\sum_{i=1}^{k}\varepsilon_i^2 < \frac{1-\frac{1}{k}}{\lambda}$), pelo que

$$R_2^* < R_1 .$$

Comparemos, agora, a variabilidade global das estimativas obtidas para os modelos $p = \frac{\lambda_i}{\lambda} + \varepsilon_i$ e $\frac{N_i'}{N} = \frac{\lambda_i}{\lambda} + \varepsilon_i$, i=1,...,k , (sem perca de generalidade, podemos considerar o mesmo ε_i em ambos os modelos)

Vimos que

$$R_2^* = \lambda \sum_{i=1}^{k} \left[\lambda \varepsilon_i^2 + \left(\frac{\lambda_i}{\lambda} + \varepsilon_i \right)^2 \right] = \lambda \left(p + \lambda \sum_{i=1}^{k} \varepsilon_i^2 \right)$$

$$R_2 = \lambda \sum_{i=1}^{k} \left[\left(\frac{N_i'}{N} \right)^2 + \lambda \varepsilon_i^2 \right] = \lambda \left[\sum_{i=1}^{k} \left(\frac{N_i'}{N} \right)^2 + \lambda \sum_{i=1}^{k} \varepsilon_i^2 \right]$$

Como vimos em (3.2), $\sum_{i=1}^{k} \left(\frac{N_i'}{N} \right)^2 > p$, pelo que se conclui que

$$R_2^* < R_2 .$$

(iv) Admitamos, agora, que os números de habitantes dos concelhos são, apenas, aproximadamente correctos e que, além disso, o modelo não é exactamente correcto, ou seja,

$$p + \delta_i = \frac{\lambda_i}{\lambda} + \varepsilon_i \ , \ i=1,\ldots,k .$$

Neste caso tem-se

$$R_2^* = \sum_{i=1}^{k} \left\{ \left[\lambda(p+\delta_i) - \lambda_i \right]^2 + \lambda(p+\delta_i)^2 \right\} = \lambda \left[\sum_{i=1}^{k} (p+\delta_i)^2 + \lambda \sum_{i=1}^{k} \varepsilon_i^{\,2} \right]$$

Tendo em conta (3.3), podemos, também escrever $\sum_{i=1}^{k} \left(\frac{N_i'}{N} \right)^2 > \sum_{i=1}^{k} (p+\delta_i)^2$ e conclui-se que

$$R_2^* < R_2 .$$

3.2 — *Conclusão*

Conclui-se que o melhor estimador do número médio de doentes num concelho, no que diz respeito à variabilidade global das estimativas, é o estimador baseado na proporção de habitantes do concelho, desde que o modelo proposto inicialmente esteja próximo do verdadeiro modelo. Além disso, a variabilidade global das estimativas é menor se trabalharmos com concelhos com aproximadamente o mesmo número de habitantes.

Como alternativa, podemos, então, agrupar os concelhos de acordo com o número de habitantes.

4 — Sobre a sensibilidade do teste de ajustamento do Qui-Quadrado

Neste parágrafo analisaremos a sensibilidade do teste do qui-quadrado à detecção de discrepâncias entre as frequências observadas e as frequências esperadas, bem como a capacidade do mesmo teste para rejeitar a hipótese nula quando as verdadeiras

probabilidades diferem das apresentadas em H_0 de uma quantidade que é considerada importante em termos da incidência de uma doença.

4.1 — *Discrepâncias entre as frequências observadas e as frequências esperadas*

Consideremos, novamente, a hipótese nula

$$H_0 : \pi_{0i} = \frac{N_i}{N.} \ , i = 1,..., k$$

que estabelece que a incidência da doença é a mesma nos k concelhos. N_i representa o número de habitantes do i-ésimo concelho e N. representa o número total de habitantes nos k concelhos. Para testar H_0 utilizámos a estatística de teste de Pearson, como vimos no parágrafo 2 deste capítulo. Se o modelo que estabelecemos naquele parágrafo fosse exacto, teríamos $X_i = n\frac{N_i}{N.}$, em que X_i representa o número de doentes observado no i-ésimo concelho, e n o número total de doentes observado nos k concelhos.

Admitamos que o modelo não é exactamente correcto, isto é, que

$$\frac{X_i}{n} - \frac{N_i}{N.} = \delta_i \ , i = 1,..., k \ .$$

Isto é equivalente a escrevermos $X_i = n\frac{N_i}{N.} + n\delta_i$. Se as discrepâncias δ_i, i = 1,...,k, forem suficientemente pequenas, o

modelo continua a não ser rejeitado.

A contribuição do i-ésimo concelho para o valor da estatística de teste de Pearson, é

$$\frac{(O_i - E_i)^2}{E_i} = \frac{\left(X_i - n\frac{N_i}{N.}\right)^2}{n\frac{N_i}{N.}} = n\frac{\delta_i^2}{\frac{N_i}{N.}} .$$

Se algumas das discrepâncias δ_i, i=1,...,k, apresentam valores que são considerados importantes no contexto prático em que os dados se inserem e se os concelhos correpondentes são os que apresentam as maiores proporções de habitantes, corremos o risco de aquelas discrepâncias serem "camufladas" pelos valores de $\frac{N_i}{N.}$, sendo pequenas as contribuições destes concelhos para o valor da estatística de teste e não sendo assim detectadas, como desejaríamos, as referidas discrepâncias, para um determinado nível de significância fixo. Se estas discrepâncias fossem observadas em concelhos com proporções de habitantes ligeiramente inferiores o valor obtido para a estatística de teste poderia, agora, levar à rejeição da hipótese nula considerando o mesmo nível de significância, apesar de, no contexto em que os dados se inserem, as discrepâncias a que nos referimos serem consideradas importantes em ambas as situações. Mudar o nível de significância do teste levaria a que discrepâncias pouco importantes observadas em concelhos com proporções de habitantes extremamente baixas provocassem a rejeição da hipótese nula.

No que se refere aos dados que temos vindo a analisar, se a discrepância observada no concelho de Câmara de Lobos for 0.08 (8%), e nos restantes concelhos -0.01, o valor obtido para a

estatística de teste, $x^2 = 13.8079$, levaria à não-rejeição da hipótese nula ao nível de significância de 0.05, quando, se observarmos os dados da Tabela 4.1, uma discrepância de 8% no concelho de Câmara de Lobos é, efectivamente importante sob o ponto de vista da incidência da doença.

TABELA 4.1

CONCELHO	$N_i/N.$
CALHETA	0.1118
P.MONIZ	0.0295
S.VICENTE	0.0643
R.BRAVA	0.1063
C.LOBOS	0.2062
ST.CRUZ	0.1915
MACHICO	0.1647
SANTANA	0.0881
P.SANTO	0.0375

4.2 — *Cálculo da potência do teste de ajustamento do Qui-Quadrado sob hipóteses alternativas típicas de estudos epidemiológicos*

Na sequência do que expusémos anteriormente, consideremos a seguinte hipótese alternativa a H_0

$$H_1:\ \pi_{1i} = \begin{cases} \dfrac{N_j + \omega}{N.} & , \quad i = j \\ \left[N_i - \omega / (k-1)\right] / N., & i \neq j \end{cases} \qquad i = 1,\ldots, k\,.$$

A estatística de teste tem, sob H_1, aproximadamente uma distribuição χ^2 não-central com k-1 graus de liberdade e parâmetro de não-centralidade

$$\lambda = n \sum_{i=1}^{k} \frac{(\pi_{0i} - \pi_{1i})^2}{\pi_{0i}} .$$

Se a discrepância entre π_{0i} e π_{1i} tiver um valor importante no contexto em que os dados se inserem, e se a proporção de habitantes no i-ésimo concelho for elevada, o valor de λ não reflectirá a importância de tal discrepância e, consequentemente a potência do teste será baixa para um nível de significância fixo.

Se, em relação aos dados da doença coronária, fixarmos o nível de significância $\alpha = 0.05$ e tendo em conta que k= 9, podemos observar pelos resultados apresentados na Tabela 4.2, que a capacidade do teste de ajustamento do qui-quadrado de rejeitar H_0 quando a verdadeira probabilidade de contracção das doenças em estudo, no concelho de Câmara de Lobos, difere da estabelecida em H_0 de um valor de 0.08, e as diferenças entre as probabilidades correspondentes aos restantes concelhos e as verdadeiras probabilidades forem de -0.01, é inferior ao que pensamos ser adequado em análises estatísticas de dados deste tipo. O valor da potência do teste, $1 - \beta = 0.757$ (Patnaik, 1949; p. 214), revela que na nossa ânsia de limitar o erro de primeira espécie corremos um risco considerável de não rejeitar uma hipótese nula falsa.

TABELA 4.2

ω / N.	λ	$1-\beta$
0.08	13.81	0.757
0.07	10.57	0.616
0.05	5.39	0.321

Apresentamos também, na mesma tabela, os valores obtidos para a potência do teste para os casos $\frac{\omega}{N.}=0.07$ e $\frac{\omega}{N.}=0.05$, calculados para o mesmo nível de significância $\alpha=0.05$.

A questão que acabámos de analisar sugere que, para testar a hipótese de igualdade de incidência das doenças em estudo em regiões com proporções de habitantes consideravelmente diferentes, seria adequada a aplicação de um teste de ajustamento mais sensível às discrepâncias que temos vindo a considerar.

Uma abordagem porventura interessante consiste em transformar os dados. Esta transformação terá de ser escolhida por forma a dar ênfase às discrepâncias mais elevadas que são "diluídas" por corresponderem aos concelhos com maiores proporções de habitantes. A família das transformações potência de expoente superior a 1, satisfaz estes objectivos, e adiante retomaremos a questão com mais profundidade.

4.3 — *Função característica de uma potência de uma variável aleatória Binomial*

A transformação dos dados observados através de uma transformação potência de expoente não necessariamente inteiro, levou-nos à determinação da função característica de uma potência fraccionária positiva de uma variável aleatória Binomial.

Consideremos $X \cap B(n, p)$. A distribuição de probabilidade e a função característica de uma variável aleatória Binomial com estes parâmetros são dadas, respectivamente, por

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, \ldots, n$$

e

$$\psi_X(t) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} e^{itk} = \left(q + pe^{it}\right)^n,$$

em que $q=1-p$.

Pretendemos determinar uma expressão para a função característica de X^r, que representaremos por $\psi_r(t)$, $\left(= \psi_{X^r}(t)\right)$, em que $r > 0$. A definição de função característica leva-nos a escrever

$$\psi_r(t) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} e^{itk^r}.$$

Expandindo e^{itk^r} em série de Taylor, podemos escrever

$$\psi_r(t) = \sum_{\omega=0}^{\infty} \frac{(it)^\omega}{\omega!} \alpha_{r\omega}$$

em que $\alpha_{r\omega}$ representa o momento de ordem $r\omega$, $(r\omega \geq 0)$, da variável aleatória Binomial de parâmetros n e p, podendo $r\omega$ tomar valores não inteiros.

Com o objectivo de obtermos uma expressão para $\psi_r(t)$ que nos permitisse um conhecimento mais detalhado da função característica da variável aleatória X^r, decidimos estudar os momentos fraccionários de uma variável aleatória Binomial.

Começaremos por estabelecer a relação entre a derivada de ordem r da função característica da variável aleatória Binomial de parâmetros n e p, calculada no ponto t=0, e o momento de ordem r da mesma variável aleatória.

— *Sobre as derivadas fraccionárias de uma função*

A derivada de ordem fraccionária é uma extensão da familiar derivada de ordem n a valores arbitrários (inteiros, racionais, irracionais, ou complexos) de n.

Uma abordagem ao cálculo fraccionário começa com o integral iterado de Cauchy

$$D_{x-x_0}^{-n} f(x) = \int_{x_0}^{x}\int_{x_0}^{t_n}\cdots\int_{x_0}^{t_3}\int_{x_0}^{t_2} f(t_1)\, dt_1 dt_2 \ldots dt_n$$

$$= \int_{x_0}^{x} f(t)\frac{(x-t)^{n-1}}{(n-1)!}\, dt\,.$$

Substituindo -n por α, obtem-se

$$(4.1)\quad D_{x-x_0}^{\alpha} f(x) = \frac{1}{\Gamma(-\alpha)}\int_{x_0}^{x} f(t)(x-t)^{-\alpha-1} dt\,, \quad \mathrm{Re}(\alpha) < 0\,.$$

Quando x_0=0, (4.1) é chamado o integral de Riemann-Liouville,

que é a definição mais comum de derivada fraccionária de ordem α que se encontra na literatura.

Para remover a restrição $\mathrm{Re}(\alpha) < 0$, a derivada fraccionária é representada pela equação

$$D^{\alpha}_{x-x_0} f(x) = \frac{d^m}{dx^m} D^{\alpha-m}_{x-x_0} f(x)$$

em que $m - 1 \leq \mathrm{Re}(\alpha) < m$ e $m=1, 2, 3, \ldots$.

Quando o limite inferior de integração é ∞ ou $-\infty$, obtemos as seguintes expressões que são chamadas as derivadas fraccionárias de Weyl:

$$D^{\alpha}_{x-\infty} f(x) = \frac{(-1)^{-\alpha}}{\Gamma(-\alpha)} \int_x^{\infty} f(t)(t-x)^{-\alpha-1} dt$$

(onde um ramo de $(-1)^{-\alpha}$ tem de ser especificado), e

$$D^{\alpha}_{x+\infty} f(x) = \frac{1}{\Gamma(-\alpha)} \int_{-\infty}^{x} f(t)(x-t)^{-\alpha-1} dt \ .$$

A definição apresentada por Marchaud (1927) para $0 < \alpha < 1$, é

$$D^{\alpha}_{x+\infty} f(x) = \frac{\alpha}{\Gamma(1-\alpha)} \int_{-\infty}^{x} \frac{f(x) - f(t)}{(x-t)^{\alpha+1}} dt \ .$$

Algumas das clássicas propriedades da diferenciação e da

integração generalizam-se sem grandes alterações para a diferintegração — a linearidade da diferintegração é uma consequência imediata de qualquer das definições apresentadas, tendo-se

$$D^{\alpha}_{x-x_0}[f(x)+g(x)] = D^{\alpha}_{x-x_0}f(x) + D^{\alpha}_{x-x_0}g(x);$$

e

$$D^{\alpha}_{x-x_0}[C\ f(x)] = C\ D^{\alpha}_{x-x_0}f(x), \quad C \text{ constante arbitrária.}$$

Outras propriedades requerem algumas alterações — a generalização da regra de Leibniz do produto de duas funções, para a diferintegração, apresentada por Osler (1972)

$$D^{\alpha}_{x}[f(x)g(x)] = \sum_{j=-\infty}^{\infty} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha-\gamma-j+1)\Gamma(\gamma+j+1)} D^{\alpha-j-\gamma}_{x} f(x) D^{j+\gamma}_{x} g(x)$$

em que γ é arbitrário; a fórmula da n-ésima derivada de uma função composta é generalizada para a diferenciação fraccionária

$$D^{\alpha}_{x} f(h(x)) = \sum_{j=0}^{\infty} U_j(x)\ D^{j}_{h(x)} f(h(x)) / j!$$

em que

$$U_j(x) = \sum_{k=0}^{j} \binom{j}{k} (-h(x))^k D^{\alpha}_{x}[h(x)^{j-k}]$$

desde que $h^{-1}(0) = 0$.

A relação

$$\frac{d^{\alpha}}{dx^{\alpha}}\frac{d^{\beta}}{dx^{\beta}}f(x)=\frac{d^{\alpha+\beta}}{dx^{\alpha+\beta}}f(x)$$

que é válida quando α e β são números naturais, não é sempre válida para α e β arbitrários.

Se considerarmos $\alpha = m + \lambda$ com $m \geq 0$, inteiro, e $0 \leq \lambda < 1$, poderemos escrever

$$D^{\alpha}_{x-x_0}f(x)=D^{\lambda+m}_{x-x_0}f(x)=D^{\lambda}_{x-x_0}\frac{d^m}{dx^m}f(x)$$

desde que $D^{-m}_{x-x_0}\frac{d^m}{dx^m}f(x)=f(x)$ (Oldham and Spanier, 1974, p.86).

Voltando agora ao cálculo dos momentos fraccionários de uma variável aleatória Binomial, consideremos $r=m+\lambda$, sendo $m \geq 0$ inteiro e $0 < \lambda < 1$.

A derivada fraccionária de ordem r de $\psi_X(t)$ em ordem a t, pode ser escrita na forma

$$D^{r}_{t+\infty}\psi_X(t)=D^{\lambda}_{t+\infty}\,\psi^{(m)}_X(t),$$

uma vez que $\psi^{(m)}_X(t)=\sum_{k=1}^{n}\binom{n}{k}\,p^k q^{n-k}(ik)^m e^{itk}$ e, se fizermos

$t_0 = -\infty$, $\psi_X(t)$ satisfaz a relação $D^{-m}_{t-t_0} \frac{d^m}{dx^m} \psi_X(t) = \psi_X(t)$.

Utilizando a definição de derivada fraccionária apresentada por Marchaud, obtemos a seguinte expressão para a derivada de ordem r em ordem a t da função característica da variável aleatória Binomial

$$D^r_{t+\infty} \psi_X(t) = D^{\lambda}_{t+\infty} \psi^{(m)}_X(t) = \frac{\lambda}{\Gamma(1-\lambda)} \int_{-\infty}^{t} \frac{\psi^{(m)}_X(t) - \psi^{(m)}_X(u)}{(t-u)^{1+\lambda}} du$$

Através de mudança de variável e, pelas propriedades da função Gamma, tendo em conta que $0 < \lambda < 1$, obtem-se

$$D^r_{t+\infty} \psi_X(t) = \frac{-1}{\Gamma(-\lambda)} \int_0^{\infty} \frac{\psi^{(m)}_X(t) - \psi^{(m)}_X(t-u)}{u^{1+\lambda}} du .$$

Substituindo $\psi^{(m)}_X(t)$ pela respectiva expressão, podemos escrever

$$D^r_{t+\infty} \psi_X(t) = \frac{-1}{\Gamma(-\lambda)} i^m \int_0^{\infty} \left\{ \left[\sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} k^m \left(e^{itk} - e^{i(t-u)k} \right) \right] / u^{1+\lambda} \right\}$$

$$= \frac{-1}{\Gamma(-\lambda)} i^m \sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} k^m e^{itk} \int_0^\infty \frac{1 - e^{-iuk}}{u^{1+\lambda}} \, du \, .$$

Calculemos, agora, $\int_0^\infty \frac{1 - e^{-iuk}}{u^{1+\lambda}}$. Escrevendo

$$\int_0^\infty \frac{1 - e^{-iuk}}{u^{1+\lambda}} = \int_0^\infty \frac{1 - \cos(uk) + i \, \text{sen}(uk)}{u^{1+\lambda}} \, du$$

$$= k^\lambda \int_0^\infty \frac{1 - \cos(uk)}{(uk)^{1+\lambda}} \, k \, du + i \, k^\lambda \int_0^\infty \frac{\text{sen}(uk)}{(uk)^{1+\lambda}} \, k \, du \, .$$

Tem-se (Gradshtein e Ryzhik, 1963, p.491)

$$\int_0^\infty \frac{e^{-\beta x} \cos(ax) - e^{-\gamma x} \cos(bx)}{x^\mu} \, dx$$

$$= \Gamma(1-\mu) \left\{ \left(a^2 + \beta^2\right)^{\frac{\mu-1}{2}} \cos\left[(\mu - 1) \, \text{arctg}\left(\frac{a}{\beta}\right) \right] \right.$$

$$\left. - \left(b^2 + \gamma^2\right)^{\frac{\mu-1}{2}} \cos\left[(\mu - 1) \, \text{arctg}\left(\frac{b}{\gamma}\right) \right] \right\}$$

em que Re $\beta > 0$, Re $\gamma > 0$, $\mu < 2$ e $\mu \neq 1$. Para aplicarmos este resultado (com $\mu = \lambda + 1$, $a = 0$ e $b = 1$), ao cálculo de

$\int_0^{\infty} \frac{1-\cos(uk)}{(uk)^{1+\lambda}}\,du$, teremos de verificar que

$$\lim_{\beta,\gamma\to 0}\int_0^{\infty} \frac{e^{-\beta x}-e^{-\gamma x}\cos(x)}{x^{\mu}}\,dx = \int_0^{\infty} \lim_{\beta,\gamma\to 0} \frac{e^{-\beta x}-e^{-\gamma x}\cos(x)}{x^{\mu}}\,dx$$

$$= \int_0^{\infty} \frac{1-\cos(x)}{x^{\mu}}\,dx .$$

Consideremos a sucessão de funções

$$f_n(x) = \frac{e^{-x/n}-e^{-x/n}\cos(x)}{x^{\mu}} = e^{-x/n}\,\frac{1-\cos(x)}{x^{\mu}},\quad n\in\mathbb{N} .$$

Para $x\in]\,0,\infty\,[$, tem-se

$$|f_n(x)| = e^{-x/n}\left|\frac{1-\cos(x)}{x^{\mu}}\right| < \left|\frac{1-\cos(x)}{x^{\mu}}\right|,\quad \forall n .$$

Consideremos

$$h(x) = \left|\frac{1-\cos(x)}{x^{\mu}}\right| .$$

No caso que estamos a estudar, tem-se $\mu = 1+\lambda$, pelo que $1<\mu<2$. Nestas condições tem-se que $\lim_{x\to 0} \frac{1-\cos(x)}{x^{\mu}} = 0$, pelo que poderemos escrever, para um dado $\varepsilon > 0$,

$$\int_0^{\infty} \left| \frac{1-\cos(x)}{x^{\mu}} \right| dx = \int_0^{\varepsilon} \left| \frac{1-\cos(x)}{x^{\mu}} \right| dx + \int_{\varepsilon}^{\infty} \left| \frac{1-\cos(x)}{x^{\mu}} \right| dx < \infty .$$

Sendo h(x) integrável em $]0,\infty[$ e, no mesmo intervalo $|f_n(x)| < h(x), \forall n$, pelo teorema da convergência dominada de Lebesgue podemos escrever

$$\int_0^{\infty} \frac{1-\cos(x)}{x^{\mu}} dx =$$

$$= \lim_{\beta,\gamma \to 0} \Gamma(1-\mu) \left\{ \left(\beta^{\mu-1}\right) - \left(1+\gamma^2\right)^{\frac{\mu-1}{2}} \cos\left[(\mu-1) \operatorname{arctg}\left(\frac{1}{\gamma}\right) \right] \right\}$$

$$= -\Gamma(1-\mu) \cos\left[(\mu-1)\frac{\pi}{2} \right] .$$

Recorrendo ao resultado semelhante obtido para a função seno (Gradshtein e Ryzhik, 1963, p.491), teríamos

$$\int_0^{\infty} \frac{\operatorname{sen}(x)}{x^{\mu}} dx = -\Gamma(1-\mu) \operatorname{sen}\left[(\mu-1)\frac{\pi}{2} \right].$$

Podemos, finalmente, escrever

$$\int_0^{\infty} \frac{1-e^{-iuk}}{u^{1+\lambda}} du = -k^{\lambda}\Gamma(-\lambda)\cos\left(\lambda\frac{\pi}{2}\right) - i\, k^{\lambda}\Gamma(-\lambda)\operatorname{sen}\left(\lambda\frac{\pi}{2}\right)$$

$$= -k^{\lambda}\Gamma(-\lambda)\ e^{i\lambda\frac{\pi}{2}}.$$

Voltando à expressão de $D^{r}_{t+\infty}\psi_X(t)$, tem-se, então

$$D^{r}_{t+\infty}\psi_X(t) = \frac{-1}{\Gamma(-\lambda)} i^{m} \sum_{k=1}^{n} \binom{n}{k}\ p^{k}q^{n-k}k^{m}e^{itk}\left[-k^{\lambda}\Gamma(-\lambda)\ e^{i\lambda\frac{\pi}{2}}\right]$$

$$= i^{m}e^{i\lambda\frac{\pi}{2}} \sum_{k=1}^{n} \binom{n}{k}\ p^{k}q^{n-k}k^{r}e^{itk},$$

e

$$D^{r}_{t+\infty}\psi_X(t)\Big|_{t=0} = i^{m}e^{i\,\lambda\frac{\pi}{2}} \sum_{k=1}^{n} \binom{n}{k}\ p^{k}q^{n-k}k^{r}.$$

Tem-se, então,

$$(-i)^{m} D^{r}_{t+\infty}\psi_X(t)\Big|_{t=0} = e^{i\lambda\frac{\pi}{2}} \sum_{k=1}^{n} \binom{n}{k}\ p^{k}q^{n-k}k^{r}$$

$$= \left[\cos\left(\lambda\frac{\pi}{2}\right) + i\ \mathrm{sen}\left(\lambda\frac{\pi}{2}\right)\right] \sum_{k=1}^{n} \binom{n}{k}\ p^{k}q^{n-k}k^{r}$$

e, finalmente,

$$\mathrm{Re}\left[(-i)^m D^r_{t+\infty}\psi_X(t)\Big|_{t=0}\right] = \alpha_r \cos\left(\lambda\frac{\pi}{2}\right),$$

obtendo-se, assim, a relação

$$\alpha_r = \frac{1}{\cos\left(\lambda\frac{\pi}{2}\right)}\mathrm{Re}\left[(-i)^m D^r_{t+\infty}\psi_X(t)\Big|_{t=0}\right].$$

Tentámos, a partir deste resultado, obter uma expressão para o momento fraccionário de uma variável aleatória Binomial de parâmetros n e p.

Se escrevermos $\psi_X(t)$ na forma

$$\psi_X(t) = q^n\left(1+\frac{p}{q}e^{it}\right)^n,$$

obtem-se, de acordo com o exposto anteriormente,

$$D^r_{t+\infty}\psi_X(t) = q^n D^\lambda_{t+\infty}\left[\frac{d^m}{dt^m}\left(1+\frac{p}{q}e^{it}\right)^n\right]$$

em que $r = m+\lambda$, $m \geq 0$ inteiro e $0 < \lambda < 1$.

Calculando a derivada de ordem inteira do segundo membro da relação anterior, obtem-se

$$\frac{d^m}{dt^m}\left(1+\frac{p}{q}e^{it}\right)^n = \sum_{k=0}^{n} u_k(t)\binom{n}{k}\left(1+\frac{p}{q}e^{it}\right)^{n-k}$$

em que

$$u_k(t) = \sum_{j=0}^{k}\binom{k}{j}(-1)^j\left(\frac{p}{q}\right)^k e^{itj}\frac{d^m}{dt^m}e^{it(k-j)}$$

$$= \left(\frac{p}{q}\right)^k i^m e^{itk}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(k-j)^m.$$

Tem-se, então,

$$D^r_{t+\infty}\psi_X(t) = q^n i^m\left\{\sum_{k=0}^{m}(-1)^k\binom{n}{k}\left(\frac{p}{q}\right)^k\left[\sum_{j=0}^{k}\binom{k}{j}(-1)^j j^m\right]\right.$$

$$\left. D^\lambda_{t+\infty}e^{itk}\left(1+\frac{p}{q}e^{it}\right)^{n-k}\right\}.$$

Temos, pois, que calcular

$$D^\lambda_{t+\infty}e^{itk}\left(1+\frac{p}{q}e^{it}\right)^{n-k} = \sum_{v=0}^{n-k}\binom{n-k}{v}\left(\frac{p}{q}\right)^v D^\lambda_{t+\infty}e^{it(k+v)}.$$

Aplicando a definição de derivada fraccionária de Marchaud ao cálculo de $D^{\alpha}_{t+\infty}e^{itk}$ em que $\alpha = m + \lambda$, $m \geq 0$ inteiro e $0 < \lambda < 1$, e utilizando resultados anteriormente apresentados, obtem-se

$$D^{\alpha}_{t+\infty}e^{itk} = \frac{-1}{\Gamma(-\lambda)}\int_0^{\infty} \frac{(ik)^m e^{itk} - (ik)^m e^{ik(t-u)}}{u^{1+\lambda}}\, du$$

$$= \frac{-1}{\Gamma(-\lambda)}(ik)^m e^{itk}\int_0^{\infty} \frac{1 - e^{-iuk}}{u^{1+\lambda}}\, du$$

$$= i^m k^{m+\lambda} e^{itk}\left[\cos\left(\lambda\frac{\pi}{2}\right) + i\ \mathrm{sen}\left(\lambda\frac{\pi}{2}\right)\right]$$

$$= i^m k^{\alpha} e^{i\left(tk+\lambda\frac{\pi}{2}\right)}$$

Podemos, agora, escrever

$$D^{\lambda}_{t+\infty}e^{itk}\left(1 + \frac{p}{q}e^{it}\right)^{n-k} = \sum_{v=0}^{n-k}\binom{n-k}{v}\left(\frac{p}{q}\right)^v (k+v)^{\lambda} e^{i\left[t(k+v)+\lambda\frac{\pi}{2}\right]}$$

Finalmente, obtem-se

$$D^r_{t+\infty}\psi_X(t) = i^m \sum_{k=0}^{m} \binom{n}{k} (-1)^k p^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j j^m$$

$$\sum_{v=0}^{n-k} \binom{n-k}{v} p^v q^{n-k-v} (k+v)^\lambda e^{i\left[t(k+v)+\lambda\frac{\pi}{2}\right]}$$

e

$$D^r_{t+\infty}\psi_X(t)\Big|_{t=0} = i^m e^{i\lambda\frac{\pi}{2}} \sum_{k=0}^{m} \binom{n}{k} (-1)^k p^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j j^m$$

$$\sum_{v=0}^{n-k} \binom{n-k}{v} p^v q^{n-k-v} (k+v)^\lambda .$$

Uma expressão para o momento de ordem $r > 0$ de uma variável aleatória Binomial de parâmetros n e p será, então,

$$\alpha_r = \sum_{k=0}^{m} \binom{n}{k} (-1)^k p^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j j^m$$

$$\sum_{v=0}^{n-k} \binom{n-k}{v} p^v q^{n-k-v} (k+v)^\lambda$$

em que $r = m + \lambda$, $m \geq 0$ inteiro e $0 \leq \lambda < 1$, uma vez que esta

expressão continua a ser válida para $\lambda = 0$.

Uma expressão para a função característica de uma potência não necessariamente inteira de uma variável aleatória Binomial de parâmetros n e p, poderá ser obtida a partir desta expressão do momento fraccionário

$$\psi_r(t) = \sum_{\omega=0}^{\infty} \frac{(it)^{\omega}}{\omega!} \sum_{k=0}^{\lfloor r\omega \rfloor} \binom{n}{k} (-1)^k p^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j j^{\lfloor r\omega \rfloor} \sum_{v=0}^{n-k} \binom{n-k}{v} p^v q^{n-k-v} (k+v)^{r\omega - \lfloor r\omega \rfloor}$$

ou, ainda,

$$\psi_r(t) = \sum_{\omega=0}^{\infty} \frac{(it)^{\omega}}{\omega!} \sum_{k=0}^{\lfloor r\omega \rfloor} \delta_{k,\omega} \, E\left[(Y+k)^{\lambda\omega - \lfloor \lambda\omega \rfloor}\right]$$

sendo $r = m + \lambda$, $m \geq 0$ inteiro e $0 \leq \lambda < 1$, e em que Y é uma variável aleatória Binomial de parâmetros n-k e p e os coeficientes $\delta_{k,\omega}$, $k = 0, 1, \ldots, \lfloor r\omega \rfloor$, são dados por

$$\delta_{k,\omega} = \binom{n}{k} (-1)^k p^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j j^{\lfloor r\omega \rfloor}.$$

Limitações temporais impedem-nos de prosseguir esta questão. O rumo da nossa investigação orienta-se agora para o estudo de velocidades de convergência, tentando dilucidar o problema de que expoente escolher na família

$$\{Y^r; Y \cap \mathcal{M}(n,\mathbf{p}) \text{ e } r \in \mathbb{R}\}$$

por forma a que a convergência para a distribuição limite — que é sempre a distribuição do Qui-Quadrado — seja tão rápida quanto possível. Os métodos que temos vindo a ensaiar são os que permitiram a Gomes (1978) investigar o problema da "penultimate approximations" em extremos, que posteriormente (Gomes e Pestana, 1984) foi investigada com toda a generalidade para variáveis no domínio de atracção da lei generalizada de von Mises—Jenkinson, e não apenas na perspectiva clássica de Fisher e Tippett (1928) e seus seguidores, que apenas tinham investigado o domínio de atracção da lei de Gumbel, e a Canto e Castro (1992) investigar em profundidade velocidades de convergência.

Assim, a perspectiva é começar por analisar a velocidade de convergência da potência da Binomial para o limite $\chi^2_{(1)}$, tentando usar expansões em série de Taylor — de onde a importância da abordagem "diferintegral" que atrás apresentámos — procurando que para potências especiais se anulem os termos de ordem mais baixa. Mas só o futuro nos mostrará a eventual riqueza desta nova perspectiva nos estudos sobre a "liberalidade" do Qui-Quadrado, que na última década tem sido tão discutida.

É de realçar que mesmo que a velocidade de convergência não melhore será interessante abordar a questão da relação entre o expoente a utilizar na transformação dos dados e a potência do teste de ajustamento do Qui-Quadrado.

Referências

Andersen, E. (1990) *The Satistical Analysis of Categorical Data.* Springer Verlag.

Canto e Castro, L. (1992) *Sobre a Teoria Assintótica de Extremos*, Tese de Doutoramento, F.C.U.L.

Cressie, N. and Read, T. (1984) *Multinomial Goodness-of-fit Tests*. R. Statist.Soc.B, 46, pp. 440-464

Fisher, R. e Tippett, L. (1928) Limiting forms of the frequency distribution of the largest and smallest member of a sample. Proc. Camb. Phil. Soc. 24, 180.

Gomes, I. (1978) *Some Probabilistic and Statistical Problems in Extreme Value Theory.* Tese de Doutoramento, Universidade de Sheffield.

Gomes, I. e Pestana, D. (1984) *Domains of Attraction and Penultimate Behaviour*. Abs. 16th European Meeting Statisticians, Marburg.

Gradshteyn, I. e Ryzhik, I. (1963) *Table of integrals, series and products.* Academic Press.

Kolmogorov, A. e Fomin, S. (1982) *Elementos da Teoria das Funções e de Análise Funcional*. Editora Mir, Moscou.

Laha, R. e Rohatgi, V. (1979) *Probability Theory.* Wiley.

Laue, G. (1980) *Remarks on the relation between fractional moments and fractional derivatives of characteristic functions.* J. Appl. Prob., 17, pp.456-466.

Lavoie, J. , Osler, T. e Tremblay, R. (1976) *Fractional Derivatives and Special Functions*. SIAM Review, 18, pp. 240-268.

Marchaud, A. (1927) *Sur les dérivées et sur les différences des fonctions de variables réelles* . Journ. de Math. , tome VI.—Fasc.IV, pp. 337-425.

Oberman, A. (1992) *Epidemiology and Prevention of Cardiovascular Disease.* Text Book of Internal Medicine, Vol.1, Chap.34, J.B. Lippincoh Company.

Oldham, K. e Spanier, J. (1974) *The Fractional Calculus.* Academic Press

Osler, T. (1970) *Leibniz rule for fractional derivatives generalized and an application to infinite series* . SIAM J. Appl. Math. , 18, pp. 658-674.

Osler, T. (1970) *The fractional derivative of a composite function* . SIAM J. Math. Anal. 1, pp. 288-293.

Osler, T. (1971) *Taylor's series generalized for fractional derivatives and applications* . SIAM J. Math. Anal., 2, pp. 37-48.

Osler, T. (1972) *A further extension of the Leibniz rule to fractional derivatives and its relation to Parseval 's formula.* SIAM J. Math. Anal., 3, pp. 1-15.

Osler, T. (1972) *An Integral Analogue of Taylor's Series and Its Use in Computing Fourier Transforms* . Mathematics of Computation, 26, pp. 449-459.

Osler, T. (1972) *The Integral Analog of the Leibniz Rule.* Mathematics of Computation, 26, pp. 903-915.

Patnaik, P.B. (1949) *The Non-Central* χ^2 *- and F-Distributions and their Applications.* Biometrika, 36, pp. 202-232.

Read, T. (1984) *Small-Sample Comparisons for the Power Divergence Goodness-of-Fit Statistics.* JASA, 79, pp. 929-935.

Tsui, Kam-Wah and Press, S.J. (1982) *Simultaneous Estimation of several Poisson Parameters under K-Normalized Squared Error Loss.* The Annals of Statistics, 1, pp. 93-100.

Yosida, K. (1966) *Functional Analysis.* Springer Verlag.

Vasconcelos, Rita (1993) *The relevance of spatial statistics on the statistical model building for coronary heart disease.* 2nd European Conference on

Mathematics Applied to Biology and Medicine, Lyon.

Vasconcelos, Rita (1994) *Análise Exploratória de uma base de dados sobre Cardiologia. Atribuição de uma dimensão espacial à doença coronária?* A Estatística e o Futuro e o Futuro da Estatística, D. Pestana et al, eds. Salamandra, Lisboa, pp.197-228.

APENDICE I

DESCRIÇÃO DAS VARÁVEIS CONSTANTES DO FICHEIRO DO SERVIÇO DE CARDIOLOGIA DO C.H.F.

PC$
(Número ficha hospitalar)

DATANASC$

IDADE

SEXO — V. qualitativa, 2 categorias:
1.- MASCULINO
2.- FEMININO

MORADA$

CONCELHO — V. qualitativa, 12 categorias:
1.- FUNCHAL
2.- STA. CRUZ
3.- MACHICO
4.- C. LOBOS
5.- CALHETA
6.- PONTA SOL
7.- P. MONIZ
8.- RIB. BRAVA
9.- S. VICENTE
10.- SANTANA
11.- PORTO SANTO
12.- ESTRANGEIROS

PROFISSAO$

TEMPSEG

ANGPREV
(Angina prévia) — V. qualitativa, 4 categorias:
1.- NÃO
2.- S.ESFORÇO

	3.- S.INSTÁVEL 4.- S.PRINZMET
INTANGPREV (Internamento por ang. prév.)	V. qualitativa, 4 categorias: 1.- SIM 2.- NÃO
EAMPREV (Enfarte prévio)	V. qualitativa, 8 categorias: 1.- NÃO 2.- S. ANTERIOR 3.- S. INFERIOR 4.- S. LATERAL 5.- COMBINADO 6.- NÃO Q. 7.- INDETERM.
NUMEAMPRE	V. qualitativa, 4 categorias: 1.- NÃO 2.- SIM.1 3.- SIM.2 4.- SIM.3
HIPERTPREV (Hipertensão prévia)	V. qualitativa, 4 categorias 1.- NÃO 2.- S.MEDIC 3.- S.NÃO.MED 4.- NÃO SABE
DIABETES	V. qualitativa, 5 categorias: 1.- NÃO 2.- S.DIETA 3.- S.ADO 4.- S.INSUL 5.- OUTRA
DISLIPPREV (Dislipidémia prévia)	V. qualitativa, 9 categorias 1.- NÃO

2.- C.NORMAL
3.- C.200.240
4.- C.240.300
5.- C.MA.300
6.- TRIG. ALTOS
7.- COMB.2.6
8.- COMB.6.OUT
9.- OUTROS

TABACO — V. qualitativa, 3 categorias:
1.- NÃO
2.- SIM, MENOS DE 20
3.- SIM, MAIS DE 20

DIAGNOST — V. qualitativa, 8 categorias:
1.- ANG. INST
2.- EAM
3.- ARRITMIA
4.- BAV
5.- EAP
6.- TAMPON
7.- OUTRO
8.- EMB. PULM

COMENT$

MORTE — V. qualitativa, 4 categorias:
1.- NÃO
2.- SIM.UTIC
3.- SIM.ENF
4.- SIM.FOLLOW

DATAMORT$

DATAINTER$

IDADEINTER
(Anos,dias)

IDADEINTER
(Idade de internamento transformada em dias)

SOBREVIDA
(Data morte - data internamento)

COMENTAR$

LOCEAMACT

(Localização do enfarte actual)

V. qualitativa, 8 categorias:

1.- ANTERIOR
2.- INFERIOR
3.- LATERAL
4.- COMBINADO
5.- INDETERM
6.- NÃO.Q.
7.- OUTRO
8.- NÃO

KILLIP
(Classe de enf.)

V. qualitativa, 5 categorias:
1.- NÃO
2.- I
3.- II
4.- III
5.- IV

TROMBOLISE

V. qualitativa, 3 categorias:
1.- NÃO
2.- SIM.STK
3.- SIM.APSAC

ARRITISQ
(Arritmia isquémica)

V. qualitativa, 7 categorias:
1.- NÃO
2.- TV
3.- FA
4.- BAVC
5.- OUT.BAV
6.- FV
7.- ASSISTOLIA

ETIMORTE	V. qualitativa, 6 categorias: 1.- SÚBITA 2.- ASSISTOLIA 3.- CHOQUE 4.- DIS. EL-MEC 5.- EAP. IRREV 6.- OUTRA
PACEPROV	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
COMPLEAM (Complicações do enfarte)	V. qualitativa, 6 categorias: 1.- NÃO 2.- CIV. 3.- IM 4.- EMBOL. PER 5.- PERICARD. 6.- OUTRA
LIDOCAINA (Tratemento)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
ANTIARRT (Tratamento)	V. qualitativa, 5 categorias: 1.- NÃO 2.- PROPAFENO 3.- AMIODARONA 4.- QUINIDINA 5.- OUTRO
VERAPAMEV (Tratamento)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
DNI (Tratamento)	V. qualitativa, 4 categorias: 1.- NÃO

2.- SIM . EV
3.- SIM . ORAL
4.- COMB. 2.3

HEPARINA (Tratamento)

V. qualitativa, 3 categorias:
1.- NÃO
2.- E.V.
3.- S.C.

ASPIRINA (Tratamento)

V. qualitativa, 6 categorias:
1.- NÃO
2.- SIM . EV
3.- S. ORAL 100
4.- S. ORAL 250
5.- COMB. 2.3
6.- COMB. 2.4

ANTCALCIO (Tratamento)

V. qualitativa, 4 categorias:
1.- NÃO
2.- DILTIAZEM
3.- NIFEDIPINA
4.- VERAPAMIL

BETABLOCK (Tratamento)

V. qualitativa, 3 categorias:
1.- NÃO
2.- SIM . ORAL
3.- SIM . EV

DIGOXINA (Tratamento)

V. qualitativa, 2 categorias:
1.- NÃO
2.- SIM

DIURETICOS (Tratamento)

V. qualitativa, 3 categorias:
1.- NÃO
2.- FUROSEMIDA
3.- TIAZIDA

CARGAH20 (Tratamento)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
INIBECA (Tratamento)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
VENTILAÇAO (Tratamento)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
CARDIOVER	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
CPKMAX	
FC	
ANGPOSEAM (Angina pós enfarte)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
TIPFOLLOW (Tipo de follow-up)	V. qualitativa, 6 categorias: 1.- SEM SABER 2.- HOSP. ACT 3.- HOSP. PASS 4.- PRIVADO 5.- MISTO 6.- OUTRO
CCARDTARD (Complicações cardíacas tardias)	V. qualitativa, 9 categorias: 1.- NÃO 2.- ANGINA 3.- ICE 4.- ARRITMIAS 5.- COMB. 2.3 6.- COMB. 2.4

7.- COMB. 3.4
8.- COMB. 3.5
9.- OUTRA

CNAOCARD
(Complicações não-cardíacas)

V. qualitativa, 4 categorias:
1.- NÃO
2.- AVC
3.- CIRUR. MI
4.- OUTRA

TIPARRITM
(Tipo de arritmia)

V. qualitativa, 6 categorias:
1.- NÃO
2.- TPSV
3.- FA. RAP
4.- TV
5.- FV
6.- OUTRA

CORONARIAS

V. qualitativa, 2 categorias
1.- NÃO
2.- SIM

OPERADO

V. qualitativa, 2 categorias:
1.- NÃO
2.- SIM

COMENTF$

ACPREV1IN
(Acidentes prévios ao 1° internam.)

V. qualitativa, 9 categorias:
1.- NÃO
2.- A. ESTÁVEL
3.- A. INSTÁVEL
4.- EAM. ANT
5.- EAM. INF
6.- EAM. LAT
7.- EAM. NÃO-Q
8.- EAM. IND
9.- EAM. COMB

MORTNAOCAR
(Mortalidade não-cardíaca)

V. qualitativa, 4 categorias:
1.- NÃO
2.- AVC
3.- NEO
4.- OUTRA

DAT1INT$
(data do 1° internamento)

DIAGN1IN
(Diagnóstico do 1° internam.)

V. qualitativa, 15 categorias:
1.- A. INSTÁVEL
2.- EAM. ANT
3.- EAM. INF
4.- EAM. LAT
5.- EAM. NÃO-Q
6.- EAM. INDET
7.- EAM. COMB
8.- ARR. FA. RAP
9.- ARR. FA. LEN
10.- ARR. FLUT
11.- ARR. TPSV
12.- ARR. TV
13.- ARR. FV
14.- EAP
15.- ICC

EAMKILL1
(Kill do 1° enfarte)

V. qualitativa, 4 categorias:
1.- I
2.- II
3.- III
4.- IV

EAMTROM1

V. qualitativa, 4 categorias:
1.- NÃO
2.- STK
3.- APSAC
4.- RT-AP

COMIS1IN (Complicações 1° intern)	V. qualitativa, 2 categorias: 1.- NÃO 2.- SIM
COMME1IN (complicações 1° intern)	V. qualitativa, 6 categorias: 1.- NÃO 2.- PERICÁRDIO 3.- CIV 4.- IM 5.- EMB.PERIF 6.- OUTRA
COMAR1IN (Complicaçõs 1° intern)	V. qualitativa, 7 categorias: 1.- NÃO 2.- FA 3.- TV 4.- FV 5.- BAVC 6.- BAV 1° 2° 7.- OUTRA
COFOLCAR1 (Complicações)	V. qualitativa, 9 categorias: 1.- NÃO 2.- ANGINA 3.- ICE 4.- ARRITMIA 5.- COM. 2.3 6.- COM. 2.4 7.- COM. 3.4 8.- COM. 2.3.4 9.- OUTRA
COFOLVAS1 (Complicações)	V. qualitativa, 4 categorias: 1.- NÃO 2.- AVC 3.- EMB. PERIF 4.- OUTRA

As variáveis para os internamentos seguintes definem-se de forma idêntica às variáveis correspondentes definidas para o 1º internamento.

DATA2IN$
(Data 2º internamento)

DIAGN2IN
(diagnóstico 2º intern)

EAMKILL2

EAMTROM2

COMPLIS2

COMPLME2

COMPLAR2

COFOLCA2

COFOLVA2

DATA3IN$

DIAGN3IN

EAMKILL3

EAMTROM3

COMIS3IN

COMME3IN

COMAR3IN

COFOLCA3

COFOLVA3

DATA4IN$

DIAGN4IN

EAMKILL4

EAMTROM4

COMIS4IN

COMME4IN

COMAR4IN

COFOLCA4

COFOLVA4

DIAGNUIN

EAMKILLN

EAMTROMN

COMPISNU

COMPMENU

COMPARNU

COFOLCAN

COFOLVAN

DATNUIN$

APENDICE II

Date: 13-JUN-94
Time: 15:47:22
File: HD 80:Systat:MADEIRA.SYS:PORTO MONIZ Sys
has 8 variables and 1 cases.

OBS	CONCELHO$	MAPNUM	MAXLAT	MINLAT	MAXLON
1	P.Moniz	3.000	32.950	32.699	-17.062

OBS	MINLON	LABLAT	LABLON
1	-17.225	32.815	-17.140

1 cases printed out of 1 cases in the file.

Date: 13-JUN-94
Time: 15:44:16
File: HD 80:Systat:MADEIRA.SYS:madeira
has 8 variables and 10 cases.

OBS	CONCELHO$	MAPNUM	MAXLAT	MINLAT	MAXLON
1	Calheta	1.000	32.838	32.699	-17.072
2	Ponta Sol	2.000	32.768	32.674	-17.042
3	P.Moniz	3.000	32.950	32.699	-17.062
4	S.Vicente	4.000	32.827	32.700	-16.950
5	R.Brava	5.000	32.754	32.649	-16.980
6	C.Lobos	6.000	32.759	32.641	-16.935
7	Funchal	7.000	32.742	32.620	-16.700
8	St.Cruz	8.000	32.683	32.634	-16.770
9	Machico	9.000	32.788	32.698	-16.725
10	Santana	10.000	32.828	32.757	-16.834

OBS	MINLON	LABLAT	LABLON
1	-17.261	32.758	-17.160
2	-17.137	32.690	-17.090
3	-17.225	32.815	-17.140
4	-17.200	32.786	-16.997
5	-17.087	32.705	-17.018
6	-17.018	32.668	-16.970
7	-16.969	32.670	-16.897
8	-16.854	32.680	-16.815
9	-16.900	32.740	-16.795
10	-16.962	32.780	-16.888

10 cases printed out of 10 cases in the file.

APENDICE III

P.Mon1	P.mon1	388
-17.1322	32.8302	
-17.132	32.8302	
-17.1314	32.8297	
-17.131	32.8297	
-17.1308	32.8296	
-17.1306	32.8296	
-17.1304	32.8294	
-17.1303	32.8294	
-17.1301	32.8292	
-17.1299	32.8292	
-17.1293	32.8288	
-17.1291	32.8288	
-17.1289	32.8286	
-17.1286	32.8286	
-17.1284	32.8284	
-17.1282	32.8284	
-17.128	32.8283	
-17.1276	32.8283	
-17.1274	32.8281	
-17.1272	32.8281	
-17.127	32.828	
-17.1269	32.828	
-17.1267	32.8278	
-17.1265	32.8278	
-17.1263	32.8276	
-17.1257	32.8276	
-17.1255	32.8275	
-17.1254	32.8275	
-17.1252	32.8273	
-17.1248	32.8273	
-17.1242	32.8268	
-17.1241	32.8268	
-17.1237	32.8265	
-17.1235	32.8265	
-17.1233	32.8263	
-17.1233	32.8262	
-17.1229	32.8259	
-17.1229	32.8257	
-17.1223	32.8252	
-17.1223	32.825	
-17.1222	32.8249	
-17.122	32.8249	

-17.1214	32.8244
-17.1197	32.8244
-17.1195	32.8242
-17.1191	32.8242
-17.1189	32.8241
-17.1188	32.8241
-17.1186	32.8239
-17.1184	32.8239
-17.1178	32.8234
-17.1171	32.8234
-17.1171	32.8236
-17.1165	32.8241
-17.1152	32.8241
-17.1146	32.8236
-17.1144	32.8236
-17.1142	32.8234
-17.114	32.8234
-17.1139	32.8233
-17.1126	32.8233
-17.1124	32.8234
-17.1114	32.8234
-17.111	32.8238
-17.1108	32.8238
-17.1107	32.8239
-17.1107	32.8241
-17.1105	32.8242
-17.1105	32.8244
-17.1103	32.8246
-17.1101	32.8244
-17.1101	32.8241
-17.1099	32.8239
-17.1099	32.8238
-17.1095	32.8234
-17.1093	32.8234
-17.1092	32.8236
-17.1086	32.8236
-17.1082	32.8239
-17.1082	32.8241
-17.108	32.8242
-17.108	32.8249
-17.1078	32.825
-17.1078	32.8254
-17.1077	32.8255

-17.1075	32.8255
-17.1073	32.8254
-17.1073	32.8249
-17.1071	32.8247
-17.1071	32.8246
-17.1069	32.8244
-17.1069	32.8241
-17.1065	32.8238
-17.1054	32.8238
-17.105	32.8234
-17.1039	32.8234
-17.1037	32.8233
-17.1035	32.8233
-17.1031	32.8236
-17.1031	32.8238
-17.1026	32.8242
-17.1024	32.8242
-17.1022	32.8244
-17.102	32.8242
-17.1018	32.8242
-17.1016	32.8241
-17.1016	32.8231
-17.1014	32.8229
-17.1014	32.8228
-17.1012	32.8226
-17.1009	32.8226
-17.1003	32.8221
-17.1003	32.8218
-17.1001	32.8217
-17.1001	32.8215
-17.0999	32.8215
-17.0997	32.8213
-17.0996	32.8213
-17.0992	32.821
-17.0992	32.8208
-17.099	32.8207
-17.099	32.8202
-17.0986	32.8199
-17.0986	32.8197
-17.0984	32.8196
-17.0982	32.8196
-17.098	32.8194
-17.098	32.8183

-17.0982	32.8181
-17.0982	32.8171
-17.098	32.817
-17.0977	32.817
-17.0975	32.8168
-17.0971	32.8168
-17.0969	32.8166
-17.0962	32.8166
-17.096	32.8168
-17.0958	32.8166
-17.0958	32.8162
-17.0956	32.816
-17.0952	32.816
-17.0948	32.8157
-17.0948	32.8149
-17.0946	32.8147
-17.0943	32.8147
-17.0941	32.8145
-17.0939	32.8147
-17.0935	32.8147
-17.0931	32.815
-17.0931	32.8152
-17.093	32.8153
-17.093	32.8155
-17.0928	32.8157
-17.0928	32.8158
-17.0926	32.816
-17.0924	32.816
-17.0918	32.8155
-17.0918	32.8152
-17.0916	32.815
-17.0916	32.8149
-17.0918	32.8147
-17.0918	32.8139
-17.0915	32.8139
-17.0913	32.8137
-17.0911	32.8137
-17.0907	32.8134
-17.0907	32.8133
-17.0905	32.8131
-17.0905	32.8129
-17.0903	32.8128
-17.0901	32.8128

-17.0896	32.8123
-17.0896	32.8121
-17.0894	32.8121
-17.089	32.8118
-17.0879	32.8118
-17.0877	32.8116
-17.0875	32.8118
-17.0864	32.8118
-17.0862	32.812
-17.086	32.812
-17.0854	32.8124
-17.0852	32.8124
-17.085	32.8123
-17.0849	32.8123
-17.0847	32.8121
-17.0845	32.8123
-17.0839	32.8123
-17.0837	32.8124
-17.0835	32.8124
-17.0833	32.8126
-17.0824	32.8126
-17.0822	32.8128
-17.0816	32.8128
-17.0811	32.8123
-17.0809	32.8123
-17.0807	32.8121
-17.0805	32.8121
-17.0803	32.812
-17.0798	32.812
-17.0796	32.8118
-17.0786	32.8118
-17.0784	32.8116
-17.0773	32.8116
-17.0771	32.8115
-17.0762	32.8115
-17.076	32.8113
-17.0756	32.8113
-17.0754	32.8111
-17.0747	32.8111
-17.0743	32.8108
-17.0737	32.8108
-17.0735	32.8107
-17.0728	32.8107

-17.0726	32.8108
-17.0724	32.8108
-17.0722	32.8107
-17.0722	32.8105
-17.072	32.8105
-17.0713	32.8099
-17.0709	32.8099
-17.0707	32.8097
-17.0703	32.8097
-17.0701	32.8095
-17.07	32.8095
-17.0698	32.8097
-17.0688	32.8097
-17.0686	32.8099
-17.0681	32.8099
-17.0679	32.81
-17.0671	32.81
-17.0669	32.8099
-17.0664	32.8099
-17.0662	32.8097
-17.066	32.8097
-17.0658	32.8095
-17.0656	32.8095
-17.0654	32.8094
-17.0651	32.8094
-17.0649	32.8092
-17.0647	32.8092
-17.0643	32.8089
-17.0641	32.8089
-17.0639	32.8087
-17.0638	32.8087
-17.0636	32.8086
-17.0632	32.8086
-17.0626	32.8081
-17.0628	32.8079 *
-17.0651	32.8079
-17.0653	32.8078
-17.0909	32.8078
-17.0911	32.8076
-17.1263	32.8076
-17.1265	32.8074
-17.1269	32.8074
-17.127	32.8076

-17.1272	32.8076
-17.1274	32.8074
-17.1276	32.8074
-17.1278	32.8076
-17.128	32.8076
-17.1282	32.8074
-17.1284	32.8074
-17.1286	32.8076
-17.1288	32.8076
-17.1289	32.8074
-17.1291	32.8076
-17.1293	32.8076
-17.1295	32.8074
-17.1365	32.8074
-17.1367	32.8076
-17.1384	32.8076
-17.1385	32.8074
-17.1495	32.8074 *
-17.1499	32.8078
-17.1499	32.8095
-17.15	32.8097
-17.1499	32.8099
-17.1499	32.8107
-17.15	32.8108
-17.1499	32.811
-17.15	32.8111
-17.15	32.8113
-17.1499	32.8115
-17.15	32.8116
-17.1499	32.8118
-17.15	32.812
-17.15	32.8145
-17.1499	32.8147
-17.1499	32.8158
-17.15	32.816
-17.15	32.8181
-17.1502	32.8183
-17.1502	32.8191
-17.15	32.8192
-17.15	32.8452
-17.1502	32.8454
-17.15	32.8456
-17.1502	32.8457

-17.1502	32.8461
-17.15	32.8462
-17.15	32.8478
-17.1502	32.848
-17.1502	32.8482
-17.15	32.8483
-17.15	32.8514
-17.1499	32.8515 *
-17.1493	32.8515
-17.1491	32.8514
-17.1489	32.8514
-17.1476	32.8503
-17.1476	32.8501
-17.1474	32.8499
-17.1474	32.8496
-17.1472	32.8494
-17.1472	32.8469
-17.1474	32.8467
-17.1474	32.8462
-17.1476	32.8461
-17.1476	32.8456
-17.1478	32.8454
-17.1478	32.8449
-17.1476	32.8448
-17.1476	32.8444
-17.1474	32.8443
-17.1474	32.8441
-17.1472	32.8441
-17.1468	32.8438
-17.1468	32.8436
-17.1466	32.8435
-17.1466	32.8431
-17.1463	32.8428
-17.1459	32.8428
-17.1451	32.8422
-17.145	32.8422
-17.1446	32.8419
-17.1444	32.8419
-17.144	32.8415
-17.1438	32.8415
-17.1436	32.8414
-17.1433	32.8414
-17.1431	32.8412

-17.1429	32.8412
-17.1427	32.841
-17.1423	32.841
-17.1421	32.8409
-17.1418	32.8409
-17.1416	32.8407
-17.1414	32.8407
-17.1412	32.8406
-17.1404	32.8406
-17.1401	32.8402
-17.1399	32.8402
-17.1393	32.8398
-17.1391	32.8398
-17.1384	32.8391
-17.1382	32.8391
-17.1374	32.8385
-17.1374	32.8375
-17.1372	32.8373
-17.1372	32.8372
-17.137	32.837
-17.137	32.8365
-17.1372	32.8364
-17.1372	32.8362
-17.1374	32.836
-17.1374	32.8352
-17.1372	32.8351
-17.1372	32.8346
-17.137	32.8344
-17.137	32.8343
-17.1367	32.8339
-17.1367	32.8338
-17.1363	32.8335
-17.1363	32.8331
-17.1357	32.8326
-17.1357	32.8325
-17.1355	32.8325
-17.1353	32.8323
-17.135	32.8323
-17.1348	32.8322
-17.1344	32.8322
-17.1335	32.8313
-17.1335	32.8312
-17.1325	32.8304

17.1323	32.8304
17.1322	32.8302

APENDICE IV

RESULTADOS PRÉ-DEFINITIVOS DOS CENSOS DE 1991

	0-14			15-24			25-64			65 ou mais			Total
	H+M	H	M	H+M	H	M	H+M	H	M	H+M	H	M	
R.A.M. Total	62002	31532	30470	46800	23000	23800	115206	51596	63610	29419	11417	18002	253427
Calheta	3092	1518	1574	1980	872	1108	5462	2145	3317	2471	977	1494	13005
C. Lobos	9997	5068	4929	6853	3371	3482	12362	5441	6921	2265	908	1357	31477
Funchal	26197	13203	12994	20640	10262	10378	55648	25567	30081	12918	4784	8134	115403
Machico	5551	2888	2663	4783	2416	2367	9571	4450	5121	2111	852	1259	22061
P. Sol	2072	1027	1045	1567	713	854	3707	1525	2182	1410	547	863	8756
P. Moniz	858	432	426	482	214	268	1519	589	930	573	241	332	3432
P. Santo	1211	626	585	833	431	402	2289	1139	1150	373	175	198	4706
R. Brava	3165	1598	1567	2462	1148	1314	5610	2201	3409	1933	766	1167	13170
St. Cruz	5679	2971	2708	4198	2121	2077	10748	4957	5791	2840	1103	1737	23465
Santana	2404	1272	1132	1645	785	860	4781	2092	2689	1472	633	839	10302
S. Vicente	1776	929	874	1357	667	690	3509	1490	2019	1053	431	622	7695