XIV Congresso Anual da Sociedade Portuguesa de Estatística

Covilhã, 27 a 30 de Setembro de 2006

Outliers em Dados Estatísticos

Fernando Rosado

Edições SPE

FICHA TÉCNICA:

Título: Outliers em Dados Estatísticos

Autor: Fernando Rosado

Editora: Sociedade Portuguesa de Estatística

Concepção gráfica da capa: lupim

Produção gráfica e Impressão: Instituto Nacional de Estatística

Tiragem: 450 exemplares

ISBN: 972-8890-07-9

Depósito legal: nº 244656/06

Edições SPE

Manuais

✓Introdução à Probabilidade e à Estatística - com complementos de Excel por Maria Eugénia Graça Martins

Mini Cursos

✓Tópicos de Sondagens por Paulo Gomes
✓Controlo Estatístico de Qualidade por Ivette Gomes e Isabel Barão
✓Modelos Lineares Generalizados por Antónia Turkman e Giovani Silva
✓Inferência sobre Localização e Escala por Fátima Brilhante, Dinis Pestana, José Rocha e Sílvio Velosa
✓Modelos Heterocedásticos. Aplicações com o software Eviews por Daniel Muller
✓Séries Temporais - Modelações Lineares e Não Lineares por Esmeralda Gonçalves e Nazaré Mendes Lopes
✓Uma Introdução à Análise de Clusters por João Branco
✓Introdução às Equações Diferenciais Estocásticas e Aplicações por Carlos Braumann
✓Outliers em Dados Estatísticos por Fernando Rosado

Actas

✓Afirmar a Estatística. Um Desafio para o Século XXI - Actas do VI Congresso Anual da SPE. C. Paulino, A. Pacheco, A. Pires e F. da Cunha (Ed.)
✓Um Olhar sobre a Estatística - Actas do VII Congresso Anual da SPE. P. Oliveira e E. Athayde (Ed.)
✓A Estatística em Movimento - Actas do VIII Congresso Anual da SPE. M. M. Neves, J. Cadima, M. J. Martins e F. Rosado (Ed.)
✓Novos Rumos em Estatística - Actas do IX Congresso Anual da SPE. L. Carvalho, F. Brilhante e F. Rosado (Ed.)
✓Literacia e Estatística - Actas do X Congresso Anual da SPE. P. Brito, A. Figueiredo, F. Sousa, P. Teles e F. Rosado (Ed.)
✓Estatística com Acaso e Necessidade - Actas do XI Congresso Anual da SPE. P. Rodrigues, E. Rebelo, e F. Rosado (Ed.)
✓Estatística Jubilar - Actas do XII Congresso Anual da SPE. C. A. Braumann, P. Infante, M. M. Oliveira, R. Alpizar-Jara e F. Rosado (Ed.)

História da Estatística

✓Memorial da Sociedade Portuguesa de Estatística. F. Rosado (Ed.)

À Lígia e
aos nossos netos
Carolina, Margarida e Santiago

Prefácio

Com esta edição SPE, *Outliers em Dados Estatísticos*, deseja-se uma ampla divulgação desta importante temática da Ciência Estatística junto da comunidade científica dos estatísticos portugueses. Ela tem, também, o objectivo de servir de apoio ao Curso, com o mesmo título, que será ministrado no XIV Congresso Anual da Sociedade Portuguesa de Estatística.

Neste livro vamos (tentar) falar de outliers e "outliers" - os primeiros numa "perspectiva da teoria tradicional" e os segundos numa visão de "terminologia em português". Esta engloba uma grande parte da contribuição portuguesa para o estudo de observações discordantes.

O "problema outlier" sempre fascinou todos aqueles que trabalham com (e têm de interpretar) dados. Até ao presente, sem grande polémica, podemos afirmar que a principal questão na teoria dos outliers está directamente relacionada com a própria definição de observação discordante. Revisitaremos este tema.

Os valores discordantes numa amostra são sempre objecto de estudo muito prático. Tradicionalmente, o analista selecciona e, em seguida, testa os dados estatísticos que, por alguma razão, lhe parecem suspeitos.

Fundamentalmente, baseado num modelo, ele necessita de um teste de discordância "bom" para o estudo de tais observações. Faremos o ponto da situação sobre os modelos para a detecção de outliers e os respectivos testes de discordância. Apresentaremos metodologias gerais para o estudo de observações discordantes e faremos aplicações a modelos gerados pelas populações mais comuns - nomeadamente, exponenciais e normais.

Um campo de trabalho que necessita de avanço envolve as medidas

de performance daqueles testes para estudo de observações discordantes. Abordaremos este tema, com os mais recentes resultados.

Novas perspectivas, quer no campo dos dados direccionais quer no âmbito da estatística espacial e ambiental, vêm colocar novas questões metodológicas à teoria dos outliers, nomeadamente a implementação de "outros" modelos de discordância com hipóteses envolvendo aqueles dados multivariados e populações até agora (ainda) pouco consideradas.

Aquele "problema outlier", em dados estatísticos, tem algumas semelhanças com as diarreias dos viajantes. Existem muitas designações para a sintomatologia - "rogue", "maverick" ou "spuriosities", por exemplo, em língua inglesa; discordantes, aberrantes ou atípicas e outras, em língua portuguesa. O número de remédios é tão grande que alguns podem ser suspeitos de serem ineficazes. Além disso, é muitas vezes com a mezinha popular que "a doença se cura".

Para o estudo de outliers em dados estatísticos, com este livro desejamos fornecer um método de diagnóstico e uma terapia.

Neste texto levantamos questões e oferecemos algumas soluções; ao mesmo tempo, encontramos indicações e possíveis caminhos para a procura de outras.

Mais do que em muitas outras áreas científicas, o estudo de outliers em dados estatísticos é um campo específico onde se exige a máxima informação dos valores da amostra. A existência daqueles inibe que todos os observados tenham igual valor. A excelência dos resultados depende, em parte, das estatísticas ordinais. A ordenação é um tema primordial no tratamento de dados discordantes. Como ordená-los? Quais são os mais importantes? Uma amostra é um todo com n observações; no entanto, ela é como uma orquestra. Só com primeiros violinos não se cria música. Cada obra acabada tem os seus primeiros; também assim numa amostra. No capítulo 2 reflectiremos sobre a importância dos outliers numa estatística de excelência.

Na passada década de oitenta dactilografei uma dissertação; com as facilidades oferecidas por uma máquina de escrever IBM - eléctrica, com a opção entre os diversos (dois ou três) tipos de letras e de símbolos podendo ser feita através do uso de cabeças permutáveis e que possuía uma inovadora tecla que permitia corrigir, imediatamente, qualquer erro descoberto antes de ser retirada a respectiva página. Com o avanço da informática surgiram, logo a seguir, os processadores de texto de que me tornei um habitual utilizador. Estes tornavam incomparavelmente mais simples a tarefa da escrita, muito em especial a matemática. Durante

vinte anos "pratiquei" Word com pontuais incursões em Latex para alguns artigos científicos. Perante o desafio da presente edição e com o precioso conselho de experimentados colegas, optei por uma "especialização" em Latex. Novos desafios para a arte de Gutenberg - um outlier do segundo milénio. Terminado o trabalho, fica a certeza da boa opção feita! Aos meus conselheiros deixo um grande agradecimento.

Este livro decorre no âmbito das actividades de investigação que tenho desenvolvido no CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa, financiado pela FCT - Fundação para a Ciência e a Tecnologia, entidade a quem agradecemos o apoio.

No conteúdo científico desta obra está incluída informação que, ao longo dos anos, resultou da contribuição de estudantes de mestrado e de doutoramento. Alguns assuntos atravessaram reuniões científicas a diferentes níveis onde o debate *inter pares*, decerto, muito os beneficiou. Para todos, são devidos os reconhecidos agradecimentos!

Com a realização do referido curso - sobre Outliers em Dados Estatísticos - sou devedor de um profundo agradecimento por, em ambiente de assembleia magna dos estatísticos portugueses, ter a oportunidade de divulgação e partilha científica desta relevante área do conhecimento.

No enquadramento definido para esta obra, todas as grandes questões sobre outliers foram abordadas. No entanto, como se explicará no texto, fica a certeza de que alguns campos específicos mereciam mais desenvolvimento. Para a maior parte, serão indicadas referências para uma eventual prossecução do estudo. Algumas daquelas interrogações, de facto, correspondem a temas em aberto; portanto, são potenciais e excelentes caminhos para investigação.

Sem dúvida que, tanto nas gralhas como nas incorrecções, este livro não é um outlier.

O autor agradece todas as críticas e indicação de erros - enfim, tudo o que possa melhorar esta edição.

Fernando Rosado

Universidade de Lisboa, Junho de 2006

Índice

Capítulo 1

Introdução

1.1 Preliminares

A história recente do estudo de observações discordantes numa amostra inicia-se em meados do século passado... há cinquenta anos!

Alguns artigos, de sistematização e síntese, reveladores do despertar da ciência estatística para a problemática das observações aberrantes foram publicados nesse tempo. No entanto, a sensibilidade dos Estatísticos para a problemática dos "dados estranhos", como veremos, começou muito antes.

Ao longo do tempo, muitos trabalhos científicos têm sido publicados e, como é de esperar, as primeiras publicações apareceram na sequência da necessidade de resolver problemas. E, algumas vezes, surgiram apenas como relatos de sucessos em uma (ou para uma) "análise de dados". Após algumas abordagens *ad hoc* assistiu-se a uma consolidação de métodos.

Uma obra, de 1978, com referência obrigatória, é a primeira edição de *Outliers in Statistical Data* da autoria de Vic Barnett e Toby Lewis, dois pioneiros na divulgação da teoria dos outliers. É um tratado cuja terceira edição [11] foi publicada em 1994 e que ainda mantém o primeiro lugar no pódio das citações em trabalhos científicos na área científica que se preocupa com os dados estatísticos difíceis.

O estudo de outliers (também) é uma componente fundamental na análise de dados.

Uma única observação (não detectada) pode destruir ou contrariar a conclusão de qualquer trabalho. Por isso, os analistas preocupam-se com a problemática dos mecanismos de geração dos valores aberrantes para uma amostra. Como especialistas na matéria, os estatísticos são então

solicitados, para encontrar soluções.

Intuitivamente, um outlier é um dado estatístico (tão) discordante da maioria dos restantes que se torna suspeito. A discordância é a principal motivação para o estudo.

Em relação ao seu grupo, qual deve ser a "distância" que permite medir a aberração de um dado estatístico? Por outras palavras: O que é uma observação suspeita?

No fundamental, o estudo de outliers deve sempre considerar que o valor discordante não é um erro grosseiro, isto é, a análise é desenvolvida na presunção de que os dados em presença estão "completamente controlados" de tal modo que eventuais erros de escrita ou outros não possam subsistir. Deve pois, assumir-se que apenas o acaso é responsabilizado pela geração dos dados. Num verdadeiro estudo de outliers estamos portanto em presença de dados estatísticos eventualmente produzidos por diferentes mecanismos aleatórios.

1.2 Alguns Exemplos

Estatística, numa "definição - síntese de dicionário" surge[1] como o ramo das matemáticas aplicadas que recorre ao cálculo das probabilidades para estabelecer hipóteses com base em acontecimentos reais, com o fim de fazer previsões ou ainda, a ciência que tem por objecto o agrupamento metódico dos factos sociais que se prestam a uma avaliação numérica (população, natalidade, mortalidade, rendimento de impostos, produções agrícolas, criminalidade, religião, etc.). Entre estes parêntesis está o prenúncio de um vasto mundo de trabalho para o estatístico (ou estatista como agora já se vai dizendo...). Esse mundo, como se vê, engloba "diversos mundos" de outras ciências.

Assim, a inter e a multidisciplinaridade é (também) uma exigência - e um desafio - na Estatística. Aquela porque evoca um espaço comum, um factor de coesão entre diferentes saberes onde, na maioria das vezes, o trabalho do estatístico se enquadra e esta porque justapõe disciplinas diversas, às vezes sem relação aparente entre elas. O estatístico fica portanto com a responsabilidade de resolver problemas, estabelecendo pontes entre diferentes mundos científicos.

Neste ambiente introdutório para a teoria dos outliers em dados estatísticos devemos considerar algumas situações concretas que, por um lado, ajudam a clarificar definições, critérios e objectivos e, por outro, es-

[1] A jeito das novas tecnologias, por exemplo, no Dicionário da Língua Portuguesa on-line (© 2006 Priberam Informática) em http://www.priberam.pt/dlpo/dlpo.aspx. No capítulo 2, retomaremos esta problemática das definições.

tabelecem limites de actuação, incluindo os mais recentes campos onde[2] podemos actuar como, por exemplo, na estatística forense[3].

Na literatura da especialidade existe uma grande variedade de exemplos, para explicar as possíveis causas de observações discordantes, para mostrar diferentes caminhos de tratamento para dados aberrantes, para "encontrar outliers" - como os que são apresentados nos manuais dos pacotes estatísticos - ou para mostrar possíveis modelos dos mecanismos de geração dos valores suspeitos.

Vamos usar alguns desses exemplos e criar outros ao longo do nosso texto. Com a selecção feita, desejamos sensibilizar o leitor para a problemática da análise e do estudo de observações discordantes numa amostra.

Levantaremos questões para reflexão bem como, nalguns casos, propostas de estudo continuado desses exemplos.

Para já, neste capítulo, mais do que respostas vamos encontrar perguntas.

Exemplo I:

Consideremos[4] *um estudo onde temos dados de temperaturas registadas em cada hora. Em Wick, no norte da Escócia, na passagem do ano 1960 foram registadas:*

$$43, 43, 41, 41, 42, 43, 58, 58, 41, 41.$$

Os dois valores 58 correspondem à meia noite de 31 de Dezembro de 1960 e à uma hora do dia 1 de Janeiro de 1961. São valores aberrantes quando comparados com os restantes. Se, como informação suplementar, acrescentarmos que, nessa passagem de ano, os serviços meteorológicos da zona, alteraram a metodologia graus Fahrenheit para décimos de grau Celsius, então os "verdadeiros dados" - se tudo se mantivesse como dantes - são:

$$43, 43, 41, 41, 42, 43, 42, 42, 39, 39.$$

O problema está resolvido!? Registadas, ficam algumas possíveis causas para o aparecimento de valores estranhos - erro humano ou ignorância.

[2]Alguns destes exemplos, independentemente de outros que entretanto sejam incluídos, vão ser desenvolvidos e estudados ao longo deste livro.

[3]Este talvez seja o tema mais recente para o estudo de outliers em dados estatísticos.

[4]Cf. Barnett e Lewis [10].

Outros exemplos, similares, podem ser vistos em Finney [48] ou em Barnett [5] e, obviamente, também - no tratado fundamental *Outliers in Statistical Data* - em Barnett e Lewis [11].

E ainda alguns desafios suplementares; pois o que continua a constar dos livros de registos das temperaturas são os dados acima e não estes últimos.

Exemplo II:

Consideremos o seguinte conjunto "académico" de dados estatísticos:

$$2, 2.8, 3.4.$$

Estes dados contêm algum valor discordante?

Estes dados, admitidos gerados por uma população normal, contêm algum valor discordante? A primeira questão levantada, apenas é válida quando os dados são "compatíveis" com as distribuições assumidas?

Assumida a hipótese de normalidade, abordaremos de novo este caso nos capítulos 7 e 9.

Exemplo III:

Na sequência da reflexão, proposta no exemplo anterior, consideremos o seguinte conjunto de dados estatísticos:

$$2, 2.8, 2.8, 3.4.$$

Estes dados, admitidos gerados por uma população normal, contêm algum valor discordante? Todas as perguntas já formuladas adquirem, com este exemplo, uma nova perspectiva ou dificuldade?

As diversas questões "introduzidas" nestes e nos casos seguintes - através de perguntas difíceis? - serão desenvolvidas e aprofundadas ao longo do texto[5] quando voltarmos a estes exemplos. Tentaremos, então, encontrar (algumas) respostas.

Um outlier , principalmente do ponto de vista prático, é muitas vezes caracterizado (e definido!) pelo posicionamento que ocupa numa caixa-com-bigodes. Essa apresentação gráfica, subjectiva e dependente de coeficientes a introduzir pelo utilizador é, muitas vezes, o único instrumento de análise de outliers. Muitas referências (às vezes contraditórias nas conclusões!) podemos citar. Pretendemos levantar esta questão que, como veremos, se prende também com a própria definição de outlier. Analisemos os seguintes casos.

[5]Para facilitar a pesquisa pode consultar-se o índice remissivo para localizar as páginas onde cada um destes exemplos será continuado.

Exemplo IV: Sobre a definição de outlier (I)

Suponhamos que estamos em presença de um output de uma "caixa-com-bigodes". Para ajudar na interpretação, o utilizador é remetido para o manual de "apoio", de onde se deseja que ela surja.

Por exemplo[6]*:"An* ***outlier*** *(o) is* ***defined*** *as a value more than 1.5 box-lengths away from the box;* ***an extreme*** *(*) as more than 3 box-lengths away from the box." Do ponto de vista prático fica tudo tornado muito simples com "os criminosos" referenciados por (o) e (*). Mas... quando o objectivo é apresentar soluções, quantas questões*[7] *são aqui deixadas em aberto e para as quais não é proposta qualquer resolução ou, pelo menos, a indicação de que "a coisa" não é assim tão simples e com a indicação de alguma leitura para os mais interessados?*

No contexto do exemplo anterior, podemos encontrar nalguns manuais, algumas instruções de apoio para a pesquisa de observações discordantes.

Exemplo V: Sobre a definição de outlier (II)

Citamos:[8]

"Finding Outliers - You will find the BASIC programming facility useful for finding cases with outliers or missing values after you run a statistical procedure. For example, you can trim tails by using the statistical procedures to locate lower and upper cut points and then running something like this:

LET LOWCUT = -3.1
LET HICUT = 3.2
IF X<LOWCUT OR X>HICUT THEN DELETE"

Neste exemplo, novas dificuldades ficam registadas. E, a principal é uma das fundamentais num estudo de outliers - a subjectividade.

Para o utilizador menos esclarecido, (decerto!) algumas destas propostas podem ser contra producentes.

Exemplo VI: Sobre a definição de outlier (III)

Numa primeira abordagem científica, um outlier deve ser um extremo que surpreende. Teremos oportunidade de clarificar melhor esta afirmação que, de acordo com o estudo que faremos no capítulo 4, não está

[6]Cf. SPSS for Windows Made Simple (Release 10) por Paul Kinnear e Colin Gray. Edição de 2000. Psychology Press.

[7]Na secção 2.8.3, desenvolvidamente, abordaremos de novo estes temas.

[8]Cf. Manual do package SYSTAT - The System for Statistcs for he Macintosh (versão 3), 1987, Leland Wilkinson, módulo Data, p. DATA-42. 1985, Systat, Inc.

certa. Ela regista uma visão tradicional do estudo de outliers em dados estatísticos, mas que (resume e) é um excelente ponto da situação.[9]

Neste ambiente introdutório e ainda no contexto dos dois exemplos anteriores, para além da subjectividade nos critérios - o que por si só, não será um grande mal - podemos verificar a confusão na definição de observação discordante que na maior parte das vezes é confundida com outlier e onde, além disso, se aumenta a entropia quando se induz que "para além dos outliers" ainda podemos encontrar "os extremos" - como vimos, no exemplo IV. É fundamental clarificar algumas das imprecisões anteriores.

Numa amostra, a amplitude do intervalo de variação, sendo uma diferença entre os extremos, é uma medida de dispersão e pode ser usada para referenciar observações com um comportamento estatístico não esperado mas, não é resistente. Assim, (como alternativa?) deve usar-se a amplitude inter-quartílica, entre cujos extremos devem estar incluídas 50 por cento das observações, como uma medida de discordância. É uma antiga proposta[10] *que permite determinar (a ou as) observações que ultrapassam esses limites e que pode ser usada para "definir outlier". Alguns autores*[11] usam esses dois pontos e definem então "outlier moderado" e "outlier severo".

Um estudo de outliers em dados estatísticos, podemos afirmar desde já, apenas é simples se a nossa análise ficar pela superficial procura e/ou detecção de um valor suspeito, ou - como acontece na maior parte das vezes - se apenas desejamos determinar uma justificação para qualificar como aberrante uma observação que, à partida, se tornou suspeita. O estatístico tem de ir mais além quer na análise quer na obrigação de divulgação de estudos. O aprofundamento, logo numa primeira fase, passa pela procura de um mecanismo de geração dos dados. Numa segunda fase deve proceder a análises complementares e algumas delas remetem, por exemplo, para estudos de robustez. Além disso, é fundamental a divulgação dos resultados. De novo, surge a referência aos tratados de Barnett e Lewis [11] que, desde a primeira edição em 1978, por todos são reconhecidos, como grandes impulsionadores e pioneiros na divulgação e, por consequência, no aprofundamento dos estudos de observações discordantes numa amostra - desde a mais simples colecção de dados até

[9] "An outlier is not only an extreme, but is surprisingly extreme". (Cf. Barnett e Lewis [11], p. 15).

[10] Na secção 2.8.3 retomaremos esta questão.

[11] Cf. por exemplo, Murteira *et al* ([92] p. 33 e segs) ou Pestana e Velosa ([101], p. 104-6). Estes autores, aproveitando exercícios práticos que propõem sobre este assunto, também apresentam algumas reflexões muito oportunas sobre a identificação de observações discordantes bem como a sua interpretação.

aos mais elaborados modelos onde a detecção de outliers se pode tornar extremamente difícil (senão impossível). Atente-se no exemplo seguinte; onde também se pretende salientar como os diferentes caminhos para o estudo de outliers podem influenciar as propriedades e os próprios estimadores para os diferentes parâmetros dos modelos envolvidos.

Exemplo VII: Modelo de Discordância

Admitindo que estamos a trabalhar em modelos Normais, consideremos[12] *a "pequena amostra" seguinte:*

$$1.74, 1.46, -0.28, -0.02, -0.40, 0.02, 3.89, 1.35, -1.10, 0.71$$

O valor 3.89 "evidencia-se" dos restantes e não temos explicação para o seu aparecimento. Sendo um extremo, em que sentido podemos considerá-lo como outlier? Se os dados forem gerados por uma distribuição normal $N(0,1)$*, aquele dado é "subjectivamente" surpreendente e também estatisticamente pouco provável. Com o auxílio de um teste podemos estudar o afastamento desse valor em relação à média. O que devemos concluir? A opção na decisão depende da distribuição alternativa que admitirmos para gerar esse valor.*

Podemos concluir que um outlier, além de ser uma manifestação aleatória não pode ser interpretado como um valor determinístico. A condição de discordante (até este momento!) é completamente subjectiva. Para uma necessária metodologia científica torna-se fundamental um mecanismo de geração - um modelo de discordância - que explique os valores em presença.

Neste exemplo, entre outras, também levantamos as questões que envolvem a dimensão e a sua relação com a problemática das "pequenas" e das "grandes" amostras. O extremo que naquela surpreendeu veria diluída essa condição se a amostra tivesse uma dimensão muito maior?

Estes temas, como é obvio, estão também incluídos em cada um dos restantes exemplos e carecem de algum aprofundamento.

Mais para diante, tentaremos!

Exemplo VIII: Teste de Discordância/Critério de Chauvenet

Consideremos o, bem conhecido, conjunto de dados analisados[13] *por*

[12] Este exemplo é retirado de Barnett e Lewis ([11], p. 15).

[13] Deve registar-se, desde logo, a partir dos títulos dos respectivos trabalhos, o pioneirismo e a sensibilidade dos autores para os problemas específicos de um estudo de outliers que, só muito mais tarde, viriam a ser sistematizados. O valioso critério de Chauvenet, ainda hoje, é utilizado em muitos exemplos práticos, principalmente

Peirce [98] e Chauvenet [105] e que é constituído por 15 observações de medições astronómicas específicas do planeta Vénus.

Adaptado um modelo, foram estabelecidos os seguintes resíduos:

$$\begin{array}{lllll} -0.30, & +0.48, & +0.63, & -0.22, & +0.18 \\ -0.44, & -0.24, & -0.13, & -0.05, & +0.39 \\ +1.01, & +0.06, & -1.40, & +0.20, & +0.10 \end{array}$$

Se admitirmos um mecanismo normal para a geração destes valores, o mínimo -1.40 torna-se suspeito? Um teste de discordância pode ser produzido considerando o "desvio" do mínimo $x_{(1)}$

$$d = \frac{\bar{x} - x_{(1)}}{s}$$

ou o "quociente entre afastamentos" de estatísticas ordinais

$$q = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$$

para aquela observação.

A avaliação de observações discordantes feita através do seu afastamento de, pelo menos, um desvio padrão em relação à média, costuma designar-se por critério de Chauvenet.

Estamos então introduzindo mecanismos de identificação de outliers. De entre os mais importantes salientam-se os testes de discordância que, com mais detalhe veremos no capítulo 2, são versões estatísticas avançadas deste critério centenário.

Em ambos os casos aquele valor não é discordante ao nível de significância 0.05.

No entanto, nos referidos estudos, Peirce e Chauvenet consideraram que o valor -1.40 deve ser rejeitado.

Exemplo IX: Outlier em dados estruturados

Consideremos o seguinte conjunto de dados[14] *onde x representa a carga (em Kg) aplicada a estruturas semelhantes e y a correspondente*

em estudos "menos aprofundados". Sobre este critério, veja-se ainda a secção 9.1, a propósito da influência da dimensão da amostra.

[14] Estes dados foram retirados de um exemplo, para um estudo análogo, apresentado por Barnett e Lewis (Cf. [11], p. 316).

deformação (em centímetros):

x :	11.2	21.1	29.9	34.1	43.8	53.4	59.9	61.2	68.9
y :	1.6	2.1	3.4	3.3	4.2	3.1	4.9	6.2	6.3

A relação entre carga e deformação pode ser analisada com um diagrama de dispersão (x,y).

Aí veremos que, também em dados estruturados, não há razão para considerarmos como potenciais outliers apenas as observações extremas[15] *da amostra quer nos valores de x quer nos valores de y.*

Nessa análise gráfica, decerto, a observação (53.4, 3.1) não sendo "extrema", desde o início, chama a nossa atenção.

Este exemplo pode ainda ser utilizado no sentido de conjugar um estudo de outliers com aqueloutro igualmente importante de observações influentes no modelo.

Esta abordagem pode ser vista em Barnett e Lewis ([11], p. 317). Em particular podemos utilizar este mesmo conjunto de dados para um estudo do efeito[16] *"ballooning" (de alargamento) muito importante porque envolve a tendência para determinadas observações terem uma grande influência na redução da variância dos estimadores o que reforça a perspectiva acima invocada de potenciais outliers "dentro" dos restantes valores.*

Nos capítulos 7 e 9 continuaremos este exemplo.

Neste capítulo de introdução, prossigamos com um derradeiro exemplo - que se desdobrará em dois - e que também desenvolveremos ao longo do nosso texto.

Trata-se de matéria muito importante quer pelos casos que envolve quer pela inovação que apresenta para o estudo de outliers em dados estatísticos. É baseado numa abordagem introduzida por Barnett [5] e por Barnett e Lewis [11] onde são relatados alguns casos legais, tendo o primeiro ocorrido em 1949.

É uma situação que consideramos fundamental (também!) pela síntese que o seu estudo permite.

Além disso, este exemplo introduz uma nova perspectiva do uso da teoria dos outliers na estatística forense e da importância da reflexão que, através dele, pode ser desenvolvida.

[15] Para estes modelos estruturados, algumas vezes são introduzidos os conceitos de *outlier em x* ou de *outlier em y*, pretendendo distinguir as duas componentes de cada dado estatístico e a sua contribuição para a discordância no modelo estruturado.

[16] Com mais desenvolvimento em 9.1.

Exemplo X: Estatística forense (I)

O Senhor Hadlum, cidadão britânico, foi mobilizado para prestar serviço militar fora do seu país. Em 12 de Agosto de 1945, a Senhora Hadlum deu à luz uma criança - 349 dias depois do Senhor Hadlum ter partido. O Senhor Hadlum apresentou uma queixa reclamando divórcio com base em adultério. Em primeira diligência o tribunal não lhe deu razão e ele decidiu recorrer da sentença.

Este é o ponto de partida para um excelente exemplo de "um problema outlier" que, sendo também historicamente de referência, exige e motiva uma forte reflexão sobre a problemática das observações discordantes em estatística.

A duração média da gestação humana é 280 dias.

Neste caso, chamemos-lhe H-H, na sequência do recurso, a principal argumentação está centrada no facto de que 349 comparado com 280 deve ser declarado discordante - ou será outlier? Qual a diferença?

Os juízes concordaram que "um limite de credibilidade" deve ser exigido mas, à luz da ciência médica, embora improvável, aquele valor era possível.

O recurso, apresentado pelo Senhor Hadlum, falhou.

Também em 1949 - em outro julgamento, no caso M-T - o tribunal admitiu que 340 dias era impossível de acordo com a "experiência ginecológica".

Em 1951, no caso P-J, foi definido "o limite". O valor de referência para os tribunais ingleses passou a ser 360.

No nosso estudo, vamos fundamentalmente usar o caso H-H e, em resumo, a principal questão de estatística forense prende-se com a decisão sobre o "dado estatístico" 349, como relatando a duração de uma gestação humana.

O valor 349, quando comparado com 280, deve ser declarado outlier?

Saliente-se que, no relato histórico apresentado, não existe qualquer referência à distribuição da variável em estudo - a normal.

Exemplo XI: Estatística forense (II)

Além de tudo, "em números", o exemplo anterior pode ser alterado se usarmos as semanas como dados. Assim, o caso H-H, exige que se decida se 50 é surpreendente quando comparado com 40. Podemos então reflectir se "349 versus 340" é equivalente a "50 versus 40" ou melhor, se a escala de registo dos dados tem alguma influência na suspeição que o experimentador (o colectivo de juízes?) associa a cada dado o que acrescenta subjectividade à decisão. O período de gestação é um extremo. Pode ser considerado um outlier, mas esta decisão não resolve todo o

problema pois "toda a amostra tem um extremo". Aquele valor pode ser uma observação verdadeira de uma outra distribuição e, sendo assim, ficava determinado o mecanismo de aparecimento desse dado estatístico. O tribunal considerou, apenas, que era um extremo surpreendente. Mas também pode ser um contaminante, isto é, uma observação pertencente a outra distribuição, por exemplo com uma translação na origem da gestação.

Voltaremos ao assunto.

Nesta introdução, concluiremos que numa análise de dados estatísticos podemos ter extremos que podem, ou não, ser outliers e que por sua vez estes, poderão, eventualmente, ser contaminantes.

1.3 Extremos, Outliers e Contaminantes

Admitamos $x_1, ..., x_n$ uma amostra, univariada de dimensão n, recolhida numa população onde admitimos um mecanismo aleatório de geração dos dados através de uma distribuição F.

Seja $x_{(1)}, ..., x_{(n)}$ aquela amostra, ordenada.

Tradicionalmente, no estudo de observações discordantes, num primeiro (e único?!) passo o analista julga (e decide!) se o mínimo $x_{(1)}$ ou o máximo $x_{(n)}$ da amostra devem ser declarados outlier. A escolha dessa eventual observação discordante é, portanto, subjectiva. No entanto, essa escolha não é arbritrária. Ela depende, em primeira análise, da distribuição F.

Na teoria, algumas vezes, distingue-se um outlier - aquela observação que se "afasta" das restantes - de um contaminante - que não é gerado pelo mesmo mecanismo. Registe-se a subjectividade, sempre envolvida no estudo, quando este é feito, apenas, através de um "critério de afastamento". Assim, um outlier pode considerar-se uma observação que "fica de fora" dos dados. Mas, essa observação pode "ficar dentro" e ser gerada por um mecanismo diferente do das restantes.

Então, um outlier pode "ficar dentro" dos dados mas ser discordante - será, então, contaminante.

Numa amostra, extremos existem sempre. Outlier "pode ser" uma observação (ou um conjunto de observações) que parece(m) ser inconsistente(s) com as restantes. Por sua vez, os contaminantes "vêm de" ou são "gerados por" outra distribuição. Então, neste contexto, um dado estatístico é outlier não porque "fica de fora" nos dados mas, porque "fica de fora" no modelo assumido para esses dados.

Um contaminante pode parecer uma observação genuína enquanto que um outlier pode ser, mas não é necessariamente, um contaminante. Nos exemplos anteriores isto já se verificou.

Por outro lado, os outliers podem ser contaminantes gerados por um mecanismo estatístico diferente. Se as sementes usadas para uma experimentação de crescimento de uma planta contiverem algumas sementes "diferentes" então as plantas por elas geradas serão contaminantes e podem ser outliers.

Obviamente, como já vimos, uns e outros também podem ser erros grosseiros!

Pode usar-se um modelo errado, por exemplo, admitindo que os dados produzidos em determinado aparelho são normais quando, de facto, são Cauchy. Assim, poderão surgir alguns valores que perturbam a análise. Podem também surgir valores discordantes em qualquer análise estruturada de dados como por exemplo em análise de regressão ou de variância. A heterocedasticidade pode gerar dados perturbadores do estudo. A perturbação na variância é um caso muito importante no estudo de outliers e que será, aprofundadamente analisado nos capítulos 4 e 7.

Uma outra questão muito importante é a identificação de outliers. Esta tarefa pode ser mais ou menos difícil. Eventualmente, torna-se mais complicada à medida que se aprofunda a estruturação dos modelos. Podemos concluir que, o estudo de outliers em dados estatísticos de uma amostra de uma população univariada, pode-se considerar uma tarefa científica menos complexa que a correspondente análise para dados relativos a, por exemplo, uma série temporal. Obviamente, a dimensão também é um factor de dificuldade.

O estudo de outliers em dados multivariados é mais complexo, desde logo porque a ordenação, que é basilar na discordância, é uma característica das populações univariadas que se perde com a dimensão da população.

A estimação de parâmetros de um modelo sem risco de grave perturbação pela presença de outliers ou de contaminantes é um problema de robustez. O exemplo mais familiar é a caixa-com-bigodes com as suas regras de identificação de outliers e que mais adiante retomaremos. Nos objectivos também a estimação robusta e a identificação de outliers[17] estão, logicamente, relacionadas.

Se os possíveis outliers forem localizados e removidos então o modelo ajustado aos restantes dados neutraliza-os e, portanto, produzem-se estimadores robustos.

[17] Para uma breve síntese acompanhada de um estudo sobre a contribuição dos portugueses para este assunto pode consultar-se [19].

Sabemos, portanto, que outliers podem ser, ou não ser, contaminantes e estes, se existirem, serão ou não, outliers!

E mais!

Não temos possibilidade de conhecer se esta ou aquela observação é um contaminante. Apenas podemos decidir se estamos perante um caso de contaminação de dados estatísticos. Nesta perspectiva, em todo o nosso estudo, admitiremos a impossibilidade de distinguir entre outlier e contaminante e, sem perigo de confusão, falaremos apenas de outliers para caracterizar as observações que detectaremos como suspeitas e confirmaremos responsáveis pela perturbação da nossa análise. Muitos autores seguem esta metodologia. Barnett e Lewis ([11], p. 9), relevando a fundamental detecção de contaminantes, após reflexão sobre esta questão, concluem que a atenção deve concentrar-se em "*outliers as the possible manifestation of contamination*". Esta é (também) uma opção prática. Todo o capítulo 6 da citada obra de Barnett e Lewis [11] apresenta inúmeros "testes de discordância para outliers" em amostras univariadas e, com isto, significando estudar "observações responsáveis por alguma contaminação".

1.4 Causas de aparecimento de outliers

Os exemplos que apresentámos, eventualmente, complementados por outros[18] permitem registar algumas possíveis causas para o aparecimento de observações discordantes numa amostra - erros, eventualmente grosseiros e passíveis de correcção, nas medições ou variabilidade intrínseca. Os primeiros em certas situações poderão ser facilmente corrigidos. No entanto, a maior parte das vezes isso não é possível e então estamos na verdadeira análise estatística de outliers numa amostra onde "o acaso deve ser estudado" mesmo que, por agora, se considere, bastante subjectiva, a noção de outlier.

1.5 Tratamento de observações discordantes

Em termos gerais podemos considerar que, para "trabalhar com" outliers, existem quatro métodos.

Podemos admitir que, a presença de uma observação que perturba a análise e se torna motivo de suspeição, é uma indicação que, de facto, devemos considerar uma diferente distribuição geradora da amostra e perante a qual, esse valor deixa de ser discordante. Estamos, neste caso, incorporando esse dado no estudo.

[18] Além das propostas anteriores, podemos sugerir os trabalhos de Anscombe [2] e Grubbs [56], como referências históricas.

Se admitirmos que a observação suspeita deve continuar no nosso estudo e, sob esta condição, praticarmos os habituais estudos de inferência estatística então devemos proteger as nossas conclusões através de métodos[19] de acomodação. Neste caso, não é necessário considerar modelos alternativos de discordância uma vez que o outlier é "incluído" na análise.

Numa terceira atitude perante a suspeição, podemos "identificar" essa observação e, por consequência, estudar um modelo de geração dos dados, por exemplo, uma mistura.

Num estudo de outliers em dados estatísticos, podemos concluir, é fundamental a identificação de observações discordantes e, para a executar, deve ser definido um modelo de discordância. A fase seguinte exige um teste que possa confirmar a suspeição - um teste de discordância.

Finalmente, a decisão pode ser a rejeição e continuar o estudo "com menos uma observação" e a partir daqui nada teremos a recear[20] sobre a perturbação de outliers.

[19] Por exemplo, para uma inferência na localização, usando os conhecidos métodos "trimming" ou "winsorizing" ou, para a escala, com o auxílio da mediana - não influenciada pelos eventuais extremos suspeitos.

[20] Teremos muitas oportunidades para verificar que esta decisão não é tão simples e inofensiva como, para já, nos parece. Basta admitir que algumas observações podem "camuflar" a presença de outras - *masking effect* - ou, o seu desaparecimento fazer "aparecer" suspeitos - *swamping*.

Capítulo 2

Como criar uma teoria de outliers

...quos fama obscura recondit[1]*...*

2.1 Introdução

A Estatística afecta a todos e atinge a vida em muitas situações. Como cidadãos ajudamos a fornecer informações estatísticas - a nossa própria entrada no mundo e a saída dele são registadas, para criar índices e taxas - e (através da publicidade) todos os dias nos tentam convencer de qualquer coisa ou mesmo enganar-nos à custa de factos e argumentos estatísticos.

A administração de uma comunidade, através das suas instituições de governo e comércio, depende muito das informações estatísticas e essa dependência aumenta à medida que o comércio intervém e cada vez mais, no planeamento da vida económica e social. Os propagandistas,

[1]Expressão primorosa de Virgílio (Eneida, V, 302) glosada, entre outros, por Santo Agostinho (A Cidade de Deus, volume I, Livro VII, Capítulo III, p. 611 e seguintes (1991). Serviço de Educação. Fundação Calouste Gulbenkian).

Na dicotomia entre a "razão menor" e uma "razão mais alta" deve o estatístico ter como objectivo (apenas) o conhecimento que lhe permite cobrir as suas necessidades científicas básicas? Em alternativa, esse deve ser um estádio inicial tendo por objecto a sabedoria estatística onde (ainda) admite a (enorme) importância dos "detalhes científicos" daqueles a quem uma obscura fama esconde - chamemos-lhes outliers; que são estimuladores da investigação e podem ser originados pelos valores discordantes de uma amostra - uma minoria.

São esses "menores" que fazem avançar a ciência?! Neles está a força!

administradores e dirigentes administrativos que utilizam (e deturpam) as estatísticas são bastante numerosos e a eles podem juntar-se os mais diversos utilizadores de estatísticas; desde os estudantes de ciências sociais aos políticos. Todos empregam factos e métodos estatísticos para fornecer bases para políticas. Tais factos e métodos também têm um lugar muito importante no desenvolvimento da sociologia e economia como ciências e ainda são muito importantes para os experimentadores na maior parte dos ramos da biologia e são usados por aqueles que trabalham nas ciências mais exactas como a física, química ou engenharia.

As ideias estatísticas estão pois, na base de muitas teorias e, de facto, uma "abordagem estatística" é talvez uma das facetas mais características da ciência moderna. Finalmente, a estatística como matéria é naturalmente do maior interesse para o grupo relativamente pequeno dos estatísticos profissionais.

Como resultado das várias maneiras de encarar o assunto, a palavra estatística e as suas associadas, estatístico (adjectivo) e estatístico (substantivo), têm vários significados. Em primeiro lugar temos as definições dos dicionários, em que estatística se refere, no singular, ao assunto como um todo e, no plural, aos dados numéricos.

No senso comum - para o vulgar utilizador - as estatísticas são apenas números. Ele, tem tendência a pensar que um estatístico é principalmente uma pessoa que conta o número das coisas.

Para um economista, habituado às ideias qualitativas da teoria económica, estatístico é quase sinónimo de quantitativo.

Para um físico, estatístico é o oposto de exacto, visto que, para ele, a estatística é uma matéria que acima de tudo diz respeito a grupos e possibilidades mais do que a certezas.

Para o cientista e investigador que está habituado a obter conhecimentos através da realização de experiências em condições que pode controlar, os métodos estatísticos são aqueles que se empregam quando é impraticável ou impossível um controle experimental rigoroso.

O campo de aplicação da estatística, na sua maior parte mas de modo nenhum totalmente, é económico e assim, o estatístico às vezes é considerado como uma espécie de economista.

Por outro lado, como os métodos estatísticos são basicamente matemáticos, muitas pessoas pensam, ainda hoje, que o estatístico é uma espécie de matemático.

Quase se poderia dizer que o matemático aceita o estatístico como economista e que o economista o considera um matemático. Alguns (poucos?!) pensam que os métodos estatísticos são tão poucos rigorosos que qualquer pessoa pode "provar" seja o que for e outros admitem que são tão rigorosos que nada provam. No outro extremo situam-se aqueles

que defendem que, como meio de aumentar os conhecimentos, o poder da estatística é ilimitado e quase mágico.

É habitual começar um livro, por exemplo sobre Estatística, definindo e ilustrando o assunto que vai versar. Um livro onde a estatística seja o tema primeiro, obviamente, não é excepção.

Uma leitura (ao acaso) das páginas iniciais de um livro sugerem duas perspectivas para as definições preliminares. Muitas vezes são breves e superficiais e outras inserem-se em campos particulares que limitam o próprio texto. Abordemos esta questão na sua maior generalidade. Ao encarar o tema da Estatística várias considerações se podem formular. E a primeira é a de considerar que é, ao mesmo tempo, uma ciência e uma arte. É uma ciência pelo facto de os seus métodos serem basicamente sistemáticos e terem aplicação geral e é uma arte porque o êxito da sua aplicação (também) pode depender da experiência e do engenho do estatístico e do seu conhecimento do campo onde actua. Contudo, não é necessário ser-se estatístico para apreciar os princípios gerais que lhe estão subjacentes. Como ciência, a estatística e, em particular, os métodos estatísticos fazem parte do método científico em geral e baseiam-se nas mesmas ideias e processos. Assim, a Estatística, tal como as outras disciplinas, está sempre em evolução. Ela é suportada por uma teoria... e, portanto, também e acima de tudo, é evolutiva!

Uma teoria é um conjunto de princípios fundamentais de uma ciência ou arte com formulação de uma doutrina acerca desses princípios.

A Estatística é pois, uma ciência porque, em síntese, desenvolve o conhecimento rigoroso e racional de um vasto ramo do saber e com as mais diversas aplicações. Por isso deve ser um conjunto organizado de conhecimentos baseados em relações objectivas verificáveis e dotadas de valor universal.

É pacífico e vulgarmente aceite que - numa apreciação geral - a Estatística é uma ciência que tem por objecto o agrupamento metódico dos factos sociais que se prestam a uma avaliação numérica - população, natalidade, mortalidade, rendimento de impostos, produções agrícolas, criminalidade, religião, etc.

Numa perspectiva mais restritiva, por vezes também se elege a palavra Estatística para designar um ramo das matemáticas aplicadas que recorre ao cálculo das probabilidades para estabelecer hipóteses com base em acontecimentos reais, com o fim de fazer previsões.

O avanço tem feito afirmar mais a primeira em detrimento da segunda.

A Estatística é a ciência dos dados, também aplicada porque a pesquisa, muitas vezes, visa também uma aplicação.

A Estatística é interessante e útil porque fornece estratégias e instrumentos para trabalhar os dados de modo a melhor "entrar" em problemas reais. Dados são números (ou a falta deles) inseridos num determinado contexto ou experiência. Mas, determinar a média de 50 números é puro cálculo aritmético, não é Estatística. Discernir sobre aquele valor 50 e decidir se temos uma pequena ou grande amostra e, em cada caso, concluir sobre a discrepância de determinado valor (mesmo que usando a média atrás calculada!) já é Estatística. Embora a Estatística se possa considerar como uma ciência matemática, ela não é um ramo da matemática e não deve ser ensinada como tal. Cada vez mais, podemos falar em pensamento estatístico que suporta e se apoia na teoria da decisão.

2.2 Estatística: Inferência e Decisão

A Estatística, na prática, exige julgamentos. É fácil listar as hipóteses matemáticas que justificam o uso de determinada metodologia, mas não é tão fácil decidir quando esse método pode ser "com segurança" utilizado na prática. A experiência torna-se aqui fundamental.

Mesmo na vertente científica mais simples - e menos polémica? - podendo ser, como já vimos, a Estatística admitida como um ramo das matemáticas aplicadas que recorre ao cálculo das probabilidades para estabelecer hipóteses com base em acontecimentos reais, o objectivo final, na maior parte das vezes, está relacionado com predições. Assim encarada é a vertente prática da Estatística que se está a salientar. Mas, a predição está directamente ligada à Inferência e à Decisão.

Toda a teoria ilumina a prática e esta informa a teoria, numa relação dialéctica. Sempre que se questionam os utentes (principalmente os de maior interesse prático), surgem "sugestões" que se desejam "com (muitos) estudos de casos" e "propostas de trabalho" com "menos teoria e mais prática". No entanto, o apoio teórico é sempre reconhecido e... deve estar sempre disponível e por perto!

Chegamos assim, à Teoria da Decisão Estatística. É nela que se fundamenta e onde (cada vez mais) está a génese da "profissão estatístico". Sobre este tema - teoria - respigamos algumas ideias basilares de Murteira [90].

"Apesar de o homem ser chamado diariamente a tomar decisões só muito recentemente os problemas que estes suscitam começaram a ser tratados segundo uma óptica científica" (*ib.* p. 97).

Historicamente, "a teoria da decisão estatística deve-se essencialmente a A. Wald que seguindo a tradição de Neyman-Pearson alargou consideravelmente os horizontes abertos por estes, tirando partido do

desenvolvimento da teoria dos jogos realizada por von Neumann e Morgenstern. O grande mérito de Wald (...) consistiu na contribuição para uma argumentação onde (...) em termos gerais, os procedimentos clássicos são casos particulares da decisão estatística" (*ib.* p. 108-9).

No entanto, para evitar confusões, importa que se clarifique que "(...)a teoria que vai desenvolver-se diz respeito à decisão individual e não de grupo. (...) a teoria que vai estudar-se não pretende substituir o decisor - mas, sim fornecer um conjunto de regras que auxiliem o decisor (...) Em termos gerais pode dizer-se que se está perante um problema de decisão quando se torna imperioso escolher ou optar entre, pelo menos, dois cursos de acção" (*ib.* p. 97).

Mas, (haverá ainda quem considere que) a Estatística não é uma teoria? Poderemos admitir que se trata de[2] "um instrumento ou alfaia cujas aplicações mais relevantes se situam, naturalmente, no domínio da investigação científica"?

"Nos problemas de inferência estatística ou de decisão estatística trabalha-se, quase sempre, no quadro de um modelo probabilístico ou, pelo menos, com uma forte componente probabilística" (*ib.* p. 23).

2.3 A necessidade de uma Teoria dos Outliers

Juntando esta componente, os modelos são muito importantes (fundamentais!) na investigação científica. Além disso, (também) na modelação surge de capital importância a amostra e cada um dos seus constituintes - (em particular) as observações e a sua dimensão. Na pesquisa de outliers numa amostra é, pois, de um problema de investigação que se trata...

E, como tal, vai gerar (ou criar a necessidade de se apoiar em, pelo menos) uma Teoria!

Alguns cientistas tomam Teoria como sinónimo de hipótese. Mas, a Teoria é distinta da hipótese e da ciência como sistema total. Porque, no método científico, a hipótese é uma fase anterior à Teoria e é parte integrante da ciência, teórica ou aplicada. A Teoria opõe-se à praxis ou acção mas são complementares. Como sabemos, o método científico começa por várias fases: observação e experiências, hipótese e formulação da lei geral ou Teoria. É, pois, uma hipótese já confirmada pela experiência e que faz parte integrante da Ciência. Mas há vários tipos de Teoria e Ciências. No entanto, só são possíveis dois tipos basilares de Teoria: as dedutivas e as indutivas. Nas dedutivas, existem um conjunto de proposições válidas ou verdadeiras (teorema) que se constrói a partir de um grupo de proposições primitivas (axiomas) pela aplicação de certas

[2]Murteira [90], citando Gustavo de Castro, A Estatística Matemática uma Alfaia Científica. *Rev. Med. Veterinária*. 1952, p. 52-64.

regras de inferência. Nas indutivas temos um conjunto de proposições verdadeiras ou prováveis (teoremas, axiomas e definições) que se elabora a partir de vários casos particulares por um processo de inferência imediata e generalizante.

Para muitos pensadores a conclusão da indução é apenas uma probabilidade. E, a "probabilidade de uma lei" cresce com o número de casos que a confirmam.

Se, por diversos testes "confirmamos um outlier" temos o conjunto, isto é, o suporte das regras que podem formar - uma ou a - teoria dos outliers.

Uma das formas mais gerais de definir Estatística é aquela que a considera um método de proceder à escolha de alternativas de acção em face da incerteza, pelo recurso à recolha e interpretação de dados sobre os fenómenos em estudo. Assim, nesta generalidade, a Teoria dos Outliers em Dados Estatísticos torna-se uma mais valia para a Ciência Estatística. Na sua construção algumas dificuldades surgirão. Contudo os obstáculos não parecem intransponíveis e é natural que se assista a grandes desenvolvimentos futuros. Evidentemente que, os métodos de interpretação estatística, na sua concepção teórica, são rigorosos e conduzem a conclusões válidas do ponto de vista científico. Além disso, a própria qualidade dos dados utilizados, afectando as conclusões, não pode constituir base para a acusação desses métodos. A má qualidade pode pôr em causa a metodologia.

A boa qualidade dos dados é um bem estatístico que se deseja. E, decerto, a qualidade melhora-se com a Teoria dos Outliers.

Em conclusão - uma ou a - teoria dos outliers deve fornecer um conjunto de regras que auxiliem o decisor além de construir instrumentos que permitam avaliar a qualidade da decisão.

Com esse desiderato... prosseguiremos!

Mas, para crescer e se desenvolver, a teoria deve ser transmitida a outrem que a expandirá com a sua contribuição!

2.4 Sobre o Ensino da Estatística

É um facto! O ensino e aprendizagem da Matemática tem-se tornado (cada vez mais) problemático e, são verificadas (cada vez mais) dificuldades na capacidade dos estudantes nesta área do saber. Por arrasto... a Estatística padece da mesma doença![3] Mas, não é este o local certo para

[3]Esta é uma temática que tem envolvido personalidades e instituições em diversas acções. Pelos conteúdos e temas registamos o *Colóquio sobre o Ensino e Aprendizagem da Estatística* realizado no ano 2000 e cujas principais intervenções e conclusões foram organizadas num volume com o título *Ensino e Aprendizagem da Estatística*

apresentar reflexões sobre esta problemática. Debrucemo-nos (apenas!?) sobre aquelas questões que mais de perto se prendem com "os outliers".

A Estatística tem-se difundido de forma rápida por todos os ramos do saber. É cada vez mais difícil indicar uma actividade humana onde a sua aplicação não se tenha revelado fecunda ou mesmo indispensável. A *Encyclopedia of Statistical Sciences* editada por Kotz e Johnson [77] ou *Handbook of Statistics* (cf. [108] e [83]) são boas e recentes referências onde podemos verificar os mais diversos campos de aplicação da estatística - também no que diz respeito aos outliers. Assim, o ensino da estatística em geral e dos outliers em particular, exige a versatilidade e a motivação para se implantar nos mais diversos domínios científicos.

Um livro de texto sobre introdução à estatística - desde o ensino mais elementar até ao mais alto nível - deve (também e acima de tudo?) motivar os estudantes. É um desafio aos autores que, para além do conteúdo científico, assim também devem incluir uma componente prática e de resolução de exercícios e problemas de Estatística. E, é neste momento que, os diferentes livros podem diversificar a oferta... apresentando "truques e tácticas" para resolver "problemas de estatística" ou ilustrando as diversas metodologias com exemplos que complementem a argumentação utilizada e a respectiva fundamentação. Aqueles são de mais fácil implantação no mercado livreiro e estes destinam-se à formação dos especialistas. Mas, como já referimos, os utilizadores da estatística - do tipo faça você mesmo - são muitos e, se não houver cuidado, o manual pode tornar-se numa potente arma para construir erros...

Na verdade a conjugação de três vertentes, ser generalista, fazer aplicações e utilizar pacotes estatísticos, pode fazer aumentar o risco de criar uma edição que pouco contribua para o ensino e principalmente para a aprendizagem de estatística. As demonstrações devem ser adaptadas aos estudantes a quem se destinam. Os exemplos devem ser eloquentes. O texto deve ser rigoroso e muitas vezes o rigor não é a principal preocupação, pois - na maior parte das vezes? - o objectivo[4] é "a receita prática".

"Muito[5] provavelmente, as dificuldades dos alunos tenderão a variar de acordo com as experiências educativas que lhes são proporcionadas. De facto, não é crível que os alunos manifestem as mesmas dificuldades

editado por Cristina Loureiro, Fernanda Oliveira e Lina Brunheira e numa edição da Sociedade Portuguesa de Estatística, a Associação de Professores de Matemática e os Departamentos de Educação e de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa.

[4]Já referenciámos esta questão, em particular, nos exemplos de 1.2.

[5]Cf. a obra citada *Ensino e Aprendizagem da Estatística* p. 7.

quando sujeitos a um ensino centrado na realização de procedimentos de cálculo ou a um ensino que encara a Estatística interligada com a análise de dados, a aprender no quadro da realização de trabalhos de pesquisa."

Como é de admitir, o sucesso de qualquer disciplina depende da competência adquirida aos mais diversos níveis da aprendizagem - também de outliers.

Na Estatística, além do já referido, pelo pioneirismo e pela excelência deve registar-se (e consultar!) o Projecto ALEA.

Do seu sítio *http://alea-estp.ine.pt* podemos retirar a motivação para uma cuidada abordagem: "O ALEA - Acção Local Estatística Aplicada - constitui-se no âmbito da Educação, da Sociedade da Informação, da Informação Estatística, da Formação para a Cidadania e da Literacia Estatística como um contributo para a elaboração e disponibilização de instrumentos de apoio ao ensino da Estatística para os alunos e professores do Ensino Básico e Secundário, tendo como principal suporte um sítio na web. Melhorar a literacia estatística é, assim, uma condição importante para, por um lado, garantir uma melhor prestação de um serviço de utilidade pública e, por outro lado, fomentar ambientes e experiências de aprendizagem diversificados recorrendo às novas tecnologias de informação. Intervenientes O ALEA nasceu de um projecto conjunto da Escola Secundária Tomaz Pelayo e do Instituto Nacional de Estatística, tendo evoluído para uma realidade em que a Direcção Regional de Educação do Norte incorporou o núcleo das entidades que o dinamizam. A supervisão científica é assegurada pela Prof[a] Doutora Maria Eugénia da Graça Martins, docente na Faculdade de Ciências da Universidade de Lisboa. Em particular, é de destacar, os cursos de Noções de Estatística e de Noções de Probabilidades (em desenvolvimento), cujos conteúdos são da sua autoria e seguem o programa oficial de Matemática do ensino secundário."

2.5 Um exercício pedagógico

Na era da informação, todas as áreas científicas (cada vez mais) reconhecem a necessidade de aumentar a literacia estatística.

Aos estudantes universitários de economia, psicologia, ciências sociais, medicina ou biologia é oferecido (pelo menos) um curso introdutório[6] de estatística. No entanto, entre eles - nas diversas ciências - observamos uma diferente motivação para a abordagem das questões estatísticas e, em particular, a problemática dos outliers.[7]

[6]Na diversidade de formações universitárias é difícil encontrar uma onde os métodos estatísticos não sejam incluídos nos currículos académicos.

[7]E, sabemos como os outliers estão na moda e na linguagem do senso comum.

Na vida profissional, mais tarde ou mais cedo, vão estar em contacto com dados estatísticos e... "vão decidir"!

Pretendemos resolver um problema de outliers? Estamos preocupados com alguma observação? A primeira motivação é o aparecimento de algum valor suspeito. Registemos esse facto.

Que fazer ou qual o caminho a seguir? Procuramos a ajuda de um pacote estatístico "à mão"? Alguns, mais afortunados, poderão ter acesso a alguma biblioteca que lhes forneça uma disponibilidade com maior oferta, com a possibilidade de escolha de algum livro sobre a matéria e onde possam procurar uma solução para o "problema outlier" (que criaram?) e com o qual se deparam. Alguns (muitos?) procurarão o especialista!

Mas, se "não especialistas" teremos "a maior probabilidade" de ficar apenas pela consulta de um manual.

Numa pesquisa inicial, percorremos o índice!

Em outliers - na maior parte das vezes - somos remetidos para gráficos de dispersão salientando - como exemplo! - uma observação discordante. Ficamos assim endossados para uma análise de dependência eventualmente terminando numa análise de regressão onde o "produto final" é um estudo sobre discordância entre resíduos.[8]

Estamos, de novo, num início.

Mas, não era exactamente este o ponto de partida nem o objectivo!

Que fazer?

2.6 Por onde começar?

Desde sempre, os investigadores interrogam-se sobre a existência de observações (aparentemente?!) não consistentes com os dados em estudo.

Na evolução da metodologia estatística, essa preocupação, inicialmente envolvia mais a perspectiva da rejeição e tem evoluído no sentido da justificação da presença de tais valores. A própria noção de outlier, (como já sabemos!) ainda necessita muito de ser clarificada. Nesse sentido são fundamentais os modelos e métodos de selecção e detecção de valores aberrantes numa amostra que são grandes domínios de investigação - já no presente, mas também no futuro!

Um grande número de temas se perfilam quando investigamos outliers em dados estatísticos. Desde uma abordagem meramente teórica e probabilística envolvendo distribuições com diferentes tendências para o aparecimento de outliers[9] até à análise prática dos mais variados pa-

[8]Na componente prática, o estudo de outliers em análise de resíduos será, entre outros, abordado no exemplo IX.

[9]"Proneness" - para produzir outliers e "resistance" - para se proteger deles.

cotes estatísticos e o modo como executam e os critérios que usam para efectuar o estudo de valores discordantes num conjunto de dados.

A identificação de outliers é muito mais importante do ponto de vista prático do que, sob possíveis abordagens teóricas, sobre as divergências em relação a um modelo. Na sequência do estudo deve admitir-se e ser elaborado um processo de acomodação.

Mas... toda a prática supõe uma teoria que a suporta donde emana e, para a qual, com reciprocidade natural contribui.

A teoria ilumina a prática e esta informa a teoria, numa relação dialéctica fundamental!

Numa situação extrema decidimos rejeitar algumas observações antes de elaborar qualquer inferência ou porque encontramos explicações aceitáveis que justifiquem a sua presença ou porque aplicamos um teste construído a partir de um modelo de discordância assumido. Em alternativa, devemos combinar todo o estudo de outliers com as correspondentes questões de robustez.[10]

Acomodação, influência e robustez são palavras-chave[11]para o aprofundamento de outras tantas questões igualmente importantes no estudo de outliers e que muitas vezes se considera serem apenas para os especialistas e portanto ficam (quase) como temas de investigação e não como pontos concretos e de índole aplicada.

E... não o esqueçamos, a questão primordial (ainda) é:

O que é um outlier?

O que é um outlier e como trabalhar estatisticamente (com) essa observação é uma questão que cada vez mais preocupa os investigadores. Está mais longe a perspectiva de rejeição pura e simples de qualquer observação não consistente com as restantes. Devemos portanto ter a mínima sensibilidade para a análise e descoberta de valores discordantes numa amostra. E essa competência adquire-se pela aprendizagem.

O ensino da estatística nas nossas escolas muito tem a fazer neste domínio - pela excelência, sensibilizando cada vez mais o utilizador comum. E aqui (também) aparece a importância da informática como

Cf. Barnett e Lewis ([11] p. 52-4).

[10]Como vimos no capítulo 1, vários métodos se perfilam. Por exemplo, "trimming e winsorizing" dão bons resultados para os outliers.

Cf. Barnett e Lewis ([11] p. 143-6)

[11]Além de referências pontuais por opção, não desenvolveremos estes importantes temas neste texto. De facto, a intenção generalista a que nos propusemos torna-se incompatível com o aprofundamento destes temas específicos que, por si só, poderão preencher o conteúdo de uma obra sobre outliers. Registamos a proximidade e a interligação destes assuntos. Sobre a influência sugere-se [29]. No entanto, para um melhor esclarecimento do leitor - também para a justificação da nossa opção - e para uma leitura comparada, sugerimos a consulta do capítulo 8 de Barnett e Lewis [11].

primeiro veículo de transmissão de informação nesse domínio. A maioria dos meios computacionais à disposição remete, o estudo de outliers, sistematicamente para abordagens (apenas) gráficas. Abordaremos esta problemática. No que se segue, optamos também por uma exposição generalista sobre dois (principais) instrumentos de trabalho. Salientamos também as importantes contribuições históricas, à luz da actual metodologia. Dando especial ênfase ao carácter teórico-prático (desde já e para que conste!) introduzimos as duas principais ferramentas - as caixas-com-bigodes e os testes de discordância.

Why do outlying observations arise and what should one do about them? - é o título do capítulo 2 da fundamental obra sobre outliers em dados estatísticos publicada por Barnett e Lewis, como sabemos, inicialmente em 1978 e que já se desenvolveu até à 3ª edição em 1994. Quase trinta anos depois, o título acima é uma boa síntese do principal problema no estudo de outliers. De facto, as primeiras preocupações com valores discordantes numa amostra surgem logo que os investigadores se questionam sobre "o peso a atribuir a cada uma das observações" ou suspeitam e especulam sobre a "menor ou maior tendência para surgir um valor aberrante".

Este é o início!

2.7 Porquê "Outliers e Contaminantes"?

Aprofundemos, um pouco mais, a questão da contaminação dos dados estatísticos que já anteriormente abordámos na secção 1.3.

Seja $x_{(1)}, ..., x_{(n)}$ uma amostra ordenada. Tradicionalmente no estudo de observações discordantes, num primeiro e único ponto da sua análise, o experimentador julgará se o mínimo $x_{(1)}$ (ou o máximo $x_{(n)}$) da amostra deve ser declarado outlier. A escolha, para estudo e teste, dessa eventual observação discordante - ou, conjuntamente, dessas eventuais observações - é, portanto, bastante (completamente?) subjectiva. Quando devemos analisar $x_{(1)}$? E $x_{(n)}$? Quando deve ser estudado[12] o par $(x_{(1)}, x_{(n)})$?

De um ponto de vista informal, a selecção depende do modo como as observações se "dispersam" e do "afastamento" entre o máximo (ou o mínimo) e as restantes. Além disso, essa escolha não é arbritrária. Ela depende, em primeiro lugar, da distribuição F, geradora da amostra. Decerto que, pela "normalidade" da natureza estatística e probabilística

[12]Esta é uma questão de "excelência" no estudo de outliers e para a qual procuraremos dar resposta nos capítulos 7 e 9 e, principalmente, do ponto de vista prático, com os exemplos da secção 9.2.

dos dados, em "a maior parte dos casos", nem o mínimo nem o máximo "ficam de fora"[13] e o problema outlier não é colocado.

Mas, o que significa "estar de fora"? O problema apenas se coloca (ou é colocado?!) quando algum dado "perturba" a análise. Esta ideia induz uma primeira tentativa de solução e uma "intuitiva" definição de outlier, apresentada por alguns especialistas, onde esse termo "apenas" tem o significado de uma "observação que está de fora".[14] Não se dando um significado para essa noção, assim, de novo, é imposta uma adicional "carga subjectiva" na análise estatística. Sobre este tema e também de um ponto de vista histórico deve consultar-se o importante estudo experimental sobre a natureza subjectiva dos processos de rejeição de outliers apresentado por Collett e Lewis [28] em 1976. Alguns anos depois, em 1984, aprofundámos esta problemática e apresentámos uma tese[15] clarificando a noção de outlier. Complementarmente e numa sequência de datas[16] devemos acrescentar o trabalho apresentado por Muñoz Garcia *et al* [89] em 1990 - com uma abordagem formal da problemática dos outliers - além de, obviamente, as três edições de *Outliers in Statistical Data* de Barnett e Lewis.

Portanto, nesta "perspectiva tradicional" temos, diversas acções perante uma abordagem do estudo de outliers em dados estatísticos. Assim, podemos "concluir" que o mínimo $x_{(1)}$ ou o máximo $x_{(n)}$ é um outlier. Mas ficam, em aberto, algumas dificuldades que carecem de resolução. A primeira é aquela que permite discernir entre uma decisão considerando aqueles valores como outliers e aqueloutra que admite o par $(x_{(1)}, x_{(n)})$ como discordante. A resposta não é simples e, como já referimos, mais adiante sobre ela dedicaremos alguma reflexão.

Qual é a diferença entre os dois casos? A resposta está directamente ligada à questão da contaminação dos dados. Nessa perspectiva, os valores extremos podem ser ou não ser outliers mas, qualquer outlier é uma observação extrema.

Para este tema, é importante o estudo apresentado por Barnett [8]. Como distinguir entre outliers - que ficam de fora - e contaminantes - que são gerados por outra distribuição?

Analise-se a figura 2.1 numa perspectiva de outliers e contaminantes.

Para isso, suponhamos que nem todas as observações são geradas pela mesma distribuição F. Admitamos que um valor (ou mais?) da amostra

[13] Expressão que podemos usar para concretizar, em português, a imagem "outlying" muito aplicada em textos de língua inglesa.

[14] É o que acontece, por exemplo, com a "solução" caixa-com-bigodes.

[15] Referimo-nos à tese de doutoramento [112] apresentada na Universidade de Lisboa.

[16] Na secção 2.8 abordaremos outros caminhos e referências históricas fundamentais para a criação de uma teoria dos outliers.

provém de uma outra população cuja lei de probabilidade é G, diferente da anterior F, e que dela resulta, por exemplo, por uma translação.

As observações que vêm de G são chamadas contaminantes.

Portanto, os contaminantes terão um valor médio diferente de F. Será maior ou menor?

Algumas das observações "o" é outlier?

E qual delas é "mais discordante"?

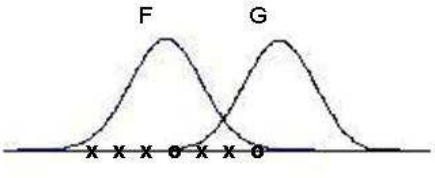

Figura 2.1: outliers e contaminantes

O que devemos "alterar" ou, por outras palavras bem ao jeito da teoria dos outliers, como devemos "acomodar os dados" - encontrar as respostas a estas questões para cada um das observações "o"?

Quantos outliers existem nos dados em análise?

Em conclusão, nesta abordagem, podemos afirmar que os outliers (porque ficam de fora) podem, ou não, ser contaminantes e que estes também podem ser ou não ser outliers. Como já afirmámos, esta é a "perspectiva tradicional" no estudo de outliers em dados estatísticos onde[17] "*all we can do is concentrate attention on outliers as the possible manifestation of contamination and we shall see how statistical methods for examining outliers aim specifically at this prospect*". Para estes autores, um outlier é[18] uma observação (ou um conjunto de observações) que parecem ser inconsistentes com as restantes.

[17] Citamos Barnett e Lewis ([11], p. 9).

[18] *ib.* p. 7.

Como também afirmam Barnett e Lewis[19] outros autores usam o termo outlier com um significado diferente.

Desde já, esclarecemos o leitor que, é entre estes que nos colocamos.

Nesta "nova" perspectiva, um outlier é (também) uma observação surpreendente mas que é confirmada como "estranha" por algum teste, dito, de discordância.

Portanto, temos dois possíveis caminhos para o trajecto da construção da noção de outlier. Um, que considera outlier aquela observação que "fica de fora" e que depois a declara discordante e outro onde outlier - vale mais do que a informação (definição tradicional!) de que "está de fora" - é uma observação que, estatisticamente, contém a informação de "porque fica de fora".

Assim, é outlier não porque fica de fora nos dados mas porque fica de fora no modelo de contaminação que gerou "esses dados". É uma observação surpreendente que é discordante (contaminante) e, por isso, se diz outlier. Representaremos esta observação por "outlier" em alternativa a outlier para a perspectiva tradicional.

A opção feita exige mais e nova sistematização pois a declaração "outlier na amostra" depende do mecanismo de geração - o modelo de discordância - admitido e da decisão através de um teste específico - o teste de discordância. Este avanço permite clarificar a noção e torná-la menos subjectiva.

2.8 Perspectiva histórica

2.8.1 Sobre a evolução da Teoria

Como vimos, no primeiro capítulo, os mais antigos critérios estatísticos para estudo de observações discordantes, foram propostos por Chauvenet há mais de um século e ainda hoje são utilizados como primeira abordagem (e muitas vezes, única?).

Um estudo enciclopédico dos métodos para estudo de outliers nos primeiros tempos está incluído em Harter [63] elaborado no início da década de setenta do século passado apresentando uma completa discussão dos primeiros escritos sobre rejeição de outliers. Aliás, Harter tem desenvolvido excelente trabalho no campo da pesquisa histórica sobre critérios para teste e rejeição ou modificação e ponderação daquelas observações que possam contaminar os dados. Harter e Balakrishnan [64], já no virar do milénio, apresentam os principais problemas, que ainda são actuais e as mais recentes contribuições para a teoria geral das estatísticas ordinais.

[19] *ib.* p. 7 e 38.

Pesquisas na literatura dos séculos XVIII e XIX têm provado ser bastante antiga a sensibilidade dos estatísticos relativamente ao problema "outlier".

Um dos primeiros trabalhos salientando a necessidade de tratar de modo especial as observações surpreendentes foi elaborado por Bernoulli em 1777.

O estudo de Beckman e Cook [16], em 1983, faz uma excelente síntese sobre o tratamento estatístico de outliers, quer do ponto de vista histórico quer das aplicações aos modelos padrão na estatística, mas concluem que: "*Although much has been written, the notion of an outlier seems as vague today as it was 200 years ago*". Passado mais um quarto de século é o momento de olhar de novo para essa afirmação e concluir sobre a sua actualidade. Com estas páginas também desejamos ajudar nessa tomada de decisão.

Edgeworth, em 1887, escreveu: "*Discordant observations may be defined as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined*"; e Grubbs [56], em 1969, oitenta e dois anos mais tarde, afirma que "*An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs*". Estas afirmações são relevantes no sentido de que um outlier é de facto um conceito subjectivo e "pós-data".

A diversidade de utilizadores dos modernos meios - principalmente os computacionais - para selecção de observações discordantes torna bastante (e ainda mais) pertinente a questão da subjectividade. Este tema está directamente relacionado com a origem da discordância de um ou vários valores num conjunto de dados e com os diferentes métodos estatísticos disponíveis para estudo de outliers.

Aprofundemos, um pouco mais, esta questão.

2.8.2 Sobre a evolução do conceito de outlier

O conceito de outlier tem sido abordado nos mais variados estudos. Após a identificação de observações discordantes, a decisão mais fácil é sempre a rejeição desses dados perturbadores. Mas isso implica, pelo menos, a diminuição da dimensão da amostra. Deve abandonar-se o estudo de toda a observação que "se afaste" do conjunto dos dados? Qual a influência dessa decisão no estudo e na identificação de novos outliers? Estas são questões muito importantes na teoria dos outliers. Analisar a dependência do conceito outlier em função do número de observações na amostra e a sua influência nas mais importantes conclusões é um ponto fundamental e não tratado na literatura especializada. Este é

um assunto que abordaremos neste texto. Além disso, dado o relevo nas aplicações, desenvolveremos exemplos nas distribuições mais importantes na estatística - exponencial e normal.

Quando se pretende avaliar o mecanismo de geração de outliers numa amostra, podemos ainda afirmar, como Barnett e Lewis ([11], p. 49):

"*by far, the most common type of alternative hypothesis as a model for contamination is what we shall refer to as the slippage alternative*".

Quanto à justificação da presença de observações discordantes, este é o ponto da situação na teoria dos outliers, registado nessa referência fundamental em todos os estudos desta área - o tratado de Barnett e Lewis já sobejamente referido. Passada mais de uma década desde a terceira edição, seguramente, também sobre este tema mantém actualidade. De facto, aquela é a hipótese alternativa mais comum. No entanto, ao ser assumida vai restringir (um pouco?) a generalidade da análise pois (subjectivamente!) limita o estudo de outliers "apenas" às observações que se tornaram suspeitas ao investigador. Como veremos no capítulo 4, a nossa perspectiva[20] é mais geral e, pelo modelo assumido, é não subjectiva.

É vasta a lista de referências onde tem sido discutida e usada aquela hipótese (justificativa de contaminação) como alternativa a uma hipótese nula onde se admite que não existem outliers na amostra.

Os modelos[21] (ditos) de tipos A e B de Ferguson [44], elaborados há cerca de cinquenta anos, constituem uma expressão geral para a alternativa por deslizamento indexado na localização ou na escala para a distribuição normal.

O modelo A admite alteração na localização em uma das observações; enquanto que no modelo B essa perturbação observa-se na escala. A alternativa "slippage" é então definida embora ainda com algumas restrições. A reflexão sobre esta última consideração é muito importante para se compreender como tradicionalmente a teoria dos outliers envolve um julgamento subjectivo com selecção *a priori* do candidato.

Barnett e Lewis ([11], p. 99) generalizam o modelo de discordância por deslizamento e consideram testes para uma hipótese alternativa onde se admite - e se fixa desde o início do estudo - por exemplo, o

[20] O estudo que apresentaremos e que, em particular, aplicaremos a populações exponenciais no capítulo 5, ilumina e clarifica a noção de outlier em amostras dessas distribuições - inclusive na problemática dos outliers múltiplos, como é reconhecido (também) por Barnett e Lewis.
(Cf. [11] p. 204-5).

[21] No capítulo 4 estudaremos em detalhe estes modelos bem como a sua importância para uma perspectiva histórica do estudo estatístico de outliers.

máximo $x_{(n)}$ (ou o mínimo $x_{(1)}$) com uma diferente distribuição. Assim, esta hipótese, que chamam de "*labbelled slippage*" também identifica - subjectivamente e *a priori* - uma observação extrema como único valor discordante na amostra. Um modelo mais geral, baseado em metodologia de máxima verosimilhança, foi proposto, em 1984, por Rosado [112] onde não é fixada *a priori* a observação que implicitamente determina a hipótese alternativa. Este modelo[22] generativo com alternativa natural - método GAN - contribui para consolidar a definição de outlier.

Na teoria dos outliers, a decisão entre as várias hipóteses apresentadas, envolve directamente os chamados testes de discordância. Sobre este vasto assunto é fundamental o capítulo 6 de Barnett e Lewis [11]. Rosado [112] estudou alguns desses testes de discordância envolvendo alternativas por deslizamento (slippage) e apresentou um modelo de discordância geral utilizando a hipótese alternativa natural. Rosado [115] usa esse mesmo modelo de discordância generativo e compila os principais resultados consequentes desse método para interpretação de outliers e os respectivos testes de discordância nas populações mais comuns.

Em quase todos os livros (e também nos) de estatística é incluído um índice remissivo.

Não é raro constatar que a palavra outlier é das mais referidas.

A consulta desses índices é uma maneira para um possível modo de, em cada obra, podermos encontrar (uma) solução para questões como:

Rejeição de outliers;

Outliers em Regressão;

O que é um outlier;

(...)

Outliers Multivariados.

Mas, muitas vezes, esta facilidade de pesquisa e oferta de soluções entra em forte conflito com o rigor da terminologia. Ao leitor menos sabedor pode parecer que todos os autores lhe estão a propor o mesmo "produto científico", idêntico "tema de trabalho" ou a mesma ferramenta estatística. De facto, não é assim e a profundidade a que o assunto é tratado varia muito de texto para texto. Também aqui é fundamental a uniformização (também) da teoria dos outliers.

Ao iniciar uma (mesmo que seja) breve reflexão sobre um conceito estatístico, de imediato pensamos na respectiva definição.

O que é um outlier?

[22] Que é um dos objectivos deste estudo e que será desenvolvido e aplicado a partir do capítulo 4.

Uma vista rápida por dicionários[23] forma uma boa síntese da problemática.

Ao acaso, consultamos três de estatística e registamos:

Outlier - A subject or other unit of analysis that has extreme values on a variable; encontramos em *Dictionary of Statistics and Methodology* de Paul Vogt.

Outlier - an observation which is far removed from the others in the set; propõe Roger Porkess em *Dictionary of Statistics*.

Outlier - an observation that appears to deviate markedly from the others members of the sample in which it occurs; é a definição de Everitt (*Dictionary of Statistics in the Medical Sciences*).

Facilmente podemos encontrar novos exemplos (e que pouco mais avançam?) sobre a definição de outlier. Todos consideram outlier como observação extrema. Não é esta a perspectiva dos mais recentes estudos onde se admite que a discordância de um valor não implica que seja extremo (por exemplo em modelos estruturados) . Ficamos com a certeza de que é uma noção que necessita ser clarificada. Toda a definição de outlier que encontramos, com mais ou menos rigor, se estabelece à volta das seguintes: "uma observação que parece ser inconsistente com o restante conjunto de dados "(assumida, como já vimos, por exemplo por Barnett e Lewis [11]) ou "um valor que com base nalgum critério objectivo é inconsistente com os dados" (considerada por Collett e Lewis [28]). O modelo de discordância com alternativa natural apresentado em Rosado [112] e que utilizaremos, define outlier através da segunda daquelas perspectivas. Assim, podemos concluir que, a definição de outlier "está entre dois extremos" - um que apenas assume essa observação como "estando de fora dos dados" e o outro exigindo uma decisão estatística de rejeição através de algum instrumento estatístico especializado, por exemplo, um teste de discordância.

Além desta questão e talvez muito mais importante encontramos a natureza subjectiva dos processos de rejeição de outliers. Estes, como foi provado por Collett e Lewis [28], são essencialmente baseados em duas fases, envolvendo a princípio um julgamento individual de que um determinado valor é surpreendente e só depois testado como discordante. Ao escolher-se um nome - outlier - para aquela observação que "ficando de fora em relação às restantes" é, de imediato, assumida e classificada como relevante para o estudo.

[23] E, a talho de foice, podemos desde já, afirmar a importância da existência de um bom Dicionário de Estatística em Português. Mais adiante, voltaremos a esta questão.

Outlier é toda a observação que estatisticamente se comporta de modo diverso das restantes. Deve pois, ter "um único nome" e consolidada a sua definição. Assim, seria evitado que em artigos e livros científicos se encontre terminologia confusa e pouco rigorosa quando se quer apresentar um estudo de outliers.

Em elucidativa selecção de "sinónimos" para observações discordantes, podemos escolher:

"*Outlying observations*",
"*far extremes*",
"*extremos*",
"*outlier severo*",
"*outlier moderado*",
"*outside value*",
"*far ouside value*",
(...)

que são (alguns dos) termos que encontramos ligados à "problemática dos outliers".

Alguns autores consideram como outlier toda a observação tendo um valor que marcadamente difere dos restantes - fica "de fora"; outros, por sua vez, propõem que um outlier é uma observação tal que "nalgum sentido" se afasta das restantes e que, após rejeição por um teste, se intitula de observação discordante. Nesta perspectiva - poderemos definir? - outliers discordantes e outliers não discordantes. Outliers são também observações que causam problemas na modelação estatística, por exemplo, num ajustamento por mínimos quadrados. A diversidade na nomenclatura exige que em cada trabalho sobre outliers cada autor tem necessidade de apresentar a "sua definição". Deve consolidar-se um termo para significar uma observação estatisticamente relevante. É o que propomos, declarando como "outlier" toda a observação que através de um critério objectivo seja responsabilizada pela não homogeneidade de uma dada amostra.

Abordemos, um pouco mais, a evolução da sensibilidade dos estatísticos, como veremos, sempre com a intenção de clarificar a noção de outlier.

Vejamos algumas citações[24] retiradas da vasta literatura estatística sobre observações discordantes numa amostra.

"*The problem of how to deal with data which contain outliers, i.e. observations which look suspicious in some way, has long been a source*

[24]Para uma melhor compreensão do nosso objectivo, sublinhamos alguns temas muito interessantes, também numa perspectiva histórica.

of concern to experimenters and data analysts."

(Guttman e Smith [58])

"*It is well recognised by those who collect or analyse data that values occur in a sample... which are so far removed from the remaining values that the analyst is not willing to believe that these values have come from the same population. Many times values occur which are dubious in the eyes of the analyst and he feels that he should make a decision as to whether to accept or reject these values as part of his sample.*"

(Dixon [35])

"*The general problem (of rejection of outliers) is a very old and common one. In its simplest form, it may be stated as follows.*

In a sample of moderate size taken from a certain population, it appears that one or two values are surprisingly far away from the main group..."

(Ferguson [44])

"*The problem (...) is to introduce some degree of objectivity into the rejection of the outlying observations.*"

(Ferguson [45])

"*The outliers are values which seem either too large or too small as compared to the rest of the observations.*"

(Gumbel [57])

"*One or more of the observations may have the appearance of being outliers and we are interested here in determining ... whether such observations should be retained in the sample for interpreting results or whether they should be regarded as being inconsistent with remaining observations.*"

(Grubbs [55])

"*An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.*"

(Grubbs, [56])

Estas citações são suficientemente elucidativas da natureza subjectiva - e também histórica - dos métodos de rejeição e identificação de outliers numa amostra.

Em síntese, todas relevam um outlier como uma observação inconsistente com os dados.

Além disso, como consequência, delas se pode inferir que, para consolidação, se torna imperioso criar uma teoria dos outliers, isto é, de um conjunto de princípios fundamentais.

Essas opiniões sistematizadas, aumentando o nível da literacia estatística, também podem contribuir, para a consolidação de uma "prática de excelência" em toda a análise estatística de outliers.

2.8.3 Os principais instrumentos: caixa-com-bigodes - um instrumento prático

Como sabemos, os outliers são vulgarmente definidos como aquelas observações que parecem ser inconsistentes com o resto dos dados. Elas (também) podem ser causadas por erro nos registos. Mas é importante saber se uma observação discordante é genuína ou indica a presença de uma outra distribuição nos dados. Muitas vezes são utilizados métodos gráficos para detectar esses valores aberrantes. As diversas abordagens e os mais variados tipos de gráficos tornam fortemente subjectiva essa análise preliminar de outliers e podem fazer concluir em sentidos errados sobre a presença ou ausência de valores discrepantes numa amostra. Nessa primeira análise dos dados há, desde logo, uma influência do método de procura de eventuais observações discordantes. O estudo Rosado e Mendes [123] aborda e analisa essa questão. A noção de outlier é evidentemente influenciada pelo modelo de discordância considerado. Vários têm sido os modelos que permitem justificar a presença de observações discordantes numa amostra e deles dependendo a teoria dos outliers.

O problema do estudo - detecção e selecção - de observações discordantes numa amostra está directamente associado à experimentação. Desde sempre os cientistas se questionaram sobre a confiança nos dados (principalmente os suspeitos) por eles recolhidos e por eles estudados. Assim, sempre que o problema surge, os autores devem concluir sobre a presença/ausência de observações aberrantes (discordantes, discrepantes, erróneas, etc.).

A terminologia para dados incoerentes numa amostra tem sido bastante variada. Como já vimos, admite-se que desde os finais do século XVIII se discute a questão outlier. No entanto, a própria definição de outlier ainda não se encontra única e estabilizada. Será ela mesma um outlier? Para alguns autores, outlier é uma observação (ou conjunto de observações) que parece ser inconsistente com o restante do conjunto dos dados. Para outros é além disso exigido que, perante algum critério, aquela observação seja também discordante. Podemos assim falar em "valores surpreendentes" como aqueles ou em "valores suspeitos" como estes últimos. Mas desde quando se preocupam os experimentadores com uma detecção gráfica de outliers? A resposta, como muitas outras em estatística, está directamente relacionada com o avanço das tecnologias computacionais e informáticas.

Assim, a caixa-com-bigodes, numa perspectiva prática, é o principal instrumento para um estudo preliminar de uma amostra. É o primeiro (e muitas vezes o último?) método de selecção e detecção de outliers num conjunto de dados; sendo tanto mais vezes usado quanto maior é a disponibilidade de meios electrónicos de cálculo.

Em todos os pacotes estatísticos[25] se remete para a caixa-com-bigodes quando, nos respectivos manuais, procuramos o que fazer para detectar observações discordantes e apenas os mais sofisticados oferecem alternativas envolvendo, por exemplo, métodos para detecção de outliers em modelos estruturados. Este tema foi, por nós abordado, em 2001, em comunicação convidada apresentada no VII Congresso da Sociedade Portuguesa de Estatística (Cf. [118]) revelando as preocupações do autor sobre o uso de métodos gráficos para detecção de outliers.

Tukey [134], em 1977, definiu inicialmente as caixas-com-bigodes como "gráficos esquemáticos" e Velleman e Hoaglin [136] apresentaram-nas para os utilizadores "não técnicos" e demonstraram a sua potência para uma grande diversidade de situações, a que se costuma chamar, da análise exploratória de dados.

Partindo do resumo de 5 números, Hoaglin *et al* [68], em 1983 (e em 1992 na edição portuguesa, Cf. [69]) argumentam que "os dados que estão bastante para além dos quartos" deverão ser considerados potenciais outliers. Usam então a dispersão-quartal Dq, isto é, a amplitude dos dados definida pelos quartos superior Q_3 e inferior Q_1, para tornar mais preciso e dar um significado ao termo outlier. Esta abordagem pressupõe que uma observação discordante se deve "afastar" das restantes. É um pressuposto que carece de justificação.

Murteira [91], em 1993, considera também, a dispersão-quartal como o "padrão de variação ou gabarito" que logicamente se impõe quando pretendemos examinar uma colecção de dados com vista à identificação de valores estranhos.

As "caudas" da caixa-com-bigodes são essencialmente obtidas a partir dos dados mais extremos dentro das barreiras de outliers.

Não são pois muito perturbadas por mudanças nos valores de qualquer dos outliers e só sofrerão alterações modestas por mudanças dos valores que estão dentro das barreiras de outliers.

Note-se que essas barreiras de outliers são elas próprias definidas através dos quartos da colecção de dados.

Conseguem pois resistir a perturbações de cerca de 25% desses dados.

[25] Algumas questões sobre este assunto foram focadas nos exemplos IV e V.

Pode assim argumentar-se sobre a resistência da caixa-com-bigodes. Mas, de facto, existem condicionantes pela redundância das observações discordantes na construção das suas próprias barreiras.

Essa deficiência está directamente relacionada com a própria noção de valor discordante onde apenas se admite ser um extremo.

Tudo isto deve ser clarificado e necessita de princípios e métodos.

Exige uma teoria!

Mas, na realidade, a caixa-com-bigodes é o grande instrumento prático.

Em resumo, com as devidas reservas já formuladas, podemos construir aquela que é a fundamental regra prática para decidir sobre outliers, através do uso de caixa-com-bigodes:

i) Definir $Q_1 - 1.5Dq$ *e* $Q_3 + 1.5Dq$ como barreiras de outliers, sendo Dq a dispersão-quartal.

ii) Se $x_i < Q_1 - 1.5Dq$ *ou* $x_i > Q_3 + 1.5Dq$ então x_i é outlier.

Neste momento, é um bom exercício, teórico e prático, reflectir sobre o significado da conclusão: "x_i é outlier".

Alguns autores avançam nesta problemática[26], e distinguem entre aquela observação que é um *outlier severo* quando se tem $x_i < Q_1 - 3Dq$ ou $x_i > Q_3 + 3Dq$ e um *outlier moderado* quando, pelo contrário, se tem $Q_1 - 3Dq < x_i < Q_1 - 1.5Dq$ *ou* $Q_3 + 1.5Dq < x_i < Q_3 + 3Dq$.

Nesta aplicação são introduzidas as chamadas barreiras externas, sendo $Q_1 - 3Dq$ (inferior) e $Q_3 + 3Dq$ (superior) e as barreiras internas com $Q_1 - 1.5Dq$ (inferior) e $Q_3 + 1.5Dq$ (superior).

Conta-se que Paul Velleman, sendo discípulo de Tukey - "inventor" da "regra" 1.5Dq - lhe perguntou:

Porquê 1.5?

Tukey terá respondido:

Porque 1 é muito pequeno e 2 é demasiado grande.

Claro que este diálogo é motivado por uma das principais questões da teoria dos outliers e que se prende com a respectiva identificação.

Aquela ingénua pergunta envolve uma difícil resposta.

A "tendência estatística para a normalidade" facilmente nos fornece alguma justificação para aquela constante 1.5, pois este valor permite verificar que é aproximadamente 1% a probabilidade de que um valor seja outlier enquanto que ao factor 2 corresponderia uma possibilidade de 1 para 1000 de ser classificado como discordante.

Numa perspectiva histórica é interessante registar que Tukey não refere o termo outlier. De facto, ao introduzir "*box-and-whisker plots*" diznos que "*it is convenient to have a rule of thumb that picks out certain*

[26]Por exemplo, Murteira *et al* (Cf. [92] p. 33-6).

values as "outside" or "far out", ([134], p. 43) e, de modo a obter uma abordagem gráfica para ter a "*identification of individual values that may be unusual*" (*ib.* p. 55).

Usando a caixa-com-bigodes podemos, como Murteira ([91], p. 100) questionar e concluir que "a identificação (é ou não é?) e a interpretação (se é, porque é?) de outliers" são "tarefas complexas e altamente subjectivas".

De facto, a selecção de valores discordantes numa amostra é um velho problema e cuja solução, na maior parte das vezes, depende do analista.

Devemos pois introduzir objectividade na metodologia de detecção / rejeição de outliers.

2.8.4 Os principais instrumentos: Máxima Verosimilhança - um instrumento teórico e prático

Admitindo que se pretende aprofundar o estudo de outliers no sentido da pesquisa de um maior e mais rigoroso suporte teórico várias alternativas se podem colocar. A escolha pode ser feita entre alguns testes *ad hoc* utilizando estatísticas "específicas" para determinados campos da ciência ou ser baseada em modelos e testes de discordância. Estes, do ponto de vista teórico assentam fundamentalmente em dois princípios - máxima verosimilhança e de optimização local - sendo aquele, sem dúvida, muito mais importante.

Historicamente, como sabemos, aplicando o princípio de máxima verosimilhança, Neyman e Pearson sugeriram um método para obter funções das observações para testar o que chamaram de hipóteses estatísticas compostas.

Mais tarde, Wilks construiu a razão de verosimilhanças para testes daquelas hipóteses e encontrou a respectiva distribuição assimptótica. O, assim chamado, teste da razão de verosimilhanças tem desde então conhecido larga aplicação na inferência estatística e, também, na teoria dos outliers. São inúmeros os trabalhos publicados envolvendo aquela metodologia.

Primeiro que tudo, naturalmente, a construção dos testes estatísticos depende, das hipóteses formuladas.

Então quais são as hipóteses estatísticas em presença num análise de valores discordantes?

Se considerarmos que, qualquer uma das n observações de uma amostra, pode ser discordante então temos em estudo uma formulação estatística com hipóteses compostas. A teoria da máxima verosimilhança torna-se pois um bom instrumento teórico - pelo suporte que fornece - mas também prático - pela diversidade de aplicações que potencia.

Figura 2.2: Os outliers

Com base em metodologia de máxima verosimilhança, no capítulo 4, formularemos uma teoria geral para o estudo de valores discordantes em dados estatísticos. É uma contribuição para a teoria dos outliers que - cada vez mais o sabemos! - deve ser criada.

2.9 O que são "Os Outliers"?

2.9.1 What Is An Outlier, Anyway?

Estamos (e também já vivemos?) na "blogosfera". Tudo "está na internet"... e, nessa liderança de comunicação, os outliers não podem ser outliers!

Com alguma (muita!) ironia que, pode servir para depreciar (mas também engrandecer?) a teoria dos outliers, atente-se na linguagem de senso comum do diálogo seguinte... aquela que todos usamos - inclusive para argumentação científica?

Pesquisemos "na net" o que é um outlier. Para ter o maior sucesso, devemos render-nos à subserviência linguística e, a língua de Camões[27] pouco avançaria nessa tarefa internacional - neste domínio uma epopeia lusíada decerto seria não gloriosa - pelo que deve ceder o passo ao idioma do british Shakespeare.

Formulada a questão titular desta subsecção somos, por exemplo, conduzidos ao "sítio": http://www.fuzzyco.com/outliers/.

Recebidos "no hall de entrada" ficamos a saber que chegámos, de facto, aos outliers e que, muito esclarecidos, nos apresentam o seu "cartão de visita" (ver figura 2.2).

E, aí chegados "tentamos" a procura da resposta para a nossa "questão fundamental".

O diálogo segue-se (obviamente em inglês):

[27] No capítulo 3 aprofundaremos, um pouco mais, a problemática da linguagem científica.

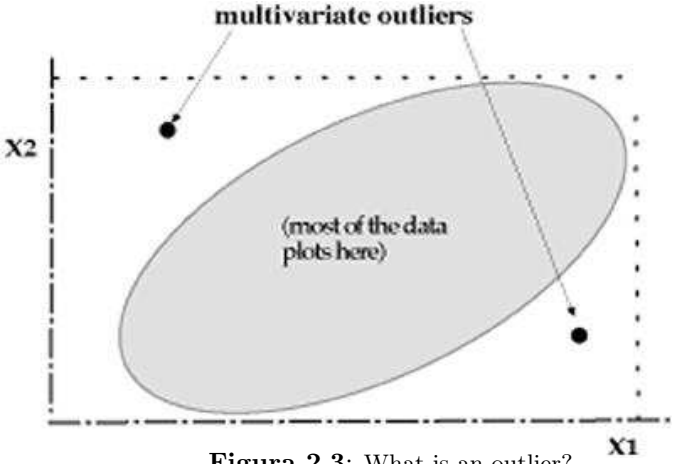

Figura 2.3: What is an outlier?

"Whenever someone approaches us at a gig, the following conversation usually ensues:

Person: Hey, you guys are really great. Blah blah, blah blah diddy blah...

Outlier: Thank you.

Person: What's the name of your band?

Outlier: The Outliers.

Person: The what?

Outlier: The Outliers

Person: The Outliners?

Outlier: No, The Outliers.

Person: What's an Outlier?

Outlier: Well...

So, here's what an Outlier is (not that we want to appear too geeky, but):

An Outlier is a statistical term. It refers to observations in a distribution of data that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism, and therefore discarding of the observations might be considered. Here's a graphical representation of what this means". (ver figura 2.3)

Ficámos elucidados? E, alguma das figuras valeu... mais do que mil palavras? A literacia estatística aumentou?

2.9.2 Sobre a literacia

Já relevámos a importância do ensino e aprendizagem... (também) de Estatística.

Todos devemos saber de (e falar sobre) outliers? Esse tema deve ficar reservado aos especialistas? Não é fácil (nem pacífica?) a opção! porque, de médico... todos temos um pouco...

E também de estatístico?!

Para toda a argumentação, sempre a literacia deve - entenda-se, tem de - ajudar.

E ela pressupõe aprendizagem. Mas será a aprendizagem total?

"Professor[28]: Não podemos estar seguros de nada neste mundo, Menina

Aluna: A neve cai no Inverno. O Inverno é uma das quatro estações. As outras três são... euh... Pri...

Professor: Sim?

Aluna: ...mavera e depois Verão... e... euh...

Professor: Começa com o...

Aluna: Ah, sim. Outono...

Professor: É exactamente isso, Menina, muito bem respondido, está perfeito. Estou convencido de que será uma boa aluna. Fará progressos. A Menina é inteligente, parece-me instruída e tem uma boa memória.

Aluna: Sei as estações, não é verdade, senhor?

Professor: Sim, Menina... ou quase. Mas isso virá com o tempo. De qualquer maneira, já não é mau. Vai ficar a saber todas as estações de olhos fechados. Como eu.

Aluna: É difícil.

Professor: Não é, não. Basta um pequeno esforço, um pouco de boa vontade, Menina. Verá. Isto vai com o tempo, fique tranquila.

Aluna: Oh, espero que sim, Professor. Tenho uma grande sede de instrução. Os meus pais também desejam que eu aprofunde os meus conhecimentos. Querem que eu me especialize. Pensam que uma simples cultura geral, mesmo que sólida, não é suficiente na nossa época.

Professor: Os seus pais estão cheios de razão. Deve continuar com os seus estudos. Desculpe que o diga, mas é uma coisa necessária. A vida contemporânea tornou-se muito complexa.

Aluna: E também tão complicada. Os meus pais têm bens, nisso tenho sorte. Poderão ajudar-me no meu trabalho, nos meus estudos muito superiores.

Professor: E pretende preparar-se para...?

[28]Transcrevemos um diálogo, numa tradução portuguesa, em *A Lição* de Ionesco.

Aluna: Tão cedo quanto possível, para o primeiro concurso de doutoramento. É daqui a três semanas.

Professor: Desculpe a pergunta, mas já fez o curso dos liceus?

Aluna: Sim, Professor. Tenho o meu diploma de ciências e o meu diploma de letras.

Professor: Oh, mas está muito avançada, mesmo demasiado avançada para a sua idade. E para que doutoramento pretende preparar-se? Ciências físicas ou filosofia?

Aluna: Os meus pais gostariam muito - caso o senhor pense que isto é possível em tão pouco tempo - que eu fizesse o doutoramento total.

Professor: O doutoramento total?... Tem muita coragem, Menina, felicito-a sinceramente. Tentaremos, Menina, fazer o nosso melhor. Aliás, já sabe bastantes coisas, sendo tão nova ainda.

Aluna: Oh, professor.

Professor: Então, se me permite, peço-lhe perdão, mas temos que nos deitar ao trabalho. Não temos tempo a perder."

Regressemos[29] para o mundo dos homens comuns onde "nenhum dos quais nasceu sem as suas imperfeições, e feliz é aquele que apresenta menos(...)" mas, onde é possivel que "os loucos fiquem infectados de sabedoria (...) E a este propósito, o testemunho é constituído pelo provérbio que diz: A loucura é a única coisa que mantém a juventude em suspensão e a velhice afastada. Como se verifica nos Brabantes de quem habitualmente se diz: Aquela idade que usualmente torna os outros homens mais sábios, torna-os mais loucos".

Mas... onde estão os outliers?

2.9.3 A descoberta de outliers

Nesta viagem pela ciência estatística em direcção à descoberta dos "verdadeiros" outliers prossigamos, de imediato, através de um continuado exemplo com um modelo de discordância relevando (talvez) aquela que é a principal dificuldade num estudo de outliers: a sua definição.

Exemplo VII: Modelo de Discordância (conclusão)

Consideremos, de novo, a "pequena amostra" seguinte, introduzida na secção 1.2:

$$1.74, 1.46, -0.28, -0.02, -0.40, 0.02, 3.89, 1.35, -1.10, 0.71$$

Como vimos, numa primeira abordagem, o valor 3.89 pode atrair a atenção do analista, embora não se tenha uma explicação determinística

[29] Respigamos (ao acaso?) algumas passagens de *O Elogio da Loucura* de Erasmo.

para tal suspeição. Como pode então ser visto como um outlier? Aquele dado, para além de ser um extremo - todas as amostras possuem esse facto estatístico - aparece como "extremamente extremo".

Se aceitarmos normalidade - por exemplo $N(0,1)$ - para os dados, então 3.89 passa também a ser um valor cujo aparecimento é estatisticamente pouco razoável. Esta razoabilidade é também condicionada pela eventual alteração, por exemplo em translação, da distribuição normal que o gerou.

Na teoria e na prática, a decisão sobre o modelo de discordância a adoptar é sempre bastante difícil. Para se perceber o problema na sua globalidade basta acrescentar que os dados do exemplo em análise foram gerados por duas normais. Além da $N(0,1)$, que gerou a maior parte, as observações 1.74, 1.35 e 3.89 são de uma normal $N(3,1)$. Naquela amostra temos 3 outliers. O grande desafio estatístico está portanto no modelo de discordância.

Temos assim a introdução de diversas componentes de um modelo cujo objectivo principal é explicar a presença de valores discordantes.

Mais do que muitos outros, o estudo de outliers em dados estatísticos é bastante exigente e fortemente condicionado pelos pressupostos que nos diversos exemplos temos vindo a introduzir. Torna-se então evidente a dificuldade para a excelência. Uma pesquisa de outliers não pode ser feita apenas como uma aplicação de um qualquer pacote estatístico que, com o auxílio de um simples gráfico, permite decidir. Mais do que uma realização do acaso, a presença de outliers numa amostra pode ser uma informação primeira de uma alteração do modelo até aí assumido.

Prosseguindo esta descoberta de outliers, vejamos o exemplo seguinte que, (estatisticamente também) se enquadra num estudo de dados em modelos estruturados. Este exemplo, diz ao estatístico e investigador que, em ciência, a surpresa pode ser causada pelas variáveis de onde menos se espera que, pela inovação, devam contribuir para o avanço - qualquer que seja a forma como o entendamos...

A terceira idade, é uma das mais importantes etapas da vida e que chega "pela calada da noite". Quando menos se espera, descobrimos que "não somos eternos" e que "estamos velhos" ou, como tal somos considerados pelas outras pessoas; e não conseguimos identificar bem o momento dessa importante descoberta - somos outliers evolutivos?!

Para cada um de nós o sentimento de envelhecimento é muito diferente. Existem pessoas com mais de oitenta anos que pensam e agem como jovens, e pessoas jovens que são verdadeiros velhos. São frases comuns - não outliers!

Ficar jovem leva tempo!

Nesta frase[30] podemos subentender que o processo de amadurecimento leva-nos a valorizar o jovem e, na idade adulta, o comportar-se jovialmente é uma coisa positiva. Transportar para a velhice valores da juventude significa viver como jovem, viver melhor... Mas isto leva tempo para se compreender e, infelizmente, às vezes nunca chega a ser compreendido...

O tempo que levamos para a compreensão da importância que é o sentir-se jovem é a chave do processo e, evidentemente, quanto mais curto for, melhor. Este processo fica mais fácil de ser transposto se tivermos informações substanciais sobre a velhice - que assim, nunca surgirá como outlier - e quanto mais subsídios conseguirmos sobre o processo de envelhecimento, mais rapidamente o compreenderemos e chegaremos à "juventude", isto é, ao seu estado de espírito. Sentir-se jovem é, sem qualquer dúvida, um sinal básico de boa qualidade de vida.

Deve-se então ignorar a palavra velho. Este é, com toda a certeza, um termo "corrompido" pelo uso e, portanto mal interpretado, carregando uma dose alta de aspectos negativos. O termo Terceira Idade deve ser o escolhido, pois é mais suave, dá ideia de coisa natural que se segue à Segunda Idade.

No final da Segunda Idade, tempo eleitor dos que não serão outliers - entre os egrégios avós de que se há-de sentir a voz que guiará a Pátria à vitória - encontramos (mais) uma dádiva da mãe natureza: os netos.

Nesta idade, com a aturada visão microscópica da experimentada vida de pais mas, com uma acrescida perspectiva macroscópica adquirida por essa vivência, "novas vidas" vemos florescer! Nesta "bela continuidade na mãe natureza" elas são faróis de primeira categoria. E, como tal, exigem excelente manutenção. Também por isso, o seu crescimento, cada vez mais, é estudado - desde o "momento zero".

No exemplo seguinte olhamos, também cientificamente, para os filhos dos nossos filhos. Mas, não apenas para os números...

Exemplo XII: Um "novo" Modelo de Discordância.

Consideremos[31] *os dados da* ***Tabela 1*** *onde, a par dos nomes, se*

[30] Atribuída a Bernard Shaw (1856-1950), um dos mais importantes dramaturgos da língua inglesa que - na terceira idade? - recebeu o prémio Nobel de Literatura em 1925. Pela fina ironia, fica bem neste contexto!

[31] Ao introduzir este exemplo, também se faz um alerta para o uso indiscriminado de tabelas de crescimento e para os eventuais erros que, pela rotina de utilização, se podem cometer na análise, apenas pelo facto de os pressupostos que permitem comparar dados não se verificarem. Para aprofundar e melhor compreender esta "interpretação da evidência" pode consultar-se, para este caso, por exemplo alguns estudos (Cf. http:// www.fao.org) das Nações Unidas onde são fornecidos elementos

Tabela 1: Comprimento e peso de 17 bebés aos 12 meses

	comprimento X	peso Y
Ana Maria	73	7.720
Carlota	73	9.400
Carolina	78	8.950
Constança	79	11
Duarte	84	11.500
Francisco	77	11.800
Guilherme	76	9.600
João Pedro	75	9.600
Laura	73.5	11
Madalena	71	9.850
Margarida	75.5	8.975
Mariana	74	9.450
Mariana Rita	78.5	11.250
Rafael	74	10.200
Rodrigo	76	11.300
Santiago	80	11.200
Vasco	78.5	9.800

registam os respectivos comprimentos e pesos aos 12 meses.

Segundo as mais recentes tabelas (portuguesas?), a associação média entre o comprimento (em centímetros) e o peso (em Kg) das meninas com 12 meses deve ser (74, 9.450). *Para os meninos é* (75.5, 10.250). *Assim, no registo daquelas variáveis devemos esperar dados que pouco se afastem dos "valores médios" referidos.*

Todos os dados recolhidos são recentes e portanto comparáveis através das mesmas tabelas de crescimento.

---

que permitem concluir que para a maioria dos países da europa se podem usar tabelas americanas, salvo na estatura dos meninos finlandeses, alemães e suecos até aos 5 anos que "em média" devem ser comparados com o percentil 80 dos americanos. Por sua vez aos japoneses corresponde o percentil 30. Para as meninas, são as alemãs, as holandesas e as norueguesas que devem utilizar o percentil 70 dos americanos. Em qualquer dos sexos, quer para a estatura quer para o peso, os referidos estudos aconselham alguma prudência na aplicação aos países Ibéricos, já que algumas regiões de Espanha, como Burgos e Guadalajara, forneceram os meninos mais altos e mais baixos. Então, com a devida reserva, analisemos estes dados mesmo sem fazer intervir algumas outras variáveis que os podem condicionar como a estatura da mãe ou a duração da gestação e que já constam em algumas tabelas de crescimento como informações complementares.

Não nos esqueçamos - já se diz - que, no último século, os portugueses "cresceram em média" 8 centímetros. Daí que talvez tenha pouco significado "comparar" bisavós e bisnetos.

Um gráfico de associação entre comprimento e peso, com os dados da tabela acima, salienta umas observações mais do que outras. Qual delas é mais discordante?

Estas duas variáveis em análise - assumidas e bem estudadas como normais - são reguladas por um modelo natural de dependência. Este, é o responsável pela comparação entre pares e, por consequência a partir dele, é costume "ordenar" os bebés pelos quantis onde "são inseridos". Como interpretar os resultados desta primeira "tarefa científica"?

Cada uma das crianças que tem peso abaixo do percentil 25, poderá ser um outlier no peso. Como se diz em análise regressão, neste caso, é um outlier em y! É um[32] outlier inferior? É um outlier moderado?

E o mesmo se pode dizer[33] para o comprimento.

E uma criança que tem comprimento acima do percentil 75, poderá ser um outlier para essa variável? É um outlier em x? É um[34] outlier superior?

E o mesmo se pode dizer para o peso.

Para ambos os casos, na língua portuguesa algumas vezes ainda se fala em outliers severos quando o "afastamento" é muito grande.

Qual destes acontecimentos - "outlier superior" ou "outlier inferior" - é mais provável?

E assim poderemos, subjectivamente, "produzir" outliers para um modelo assumido. Se este for alterado, decerto a conclusão será diferente. Basta que se considere uma tabela de crescimento de uma outra população, por exemplo, nórdica?

Quais são as diferenças entre as duas análises? Como comparar uma observação outlier em x e não outlier em y com uma outra, outlier em y e não outlier em x? Será "apenas" outlier? E, não é cada filho um outlier para os pais?

E ainda não terminámos as dúvidas.

Um dado com duas medidas outlier superior (ou inferior) ou com uma superior e a outra inferior é mais provável do que um outro com "medidas na média"? Por outras palavras, uma observação média é "menos outlier" do que uma outra onde algum dos valores se afaste da média? Como comparar afastamentos da média?

[32] Lower outlier, na terminologia inglesa de Barnett e Lewis [11].

[33] No índice remissivo, para cada um destes termos, podem consultar-se indicações sobre os respectivos desenvolvimentos ao longo do texto.

[34] Upper outlier, na terminologia inglesa de Barnett e Lewis [11].

Lá mais para a frente, encontraremos respostas para algumas destas questões.

Com as hipóteses assumidas neste exemplo podemos concluir que o peso médio dos bebés de 12 meses é $9.850Kg$. Além disso, o seu comprimento médio deve ser $74.75cm$.

Por sua vez, os dados da tabela anterior fornecem a média 76.3 para o comprimento, com o desvio padrão 3.29. Para o peso obtemos 10.170 com o desvio padrão 1.12. Salientem-se as discrepâncias entre os desvios padrão obtidos.

Para as meninas, aqueles dados têm a média 75 no comprimento com o desvio padrão empírico igual a 2.84. A média dos pesos é 9.73 com o desvio padrão 1.17.

Para os meninos, a média dos comprimentos é 77.8 com um desvio padrão 3.32 e para o peso esses valores são 10.670 e 0.88, respectivamente.

Em todos os casos - para os meninos, para as meninas e para os dados globais - as médias das variáveis são superiores aos respectivos valores "retirados das tabelas de crescimento". Assim, estas meninas têm comprimento médio com mais 1 centímetro e peso com mais (cerca de) 300 gramas, do que "o que diz" a tabela de crescimento; por sua vez, os meninos têm 2.5 centímetros acima do valor tabelado e o seu peso também é superior em (cerca de) 500 gramas.

Registamos ainda os grandes valores dos desvios padrão para os comprimentos quando comparados com os dos pesos.

Para um estudo estatístico, este facto pode indiciar maior fiabilidade dos pesos dos bebés quando comparados com os respectivos comprimentos.

Para as duas variáveis em causa existe um modelo estatístico ao qual elas "se adaptam". Adoptado esse modelo, a cada par fica associado um resíduo que "mede o afastamento" da observação correspondente. Este resíduo - parte intrínseca *a posteriori* de cada observação - condiciona e pode ser usado para definir um "par outlier". E um resíduo nulo pode ser menos provável que um outro diferente de zero. Daí, a correspondente observação, surge como mais discordante do que qualquer outra.

Com estes "dados dos amigos" conseguimos um rol com os mais diversos "termos científicos" associados a uma observação discordante e já com alguma generalidade; pois estamos a tratar de dados estruturados.

Um outlier é portanto uma singularidade e indica um dado que, de todo, não é típico em relação aos restantes componentes da "sua" amostra. Deve encontrar-se a razão dessa singularidade. E um método estatístico que o permite fazer, neste caso, é o estudo de outlier em análise de regressão. E se juntarmos mais variáveis ao modelo, atingindo portanto

um estudo multidimensional de dimensão p - com p maior do que o 2 em que nos colocámos neste exemplo - como veremos no capítulo 10, somos levados para métodos da análise de dados multivariados, muito em voga, e cujo principal instrumento é a análise em componentes principais, que estudaremos. Neste caso - de modelo não estruturados - encontraremos novos métodos para estudar outliers num contexto diferente.

Em resumo, numa primeira abordagem - focando dados onde se assume alguma modelação entre as diversas variáveis - nada de novo, nos traz este exemplo, sobre uma pesquisa de outliers em dados estruturados. No entanto, as pequenas comparações estabelecidas devem permitir concluir que as variáveis comprimento e peso não são igualmente fiáveis para um estudo estatístico daquelas características dos nossos bebés. O peso dá mais garantia. Será esta variável "mais importante" do que aquela?

Mas, a Carolina, a Margarida e o Santiago são os netos do autor. Portanto, uma "variável sentimental" acrescentada ao modelo de discordância torná-los-ia em outliers do coração. Ela ensinaria que, como em muitos outras coisas da vida (profissional), também a independência é factor fundamental para o avanço científico.

É a descoberta da "variável fulcral" que, em cada exemplo, fornece a reconhecida capacidade do estatístico para poder (sempre) provar a sua tese.

2.10 A Fortuna / O Acaso decide!

Admitamos uma análise de outliers em dados estatísticos. Dividamos - separemos - os dados em estudo em duas partes (dois e só dois grupos?): "os eleitos", que admitimos em maior número - a maior parte - e "os suspeitos". Estes - sempre presentes?! - estão em muito menor quantidade - pois vulgarmente[35] consideramos apenas um ou dois valores discordantes. Não há qualquer motivo assinalável para aquela escolha. No entanto, ela é (quase sempre) feita. A confirmação está na (cada vez maior) utilização dos testes de outliers nos mais diversos domínios científicos e pelos diferentes especialistas também nas aplicações, inclusive dos pacotes estatísticos. Todos desejam melhorar a qualidade dos seus trabalhos e conclusões através de uma "purificação dos dados". Mas, poderão os suspeitos conter mais e melhor informação que os eleitos? Qual é a causa que faz eleger estes em detrimento daqueles? Porque é que os suspeitos - a quem uma obscura fama esconde - não são os eleitos

[35] Também por razões científicas!

das análises estatísticas mais eloquentes? É que se trata de um assunto muito importante - este de mostrar quais são "os verdadeiros" embora possa não ser deles que nos vêm (as melhores!) pistas - e também todas as fragilidades - para as conclusões.

Divididos os dados estatísticos entre eleitos e suspeitos, devemos questionar "quem é" ou "porque é" eleito.

Quem atribui essa "condição estatística"?

O Acaso[36] é a única coisa que não acontece por acaso.

A Estatística[37] é muito antiga mas tem uma história curta. Só entrou nas academias[38] no segundo quartel do século XX e o principal arquitecto foi Fisher - justamente apelidado de fundador da moderna ciência estatística.

Com certeza, é a Fortuna que, mais conforme o seu capricho do que conforme a justiça, assegura a todos "a eleição" ou, pelo contrário, "a obscuridade".

Mas, quantos eleitos e quantos suspeitos?

Se a Fortuna decide quem são os obscuros - os que têm "a verdadeira força" - porque não consegue essa honra para si própria? É caso para pensar? Porque sofreu uma fortuna adversa? Nesse caso ela, que nobilita os outros mas a si própria não pode nobilitar, é adversária de si própria! Produz assim os "outliers"!

Os eleitos devem estar sempre em maioria? E, merecem mais atenção? Muitas vezes são escolhidos em razão da maior importância ou da contribuição para o estudo.

E de entre os escolhidos, quais são "os mais fracos"? São todos igualmente bons? Qual é então a razão (ou a causa!) que obriga alguns eleitos a terem menor importância (por que estes devem existir!). E a comparação entre os "eleitos menores" e "os suspeitos"? Estes, chamemos-lhes outliers, podem ter muito mais valor. É um outlier que abre o acesso a uma análise estatística mais profunda - que pode ser a origem de um trabalho de excelência. É a excelência em qualquer estudo de outliers que pode fazer a diferença entre um estatístico e um utilizador da estatística.

[36] A propósito de outliers, revisitemos este tema. A frase é de Almada Negreiros (Cf. p. 125 de *Matemática e Cultura*. Furtado Coelho *et al.* 1992. Edições Cosmos). Já foi tema de conferência de Tiago de Oliveira (*ib.* p. 125-49).

A estatística casa bem com o acaso e ambos criam necessidade. É um tema recorrente que também já gerou um *leitmotiv* para uma edição SPE - *Estatística com Acaso e Necessidade*; Actas do XI Congresso Anual.

[37] Neste contexto, leia-se a "pequena excursão" apresentada por Tiago de Oliveira (*ib.* p. 125-8).

[38] Sobre este assunto são importantes os artigos de Efron e Rao em [107].

São os outliers - a quem uma obscura fama esconde - que conferem a vida aos dados.

Seguramente, o acaso é quem mais garante - também a cada dado estatístico - a notoriedade ou a obscuridade.

Não se julgue mais digno de honra aquele que é agrupado nos eleitos.

Mais fortes, são aqueles a quem o acaso - a mãe natureza - deu a condição de possuir muito mais informação (estatística).

Para os eleger crie-se (pelo menos) uma teoria dos outliers!

É o que tentamos com este livro!

Capítulo 3

Outliers em Português

3.1 Nota prévia

Este capítulo, com pequenas alterações, corresponde ao texto, com o mesmo nome, publicado no livro *Memorial da Sociedade Portuguesa de Estatística* editado no âmbito das comemorações dos 25 anos desta associação científica que congrega os interesses da estatística e dos estatísticos portugueses.

Independentemente de algumas das referências apresentadas em anexo deste capítulo serem escritas em inglês, de facto, elas correspondem a investigação de autores portugueses no âmbito da teoria dos outliers e a sua listagem é fundamental, por um lado, para uma análise da evolução dos interesses dessa área científica entre os autores portugueses e por outro, fornecendo informação, numa perspectiva de trabalho futuro.

Trata-se portanto de um "texto memorial" onde, para além de um relato sobre "A Ciência Estatística em Portugal" nos últimos 25 anos, se pretende salientar a contribuição portuguesa para o estudo estatístico de outliers num contexto histórico que, (também) consideramos, deve ser registado.

3.2 Na década de 70 - parte I

A Estatística tornou-se conhecida no século XX como um instrumento matemático para analisar os dados e afirmou-se como ciência na sequência da contribuição inovadora desenvolvida por grandes cientistas do século XIX que foram introduzindo o "pensamento estatístico" nas diversas áreas do saber. A (criação da) Estatística é, portanto, o culminar

de uma "tradição de pensamento científico". Quando e como, despertou Portugal para essa realidade (estatística) mundial?

Situemo-nos na década de 70 do século passado. Em Portugal viviam-se "momentos novos" nos mais diversos caminhos e também no campo da ciência. Com especial incidência a partir dos anos 60, muitos portugueses migravam no seu país, procurando melhores condições de vida ou, (pelo menos,) trabalho. Alguns, não poucos, partiam para países terceiros onde "a certeza" de "melhor vida" ajudava a enfrentar esse desafio. Este estado de espírito português também arrastava e desafiava os jovens para novos caminhos. Alguns, principalmente os jovens universitários, descobriam (assim) o percurso para dizer "não à guerra colonial". Muitos jovens de Portugal, terão aproveitado esta "motivação" que lhes trouxe a feliz consequência de progressão (também) no saber. Na sua grande maioria eram oriundos das grandes cidades e das zonas "mais evoluídas". Ser jovem no interior do país e com ambições de "estudos avançados" implicava a migração - eventualmente de toda a família - para uma das três cidades, à época, com universidade. Para essa minoria, na maioria dos casos, os percursos eram bastante sinuosos (e aleatórios?) tornando difícil atingir o "grande objectivo familiar" - completar "um curso superior".

Nessa época, na língua portuguesa, as palavras mestrado e doutoramento tinham um significado não muito bem definido e, sempre se relacionavam com graus científicos "do estrangeiro".

No início dos anos 80 começaram a ser criados os Mestrados em Probabilidades e Estatística. Doutoramento em Portugal era um acontecimento raro e a especialidade de Probabilidades e Estatística não existia. Muito se avançou nos últimos 20 anos. Os estudos pós-graduados já estão a dar os primeiros passos.

Nesses anos - de pós-guerra colonial - Portugal iniciava-se pois, para um lugar na Ciência. É uma excelente referência temporal para se iniciar um Memorial[1] que - como "aquele outro" de enorme sucesso - também envolve "homens e formigas" que levam "isto daqui para ali porque as forças não dão para mais, e depois vem outro homem que transportará a carga até à próxima formiga, até que, como de costume, tudo termina num buraco..." (citando *Memorial do Convento* de José Saramago).

Na década em referência, alguns obreiros, despertavam, tentavam transformar... Uns, com a ajuda de bolsas de estudo peregrinavam "lá para fora", às vezes para bem longe, por outras universidades dando passos fundamentais... Outros, poucos e em muito menor número, avançaram "dentro de portas"... com diferentes dificuldades! Referindo uns e outros, de pioneiros estamos a falar!

[1] Referimo-nos ao *Memorial da Sociedade Portuguesa de Estatística*, de onde este texto foi compilado.

Nos "idos anos oitenta" surgiu pois o despertar português para a investigação científica - nessa época incipiente (também) em estatística.

Alguns históricos estatísticos portugueses - na sua maior parte incentivados por Tiago de Oliveira - juntavam-se à diáspora lusitana. Foram, com sucesso, até outras universidades aprender; para até nós trazer ciência estatística. Outros - e neste grupo me integro - pelas razões mais diversas, decidiram ficar e, acumulando ensino e investigação, também ajudaram a sedimentar novos cursos universitários ajudando a Matemática Aplicada a "dar à luz a Estatística" nas Universidades Portuguesas e aqui, é justo referir o protagonismo da Faculdade de Ciências da Universidade de Lisboa. Uns e outros, alguns anos mais tarde, congregavam esforços para alcançar "novas perspectivas" científicas em Portugal, também na Ciência Estatística que entretanto vinha sendo implantada a partir de pioneiros como Tiago de Oliveira , Bento Murteira e outros. Estava portanto "acontecendo o acaso" que seria a génese da moderna Estatística em Portugal.

Um primeiro grande fruto, em 1980, foi a fundação da Sociedade Portuguesa de Estatística - SPE, (durante alguns anos incluindo também a Investigação Operacional) com a designação de Sociedade Portuguesa de Estatística e Investigação Operacional - SPEIO. No início da década de 90, Ivette Gomes liderou o grupo que sedimentaria a SPE - associação de onde a Investigação Operacional se tinha separado pois entretanto tinha sido criada a APDIO - Associação Portuguesa para o Desenvolvimento da Investigação Operacional, onde os investigadores dessa área se congregaram.

3.3 Na década de 70 - parte II

Em 1978, Barnett e Lewis (Cf.[11]) publicaram a primeira edição de *Outliers in Statistical Data* - livro de base para o estudo de outliers em dados estatísticos tanto do ponto de vista teórico como prático. Nesta obra fundamental foi, pela primeira vez, agregada e sistematicamente organizada toda a vasta literatura sobre outliers. Na segunda edição, em 1984, os autores incluíram novos temas do estudo estatístico de outliers e outros que sofreram grande evolução metodológica desde a publicação da edição anterior. Em 1994 foi publicada a terceira edição e nela foram incluídas novas abordagens para dados univariados e multivariados, apresentando ainda tópicos especiais nos métodos bayesianos e em sucessões cronológicas com os aditivos e os inovadores.

As "observações difíceis" de uma amostra sempre desafiaram os estatísticos. O conceito de outlier tem fascinado (em especial) os cientistas que numa primeira abordagem querem interpretar os dados. Os mais di-

versos nomes têm sido aplicados a uma observação (ou a um grupo de observações) que se apresenta diferente; desde "não representativa" até "espúria" ou "discordante", numa terminologia tão vaga quanto as outras. De facto, para uma observação ser discordante, é fundamental que se indique o modelo do qual discorda... relevando portanto o modelo de discordância. Na época em que estamos, o registo da informação, ainda com mais ênfase permitia admitir como erros todas as observações que ao experimentador parecessem mal vindas. E as reacções foram desde os seguidores da "incondicional inclusão" - como admitem Barnett e Lewis na primeira edição da obra acima referenciada - porque "nunca devemos violar a santidade dos dados" atrevendo-nos a julgar as suas propriedades, até aqueles que sempre usam a "metodologia" "na dúvida deita-se fora" como regra prática.

Em 1976, (Cf.[4]) Barnett publicou "*The Ordering of Multivariate Data*", um estudo fundamental cujo lema é: "*order properties... exist only in one dimension*" e com discussão pelos melhores especialistas. É um artigo de referência[2]que desperta para a importância da ordenação na detecção de observações discordantes. Conjugado com a dimensão dos dados estatísticos esse artigo "atravessa" muitos domínios, novos à época, como o estudo de dados multivariados e a sua relação com as subordens. Li esse artigo. O "termo outlier" surge "no contexto" onde vai adquirindo cada vez mais importância à medida que se avança no estudo desse texto. Este pode ser um sinal, a palavra-chave, para o despertar de um novo campo de investigação - "observações discordantes em dados estatísticos" - (nesta década ainda) sem história em Portugal e muito novo no mundo científico de então!

E assim pôde acontecer (mais) um acaso científico!

Este, (verificado em 1982) conduziria à elaboração de uma tese [112] *Existência e Detecção de Outliers - Uma Abordagem Metodológica*, para obtenção de doutoramento na área dos outliers - o primeiro em Portugal. Numa perspectiva actual os pontos de vista são mais sofisticados. A teoria estatística dos outliers já possui diversas metodologias de tratamento de observações discordantes ou contaminantes; têm sido propostos modelos de discordância que permitem explicar a geração dos dados; os procedimentos robustos têm tido bastante avanço. Introduzindo diversos mecanismos de geração de outliers numa amostra, nesta época e com bastante interesse do ponto de vista prático, foi publicado [66].

Outras referências históricas, também sobre este assunto, podem ser consultadas em [11].

[2]No capítulo 10 enquadraremos este estudo numa análise multidimensional dos dados estatísticos.

3.4 Um relato na primeira pessoa

No contexto deste Memorial, à minha modesta contribuição não pode deixar de ficar associado o nome do Professor Tiago de Oliveira que foi o meu orientador de doutoramento e que me fez descobrir o caminho, também para ele novo, da teoria dos outliers. Recordando o Professor Tiago desloco-me no tempo e revivo bons momentos que têm início nas aulas da, na altura recém criada, Licenciatura em Matemática Aplicada na Faculdade de Ciências da Universidade de Lisboa - na "velha" Escola Politécnica e que se prolongam até às sessões de acompanhamento do meu trabalho de investigação conducente ao doutoramento que regular e semanalmente mantínhamos como agenda onde, na maior parte das vezes, eu era apenas um ouvinte atento da sua vasta cultura e eloquência que me deram a oportunidade de muito aprender e de muito crescer.

O Prof. Tiago foi o meu Mestre desde os tempos da Faculdade, onde me iniciei como estudante universitário e de onde, até hoje, apenas me "afastei" para cumprir o serviço militar obrigatório nos anos "de referência" - 1973/75.

O Prof. Tiago tinha grande capacidade para o cálculo científico e era enorme a rapidez como manobrava as mais intrincadas expressões matemáticas. Quando lhe apresentei aquele que viria a ser um dos meus primeiros resultados, também para ele inesperados e inovadores, o seu grau de surpresa foi tal que replicou: "Os cálculos estarão certos?" Felizmente estavam e tive a oportunidade de ver e viver a (primeira) alegria da descoberta na presença de um grande cientista. Estes resultados iniciais conduziriam à inovação científica no estudo de observações discordantes "no meio da amostra" a que (com alguma lógica) chegámos a admitir chamar "inliers". Mas, o *modelo de discordância* é o instrumento estatístico fundamental para o estudo, que descrimina, (e condiciona!) a condição outlier (de uma ou várias observações). Tal como não devemos distinguir entre outliers superiores e outliers inferiores pois ambos são discordantes em relação a um modelo e essa condição em nada os distingue, também abandonámos a designação inlier. É o modelo de discordância - que apelidei de generativo com alternativa natural[3] - que "condiciona" e "permite definir" uma observação que deve ser declarada outlier; depois de esta ser descoberta através de um teste de homogeneidade à amostra. Outliers[4] são observações que "estatisticamente" nos surgem diferentes. No entanto, a condição outlier é fortemente condicionada pelo modelo de discordância que admitimos, como

[3]Este modelo de discordância inserido numa abordagem geral do estudo de Outliers em Dados Estatísticos será apresentado no Capítulo 4.

[4]Por enquanto ainda usamos esta palavra num contexto geral de escrita. Na secção seguinte, explicitando a nossa perspectiva passaremos a usar "outlier" (com aspas).

mecanismo gerador dos dados em presença. E cada vez mais é uma noção usada nos computadores. Todo o pacote estatístico invoca a sua actualidade com variadas aplicações na detecção de outliers nos diferentes ramos. Sabemos como o Prof. Tiago pouco simpatizava com os computadores. Talvez as suas críticas fossem bem mais mordazes com o avanço que nos levou até à Internet, onde aparentemente qualquer leigo se pode "cultivar", incluindo na teoria dos outliers. Basta saber "navegar" e escolher um bom "site". Pois bem! Desde grupos artísticos e musicais (http://www.fuzzyco.com/outliers/) - como já referimos no capítulo 2 - até outliers arqueológicos (http://ecolan.sbs.ohio-state.edu/jhm/arch/outliers.html) podemos encontrar nos momentos seguintes ao "toque do rato" no sítio certo da respectiva "home-page".

São os sinais dos tempos que nos fazem reflectir sobre caminhos percorridos e percursos vindouros. Como se deturpará uma noção pelo seu mau uso, não rigoroso e completamente vago!? Não podemos confundir a divulgação científica com o marketing. Mas, de facto, já nesta nossa época, procurando outliers na "rede global de informação" chegamos primeiro ao acessório e só os especialistas conseguem (não o necessitando) analisar onde estão os verdadeiros outliers (o fundamental!).

Com alguma dificuldade conseguimos encontrar as referências à obra base [11] de Barnett e Lewis sobre o estudo de outliers:
(http://www.amazon.com/exec/obidos/tg/detail/-/0471930946/002-4576153-0685610?v=glance).

Numa linguagem para todos compreensível, o Prof. Tiago foi um outlier. Tal como na estatística, que tanto amou e tão apaixonadamente fez crescer e criar escola em Portugal, qualquer observação discordante só é confirmada na sua condição outlier desde que assumido algum modelo de discordância. Cientificamente perfeito, o modelo da vida não nos permite construir o respectivo teste de discordância. Com o seu desaparecimento prematuro, a mãe natureza (que costumava invocar) não lhe permitiu ver reconhecida muita da sua obra. Nesta época jubilar para a Sociedade Portuguesa de Estatística que ele tanto quis e da qual foi o principal dinamizador, em breves palavras registo, a "mais sincera homenagem"!

3.5 "Outliers" em português!

Outlier em portugês deve ser "outlier". O glossário de termos estatísticos disponível na página web do International Statistical Institute - ISI, associação prestigiada de congregação mundial de estatísticos e onde a SPE é associação filiada desde 1988, pode ser um ponto de partida. Numa

consulta àquele documento - e cada vez mais este gesto se tornará trivial - é proposta a seguinte correspondência para a palavra outliers:

- valeurs aberrantes, observations aberrantes para a língua francesa,
- valori anomali para a língua italiana,
- valores extremos, valores atípicos para a língua espanhola,
- valores de exceção (sic!) para a língua portuguesa.

Destes exemplos que, em termos linguísticos, nos são "mais próximos", podemos concluir que é pouco eficaz o efeito prático da existência daquele glossário com a agravante de existirem várias sugestões de tradução que, naturalmente, obrigam a um esclarecimento pormenorizado do sentido que se pretende dar - valores extremos não serão sempre valores atípicos. No glossário em análise, nalguns casos, entenda-se nalgumas línguas, não há propostas de tradução para outliers - por exemplo, em dinamarquês, norueguês ou sueco. Devemos registar ainda que a "navegação" no glossário apenas permite a correspondência num sentido. Por exemplo, não conseguiremos facilmente concluir que a "valeurs aberrantes" da língua francesa deveria corresponder "valores de exceção" de um texto em português e que de outliers se tratava nessa pesquisa. Para resolver essa questão teremos sempre de passar pela palavra outlier - "intermédia e de ligação". Assim, a consulta do referido glossário consolida a opção pela não tradução, em palavras como outliers, embora seja muito útil na correspondência para outros termos estatísticos na língua de Camões (e aqui os exemplos serão muitos). Desenvolvendo esta questão podemos questionar sobre a vantagem de traduzir "Bootstrap" ou "p-value" ou outros "termos estatísticos" internacionalmente esclarecidos e por todos usados como pertencendo a uma "linguagem comum" dos estatísticos; com evidentes vantagens se for universal?

Sempre que, em qualquer texto em português, lemos "valores de exceção" ou "valores atípicos" ou "valores discordantes" - considerando apenas 3 alternativas - aparece (e é exigida?) a (necessária?) explicação de que aquela tradução corresponde a outliers. Assim, cada uma dessas expressões, num texto científico, não é mais do que um código de palavras que faz corresponder "valores aberrantes" a outliers para (apenas) um determinado texto e não para o mesmo autor que, noutro artigo, usa (ou pode usar) outra terminologia. Esta questão passa perto (ou será que não?) da polémica surgida em Portugal há quase 20 anos, entre os que defendem a "... obrigatoriedade do uso do português nas... dissertações..." como a proposta ao Governo e ao Conselho de Reitores das Universidades Portuguesas em 1988 pela Comissão Nacional da Língua Portuguesa e os que, no campo oposto, asseguram desde logo que, como primeira consequência, muito nefasta, essa será uma grave intromissão na autonomia universitária. A questão está em aberto e enquanto assim

estiver será decerto um alento para a ciência. O fundamental, de facto, é o dinamismo da investigação e a publicação científica em português, sem prejuízo e com o maior incentivo à sua internacionalização. A Sociedade Portuguesa de Estatística, nos debates, também já aflorou esta polémica e, em especial, registam-se os artigos publicados nos Boletins Informativos 2 e 3/99 e Jan/Abr 2000. Uma tão interessante quanto importante tarefa (que lhe cabe?) é a criação de um Dicionário de Estatística. É um grande desafio, dada a vastidão de assuntos e a variedade de termos e temas mas que tem garantido a divulgação enciclopédica da estatística, o que assegura o sucesso de uma obra com esse objectivo. Para já, com a certeza de que se aumenta a sistematização e, para que mais facilmente se possa concluir em que área se inclui um determinado artigo científico através de alguma das suas "palavras - chave", aceitemos que, em português, se escreva "outliers".[5]

Com a utilização desta opção, perante um texto, todos saberemos em que domínio se integra e, decerto, será muito mais fácil "buscar" artigos do nosso interesse científico; principalmente entre as publicações científicas compostas "na língua de Camões".

3.6 E o futuro?

Apesar da sua longa história, "o problema outlier" continua a despertar o maior interesse tanto do ponto de vista teórico como prático. Nos mais diversos campos e aplicações, sendo uma eventual explicação para a proliferação na terminologia da teoria dos outliers, as revistas científicas internacionais contêm cada vez mais contribuições nessa área de estudo. Vejam-se os mais importantes, por exemplo, em *Applied Statistics*, *Technometrics*, *Biometrika* ou *Journal of the American Statistical Association*. A investigação mais recente desenvolve ainda alguns métodos informais para pesquisa de observações discordantes em modelos estruturados e apresenta questões do maior relevo para amostras multivariadas.

Da etimologia da palavra estatística resulta que o seu uso (mais ou menos) sempre se associa à colheita e ao uso de dados de modo a apoiar a administração de um estado. O sistema de justiça é, na realidade, um dos pilares fundamentais de um moderno estado e é basilar na política da maior parte dos países. Metodologias probabilísticas já são usadas desde

[5] As razões invocadas para a utilização desta terminologia no tratamento estatístico de observações discordantes são, simultaneamente, linguísticas e de metodologia estatística em português. Assim, no texto que se segue, usaremos outliers no sentido tradicional e, escreveremos "outliers" sempre que desejarmos dar ênfase a esses argumentos.

o século XIV para modelar e apoiar a decisão na aplicação da justiça. Os mais recentes avanços da teoria dos "outliers" têm surgido baseados na inferência estatística para interpretar dados de um ponto de vista legal. Os tribunais estão introduzindo novos desafios para os estatísticos que assim são solicitados a pronunciar-se em domínios de trabalho não tradicionais - por exemplo a correcta aplicação da legislação envolvendo os direitos de autor ou, com muito maior impacto, as evidências bioestatísticas ou genéticas em determinada prova. Toda a prova admissível, e não apenas a prova científica, pode desempenhar um papel fundamental em tribunal. Torna-se aqui fulcral o termo "admissível". O "julgamento" feito por um estatístico poderá ser o apoio (também científico) na decisão do tribunal. São novos temas para a estatística e, por consequência, para a teoria dos "outliers". Este é decerto o mais recente desafio para os "estatísticos dos outliers" e que se vem juntar a alguns outros objectivos científicos ainda por atingir tais como os que envolvem as metodologias multivariadas e todos os que mais directamente se relacionam com questões de modelação estatística e inferência robusta. Esse desafio envolve a própria designação e terminologia pois se poderá seguir para a "nomo-estatística" (se optarmos pela etimologia do latim) ou "dicometria" (se usarmos as origens gregas), dando pois a possibilidade de, em breve, se começar a usar "nomo-outliers" ou dico-outliers". E esse futuro dos "outliers em tribunal" já começou. É bastante a referência histórica dos exemplos enunciados por Barnett e Lewis ([11], p. 4-7).

No futuro, cada vez mais, os "outliers" continuarão a ocupar um lugar do centro na ciência estatística e nos métodos estatísticos, pois sempre uma observação discordante será um desafio para o analista e dela poderá depender o seu relatório final para a mais importante tomada de decisão. Mas, quando tudo está dito e feito, o principal problema no estudo de observações eventualmente suspeitas, continua a ser aquele que desafiou os primeiros investigadores - O que é um "outlier" e como se deve trabalhar com essa observação? Nos próximos capítulos desenvolveremos, também, esta problemática.

3.7 Anexo

Segue-se uma Lista Bibliográfica, (obviamente incompleta) com "Outliers em português":

1. Almeida, C. (2001) - *Máxima Verosimilhança e Detecção de Outliers.* Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

2. Alpiarça, I. (1995) - *Outliers em Localização e Escala para Populações Exponenciais e Gama*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

3. Branco, J. (2005) - Estatística Robusta:Contribuição Portuguesa. *Memorial da Sociedade Portuguesa de Estatística*. Edições SPE.

4. Braumann, M.M. (1989) - *Testes de Discordância para "Outliers" - em Populações Normais e Gama*. Provas de Aptidão Pedagógica e Capacidade Científica. Universidade de Évora.

5. Braumann, M. M. (1994) - *Sobre Testes de Detecção de "Outliers" em Populações Exponenciais*. Dissertação de Doutoramento. Universidade de Évora.

6. Costa, S. (2005) - *Análise Estatística Multivariada na Segmentação de uma Companhia de Seguros*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

7. Jorge, A. (1999) - *Sobre a Definição de Outlier no Domínio Específico dos Modelos Lineares e Séries Temporais*. Dissertação de Mestrado. Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão.

8. Martins, S. (2000) - *Medidas de "Performance" em Modelos de Discordância Exponenciais*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

9. Mendes, Z. (1993) - *Modelos e Testes de Discordância para Outliers em Populações Normais*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

10. Oliveira, P. (1988) - *Tratamento Estatístico de "Outliers"*. Provas de Aptidão Pedagógica e Capacidade Científica. Universidade do Minho.

11. Palma, J. (1998) - *Outliers em Séries Temporais. Uma abordagem no domínio dos modelos ARMA*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

12. Palma, J. (2006) - *Medidas de Desempenho para os Testes de Discordância em Populações Normais*. Tese de Doutoramento. Universidade de Lisboa.

13. Passos, J. (1992) - *Influência das Observações nos Coeficientes Estimados no Modelo de Regressão Múltipla*. Dissertação de Mestrado.

Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão.

14. Figueira, M.M. (1995) - *Identificação de Outliers: uma aplicação ao conjunto das maiores empresas com actividade em Portugal*. Dissertação de Mestrado. Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão.

15. Rosado, F. (1982) - Análise Qualitativa de Densidades de Gram - Charlier e de Edgeworth como Modelos de Alternativas Inerentes para Outliers. *Nota nº 27 do CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa.*

16. Rosado, F. (1982) - Distribuições Assintóticas sobre Testes de Discordância para Outliers em Populações Exponenciais. *Nota nº 31 do CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa.*

17. Rosado, F. (1982) - Testes de Discordância para Outliers em Distribuições Exponenciais - Resultados Assintóticos. *Actas ds IX Jornadas Hispano - Lusas de Matemática*, vol. II, p.623-26.

18. Rosado, F. (1984) - The Null Distribution Function of Discordancy Tests for Outlier in Exponential Populations. *METRON, Rivista Internazioanale di Statistica*, vol.XLII, nº 1-2, p.51-7.

19. Rosado, F. (1984) - *Existência e Detecção de Outliers - Uma Abordagem Metodológica*. Tese de Doutoramento. Universidade de Lisboa.

20. Rosado, F. (1984) - Outliers a posteriori. *Actas do III Colóquio de Estatística e Investigação Operacional*, p.273-9.

21. Rosado, F. (1986) - Identificação de Outliers. Conferência apresentada no VII Simpósio Nacional de Probabilidades e Estatística na Universidade Estadual de Campinas (Brasil); pré-print na *Nota 31/86 do CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa.*

22. Rosado, F. (1987) - Outliers in Exponential Populations. *METRON, Rivista Internazionale di Statistica*, vol.XLV, n.1-2, p.85-91.

23. Rosado, F. (1987) - Algumas Reflexões sobre a Condição Outlier. *Actas das XII Jornadas Luso-Espanholas de Matemática*, volume III, p. 175-80.

24. Rosado, F. (1990) - Outliers, Inliers e Observações Influentes. *Actas das XV Jornadas Luso-Espanholas de Matemática* vol. IV p.227-229.

25. Rosado, F. (1996) - Detecção de Outliers e Análise em Componentes Principais. *Nota 2/96 do CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa.*

26. Rosado, F. (1996) - Testes de Discordância para Outliers e sua Dependência da Dimensão da Amostra in A Estatística a Decifrar o Mundo (R. Vasconcelos *et al* editores). *Actas do IV Congresso Anual da Sociedade Portuguesa de Estatística* p. 173-181. Edições Salamandra.

27. Rosado, F. (1997) - Sobre a influência de observações discordantes na modelação estatística. *Nota CEAUL 2/97 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

28. Rosado, F. (1997) - Sobre a Detecção de Outliers Utilizando Meios Computacionais. in Estatística: a diversidade na unidade (M. Souto de Miranda e I. Pereira editores). *Actas do V Congresso Anual da Sociedade Portuguesa de Estatística* p.199-206. Edições Salamandra.

29. Rosado, F. (1998) - Efeitos de uma Única Observação na Estimação de Parâmetros em Modelos Lineares. *Nota CEAUL 1/98 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

30. Rosado, F. (1998) - Outlier(s). Boletim Informativo da Sociedade Portuguesa de Estatística. Número Especial de Homenagem a Tiago de Oliveira, p.39-40.

31. Rosado, F. (2000) - A Note on Detection of Discordant Observations. *Nota CEAUL 18/2000 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

32. Rosado, F. (2000) - O que é a Estatística. Dossier Especial "Ano Mundial da Matemática". *Jornal Primeiro de Janeiro* de 2 de Outubro.

33. Rosado, F. (2001) - Outliers em Dados Estatísticos - o passado e o presente. E o futuro? *Actas do VII Congresso Anual da Sociedade Portuguesa de Estatística - Um Olhar sobre a Estatística*, p.90-110.

34. Rosado, F. (2001) - Using Maximum Likelihood in the Detection of Outliers. Comunicação apresentada na 53ª Sessão do ISI - International Statistical Institute realizada em Seoul (Coreia).

35. Rosado, F. (2005) - On the Statistical Interpretation of Outlier on Forensic Statistics. Comunicação apresentada na 55ª Sessão do ISI realizada em Sidney (Austrália).

36. Rosado, F. e Almeida, C. (2001) - Máxima Verosimilhança e a Detecção de Outliers. *Nota CEAUL 13/2001 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

37. Rosado, F. e Alpiarça, I. (1994) - Outliers Múltiplos em Modelos de Discordância para Populações Exponenciais. *Nota 7/94 do CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa.*

38. Rosado, F. e Alpiarça, I. (1994) - Sobre Modelos de Discordância para Outliers em populações Exponenciais e Gama. *Actas do II Congresso Anual da Sociedade Portuguesa de Estatística* p.67-71.

39. Rosado, F. e Braumann, M.M. (1990) - Critical Values for a Lower Outlier in a Gamma Sample. *METRON, Rivista Internazionale di Statistica*, vol.XLVIII, n.1-4, p.19-25.

40. Rosado, F. e Mendes, Z. (1993) - Sobre a Detecção de Observações Discordantes em Populações Normais. *Actas do I Congresso Anual da Sociedade Portuguesa de Estatística - A Estatística e o Futuro e o Futuro da Estatística* (p.271-286). Edições Salamandra.

41. Rosado, F. e Oliveira, I.(1996) - Selecção Multivariada de Outliers - Uma Aplicação às Variedades de Castanheiro em Trás-os-Montes. in A Estatística a Decifrar o Mundo (R. Vasconcelos et al editores). *Actas do IV Congresso Anual da Sociedade Portuguesa de Estatística p.163-173.* Edições Salamandra.

42. Rosado, F. e Palma, J. (2001) - Measures of Performance for Discordancy Tests in Normal Populations. *Revista de Estatística.* Contributed Papers, vol.II, 2º Quad. p.357-358.

43. Rosado, F. e Palma, J. (2001) - Sobre a Qualidade de Testes de Discordância em Populações Normais. *Nota CEAUL 12/2001 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

44. Rosado, F. e Palma, J. (2003) - Problemas e Limitações da Detecção de Outliers. *Nota CEAUL 2/2003 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

45. Rosado, F. e Palma, J. (2005) - Outliers em Dados Circulares. *Nota CEAUL 1/2005 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

46. Rosado, F. e Santos, E. (2000) - Outliers em Regressão Linear Simples - O Efeito de Alargamento. *Nota CEAUL 19/2000 do Centro de Estatística e Aplicações da Universidade de Lisboa.*

Capítulo 4

O Método Generativo com Alternativa Natural

4.1 Preliminares

Como já salientámos nos capítulos anteriores, para além dos critérios de rejeição de "outliers" fundamentados em processos empírico-intuitivos, todos os testes para análise de observações discordantes numa amostra que têm sido considerados pelos vários autores são, fundamentalmente, baseados no método da razão de verosimilhanças ou em propriedades localmente óptimas.[1]

Naturalmente, a construção de qualquer teste de discordância depende, em primeira análise, da hipótese alternativa formulada no modelo de discordância. Essa hipótese permite, de facto, introduzir a própria noção de outlier como observação surpreendente. Uma observação discordante para uma determinada alternativa poderá não o ser para outra.

4.2 Modelo de Discordância

Consideremos uma hipótese H_0 onde é assumida a não existência de observações discordantes, isto é, segundo a qual a amostra em estudo é extraída de uma determinada população onde F_0 é a distribuição de interesse e todas as observações são por ela geradas.

[1] Barnett e Lewis (Cf.[11]) apresentam um excelente resumo, principalmente numa perspectiva histórica, dos vários estudos sobre os princípios estatísticos em que se fundamentam os testes de discordância. Sobre este tema, embora pouco aplicado, mas pela originalidade matemática da abordagem, pode também consultar-se [26].

Na formulação de um *modelo de discordância*, para além de H_0 deve ser introduzida a hipótese H_1 de existência de "outliers" na amostra.

Diversas hipóteses alternativas - que condicionam a discordância de uma observação - têm sido consideradas.[2] A cada modelo alternativo corresponde uma situação bastante particular para os testes que têm sido formulados. De tal modo assim é que, para as várias hipóteses alternativas, os correspondentes modelos de discordância, não é possível apresentar testes de discordância que utilizem tais hipóteses em pleno, sem qualquer restrição, no que se refere às observações a testar. É, pois, fundamental a introdução de uma metodologia geral.

Para os objectivos a que nos propomos e, em jeito de resumo, podemos afirmar que, os testes de discordância mais estudados consideram modelos onde são formuladas hipóteses alternativas inerentes, por contaminação ou mistura, por deslizamento[3], por deslizamento indexado[4] e, com variáveis permutáveis (de origem bayesiana).

O método que definiremos na secção 4.4, apresentando uma nova abordagem ao estudo de "outliers", formula um modelo que podemos considerar geral, no sentido da não restrição pela hipótese alternativa que, em todos os outros, fixa a observação que vai ser analisada pelo teste.

Os modelos de discordância com *alternativas inerentes*, contrapõem genericamente duas distribuições para a população. Fixado um modelo básico com função de distribuição F_0, para a população, na ausência de "outliers" e, detectado/seleccionado um valor discordante, admite-se com uma alternativa inerente que, todas as observações seguem uma mesma lei F_1, sob a qual o "outlier" perde a condição de valor surpreendente. Na década de 70 do século passado, surgiram (Cf. [130])os primeiros testes de discordância onde são consideradas hipóteses alternativas inerentes para populações normais. De igual modo, (pode consultar-se [131]) foram também admitidos modelos de discordância exponenciais para a hipótese nula e para a respectiva alternativa inerente.

Os estudos referenciados apresentam e fixam modelos de discordância com distribuições da mesma família para a hipótese nula e para a respectiva alternativa inerente. Numa situação mais geral, pode fixar-se a não normalidade como hipótese alternativa inerente opondo-se a uma

[2]Para uma perspectiva global e também histórica, ainda actualizada, pode consultar-se o capítulo 2 de [11].

[3]Designação correspondente à tradução de "slippage alternative".

[4]Designação correspondente à tradução de "labelled slippage alternative" e utilizada, principalmente, em [11]. Adiante nas secções (4.3 e 4.4) deste capítulo será clarificada a distinção entre o deslizamento indexado e a alternativa natural.

hipótese nula de normalidade para a população. Nesse sentido e, admitindo ainda normalidade na hipótese nula, pode ser feita uma análise em modelos de discordância que consideram alternativas inerentes que possam "incluir" a distribuição na hipótese nula, por exemplo baseadas em densidades de Gram-Charlier do tipo A e de Edgeworth.

Estaríamos assim, criando modelos menos restritivos no sentido de, na hipótese alternativa serem contempladas um maior número de distribuições inerentes. Neste contexto, Rosado [109] provou que, não são admissíveis tais hipóteses alternativas inerentes para a distribuição normal, do ponto de vista de "outliers" em extremos.

Nos modelos de discordância formulados com hipóteses *alternativas por contaminação*, a uma distribuição fixada pela hipótese nula H_0, opõe-se uma distribuição $(1-\theta)\,F+\theta\,G$, contaminação de duas funções de distribuição F e G que permite justificar a presença de "outliers" e sendo θ o coeficiente de contaminação que introduz no modelo as possíveis observações contaminantes[5] vindas da população G.

Formulado o modelo de discordância

$$H_0 : F \qquad vs \qquad H_1 : (1-\theta)\,F+\theta\,G$$

a alternativa por contaminação, opondo duas distribuições, torna-se semelhante, mas não igual (em virtude da presença do parâmetro θ), a uma alternativa inerente. A grande maioria dos testes de discordância inicialmente propostos, consideram hipóteses alternativas de contaminação. De igual modo, grande parte da teoria da acomodação até hoje considerada, pressupõe modelos alternativos de contaminação. Nesse sentido, também do ponto de vista histórico, ainda deve considerar-se fundamental o estudo apresentado em [71].

Os modelos de discordância que utilizam alternativas por deslizamento representam uma evolução na perspectiva da moderna teoria dos "outliers". Estes modelos são baseados, fundamentalmente, nos modelos A e B formulados nos trabalhos de Ferguson (consultar [44] e [45]). Os modelos A e B foram construídos de modo a permitir justificar a presença de uma observação discordante e, em síntese, o modelo A admite que foi originada por uma translação enquanto que no modelo B é suposto que houve um aumento na variância da observação "aparentemente errónea". Os modelos de discordância formulados por Ferguson podem, em certo sentido, ser considerados bastante gerais no que respeita às hipóteses alternativas que apresentam, embora com a restrição de ser fixada a observação que deve ser testada. Fundamentando o trabalho, motivado por

[5] É fundamental o esclarecimento da distinção entre observação contaminante e "outlier".

Grubbs [55] e Dixon [35], Ferguson (Cf. [44] e [45]) admite que:

"... *The problem is to introduce some degree of objectivity into the rejection of the outlying observations*".

Nesse sentido, "... *to give a structure to the outlier problem*", Ferguson propõe que se considerem hipóteses alternativas por deslizamento em distribuições normais (única situação considerada) formulando os seguintes modelos:

MODELO A (efeito em μ) - $x_1, ..., x_n$ vêm independentemente de populações normais com variância comum σ^2. Sob a hipótese nula H_0 têm um valor médio comum μ. Há constantes conhecidas $a_1, ..., a_n$ (a maioria das quais são zero), um parâmetro desconhecido Δ e uma permutação desconhecida $(\nu_1, ..., \nu_n)$ de (1,2,...,n) tal que a distribuição normal correspondente a x_i tem valor médio $\mu_i = \mu + \sigma\Delta a_{\nu_i}$ (i=1,2,...,n). A alternativa $\bar{H}$ admite $\Delta \neq 0$ (ou só $\Delta > 0$ quando os a_i forem todos positivos).

MODELO B (efeito em σ) - $x_1, ..., x_n$ são observações independentes e normais com valor médio comum μ. Sob a hipótese nula H_0 têm variância comum σ^2. Há constantes conhecidas $a_1, ..., a_n$ (algumas serão nulas), um parâmetro desconhecido Δ e uma permutação desconhecida $(\nu_1, ..., \nu_n)$ de (1,2,...,n) tal que a distribuição normal de x_i tem variância $\sigma_i^2 = \sigma^2\ exp\ (\Delta a_{\nu_i})$ (i=1,2,...,n). Pela alternativa $\bar{H}$ é fixado $\Delta > 0$.

É fundamental salientar a restrição, $\Delta > 0$, que é introduzida no MODELO B. Ela, de facto, determina *a priori* a observação que deve ser analisada.

A propósito da definição desse modelo, (Cf. [44], p. 268) Ferguson assume que, pela hipótese nula H_0 se fixa $\Delta = 0$, enquanto que, pela alternativa $\bar{H}$, se determina que o problema outlier tem como origem apenas observações com maior variância pelo que, as constantes a_i não nulas serão positivas.

Por sua vez, Barnett e Lewis seguem a mesma metodologia, formulando um modelo análogo (Cf. [11], p. 49) consideram que $\Delta < 0$ é irrelevante para o estudo de observações discordantes numa amostra.

Este é o "ponto da situação".

Esta metodologia, embora correspondendo a uma perspectiva bastante geral no que respeita à selecção de "outliers" em dados estatísticos ainda apresenta restrições que se reflectem na subjectividade do estudo.

O método geral que introduziremos neste capítulo permite, de facto, avançar com inovação não limitando *a priori* a hipótese alternativa. No capítulo 7, estudando "outliers" em populações normais abordaremos de

novo esta questão e verificaremos que, a hipótese $\Delta < 0$ não deve ser excluída.

Os MODELOS A e B acima considerados são, no entanto, bastante gerais no sentido de que permitem a sua utilização em testes de discordância sem qualquer restrição quanto ao número de observações suspeitas na amostra e ainda, não sendo restritivos quanto à dimensão das variáveis aleatórias. Estes modelos[6] foram utilizados, alguns anos mais tarde (ver, por exemplo, [129]) também para populações normais multivariadas.

Algumas tentativas (consultar [73] e [135]) de generalização destes modelos de discordância por deslizamento indexado na observação x_i, foram feitas introduzindo hipóteses com variáveis permutáveis. No seu trabalho, Kale e Sinha [73] consideram a estimação da vida média de uma população com base numa amostra onde um outlier está presente. Assim, formulam um modelo de discordância onde, como alternativa, admitem que (n-1) das observações seguem uma mesma distribuição fixada pela hipótese nula, enquanto que a outra observação tem uma vida média muito superior. É, além disso, suposto que o outlier pode ser qualquer uma das observações da amostra e não sendo previamente conhecida a observação discordante. Anos mais tarde, Barnett e Lewis [11], apresentam esta mesma situação, definindo assim o modelo de discordância com alternativa onde consideram varaiáveis permutáveis - as observações do modelo de Kale e Sinha (Cf. [73]) que entretanto também tinha sido utilizado por Joshi [72]. Temos então, também do ponto de vista histórico, uma perspectiva de construção de modelos de discordância.

Uma generalização diferente dos modelos de discordância com hipóteses *alternativas por deslizamento* é introduzida na secção seguinte, onde formalizaremos um modelo com alternativa por deslizamento indexado, tornando mais simples a exposição desse modelo. Além disso, este mesmo exemplo introdutório permitirá apresentar uma primeira motivação para a construção do método generativo com alternativa natural que propomos para tratamento e estudo de "outliers" em dados estatísticos.

4.3 Exemplo introdutório

Como referimos na secção anterior, uma das bases estatísticas para a construção de testes de discordância para "outliers" é o princípio da razão de verosimilhanças. Porque o estudo que apresentamos se baseia em critérios de máxima verosimilhança e porque, além disso, as hipóteses

[6]No capítulo 10, abordaremos a problemática específica de um estudo de outliers em dados estatísticos multivariados.

formuladas se interligam com as alternativas por deslizamento, muito em especial com deslizamento indexado, apresentamos o seguinte exemplo, de começo preliminar, para a construção de um teste de discordância em populações exponenciais e que utiliza aquela formulação.

Consideremos uma amostra aleatória simples $x_1, ..., x_n$ com dimensão n, de uma população exponencial com densidade probabilidade

$$f(x; \lambda, \delta) = \frac{1}{\delta} \exp\left(-\frac{x-\lambda}{\delta}\right) \qquad x \geq \lambda. \tag{4.1}$$

Seja $x_{(1)}, ..., x_{(n)}$ a correspondente amostra ordenada. Suponhamos conhecido o parâmetro de localização $\lambda = \lambda_0$. Por translação podemos então admitir que, sem perda de generalidade, a população tem o parâmetro de localização $\lambda = 0$.

4.3.1 Modelo de Discordância

De acordo com a hipótese H_0 de não existência de "outliers", admitimos que todas as observações $x_1, ..., x_n$ pertencem a uma distribuição exponencial com função densidade de probabilidade $f(x; 0, \delta)$, como introduzida em (4.1). A correspondente hipótese alternativa $\bar{H}$ por *deslizamento do parâmetro de escala* δ, admite que $(n-1)$ daquelas observações têm função de densidade $f(x; 0, \delta)$ enquanto que a outra observação que designaremos por x_j, tem densidade $f(x; 0, k\delta)$.

O modelo de discordância que, como exemplo apresentamos, segue, muito de perto, a formulação de Barnett e Lewis (Cf. [11], p. 98-101) e, nesse sentido, vamos também fixar $k > 1$. Vamos portanto colocar-nos naquela que apelidamos de perspectiva tradicional para o estudo de outliers em dados estatísticos. Aliás, o exemplo que estamos construindo vai permitir, por um lado, fazer um estudo comparado entre aquela e o nosso novo método para estudo de "outliers", no que respeita à construção de testes de discordância e, por outro, de modo imediato, verificar que determinadas situações (da exponencial fixando $k > 1$, por exemplo) resultam como casos particulares do estudo que propomos.

Admitindo a formulação geral por deslizamento do tipo B apresentada por Ferguson e acima descrita, o modelo de discordância que estamos construindo fixa x_j como observação discordante determinada pela hipótese alternativa $\bar{H}$. De acordo com este modelo, a verosimilhança da amostra, na hipótese alternativa é

$$L(x_1, ..., x_n; \delta, k) = \frac{1}{k\delta^n} \exp\left(-(\frac{n\bar{x} - x_j}{\delta} + \frac{x_j}{k\delta})\right)$$

sendo x_j - o valor discordante na amostra - fixado pela alternativa por deslizamento.

Na hipótese de não existência de "outliers" ($k = 1$), a verosimilhança será então $L(x_1, ..., x_n; \delta, 1)$.

Os estimadores $\widehat{\delta}_0$ e $\widehat{\delta}_1$, de máxima verosimilhança para o parâmetro δ, na hipótese H_0 de não existência de "outliers" e na alternativa $\bar{H}$ por deslizamento de x_j são, respectivamente,

$$\widehat{\delta}_0 = \bar{x}$$

e

$$\widehat{\delta}_1 = \bar{x} - \frac{(k-1)x_j}{kn}.$$

4.3.2 Teste de Discordância

Para construir este teste, vamos utilizar o critério da razão de verosimilhanças para decidir sobre a discordância de x_j. Devemos pois calcular o quociente

$$l_n = \frac{max_{H_0}\ L(x_1, ..., x_n; \delta, 1)}{max_{H_0 \cup \bar{H}}\ L(x_1, ..., x_n; \delta, k)}. \tag{4.2}$$

Com a utilização de $\widehat{\delta}_0$ e $\widehat{\delta}_1$, obtemos

$$max_{H_0}\ L(x_1, ..., x_n; \widehat{\delta}_0, 1) = \frac{exp\ (-n)}{\bar{x}^n} \tag{4.3}$$

$$max_{\bar{H}}\ L(x_1, ..., x_n; \widehat{\delta}_1, k) = \frac{exp\ (-n)}{k\ (\bar{x} - \frac{(k-1)x_j}{kn})^n}. \tag{4.4}$$

Se usarmos os máximos em (4.3) e (4.4) então (4.2) pode-se escrever

$$l_n = \frac{\frac{exp\ (-n)}{\bar{x}^n}}{max_j\ (\frac{exp\ (-n)}{\bar{x}^n}, \frac{exp\ (-n)}{k(\bar{x} - \frac{(k-1)x_j}{kn})^n})} \tag{4.5}$$

uma vez que o cálculo em todo o espaço exige a verificação para todos os índices das observações da amostra. A simplificação dos cálculos em (4.5) conduz a

$$l_n = \frac{1}{max_j\ (\ 1\ ;\ \frac{1}{k(1-\frac{(k-1)x_j}{kn\bar{x}})^n})}$$

$$= min_j\ (\ 1\ ;\ k\ (1 - \frac{(k-1)x_j}{kn\bar{x}})^n). \tag{4.6}$$

Da definição do quociente l_n resulta imediatamente

$$0 \ \leq \ l_n \ \leq 1$$

e, além disso, para grandes valores de l_n, estaremos numa situação a que deve corresponder a aceitação da hipótese H_0 de não existência de "outliers", uma vez que, nesse caso, os máximos no numerador e no denominador de l_n em (4.2) serão "aproximadamente" iguais.

Para a hipótese H_0, podemos então construir uma região de rejeição

$$l_n \ < \ c \ (< 1)$$

ou, tenda em conta (4.6),

$$min_j\ (\ 1\ ;\ k\ (1 - \frac{(k-1)x_j}{kn\bar{x}})^n) \ < \ c \ < \ 1.$$

Porque $c < 1$, temos então a seguinte regra de decisão para estudo de uma observação discordante:

$$min_j\ (\ k\ (1 - (1 - \frac{1}{k})\frac{x_j}{n\bar{x}})^n) \ < \ c. \tag{4.7}$$

4.3.3 Região de Rejeição

Uma vez que no modelo de discordância formulado por deslizamento na observação x_j, se fixou $k > 1$, o teste da razão de verosimilhanças determina em (4.7) que poderá ser utilizada a estatística $x_j / \sum_i x_i$ para estudar a observação $max\ x_i = x_{(n)}$ como "outlier". A consequente região de rejeição para $x_{(n)}$ será então:

$$\frac{x_{(n)}}{\sum_i x_i} \ \geq \ c' \tag{4.8}$$

Assim, a regra de teste (4.8) permite decidir sobre o máximo $x_{(n)}$ da amostra de uma população exponencial que foi seleccionado *a priori* uma vez que é correspondente a uma observação com deslizamento no

parâmetro de escala $k > 1$; supondo conhecido o parâmetro de localização, igual para todas as observações. Esta é a situação prática tradicional (Cf. [11], p. 99). Nela, deseja construir-se um teste de discordância para o estudo de uma determinada observação que causou surpresa ao investigador; no caso vertente, nomeadamente, o máximo $x_{(n)}$.

O **exemplo introdutório** que acabamos de construir, independentemente do estudo que, no capítulo seguinte, faremos sobre distribuições exponenciais, é muito importante, também do ponto de vista histórico na teoria dos "outliers" porque, por um lado, formaliza uma hipótese alternativa por deslizamento indexado introduzida para justificar a surpresa de $x_{(n)}$ e, por outro, levanta (e confirma!) a importante questão da subjectividade no estudo de observações discordantes.

Este exemplo, por raciocínio análogo, conduziria ao estudo do mínimo $x_{(1)}$ da amostra se fixássemos $k < 1$.

O método generativo com alternativa natural, que é objectivo principal deste capítulo e que definiremos na secção seguinte permite, como veremos, que se possa considerar, como caso particular, o exemplo introdutório acima construído na medida em que ele levará à escolha do máximo $x_{(n)}$ (ou ao mínimo $x_{(1)}$) sempre que, em termos de máxima verosimilhança essa observação ($x_{(1)}$ *ou* $x_{(n)}$) deva ser estudada como discordante. Note-se e saliente-se desde já que, nesta metodologia se admite que eventualmente nem uma nem outra dessas observações seja estudada como "outlier".

Será portanto eliminada a influência do factor de proporcionalidade k - que não será fixado. Neste sentido é introduzida objectividade no estudo de "outliers" em dados estatísticos.

4.4 O Método GAN

4.4.1 Generalidades

Por comodidade de simbologia e maior facilidade na exposição do método vamos admitir apenas um parâmetro δ (de dispersão, por exemplo) na função densidade de probabilidade

$$f(x;\delta) \;=\; \frac{1}{\delta}\; f\;(x/\delta) \tag{4.9}$$

para a população.

A generalização do método para populações com função densidade com mais de um parâmetro, como veremos, é imediata e sem qualquer

dificuldade de aplicação. Não é pois restrição no método, a consideração de um só parâmetro.

Como se sabe, uma "alteração" na dispersão δ significa, na maior parte das vezes, mudança de aparelho experimental e/ou de observador enquanto que, uma "mudança" no parâmetro de localização pode sugerir alterações no fenómeno em estudo. Assim, consideraremos na generalidade um parâmetro de dispersão δ em $f(x, \delta)$ por poder ser, o aparecimento de "outliers", devido fundamentalmente à maior ou menor dispersão de uma dada observação na amostra. Além disso, estaremos também a incluir a maioria das situações práticas que, bem poucas vezes, apresentam valores discordantes devidos (apenas) a um parâmetro de localização.

Existem muitas situações práticas - formalizadas na maior parte dos estudos conhecidos - e que motivam os diversos testes de discordância para "outliers" cuja distribuição se considera ter uma dispersão proporcional à das restantes observações na amostra. Concretamente, a maioria dos trabalhos apresentados sobre testes de discordância para "outliers" em dispersão, consideram modelos onde se admite coeficiente δ para (n-1) das observações e k δ para a observação discordante que, entretanto fica fixada *a priori* quando é condicionado o valor de k, quer seja superior ou inferior a 1. Historicamente, um primeiro avanço, nesta problemática foi feito por Fieller em 1976, utilizando a formalização com modelo por deslizamento indexado (Cf. [46]). Esta mesma perspectiva foi, como referimos anteriormente, também seguida por Barnett e Lewis [11] embora com a restrição de se ver cada estatística de teste a ser justificada para ser usada para testar esta ou aquela observação fixada *a priori*. Isto é, em primeiro lugar o experimentador suspeita de uma observação e, em seguida, usa o teste que é fornecido para estudar a observação seleccionada. Assim, apenas são analisadas as observações "de que o analista suspeita". É, de facto, uma restrição na teoria geral para o estudo de "outliers" em dados estatísticos.

O método generativo com alternativa natural, que adiante definiremos, produz uma metodologia mais geral e, portanto, menos restritiva! Na sequência do exemplo introdutório que acima utilizámos, o método considera um parâmetro de dispersão δ' para uma possível observação discordante e δ para as restantes (n-1) observações. É uma abordagem mais geral do que aqueloutra de, digamos, δ e kδ. Teremos oportunidade de fazer um estudo comparativo entre as duas formulações.

4.4.2 Definição

O método generativo com alternativa natural - adiante designado por método GAN - é fundamentado em princípios de máxima verosimilhança.

O método GAN para pesquisa, selecção e tratamento estatístico de "outliers" consta das três fases seguintes:

- Formulação do modelo de discordância natural.
- Teste de homogeneidade da amostra.
- Selecção objectiva do (ou dos) "outlier(s)".

Depois de numa primeira fase se construírem as hipóteses do modelo e onde se assume a *alternativa natural*, o nosso *método generativo* propõe um *teste de homogeneidade* das observações onde é tomada a decisão sobre a existência de "outliers" na amostra.

A amostra é homogénea sempre que as observações forem geradas pela mesma distribuição, isto é, quando o modelo que assumimos está correcto e portanto não existem "razões estatísticas" para aceitar a presença de valor(es) discordante(s).

Se no teste de homogeneidade for decidida a aceitação, o método termina nessa segunda fase negando, portanto, a existência de qualquer observação discordante. Pelo contrário, a rejeição da homogeneidade das observações conduz à terceira fase do método onde é seleccionada a observação que, de acordo com este critério objectivo, deve ser então declarada "outlier". Assim, na segunda fase do método, temos a decisão fundamental sobre a eventual existência de "outlier" que, apenas na fase seguinte poderá ser seleccionado. Temos, portanto, uma selecção *a posteriori*. Contrariamente aos "métodos tradicionais", este estudo que propomos, selecciona o "outlier" apenas na última fase e só após uma decisão sobre algo que vai "estatisticamente mal" na geração dos dados - a homogeneidade. De facto, tradicionalmente o "outlier" é seleccionado *a priori* e só depois é "usado" um teste de discordância para decidir sobre se essa observação - previamente suspeita aos olhos do investigador - deve ser considerada discordante. A selecção do valor eventualmente discordante é então fortemente condicionada pela experiência do investigador. Diferentes analistas poderão suspeitar de diversos valores.

Modelo de discordância natural

Consideremos uma amostra aleatória simples $x_1, ..., x_n$, com dimensão n, duma população com funções densidade de probabilidade dependentes de parâmetros δ_i, $f(x; \delta_i)$, $(i = 1, ..., n)$.

A verosimilhança da amostra é $L(x_1, ..., x_n; \delta_1, ..., \delta_n) = \prod_i^n f(x_i; \delta_i)$.

De um ponto de vista teórico, podemos admitir que todas as variáveis aleatórias em estudo têm densidades com diferentes parâmetros.

De um ponto de vista prático - o que mais interessa para uma metodologia de pesquisa de "outliers" - essa formulação apenas complica, e eventualmente torna impossíveis, os cálculos envolvidos na estimação e testes estatísticos de que necessitamos. Além disso, a diversidade de valores para os parâmetros δ_i, não permitiria ser coerente com algum mecanismo de geração dos valores discordantes que, evidentemente, devem "todos" ser oriundos da mesma distribuição; pois, no caso contrário, o caminho a seguir é a reformulação de todo o problema em estudo. Portanto, naquela hipótese a generalidade entra em conflito com a aplicabilidade.

Assim, no método generativo com alternativa natural, utilizamos o seguinte modelo de discordância:

- Pela hipótese H_0 - de *homogeneidade* - admitimos, como aliás na generalidade dos estudos sobre "outliers", que todas as observações $x_1, ..., x_n$ têm a mesma densidade $f(x_i; \delta)$, $(i = 1, ..., n)$. Nesta hipótese H_0, a verosimilhança é

$$L(x_1, ..., x_n; \delta) = \prod_i^n f(x_i; \delta).$$

- Pela hipótese $\bar{H}$ - a *alternativa natural* - admitimos a presença de um valor discordante na amostra[7] e que, em princípio, pode ser uma qualquer das n observações. Seja então $\bar{H}_j$ a hipótese que admite x_j como observação discordante, isto é, tal que:

  - x_j tem densidade de probabilidade $f(x_j; \delta')$, para algum índice j∈(1,...,n)
  - $x_1, ..., x_{j-1}, x_{j+1}, ..., x_n$ seguem a mesma distribuição com densidade $f(x_i; \delta)$ para $i \neq j$.

Se for válida a hipótese alternativa natural $\bar{H}$, existindo portanto uma observação discordante x_j isso significa que, pelo modelo, é assumida a hipótese $\bar{H}_j$ responsável pela geração desse "outlier". Sendo assim, a hipótese alternativa natural $\bar{H}$ pode-se considerar como uma reunião das n hipóteses $\bar{H}_j$ acima enunciadas. Neste sentido tem-se $\bar{H} = \cup_{j=1}^n \bar{H}_j$.

[7] Embora fixando apenas um valor (eventualmente) discordante, vemos que este modelo é facilmente generalizável para uma formulação onde, em termos gerais, se podem admitir k "outliers" na amostra.

Na alternativa natural, sendo assumida a hipótese $\bar{H}_j$, temos a verosimilhança

$$L_{\bar{H}_j}(x_1, ..., x_n; \delta, \delta') = \prod_{i \neq j} f(x_i; \delta)\ f(x_j; \delta').$$

Se representarmos por $\widehat{\delta}$ o estimador de máxima verosimilhança para δ na hipótese H_0 e por $\widehat{\delta}_j$ *e* $\widehat{\delta}'_j$ os estimadores de máxima verosimilhança para δ e δ' na hipótese $\bar{H}_j$ temos, para máximos dessas funções, sob H_0 e $\bar{H}_j$, respectivamente $\widehat{L}_0\ (x_1, ..., x_n; \widehat{\delta})$ e $\widehat{L}_j\ (x_1, ..., x_n; \widehat{\delta}_j, \widehat{\delta}'_j)$ onde, como vimos, o índice j é representativo de que foi admitida a observação x_j como discordante. Estes dois máximos, para simplificação de escrita, serão a seguir representados por $\widehat{L}_0$ e $\widehat{L}_j$.

Teste de homogeneidade

Para a formulação de um *teste de homogeneidade* na amostra $x_1, ..., x_n$ em estudo, comecemos por construir o quociente dos máximos das funções de verosimilhança

$$l_n = \frac{\max\limits_{H_0} L(x_1, ..., x_n; \delta)}{max_{H_0 \cup \bar{H}}\ L(x_1, ..., x_n; \delta, \delta')}.$$

Porque, como foi referido, a hipótese alternativa natural $\bar{H}$ se pode considerar uma união de n hipóteses $\bar{H}_j$, então pode escrever-se

$$l_n = \frac{\widehat{L}_0}{max\ (\widehat{L}_0, max_j \widehat{L}_j)}. \tag{4.10}$$

A razão de verosimilhanças l_n é fundamental neste estudo, por um lado, para a construção da regra de teste de H_0, de homogeneidade das observações e, por outro, na fase seguinte, para seleccionar o valor discordante que, se tal for o caso, deve ser considerado como responsável pela rejeição dessa hipótese.

Construamos então o teste de homogeneidade da amostra. Da própria definição do quociente l_n, resulta imediatamente que

$$0 \ \leq \ l_n \ \leq \ \ 1$$

e que, além disso, para grandes valores de l_n, estaremos na zona de aceitação da hipótese de homogeneidade formulada para as observações $x_1, ..., x_n$ uma vez que, nesse caso, o máximo no denominador de (4.10)

será "aproximadamente" igual a $\widehat{L}_0$, permitindo concluir que todas as observações seguem a mesma distribuição com densidade $f(x;\delta)$.

Assim, com base em princípios de verosimilhança máxima, podemos formular a regra de teste

$$l_n < c \quad (com \;\; c < 1) \tag{4.11}$$

A constante c, a calcular, "separa" as duas hipóteses em estudo e permite construir uma região de rejeição para H_0 - testa a homogeneidade. Nessa região estão incluídas todas as amostras $x_1, ..., x_n$ para as quais se verifica a desigualdade anterior.

Para determinar a região de rejeição, a partir de (4.10) podemos obter

$$l_n = \frac{1}{max\ (1, T(x_1, ..., x_n))}$$

onde

$$T(x_1, ..., x_n) \;=\; \frac{max_j\ \widehat{L}_j}{\widehat{L}_0}. \tag{4.12}$$

A região de rejeição que se retira de (4.11) exige, portanto, que

$$\frac{1}{max\ (1, T(x_1, ..., x_n))} \;<\; c.$$

Porque $c < 1$, então deve ter-se

$$T(x_1, ..., x_n) \;>\; c' \tag{4.13}$$

com $c' = 1/c$.

Se, para uma amostra $x_1, ..., x_n$, a correspondente estatística de teste $T(x_1, ..., x_n)$ *não verifica* a condição (4.13) então, o método generativo com alternativa natural conduz à aceitação da homogeneidade dessa amostra, negando portanto a existência de qualquer observação discordante e terminando, nesta segunda fase, a pesquisa de "outliers".

Selecção do "*outlier*"

Supondo rejeitada a hipótese de homogeneidade, aceitando portanto a existência de uma observação discordante, somos conduzidos, pelo próprio método, ao valor x_j que deve ser declarado "outlier". Neste sentido, o *método é generativo* porque, na última fase, "faz aparecer" a observação responsável pela não homogeneidade da amostra em estudo.

O "outlier" que, *a posteriori*, deve ser seleccionado é aquela observação que corresponde ao índice j onde a estatística $T(x_1, ..., x_n)$ atinge o máximo, conforme referido em (4.12). Esta observação "outlier" assim descoberta é, vulgarmente, o máximo ou o mínimo da amostra.[8]

4.5 Exemplo introdutório (continuação)

Como vimos, na secção 4.3, neste exemplo introdutório, já encontrámos a região

$$\min_j \left(k \left(1 - (1 - \frac{1}{k})\frac{x_j}{n\bar{x}}\right)^n\right) < c \tag{4.14}$$

para decidir sobre a rejeição da homogeneidade na amostra.

A mesma estatística de teste pode ser encontrada por aplicação do método GAN neste exemplo introdutório para populações exponenciais.

Se admitirmos alguma informação suplementar sobre k, de modo que possamos, por exemplo, garantir que $k > 1$, então a região anterior é equivalente a

$$max_j \frac{x_j}{\sum_i x_i} > c'$$

sendo portanto, na terceira fase do método GAN, seleccionado o máximo $x_{(n)}$ como "outlier". Assim, podemos verificar que já estamos numa situação mais geral do que aquela que foi assumida no exemplo introdutório. De facto, no exemplo introdutório, o modelo e, portanto, o teste foram construídos para "justificar" o máximo. Porque assumimos informação suplementar sobre k, que mantemos desconhecido, o método GAN selecciona *a posteriori* esse mesmo máximo.[9]

Se, ao contrário, soubermos que $k < 1$, a região (4.14) para rejeição da homogeneidade na amostra é equivalente a

$$min_j \frac{x_j}{\sum_i x_i} < c''$$

sendo, neste caso, encontrado o mínimo $x_{(1)}$ como "outlier", na última fase do método GAN.

[8] A discordância estatística de valores numa amostra pode não se manifestar apenas nos extremos. Nalgumas situações que estudaremos, por exemplo no capítulo 7, podemos encontrar outros valores discordantes que não esses.

[9] A estatística $x_{(n)}/(\sum_i x_i)$, aqui encontrada, é o principal - e muitas vezes único - instrumento de trabalho para o estudo tradicional do máximo de uma amostra de uma população exponencial (Veja-se, por exemplo, [11] p. 195).

Na sequência deste exemplo introdutório, no capítulo 5, em termos gerais, será desenvolvido um estudo de "outliers" em dados estatísticos para amostras de populações exponenciais.

4.6 Considerações sobre o Método GAN e apresentação de algumas propriedades

O método GAN para análise de "outliers" em dados estatísticos, que foi apresentado neste capítulo, também como nos estudos tradicionais, permite testar uma determinada observação x_j como discordante mas desta vez, com a indicação de qual o índice j a considerar e, por consequência, indigitando uma "observação-candidata".

Como vimos, os métodos tradicionais para tratamento de um valor discordante numa amostra pressupõem sempre que, a observação aberrante a ser testada, é um extremo - máximo ou mínimo. Essa observação é seleccionada *a priori* pelo analista e, portanto, podemos afirmar que o correspondente teste de discordância "serve apenas" para estudo dessa observação extrema. Suspeita-se de um valor e, em seguida, escolhe-se o teste apropriado para a confirmar como outlier. Trata-se então de um teste *ad hoc* que tem de ser justificado[10] porque - e em que situações - é interessante. Esta dificuldade continua e é agravada quando se estudam "outliers" múltiplos. Neste caso a metodologia tradicional também conduz à escolha de extremos - o máximo e o mínimo da amostra, no caso de "outliers" duplos. Mas, serão estes valores admissíveis? Mais adiante abordaremos de novo esta questão.

No método GAN, como vimos, a observação que deve ser declarada "outlier", se tal for o caso, é definida *a posteriori*. Esta propriedade torna-se particularmente importante na pesquisa e tratamento de algum "outlier" inesperado, isto é, de um valor que não surpreenda o estatístico. Qual a razão que leva a estudar o máximo em vez de testar o mínimo de uma amostra? Além disso, o grau de surpresa provocada por qualquer dado estatístico é uma forte condicionante do estudo. Desde os anos sessenta do século passado[11] que, sempre os investigadores se preocuparam, com a natureza subjectiva dos procedimentos utilizados para a rejeição de outliers.

No método GAN é retirada a "tradicional subjectividade" na selecção da observação a testar como discordante. Esta importante propriedade

[10] Sobre este importante assunto na teoria outliers, veja-se o capítulo 4 do livro de Barnett e Lewis definindo os princípios e os critérios para a construção de testes de discordância. (Cf. [11] p. 94-121).

[11] (Cf. [28]).

pode ser invocada como contribuição para a resolução do mais velho e simultaneamente o maior problema em todo o estudo de "outliers".

De facto, como afirmam[12] Barnett e Lewis, no seu excelente tratado sobre o estudo de outliers em dados estatísticos:

"*... when all is said and done, the major problem in outlier study remains the one that faced the very earliest workers in the subject - what is an outlier?*"

Por sua vez, o trabalho de Beckman e Cook [16], salientando esta mesma dificuldade e que foi discutido por Barnett [7], Mc Culloch e Meeter [80], Hawkins [67], Prescott [105], Hogg [70] e Drapper [39] é uma boa referência, também histórica, sobre esta problemática.

Com efeito, na introdução ao seu estudo, Beckman e Cook, depois de apresentarem uma definição, não única, de outlier mas evolutiva no estudo, salientam a antiguidade deste problema e concluem:

"*Although much has been written, the notion of an outlier seems as vague today as it was 200 years ago*".[13]

Sobre esta questão pode também consultar-se o estudo de Muñoz-Garcia *et al* [89].

Outra importante propriedade do método GAN relaciona-se com as observações candidatas a outliers. Assim, como vimos anteriormente, também nos modelos de discordância tradicionais para o "estudo" de outliers, pelo observador, são fixadas observações extremas como eventualmente discordantes e que devem ser testadas. Relevando esta questão, diversos autores afirmam que, pelo estatístico, devem ser estudados aqueles valores que parecem demasiado grandes ou demasiado pequenos quando comparados com as restantes observações. Numa síntese desta perspectiva, Gumbel [57] afirma:

"*The outliers are values which seem either too large or too small as compared to the rest of the observations. Thus, they are extremes*".

Na maioria das situações que estudaremos nos capítulos seguintes em pormenor, a observação discordante, se existir, seleccionada pelo método GAN, é também um extremo da amostra; mas nem sempre, necessariamente, assim será. Encontraremos exemplos, em modelos de discordância natural, para populações normais.
Além disso pode, também e desde já, colocar-se a questão da distinção entre os extremos, isto é, quando é que um máximo se torna "demasiado máximo" ou um mínimo é "demasiado mínimo"? E ainda, quando é que, numa amostra, o "grau de surpresa" do máximo é maior (ou menor) do

[12](Cf. [11], p. 459).
[13](Cf. [16], p.120).

que o "grau de surpresa" do mínimo? O método GAN permite estudar candidatos a "outlier" que tradicionalmente não são considerados. Neste sentido, o método é generativo.

O método GAN - metodologia para estudo de "outliers" em dados estatísticos - introduzido na secção 4.4 permite-nos, em síntese, clarificar a própria noção de observação discordante. Podemos, então, concretizar a seguinte:

Definição[14]

"Outlier" numa amostra de dados estatísticos é a observação que, perante o modelo de discordância natural formulado e após rejeição da homogeneidade, na terceira fase do método GAN, for seleccionada como responsável por essa decisão.

No método GAN existe uma implicação entre a rejeição da homogeneidade das observações e a decisão sobre a detecção de "outliers". Portanto, estes, só existem em amostras não homogéneas. E, ao contrário, a presença de "outliers" numa amostra, é razão e garantia para que ela seja detectada como não homogénea.

4.7 O Método GAN para p "outliers"

O método GAN foi introduzido na secção 4.4 apenas para um "outlier". Esta opção baseou-se na comodidade de simbologia na exposição inicial mas, além disso e principalmente, por considerarmos ser, nesse caso, muito mais fácil a compreensão e análise não só da própria metodologia mas também das propriedades indicadas e de um estudo comparado com os métodos tradicionais.

Como já foi referido, o método pode ser generalizado para uma aplicação a qualquer amostra em cujo modelo de discordância natural se admita a presença de p "*outliers*".

Antes de apresentar essa generalização devemos salientar que são raros os trabalhos, com e sobre outliers, que consideram mais do que duas observações, para estudar como discordantes.

Na realidade, embora sendo de formalização geral, os vários estudos existentes são depois concretizados, também por condicionantes de índole teórica, apenas para um ou dois outliers.

[14] Esta noção, apresentada para um "outlier" é facilmente adaptada para o caso de se pretender definir "outliers" múltiplos sendo, evidentemente, condicionada pelo modelo de discordância natural que for assumido no estudo. Sobre esta questão pode consultar-se a secção seguinte deste capítulo.

Esta decisão prática com a qual genericamente concordamos poderá (e deverá!) ser aplicada também na nossa metodologia.

É claro que, de qualquer ponto de vista, considerar um outlier numa amostra de dimensão 3, não é o mesmo do que considerar um outlier se tivermos a dimensão 30, ou 300. E, o mesmo pode afirmar-se sobre múltiplos outliers. Quantos outliers se devem estudar uma amostra de dimensão n? De facto, algo vai mal nos dados se tivermos mais observações suspeitas do que aquelas que consideramos genuínas.

Estas dificuldades são, obviamente agravadas se adicionarmos a vertente multiplicidade na discordância das observações. Como distinguir, para optar, entre um estudo admitindo um outlier ou uma análise com dois valores aberrantes?

Estas são, ainda, algumas questões em aberto e onde a objectividade, ou a falta dela, mais se sente.

Prossigamos então, para a formalização do método GAN para p "*outliers*". Também agora, para simplificação na exposição embora sem perda de generalidade, vamos admitir que as funções de densidade envolvem apenas um parâmetro. Por generalização, de toda a metodologia e notação apresentada na secção 4.4, o método consta das três fases seguintes:

- Formulação do modelo de discordância natural.
- Teste de homogeneidade da amostra.
- Selecção objectiva de p "outliers".

Modelo de discordância natural

- Pela hipótese H_0 - de *homogeneidade* - admitimos, como aliás na generalidade dos estudos sobre "outliers" que, todas as observações $x_1, ..., x_n$ têm a mesma densidade $f(x_i; \delta)$, $(i = 1, ..., n)$. Nesta hipótese H_0, a verosimilhança é

$$L(x_1, ..., x_n; \delta) = \prod_i^n f(x_i; \delta).$$

- Pela hipótese $\bar{H}$ - a *alternativa natural* - admitimos a presença de p valores discordantes na amostra e que, em princípio, podem ser quaisquer p das n observações. Seja então $\bar{H}_{j_1,..,j_p}$ a hipótese que admite x_{j_1},...,x_{j_p} como observações discordantes, para alguma combinação $(j_1, ..., j_p)$ dos índices $(1, ..., n)$, isto é, tal que:

- x_{j_1},...,x_{j_p} têm densidade $f(x_j;\delta')$, para j$\in(j_1,...,j_p)$
- as restantes observações seguem a mesma distribuição com densidade $f(x;\delta)$.

Se for válida a hipótese alternativa natural $\bar{H}$, existindo portanto, p observações discordantes x_{j_1},...,x_{j_p} isso significa que, pelo modelo, é assumida a hipótese $\bar{H}_{j_1,..,j_p}$ responsável pela geração destes "outliers".

Sendo assim, a hipótese alternativa natural $\bar{H}$ pode-se considerar como uma reunião das $\binom{n}{p}$ hipóteses $\bar{H}_{j_1,..,j_p}$ acima enunciadas. Neste sentido tem-se $\bar{H} = \bigcup \bar{H}_{j_1,..,j_p}$.

Na alternativa natural, sendo assumida a hipótese $\bar{H}_{j_1,..,j_p}$, temos a verosimilhança

$$L_{\bar{H}_{j_1,..,j_p}}(x_1,...,x_n;\delta,\delta') = \prod_{i\in(j_1,..,j_p)} f(x_i;\delta') \prod_{i\notin(j_1,..,j_p)} f(x_i;\delta).$$

Se representarmos por $\widehat{\delta}$ o estimador de máxima verosimilhança para δ na hipótese H_0 e por $\widehat{\delta}_{j_1,..,j_p}$ e $\widehat{\delta}'_{j_1,..,j_p}$ os estimadores de máxima verosimilhança para δ e δ' na hipótese $\bar{H}_{j_1,..,j_p}$ temos, para máximos dessas funções, sob H_0 e $\bar{H}_{j_1,..,j_p}$, respectivamente $\widehat{L}_0\ (x_1,...x_n;\widehat{\delta})$ e $\widehat{L}_j\ (x_1,...x_n;\widehat{\delta}_j,\widehat{\delta}'_j)$ onde, como vimos, o índice j é representativo de que foi admitida a observação x_j como discordante. Estes dois máximos, para simplificação de escrita, serão a seguir representados por $\widehat{L}_0$ e $\widehat{L}_j$.

Teste de homogeneidade

Para a formulação de um *teste de homogeneidade* na amostra em estudo, por generalização imediata de (4.10), também agora devemos construir o quociente

$$l_n = \frac{max_{H_0}\ L(x_1,...,x_n;\delta)}{max_{H_0\cup\bar{H}}\ L(x_1,...,x_n;\delta,\delta')}.$$

Porque, como foi referido, a hipótese alternativa natural $\bar{H}$ se pode considerar uma união de $\binom{n}{p}$ hipóteses, então pode escrever-se

$$l_n = \frac{\widehat{L}_0}{max\ (\widehat{L}_0, max_{j_1,...,j_p}\widehat{L}_{j_1,...,j_p})}. \tag{4.15}$$

Tal como na secção 4.4 para o caso onde se admite a presença de um "outlier" na amostra, também agora, com base em princípios de verosimilhança máxima, podemos formular a regra de teste

$$l_n < c \tag{4.16}$$

para decidir sobre a hipótese H_0 de homogeneidade nas observações. Na correspondente região de rejeição, estão incluídas todas as amostras $x_1, ..., x_n$ para as quais se verifica a desigualdade anterior.

Para determinar a região de rejeição, a partir de (4.15), podemos obter

$$l_n = \frac{1}{max\ (1, T(x_1, ..., x_n))}$$

onde

$$T(x_1, ..., x_n) \ = \ max_{j_1,...,j_p} \ \frac{\widehat{L}_{j_1,...,j_p}}{\widehat{L}_0} \tag{4.17}$$

é a estatística de teste.
A região de rejeição definida em (4.16) exige, portanto, que

$$T(x_1, ..., x_n) \ > \ c' \tag{4.18}$$

com $c' = 1/c$.

Se, para uma amostra $x_1, ..., x_n$, a correspondente estatística de teste $T(x_1, ..., x_n)$ *não verifica* a condição (4.18) então, o método generativo com alternativa natural conduz à aceitação da homogeneidade dessa amostra, negando portanto a existência de qualquer observação discordante e terminando, nesta segunda fase, a pesquisa de "outliers".

Selecção dos p "*outliers*"

Supondo rejeitada a hipótese de homogeneidade, isto é, sendo verificada a condição (4.18), aceitando-se portanto a existência de p observações discordantes, somos conduzidos, pelo próprio método, aos valores $x_1, ..., x_p$ que devem ser declarados "outliers". Neste sentido, como já vimos, o *método é generativo* porque, na última fase, "aponta" os responsáveis pela rejeição da homogeneidade na amostra em estudo. Os "outliers" que, *a posteriori*, são seleccionados são aquelas observações que correspondem à combinação $(j_1, ..., j_p)$ dos índices onde atingido o máximo da estatística $T(x_1, ..., x_n)$ acima definida.

4.8 Um exemplo para o caso geral

Como aplicação imediata do método GAN, em geral para p "outliers", vamos apresentar um exemplo para populações exponenciais, independentemente do estudo que faremos em diversos modelos de discordância, no capítulo 5.

Consideremos pela hipótese H_0 de homogeneidade das observações de uma amostra exponencial com densidade $f(x;\delta) = \frac{1}{\delta}\ exp(-\frac{x}{\delta})$. A aplicação da metodologia do método GAN, permite determinar, na hipótese H_0, o estimador de máxima verosimilhança $\widehat{\delta} = \bar{x}$, obtendo-se então, para máximo da função de verosimilhança

$$L_0\ (x_1, ..., x_n)\ =\ \frac{1}{\bar{x}^n}\ exp\ (-n).$$

Suponhamos em seguida que, pela hipótese alternativa natural, admitimos a presença de p "outliers" na amostra que seguem uma distribuição também exponencial mas, com diferente parâmetro de dispersão δ', enquanto que, as restantes (n-p) observações têm a anterior densidade exponencial $f(x;\delta)$. Admitamos que aquelas p observações discordantes podem ser quaisquer na amostra e seja $(j_1, ..., j_p)$ a combinação dos índices correspondentes. Sendo $L_{\bar{H}_{j_1,...,j_p}}(x_1, ..., x_n; \delta, \delta')$ a função de verosimilhança da amostra na hipótese alternativa $\bar{H}_{j_1,...,j_p}$. Como é comum nestes exemplos, com vista à obtenção de estimadores de máxima verosimilhança para os estimadores, para simplificação de cálculos, podemos utilizar os logaritmos das respectivas funções.

Temos então

$$log\ L_{\bar{H}_{j_1,...,j_p}}(x_1, ..., x_n; \delta, \delta')\ =\ -p\ log\ \delta'\ -\ \frac{\sum_{i\in(j_1,..,j_p)}}{\delta'}$$

$$-\ (n-p)\ log\ \delta\ -\ \frac{\sum_{i\notin(j_1,..,j_p)} x_i}{\delta}$$

donde, para δ e δ', podemos obter os estimadores de máxima verosimilhança na hipótese alternativa natural $\bar{H}_{j_1,...,j_p}$

$$\widehat{\delta}_{j_1,...,j_p}\ =\ \frac{1}{n-p}\ \sum_{i\notin(j_1,..,j_p)}\ x_i$$

$$=\ \frac{1}{n-p}\ (n\ \bar{x} - \sum_{i\in(j_1,..,j_p)}\ x_i)$$

e

$$\widehat{\delta'}_{j_1,...,j_p} \;=\; \frac{1}{p} \sum_{i\in(j_1,..,j_p)} x_i.$$

Por aplicação do método GAN, para este caso geral, devemos construir a estatística $T(x_1, ...x_n)$ definida em (4.17).
Assim, tem-se

$$T(x_1, ..., x_n) \;=\; max_{j_1,...,j_p} \frac{\widehat{L}_{j_1,...,j_p}}{\widehat{L}_0} \tag{4.19}$$

$$= \frac{p^p\ (n-p)^{n-p}}{n^n}\ max_{j_1,...,j_p} \frac{1}{(1-S_{j_1,...,j_p})^{n-p}\ (S_{j_1,...,j_p})^p}$$

onde

$$S_{j_1,...,j_p} \;=\; \frac{\sum_{i\in(j_1,..,j_p)} x_i}{n\bar{x}}$$

salienta a influência, das possíveis somas dos p valores, eventualmente discordantes, na estatística de teste $T(x_1, ..., x_n)$.

A partir de (4.19) podemos facilmente obter

$$S(x_1, ..., x_n) \;=\; min_{j_1,...,j_p}\ (1-S_{j_1,...,j_p})^{n-p}\ (S_{j_1,...,j_p})^p \tag{4.20}$$

como estatística equivalente a $T(x_1, ..., x_n)$ para teste de homogeneidade da amostra, neste exemplo.
A correspondente região de rejeição é, portanto,

$$S(x_1, ..., x_n) \;<\; c$$

e onde c é um ponto crítico a calcular, mais ou menos facilmente, conforme o número p de "outliers" admitidos no modelo de discordância natural. O método GAN neste exemplo para o caso geral permite-nos, também, concluir que, para a selecção objectiva das p observações discordantes, basta analisar as possíveis somas em $S_{j_1,...,j_p}$ e a sua influência em $S(x_1, ..., x_n)$. No capítulo seguinte apresentaremos alguns casos de interesse, com os respectivos cálculos dos pontos críticos.

Capítulo 5

"Outliers" em Populações Exponenciais

5.1 Introdução

Consideremos uma população exponencial com densidade de probabilidade

$$f(x;\lambda,\delta) = \frac{1}{\delta}\, exp\,(-\frac{x-\lambda}{\delta}) \qquad x \geq \lambda \qquad (5.1)$$

$$= 0 \qquad x < \lambda$$

Conforme anteriormente anunciámos, em termos gerais, consideramos que o aparecimento de *"outliers"* numa amostra se deve, fundamentalmente, à introdução de uma observação que, embora proveniente de uma mesma família (exponencial, normal, etc.) tem, no entanto, alguma alteração paramétrica que justifica esse surgimento - por exemplo, uma modificação de escala por mudança de observador experimental.

No estudo que a seguir apresentamos para *"outliers"* em populações exponenciais vamos dar especial relevo à influência do parâmetro de dispersão δ na eventual presença de valores discordantes na amostra.

A generalidade é inimiga da funcionalidade e da aplicabilidade de qualquer metodologia. Embora, em termos gerais, possamos considerar os dois parâmetros para estudo, vamos admitir que λ é conhecido e, então, para facilidade na exposição será admitido nulo o parâmetro de localização. Estas são, de facto, as hipóteses vulgarmente consideradas para os parâmetros λ e δ.

Para, numa perspectiva tradicional, testar um *outlier*, seleccionado *a priori*, como discordante em modelos que consideram desconhecido o parâmetro de localização existem apenas os testes "de tipo Dixon", baseados em estatísticas ordinais tais como

$$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}.$$

Os testes Dixon[1] - instrumentos práticos com alguma utilização em situações concretas - representam a tradição no estudo de observações discordantes nos anos sessenta do século passado, quando a estatística moderna estava crescendo. Os testes deste tipo foram utilizados principalmente[2] para amostras de populações exponenciais e normais. Esses são, de facto, testes *ad hoc* que se inserem numa abordagem diversa. Embora muito gerais pois nada exigem sobre a localização das distribuições são restritivos, pela dificuldade em ser "justificados e construídos" não tendo apoio teórico de modo, por exemplo, a poderem ser baseados em alguma metodologia de inferência estatística. Além disso são muito sensíveis ao "efeito de mascaramento".[3]

Consideremos o seguinte exemplo, elucidativo da tradicional subjectividade no estudo de outliers em populações exponenciais e que desenvolveremos ao longo deste capítulo.

Exemplo XIII:

Consideremos uma análise de outliers em dados exponenciais, baseada num estudo apresentado por Kimber e Stevens [76]. Foram registados os tempos de passagem de veículos automóvel num determinado cruzamento e na mesma direcção. Verificaram-se os seguintes valores (em segundos) para os intervalos entre os tempos de passagem:

25,52,7,61,446,34,87,76,4,17,19,240,116,45,64,141,31,503,10,181,101.

Um modelo exponencial parece adaptar-se aos dados, mas as observações 446 e 503 parecem demasiado grandes - como sugerem Kimber e Stevens (ib. p.156) que, no referido trabalho estudam essas observações bem como o eventual efeito de mascaramento por elas produzido.
Nestes dados, temos observações suspeitas?
Continuaremos, já a seguir...

O estudo que, neste capítulo, vamos apresentar para populações exponenciais, está dividido em duas partes.

[1] Sobre estes testes podem consultar-se as principais referências de Dixon sobre este assunto (Cf. [35], [36], [37] e [38]).

[2] Pode analisar-se a longa lista de testes "tipo Dixon" apresentada a páginas 195-6 e 219-21 de [11].

[3] Expressão que escolhemos para traduzir *masking effect*.

Numa primeira parte, na formulação do problema - na secção seguinte - estudaremos o método GAN no modelo de discordância tradicional e generalizando a problemática já analisada nas secções 4.3 e 4.5.

De facto, na abordagem vulgarmente utilizada, o objectivo principal - diverso do nosso como já sabemos - é a construção de um teste de discordância para determinada observação que, por alguma razão, se tornou suspeita ao analista e que *a priori* considera que deve testar. Essa observação, previamente fixada, para populações exponenciais é sempre o máximo ou o mínimo da amostra. Então, para dar justificação ao objectivo, um modelo de discordância tradicional admite na densidade (5.1) um parâmetro kδ para o valor a testar enquanto que todas as restantes observações ficam com o coeficiente de dispersão δ. Portanto será $k = 1$ na hipótese de não discordância desse valor.

E, como o objectivo é justificar um teste para o máximo $x_{(n)}$ ou para o mínimo $x_{(1)}$, fixa-se também o k - k>1 para $x_{(n)}$ e k<1 para $x_{(1)}$.

Esta é formulação tradicional do estudo de outliers em populações exponenciais e é utilizada na maior parte dos trabalhos apresentados. Uma boa referência (ainda) é o capítulo 4 de [11], embora apenas contemplando o caso k>1. Para uma situação 0<k<1 pode consultar-se [75] e [76]; este último com o interesse acrescido de abordar um modelo com dois *outliers* e que estudaremos em 5.3.

Exemplo XIII (cont.)**:**

Consideremos os dados acima introduzidos. A amostra ordenada é

4,7,10,17,19,25,31,34,45,52,61,64,76,87,101,116,141,181,240,446,503.

As observações extremas são agora igualmente suspeitas? A ordenação da amostra fez aparecer suspeitos?
São mais ou são menos discordantes?
Porque consideramos os "dois máximos"?
O par (4,7) é menos suspeito do que o par (446,503)?
E o máximo 503 é mais ou menos inesperado do que o mínimo 4?

Numa segunda parte, também na secção seguinte, estudaremos na forma mais geral, o método GAN para populações exponenciais com um modelo de discordância natural. Nesta aplicação, iremos considerar para as observações, discordantes e não discordantes, as densidades (5.1) $f(x;\delta')$ e $f(x;\delta)$ respectivamente, supondo nulo o parâmetro de localização. Esta abordagem, pouco usada no estudo de outliers, foi apresentada por Fieller [46] e por Lewis e Fieller [79], numa perspectiva próxima da nossa, considerando inicialmente este modelo de discordância. Mas, em seguida, pretendendo construir testes para observações seleccionadas *a priori*, estes autores, fixam e separam as duas situações $\delta > \delta'$ e $\delta < \delta'$.

Colocam-se, portanto, no modelo tradicional perdendo então a generalidade inicial em que se tinham colocado. O estudo que propõem não permite escolher entre esses dois modelos.

Também neste ponto o método GAN é geral.

5.2 Método GAN - um "outlier"

5.2.1 Modelo de discordância com formulação tradicional

Consideremos uma amostra $x_1, ..., x_n$ de uma população exponencial com densidade $f(x; \lambda, \delta)$ definida em (5.1) e de acordo com os pressupostos da secção anterior.

O modelo de discordância proposto pelo método GAN no capítulo 4 é, de acordo com a formulação tradicional, definido como segue:

- Pela hipótese H_0 - de *homogeneidade* - admitimos, como aliás na generalidade dos estudos sobre *"outliers"*, que todas as observações $x_1, ..., x_n$ têm aquela mesma densidade $f(x_i; \delta)$, $(i = 1, ..., n)$.
- Pela hipótese $\bar{H}$ - a *alternativa natural* - admitimos a presença de um valor discordante na amostra e que pode ser uma qualquer das n observações. Seja então $\bar{H}_j$ a hipótese que admite x_j como observação discordante, isto é, tal que:
  - x_j tem densidade de probabilidade $f(x_j; k\delta)$, para algum índice j$\in$(1,...,n) e com k>0
  - as restantes observações x_i seguem a mesma distribuição exponencial com densidade $f(x_i; \delta)$, para $i \neq j$.

Como sabemos, da definição do método GAN, a hipótese alternativa natural $\bar{H}$ pode considerar-se uma reunião de n hipóteses $\bar{H}_j$, correspondentes a um modelo onde, para cada j, x_j é assumida com densidade $f(x_j; k\delta)$.

Embora aplicando o método generativo com alternativa natural, temos ainda assim, neste caso, construído um modelo de discordância utilizando a formulação tradicional onde, através do parâmetro k se salienta a motivação para - em seguida e seleccionado *a priori* - se testar o máximo $x_{(n)}$ - se k>1 - ou o mínimo $x_{(1)}$, no caso contrário.

A verosimilhança da amostra, na hipótese H_0 de homogeneidade é

$$L_0\ (x_1, ..., x_n; \delta)\ =\ \frac{1}{\delta^n}\ exp\ (-\ \frac{n\bar{x}}{\delta}) \tag{5.2}$$

enquanto que, na alternativa $\bar{H}_j$, de discordância na observação x_j, se tem

$$L_j\ (x_1, ..., x_n; k, \delta)\ =\ \frac{1}{k\ \delta^n}\ exp\ (-(\ \frac{n\bar{x}\ -\ x_j}{\delta}\ +\ \frac{x_j}{k\ \delta})).$$

Representemos por $\widehat{\delta}_0$ e $\widehat{\delta}_j$ os estimadores de máxima verosimilhança para δ nas hipóteses H_0 e $\bar{H}_j$. Os correspondentes máximos serão denotados por $\widehat{L}_0$ e $\widehat{L}_j$.

Formulado um modelo de discordância tradicional, nesta aplicação do método GAN, que estamos construindo para o estudo de outliers em populações exponenciais devemos, em seguida, determinar as estatísticas para os testes de homogeneidade nas observações. Vamos considerar essas estatísticas, nas várias situações sobre o conhecimento de k e δ. As regras de teste serão apresentadas mais tarde.[4] Com evidentes vantagens metodológicas vamos, em primeiro lugar, considerar o caso mais geral.

5.2.1.1 k e δ desconhecidos

Como é proposto pelo método GAN, para formulação de um teste de homogeneidade nas observações, devemos obter a estatística

$$T\ (x_1, ..., x_n)\ =\ \frac{max_j\ \ \widehat{L}_j(x_1, ..., x_n; k, \delta)}{\widehat{L}_0(x_1, ..., x_n; \delta)}. \tag{5.3}$$

Ora, a verosimilhança da amostra, nessa hipótese é

$$L_0\ (x_1, ..., x_n; \delta)\ =\ \frac{1}{\delta^n}\ exp\ (-\frac{n\bar{x}}{\delta})$$

e, portanto, o estimador de máxima verosimilhança para δ é $\widehat{\delta}_0 = \bar{x}$. O correspondente máximo para L_0 é

$$\widehat{L}_0\ =\ \bar{x}^n\ exp\ (-n).$$

Na hipótese alternativa $\bar{H}_j$, a verosimilhança da amostra é

$$L_j\ (x_1, ..., x_n; k, \delta)\ =\ \frac{1}{k\delta^n}\ exp\ (-(\frac{n\bar{x} - x_j}{\delta} + \frac{x_j}{k\delta})) \tag{5.4}$$

o estimador para δ é

[4] Pela semelhança com o modelo de discordância natural, que estudaremos a seguir, as respectivas regras de teste e a (eventual) selecção do "outlier" apenas serão apresentadas, em conjunto para os dois casos, na sub-secção 5.2.4.

$$\widehat{\delta}_j \;=\; \bar{x} \;-\; (1 \;-\; \frac{1}{k})\;\frac{x_j}{n}$$

pelo que, para o máximo para L_j, temos

$$\widehat{L}_j = \;\frac{1}{k\;\bar{x}_n\;(1-(1-\frac{1}{k})\frac{x_j}{n\bar{x}})^n)}\;exp\;(-n).$$

A estatística $T(x_1,..,x_n)$ a calcular e acima referida, permite a determinação de

$$S(x_1,...,x_n) \;=\; max_j\;\frac{1}{k\;(1-(1-\frac{1}{k})\;\frac{x_j}{n\bar{x}})^n} \tag{5.5}$$

como estatística equivalente [5] para o estudo da homogeneidade da amostra. Porque não foi estimado, o parâmetro k, de facto, ainda surge em S.

Suponhamos que, embora desconhecendo k, existe alguma informação suplementar[6] que nos permite garantir, por exemplo, $k > 1$.
Neste caso a estatística obtida em (5.5) permite determinar que o máximo é atingido em $x_{(n)}$. Podemos, portanto usar a estatística

$$S(x_1,...,x_n) \;=\; \frac{x_{(n)}}{n\bar{x}}$$

para avaliar a homogeneidade na amostra.

O método GAN permite então resolver este caso, clarificando o significado de observação discordante e, sem ser necessário estimar k. Sobre este assunto é importante salientar o estudo, acima referido, de Barnett e Lewis (*ib.* p. 99) e o seu relacionamento com o método GAN.

De modo análogo, se $0 < k < 1$, podemos concluir que $x_{(1)}$ é a observação que, eventualmente, deve ser testada como outlier, devendo para tal ser usada a estatística equivalente

$$S(x_1,...,x_n) \;=\; \frac{x_{(1)}}{n\bar{x}}.$$

Estes dois casos que acabámos de apresentar correspondem à grande maioria dos estudos para outliers em populações exponenciais; como

[5] Para simplificação da escrita e sempre que não haja risco de confusão, de acordo com a simbologia já introduzida na apresentação geral do método, $S(x_1,...,x_n)$ representa sempre uma estatística equivalente a $T(x_1,...,x_n)$.

[6] Esta hipótese, de existência de informação suplementar sobre k é, de facto, utilizada nos testes de discordância tradicionais para outliers. Veja-se, por exemplo, Barnett e Lewis ([11], p. 98 e seguintes).

pode verificar-se em ([11] cap. 6). Podemos afirmar que, a partir daqui quase tudo é novo quando comparado com a abordagem tradicional para o estudo de *outliers* em dados estatísticos para amostras de populações exponenciais.

O primeiro caso "mais geral", embora ainda seguindo uma perspectiva de modelo de discordância com formulação tradicional, corresponde à situação onde não dispomos da, acima admitida, informação suplementar sobre k.

Nesta situação, a partir de (5.4) e considerando o respectivo logaritmo, temos

$$log\ L_j(x_1, ..., x_n) \ = \ -\ log\ k \ - \ n\ log\delta \ - \ (\frac{n\bar{x} - x_j}{\delta} + \frac{x_j}{k\delta});$$

e, daqui podemos obter, na hipótese alternativa $\bar{H}_j$, os estimadores de máxima verosimilhança

$$\widehat{\delta}_j \ = \ \frac{n\bar{x} - x_j}{n-1}$$

e

$$\widehat{k} \ = \ \frac{(n-1)x_j}{n\bar{x} - x_j}$$

pelo que,

$$\widehat{L}_j \ = \ \frac{1}{x_j}\ (\frac{n-1}{n\bar{x} - x_j})^{n-1}\ exp\ (-n).$$

A partir da estatística (5.3) para teste de homogeneidade facilmente obtemos

$$S(x_1, ...x_n) \ = \ min_j \ \frac{x_j}{n\bar{x}}\ (1 \ - \ \frac{x_j}{n\bar{x}})^{n-1}. \qquad (5.6)$$

Um breve estudo da estatística geral obtida em (5.6) e, uma vez que, nos modelos exponenciais temos

$$0 \ \leq \ \frac{x_j}{n\bar{x}} \ \leq \ 1$$

desde já, podemos ver que o método GAN nos indica que se deve analisar apenas o mínimo $x_{(1)}$ ou o máximo $x_{(n)}$ da amostra, como candidatos a *"outlier"*. Esta é uma importante conclusão que devemos registar pela

vantagem prática que daqui nos advém. De facto, este resultado coloca, sob eventual suspeita de discordância na amostra, apenas duas observações.

Ao terminar esta primeira aplicação do método GAN a populações exponenciais utilizando a formulação tradicional para o modelo de discordância devemos salientar, sem prejuízo de outras que adiante faremos, duas breves comparações entre os resultados aqui obtidos e as correspondentes situações nos testes de discordância que têm sido apresentados para outliers.

Em primeiro lugar, verificamos que na hipótese acima admitida de, embora desconhecendo k, se dispor de alguma informação suplementar, então o método GAN dispensa o cálculo do respectivo estimador. Esse estimador é normalmente exigido para a construção do teste de discordância[7] - a não confundir com teste de homogeneidade.

Em segundo lugar devemos salientar que, no caso mais geral (com k e δ desconhecidos), não se dispondo de qualquer informação suplementar sobre o parâmetro k, o método GAN ainda resolve o problema e propondo, de igual modo, os extremos da amostra como candidatos a outlier, mas não sendo, como tal, escolhidos previamente. Esta é, portanto, uma abordagem mais geral, não só nas hipóteses que formulamos mas também no método de pesquisa utilizado. Neste sentido, é uma metodologia não restritiva.

5.2.1.2 k conhecido e δ desconhecido

Esta hipótese, na formulação tradicional do modelo de discordância para populações exponenciais, de imediato nos conduz aos pressupostos da última parte do caso anteriormente analisado. De facto, se conhecermos $k = k_0$, esta situação corresponde a "ter informação suplementar" conforme considerado, isto é, saberemos da relação entre k e 1 e, portanto podemos continuar como acima se fez.

Assim:

- Se $k_0 > 1$ então devemos utilizar a estatística

$$S(x_1, ..., x_n) \ = \ \frac{x_{(n)}}{n\bar{x}}$$

  para testar a homogeneidade da amostra, sendo o máximo $x_{(n)}$ o possível candidato.

[7] Podem consultar-se, para populações exponenciais, as diversas referências apresentadas por Barnett e Lewis (Cf. [11], cap.6).

- Se $0 < k < 1$, a decisão deve ser tomada usando a estatística

$$S(x_1, ..., x_n) \ = \ \frac{x_{(1)}}{n\bar{x}}.$$

e o mínimo $x_{(1)}$ será considerado como candidato a *"outlier"*.

5.2.1.3 δ conhecido e k desconhecido

Suponhamos agora, apenas conhecido o parâmetro $\delta = \delta_0$. Neste modelo de discordância se, tal como fizemos em 5.2.1.1, admitirmos alguma informação sobre k e considerarmos (5.5) podemos concluir que:

- Se $k > 1$ então deve ser testada a observação $x_{(n)}$
- Se $k < 1$ então deve ser testada a observação $x_{(1)}$.

Se, pelo contrário, não estiver disponível aquela informação sobre o parâmetro k, devemos proceder ao cálculo do estimador na hipótese $\bar{H}_j$. Cálculos semelhantes aos anteriores conduzem a

$$\widehat{k} \ = \ \frac{x_j}{\delta_0}$$

para estimar k.

Temos portanto

$$\widehat{L}_0 \ = \ \frac{1}{\delta_0^n} \ \ exp\ (-\frac{n\bar{x}}{\delta_0})$$

e

$$\widehat{L}_j \ = \ \frac{1}{x_j \delta_0^{n-1}}\ exp\ (-\frac{n\bar{x} - x_j}{\delta_0} - 1)$$

e o teste de homogeneidade, usando (5.3), conduz à estatística

$$S(x_1, ..., x_n) \ = \ min_j \ \frac{x_j}{\delta_0}\ exp\ (-\frac{x_j}{\delta_0})$$

podendo, também aqui, ser obtidos apenas o mínimo ou o máximo da amostra como candidatos a *"outlier"*.

5.2.1.4 k e δ conhecidos

Pelas razões apresentadas no estudo do caso anterior, a situação agora admitida pode ser considerada como particular, pois o conhecimento de k e δ assim o determina.

Portanto, supondo rejeitada a hipótese H_0 de homogeneidade na amostra temos, neste caso, que:

- Se $k > 1$ então $x_{(n)}$ deve ser testada como discordante.
- Se $k < 1$ então deve ser testada a observação $x_{(1)}$.

Saliente-se que o método GAN mostra que a escolha da observação eventualmente *"outlier"*, como seria de esperar, depende apenas do parâmetro envolvido no mecanismo de geração desse valor, isto é, de k.

Terminamos este primeiro estudo de aplicação do método GAN aos vários casos componentes do modelo de discordância com formulação tradicional para *"outliers"* em populações exponenciais, com uma síntese no Quadro Resumo abaixo apresentado. Em seguida vamos abordar a formulação natural e, no final, apresentaremos uma síntese das duas abordagens.

Quadro Resumo - Formulação Tradicional

	δ	k	Inf.	Cand.	Estatística S
I	desc	desc*	$k > 1$	$x_{(n)}$	$x_{(n)}/n\bar{x}$
			$k < 1$	$x_{(1)}$	$x_{(1)}/n\bar{x}$
	desc	desc	-	$x_{(1)}$ ou $x_{(n)}$	$min_j \ \frac{x_j}{n\bar{x}} \ (1 - \frac{x_j}{n\bar{x}})^{n-1}$
II	desc	conh	$k > 1$	$x_{(n)}$	$x_{(n)}/n\bar{x}$
			$k < 1$	$x_{(1)}$	$x_{(1)}/n\bar{x}$
III	desc	desc*	$k > 1$	$x_{(n)}$	$x_{(n)}$
			$k < 1$	$x_{(1)}$	$x_{(1)}$
	conh	desc	-	$x_{(1)}$ ou $x_{(n)}$	$min_j \ \frac{x_j}{\delta} \ exp(-\frac{x_j}{\delta})$
IV	conh	conh	$k > 1$	$x_{(n)}$	$x_{(n)}$
			$k < 1$	$x_{(1)}$	$x_{(1)}$

* - dispondo de informação suplementar sobre K

5.2.2 Modelo de discordância com formulação natural

Consideremos uma amostra $x_1, ..., x_n$ de uma população exponencial com densidade $f(x; \lambda, \delta)$ tal como definida em (5.1) e de acordo com os pressupostos estabelecidos.

O modelo de discordância proposto pelo método GAN no capítulo 4 é, de acordo com a formulação natural, definido como segue:

- Pela hipótese H_0 - de *homogeneidade* - admitimos, como aliás na generalidade dos estudos sobre *"outliers"*, todas as observações $x_1, ..., x_n$ com a mesma densidade $f(x_i; \delta)$, $(i = 1, ..., n)$.
- Pela hipótese $\bar{H}$ - a *alternativa natural* - admitimos a presença de um valor discordante na amostra e que pode ser uma qualquer das n observações. Seja então $\bar{H}_j$ a hipótese que admite x_j como observação discordante, isto é, tal que:
  - x_j tem densidade de probabilidade $f(x_j; \delta')$, para algum índice j∈(1,...,n)
  - as restantes observações x_i seguem a mesma distribuição exponencial com densidade $f(x_i; \delta)$, para $i \neq j$.

Tal como se fez no estudo da formulação tradicional, vamos considerar os vários casos relativamente ao conhecimento dos parâmetros δ *e* δ'. Isto também permite um estudo comparativo entre as formulações - tradicional e natural.

5.2.2.1 δ e δ' conhecidos

Admitamos δ=δ_0 e δ'=δ'_0. Nesta formulação do modelo de discordância com alternativa natural temos, para máximos das verosimilhanças em H_0 e $\bar{H}_j$, respectivamente

$$\widehat{L}_0 \;=\; \frac{1}{\delta_0^n}\; exp\; (-\; \frac{n\bar{x}}{\delta_0})$$

e

$$\widehat{L}_j \;=\; \frac{1}{\delta'_0\delta_0^{n-1}}\; exp\; (-\; \frac{n\bar{x}}{\delta_0})\; exp\; (-\; x_j \frac{\delta_0 - \delta'_0}{\delta_0\delta'_0}).$$

A estatística (4.12) é

$$T(x_1, ..., x_n) \;=\; \frac{\delta_0}{\delta'_0}\; max_j\; exp\; (-\; x_j \frac{\delta_0 - \delta'_0}{\delta_0\delta'_0}).$$

e, portanto, podemos utilizar, para teste de homogeneidade nas observações, a estatística equivalente

$$S(x_1, ..., x_n) \;=\; min_j\; (\delta_0 - \delta'_0)\; x_j.$$

Se existir uma observação discordante, desta estatística podemos concluir que será:

- o mínimo $x_{(1)}$ da amostra se $\delta_0 > \delta_0'$
- o máximo $x_{(n)}$ da amostra se $\delta_0 < \delta_0'$.

Além disso, o método GAN propõe o uso das estatísticas de teste $X_{(1)}$ *ou* $X_{(n)}$ (conforme o caso) para avaliar a homogeneidade nas observações e consequente selecção do *"outlier"*, na terceira fase do método (se aplicável).

5.2.2.2 δ desconhecido e δ' conhecido

Embora com muito pouco interesse do ponto de vista das aplicações, admitamos δ'=δ_0' conhecido. Nesta situação, é necessário obter estimadores de máxima verosimilhança para δ nas hipóteses H_0 e $\bar{H}_j$. Assim, na primeira hipótese, temos

$$\widehat{\delta_0} \;=\; \bar{x}$$

pelo que,

$$\widehat{L_0} \;=\; \frac{1}{\bar{x}^n}\; exp\;(-n).$$

Por sua vez, na hipótese alternativa $\bar{H}_j$ obtemos

$$\widehat{\delta_j} \;=\; \frac{n\bar{x} - x_j}{n-1}$$

$$\widehat{\delta_0} \;=\; \frac{1}{\bar{x}^n}\; exp\;(-n)$$

e, portanto

$$\widehat{L_j} \;=\; \frac{1}{\delta_0'\;(\frac{n\bar{x}-x_j}{n-1})^{n-1}}\; exp\;(-(n-1) - \frac{x_j}{\delta_0'}).$$

A estatística $T(x_1, .., x_n)$ de (4.12) é agora equivalente a

$$S(x_1, ..., x_n) \;=\; \min_j\;(n - \frac{x_j}{\bar{x}})^{n-1}\; exp\;(\frac{x_j}{\delta'}).$$

5.2.2.3 δ conhecido e δ' desconhecido

Consideremos δ=δ_0 conhecido. Neste caso, o estimador para δ' é

$$\widehat{\delta'} \;=\; x_j$$

e então a referida estatística para o teste da homogeneidade das observações é agora equivalente a

$$S(x_1, ..., x_n) \;=\; \min_j \frac{x_j}{\delta_0}\; exp\; (-\frac{x_j}{\delta_0}).$$

Um breve estudo de $S(x_1, .., x_n)$ permite-nos concluir que, também neste caso, o método GAN seleccionará *a posteriori*, como candidato a *"outlier"*, apenas o mínimo ou o máximo da amostra. Como se devia esperar, esta situação corresponde, com formulação natural, ao anterior caso da secção 5.2.1.3 acima estudado no modelo de discordância com formulação tradicional.

5.2.2.4 δ e δ' desconhecidos

Neste caso, mais geral, para os parâmetros δ e δ' temos

$$\widehat{\delta} \;=\; \bar{x}$$

donde, para máximo de L_0, vem

$$\widehat{L}_0 \;=\; \frac{1}{\bar{x}^n}\; exp\; (-n).$$

Pela hipótese alternativa $\bar{H}_j$, a verosimilhança da amostra é

$$L_j(x_1, ..., x_n; \delta, \delta') \;=\; \frac{1}{\delta'\;\delta^{n-1}}\; exp\; (-\frac{n\bar{x} - x_j}{\delta} - \frac{x_j}{\delta'})$$

e portanto obtemos

$$\widehat{\delta}_j \;=\; \frac{n\bar{x} - x_j}{n-1}$$

e

$$\widehat{\delta'_j} \;=\; x_j$$

para estimadores de δ e δ', respectivamente.

Então tem-se

$$\widehat{L}_j \;=\; (n-1)^{n-1} \frac{1}{(n\bar{x} - x_j)^{n-1}\; x_j}\; exp\; (-n).$$

Para testar H_0 devemos usar a estatística (4.12) que nos permite encontrar

$$S(x_1, ..., x_n) \;=\; \min_j \frac{x_j}{n\bar{x}} \,(1 - \frac{x_j}{n\bar{x}})^{n-1}.$$

De imediato verificamos que, neste caso, temos uma situação equivalente ao estudado em 5.2.1.1, para a formulação tradicional, no caso mais geral onde não existe informação suplementar sobre os parâmetros, apenas sendo formalmente diferentes. Como em 5.2.1.1, podemos então concluir que, também agora, apenas o mínimo ou o máximo da amostra serão candidatos a *"outlier"*.

Como se fez para o modelo de discordância com formulação tradicional, apresentamos abaixo um quadro resumo dos resultados já obtidos com a metodologia natural.

Quadro Resumo - Formulação Natural

	δ	δ'	Cand. a *"outlier"*	Estatística S
I'	conh	conh	$x_{(1)}$ se $\delta > \delta'$	$x_{(1)}$
			$x_{(n)}$ se $\delta < \delta'$	$x_{(n)}$
II'	desc	conh	$x_{(1)}$ ou $x_{(n)}$	$min_j \;(n - \frac{x_j}{\bar{x}})^{n-1} \; exp(\frac{x_j}{\delta'})$
III'	conh	desc	$x_{(1)}$ ou $x_{(n)}$	$min_j \; \frac{x_j}{\delta} \; exp(-\frac{x_j}{\delta})$
IV'	desc	desc	$x_{(1)}$ ou $x_{(n)}$	$min_j \; \frac{x_j}{n\bar{x}} \; (1 - \frac{x_j}{n\bar{x}})^{n-1}$

5.2.3 Síntese

Do exposto nas secções 5.2.1 e 5.2.2, é oportuno fazer uma síntese dos vários casos e situações aí estudados para os dois modelos de discordância. Esta análise comparativa vai também permitir construir o quadro resumo que abaixo é apresentado e onde é sumariada a abordagem proposta pelo método GAN para o estudo de *"outliers"* em dados estatísticos exponenciais.

Salientemos, desde já, que as duas formulações não são completamente distintas. Esta síntese permite ainda comparar as estatísticas

utilizadas, principalmente, com vista à sua aplicação e estudo das respectivas distribuições e consequentes regras de decisão.

Uma primeira conclusão que permite afirmar a maior generalidade do estudo pelo método GAN em relação à formulação tradicional prende-se com o surgimento da estatística

$$S(x_1, ...x_n) \;=\; min_j \; \frac{x_j}{n\bar{x}} \; (1 - \frac{x_j}{n\bar{x}})^{n-1} \tag{5.7}$$

e correspondente ao caso mais geral onde não se dispõe de qualquer informação sobre os parâmetros envolvidos nas distribuições do mecanismo de geração dos dados em estudo.

Esta estatística $S(x_1, ..., x_n)$, acima considerada, é inovadora. Em primeiro lugar, porque nunca foi proposta para o estudo de outliers em populações exponenciais e, em segundo lugar, porque a regra de decisão que o seu uso permite construir é objectiva, isto é, a descoberta de um *"outlier"* na amostra não depende do analista que a estudar.

Se, como caso particular, admitirmos que se dispõe de alguma informação suplementar que permita ordenar os valores dos parâmetros desconhecidos então o método GAN propõe também, como na formulação tradicional, as estatísticas

$$S_1(x_1, ..., x_n) \;=\; \frac{x_{(1)}}{\sum_i \; x_i}$$

e

$$S_n(x_1, ..., x_n) \;=\; \frac{x_{(n)}}{\sum_i \; x_i}.$$

As estatísticas S_1 e S_n que (re)obtivemos têm sido o principal instrumento para o estudo de outliers em dados estatísticos exponenciais mas com a diferença que nesse caso as observações a analisar - $x_{(1)}$ ou $x_{(n)}$ - são seleccionadas *a priori*. Podemos afirmar que S_1 e S_n são as estatísticas mais gerais na formulação tradicional, enfatizando a importância da estatística $S(x_1, ..., x_n)$ acima indicada.

Uma breve nota para salientar que também pelo método GAN - embora numa perspectiva diferente como sabemos - em termos práticos S, S_1 e S_n correspondem a modelos onde se completa e generaliza o uso, obviamente mais restrito, das estatísticas $X_{(1)}$ e $X_{(n)}$ quando são

conhecidos os parâmetros.[8] Num ponto de vista de aplicação prática desta metodologia devemos ainda acrescentar a estatística para o estudo do caso em que apenas se desconhece o parâmetro da observação discordante. A situação contrária - onde se desconhece δ e se conhece δ' - é pouco recomendável pois aí se admite mais informação sobre o mecanismo aleatório de geração do *"outlier"* do que aquele que cria as "boas observações". Por esta razão não incluiremos este caso no quadro resumo que a seguir apresentamos como síntese das estatísticas e candidatos fornecidos pelo método GAN para o estudo de "outliers" em dados estatísticos exponenciais.

Exemplo XIII (cont.)**:**

Continuamos a estudar os dados exponenciais

4,7,10,17,19,25,31,34,45,52,61,64,76,87,101,116,141,181,240,446,503.

Admitamos a presença de um outlier na amostra. A estatística (5.7), calculada em $x_{(1)}$ e em $x_{(21)}$, fornece os valores 0.001708 e 0.001447 e, portanto, pelo método GAN, podemos seleccionar,nesta amostra, o máximo $x_{(21)} = 503$ como candidato a "outlier". Mas, se para esta amostra, fizermos uma alteração do mínimo $x_{(1)}$ e admitirmos que o dado registado é 3 (e não o 4 observado) então os respectivos valores para a estatística S são 0.001292 e 0.001447. Esta situação conduziria ao mínimo $x_{(1)}$=3 como candidato a "outlier" em vez do máximo 503.
Qual a diferença de suspeição entre 4 e 503? E entre 3 e 503?
Continuaremos a análise deste exemplo.

5.2.4 Regras para o Teste de Homogeneidade; seus valores críticos. Selecção do "Outlier"

5.2.4.1 δ e δ' conhecidos

Como sabemos, neste caso, o método GAN propõe, para teste da hipótese H_0 de homogeneidade nas observações, a estatística

$$S(x_1, ..., x_n) \ = \ min_j \ ((\delta - \delta')x_j)$$

e

$$S(x_1, ..., x_n) \ < \ c \tag{5.8}$$

[8] Consulte-se o capítulo 6 da obra de referência de Barnett e Lewis [11], onde também se podem comparar as diferentes abordagens.

Método GAN em Populações Exponenciais

δ	δ'	Inf.	Cand. a *"outlier"*	Estatística S
conh	conh	-	$x_{(1)}$ se $\delta > \delta'$	$x_{(1)}$
			$x_{(n)}$ se $\delta < \delta'$	$x_{(n)}$
conh	desc	-	$x_{(1)}$ ou $x_{(n)}$	$min_j \ \frac{x_j}{\delta} \ exp(-\frac{x_j}{\delta})$
desc	desc	$\delta > \delta'$	$x_{(1)}$	$x_{(1)} \ / \ n\bar{x}$
		$\delta < \delta'$	$x_{(n)}$	$x_{(n)} \ / \ n\bar{x}$
		-	$x_{(1)}$ ou $x_{(n)}$	$min_j \ \frac{x_j}{n\bar{x}} \ (1 - \frac{x_j}{n\bar{x}})^{n-1}$

é a correspondente região de rejeição. Assim, verificada esta condição, deve ser rejeitada a hipótese H_0 de que a amostra $x_1, ..., x_n$ é gerada pela mesma densidade exponencial. Além disso, conclui-se que existe uma observação responsável por essa decisão e que, *a posteriori*, é seleccionada na última fase daquele método generativo.

Pretendemos determinar os pontos críticos c.

Admitamos um nível de significância α para a nossa decisão. Teremos então,

$$Prob \ [\ S(x_1, ..., x_n) \ < \ c \ / \ H_0 \ verdadeira \ ] \ = \ \alpha.$$

Ora, na hipótese H_0, cada variável aleatória X_j tem densidade exponencial

$$f(x; \delta) \ = \ \frac{1}{\delta} \ exp \ (-\frac{x}{\delta})$$

e, portanto, a variável aleatória

$$Y_j \ = \ (\delta \ - \ \delta') \ X_j$$

tem a função de distribuição

$$F_{Y_j}(y) \ = \ \begin{cases} F_{X_j}(\frac{y}{\delta - \delta'}) & \text{se } \delta > \delta' \\ 1 \ - \ F_{X_j}(\frac{y}{\delta - \delta'}) & \text{se } \delta < \delta' \end{cases}$$

e onde F_{X_j} é a função de distribuição de X_j.

Se $\delta > \delta'$, temos

$$F_{Y_j}(y) \;=\; \begin{cases} 0 & \text{se } y \;<\; 0 \\ 1 \;-\; exp(-\frac{y}{\delta(\delta-\delta')}) & \text{se } y \;\geq\; 0. \end{cases}$$

e porque

$$F_S(c) \;=\; Prob\,[\;(\delta-\delta')\;min_j \;\leq\; c\;]$$

obtemos

$$F_S(c) \;=\; \begin{cases} 0 & \text{se } c \;<\; 0 \\ 1 \;-\; exp(-\frac{nc}{\delta(\delta-\delta')}) & \text{se } c \;\geq\; 0. \end{cases}$$

Fixado um nível de significância α e de acordo com a região de rejeição (5.8), deve ter-se $F_S(c) = \alpha$, pelo que

$$c \;=\; -\;\frac{\delta(\delta-\delta')}{n}\;log(1-\alpha)$$

permite calcular o ponto crítico para o teste de homogeneidade da amostra.

Portanto, (Cf. o quadro - resumo anterior), quando $\delta > \delta'$

$$min_j\;x_j \;<\; -\;\frac{\delta}{n}\;(log(1-\alpha))$$

é a região de rejeição da homogeneidade da amostra e, se verificada, o mínimo $x_{(1)}$, de acordo com o método GAN, será declarado *"outlier"*.

Se $\delta < \delta'$, de modo análogo, podemos obter

$$F_S(c) \;=\; \begin{cases} 1 \;-\; (1 \;-\; exp(-\frac{c}{\delta(\delta-\delta')}))^n & \text{se } c \;<\; 0 \\ 1 & \text{se } c \;\geq\; 0. \end{cases}$$

Fixado um nível de significância α e de acordo com a mesma região de rejeição (5.8), deve ter-se $F_S(c) = \alpha$, pelo que

$$c \;=\; -\;\delta(\delta-\delta')\;log(1-(1-\alpha)^{1/n})$$

permite calcular o ponto crítico para o teste de homogeneidade da amostra.

Portanto, (Cf. o quadro - resumo anterior), quando $\delta < \delta'$

$$max_j\ x_j\ >\ -\ \delta\ log(1-(1-\alpha)^{1/n})$$

é a região de rejeição da homogeneidade da amostra e, se verificada, o máximo $x_{(n)}$, de acordo com o método GAN, será declarado *"outlier"*.

5.2.4.2 δ conhecido e δ' desconhecido

Nesta hipótese para os parâmetros de dispersão da população exponencial temos, como vimos, a estatística

$$S(x_1,...,x_n)\ =\ min_j\ \frac{x_j}{\delta}\ exp(-\frac{x_j}{\delta})$$

para o teste da hipótese H_0 de homogeneidade nas observações e que,

$$S(x_1,...,x_n)\ <\ c$$

é a correspondente região de rejeição. Fixado um nível de significância α, pretendemos obter o ponto crítico c para decidir sobre H_0 e possível selecção do *"outlier"*.

Ora, se considerarmos uma nova variável aleatória

$$Z_j\ =\ \frac{X_j}{\delta}$$

então, a função de distribuição $F_S(s)$ de S é

$$F_S(s)\ =\ 1\ -\ \prod_1^n\ (1\ -\ Prob\ [Z_j\ exp(-Z_j)\ \leq\ s]).$$

Com o auxílio da figura 5.1 podemos concluir que, o acontecimento $\{z_i\ exp\ (-z_i) > s\}$ é equivalente $\{a\ \leq z_i \leq\ b\}$, onde a e b são as soluções da equação $x\ exp\ (-x) = s$ e porque, além disso, Z_i é uma variável aleatória exponencial reduzida, obtemos

$$F(s)\ =\ 1\ -\ (exp\ (-a) - exp\ (-b))^n.$$

Encontradas (numericamente) as soluções $c_1\ e\ c_2$ da equação

$$u\ exp\ (-u) = c$$

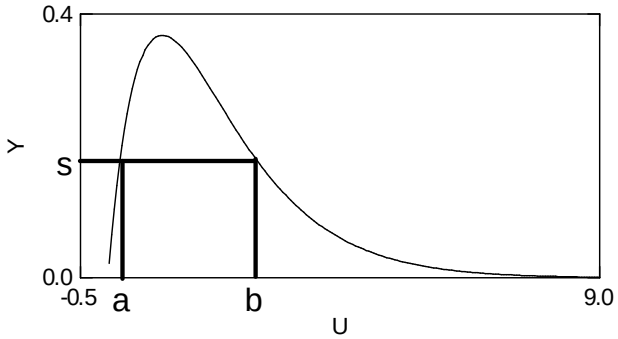

Figura 5.1: Função $f(u) = ue^{-u}$

os pontos críticos para a rejeição da homogeneidade ao nível de significância α são os pontos c para os quais se tem

$$e^{-c_1} - e^{-c_2} \;=\; (1 \;-\; \alpha)^{1/n}.$$

Abaixo apresentamos a **Tabela 1** com os valores críticos c para os níveis de significância de 1 e 5 por cento.

Sempre que, para uma amostra $x_1, ..., x_n$, se tem $S(x_1, ..., x_n) < c$, deve rejeitar-se a hipótese H_0 de homogeneidade nas observações e passar à terceira fase do método GAN - a selecção do *"outlier"*.

Como sabemos, do estudo feito nos parágrafos anteriores e de acordo com o quadro - resumo anterior, devemos decidir que a observação - $x_{(1)}$ ou $x_{(n)}$ apenas! - que minimiza x/δ exp(-x/δ) é um *"outlier"*. Nesta terceira fase do método GAN, a decisão objectiva sobre a presença de uma observação discordante na amostra, implica a análise de apenas essas duas observações extremas.

5.2.4.3 δ e δ' desconhecidos

Nesta hipótese, vamos fazer o estudo repartido por três casos distintos; dependentes do conhecimento de alguma informação suplementar sobre

Tabela 1: Pontos críticos, com níveis de significância α, para o teste de homogeneidade em amostras exponenciais e em modelos de discordância com δ **conhecido** e δ' **desconhecido**.

Dimensão da amostra	$\alpha = 0.05$	$\alpha = 0.01$
3	0.0144512	0.0029642
4	0.0109586	0.0022347
5	0.0088324	0.0017943
6	0.0074008	0.0014994
7	0.0063706	0.0012888
8	0.0055935	0.0011298
9	0.0049862	0.0010057
10	0.044985	0.0009063
15	0.0030245	0.0006070
20	0.0022802	0.0004566
25	0.0018309	0.0003660
30	0.0015300	0.0003058
40	0.0011529	0.0002298
50	0.0009248	0.0001842
60	0.0007723	0.0001537
70	0.0006632	0.0001319
80	0.0005811	0.0001156
90	0.0005172	0.0001098
100	0.0004659	0.0000927
200	0.0002346	0.0000466
300	0.0001569	0.0000311
400	0.0001179	0.0000234
500	0.0000946	0.0000188
1000	0.0000475	0.0000094

os parâmetros de dispersão, isto é, se $\delta>\delta$' ou $\delta<\delta$' ou se, pelo contrário, tal não é conhecido. O desenvolvimento destes casos I), II) e III) será também aproveitado para apresentar uma perspectiva histórica sobre a utilização das respectivas estatísticas.

Caso I) Com a informação suplementar: $\delta < \delta'$

Se $\delta < \delta'$, embora desconhecidos, então sabemos ser o máximo $x_{(n)}$ da amostra que é o candidato a *"outlier"* e que deve ser utilizada a estatística

$$S_n(x_1, ..., x_n) \;=\; \frac{x_{(n)}}{\sum_i \; x_i}$$

para o teste de homogeneidade[9].

Como temos indicado, esta estatística S_n tem sido tradicionalmente utilizada para testar a discordância do máximo $x_{(n)}$ de uma amostra exponencial, fixado *a priori* como candidato a outlier. Neste sentido é estudado em Fieller ([46] cap. II), em Lewis e Fieller ([79] p. 371) ou em Barnett e Lewis [11]. Esta mesma estatística S_n, foi por nós estudada, apresentando uma abordagem assimptótica para a respectiva função de distribuição (Cf. Rosado [109] e [110]), tornando particularmente simples o cálculo dos pontos críticos para a região de rejeição.

O estudo que apresentámos nas secções 5.2.1 e 5.2.2 e a análise comparativa formulada, permite-nos concluir da invariância de escala δ na estatística S_n. Portanto a determinação dos pontos críticos envolve apenas uma comparação do parâmetro desconhecido com 1, isto é, estão em alternativa os modelos com $\delta' < 1$ e com $\delta' > 1$.

Fixado um nível de significância α, para diversos valores da dimensão da amostra, são conhecidas tabelas de pontos críticos de rejeição do máximo.[10]

Para avaliar a discordância do máximo $x_{(n)}$ como outlier, estamos a utilizar a estatística S_n, com

$$S_n(x_1, .., x_n) \;=\; \frac{x_{(n)}}{\sum_i \; x_i}$$

cujo domínio de variação é

[9] O método prático para o teste de homogeneidade numa amostra exponencial com parâmetros de dispersão δ e δ' desconhecidos, mas sabendo que $\delta < \delta'$, é apresentado em (5.9). Pode dispensar-se o texto interposto. A leitura deste tem interesse teórico, numa análise comparativa entre as diversas metodologias que historicamente encontramos para estudar o máximo como outlier.

[10] Consultem-se, por exemplo, as tabelas de Barnett e Lewis ([11] p. 473 e 474).

$$\frac{1}{n} \leq S_n \leq 1.$$

Em alternativa, podemos usar a estatística S_n^*, com

$$S_n^*(x_1, .., x_n) = \frac{1}{S_n}$$

com a correspondente alteração na região de decisão.

Para determinação dos valores críticos tanto de S_n como de S_n^*, é necessário conhecer as distribuições exactas dessas estatísticas, quando a hipótese H_0 é verdadeira. Embora com algumas dificuldades de aplicação prática, são bem conhecidas essas distribuições, no caso de não existência de *"outliers"*, o que significa, após a hipótese suplementar δ=1 acima formulada que, $x_1, ..., x_n$ seguem uma lei exponencial com parâmetro de localização nulo e parâmetro de escala δ=1.

A função de distribuição exacta de S_n na hipótese H_0 foi obtida[11] por Fisher [49]. Este resultado teórico fundamental[12] tem sido abordado por outros autores com mais interesse para a teoria dos outliers.

São conhecidos os resultados exactos seguintes:

1. S_n tem função de distribuição

$$F_{S_n}(s) = \sum_{0}^{[1/s]} (-1)^j \binom{n}{j} (1 - js)^{(n-1)} \qquad \frac{1}{n} \leq s \leq 1$$

   onde [1/s] representa a parte inteira de $1/s$.

2. S_n^* tem função de distribuição

$$\begin{aligned} F_{S_n^*}(s) &= 1 - F_{S_n}(1/s) \\ &= \sum_{1}^{[s]} (-1)^{(j-1)} \binom{n}{j} (1 - j/s)^{(n-1)} \qquad 1 \leq s \leq n. \end{aligned}$$

As funções de distribuição $F_{S_n}(s)$ e $F_{S_n^*}(s)$ apresentam algumas dificuldades para o cálculo dos pontos críticos. Podemos verificar que, tendo

[11] Este estudo é muito importante para uma perspectiva histórica da teoria dos outliers em dados estatísticos porque, para além do relevo do resultado obtido, (também) reflecte a sensibilidade - do autor e da época - para a problemática de observações discordantes numa amostra.

[12] Citamos, a propósito, as demonstrações de Darling [32] através do estudo da função característica de S_n e também Barnett e Lewis ([11] p. 94-112).

em conta o tipo de operações a efectuar e, principalmente, com o aumento do número de parcelas, esses obstáculos crescem com a dimensão n da amostra. Torna-se importante um estudo assimptótico para essas distribuições.

Definamos as estatísticas

$$U_n \ = \ n \ S_n \ - \ log \ n$$

e

$$V_n \ = \ S_n^* \ \frac{log^2 \ n}{n} \ - \ log \ n.$$

A variável aleatória U_n, cujo domínio de variação é

$$1 - log \ n \ \leq \ u \ \leq \ n - log \ n$$

tem distribuição exacta

$$F_{U_n} \ (u) \ = \ F_{S_n} \ (\frac{u + log \ n}{n})$$

e a variável aleatória V_n, com o domínio

$$\frac{log^2 \ n}{n} - log \ n \ \leq \ v \ \leq \ log^2 \ n - log \ n$$

tem função de distribuição exacta

$$\begin{aligned} F_{V_n}(v) \ &= \ F_{S_n^*}(\tfrac{v+log \ n}{log^2 \ n} \ n) \\ &= \ 1 - F_{S_n}(\tfrac{log^2 \ n}{n(v+log \ n)}). \end{aligned}$$

Por outro lado, Rosado [109] demonstra que a estatística U_n tem distribuição assimptótica $U(x) = \Lambda(x) \ = exp \ (exp \ (-x))$ - de Gumbel para máximos - enquanto V_n tem a distribuição limite $V(x) = 1 - \Lambda(-x)$ - de Gumbel para mínimos.

Além disso, sabemos que as variáveis aleatórias

$$Y_i \ = \ 1 \ - \ exp \ (-X_i), i = 1, .., n$$

são uniformes no intervalo [0,1] e independentes.

Assim, definindo

$$S'_n \;=\; \frac{1 \;-\; exp\;(-x_{(n)})}{\sum_i\;(1 \;-\; exp\;(-x_i))}$$

com base nos estudos desenvolvidos por Darling [32], podemos concluir que a distribuição exacta de $1/S'_n$ é a de uma soma de $(n-1)$ variáveis aleatórias uniformes [0,1] e independentes, adicionadas à constante 1 e que, portanto, se tem

$$F_{1/S'_n}(s) \;=\; \frac{1}{(n-1)!}\;\sum_{i=0}^{k-2}\;(-1)^i\;\binom{n-1}{i}\;(s-1-i)^{n-1}$$

com $k-1 < s < k$ e $k = 2, ..., n$.

Finalmente, construamos a estatística

$$W_n \;=\; \sqrt{\frac{12}{n-1}}\;\frac{1}{S'_n} \;-\; (n+1)\;\sqrt{\frac{3}{n-1}}$$

obtida por conveniente transformação da anterior S'_n e com domínio de variação

$$\sqrt{\frac{12}{n-1}} - (n+1)\sqrt{\frac{3}{n-1}} \leq w \leq n\sqrt{\frac{12}{n-1}} \;-\; (n+1)\sqrt{\frac{3}{n-1}}.$$

O teorema do limite central assegura que W_n é assimptoticamente Normal $N(0,1)$.

Então temos, para a respectiva função de distribuição,

$$F_{W_n}(w) = F_{\frac{1}{S'_n}}(\frac{\sqrt{n-1}w + (n+1)\sqrt{3}}{\sqrt{12}}).$$

Na tabela 2, apresentamos um resumo dos resultados exactos e assimptóticos para as estatísticas U_n, V_n e W_n, que podem ser utilizadas em testes de homogeneidade.

Se forem a_α *e* $b_{\;\alpha}$ quantis tais que

$$F_{S_n}(a_\alpha) \;=\; F_{S'_n}(b_\alpha)\;(1 \;-\alpha)$$

então os pontos críticos, para U_n, V_n e W_n são, respectivamente:

Tabela 2: Resumo dos resultados exactos e assimptóticos para as estatísticas $\mathbf{U}_n$, $\mathbf{V}_n$ e $\mathbf{W}_n$.

	$\mathbf{U}_n$	$\mathbf{V}_n$	$\mathbf{W}_n$
Distribuição exacta	$F_{U_n}(u)$	$F_{V_n}(v)$	$F_{W_n}(w)$
Distribuição assimptótica	$\Lambda(x) = e^{-e^{(-x)}}$	$1 \ - \ \Lambda(-x)$	$N(0,1)$
Região Crítica	$U_n \geq c_1$	$V_n \leq c_2$	$W_n \leq c_3$

$$\begin{aligned} c_1 &= n\ a_\alpha - log\ n \\ c_2 &= \tfrac{log^2\ n}{n\ a_\alpha} \ - \ log\ n \\ c_3 &= \sqrt{\tfrac{12}{n-1}}\ b_\alpha \ - \ (n+1)\sqrt{\tfrac{3}{n-1}}. \end{aligned}$$

Tendo em vista uma melhor eficiência na aplicação, vamos fazer um estudo comparativo dos resultados obtidos, envolvendo as distribuições exactas e assimptóticas das estatísticas acima consideradas. Para o nosso propósito, é suficiente fazer a comparação dos três comportamentos limite apenas para valores próximos do nível de significância utilizado, uma vez que, para o teste de homogeneidade da amostra, basicamente, só é fundamental o conhecimento dos pontos críticos. Comecemos por estudar o comportamento da estatística $U_n = nS_n - log\ n$ cuja distribuição assimptótica[13] é Gumbel para máximos.

Fixado um nível de significância α, sejam $s_n,\ u_n\ e\ u_\infty$ tais que

$$\begin{aligned} s_n &- \textit{ponto crítico exacto para a estatística } S_n \\ u_n &- \textit{ponto crítico exacto para a estatística } U_n \\ u_\infty &- \textit{ponto crítico assimptótico para a estatística } U_n \end{aligned}$$

e de tal modo que

[13] Cf. Rosado [109].

$$\begin{aligned} Prob\,[S_n \leq s_n] &= Prob\,[U_n \leq u_n] \\ &= \Lambda(u_\infty) \\ &= 1-\alpha \end{aligned}$$

Tem-se, evidentemente, para os pontos exactos a relação

$$u_n \;=\; n\; s_n \;-\; log\; n.$$

Ainda para U_n, definamos a constante

$$e_n \;=\; \Lambda(u_\infty) \;-\; \Lambda(u_n)$$

que, para cada n, vai permitir fazer a comparação entre os vários pontos críticos.

Para um nível de significância α=0.05 temos o ponto crítico assimptótico $u_\infty = 2.9702$ e então, se na relação anterior para os valores exactos e assimptóticos substituirmos u_n por u_∞ obtemos a seguinte expressão limite para o correspondente quantil s_n *de* S_n

$$s_n \;=\; \frac{2.9702 \;+\; log\; n}{n}$$

que, para "grandes amostras" permite, com boa aproximação, determinar os quantis eliminando assim as dificuldades de cálculo apresentadas na aplicação da expressão exacta para $F_{S_n}(s)$. Nesse sentido, consulte-se a coluna u_n da tabela 3.

Note-se também que, para $n \geq 100$ é praticamente desprezável o erro cometido em termos de nível de significância.

Na tabela 4, verificamos ainda que, ao nível de significância α=0.01, a aproximação é melhor, vendo-se que o erro cometido é inferior a 1 por cento, logo para amostras de dimensão 40.

A expressão assimptótica para o ponto crítico exacto s_n, correspondente ao nível de significância α=0.01, é

$$s_n \;=\; \frac{4.6001 \;+\; log\; n}{n}$$

onde 4.6001 é o ponto crítico assimptótico u_∞ correspondente ao nível de significância α=0.01.

Tabela 3: Estudo do comportamento, exacto e assimptótico, de $\mathbf{S}_n$ e $\mathbf{U}_n$, para o nível de significância α=0.05.

n	s_n	u_n	$\Lambda\ (u_n)$	e_n
30	0.1980	2.5388	0.9241	0.0259
40	0.1576	2.6151	0.9294	0.0206
60	0.1131	2.6916	0.9345	0.0155
100	0.0738	2.7748	0.9395	0.0105
120	0.0632	2.7965	0.9408	0.0092
200	0.0413	2.9617	0.9496	0.0004
∞	0	2.9702	0.95	0

Tabela 4: Estudo do comportamento, exacto e assimptótico, de $\mathbf{S}_n$ e $\mathbf{U}_n$, para o nível de significância α=0.01.

n	s_n	u_n	$\Lambda\ (u_n)$	e_n
30	0.2412	3.8348	0.9786	0.0114
40	0.1915	3.9711	0.9813	0.0087
60	0.1371	4.1316	0.9841	0.0059
100	0.0888	4.2748	0.9862	0.0038
120	0.0759	4.3205	0.9869	0.0032
200	0.0495	4.6017	0.9900	-0.00001
∞	4.6601	0.99	0.95	0

Em relação à estatística V_n, sejam

v_n – *ponto crítico exacto para a estatística* V_n

v_∞ – *ponto crítico assimptótico para a estatística* V_n

e tais que α = 1-$\Lambda(-v_\infty)$ sendo portanto u_∞=$-v_\infty$.

De modo análogo definamos, também para V_n, a constante de erro

$$\begin{aligned} e_n &= (1 - \Lambda(-v_\infty)) - (1 - \Lambda(-v_n)) \\ &= \Lambda(-v_\infty) - (\Lambda(-v_n). \end{aligned}$$

Porque

$$V_n = \frac{log^2\ n}{n}\ \frac{1}{S_n} - log\ n$$

temos, também para V_n, a seguinte relação entre os valores exactos v_n *e* s_n

$$v_n = \frac{log^2\ n}{n\ s_n} - log\ n$$

e, facilmente se verifica que, entre os quantis exactos de U_n e V_n se pode estabelecer a igualdade

$$v_n = -\frac{log\ n}{n\ s_n}\ u_n.$$

Note-se a lenta convergência - no caso de V_n - do ponto crítico exacto para o assimptótico, na tabela 5. De igual modo, na mesma tabela, saliente-se que, a aproximação fornecida pela utilização da estatística V_n é "menos boa" do que a de U_n, uma vez que o erro cometido e_n é bastante grande mesmo para $n = 200$.

Para o nível de significância α=0.01 nota-se, na tabela 6, uma melhor aproximação se bem que ainda inferior à fornecida por U_n.

Comparem-se, nesse sentido, as tabelas 4 e 6.

Os estudos que fizemos para a estatística V_n, confirmam o comentário apresentado por Darling [32]: "I*t does not follow, of course, that the constants* c_n *and* d_n *are "best" in the sense that they give the "closest" aproximation to the limiting distribution functions when n is finite*".

Finalmente, para a estatística

$$W_n = \sqrt{\frac{12}{n-1}}\ \frac{1}{S'_n} - (n+1)\sqrt{\frac{3}{n-1}}$$

Tabela 5: Estudo comparativo, exacto e assimptótico, de $\mathbf{V}_n$, para o nível de significância α=0.05.

n	v_n	1 - Λ $(-v_n)$	e_n
30	-1.4537	0.2084	-0.1584
40	-1.5303	0.1946	-0.1446
60	-1.6242	0.1788	-0.1288
100	-1.7315	0.1622	-0.1122
120	-1.7656	0.1573	-0.1073
200	-1.8997	0.1389	-0.0889
∞	-2.9702	0.05	0

Tabela 6: Estudo comparativo, exacto e assimptótico, de $\mathbf{V}_n$, para o nível de significância α=0.01.

n	v_n	1 - Λ $(-v_n)$	e_n
30	-1.8025	0.1520	-0.1420
40	-1.9124	0.1373	-0.1273
60	-2.0564	0.1201	-0.1101
100	-2.2169	0.1032	-0.0932
120	-2.2710	0.0981	-0.0881
200	-2.4627	0.0816	-0.0716
∞	-4.6001	0.01	0

por nós construída, consideremos também

$$\begin{aligned} w_n &- \textit{ponto crítico exacto para a estatstica } W_n \\ w'_n &- \textit{ponto crítico exacto para a estatstica } \tfrac{1}{S'_n} \\ w_\infty &- \textit{ponto crítico assimptótico para aestatística } W_n \end{aligned}$$

e a constante de erro

$$e_n \;=\; \Phi(w_\infty) \;-\; \Phi(w_n)$$

onde Φ representa a função de distribuição da normal N(0,1).

Também neste caso, podemos estabelecer a seguinte relação entre os pontos críticos exactos

$$w_n \;=\; \sqrt{\frac{12}{n-1}}\; w'_n \;-\; (n+1)\,\sqrt{\frac{3}{n-1}}$$

que, ao nível de significância α=0.05 (portanto com w_∞=-1.645), para o ponto crítico exacto w'_n de $1/S_n$ dá

$$w'_n \;=\; \frac{-1.645\sqrt{n-1} \;+\; (n+1)\,\sqrt{3}}{\sqrt{12}}.$$

Para analisarmos o comportamento assimptótico de W_n - tabelas 7 e 8 - salientamos que, ao nível de significância α=0.05, logo para amostras de dimensão 10, é praticamente nulo o erro cometido, pela substituição do ponto crítico exacto w_n pelo quantil assimptótico w_∞=-1.645.

Em conclusão, analisando na globalidade os quadros comparativos de U_n, V_n e W_n, podemos afirmar que, o resultado assimptótico correspondente ao uso da estatística W_n fornece a melhor aproximação. De igual modo, para o estudo das distribuições exactas, também W_n é preferível, uma vez que mesmo para as "pequenas amostras" os quantis exactos são praticamente iguais aos assimptóticos. Para a determinação dos quantis exactos para W_n, pela enorme simplificação nos cálculos, devem então usar-se os quantis da normal N(0,1).

O esudo comparativo que apresentámos, para teste de homogeneidade numa amostra $x_1, ..., x_n$ obtida numa população exponencial com

Tabela 7: Estudo do comportamento, exacto e assimptótico, de $\mathbf{W}_n$ para o nível de significância α=0.05.

n	w	w_n	$\Phi\ (w_n)$	e_n
10	4.0731	-1.6476	0.0502	0.0002
20	8.4287	-1.6461	0.0501	0.0001
30	12.9412	-1.6459	0.0500	0.0000
∞	∞	-1.6450	0.0500	0.0000

Tabela 8: Estudo do comportamento, exacto e assimptótico, de $\mathbf{W}_n$ para o nível de significância α=0.01.

n	w	w_n	$\Phi\ (w_n)$	e_n
10	3.5139	-2.2934	0.0110	0.0010
20	7.5931	-2.3102	0.0104	0.0004
30	11.9007	-2.3153	0.0102	0.0002
∞	∞	-2.3250	0.0100	0.0000

parâmetros de dispersão δ e δ' desconhecidos, mas com $\delta < \delta'$, permite formular o seguinte:

Método prático de decisão (5.9)

1. Determinar
$$y_i \;=\; 1 \;-\; exp\,(-x_i) \;\; i = 1, ..., n$$
para obter uma amostra uniforme.

2. Calcular
$$W_n \;=\; \sqrt{\frac{12}{n-1}}\;\frac{\sum_i\; y_i}{y_{(n)}} \;-\; (n+1)\;\sqrt{\frac{3}{n-1}}.$$

3. Para o nível de significância 0.05, rejeitar a homogeneidade da amostra quando se tem
$$W_n \;<\; -1.645.$$

4. Para o nível de significância 0.01, rejeitar a homogeneidade da amostra quando se tem
$$W_n \;<\; -2.325.$$

5. Decidida a rejeição da homogeneidade seleccionar o máximo $x_{(n)}$ como *"outlier"*.

Caso II) Com a informação suplementar: $\delta > \delta'$

Se $\delta > \delta'$, embora desconhecidos, então sabemos ser o mínimo $x_{(1)}$ da amostra que é o candidato a *"outlier"* e que deve ser utilizada a estatística
$$S_1(x_1, .., x_n) \;=\; \frac{x_{(1)}}{\sum_i\; x_i}$$
para teste de homogeneidade nas observações.

Como sabemos, esta mesma estatística S_1 tem sido tradicionalmente usada no teste de discordância da observação $x_{(1)}$, sempre que esta seja, *a priori*, seleccionada pelo analista como candidata a outlier.

Neste sentido, referimos os excelentes estudos propostos por Barnett e Lewis [11], Fieller [46] e Lewis e Fieller [79].

Os pontos críticos para a estatística S_1 são conhecidos e podem ser consultados, por exemplo, em Barnett e Lewis ([11], p. 476). Os estudos apresentados por Rosado ([109] e [110]), numa abordagem assimptótica para a respectiva função de distribuição permitem, também neste caso, simplificar a obtenção dos pontos críticos para a região de rejeição

$$S_1(x_1, ..., x_n) \ < \ c.$$

O estudo que apresentámos em 5.2.1 e 5.2.2, para determinação das estatísticas de teste pelo método GAN, particularmente a análise comparativa entre as duas formulações, permite concluir da invariância de escala δ na estatística S_1. Em consequência, para determinação dos valores c, vamos introduzir a hipótese suplementar $\delta = 1$.

Neste caso, temos para $x_{(1)}$ a função de distribuição

$$Prob \ \{ \ X_{(1)} \leq x \ \} \ = \ 1 \ - \ e^{-nx}$$

e portanto $n \ X_{(1)}$ tem distribuição exponencial, com localização nula pelo modelo que admitimos e parâmetro $\delta = 1$, donde

$$Prob \ \{ \ \frac{X_{(1)}}{n} \leq x \ \} \ = \ 1 \ - \ e^{-n^2 x}.$$

Porque

$$S_1(x_1, .., x_n) \ = \ \frac{x_{(1)}}{n} \ \frac{1}{\bar{x}}$$

e como, $\bar{X}$ converge em probabilidade para 1, do estudo apresentado por Rosado (Cf. [109], p. 8-11), para cada n, podemos usar

$$F(x) \ = \ 1 \ - \ e^{-n^2 x}$$

como aproximação da função de distribuição de S_1.

Portanto, fixado um nível de significância α, temos

$$c \ = \ - \ \frac{log \ (1 \ - \ \alpha)}{n^2}$$

para calcular os pontos críticos para este caso com os parâmetros δ e δ' desconhecidos, mas admitindo $\delta > \delta'$. Valores da estatística S_1 inferiores a c conduzem à rejeição da homogeneidade na amostra e à consequente aceitação do mínimo $x_{(1)}$ como *"outlier"*.

Caso III) Sem informação suplementar sobre as escalas

Se os parâmetros δ e δ', são desconhecidos, e não dispomos de qualquer informação suplementar, estamos no caso mais geral para este estudo[14] de *"outliers"* em populações exponenciais. Como vimos, a partir de (4.12), o método GAN propõe que se use a estatística

$$S(x_1, ..., x_n) \;=\; \min_j \frac{x_j}{\sum_i \; x_i} \; (1 \;-\; \frac{x_j}{\sum_i \; x_i})^{n-1} \qquad (5.10)$$

que já estudámos em 5.2.3, e donde se pode concluir que

$$S(x_1, ..., x_n) \;<\; c \qquad (5.11)$$

é a correspondente região de rejeição da hipótese H_0.

Se, para uma amostra aquela desigualdade for verificada, rejeitaremos a hipótese de homogeneidade e, na fase seguinte, seleccionaremos, como *"outlier"*, a observação responsável por essa rejeição.

Para uma completa aplicação do teste necessitamos então que se estude a função de distribuição de S e a consequente determinação dos valores críticos de rejeição c.

Sejam

$$T_j \;=\; \frac{X_j}{\sum_i \; x_i} \qquad j = 1, ..., n$$

e seja $T_{(1)} \leq ... \leq T_{(n)}$ a correspondente amostra ordenada.

Para a estatística $S(x_1, ..., x_n)$ pode então escrever-se

$$S(x_1, ..., x_n) \;=\; \min_j \; T_j \; (1 \;-\; T_j)^{n-1}$$

[14] A nova abordagem aqui proposta, para o estudo de "outliers" em dados estatísticos exponenciais, foi referenciada e discutida no tratado fundamental *Outliers in Statistical Data* por Barnett e Lewis (Cf. [11], p. 204-5).

Este caso, mais geral para o estudo de *"outliers"* em amostras exponenciais, foi inicialmente proposto por Rosado [109] e [110], sendo em seguida investigado e trabalhado numa tese de doutoramento em Rosado [112] e continuado com Rosado [114], Rosado e Braumann [122], Rosado e Alpiarça [121] e Braumann [21].

Para um aprofundamento teórico, na área das amostras exponenciais, mas também com aplicações em diversos domínios científicos é apropriado referenciar, por ordem cronológica: Oliveira [94], Passos [96], Figueira [47], Martins [87] e Costa [30].

Para uma aplicação imediata deste resultado com o teste a seguir formulado use-se a estatística de (5.10) e consulte-se a tabela 9 onde são apresentados os pontos críticos. Para valores da estatística S de (5.10) inferiores aos tabelados deve-se rejeitar a homogeneidade na amostra e também aceitar o candidato - $x_{(1)}$ ou $x_{(n)}$ apenas! - como *"outlier"*.

com

$$0 \leq T_j \leq 1.$$

Seja F_S a função de distribuição daquela estatística S. Tem-se

$$\begin{aligned} F_S(s) &= Prob\ \{\ S \leq s\} \\ &= 1\ -\ Prob\ \{\ S > s\}. \end{aligned}$$

Com um raciocínio análogo ao que se fez nos casos anteriores, podemos também agora concluir que, o acontecimento $\{S > s\}$ é equivalente a $\{s_1 \leq T_{(1)} \leq T_{(n)} \leq s_2\}$, onde s_1 e s_2 são , desta vez, as soluções da equação $x\ (1-x)^{n-1} = s$.

Temos então

$$\begin{aligned} Prob\ \{\ S > s\ \} &= Prob\ \{\ T_{(1)}(1-T_{(1)})^{n-1} > s \wedge \\ &\qquad \wedge\ T_{(n)}(1-T_{(n)})^{n-1} > s\ \} \\ &= Prob\ \{\ s_1 \leq T_{(1)} \leq T_{(n)} \leq s_2\ \}. \end{aligned}$$

Fixado um nível de significância α, como vimos, a região de rejeição (5.11) exige que se tenha $F_S(c) = \alpha$.

O comportamento da função envolvida na determinação da estatística S, principalmente nos pontos de interesse, cria diversos problemas no cálculo.

Para contornar essas dificuldades e após a introdução das variáveis aleatórias T_i acima definidas, podemos concluir que a distribuição conjunta de $(T_{(1)}, T_{(n)})$ pode ser usada para calcular c.

O estudo desta distribuição conjunta envolvendo "metodologia dos envelopes" foi inicialmente feito, embora com alguma dificuldade na implementação prática, por Fieller [46].

Esse resultado foi, mais tarde, melhorado por Rosado [113].

Com base neste estudo sabemos que

$$Prob\ \{\ S > s\ \} = \sum_{i=0}^{k} (-1)^i\ \binom{n}{i}\ (\ i\ c_1 + (n-i)c_2 - 1\ )^{n-1}$$

onde[15]

$$k = [\ \frac{nc_2 - 1}{c_2 - c_1}\ ].$$

[15]Representamos por [x] a parte inteira de x.

Os valores críticos c, obtidos por aplicação desta metodologia, são apresentados na tabela 9.

Assim, em resumo, uma vez calculada a estatística

$$S(x_1, ..., x_n) \;=\; \min_j \frac{x_j}{\sum_i x_i} \;(1 \;-\; \frac{x_j}{\sum_i x_i})^{n-1}$$

e sempre que se tenha $S < c$, para um determinado nível de significância, deve rejeitar-se a homogeneidade na amostra e seleccionar como *"outlier"* a observação responsável por essa rejeição - $x_{(1)}$ *ou* $x_{(n)}$.

Em jeito de conclusão registe-se as discrepâncias (esperadas!) dos valores fornecidos pelas tabelas 8 e 9, construídas com base em modelos que admitem diferenças estatísticas significativas na informação exigida. Esse é o ponto de partida para um estudo do desempenho[16] dos diversos testes de discordância, neste caso, para populações exponenciais. Esse trabalho original, foi feito por Braumann [21].

Através da aplicação do método generativo com alternativa natural verificaremos, objectivamente, como aprofundar o estudo de outliers em duas direcções - a selecção e o tratamento de "outliers" - ambas com base em metodologia de máxima verosimilhança.

O suporte teórico envolvido no método GAN torna especialmente relevantes as reflexões que apresentaremos (também) no sentido da clarificação da noção de outlier.

Exemplo XIII (cont.):

Consideremos, de novo, os dados exponenciais, de Kimber e Stevens [76].

Temos então uma amostra (n=21) de uma população exponencial:

4, 7, 10, 17, 19, 25, 31, 34, 45, 52,
61, 64, 76, 87, 101, 116, 141, 181, 240, 446, 503.

que foi estudada por Kimber e Stevens [76] e para a qual, em particular, formularam um teste para avaliar a discordância da observação 503 que lhes pareceu ser surpreendentemente grande.

Porquê esta observação?

Um estudo mais aprofundado[17] *e objectivo através do método GAN permite concluir que apenas o mínimo ou o máximo da amostra são candidatos a outlier naquela amostra.*

Para este exemplo, admitindo um modelo de discordância geral com todos os parâmetros de escala desconhecidos, podemos verificar que o

[16] No capítulo 8, apresentaremos um estudo detalhado para populações normais.

[17] Cf. [114]. Este estudo foi citado por Barnett e Lewis ([11], p. 204-5) e originou uma discussão clarificadora sobre este assunto.

Tabela 9: Pontos críticos, com níveis de significância α, para o teste de homogeneidade em amostras exponenciais e em modelos de discordância com δ e δ' **desconhecidos**.

Dimensão da amostra	$\alpha = 0.05$	$\alpha = 0.01$
3	0.0059696422	0.0011481481
4	0.0031496759	0.0006209208
5	0.0019650620	0.0003908917
6	0.0013455908	0.0002687030
7	0.0009798027	0.0001960099
8	0.0007455544	0.0001492734
9	0.0005864513	0.0001174596
10	0.0004734290	0.0000948314
11	0.0003902464	0.0000781654
12	0.0003272424	0.0000655372
13	0.0002783745	0.0000557440
14	0.0002397056	0.0000479879
15	0.0002085806	0.0000417479
16	0.0001831558	0.0000366511
17	0.0001621186	0.0000324343
18	0.0001445136	0.0000289060
19	0.0001296324	0.0000259240
20	0.0001169402	0.0000233810
25	0.0000747360	0.0000149291
30	0.0000518693	0.0000103535
40	0.0000291723	0.0000058162
50	0.0000186777	0.0000037207
60	0.0000129779	0.0000025836
70	0.0000095406	0.0000018983
80	0.0000073090	0.0000014536
90	0.0000057784	0.0000011488
100	0.0000046831	0.0000009307
150	0.0000020863	0.0000004142
200	0.0000011757	0.0000002332
250	0.0000007535	0.0000001494

máximo $x_{(21)}$ é candidato mas não é outlier. No caso mais geral com parâmetros desconhecidos obtemos, para a anterior estatística de teste, os valores 0.001708 e 0.001447 originados pelas observações 4 e 503 respectivamente.

O correspondente ponto crítico é 0.0001169 (Cf. a tabela 9).

Consideremos os dados anteriores com o mínimo $x_{(1)}$ "alterado" para o valor 3. Aparentemente nada se modificaria nas conclusões ao estudarmos a presença de outliers na amostra modificada.

Ora, o método GAN leva-nos agora, objectivamente, a considerar o mínimo $x_{(1)} = 3$ como candidato a outlier (os valores da estatística de teste são agora 0.001293 e 0.001445 para as observações 3 e 503, respectivamente).

Na secção 9.2 continuaremos o estudo deste exemplo.

5.3 Método GAN - dois "outliers"

5.3.1 Modelo de discordância

Sabemos que o método GAN é facilmente aplicável também a modelos de discordância onde se admite a presença de mais do que um *"outlier"* na amostra.

Porque esta metodologia pode ser aproveitada para salientar algumas importantes propriedades vamos abordar o modelo de discordância com 2 "outliers", para populações exponenciais.

Como foi anteriormente referido, pelo interesse do ponto de vista prático vamos, também aqui, considerar que as observações são geradas por uma densidade exponencial onde admitimos uma localização em zero. Além disso, nesta problemática de múltiplos *"outliers"*, como na generalidade dos trabalhos, vamos também admitir que as observações discordantes têm o mesmo parâmetro de escala. Tal é uma opção apenas do ponto de vista das aplicações pois, do ponto de visto teórico, o método GAN pode ser formulado supondo que as escalas são diferentes. A densidade é, portanto,

$$f(x;\delta) \;=\; \frac{1}{\delta}\; exp\; (-\frac{x}{\delta}.)$$

Na alternativa natural $\bar{H}$ admitiremos a presença de dois valores discordantes na amostra e que podem ser quaisquer duas das n observações. Neste sentido, seja $\bar{H}_{jk}$ a hipótese que formula x_j *e* x_k $(j \neq k)$ como valores discordantes, isto é, tal que:

- x_j *e* x_k têm função densidade de probabilidade $f(x;\delta')$ para algum par (j,k) de índices;
- as restantes (n-2) observações seguem uma mesma distribuição com densidade de probabilidade $f(x_i;\delta), i = 1, ..., n \ \ (i \neq j, k)$.

5.3.1.1 δ e δ' conhecidos

Neste caso - o mais fácil nos cálculos, mas que podemos considerar de pouca aplicação, por exigir o conhecimento de todos os parâmetros - o método GAN, através de (4.17) propõe que se use a estatística

$$T(x_1, ..., x_n) \ = \ max_{j \neq k} \ \frac{\widehat{L}_{jk}}{\widehat{L}_0} \tag{5.12}$$

onde $\widehat{L}_{jk}$ *e* $\widehat{L}_0$ são os máximos da verosimilhança na hipótese alternativa natural e de homogeneidade, respectivamente.

A partir de (5.12) determinamos a estatística equivalente

$$S(x_1, ..., x_n) \ = \ min_{j \neq k} \ (\delta \ - \ \delta') \ (x_j \ + \ x_k)$$

para testar a homogeneidade da amostra.

Podemos daqui concluir que, como par candidato a *"outlier"*, temos:

$$(x_1, x_2) \qquad \text{se} \quad \delta \ > \delta'$$

$$(x_{(n-1)}, x_{(n)}) \qquad \text{se} \quad \delta < \delta'.$$

O método GAN permite ainda concluir que, para este modelo de discordância e em termos de máxima verosimilhança, o par $(x_{(1)}, x_{(n)})$ não é candidato a *"outlier"* qualquer que seja a relação de ordem entre os parâmetros de escala.

Esta é uma importante conclusão que é uma consequência da metodologia geral e objectiva utilizada no método GAN.

5.3.1.2 δ conhecido e δ' desconhecido

Tal como fizemos para o estudo de um, também neste modelo com dois *"outliers"* vamos considerar, eventualmente desconhecido, o parâmetro de escala das observações discordantes.

Admitindo então que conhecemos $\delta = \delta_0$, o método GAN através da estatística (5.12) propõe que se use

$$S(x_1, ..., x_n) \;=\; min_{j \neq k} \; (\frac{x_j + x_k}{\delta_0})^2 \; exp \; (-\; \frac{x_j + x_k}{\delta_0}).$$

Com pouco interesse do ponto de vista das aplicações e neste caso com maior complexidade nos cálculos envolvidos, não analisaremos a situação contrária, isto é, onde o desconhecimento existe nas observações em que "mais acreditamos" e onde saberíamos o parâmetro de escala daquelas que são responsáveis pela não homogeneidae da amostra.

Consideremos então o caso mais geral com:

5.3.1.3 δ e δ' desconhecidos

Neste caso temos, na hipótese H_0, de homogeneidade, o estimador

$$\widehat{\delta}_0 \;=\; \bar{x}$$

a que corresponde

$$\widehat{L}_0 = \frac{1}{\bar{x}^n} \; exp \; (-n)$$

para máximo da função de verosimilhança.

Na hipótese alternativa $\bar{H}_{jk}$ temos os estimadores

$$\widehat{\delta}_{jk} \;=\; \frac{n\bar{x} - (x_j + x_k)}{n-2}$$

$$\widehat{\delta'}_{jk} \;=\; \frac{x_j + x_k}{2}$$

para δ e δ', respectivamente.

E, para a verosimilhança, temos o máximo

$$\widehat{L}_{jk} \;=\; \frac{(n-2)^{n-2} \; 2^2}{(n\bar{x} - (x_j + x_k))^{n-2}(x_j + x_k)^2} \; exp \; (-n).$$

Então, a estatística (5.12) para o teste de homogeneidade é

$$T(x_1, ..., x_n) = (n-2)^{n-2} \; 2^2 \; \bar{x}^n \; max_{j \neq k} \frac{1}{(n\bar{x} - (x_j + x_k))^{n-2}(x_j + x_k)^2}$$

ou, como equivalente e de muito mais fácil manuseamento,

$$S(x_1,...,x_n) \;=\; min_{j\neq k}\;(\frac{x_j+x_k}{n\bar{x}})^2\;(1\;-\;\frac{x_j+x_k}{n\bar{x}})^{n-2}. \qquad (5.13)$$

Com um estudo semelhante ao que fizemos em 5.2.4.2, desta vez utilizando a função $g(x)=x^2(1-x)^{n-2}$, podemos concluir que:

Pelo método GAN, em populações exponenciais onde se formula um modelo de discordância com uma alternativa natural com duas observações discordantes, apenas os pares $(x_{(1)},x_{(2)})$ - com os dois mínimos - ou, $(x_{(n-1)},x_{(n)})$ - com os dois máximos - podem ser candidatos a *"outliers"*.

Salientamos esta importante propriedade do método GAN que, em termos de máxima verosimilhança, exclui a possibilidade dos extremos $x_{(1)}$ *e* $x_{(n)}$ poderem ser outliers em modelos onde a escala é a eventual causa de discordância. Devemos, de igual modo, realçar a propriedade generativa desta metodologia onde, de facto, a objectividade é um critério de trabalho![18]

Tal como fizemos em 6.2.2.3, para terminar esta aplicação do método GAN a modelos exponenciais com dois *"outliers"* podemos analisar os casos em que se dispõe de informação suplementar sobre a relação entre os parâmetros de escala δ e δ'. É um estudo, que deixamos ao cuidado do leitor, para concluir que:

1. Se admitirmos alguma informação suplementar de modo a poder garantir que $\delta<\delta'$, com a proporcionalidade k, então a estatística para o teste de homogeneidade

$$T(x_1,..,x_n) \;=\; \frac{1}{k^2\;min_{j\neq k}\;(1\;-\;(1-\frac{1}{k})\;\frac{x_j+x_k}{n\bar{x}})^n}$$

permite concluir que $(x_{(n-1)},x_{(n)})$ é o par candidato. Pode usar-se a estatística equivalente

$$S(x_1,..,x_n) \;=\; \frac{x_{(n-1)}\;+\;x_{(n)}}{n\bar{x}}$$

[18] Como comparação com os, assim chamados, métodos tradicionais é oportuno citar Fieller ([46], p.2.14): "In order to consider the null distribution of certain statistics appropriate for testing the largest and smallest observations simultaneously as outliers, it is necessary to find the joint distribution...".

É fundamental esclarecer o modelo.

Também Barnett e Lewis ([11] p. 201-2) consideram vários testes, por exemplo Ga5(Ea5), para testar os dois máximos como outliers.

para decidir se aquele par é *"outlier"*.

2. Se, pelo contrário, admitirmos que $\delta > \delta'$ então a decisão é tomada com base em

$$S(x_1, .., x_n) \ = \ \frac{x_{(1)} \ + \ x_{(2)}}{n\bar{x}}$$

sendo agora $(x_{(1)}, x_{(2)})$ o par[19] candidato a *"outlier"*.

5.4 Método GAN - p "outliers"

Compare-se a estatística (5.13) com (5.7) e (5.11) utilizadas em 5.2.3 para se verificar a semelhança nas expressões e na interligação entre os expoentes e o número de *"outliers"* admitidos nos respectivos modelos de discordância natural.

Numa generalização imediata da metodologia proposta ao longo deste capítulo, podemos formular um modelo de discordância natural onde admitimos a presença de p *"outliers"*. Pelas mesmas razões, tal como fizemos anteriormente, também aqui devemos supor que todas as observações discordantes têm o mesmo parâmetro de escala.

No caso mais geral, com todos os parâmetros desconhecidos, a estatística para o teste de homogeneidade é, neste caso

$$S(x_1, ..., x_n) \ = \ min_{i_1 \neq ... \neq i_p} \ (\frac{x_{i_1} + ... + x_{i_p}}{n\bar{x}})^p \ (1 \ - \ \frac{x_{i_1} + ... + x_{i_p}}{n\bar{x}})^{n-p}.$$

Com base nesta estatística de teste teremos, tal como nos estudos anteriores, a indicação *a posteriori* das observações *"outliers"* e que no caso em análise vão ser escolhidas entre as possíveis combinações dos p índices. Com um estudo semelhante ao que fizemos em 5.2.4.2 , desta vez utilizando a função $g(x) = x^p(1-x)^{n-p}$, podemos concluir que:

Pelo método GAN, em populações exponenciais onde se formula um modelo de discordância com uma alternativa natural com p observações discordantes, apenas os vectores $(x_{(1)}, ..., x_{(p)})$ - com os p mínimos - ou, $(x_{(n-p+1)}, ..., x_{(n)})$ - com os p máximos - podem ser candidatos a *"outliers"*.

[19] Para um aprofundamento do estudo pode sugerir-se, como leitura suplementar, a abordagem feita por Barnett e Lewis, para o mesmo par aleatório, com o teste Ga11(Ea11) para este modelo (Cf. [11], p. 207).

Capítulo 6

"Outliers" em Populações Gama

6.1 Introdução

É reconhecida e fundamentada em todos os trabalhos, a importância do estudo estatístico de populações Gama. São inúmeros os resultados obtidos para a detecção e teste de *"outliers"* em populações Gama.

Para estas distribuições, no campo das aplicações, são importantes os resultados iniciais de Epstein [42] e [43] e Basu [12], [13] e [14]; bem como Joshi [72] e Kale e Sinha [73] aos quais se seguiram Fieller [46], Lewis e Fieller [79] e Kimber e Stevens [76], entre outros.

As principais áreas de aplicação[1] envolvem o campo "life testing" com a consequente complexidade no estudo de *"outliers"* em dados estatísticos nesses domínios, onde muito há para investigar, com um recente avanço para a área[2] da estatística ambiental.

Em muitos casos, os resultados nessas distribuições Gama são válidos em famílias de distribuições onde as exponenciais se incluem e portanto podendo considerar-se como casos particulares. No entanto a complexidade dos cálculos, muitas vezes, não permite a obtenção de todos os resultados exigindo que o estudo fique "pelo particular" - entenda-se pelas exponenciais - não se conseguindo a "solução geral". É neste sentido que se justifica a nossa opção de, para esta família de distribuições, escrever dois capítulos separados. O estudo, que a seguir apresentamos para "outliers" em dados estatísticos de populações gama, na sua maior parte,

[1] Cf. [3].
[2] Cf. [88].

constitui uma generalização imediata da aplicação do método GAN a exponenciais, feita no capítulo anterior. Teremos, desta vez, a possibilidade de verificar diferentes análises específicas daquelas distribuições.

Consideremos uma densidade de probabilidade para a distribuição gama $G(x;\phi,\delta,\lambda)$, na forma mais geral:

$$f(x;\phi,\delta,\lambda)) \;=\; \frac{1}{\delta}\; f_\phi(\frac{x-\lambda}{\delta}) \qquad \phi,\delta \;>\; 0$$

onde

$$f_\phi(z) \;=\; \frac{z^{\phi-1}}{\Gamma(\phi)}\; exp\;(-z) \qquad z \geq 0$$

e sendo $\phi, \delta \; e \; \lambda$ os parâmetros de forma, dipersão e localização, respectivamente.

Tal como se fez no capítulo anterior para populações exponenciais, também agora vamos considerar todas as observações $x_1, ..., x_n$ com parâmetros iguais na hipótese H_0 de homogeneidade. Como já foi dito, o modelo pode generalizar-se para amostras geradas por mecanismos aleatórios com todos os parâmetros diferentes mas a complexidade dos cálculos e, principalmente, o pouco interesse com vista às aplicações, sustentam esta opção.

Vamos considerar o modelo de discordância com alternativa natural admitindo conhecidos os parâmetros de localização[3] e de forma. Assim, as densidades são da forma seguinte:

$$f(x;\phi,\delta)) \;=\; \frac{1}{\delta^\phi\;\Gamma(\phi)}\; x^{\phi-1}\; exp\;(-\frac{x}{\delta}) \qquad \phi,\delta>0; x\geq 0. \qquad (6.1)$$

6.2 Método GAN - um "outlier"

6.2.1 Modelo de discordância com alternativa natural

Consideremos uma amostra $x_1, ..., x_n$ de uma população gama com densidade $f(x;\phi,\delta)$ tal como definida em (6.1).

O modelo de discordância com alternativa natural é definido por:

[3]Também aqui, sem perda de generalidade, vamos admitir que $\lambda = 0$.

- Pela hipótese H_0 de homogeneidade, todas as observações $x_1, ..., x_n$ têm densidade de probabilidade $f(x_i; \phi_0, \delta)$, $(i = 1, ..., n)$, supondo o parâmetro de forma $\phi = \phi_0$ conhecido e δ o parâmetro de dispersão.

- Pela hipótese alternativa natural, vamos admitir a presença de um valor (eventualmente) discordante na amostra e que - como sabemos pela metodologia GAN - pode ser qualquer uma das n observações. Seja então $\bar{H}_j$ a hipótese que admite x_j como discordante com densidade $f(x_i; \phi_0, \delta')$ para algum $j \in (1, ..., n)$ e as restantes observações com densidade $f(x_i; \phi_0, \delta)$.

6.2.1.1 δ e δ' conhecidos

Nestas hipóteses o método GAN - introduzido em 4.4 - determina a estatística

$$T\ (x_1, ..., x_n)\ =\ \frac{max_j\ \ \widehat{L}_j}{\widehat{L}_0} \tag{6.2}$$

que deve ser calculada.

No caso que analisamos, a partir de $T(x_1, ..., x_n$, podemos encontrar

$$S(x_1, ..., x_n)\ =\ min_j\ (\delta_0\ -\ \delta_0')\ x_j \tag{6.3}$$

como estatística equivalente.

Esta estatística para o teste de homogeneidade da amostra não depende do parâmetro de forma ϕ.
Tal como aconteceu para as populações exponenciais, também agora, daquela estatística podemos concluir que, se existir uma observação discordante, é:

- o mínimo $x_{(1)}$ da amostra se $\delta_0 > \delta_0'$

- o máximo $x_{(n)}$ da amostra se $\delta_0 < \delta_0'$.

Portanto, o método GAN propõe o uso das estatísticas de teste $X_{(1)}$ ou $X_{(n)}$ (conforme o caso) para decidir sobre a homogeneidade e consequente selecção da observação *"outlier"*, na terceira fase (se aplicável).

6.2.1.2 δ conhecido e δ' desconhecido

Nas hipóteses consideradas, para este caso, a partir de (6.2) podemos determinar

$$S(x_1, ..., x_n) \ = \ min_j \ \frac{x_j}{\delta} \ exp \ (-\frac{x_j}{\delta}) \tag{6.4}$$

para fazer o teste de homogeneidade e, de novo, podemos concluir que: **A estatística para o teste de homogeneidade da amostra não depende do parâmetro de forma** ϕ.

Como candidatos a *"outlier"* surgem $x_{(1)}$ ou $x_{(n)}$ e a condição de discordante, também neste caso, não depende do parâmetro de forma, que admitimos conhecido.

6.2.1.3 δ e δ' desconhecidos

Para este caso, a estatística de teste é:

$$S(x_1, ..., x_n) \ = \ max_j \ \frac{1}{k^{\phi_0}(1 \ - \ (1 - \frac{1}{k}) \ \frac{x_j}{n\bar{x}})^{n\phi_0}} \tag{6.5}$$

onde k é a razão δ'/δ.

Se pudermos garantir que, por exemplo, $k \ > \ 1$ então

$$S(x_1, ..., x_n) \ = \ [k \ min_j \ (1 \ - \ (1 - \frac{1}{k}) \ \frac{x_j}{n\bar{x}})^n]^{-\phi_0}$$

ou, como equivalente,

$$S(x_1, ..., x_n) \ = \ \frac{x_{(n)}}{n\bar{x}}$$

para o teste de homogeneidade. A partir da estatística S podemos pois concluir se o máximo $x_{(n)}$ é *"outlier"*.

De modo análogo, se $0 \leq k \leq 1$, temos o mínimo $x_{(1)}$ como candidato e a decisão sobre a homogeneidade da amostra pode ser tomada através de

$$S(x_1, ..., x_n) \ = \ \frac{x_{(1)}}{n\bar{x}}.$$

Se não tivermos disponível qualquer informação sobre k, então estamos no caso mais geral deste modelo de discordância e os cálculos em torno da estatística (6.5) conduzem aos estimadores

$$\widehat{\delta}_j \ = \ \frac{n\bar{x} \ - \ x_j}{\phi_0(n-1)}$$

e

$$\widehat{\delta}'_j \;=\; \frac{x_j}{\phi_0}$$

sendo

$$S(x_1,...,x_n) \;=\; min_j \; \frac{x_j}{n\bar{x}} \; (1 \;-\; \frac{x_j}{n\bar{x}})^{n-1}$$

a estatística para o teste e decisão sobre x_j - escolhido entre $x_{(1)}$ e $x_{(n)}$ - como *"outlier"*.

Do estudo que efectuámos para populações gama podemos concluir da semelhança com as exponenciais e, em particular, as conclusões com a intervenção dos diversos parâmetros na decisão a tomar sobre a existência de valores discordantes numa amostra. Em resumo temos o seguinte quadro com as diferentes estatísticas para o teste de homogeneidade nos diversos casos correspondentes ao conhecimento dos parâmetros.

Método GAN em Populações Gama

δ	δ'	Inf.	Cand. a *"outlier"*	Estatística S
conh	conh	-	$x_{(1)}$ se $\delta > \delta'$	$x_{(1)}$
			$x_{(n)}$ se $\delta < \delta'$	$x_{(n)}$
conh	desc	-	$x_{(1)}$ ou $x_{(n)}$	$min_j \; \frac{x_j}{\delta} \; exp(-\frac{x_j}{\delta})$
desc	desc	$\delta > \delta'$	$x_{(1)}$	$x_{(1)} \;/\; n\bar{x}$
		$\delta < \delta'$	$x_{(n)}$	$x_{(n)} \;/\; n\bar{x}$
		-	$x_{(1)}$ ou $x_{(n)}$	$min_j \; \frac{x_j}{n\bar{x}} \; (1 - \frac{x_j}{n\bar{x}})^{n-1}$

6.2.2 Regras para o Teste de Homogeneidade; seus valores críticos. Selecção do "Outlier"

6.2.2.1 δ e δ' conhecidos

Se $\delta > \delta'$, o método GAN propõe, para teste da hipótese H_0 de homogeneidade nas observações, a estatística

$$S(x_1,...,x_n) \;=\; min_j \; ((\delta - \delta')x_j)$$

e

$$S(x_1, ..., x_n) \ < \ c \tag{6.6}$$

é a correspondente região de rejeição. Assim, verificada esta condição, deve ser rejeitada a hipótese H_0 de que a amostra $x_1, ..., x_n$ é gerada pela mesma densidade gama.

Pretendemos determinar os pontos críticos c.

Admitamos um nível de significância α para a nossa decisão. Teremos então,

$$F_S \ (c) \ = \ \alpha$$

e, portanto, para o ponto crítico c é

$$F_S \ (\frac{c}{\delta - \delta'}) \ = \ 1 \ - \ (1 \ - \ \alpha)^{\frac{1}{n}}.$$

As tabelas da função gama incompleta, por exemplo em Pearson e Hartley [97] ou Abramowitz e Stegun [1], permitem determinar os pontos críticos de rejeição para este modelo de discordância.

Se $\delta < \delta'$, temos

$$F_S \ (\frac{c}{\delta - \delta'}) \ = \ (1 \ - \ \alpha)^{\frac{1}{n}}$$

e, como no caso anterior, as tabelas da função gama incompleta devem ser usadas para calcular c.

6.2.2.2 δ conhecido e δ' desconhecido

Nesta hipótese para os parâmetros de dispersão da população gama temos a estatística

$$S(x_1, ..., x_n) \ = \ min_j \ \frac{x_j}{\delta} \ exp(-\frac{x_j}{\delta})$$

para o teste da hipótese H_0 de homogeneidade nas observações e

$$S(x_1, ..., x_n) \ < \ c$$

é a correspondente região de rejeição. Fixado um nível de significância α, pretendemos obter o ponto crítico c para decidir sobre H_0 e possível selecção do *"outlier"*.

Cálculos análogos aos que fizemos em 5.2.4.2 conduzem a

$$F_S(s) \;=\; 1 \;-\; [\sum_{j=0}^{\phi-1} \frac{1}{j!}\,(s_1^j\; exp\;(-s_1) \;-\; s_2^j\; exp\;(-s_2))]^n$$

onde estamos a supor que ϕ é inteiro.

Os pontos críticos para a rejeição da homogeneidade ao nível de significância α são os pontos c para os quais, encontradas (numericamente) as soluções c_1 *e* c_2 da equação

$$x\; exp\;(-x) = c$$

se tem

$$\sum_{j=0}^{\phi-1} \frac{1}{j!}\,(s_1^j\; exp\;(-s_1) \;-\; s_2^j\; exp\;(-s_2)) \;=\; (1 \;-\; \alpha)^{\frac{1}{n}}.$$

Abaixo apresentamos as **Tabelas 1 e 2** com os valores críticos c para os níveis de significância de 1 e 5 por cento e para alguns valores do parâmetro de forma.

Sempre que, para uma amostra $x_1, ..., x_n$, se tem $S(x_1, ..., x_n) < c$, deve rejeitar-se a hipótese H_0 de homogeneidade nas observações e passar à terceira fase do método GAN - a selecção do *"outlier"*.

Como sabemos, do estudo feito nos parágrafos anteriores e de acordo com o quadro-resumo (5.2.3) devemos decidir que a observação - $x_{(1)}$ ou $x_{(n)}$ apenas! - que minimiza x/δ exp(-x/δ) é um *"outlier"*. Nesta terceira fase do método GAN, a decisão objectiva sobre um dado discordante na amostra, implica a análise de apenas essas duas observações extremas.

6.2.2.3 δ e δ' desconhecidos

Do mesmo modo e, pelas mesmas razões, que foi feito para as exponenciais, nesta hipótese, vamos repartir o estudo por três casos distintos; dependentes do conhecimento de alguma informação suplementar sobre os parâmetros de dispersão.

Caso I) Com a informação suplementar: $\delta < \delta'$

Se $\delta < \delta'$, embora desconhecidos, então sabemos ser o máximo $x_{(n)}$ da amostra que é candidato a *"outlier"* e que deve ser utilizada a estatística

Tabela 1: Pontos críticos, ao nível de significância $\alpha = \mathbf{0.05}$, para o teste de homogeneidade em amostras gama e em modelos de discordância natural, com parâmetros de forma ϕ e de escala δ **conhecido** e δ' **desconhecido**.

Dimensão da amostra	$\phi = 2$	$\phi = 4$	$\phi = 6$	$\phi = 8$
3	0.0144512	0.000837938	0.0000565468	0.0000041632
4	0.0109586	0.000587632	0.0000374987	0.0000026322
5	0.0088324	0.000447002	0.0000273531	0.0000018590
6	0.0074008	0.000357851	0.0000211926	0.0000014009
7	0.0063705	0.000296695	0.0000170988	0.0000011043
8	0.0055934	0.000252378	0.0000142115	0.0000008989
9	0.0049862	0.000218899	0.0000120804	0.0000007493
10	0.0044985	0.000192790	0.0000104518	0.0000006357
15	0.0030245	0.000118545	0.0000060186	0.0000003496
20	0.0022802	0.000084151	0.0000040848	0.0000002263
25	0.0018309	0.000064585	0.0000030291	0.0000001594
30	0.0015300	0.000052066	0.0000023724	0.0000001173
40	0.0011522	0.000037107	0.0000016271	0.0000000828
50	0.0009245	0.000028562	0.0000012136	0.0000000588
60	0.0007721	0.000023084	0.0000009551	0.0000000436
70	0.0006630	0.000019288	0.0000007794	0.0000000330
80	0.0005808	0.000016511	0.0000006523	0.0000000251
90	0.0005172	0.000014402	0.0000005562	0.0000000190
100	0.0004660	0.000012746	0.0000004938	0.0000000141
200	0.0002346	0.000005727	0.0000001999	0.0000000057
300	0.0001569	0.000003597	0.0000001139	0.0000000000
400	0.0001179	0.000002586	0.0000000840	0.0000000000
500	0.0000946	0.000002010	0.0000000623	0.0000000000
1000	0.0000475	0.000000914	0.0000000198	0.0000000000

Tabela 2: Pontos críticos, ao nível de significância $\alpha = \mathbf{0.01}$, para o teste de homogeneidade em amostras gama e em modelos de discordância natural, com parâmetros de forma ϕ e de escala δ **conhecido** e δ' **desconhecido**.

Dimensão da amostra	$\phi = 2$	$\phi = 4$	$\phi = 6$	$\phi = 8$
3	0.0029642	0.000115678	0.0000058538	0.0000003390
4	0.0022347	0.000082124	0.0000039739	0.0000002193
5	0.0017944	0.000063038	0.0000029470	0.0000001542
6	0.0014994	0.000050819	0.0000023077	0.0000001139
7	0.0012879	0.000042381	0.0000018885	0.0000000980
8	0.0011291	0.000036226	0.0000015837	0.0000000803
9	0.0010051	0.000031549	0.0000013565	0.0000000671
10	0.0009059	0.000027885	0.0000011814	0.0000000569
15	0.0006069	0.000017381	0.0000006920	0.0000000276
20	0.0004566	0.000012443	0.0000004807	0.0000000132
25	0.0003662	0.000009612	0.0000003605	0.0000000103
30	0.0003056	0.000007782	0.0000002847	0.0000000102
40	0.0002298	0.000005595	0.0000001944	0.0000000053
50	0.0001842	0.000004331	0.0000001424	0.0000000021
60	0.0001538	0.000003514	0.0000001138	0.0000000000
70	0.0001319	0.000002945	0.0000000973	0.0000000000
80	0.0001156	0.000002523	0.0000000818	0.0000000000
90	0.0001028	0.000002204	0.0000000699	0.0000000000
100	0.0000927	0.000001964	0.0000000607	0.0000000000
200	0.0000466	0.000000891	0.0000000188	0.0000000000
300	0.0000312	0.000000558	0.0000000102	0.0000000000
400	0.0000233	0.000000409	0.0000000089	0.0000000000
500	0.0000188	0.000000318	0.0000000052	0.0000000000
1000	0.0000093	0.000000139	0.0000000000	0.0000000000

$$S(x_1, ..., x_n) \ = \ \frac{x_{(n)}}{\sum_i \ x_i}$$

para o teste de homogeneidade das observações e

$$S(x_1, ..., x_n) \ > \ c \tag{6.7}$$

é a correspondente região de rejeição.

Pontos críticos são apresentados na tabela III de Barnett e Lewis (Cf. [11] p. 473-4) para os níveis de significância habituais e para vários valores do parâmetro de forma.

Assim, sempre que se verifique a desigualdade (6.7), deve rejeitar-se a homogeneidade da amostra e seleccionar o máximo $x_{(n)}$ como *"outlier"*.

Caso II) Com a informação suplementar: $\delta > \delta'$

Se $\delta > \delta'$, embora desconhecidos, então sabemos ser o mínimo $x_{(1)}$ da amostra que é candidato a *"outlier"* e que deve ser utilizada a estatística

$$S(x_1, .., x_n) \ = \ \frac{x_{(1)}}{\sum_i \ x_i}$$

para teste de homogeneidade nas observações e agora temos

$$S(x_1, ..., x_n) \ < \ c$$

como região de rejeição. Para calcular os pontos críticos para a estatística S surgem, neste caso, grandes dificuldades. O algoritmo recursivo construído por Fieller [46] permite determinar a distribuição de S. Mas, quando n e ϕ aumentam, as expressões de recorrência tornam-se muito difíceis de aplicar. Então este caso tem grandes limitações na sua aplicação logo para valores de n igual a 5 ou 6.

Caso III) Sem informação suplementar

Se δ e δ', são desconhecidos, e não dispomos de qualquer informação suplementar, estamos no caso mais geral para este estudo de *"outliers"* em populações exponenciais.

Como vimos, a partir de (4.12), o método GAN propõe que se use a estatística

$$S(x_1, ..., x_n) \ = \ min_j \ \frac{x_j}{\sum_i \ x_i} \ (1 \ - \ \frac{x_j}{\sum_i \ x_i})^{n-1} \tag{6.8}$$

que já estudámos em (5.2.3), e donde se pode concluir que

$$S(x_1, ..., x_n) \ < \ c \tag{6.9}$$

é a correspondente região de rejeição da hipótese H_0. Se, para uma amostra aquela desigualdade for verificada, rejeitaremos a hipótese de homogeneidade e, na fase seguinte, seleccionaremos, como *"outlier"*, a observação responsável por essa rejeição. Para uma completa aplicação do teste necessitamos então que se estude a função de distribuição de S e a consequente determinação dos valores críticos de rejeição c.

Sejam

$$T_j \ = \ \frac{X_j}{\sum_i \ x_i} \qquad j = 1, ..., n$$

e seja $T_{(1)} \leq ... \leq T_{(n)}$ a correspondente amostra ordenada. Um estudo semelhante ao que foi feito em (5.2.3) permite concluir que a distribuição conjunta do par $(T_{(1)}, T_{(n)})$ é fundamental para calcular c. Esta distribuição foi estudada por Fieller [46], mas com as dificuldades já enunciadas para o caso anterior. Portanto, embora resolvido do ponto de vista teórico, surge o problema prático da determinação dos valores críticos c - eventualmente por estudos de simulação.

6.3 Método GAN - p "outliers"

Numa generalização imediata da metodologia proposta ao longo do capítulo anterior para populações exponenciais, podemos formular um modelo de discordância natural onde admitimos a presença de p *"outliers"*.

Com um estudo semelhante ao que fizemos em (5.2.4.2), podemos concluir que:

Pelo método GAN, em populações gama onde se formula um modelo de discordância com uma alternativa natural com p observações discordantes, apenas os vectores $(x_{(1)}, ..., x_{(p)})$ **- com os p mínimos - ou,** $(x_{(n-p+1)}, ..., x_{(n)})$ **- com os p máximos - podem ser candidatos a** *"outliers"*.

Capítulo 7

"Outliers" em Populações Normais

7.1 Introdução

O problema da detecção e do tratamento de dados que, aos olhos do estatístico, parecem discordantes - aberrantes, espúrios, surpreendentes ou atípicos - surgiu no momento em que o experimentador começou a efectuar análise de dados. Desde logo se apercebeu que as informações contidas nesses dados poderiam ser distorcidas ou alteradas profundamente devido à existência dos chamados outliers e daí surgiu a necessidade de os detectar e tratar. No entanto a forma de o fazer tem sido polémica e sujeita a muitas divergências apresentando, ainda hoje como já vimos, alguns problemas e limitações.

É claro que, a importância do estudo de outliers em populações normais é tão grande como a frequência em que os dados são por elas produzidos para situações reais. Numa pequena polémica - podemos alinhar na perspectiva prática - considerando que "tudo na vida é normal" ou para lá tende...

Se o aceitarmos temos uma explicação simples e primeira para a importância do estudo de outliers em dados estatísticos normais e nas mais diversas situações e modelos - desde os mais simples até aos mais estruturados.

Historicamente, a principal motivação para um problema de outliers - o tratamento estatístico de valores surpreendentes - surgiu, como temos referido, dos problemas práticos relacionados com observações em Astronomia e com a repetição e determinação de dados fundamentais.

Os "erros" verificados, muitas vezes, podem ser originados pelos di-

ferentes registos de diversos operadores e/ou por terem sido utilizados vários aparelhos. Estes pressupostos também permitem concluir que aquele problema pode envolver metodologias estatísticas onde o estudo da influência da escala se torna muito importante. Também nesta perspectiva aparece a relevância estatística da distribuição normal na teoria dos erros e assim, desde sempre e mais uma vez, foi considerada da maior importância a resolução do problema da existência de valores aberrantes em dados normais.

Neste contexto surgiram os mais variados critérios para rejeição de observações que, empiricamente, parecem suspeitas ao experimentador. Como sabemos, o primeiro critério não-subjectivo para rejeição de uma observação discordante é atribuído ao astrónomo Chauvenet [25].

O "teste de discordância" definido por Chauvenet[1] não se aplicando a uma lei de probabilidade específica, aconselha a rejeição de qualquer observação que se "afaste muito" da média das restantes.

Nas mais diversas obras, em particular para as populações normais, são inúmeros os testes de discordância formulados com base em critérios de rejeição do tipo daquele introduzido por Chauvenet.

Beckman e Cook (Cf. [16] p. 124-32) apresentam um excelente estudo (também) histórico sobre outliers em populações normais, também argumentando (*ib.* p. 119) que, a literatura sobre outliers é muito vasta e, além disso, tem bastante em comum com muitas outras áreas.

A presença de observações discordantes também pode gerar problemas de falta de robustez. Este é um tema[2] concomitante com a "estatística dos outliers".

Na génese da moderna teoria de *"outliers"* em dados estatísticos gerados por populações normais são considerados fundamentais os modelos de discordância definidos por Ferguson [44] e [45] e os testes apresentados por Dixon [35] e [37].

Na realidade, também do ponto de vista histórico, os modelos de discordância de Ferguson - e que apresentámos no capítulo 4 - constituem uma primeira formulação onde é introduzida uma hipótese alternativa geral para populações normais e que reflecte mudanças nos parâmetros de localização e de escala. É também bastante geral quanto ao número de

[1] O critério de Chauvenet, fundamentalmente prático, é "aconselhado" em muitas situações práticas das diversas ciências. De acordo com esse autor, como também se refere em Barnett e Lewis ([11] p. 4): "*Any result of a series containing n observations shall be rejected when the magnitude of its deviation from the mean of all measurements is such that the probability of occurrence of all deviations as large or larger is less than* $1/2n$".

[2] Sobre este assunto - também salientando a contribuição portuguesa - remete-se o leitor para o recente estudo de Branco [19].

outliers que se podem admitir na amostra. Com esta nova metodologia inicia-se a moderna teoria para estudo de outliers em dados estatísticos.

Para as populações normais devem também referir-se os trabalhos de Dixon [37] onde são formulados aqueles a que ainda se hoje chama "testes de tipo Dixon". Nesses estudos são, acima de tudo, propostos alguns objectivos fundamentais: "*Many authors have written on the subject of the rejection of outlying observations. Apparently none have been successful in obtaining a general solution to the problem. Nor has been success in the development of a criterion for discovery of outliers by means of a general statistical theory: e. g. maximum likelihood*".

Conforme salientámos em 2.8.4, também para as populações normais, esta (ainda) é uma questão fundamental - quiçá um dos principais problemas teóricos da análise estatística de *"outliers"*.

Na sequência destes resultados iniciais, foram avançadas[3] diversas propostas metodológicas. Pela discrepância registada nos diferentes estudos, ao longo do tempo propostos, devemos realçar que, a grande maioria dos testes apresentados pelos mais diversos autores, são destinados a observações previamente seleccionadas pelo analista, mais concretamente, para o máximo $x_{(n)}$ da amostra. É interessante[4] registar este facto.

Além disso, perante um caso concreto, com essas formulações, ficará sempre por resolver a escolha, para estudo, entre uma ou outra daquelas observações extremas.

Numa amostra de dados estatísticos normais, quando é que um máximo é "demasiado máximo" ou um mínimo é "demasiado mínimo"?

Por qual das discrepâncias nos devemos decidir - a do máximo ou a do mínimo?

Também neste ponto do estudo de "outliers" em populações normais, e de novo, é oportuno referir o trabalho [28] de Collett e Lewis sobre subjectividade.

---

[3]Um primeiro estudo, de referência, é a tese de doutoramento de Fieller [46]. Em seguida surgiu a publicação do excelente tratado *Outliers in Statistical Data* de Barnett e Lewis [11], que já diversas vezes referimos e cuja primeira edição foi publicada em 1978.

Em 1984, Rosado [112] apresentou uma nova abordagem metodológica para o estudo de "outliers" (também) para populações normais e que desenvolveremos neste capítulo.

Pela generalidade, deve também referir-se o trabalho [89] de Munoz Garcia e outros.

De um ponto de vista das aplicações e com uma extensa lista de testes de discordância para populações normais, deve-se consultar Barnett e Lewis ([11], p. 216-50).

[4]Consulte-se ([11], p. 216-50) onde podemos encontrar mais de uma dezena de testes para o máximo de uma amostra de normais sem que se tenha algum teste específico para o mínimo. Esta observação é apenas estudada em múltiplos outliers, subjectivamente ligada àquela.

Nas secções seguintes tentaremos encontrar respostas para estas questões.

No capítulo 8 avaliaremos o desempenho, das diversas propostas, num estudo de outliers em populações normais.

7.2 Método GAN - um "outlier"

7.2.1 Modelo de Discordância com Alternativa Natural

Além das dificuldades[5] que por diversas vezes já apontámos, o estudo - a selecção e a detecção - de *"outliers"* é ainda agravado no momento da escolha da estatística de teste a adoptar face à grande variedade "da oferta". Esta agravante toma ainda mais relevo se a análise envolve a mais estudada de todas as populações.

Para amostras geradas por uma população normal, Barnett e Lewis [11], para o estudo de observações discordantes seleccionadas *a priori* propõem mais de quarenta testes. A opção por uma ou outra estatística é uma questão fundamental pois uma observação poderá ser considerada *"outlier"* por um teste e não o ser relativamente a outro.

Portanto, a problemática do estudo de *"outliers"* em dados estatísticos normais é agravada pelo excesso de oferta de abordagens científicas como é salientado por Barnett e Lewis (Cf. [11] p. 216-7) onde se reconhece a necessidade de sistematização - também quanto à informação disponível sobre os parâmetros de localização e/ou de escala para aquelas distribuições.

Para o estudo de *"outliers"* em dados estatísticos normais o método GAN introduzido no capítulo 4, no caso mais geral para populações normais, considera o seguinte modelo de discordância:

- Pela hipótese H_0 - de *homogeneidade* - admitimos, como aliás na generalidade dos estudos sobre "outliers", que todas as observações $x_1, ..., x_n$ têm a mesma distribuição normal $N(\mu, \sigma)$ com densidade

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

  Nesta hipótese H_0, a verosimilhança é

$$L(x_1, ..., x_n; \mu, \sigma) = \prod_i^n f(x_i; \mu, \sigma).$$

[5] De entre estas, no contexto actual, deve-se relevar a subjectividade - nas suas diversas vertentes, desde o analista ao modelo.

- Pela hipótese $\bar{H}$ - a *alternativa natural* - admitimos a presença de um valor discordante na amostra e que, em princípio, pode ser uma qualquer das n observações.

  Seja então $\bar{H}_j$ a hipótese que admite x_j como observação discordante, isto é, tal que:

  - x_j tem densidade de probabilidade $f(x_j; \mu', \sigma')$, para algum índice j∈(1,...,n)
  - $x_1, ..., x_{j-1}, x_{j+1}, ..., x_n$ seguem a mesma distribuição com densidade $f(x_i; \mu, \sigma)$ para $i \neq j$.

Este modelo de discordância geral reveste-se de enormes dificuldades de cálculo na obtenção das várias estatísticas para os testes de homogeneidade e, por consequência, também dos respectivos pontos críticos. Estes obstáculos poderiam ser ultrapassados com o recurso aos métodos computacionais. No entanto estaria em causa também a aplicabilidade uma vez que a interpretação prática das eventuais causas da presença de valores discordantes numa amostra evidenciam sempre a localização ou a escala como a principal razão dessa discordância. Será portanto feita uma abordagem metodológica onde optamos por um modelo geral mas dentro de um domínio mais simples de interpretação para o mecanismo gerador dos *"outliers"*. O nosso estudo vai então abordar cada uma dessas potenciais razões em separado. Assim, para as populações normais estudaremos *"outliers por μ"*, resultantes de discordância em localização e *"outliers por σ"*, se o mecanismo perturbador na geração é activado pela mudança na escala. Para ambos os modelos deve salientar-se a maior generalidade[6] na abordagem quer pelo modelo utilizado quer pela objectividade que é introduzida no estudo.

7.2.2 Modelo de discordância com "outlier" por μ

O modelo de discordância com *"outlier"* por μ admite que o parâmetro σ é conhecido.

Além disso, considera a hipótese H_0, de homogeneidade, formulada no modelo de discordância natural da secção anterior; são portanto admitidas todas as observações com a mesma densidade normal $f(x; \mu, \sigma)$.

Para hipótese alternativa $\bar{H}_j$, neste modelo, vamos supor

$$x_j \textit{ com densidade } f(x_j; \mu', \sigma) \textit{ para algum índice } j \in (1, ..., n)$$

[6]Como já referimos, também para as populações normais, esta metodologia contém os modelos "slippage" formulados por Ferguson [44] e [45].

e, as restantes observações

$$x_1, ..., x_{j-1}, x_{j+1}, ..., x_n \text{ com densidade } f(x_i; \mu, \sigma) \; i = 1, ..., n \; (i \neq j).$$

7.2.2.1 μ e μ' conhecidos

Se, neste modelo de discordância com *"outlier"* por μ, conhecemos os dois parâmetros de localização então, a função verosimilhança da amostra na hipótese H_0 de homogeneidade tem o máximo

$$\widehat{L}_0 \; = \; \frac{1}{(\sigma\sqrt{2\pi})^n} \; exp \; (- \; \frac{\sum_i (x_i - \mu)^2}{2\sigma^2})$$

e, na hipótese alternativa natural, é

$$\widehat{L}_j \; = \; \frac{1}{(\sigma\sqrt{2\pi})^n} \; exp \; (- \; \frac{\sum_i (x_i - \mu)^2 - (x_j - \mu)^2 + (x_j - \mu')^2}{2\sigma^2}).$$

O Método GAN, introduzido no capítulo 4, exige que se calcule a estatística de teste (4.12) que para o modelo de discordância em estudo conduz a

$$S(x_1, ..., x_n) \; = \; max_j \; (\mu' - \mu) \; \frac{x_j - \mu}{\sigma}$$

como estatística equivalente, para o teste de homogeneidade das observações.

Assim, se os parâmetros de localização são conhecidos, podemos concluir que:

- se $\mu < \mu'$, então $x_{(n)}$ é candidato a *"outlier"*
- se $\mu > \mu'$, então $x_{(1)}$ é candidato a *"outlier"*.

Portanto, nestes modelos, com alternativa natural, onde se admite a possibilidade de existir um valor discordante na amostra, apenas os extremos podem ser candidatos a *"outlier"*.

7.2.2.2 μ conhecido e μ' desconhecido

Neste modelo, na hipótese alternativa $\bar{H}_j$, o parâmetro desconhecido admite $\widehat{\mu}'_j = x_j$, como estimador de máxima verosimilhança.

Temos, portanto

$$S(x_1, ..., x_n) \; = \; max_j \; |\frac{x_j - \mu}{\sigma}|$$

como estatística para teste de homogeneidade.

De novo podemos concluir que apenas os extremos podem ser candidatos a *"outlier"*.

7.2.2.3 μ desconhecido e μ' conhecido

Para este caso, em termos de máxima verosimilhança obtemos

$$\widehat{\mu} \;=\; \bar{x}$$

e

$$\widehat{\mu}_j \;=\; \bar{x} \;-\; \frac{x_j - \bar{x}}{n-1}$$

como estimadores para o parâmetro desconhecido na hipótese nula e na alternativa natural respectivamente.

O método GAN propõe

$$S(x_1, ..., x_n) \;=\; max_j \; (\frac{n}{n-1}(x_j - \bar{x})^2 \;-\; (x_j - \mu')^2)$$

como estatística para o teste de homogeneidade. Este não é um caso de relevo dada a situação particular que é admitida para os parâmetros pelo que não será considerado no estudo seguinte.

7.2.2.4 μ e μ' desconhecidos

Neste caso, mais geral, desconhecendo ambos os parâmetros de localização, temos os estimadores de máxima verosimilhança

$$\widehat{\mu} \;=\; \bar{x}$$

na hipótese nula e,

$$\widehat{\mu}_j \;=\; \bar{x} \;-\; \frac{x_j - \bar{x}}{n-1}$$

$$\widehat{\mu'}_j \;=\; x_j$$

na hipótese alternativa natural.

A estatística para o teste de homogeneidade é

$$S(x_1, ...x_n) \;=\; max_j \; |\frac{x_j - \bar{x}}{\sigma}|.$$

Como candidato a *"outlier"*temos, também neste caso, apenas o mínimo ou o máximo da amostra.

Quanto à selecção do candidato a *"outlier"* potenciado por uma alteração no parâmetro de localização, podemos resumir o estudo[7] feito, até este momento, no quadro seguinte.

Método GAN em Populações Normais
Modelo de Discordância com "*outlier*" por μ

μ	μ'	Inf.Sup.	Cand. a *"outlier"*	Estatística S
conh	conh	$\mu > \mu'$	$x_{(1)}$	$\frac{X_{(1)}-\mu}{\sigma}$
		$\mu < \mu'$	$x_{(n)}$	$\frac{X_{(n)}-\mu}{\sigma}$
conh	desc		$x_{(1)}$ ou $x_{(n)}$	$max_j \mid \frac{X_j-\mu}{\sigma} \mid$
desc	desc		$x_{(1)}$ ou $x_{(n)}$	$max_j \mid \frac{X_j-\bar{x}}{\sigma} \mid$

7.2.3 Modelo de discordância com "outlier" por σ

O modelo de discordância com *"outlier"* por σ admite que o parâmetro μ é conhecido.

Além disso, considera a hipótese H_0, de homogeneidade, formulada no modelo de discordância natural da secção 7.2.1; sendo portanto admitidas todas as observações com a mesma densidade $f(x;\mu,\sigma)$.

Para hipótese alternativa $\bar{H}_j$, neste modelo, vamos supor

$$x_j \textit{ com densidade } f(x_j;\mu,\sigma') \textit{ para algum índice } j \in (1,...,n)$$

e, as restantes observações

$$x_1,...,x_{j-1},x_{j+1},...,x_n \textit{ com densidade } f(x_i;\mu,\sigma)\ i=1,...,n\ (i \neq j).$$

[7] No recente estudo elaborado por Palma [95], em particular no capítulo 2, são analisados outros diferentes casos sobre o conhecimento dos diversos parâmetros envolvidos nos modelos de discordância (Cf. [95], p. 31). É uma consulta que se recomenda ao leitor. No entanto, o fundamental está resumido nos quadros que apresentamos.

7.2.3.1 σ e σ' conhecidos

Se, neste modelo de discordância com *"outlier"* por σ, conhecemos os dois parâmetros de escala então, a função verosimilhança da amostra, na hipótese H_0 de homogeneidade, tem o máximo

$$\widehat{L}_0 \;=\; \frac{1}{(\sigma\sqrt{2\pi})^n}\; exp\;(-\;\frac{\sum_i (x_i-\mu)^2}{2\sigma^2})$$

e, na hipótese alternativa natural é

$$\widehat{L}_j \;=\; \frac{1}{\sigma^{n-1}\sigma'(\sqrt{2\pi})^n}\; exp\;(-\;\frac{1}{2}\;(\sum_{i\neq j}(\frac{x_i-\mu}{\sigma})^2+(\frac{x_j-\mu}{\sigma})^2)).$$

O Método GAN, introduzido no capítulo 4, determina que se obtenha a estatística de teste (4.12) que para o modelo de discordância em estudo conduz a

$$S(x_1,...,x_n) \;=\; max_j\;(\sigma'^2-\sigma^2)\;(\frac{x_j-\mu}{\sigma})^2$$

como estatística equivalente, para o teste de homogeneidade das observações.

Neste modelo de discordância com *"outlier"* por σ, é muito importante analisar os diferentes casos, quanto ao conhecimento dos parâmetros de escala. Esta análise vai, mais uma vez, salientar a propriedade generativa do método GAN e onde vão estudar-se como candidatos a *"outlier"* algumas observações que, aos olhos do experimentador, podem não ser suspeitas e que, tradicionalmente, não são consideradas como relevantes. É um resultado estatístico da máxima importância também porque potencia um estudo objectivo.

Se os parâmetros são conhecidos, a partir da estatística anterior, podemos determinar quais as observações onde o máximo é atingido.

Assim, se $\sigma < \sigma'$, então $x_{(1)}$ ou $x_{(n)}$ são os únicos candidatos a *"outlier"*.

E se, pelo contrário $\sigma > \sigma'$, então podemos obter, para o teste de homogeneidade da amostra, como estatística equivalente

$$S(x_1,...,x_n) \;=\; min_j\;(\frac{x_j-\mu}{\sigma})^2. \tag{7.1}$$

Esta estatística introduz uma nova informação sobre o mecanismo de geração da amostra.

De facto, o mínimo de $S(x_1, ..., x_n)$ é atingido em $x_{[\mu]}$ - a observação mais próxima do valor médio μ - e que assim é introduzida como candidata a *"outlier"*.

Esta observação é gerada pelo método GAN e, em termos de máxima verosimilhança, é a responsável pela rejeição da homogeneidade das observações. Deve, portanto, ser objecto do estudo de observações discordantes na amostra.

Nestes modelos, com alternativa natural, onde se admite a possibilidade de existir um valor discordante na amostra podemos, desde já, concluir que não são apenas os extremos que nos devem surpreender como candidatos a *"outlier"*. Deve salientar-se que, até aqui, apenas os extremos da amostra eram admitidos como surpreendentes e que este "terceiro candidato" surge como uma consequência da propriedade generativa do método GAN.

7.2.3.2 σ conhecido e σ' desconhecido

Neste modelo, na hipótese alternativa $\bar{H}_j$, o parâmetro desconhecido admite

$$\widehat{\sigma}'_j = |x_j - \mu|$$

como estimador de máxima verosimilhança.

Temos, portanto

$$S(x_1, ..., x_n) \;=\; max_j \; (\frac{|x_j - \mu|}{\sigma})^{-1} \; exp \; (\; \frac{1}{2} \; (\frac{x_j - \mu}{\sigma})^2)$$

como estatística para teste de homogeneidade.

Com um breve estudo da estatística S, o método GAN, neste caso, fornece como candidatas a *"outlier"*, as observações $x_{[\mu]}$, $x_{(1)}$ ou $x_{(n)}$, já consideradas.

7.2.3.3 σ e σ' desconhecidos

Neste caso, mais geral, desconhecendo ambos os parâmetros de localização, temos os estimadores de máxima verosimilhança

$$\widehat{\sigma}^2 \;=\; s^2[\mu] = \; \frac{1}{n} \sum_i \; (x_i - \mu)^2$$

na hipótese nula e,

$$\widehat{\sigma}^2_j \;=\; \frac{1}{n-1} \sum_{i \neq j} \; (x_i - \mu)^2$$

$$\widehat{\sigma}_j'^2 \;=\; (x_j \;-\; \mu)^2$$

na hipótese alternativa natural.

A estatística para o teste de homogeneidade é

$$S(x_1, ...x_n) \;=\; min_j\; (\frac{(x_j-\mu)^2}{\sum_i (x_i-\mu)^2})^{\frac{1}{2}}\; (1 \;-\; \frac{(x_j-\mu)^2}{\sum_i (x_i-\mu)^2})^{\frac{n-1}{2}}.$$

Considerando a estatística S como uma função de z tal que

$$g(z) \;=\; z^{\frac{1}{2}}\; (1 \;-\; z)^{\frac{n-1}{2}}$$

com

$$z \;=\; \frac{(x-\mu)^2}{\sum_i (x_i-\mu)^2}$$

e sendo $g(o) = g(1) = 0$, vemos que o mínimo de g é atingido quando z é mínimo ou máximo.

Ora, z é mínimo em $x_{[\mu]}$ e é máximo em $x_{(1)}$ ou $x_{(n)}$. Assim, neste modelo de discordância, são também estas - e, de novo, apenas estas! - as observações eventualmente discordantes.

Método GAN em Populações Normais
Modelo de Discordância com "*outlier*" por σ

σ	σ'	Cand. a *"outlier"*	Estatística S
conh	conh	$x_{(1)}$ *ou* $x_{(n)}$ se $\sigma < \sigma'$	$max_j\; \lvert\frac{x_j-\mu}{\sigma}\rvert$
		$x_{[\mu]}$ se $\sigma > \sigma'$	$min_j\; \lvert\frac{x_j-\mu}{\sigma}\rvert$
conh	desc	$x_{(1)},\; x_{(n)}$ ou $x_{[\mu]}$	$max_j\; (\frac{\lvert x_j-\mu\rvert}{\sigma})^{-1}\; exp\; (\;\frac{1}{2}\; (\frac{x_j-\mu}{\sigma})^2)$
desc	desc	$x_{(1)},\; x_{(n)}$ ou $x_{[\mu]}$	$min_j\; (\frac{(x_j-\mu)^2}{\Sigma_i (x_i-\mu)^2})^{\frac{1}{2}}\; (1-\frac{(x_j-\mu)^2}{\Sigma_i (x_i-\mu)^2})^{\frac{n-1}{2}}$

Quanto à selecção do candidato a *"outlier"* potenciado por uma alte-

ração no parâmetro de escala, podemos resumir o estudo[8], feito até este momento, no quadro acima.

7.2.4 Testes de Homogeneidade; seus valores críticos

7.2.4.1 Modelo de discordância com "outlier" por μ

Caso I) μ e μ' **conhecidos**

Admitamos $\mu < \mu'$.

Neste caso sabemos que

$$S_1(x_1, ..., x_n) \ = \ \frac{X_{(n)} - \mu}{\sigma}$$

é a estatística para o teste de homogeneidade da amostra e,

$$S_1(x_1, ..., x_n) \ > \ c$$

é a correspondente região de rejeição.

A distribuição de S_1 é bem conhecida.

As tabelas, construídas em Pearson e Hartley [97], por exemplo, permitem determinar os pontos críticos c. Estas tabelas foram reproduzidas e completadas por Barnett e Lewis (Cf. [11], p. 241 e 245), para o caso de amostras com dimensão superior a 30.

Palma [95], também determinou tabelas[9] para esta estatística.

Na tabela 1 apresentamos alguns pontos críticos.

Admitindo que, para uma determinada amostra, a correspondente estatística S_1 não verifica a condição acima então, o método GAN propõe a aceitação da homogeneidade nas observações e termina a análise de *"outliers"* nessa amostra.

Se, pelo contrário, aquela condição é verificada, então é rejeitada a homogeneidade da amostra e $x_{(n)}$ - a observação responsável por essa decisão - deve ser *a posteriori* seleccionado como *"outlier"*.

Admitamos $\mu > \mu'$.

Neste caso, temos

$$S_2(x_1, ..., x_n) \ = \ \frac{X_{(1)} - \mu}{\sigma}$$

[8]Tal como referimos para o modelo de discordância com "outlier" por μ, no recente estudo elaborado por Palma [95], em particular no capítulo 2, são analisados outros diferentes casos sobre o conhecimento dos diversos parâmetros envolvidos nos modelos de discordância (Cf. [95], p.31). É uma consulta que se recomenda ao leitor. No entanto, o fundamental está resumido no quadro que apresentamos.

[9]Cf. [95], tabela A1, p. 187.

Tabela 1: Pontos críticos para S_1.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	2.88	2.32
10	3.09	2.57
15	3.21	2.71
20	3.29	2.80
25	3.35	2.87
30	3.40	2.93
40	3.48	3.02
50	3.54	3.08
100	3.72	3.28

como estatística para testar a homogeneidade da amostra e,

$$S_2(x_1, ..., x_n) \ < \ c$$

é a correspondente região de rejeição. Também aqui podemos usar as tabelas de Barnett e Lewis (Cf. [11], p. 485 e seg) ou o estudo de Palma [95]. A simetria distribucional permite ainda que se use a tabela 1 com as devidas adaptações.

Se, para uma determinada amostra, a correspondente estatística S_2 não verifica a condição acima, então o método GAN, propõe a aceitação da homogeneidade nas observações e, nesta fase, termina a análise de *"outliers"*.

Se, pelo contrário, aquela condição é verificada, então é rejeitada a homogeneidade da amostra e $x_{(1)}$ - a observação responsável por essa decisão - é *a posteriori* selecionado como *"outlier"*.

Caso II) μ conhecido e μ' desconhecido

O método GAN conduz neste caso à estatística

$$S_3(x_1, ...x_n) \ = \ max_j \ |\frac{x_j - \mu}{\sigma}|$$

e

$$S_3(x_1, ...x_n) \ > \ c$$

é a região de rejeição da homogeneidade.

Nesta hipótese as variáveis aleatórias envolvidas são qui-quadrado com um grau de liberdade e portanto

Tabela 2: Pontos críticos para S_3.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	3.09	2.57
10	3.29	2.80
15	3.40	2.93
20	3.48	3.02
25	3.54	3.08
30	3.59	3.14
40	3.66	3.22
50	3.72	3.28
100	3.89	3.47

$$c \;=\; F^{-1}_{\chi^2_1}(1-\alpha)^{\frac{1}{n}}.$$

Na tabela 2 apresentamos alguns pontos críticos.

Admitindo rejeitada a homogeneidade da amostra surgem como candidatos a "outlier" apenas $x_{(1)}$ ou $x_{(n)}$.

Como sabemos, será *"outlier"* a observação que mais se afastar de μ.

Caso III) μ e μ' desconhecidos

No estudo que apresentámos em 7.2.2.4, o método GAN nestas hipóteses, mais gerais, propõe o uso de

$$S_4(x_1, ...x_n) \;=\; max_j \; |\frac{x_j - \bar{x}}{\sigma}|$$

para o teste de homogeneidade e,

$$S_4(x_1, ...x_n) \;>\; c$$

é a região de rejeição.

Ora, este teste é equivalente ao teste Nσ2 utilizado por Barnett e Lewis (Cf. [11] p. 246). Podemos utilizar a respectiva tabela XIIIf (*ib.* p. 486) para obter os pontos críticos c. Esta estatística também foi estudada por Palma [95].

Como referência, na tabela 3, apresentamos alguns pontos críticos aproximados a menos de 0.01.[10]

[10]Para obter mais informação sobre estes e outros valores críticos deve-se consultar o recente estudo de Palma onde são consideradas as hipóteses adicionais de se dispor

Tabela 3: Pontos críticos para S_4.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	2.765	2.276
10	3.110	2.645
15	3.285	2.822
20	3.383	2.940
25	3.472	3.023
30	3.506	3.089
40	3.598	3.175
50	3.671	3.251
100	3.877	3.453

7.2.4.2 Modelo de discordância com "outlier" por σ

Caso I) σ e σ' **conhecidos**

Admitamos $\sigma < \sigma'$.
Neste caso sabemos que

$$S_5(x_1, ..., x_n) \ = \ max_j \ |\frac{x_j - \mu}{\sigma}|$$

é a estatística para o teste de homogeneidade da amostra e,

$$S_5(x_1, ..., x_n) \ > \ c$$

é a correspondente região de rejeição.

Fixado um nível de significância, o correspondente ponto crítico c pode ser obtido através da função gama incompleta, tabelada em Pearson e Hartley [97] ou Abramowitz e Stegun [1].

Como referência, na tabela 4, apresentamos[11] alguns pontos críticos.

Admitindo que, para uma determinada amostra, a correspondente estatística S_5 não verifica a condição acima, então o método GAN propõe a aceitação da homogeneidade nas observações e termina a análise de *"outliers"* nessa amostra.

Se, pelo contrário, aquela condição é verificada, então é rejeitada a homogeneidade da amostra e surgem como candidatos apenas $x_{(1)}$ ou

---

de alguma informação sobre a ordenação dos parâmetros desconhecidos.
(Cf. [95], tabelas A3, A4 e A5; p. 189-90).

[11] Para obter mais informação sobre estes e outros valores críticos deve-se consultar o recente estudo de Palma (Cf. [95], tabelas A9, A10 e A11 p. 193-4).

Tabela 4: Pontos críticos para S_5.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	1.764	1.715
10	2.480	2.290
15	2.802	2.547
20	2.997	2.708
25	3.125	2.821
30	3.227	2.907
40	3.377	3.033
50	3.466	3.122
100	3.740	3.376

$x_{(n)}$.

Seleccionaremos, como *"outlier"*, aquela observação que dê o maior valor $(x - \mu)^2$.

Admitamos $\sigma > \sigma'$.

Neste caso sabemos que

$$S_6(x_1, ..., x_n) \ = \ min_j \ |\frac{x_j - \mu}{\sigma}|$$

é a estatística para o teste de homogeneidade da amostra e,

$$S_6(x_1, ..., x_n) \ < \ c$$

é a correspondente região de rejeição.

Como referência, na tabela 5 apresentamos alguns pontos críticos.[12]

Se, para uma determinada amostra, a correspondente estatística S_6 não verifica a condição acima então, o método GAN propõe a aceitação da homogeneidade nas observações e, nesta fase, termina a análise de *"outliers"*. Se, pelo contrário, aquela condição é verificada, então é rejeitada a homogeneidade da amostra e a observação $x_{[\mu]}$ mais perto do valor médio conhecido é responsável por essa decisão e, *a posteriori*, seleccionada como *"outlier"*.

Esta decisão, baseada em princípios de máxima verosimilhança, é inovadora no tratamento de *"outliers"* em dados estatísticos e fundamenta a definição de observação discordante numa amostra que, assim, é objectivamente introduzida.

[12] Para obter mais informação deve-se consultar o recente estudo de Palma. (Cf. [95], tabela A15, p. 197).

Tabela 5: Pontos críticos para S_6.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	0.0025	0.013
10	0.0013	0.0064
15	0.00084	0.0043
20	0.00063	0.0032
25	0.00050	0.0026
30	0.00042	0.0021
40	0.00031	0.0016
50	0.00025	0.0013
100	0.00013	0.00064

Caso II) σ conhecido e σ' desconhecido

O método GAN conduz neste caso à estatística

$$S_7(x_1, ...x_n) \ = max_j \ (\frac{\sigma}{|x_j - \mu|})^{-1} \ exp \ (\ \frac{1}{2} \ (\frac{x_j - \mu}{\sigma})^2$$

e

$$S_7(x_1, ...x_n) \ > \ c$$

é a região de rejeição da hipótese H_0 de homogeneidade.

Para determinar os pontos críticos c deve ter-se para a respectiva função de distribuição, na hipótese H_0, $F_{S_7}(c) = \alpha$.

Ora, se considerarmos a função

$$g(z) \ = \ \frac{1}{z} \ exp \ (\frac{1}{2} z^2)$$

com

$$z \ = \ \frac{|x - \mu|}{\sigma}$$

temos

$$\begin{aligned} Prob \ \{S_7 < c\} \ &= \ Prob \ \{ \ c_1 < z_i < c_2, \ i = 1, ..., n\} \\ &= \ (Prob \ \{ \ c_1 < z < c_2 \ \})^n \end{aligned}$$

e onde c_1 *e* c_2 são constantes tais que $g(c_1) = g(c_2) = c$.

Tabela 6: Pontos críticos para S_7.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	422.7	84.6
10	849.6	167.4
15	1257.3	249.8
20	1673.8	332.0
25	2088.8	414.2
30	2503.8	496.2
40	3335.0	660.0
50	4162.5	823.7
100	8296.9	1639.4

Mas,

$$\begin{aligned} Prob\,\{Z < c\} &= 0 \qquad\qquad se \quad c < 0 \\ &= 2\,\Phi(c) - 1 \quad se \quad c > 0 \end{aligned}$$

onde Φ representa a função de distribuição da Normal $(0, 1)$.

Portanto,

$$Prob\,\{\ c_1 < z < c_2\ \} \;=\; 2\,(\Phi(c_2) - \Phi(c_1)).$$

Para o ponto crítico c, tal que $g(c_1) = g(c_2) = c$, devemos ter

$$\Phi(c_2) - \Phi(c_1) \;=\; \frac{1}{2}\,(1-\alpha)^{\frac{1}{n}}.$$

Como referência, na tabela 6 apresentamos alguns pontos críticos.[13]

Se, para uma determinada amostra, o método GAN rejeita a homogeneidade, então é declarado *"outlier"* o valor - escolhido entre $x_{[\mu]}$, $x_{(1)}$ e $x_{(n)}$ - que maximiza a expressão

$$\frac{\sigma}{|x_j - \mu|})^{-1}\ exp\ (\ \frac{1}{2}\ (\frac{x_j - \mu}{\sigma})^2).$$

[13]Para obter mais informação deve-se consultar o estudo de Palma. (Cf. [95], tabela A16 p. 198).

Caso III) σ e σ' desconhecidos

No estudo que apresentámos em 7.2.3.3, para este modelo de discordância, obtivemos

$$S_8(x_1, ..., x_n) \;=\; min_j \; (\frac{(x_j - \mu)^2}{\Sigma_i (x_i - \mu)^2})^{\frac{1}{2}} \; (1 - \frac{(x_j - \mu)^2}{\Sigma_i (x_i - \mu)^2})^{\frac{n-1}{2}}$$

como estatística para testar H_0, a hipótese de homogeneidade da amostra e,

$$S_8(x_1, ..., x_n) \;<\; c$$

é a correspondente região de rejeição.

Se considerarmos a função

$$g(z) \;=\; z^{\frac{1}{2}} \; (1 - z)^{\frac{n-1}{2}}$$

com

$$z \;=\; \frac{(x - \mu)^2}{\Sigma_i (x_i - \mu)^2}$$

devemos ter, na hipótese H_0, para um nível de significância α

$$\begin{aligned} Prob\ \{\ S_8 \;>\; c \;/\; H_0\} \;&=\; 1 - \alpha \\ &=\; Prob\ \{\ z_{(1)}^{\frac{1}{2}} (1 - z_{(1)})^{\frac{n-1}{2}} > c \;\wedge \\ &\quad \wedge \;\; z_{(n)}^{\frac{1}{2}} \; (1 - z_{(n)})^{\frac{n-1}{2}} \;>\; c \;\} \\ &=\; Prob\ \{\ c_1 < z_{(1)} < z_{(n)} < c_2 \;\} \end{aligned}$$

onde c_1 *e* c_2 são constantes tais que $g(c_1) = g(c_2) = c$.

A decisão sobre a homogeneidade na amostra é então tomada a partir da análise de $z_{(1)}$ e $z_{(n)}$. Ora, na hipótese H_0 de homogeneidade, sendo cada x_i com densidade normal $N(\mu, \sigma)$ temos cada Z_i a seguir uma distribuição de um quociente de dois qui-quadrados com um e n graus de liberdade, respectivamente. Assim, da análise anterior podemos concluir que neste modelo de discordância com "outlier"por σ, o

Tabela 7: Pontos críticos para S_8.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	0.00103	0.00557
10	0.000356	0.00189
15	0.000194	0.00103
20	0.000129	0.000687
25	0.0000948	0.000505
30	0.0000779	0.000382
40	0.0000449	0.000238
50	0.0000352	0.000176
100	0.0000123	0.0000622

teste de homogeneidade é equivalente ao correspondente para uma população qui-quadrado com um grau de liberdade. Daqui, infere-se que este estudo pode ser continuado, conforme os modelos e as hipóteses já estudadas no capítulo 6, com as populações gama.

Como referência, na tabela 7 apresentamos[14] alguns pontos críticos aproximados.

Uma vez rejeitada a homogeneidade das observações, como anteriormente estudámos, o método GAN indica como "outlier", aquela observação - $x_{[\mu]},\ x_{(1)}\ ou\ x_{(n)}$ - que minimiza a função correspondente à estatística de teste.

Caso IV) Com μ, σ e σ' desconhecidos

Temos estado a considerar que todas as observações têm o mesmo valor médio μ e que este parâmetro é conhecido. Além disso, nos modelos desta secção, foi admitido que a discordância do "outlier" poderá ser explicada por uma alteração na dispersão. Nalgumas situações práticas não disporemos daquela informação sobre a localização.

Suponhamos então que, no modelo de discordância com "outlier" por σ também o parâmetro μ é desconhecido. As equações de máxima verosimilhança não permitem encontrar facilmente a estatística para o teste de homegeneidade da amostra através da metodologia GAN. No entanto, estas hipóteses correspondem a uma situação prática importante para um modelo com normais.

Ora, se a localização é comum, consideremos o parâmetro μ estimado por $\bar{x}$, com base em todas as observações. Esta "estatística *ad hoc*" pode

[14] Para obter mais informação sobre estes e outros valores críticos deve-se consultar o recente estudo de Palma (Cf. [95], tabelas A17, A18 e A19 p. 199-200).

Tabela 8: Pontos críticos para S_9.

n	$\alpha = 0.01$	$\alpha = 0.05$
5	0.00143	0.00724
10	0.000436	0.00211
15	0.000220	0.00107
20	0.000135	0.000710
25	0.0000107	0.000497
30	0.0000772	0.000371
40	0.0000491	0.000247
50	0.0000345	0.000173
100	0.0000122	0.0000591

então ser utilizada para testar a homogeneidade de uma amostra gerada por uma localização conhecida $\mu = \bar{x}$. Estaremos na situação anterior.

Pelos pressupostos agora introduzidos podemos usar

$$S_9(x_1, ..., x_n) \;=\; min_j \; (\frac{(x_j - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2})^{\frac{1}{2}} \; (1 - \frac{(x_j - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2})^{\frac{n-1}{2}}$$

como estatística para testar H_0, a hipótese de homogeneidade da amostra e,

$$S_9(x_1, ..., x_n) \;<\; c$$

é a correspondente região de rejeição.

Como referência, na tabela 8 apresentamos[15] alguns pontos críticos aproximados.

Como no anterior, também neste caso, uma vez rejeitada a homogeneidade das observações, como anteriormente estudámos, o método GAN indica como "outlier", aquela observação[16] - $x_{[\mu]}$, $x_{(1)}$ *ou* $x_{(n)}$ - que minimiza a função correspondente à estatística de teste.

Em termos do desempenho, que abordaremos no capítulo 8, mas também como estudo comparado entre metodologias propõe-se ao leitor a consulta do tratado sobre outliers de Barnett e Lewis[17] onde, para o

[15]Para obter mais informação sobre estes e outros valores críticos deve-se consultar o recente estudo de Palma (Cf. [95], tabelas A26, A27 e A28 p. 205-6).

[16]Mantendo a notação, neste caso temos $x_{[\mu]}$ a representar aquela observação que mais se aproxime de $\bar{x}$.

[17](Cf. [11], p. 218 e seg.)

caso vertente, é utilizada a estatística

$$max\ \{\frac{x_{(n)}-\bar{x}}{s}\ ,\ \frac{\bar{x}-x_{(1)}}{s}\}$$

para aquele a que chamam "*two sided discordancy test of an extreme outlier in a normal sample with* μ *and* σ^2 *unknown*". Podemos, aqui de novo, encontrar a contribuição inovadora, introduzida pela observação $x_{[\mu]}$ que gerámos para o estudo de outliers em dados estatísticos.

7.3 Método GAN - p "outliers"

7.3.1 Modelo de discordância com dois "outliers"

Consideremos um modelo de discordância onde se admite que:

- Pela hipótese H_0, de homogeneidade, todas as observações com densidade de probabilidade normal $f(x;\mu,\sigma)$.
- Pela hipótese alternativa natural $\bar{H}_{jk}$, existem duas observações x_j *e* x_k discordantes, com densidade $f(x;\mu',\sigma)$. As restantes observações da amostra seguem a anterior densidade normal $f(x;\mu,\sigma)$.

Temos assim construído um modelo de discordância com alternativa natural. Admitamos que os parâmetros μ, μ' *e* σ são conhecidos. O Método GAN introduzido no capítulo 4 determina que se obtenha a estatística de teste (4.12) que, para o modelo em análise, conduz a

$$T(x_1,...,x_n)\ =\ max_{j\neq k}\ \frac{\widehat{L}_{jk}}{\widehat{L}_0}$$

e portanto

$$S(x_1,...,x_n)\ =\ max_{j\neq k}\ (\mu-\mu')\ \frac{x_j+x_k}{\sigma^2}$$

é uma estatística equivalente, para o teste de homogeneidade das observações.

Desta estatística S podemos concluir que:

- o par candidato a *"outlier"* é $(x_{(n-1)},x_{(n)})$ *se* $\mu<\mu'$
- o par candidato a *"outlier"* é $(x_{(1)},x_{(2)})$ *se* $\mu>\mu'$.

No modelo acima estudado, em termos de máxima verosimilhança, o par $(x_{(1)},x_{(n)})$ nunca deve ser considerado como observação discordante. Esta é uma conclusão retirada a partir do uso do método GAN e que, de

novo, introduz objectividade no estudo de *"outliers"* em dados estatísticos normais clarificando a própria definição de "par discordante" através da especificação do modelo utilizado.[18]

Resultados análogos podem ser obtidos se considerarmos diferentes modelos de discordância correspondentes a diferentes informações sobre o conhecimento dos diversos parâmetros envolvidos nas distribuições.

Deixamos esse exercício ao cuidado do leitor, tendo em mente que a principal conclusão se prende com o envolvimento dos "dois extremos" - os dois máximos ou os dois mínimos - e, portanto deixando sempre o par $(x_{(1)}, x_{(n)})$ como um *"outlier"* entre os outliers.

7.3.2 Modelo de discordância com p "outliers"

Numa generalização imediata da metodologia da secção anterior, consideremos um modelo de discordância onde se admite que:

- Pela hipótese H_0, de homogeneidade, todas as observações com densidade de probabilidade normal $f(x; \mu, \sigma)$.

- Pela hipótese alternativa natural $\bar{H}_{j_1,...,j_p}$, existem p observações $x_{j_1}, ..., x_{j_p}$ discordantes, com densidade $f(x; \mu', \sigma)$.

  As restantes observações da amostra seguem a anterior densidade normal $f(x; \mu, \sigma)$.

Um estudo semelhante ao efectuado na secção anterior através da análise das respectivas estatísticas para o teste de homogeneidade da amostra, onde no caso mais simples surge

$$S(x_1, ..., x_n) \; = \; max_{j_1 \neq ... \neq j_p} \; (x_{j_1} + ... + x_{j_p} - p\mu)^2$$

podemos agora concluir que os candidatos a "outlier" envolvem as p primeiras ou as p últimas estatísticas ordinais.

Assim, introduzimos objectividade no estudo de p observações discordantes numa amostra de dados estatísticos normais quando se admitem apenas $(x_{(1)}, ..., x_{(p)})$ ou $(x_{(n-p+1)}, ..., x_{(n)})$ como candidatos a *"outlier"*.

[18] Como já foi indicado, para um estudo comparativo e de aprofundamento sugere-se a leitura da secção Testes de Discordância para Amostras Normais de Barnett e Lewis ([11], p. 216-50), desta vez, em particular os testes $N\mu\sigma 8$ *e* $N\mu\sigma 9$ (*ib.* p. 249-50).

7.4 Exemplos e Aplicações

Exemplo II (conclusão):

Consideremos, de novo, os dados do Exemplo II:

$$2,\ 2.8,\ 3.4.$$

Justifica-se um estudo de outliers?

Se admitirmos um mecanismo normal para a geração destes valores, o máximo torna-se suspeito "apenas" porque "já fica" na unidade seguinte? A "distância" do máximo à mediana é menor do que a correspondente para o mínimo. O 2 é discordante? E para o 2.8 nem "olhamos"?

Estes dados "produzem" uma média

$$\bar{x} \ = \ 2.73$$

e um desvio padrão

$$s \ = 0.7$$

o que nos dá uma situação amostral bem confortável e com uma pequena variância.

Aprofundemos a nossa análise dos dados.

Usemos o critério de Chauvenet, e um teste de discordância pode ser produzido considerando o "desvio"

$$d = \frac{\bar{x} - x_i}{s}.$$

Admitamos que se trata de uma amostra. No mínimo, a estatística d vale 1.04 e para o máximo 0.95. Portanto, o critério de máximo desvio torna suspeito o 2.

Estudemos estes dados estatísticos supondo que são gerados por um modelo de discordância com "outlier" por σ.

O teste de homogeneidade para o modelo geral considerado em 7.2.4.1 leva-nos a suspeitar do mínimo 2.

Por sua vez, um modelo de discordância com "outlier" por σ no modelo geral onde μ, σ e σ' são desconhecidos e que usámos em 7.2.4.2, utiliza a estatística de teste

$$S_9(x_1, ..., x_n) \ = \ min_j \ (\frac{(x_j - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2})^{\frac{1}{2}} \ (1 - \frac{(x_j - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2})^{\frac{n-1}{2}}.$$

Para esta estatística obtemos os valores 0.3358, 0.0668 e 0.369 em 2, 2.8 e 3.4 respectivamente. Assim, em termos de máxima verosimilhança, o valor 2.8 "no meio dos dados" torna-se suspeito.

Exemplo XIV

Na sequência desta abordagem podemos considerar um nova situação com os dados 4.9, 5.9 e 7.

Um estudo análogo ao que foi feito no exemplo anterior vai permitir salientar o valor 5.9 em detrimento dos outros dois.

Estes exemplos, a propósito do número de elementos a analisar, serão objecto de novo estudo no capítulo 9.

Exemplo IX (cont)**: outlier em dados estruturados**

Usemos, de novo, os dados do exemplo IX.

Para além das considerações já formuladas sobre estes dados e implicitamente sobre os resíduos, suponhamos que decidimos registar uma nova observação correspondente a uma carga de 102 Kg e sobre a qual a deformação foi 8.7 cm. Teremos então 10 observações de uma relação bivariada entre a carga e deformação. Em termos de modelação estatística este novo dado tem pouco efeito mas é seguramente bastante influente ao reduzir o desvio-padrão do estimador do declive duma recta de regressão que se ajuste a estes dados. No entanto, esta nova observação tem valores extremos tanto na carga como na extensão.

Numa perspectiva tradicional de pesquisa de outliers, aquele par não pode ser considerado como discordante. "So this observation is not outlying in a 'pattern breaking' respect" - é a conclusão de Barnett e Lewis (Cf. [11], p. 317).

Regressemos à situação inicial com, apenas, as primeiras 9 observações. Ajustando uma recta de regressão obtemos

$$y = 0.075x + 0.677$$

Os resíduos estimados e_i e os resíduos studentizados e_i/s_i estão apresentados na tabela 9. A observação 6, (53.4, 3.1), tem um resíduo e_6 bastante grande e também o correspondente resíduo studentizado é alto.

O critério tradicional, utilizado por Barnett e Lewis[19] permite que aquela observação seja decidida como discordante ao nível de significância de 5% (mas não a 1%).

[19](Cf. [11], p. 323) e a tabela XXXVII de Barnett e Lewis [11].

Tabela 9: Resíduos com n=9.

observação	x_i	e_i	e_i/s_i
1	11.2	0.0762	0.1335
2	21.1	-0.1727	-0.2664
3	29.9	0.4616	0.6698
4	34.1	0.0438	0.0626
5	43.8	0.2101	0.2959
6	53.4	-1.6162	-2.3254
7	59.9	-0.3079	-0.4589
8	61.2	0.8938	1.3450
9	68.9	0.4113	0.6680

Utilizemos, finalmente, o anterior conjunto de dados como exemplo de aplicação do estudo de outliers através do método generativo com alternativa natural em dados estruturados.

Exemplo IX (conclusão)**: outlier em dados estruturados**

Admitamos então os dados iniciais ao qual supomos acrescentada uma nova observação (45.3, 4.1). O conjunto de dados que vamos estudar é portanto o seguinte:

x :	11.2	21.1	29.9	34.1	43.8	53.4	59.9	61.2	68.9	45.3
y :	1.6	2.1	3.4	3.3	4.2	3.1	4.9	6.2	6.3	4.1

Saliente-se que esta observação acrescentada tem um valor y correspondente ao respectivo valor estimado pelo modelo inicial quando se tem x=45.3.

Esta modificação nos dados não tem, portanto, qualquer efeito no modelo estimado que assim se mantém.

Os resíduos estimados e_i e os resíduos studentizados e_i/s_i para este conjunto de 10 observações são apresentados na tabela 10. É muito importante, inclusive para a clarificação da noção de outlier, que se faça uma comparação entre as tabelas 9 e 10. Poderemos verificar o diferente comportamento dos resíduos para cada um dos conjuntos de dados utilizados e compará-los com os correspondentes resíduos studentizados.

Além disso, verificamos que todos os resíduos para as primeiras nove observações se mantêm e que os correspondentes resíduos studentizados são bastante diferentes. Todos os valores absolutos dos resíduos studen-

Tabela 10: Resíduos com n=10.

observação	x_i	e_i	e_i/s_i
1	11.2	0.0764	0.1424
2	21.1	-0.1724	-0.2829
3	29.9	0.4619	0.7127
4	34.1	0.0441	0.0670
5	43.8	0.2104	0.3149
6	53.4	-1.6158	-2.4666
7	59.9	-0.3075	-0.4856
8	61.2	0.8942	1.4252
9	68.9	0.4117	0.7964
10	45.3	-0.0031	-0.0046

tizados foram alterados no exemplo modificado.

A observação 6 que, na análise anterior, foi classificada como discordante perde agora essa condição pois os pontos críticos fornecidos pela referida tabela de Barnett e Lewis[20] *são, para n=10, 2.37 e 2.55 a 5% e a 1% respectivamente.*

Por sua vez, um modelo de discordância com "outlier" por σ no modelo geral onde μ, σ e σ' são desconhecidos e que usámos em 7.2.4.2, através da estatística de teste S_9 elege a observação (45.3, 4.1) como "outlier".

De facto, num modelo de discordância geral, admitindo normalidade e utilizando a correspondente estatística de teste para os resíduos studentizados obtemos o valor 0.001536.

Os respectivos pontos críticos (Cf. tabela 8) são 0.000436 para o nível 1% e 0.00211 para o nível 5% .

Pelo método generativo com alternativa natural, a observação 10, a menos suspeita aos olhos do analista, pode portanto ser declarada um "outlier" a 5%.

Exemplo X (cont.)**: Estatística Forense**

Consideremos, de novo, o caso H-H do exemplo X, já considerado no capítulo 1. Como vimos, em tribunal, tornou-se fundamental a discordância entre o valor 280 - para uma gestação normal - e aquela de 349 dias, verificada no caso em análise.

Como também já referimos no exemplo XI, esta análise em dados estatísticos pode ser "alterada" para uma outra, onde se pode reflectir

[20](Cf. tabela XXXVII de Barnett e Lewis [11].)

sobre a estratégia de apresentação do problema com vista à conclusão a retirar. Se a contagem fosse em semanas - onde, em termos práticos, se defrontavam os valores 50 e 40 - qual a influência na tomada de decisão e qual a repercussão no estudo global deste exemplo? Deixamos a questão para reflexão!

Em termos hipotéticos, vamos admitir que a prova, exigida pelo tribunal, teria de ser feita depois de observada uma gestação com 200 dias.

Estamos aptos a construir uma nova versão deste exemplo para a estatística forense no qual podemos propor que se comparem os dados 200 e 340 com 280. Assim, será 340 "mais outlier" do que 200?

Aos dados envolvidos na evidência estatística deste caso vamos então admitir que juntamos um valor 200. Passamos então a dispor de uma amostra 200, 280 e 340 de uma população normal.

Estudemos estes dados estatísticos supondo que são gerados por um modelo de discordância com "outlier" por σ, com μ conhecido e igual a 280.

O teste de homogeneidade para o modelo considerado em 7.2.4.1 leva-nos a suspeitar do mínimo 200.

A estatística de teste

$$S_8(x_1, ..., x_n) \;=\; min_j \; (\frac{(x_j - \mu)^2}{\Sigma_i (x_i - \mu)^2})^{\frac{1}{2}} \; (1 - \frac{(x_j - \mu)^2}{\Sigma_i (x_i - \mu)^2})^{\frac{n-1}{2}}$$

fornece os valores 0.3358, 0.0668 e 0.369 em 200, 280 e 340, respectivamente. Assim, em termos de máxima verosimilhança, o valor 280 "no meio dos dados" torna-se suspeito.

Quais devem ser os efeitos desta análise estatística para o tribunal?

Capítulo 8

Medidas de Desempenho

8.1 Introdução

É sabido que, o problema da detecção e tratamento das observações aberrantes, surpreendentes ou discordantes surge no momento em que o experimentador inicia qualquer análise de dados estatísticos. Desde logo, se apercebe que as informações contidas nesses dados podem ser distorcidas ou alteradas profundamente devido à existência dos chamados outliers. Estes, por sua vez, podem ser bastante influentes no estudo. E, daqui vem a necessidade de os "analisar" e proceder ao "tratamento". Mas, como resistem os dados a essa presença e quais as principais consequências?

Coloquemo-nos na perspectiva - que apelidámos tradicional - para o estudo de outliers em dados estatísticos. Sabemos que, as dificuldades começam, logo no primeiro momento, com a própria noção de outlier.

Barnett e Lewis [11] - obra de referência fundamental, também para a avaliação do desempenho (performance) - definem outlier num conjunto de dados, como uma observação que parece inconsistente com os restantes elementos da amostra, dado o seu carácter extremo.

Considerando que apenas uma observação de uma amostra não segue a distribuição F, esse dado aberrante pode ter sido gerado por uma diferente lei G , pode tratar-se de um valor "marcadamente extremo" mas que "pertence a F" ou ainda, surgir devido a um erro de medição - caso menos problemático se for descoberta a fonte do erro.

Contaminantes são as observações provenientes de uma outra distribuição que não aquela, geradora dos dados. Estes podem ou não conter outliers - nesta perspectiva tradicional - pois, uma observação

proveniente de uma outra distribuição pode não assumir um valor extremo.

Como é impossível distinguir se um outlier é uma observação de um outra distribuição ou um extremo desta, que se está a admitir geradora "da grande maioria dos dados", em todos os casos o outlier é tratado de igual modo, como contaminante. Quanto aos contaminantes que não sejam outliers - isto é, valores extremos - Barnett e Lewis e muitos outros, reconhecem a incapacidade para os detectar. Assim, nesta abordagem tradicional, valores extremos podem ou não ser outliers mas estes são sempre valores extremos.

Além das dificuldades já apontadas, a problemática de detecção de outliers é ainda agravada no momento da escolha da estatística de teste a adoptar face à grande variedade existente.

Depois da escolha particular do valor suspeito, por exemplo, para populações com distribuição normal, Barnett e Lewis [11] apresentam um conjunto de mais de quarenta testes. A opção por uma ou outra estatística é uma questão fundamental pois uma observação poderá ser considerada outlier por um teste e não o ser por outro. Ao optarmos por um teste devemos ter presente o processo de construção do mesmo. Na maior parte deles a sua construção resultou da aplicação de um princípio óbvio para amostras univariadas, a utilização de estatísticas ordinais. Posteriormente, para muitas desses testes foram descobertas propriedades óptimas, geralmente muito depois de terem sido formulados.

No entanto, na construção de um teste próprio para a detecção de outliers - vulgarmente designado teste de discordância - como em qualquer outra avaliação estatística deste género, devem ser formuladas hipóteses que se defrontam e sobre as quais se tem de decidir.

No caso específico da teoria dos outliers, o cálculo da potência e a construção de testes com certas propriedades desejáveis exige sempre que se especifique o modelo de discordância.

Então, na perspectiva tradicional para o estudo de outliers em dados estatísticos, também a avaliação do desempenho é subjectiva porque não só é condicionada pela observação a testar como também pela escolha do teste de discordância. E como cada teste "é indicado" para uma determinada situação não é possível avaliar as suas diversas capacidades porque, à partida, os pressupostos são distintos. Além disso, como se sabe, a maior parte dos testes de discordância são formulados para aplicação em (ou para) determinada observação de que, subjectivamente, o experimentador "desconfia". Poderíamos assim, eventualmente, comparar testes[1] para máximos ou para mínimos; fazendo portanto apenas

[1]Para melhor clarificar esta realidade podemos, por exemplo, analisar o capítulo 6 de Barnett e Lewis - [11], p. 195 e seguintes - onde são apresentados imensos testes

uma avaliação de desempenho entre grupos de testes e não numa análise global.

Na metodologia GAN em que nos inserimos neste livro, esta problemática não se coloca pois, desde o início, criamos uma definição objectiva de outlier e na sequência do modelo de discordância natural os diversos testes poderão ser avaliados entre si.

8.2 Sobre o Desempenho

Face à multiplicidade de métodos para a construção de testes e modelos de discordância e das estatísticas daí resultantes torna-se necessário avaliar a sua eficácia na detecção de outliers.

Uma medida do desempenho de um teste de discordância é o nível de significância.

No entanto, especificamente para (e em) a teoria dos outliers, a interpretação da significância[2] pode tornar-se problemática. Além disso, a comparação de testes com o mesmo nível de significância depende da hipótese alternativa que é proposta para explicar os outliers.

O cálculo das medidas de desempenho requer o conhecimento da distribuição da estatística de teste na hipótese alternativa que consideramos para explicar os outliers pois exige o conhecimento do comportamento da distribuição da estatística de teste nessa hipótese. Isto coloca problemas de difícil solução computacional especialmente no quadro da distribuição normal e no passado muitos autores ignoraram esta realidade ou limitaram-se a apresentar resultados muito particulares. Adiante, aprofundaremos um pouco mais este estudo para as populações normais.

8.3 Desempenho com um "outlier"

Na avaliação do desempenho vamos, de novo, colocar-nos em modelos onde, como nas situações mais comuns, se confundem outlier e contaminante.

Consideremos uma hipótese H_0 de não existência de outliers, segundo a qual a amostra é extraída de uma determinada população F.

Sabemos que, na formulação de um modelo de discordância, pelo qual se introduz a hipótese de existência de outliers, têm sido consideradas diversas hipóteses alternativas $\bar{H}$.

---

de discordância quer para populações exponenciais quer normais e onde, para cada teste, são indicadas "Properties of test".

[2]Cf. o estudo de Collet e Lewis [28], que já anteriormente referimos e que é muito importante para o esclarecimento da subjectividade no estudo de outliers em dados estatísticos.

A cada modelo alternativo corresponde uma situação bastante particular para os testes formulados.

Os testes tradicionalmente estudados[3] consideram modelos onde são formulados hipóteses alternativas inerentes, por contaminação e por deslizamento.

Os modelos de discordância com alternativas inerentes, contrapõem, genericamente, duas distribuições para a população.

Na hipótese nula:

$$H_0 : x_j \in F \ \ (j = 1, 2, \ldots, n)\,,$$

declara-se que todas as observações provêm de uma distribuição F.

Na hipótese alternativa:

$$\bar{H} : x_j \in G\,(j = 1, 2, \ldots, n)\,,$$

todas as observações provêm de uma distribuição G.

Nos modelos de discordância formulados com hipóteses alternativa por contaminação, a uma distribuição F fixada pela hipótese nula, opõe-se uma distribuição $(1 - \theta)\,F + \theta G$, contaminação de duas funções de distribuição F e G, sendo θ o coeficiente de contaminação que introduz no modelo as possíveis observações contaminantes vindas da população G:

$$H_0 : x_j \in F\,(j = 1, 2, \ldots, n)\,,$$
$$\bar{H} : x_j \in (1 - \theta)\,F + \theta G \ \ (j = 1, 2, \ldots, n)\,.$$

Na alternativa por deslizamento, como vimos, uma das hipóteses mais gerais para a discordância, aceita-se que todas as observações excepto um pequeno número k, provêm, independentemente, de um modelo inicial F indexado por parâmetros - de localização e escala, por exemplo - e que as restantes são de uma versão modificada de F na qual pelo menos um desses parâmetros sofreu alguma alteração.

Numa perspectiva tradicional para o estudo de outliers, como vimos em 2.8.2 e 4.2, os modelos A e B de Ferguson [44] e [45] constituem, provavelmente, a expressão mais geral da alternativa por deslizamento.

Da análise dos modelos referidos facilmente se conclui da extrema dificuldade em estabelecer um estudo comparativo entre eles e daqui resulta, também neste caso, a impossibilidade de uma avaliação do desempenho.

[3]Cf. a citada obra de Barnett e Lewis [11].

Além destas três perspectivas, temos a metodologia com alternativa natural que temos desenvolvido ao longo deste livro. Também neste capítulo - para aprofundar, um pouco as dificuldades e apresentar algumas das possíveis soluções para um estudo de performance numa teoria dos outliers - abordaremos, principalmente, a alternativa natural.

A abordagem através do método GAN é inovadora e, como sabemos, com generalidade, porque se pode aplicar em todas as amostras - obviamente com diversos graus de dificuldade.

8.4 Medidas de Desempenho

Ao executar uma avaliação do desempenho[4] de testes de discordância (também) verificamos a especificidade da teoria dos outliers quer pela diversidade de "hipóteses" em estudo quer pela necessidade de introduzir várias medidas.

Na presença de um único outlier na amostra e assumindo a alternativa de deslizamento ou de alternativa natural, uma das observações da amostra sob $\bar{H}$ será o contaminante. Suponha-se que é a observação x_n. Se W é uma estatística de teste, existe a correspondente medida W_n, verificada, para o contaminante.

David [33], para as estatísticas de teste genericamente

$$W = \min_{1\leq j\leq n} W_j \left(\text{ou } W = \max_{1\leq j\leq n} W_j \right),$$

define regiões de rejeição $\Re$ que, respectivamente, podem ser da forma $W < c$ (ou $W > c$), onde c é o valor crítico.

No caso de se concluir pela contaminação - existência de outlier - a observação responsável corresponde ao W_j que minimiza (ou maximiza).

São então sugeridas as seguintes cinco probabilidades como medidas da performance de W_n:

$$P_1 = P\left(W \in \Re | \bar{H}\right),$$

é a probabilidade de se concluir pela existência de contaminação quando se admite que ela existe. É função potência.

$$P_2 = P\left(W_n \in \Re | \bar{H}\right),$$

[4]Para uma perspectiva histórica, cada um à sua maneira sobre este assunto, são importantes: o capítulo 4 de Barnett e Lewis [11], o capítulo 2 de Hawkins [66] e David [33].

é a probabilidade de x_n ser "suficientemente" discordante, admitindo que há contaminação.

$$P_3 = P\left(W \in \Re \text{ e } W = W_n | \bar{H}\right),$$

é a probabilidade de o contaminante ser o outlier e identificado como tal, admitindo que à contaminação.

$$P_4 = P\left(W \in \Re \text{ e } W = W_n \text{ e } W_1\ , \ldots, W_{n-1}\ \notin \Re | \bar{H}\right),$$

é a probabilidade de decidir existência de contaminação e a única observação que satisfaz o critério de discordância ser a contaminante x_n , admitindo que há contaminação.

$$P_5 = P\left(W_n \in \Re\ | W = W_n, \bar{H}\right),$$

é a probabilidade de a observação contaminante satisfazer o critério de discordância caso tenha sido considerada responsável e haja contaminação. Esta medida é equivalente[5] a P_3/P_6, onde

$$P_6 = P\left(\ W = W_n | \bar{H}\right),$$

é a probabilidade de considerar responsável a observação contaminante, caso haja contaminação.

$$P_7 = P\left(V \in \Re, V \neq V_n | \bar{H}_n\right) = P_1 - P_3.$$

Assim, para além dos erros clássicos de primeira e segunda espécie, podemos ainda, admitindo que um dos elementos da amostra era o contaminante, cometer o erro de responsabilizar como contaminante outro elemento da amostra.

Na hipótese de existência de um outlier, P_1 é a potência do teste, P_2 é a probabilidade de rejeitar H_0 e o contaminante satisfazer o critério de rejeição, P_3 é a probabilidade de rejeitar H_0 e identificar correctamente o outlier, P_4 exige além de P_3 que o contaminante seja o único elemento da amostra a satisfazer o critério da rejeição e P_5 é a probabilidade de rejeitar H_0 condicionada à identificação correcta do suspeito de contaminação. Além destas medidas, temos duas outras que se obtêm a partir das anteriores; P_6, probabilidade de identificação correcta do suspeito de contaminação (independentemente do teste concluir ou não pela existência de contaminação) e P_7 que, ao contrário das restantes, é uma medida de falta de desempenho, ou seja a probabilidade de rejeitar H_0 mas falhando a identificação do contaminante.

[5]Cf. Dixon [35].

Para um bom teste exige-se que P_1, P_2, P_3, P_4, P_5 e P_6 sejam elevados e P_7 o mais reduzido possível.

Pretende-se ainda que $P_1 - P_3$ tenha um valor reduzido; ou seja que a probabilidade que o teste erradamente identifique uma boa observação como discordante. P_3/P_5 traduz a probabilidade de o contaminante aparecer como outlier. Deste modo exige-se que esta razão seja a maior possível.

Pretende-se ainda que P_7 seja reduzido enquanto que P_1, P_3, P_5 e P_6 tão grandes quanto possível, embora estas duas últimas medidas possam estar em conflito.

8.4.1 Medidas de Desempenho em Exponenciais

Para as populações exponenciais, Braumann[6] estudou medidas de desempenho para diversos testes de discordância.

Para ilustrar o processo de cálculo dessas medidas, correspondendo a um estudo pelo método GAN, consideremos as hipóteses seguintes:

H_0: $x_1, x_2, \ldots, x_n$ é uma amostra de uma distribuição $Exp\,(\lambda, \delta)$.

$\bar{H}_j$: a observação x_j segue uma distribuição $Exp\,(\lambda, \delta')$ e as restantes uma distribuição $Exp\,(\lambda, \delta)$.

Uma estatística de teste apropriada no caso em que λ, δ e δ'são conhecidos e $\delta > \delta'$ é

$$A = \frac{x_{(1)} - \lambda}{\delta}.$$

A região de rejeição é da forma $A < c$, e o candidato a outlier é a observação $x_{(1)}$.

Para as diferentes medidas, consideremos

$$Y_i = \frac{x_i - \lambda}{\delta}$$

e então vem:

H_0: todos os Y_i seguem uma distribuição $Exp\,(0, 1)$.

$\bar{H}_j$: Y_j tem distribuição $Exp\,(0, \sigma)$ e os restantes seguem uma distribuição $Exp\,(0, 1)$, com $\gamma = \delta'/\delta$.

Supondo

$$L = \min_{1 \le j \le n-1} Y_j,$$

resulta

$$Y_{(1)} = \min_{1 \le j \le n-1} (L, Y_n)$$

[6]Cf. [21], [22] e [23].

e podemos então obter as diferentes distribuições necessárias para o cálculo das medidas de performance:

$$\begin{aligned} F_{Y_{(1)}} &= P(Y(1) \leq y) \\ &= 1 - (1 - F_L(y))\ e^{-\frac{y}{\gamma}} \\ &= 1 - e^{-(n-1)y}\ e^{-\frac{y}{\gamma}}. \end{aligned}$$

Das distribuições das estatísticas de teste, nas hipóteses nula e alternativa, tem-se respectivamente:

$$F_A^0(a) = 1 - e^{-\frac{a}{\gamma}}, \text{ com } a > 0,$$

$$F_A^1(a) = 1 - e^{-(n-1)a}\ e^{-\frac{a}{\gamma}}, \quad \text{com } a > 0.$$

Para as medidas de performance, considerando que c é o ponto crítico, temos, por exemplo:

$$\begin{aligned} P_1 &= P(A \leq c/\bar{H}) \\ &= F_A^1(c) \\ &= 1 - e^{-(n-1)c}\ e^{-\frac{c}{\gamma}} \end{aligned}$$

$$\begin{aligned} P_3 &= P(A \leq c, A = Y_n/\bar{H}) \\ &= P(Y_n \leq c, L \geq Y_n/\bar{H}) \\ &= \frac{1}{1 + (n-1)\gamma}(1 - e^{-nc}\ e^{c(1-\frac{1}{\gamma})}). \end{aligned}$$

$$\begin{aligned} P_6 &= P\left(A = Y_n/\bar{H}\right) \\ &= P\left(L \geq Y_n/\bar{H}\right) \\ &= \frac{1}{1 + (n-1)\,\gamma}. \end{aligned}$$

As medidas P_5 e P_7 podem-se obter a partir das anteriores.

Até agora considerámos medidas de performance que se adequam às hipóteses alternativas de deslizamento ou natural, no entanto para outro tipo de modelo de discordância teremos que determinar outro tipo de medidas. Considere-se uma alternativa de mistura e um máximo da amostra $x_{(n)}$ correspondente a um outlier. Sob $\bar{H}$ o número de contaminantes na amostra não é conhecido ao contrário do que se passava na hipótese de deslizamento ou de alternativa natural. Segundo Barnett e Lewis [11] os seguintes acontecimentos são relevantes em termos de

performance:

D: O teste identifica $x_{(n)}$ como discordante.

E: Verifica-se $\bar{H}$ e a amostra contém um ou mais contaminantes.

F: $x_{(n)}$ é o contaminante.

Neste caso a potência P_1, tem uma analogia directa com $P\left(D|\bar{H}\right)$. Contudo $P\left(D|E\right)$, que é uma função potência condicional à presença de contaminação, é também uma medida útil. É representada por P_6. Medidas análogas para P_3 e P_5 são respectivamente $P\left(D \cap F|E\right)$ e $P\left(D|F\right)$.

Um bom teste deve ter um valor elevado para $P\left(D|F\right)$ e $P\left(D|E\right)$ e reduzido para $P\left(D|\bar{H}\right) - P\left(D \cap F|E\right)$, ou seja valores altos para P_5, P_6 e baixos para $P_1 - P_3$.

No caso de um teste de discordância numa alternativa inerente a situação simplifica-se, pois não existe uma observação contaminante específica. Como tal, as probabilidades P_3 e P_5 são indefinidas (e, assim também, os acontecimentos E e F). A medida de performance apropriada para o teste será a potência $P\left(D|\bar{H}\right)$.

8.4.2 Medidas de Desempenho em Normais

Considere-se o modelo generativo de alternativa natural introduzido em 4.4.

Neste estudo, na hipótese H_0 de homogeneidade, admitiremos uma amostra $x_1, x_2, \ldots, x_n$ onde as observações seguem uma distribuição normal com densidade de probabilidade

$$f\left(x; \mu, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Vamos usar no nosso estudo um modelo de discordância com outlier por σ. Pela hipótese alternativa natural iremos admitir a presença de um só outlier na amostra. Seja $\bar{H}_j$ a hipótese que supõe x_j uma observação com densidade de probabilidade $f\left(x, \mu, \sigma\prime\right)$ enquanto que, todas as restantes observações x_i $(i \neq j)$ seguem uma distribuição normal com densidade $f\left(x, \mu, \sigma\right)$. Considera-se μ conhecido.

Como vimos no capítulo 7, no caso em presença pode usar-se a estatística

$$S_1 = \max_j \left(\frac{|x_j - \mu|}{\sigma}\right)^{-1} \exp\left[\frac{1}{2}\left(\frac{x_j - \mu}{\sigma}\right)^2\right].$$

para decidir sobre a homogeneidade da amostra e, se existir, seleccionar o "outlier".

Consideremos $Z_j = |x_j - \mu| / \sigma$ e a função $\varphi(z) = \frac{1}{z}\exp\left(\frac{1}{2}z^2\right)$. O máximo da estatística S_1 é atingido quando Z_j é máximo ou mínimo, pelo que as observações candidatas a "outlier" são aquelas que minimizam Z_j, ou seja, $x_{[\mu]}$ *ou que* maximizam Z_j , isto é, $x_{(1)}$ e $x_{(n)}$. A região de rejeição da hipótese H_0 é então da forma $S_1 > c$, isto é $Z_{(1)} < c_1$ ou $Z_{(n)} > c_2$.

No caso em que σ é conhecido e $\sigma\prime$ desconhecido mas temos informação suplementar que nos indica que $\sigma\prime < \sigma$, devemos usar como estatística de teste $S_2 = \min_j \left|\frac{x_j - \mu}{\sigma}\right|$. O candidato a outlier é a observação $x_{[\mu]}$.

Tendo agora a informação suplementar que $\sigma < \sigma\prime$, obtemos como estatística de teste $S_3 = \max_j \left|\frac{x_j - \mu}{\sigma}\right|$. Os candidatos a outlier são $x_{(1)}$ e $x_{(n)}$.

Suponham-se $Y_1, Y_2, \ldots, Y_n$, onde $Y_i = (x_i - \mu)/\sigma$ e sejam ainda $Z_1, Z_2, \ldots, Z_n$ tal que $Z_i = |Y_i|$. Na hipótese nula, H_0, de homogeneidade das observações e ausência de outliers, as observações $x_1, x_2, \ldots, x_n$ seguem uma distribuição normal com parâmetros μ, σ. Então cada um dos Y_j $(j = 1, \ldots, n)$ segue uma distribuição normal com valor médio nulo e variância unitária.

Na hipótese alternativa $\bar{H}_n$, vamos supor que a observação x_n corresponde à observação outlier, e que segue uma distribuição normal com valor médio μ e desvio padrão σ', enquanto as restantes observações têm o mesmo valor médio mas desvio padrão σ. Neste caso $E[Y_n] = 0$ e $V[Y_n] = \delta^2$ com $\delta = \sigma'/\sigma$, pelo que na hipótese alternativa

$$Y_1, Y_2, \ldots, \frac{Y_n}{\delta} \sim N(0,1).$$

A correspondente função de distribuição de Z_n na hipótese $\bar{H}_n$, que nos será útil posteriormente, é dada por

$$F^1_{Z_n}(z) \quad = \quad \phi\left(\frac{z}{\delta}\right) - \phi\left(-\frac{z}{\delta}\right) = 2\phi\left(\frac{z}{\delta}\right) - 1, \; z \geq 0.$$

Consideremos agora $M = \max_{1 \leq i \leq n-1} Z_i$ e $L = \min_{1 \leq i \leq n-1} Z_i$ e vejamos qual a distribuição da estatística de teste S_1 na hipótese nula e alternativa. Supondo $0 \leq s_1 \leq s_2$

$$F^0_{S_1}(s) \quad = \quad P(s_1 \leq Z_1, Z_2, \ldots, Z_n \leq s_2) = [2(\phi(s_2) - \phi(s_1))]^n,$$

$$\begin{aligned} F^1_{S_1}(s) &= P\left(S_1 \leq s | \bar{H}_n\right) = P^1_{Z_{(1)},Z_{(n)}}(s_1, s_2) = \\ &= \left[2\left(\phi(s_2) - \phi(s_1)\right)\right]^{n-1}\left[2\left(\phi\left(\frac{s_2}{\delta}\right) - \phi\left(\frac{s_1}{\delta}\right)\right)\right], \\ & \qquad com \quad 0 \leq s_1 \leq s_2. \end{aligned}$$

Para determinar o ponto crítico, veja-se que

$$\alpha = P(S_1 > c | H_0) = 1 - F^0_{S_1}(c) \Leftrightarrow \phi(c_2) - \phi(c_1) = \frac{1}{2}(1-\alpha)^{\frac{1}{n}},$$

tal que

$$\varphi(c_1) = \varphi(c_2) = c.$$

Na determinação de c, fixado o valor de α (significância do teste) e n (dimensão da amostra) temos que encontrar os valores c_1 e c_2 que satisfaçam ambas as condições. Sob esses pressupostos o c encontrado, será o valor crítico para testar a presença de um outlier.

Seja agora a estatística S_2. As distribuições, respectivamente na hipótese nula e alternativa, são dadas por

$$F^0_{S_2}(s) = P(S_2 \leq s | H_0) = 1 - \left[1 - (\phi(s) - \phi(-s))\right]^n \quad s \geq 0,$$

$$\begin{aligned} F^1_{S_2}(s) &= P\left(S_2 \leq s | \bar{H}_n\right) = P\left(Z_{(1)} \leq s\right) = 1 - P\left(Z_{(1)} > s\right) = \\ &= 1 - \left[1 - (\phi(s) - \phi(-s))\right]^{n-1}\left[1 - \left(\phi\left(\frac{s}{\delta}\right) - \phi\left(-\frac{s}{\delta}\right)\right)\right] \quad s \geq 0. \end{aligned}$$

Valores críticos para S_2, podem-se obter a partir de

$$c = \phi^{-1}\left[\frac{2 - (1-\alpha)^{\frac{1}{n}}}{2}\right].$$

No caso de S_3, as distribuições correspondem a

$$F^0_{S_3}(s) = \left[\phi(s) - \phi(-s)\right]^n \quad s \geq 0,$$

$$F^1_{S_3}(s) = \left[\phi(s) - \phi(-s)\right]^{n-1}\left[\phi\left(\frac{s}{\delta}\right) - \phi\left(-\frac{s}{\delta}\right)\right] \quad s \geq 0.$$

Valores críticos para S_3, obtêm-se a partir de

$$c = \phi^{-1}\left[\frac{1 + (1-\alpha)^{\frac{1}{n}}}{2}\right].$$

Consideremos a estatística S_1, vamos obter as respectivas medidas de desempenho.

$$\begin{aligned} P_1 &= P\left(S_1 > c \left| \bar{H}_n \right.\right) = 1 - F_{S_1}^1(c) = \\ &= 1 - \left[2\left(\phi(c_2) - \phi(c_1)\right)\right]^{n-1} \left[2\left(\phi\left(\frac{c_2}{\delta}\right) - \phi\left(\frac{c_1}{\delta}\right)\right)\right], \end{aligned}$$

$$P_2 = P\left(\varphi(Z_n) > c \left| \bar{H}_n \right.\right) = 1 - \left[2\left(\phi\left(\frac{c_2}{\delta}\right) - \phi\left(\frac{c_1}{\delta}\right)\right)\right].$$

Para o cálculo de P_3 note-se que $S_1 = \max\left(\varphi(L), \varphi(M), \varphi(Z_n)\right)$. Atendendo a que $L \leq M$, tendo também em atenção a sua posição relativamente ao máximo de φ, considerando $\theta = \varphi(Z_n)$, vem

$$\begin{aligned} P_3 &= P\left(S_1 > c, S_1 = \varphi(Z_n) \left| \bar{H}_n \right.\right) = \\ &= P\left(\theta_1 \leq L \leq \theta_2, \theta_1 \leq M \leq \theta_2, \theta > c \left| \bar{H}_n \right.\right) = \\ &= \left(\int_0^{c_1} + \int_{c_2}^{+\infty}\right) \left[2\left(\phi(\theta_2) - \phi(\theta_1)\right)\right]^{n-1} \frac{2}{\delta\sqrt{2\pi}} \exp\left(-\left(\frac{z^2}{2\delta^2}\right)\right) dz, \end{aligned}$$

com θ_1 e θ_2 determinados de modo que $0 < \theta_1 < 1 < \theta_2$ e $\varphi(\theta_1) = \varphi(\theta_2) = \theta$, ou seja, $\frac{1}{\theta_1}\exp\left(\frac{1}{2}\theta_1^2\right) = \frac{1}{\theta_2}\exp\left(\frac{1}{2}\theta_2^2\right) = \theta$. Decompondo os integrais, no primeiro vem $\theta_1 = z$ e no segundo $\theta_2 = z$ e fazendo no segundo integral a mudança de variável de integração de θ_2 para θ_1 vem

$$\begin{aligned} P_3 &= \int_0^{c_1} \left[2\left(\phi(\theta_2) - \phi(\theta_1)\right)\right]^{n-1} \frac{2}{\delta\sqrt{2\pi}} \times \\ &\quad \times \left[\exp\left(-\left(\frac{\theta_1^2}{2\delta^2}\right)\right) - \exp\left(-\left(\frac{\theta_2^2}{2\delta^2}\right)\right) \frac{d\theta_2}{d\theta_1}\right] d\theta_1. \end{aligned}$$

Para o cálculo numérico decompõe-se $[0, c_1]$ em pontos $\theta_{1,k} = (kc_1)/l$ com $k = 0, 1, \ldots, l$ e considera-se $0 < \theta_{1,k} < 1$ e $\frac{1}{\theta_{1,k}}\exp\left(\frac{1}{2}\theta_{1,k}^2\right) = \frac{1}{\theta_{2,k}}\exp\left(\frac{1}{2}\theta_{2,k}^2\right)$. Então

$$\begin{cases} P_3 \leq \sum_{k=1}^{l} g\left(\theta_{1,k-1}\right)\left(\theta_{1,k} - \theta_{1,k-1}\right) \\ P_3 \geq \sum_{k=1}^{l} g\left(\theta_{1,k}\right)\left(\theta_{1,k} - \theta_{1,k-1}\right), \end{cases}$$

em que $g(\theta)$ corresponde à função a integrar em P_3.

Para P_4 teremos

$$P_4 = P\left(S_1 > c, S_1 = \varphi(Z_n); \varphi(Z_1), \ldots, \varphi(Z_{n-1}) \leq c \left| \bar{H}_n \right.\right),$$

note-se que $S_1 > c$ e $\varphi(Z_1), \ldots, \varphi(Z_{n-1}) \leq c$ implica que $S_1 = \varphi(Z_n)$. Assim

$$\begin{aligned} P_4 &= P\left(\varphi(Z_n) > c; \varphi(L) \leq c, \varphi(M) \leq c \left| \bar{H}_n \right.\right) = \\ &= \left[1 - 2\left(\phi\left(\frac{c_2}{\delta}\right) - \phi\left(\frac{c_2}{\delta}\right)\right)\right] \left[2\left(\phi(c_2) - \phi(c_1)\right)\right]^{n-1} \end{aligned}$$

$$P_6 = P\left(S_1 = \varphi(Z_n) \left| \bar{H}_n \right.\right) = P\left(\theta_1 \leq L \leq \theta_2, \theta_1 \leq M \leq \theta_2 \left| \bar{H}_n \right.\right).$$

Num raciocínio idêntico ao feito em P_3, substituindo $\theta > c$ por $\theta \geq e^{1/2}$, a que corresponde $0 \leq \theta_1 \leq 1$, vem $P_6 = \int_0^1 g(\theta_1)\, d\theta_1$.

As restantes medidas de desempenho podem obter-se a partir das anteriores com $P_5 = P_3 / P_6$ e $P_7 = P_1 - P_3$.

Calculemos agora as medidas de desempenho para a estatística S_2

$$\begin{aligned} P_1 &= P\left(S_2 < c \left| \bar{H}_n \right.\right) = F_{S_2}^1(c) = \\ &= 1 - \left[1 - \left(\phi(c) - \phi(-c)\right)\right]^{n-1} \left[1 - \left(\phi\left(\frac{c}{\delta}\right) - \phi\left(-\frac{c}{\delta}\right)\right)\right]; \end{aligned}$$

$$P_2 = P\left(Z_n < c \left| \bar{H}_n \right.\right) = \phi\left(\frac{c}{\delta}\right) - \phi\left(-\frac{c}{\delta}\right);$$

$$\begin{aligned} P_3 &= P\left(S_2 < c, S_2 = Z_n \left| \bar{H}_n \right.\right) = P\left(Z_n < c, L \geq Z_n \left| \bar{H}_n \right.\right) \\ &= \int_0^c \left[1 - \left(\phi(z) - \phi(-z)\right)\right]^{n-1} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{z}{\delta}\right)^2}{2}\right) dz, \end{aligned}$$

para o cálculo numérico deste integral utilizou-se a quadratura adaptativa de Lobatto (Veja-se, por exemplo, em Gander e Gautschi [50]);

$$\begin{aligned} P_4 &= P(S_2 < c, S_2 = Z_n; Z_1, \ldots, Z_{n-1} > c | \bar{H}_n,) = \\ &= \left[\phi\left(\frac{c}{\delta}\right) - \phi\left(-\frac{c}{\delta}\right)\right] \left[1 - \left(\phi(c) - \phi(-c)\right)\right]^{n-1}; \end{aligned}$$

$$\begin{aligned} P_6 &= P\left(S_2 = Z_n \left| \bar{H}_n \right.\right) = P\left(L \geq Z_n \left| \bar{H}_n \right.\right) \\ &= \int_0^{+\infty} \left[1 - \left(\phi\left(z\right) - \phi\left(-z\right)\right)\right]^{n-1} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\delta^2}\right) dz. \end{aligned}$$

Por último, vejamos as medidas de desempenho para S_3

$$P_1 = \left[\phi\left(c\right) - \phi\left(-c\right)\right]^{n-1} \left[\phi\left(\tfrac{c}{\delta}\right) - \phi\left(-\tfrac{c}{\delta}\right)\right];$$

$$P_2 = 1 - \left(\phi\left(\tfrac{c}{\delta}\right) - \phi\left(-\tfrac{c}{\delta}\right)\right);$$

$$P_3 = \int_c^{+\infty} \left(\phi\left(z\right) - \phi\left(-z\right)\right)^{n-1} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\delta^2}\right) dz;$$

$$P_4 = \left[\left(\phi\left(c\right) - \phi\left(-c\right)\right)\right]^{n-1} \left[1 - \left(\phi\left(\frac{c}{\delta}\right) - \phi\left(-\frac{c}{\delta}\right)\right)\right];$$

$$P_6 = \int_0^{+\infty} \left(\phi\left(z\right) - \phi\left(-z\right)\right)^{n-1} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\delta^2}\right) dz.$$

8.5 Desempenho com múltiplos "outliers"

Em muitas situações o número de observações discordantes numa amostra poderá ser maior do que um.

Podemos ter, por exemplo, uma amostra normal (ou exponencial) de dimensão n, com dois outliers superiores $x_{(n-1)}$ e $x_{(n)}$, ambos invulgarmente distantes para a direita da observação seguinte $x_{(n-2)}$, ou dois outliers inferiores $x_{(1)}$ e $x_{(2)}$ afastadas - para a esquerda - mais do que seria de esperar, ou ainda um par ($x_{(1)}$,$x_{(n)}$) de outliers (superior e inferior) .

Tal como já vimos em alguns exemplos, uma amostra pode conter três observações que, indistintamente, surgem como discordantes em relação

às restantes $n-3$. Esta é uma dificuldade acrescida para o estudo[7] de outliers em dados estatísticos.

Em todas estas situações, e em termos gerais, existem k (>1) observações discordantes na amostra de dimensão n e o analista pretende estudar a possibilidade de existirem k contaminantes. Deste modo são necessários testes de discordância apropriados a uma situação de múltiplos outliers.

Nestes modelos, podemos optar entre dois tipos de procedimentos de detecção: em bloco ou sequencial.

Num teste sequencial as observações $x_1, x_2, \ldots, x_n$ são estudadas numa determinada sequência. Cada observação, *de per si*, é testada em termos de discordância em relação às restantes já analisadas utilizando-se um teste para cada outlier e eliminando-o se tal for decidido. Em bloco, todos são decididos num só teste.

Consideremos um modelo com populações normais.

Suponha-se que existem dois outliers numa amostra, podendo ser qualquer par dos n elementos. Sejam x_j e x_k, com $j \neq k$ e considere-se o seguinte modelo de discordância com "outlier" por μ:

H_0 : Todas as observações têm densidade

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$\bar{H}_{j,k}$: Na hipótese alternativa as observações x_j e x_k têm respectivamente as seguintes densidades

$$f(x, \mu', \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu')^2}{2\sigma^2}\right),$$

$$f(x, \mu'', \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu'')^2}{2\sigma^2}\right).$$

Aplicando o método GAN obtemos

$$\hat{L}_0 = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right),$$

[7] Alguma análise e reflexão sobre esta problemática foi desenvolvida nos estudos elaborados por Braumann [20] e Rosado e Alpiarça [121].

$$\hat{L}_{j,k} = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i\neq j,k}^{n}(x_i-\mu)^2 + (x_j-\mu')^2 + (x_k-\mu'')^2\right)\right).$$

A estatística de teste neste caso será

$$S = \max_{j\neq k} \frac{x_j(\mu'-\mu) + x_k(\mu''-\mu)}{\sigma^2}.$$

Como candidatos a outliers temos:

$x_{(n)}$ e $x_{(n-1)}$ se $\mu' > \mu$ e $\mu'' > \mu$;
$x_{(1)}$ e $x_{(2)}$ se $\mu' < \mu$ e $\mu'' < \mu$;
$x_j = x_{(n)}$ e $x_k = x_{(1)}$ se $\mu' > \mu$ e $\mu'' < \mu$;
$x_k = x_{(n)}$ e $x_j = x_{(1)}$ se $\mu' < \mu$ e $\mu'' > \mu$.

Considerando agora uma situação de um modelo de discordância com "outlier" por σ temos

H_0 : Todas as observações têm densidade de probabilidade

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$\bar{H}_{j,k}$: Na hipótese alternativa as observações x_j e x_k têm as seguintes densidades:

$$f(x,\mu,\sigma') = \frac{1}{\sigma'\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma'^2}\right),$$

$$f(x,\mu,\sigma'') = \frac{1}{\sigma''\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma''^2}\right).$$

Neste quadro admitindo que ambos, σ' e σ'' são desconhecidos obtemos a seguinte estatística de teste:

$$S = \max_{j\neq k} \frac{\sigma^2}{|x_j-\mu|\,|x_k-\mu|} \exp\left\{\frac{1}{2\sigma''^2}\left[(x_j-\mu)^2 + (x_k-\mu)^2\right] - 1\right\}.$$

Os candidatos a "outlier" são então:

Se $\sigma > \sigma'$ e $\sigma < \sigma''$ x_j é o elemento mais próximo de μ e x_k o mais afastado;

Se $\sigma < \sigma'$ e $\sigma > \sigma''$ x_k é o elemento mais próximo de μ e x_j o mais afastado;

Se $\sigma < \sigma'$ e $\sigma < \sigma''$ x_j e x_k são os valores mais afastados de μ,

$$x_{(n)} \text{ e } x_{(n-1)}, \text{ ou } x_{(1)} \text{ e } x_{(2)}, \text{ ou } x_{(1)} \text{ e } x_{(n)};$$

Se $\sigma > \sigma'$ e $\sigma > \sigma''$ x_j e x_k são os valores mais próximos de μ.

Com este exemplo ilustrámos também a possibilidade de no caso de um teste de bloco os "outliers" poderem ser outros para além das observações extremos.

Vamos agora ilustrar o procedimento sequencial.

Suponha-se, para mais fácil exposição, $k = 2$, e que pretendemos testar a discordância de dois outliers superiores $x_{(n-1)}$ e $x_{(n)}$ numa amostra exponencial de dimensão n.

Considere-se que $x_{(1)}, \ldots, x_{(n-2)}$ pertencem a uma distribuição F com densidade $\theta e^{-\theta x}$ $(x > 0)$, e $x_{(n-1)}$, e $x_{(n)}$ pertencem a distribuições exponenciais G_1, G_2 com densidades respectivamente $\lambda\theta e^{-\lambda\theta x}$ e $\mu\theta e^{-\mu\theta x}$ $(x > 0)$.

A hipótese nula é:

$$H_0 : \lambda = \mu = 1.$$

Podemos então considerar um par de alternativas consecutivas a H_0; sejam:

$$\bar{H}' : \lambda = 1, \mu < 1,$$
$$\bar{H}'' : \lambda < 1.$$

O procedimento habitual consiste em testar primeiro H_0, contra $\bar{H}'$ utilizando um teste para um outlier superior.

Se H_0 for aceite, ambos os outliers são declarados consistentes com o resto da amostra e o teste de discordância termina.

Se H_0 for rejeitada, então $\bar{H}''$ é testada contra uma hipótese de trabalho revista confinada a $x_{(1)}, \ldots, x_{(n-1)}$

$$H'' : \lambda < 1.$$

Novamente vamos utilizar o teste para um único outlier superior. Temos então um procedimento sequencial, com três possíveis caminhos:

Aceita-se H_0 e portanto nem $x_{(n)}$ nem $x_{(n-1)}$ são considerados discordantes.

Rejeita-se H_0, então aceita-se H'' e portanto $x_{(n)}$ é considerado discordante mas não $x_{(n-1)}$.

Rejeita-se H_0, rejeita-se H'' e, neste caso, $x_{(n)}$ e $x_{(n-1)}$ são ambos considerados discordantes.

Neste procedimento, o "outlier extremo" foi testado primeiro, depois o segundo mais extremo e assim sucessivamente.

No entanto podemos utilizar outro método.

Suponha-se que, no exemplo anterior, primeiro testamos H'' contra $\bar{H}''$, utilizando um teste para $x_{(n-1)}$ como o único outlier na amostra $x_{(1)}, \ldots, x_{(n-1)}$ omitindo $x_{(n)}$. Se H'' é rejeitado, o outlier $x_{(n-1)}$ é considerado discordante em conjunto com $x_{(n)}$. Se H'' é aceite, testamos então H_0 contra $\bar{H}'$.

Os três possíveis caminhos neste procedimento sequencial são:

Rejeita-se H'', *isto*, nem $x_{(n-1)}$ ou $x_{(n)}$ são ambos considerados discordantes.

Aceita-se H'', rejeita-se H_0, isto é, $x_{(n)}$ é considerado discordante mas não $x_{(n-1)}$.

Aceita-se H'', aceita-se H_0, isto é, nem $x_{(n)}$ nem $x_{(n-1)}$ são considerados discordantes

Barnett e Lewis [11], designam estes dois tipos de procedimentos sequenciais respectivamente interno (inward) e externo (outward), sendo a utilização deste último preferível ao primeiro.

Em princípio, o procedimento sequencial não apresenta qualquer problema na sua construção, já que meramente envolve a repetição de um teste construído para a situação de um único outlier do vasto conjunto de testes disponíveis. Contudo, há que fazer importantes opções. Qual o teste para um único outlier que deve ser utilizado numa dada situação? Qual o nível de significância que deve ser utilizada em cada etapa?

A escolha entre um procedimento em bloco ou um procedimento sequencial numa situação de múltiplos outliers depende da performance relativa dos testes em relação à hipótese alternativa $\bar{H}$.

Em geral testamos k outliers - um número fixo. Se o verdadeiro número de outliers excede k então a homogeneidade não é rejeitada. Este fenómeno é geralmente designado por "masking". Por outro lado, se o verdadeiro número de outliers é inferior a k mas não zero, existem boas hipóteses que o modelo nulo seja rejeitado a favor de um modelo de k outliers. Este fenómeno é habitualmente designado por "swamping". Então, a menos que k seja escolhido correctamente, podemos chegar a resultados falsos.

O procedimento sequencial apesar de muito utilizado tem uma limitação importante, nomeadamente no caso do procedimento externo, o possível efeito de mascaramento ("masking"). Por outro lado o procedimento interno é largamente imune ao "masking", desde que o número

de contaminantes na amostra não exceda o número de outliers k que é assumido no teste.

Isto conduz-nos à questão de decidir qual o valor de k. Para um procedimento de bloco, e também para um procedimento sequencial externo, k necessita de ser especificado. No procedimento sequencial interno, k resulta naturalmente do procedimento pois uma sequência de testes é efectuado, observação a observação, até ao primeiro resultado não-discordante ser atingido, digamos o $(m+1)$-ésimo teste, e neste caso k é determinado sendo igual a m.

O valor de k escolhido no caso de um procedimento de bloco ou de testes sequenciais é de facto o número máximo de contaminantes que se assume que a amostra contém. Pode resultar de uma inspecção aos dados e da constatação do número de observações que de uma forma evidente se afastam do resto da amostra. Contudo, é preferível decidir o valor de k através de um qualquer processo de cálculo a partir dos dados, do que de inspecção visual. Este problema de decisão qual o número de outliers na amostra foi estudado por diversos autores.

Tietjen e Moore [133] propuseram um método para determinar k. Supondo que o número de outliers superiores tem que ser estimado; os autores propõem localizar a diferença mais elevada entre observações adjacentes na amostra colocada em ordem ascendente para a direita da média da amostra, e k será determinado como o número de observações situadas à direita da diferença mais elevada.

Analisemos o desempenho para o procedimento em bloco.

Consideremos uma amostra univariada $x_1, x_2, \ldots, x_n$, com a hipótese nula

$$H_0 : x_j \in F (j = 1, \ldots, n).$$

Em vez de um único outlier, digamos x_n, temos um grupo de k outliers $(k > 1)$ digamos

$$x_{(n-k+1)}, \ldots, x_{(n-1)}, x_{(n)}.$$

Utilizando uma estatística de teste de discordância de bloco, digamos Z, temos um teste que considera $x_{(n-k+1)}, \ldots, x_{(n-1)}, x_{(n)}$ discordante se $Z > z_\alpha$, onde z_α é o valor crítico para um nível de significância α definido por

$$P\,(\,Z > z_\alpha |\, H_0) = \alpha.$$

Considere-se primeiro uma hipótese alternativa de deslizamento $\bar{H}$ sob a qual, $n-k$ daquelas observações pertencem a F e as restantes k , os contaminantes, pertencem a uma distribuição diferente G.

Analogamente ao caso de um "outlier" temos três medidas de per-

formance úteis P_1, P_3 e P_5, sendo P_1 a função potência

$$P_1 = P\,(\,Z > z_\alpha |\, H_0)\,;$$

P_3 é a probabilidade dos k contaminantes serem os "outliers" e serem identificados como discordantes; e P_5 é a probabilidade de quando os contaminantes são os "outliers", serem identificados como discordantes.

Suponha-se que o desempenho do teste está a ser avaliado contra uma hipótese alternativa de mistura. Sob $\bar{H}$, o número de contaminantes é agora uma variável aleatória com distribuição binomial, designada por B.

As anteriores medidas P_1, P_3, P_5 e P_6 podem ser transpostas para testes de bloco definindo os acontecimentos D, E, F :

D: o teste declara que os k "outliers" são discordantes.
E_b: verifica-se $\bar{H}$ e a amostra contem pelo menos b contaminantes $(b = 1, 2, \ldots, k, k+1, \ldots)$.
F: verifica-se $\bar{H}$ e todos os k "outliers" são contaminantes.

Verifica-se que $P\,(\,D|\,B) = 0$ para $B < k$.

Podemos então escrever

$$\begin{aligned} P\left(\,D|\,\bar{H}\right) &= P_1 = P\,(Z > z_\alpha)\,; \\ P\,(\,D \frown F|\,E_k) &= P_3 = P\,(C = k)/\,P\,(E_k)\,; \\ P\,(\,D|\,F) &= P_5 = P\,(C = k)/\,P\,(F)\,; \\ P\,(\,D|\,E_k) &= P_6 = P\,(B \geq k, Z > z_\alpha)/\,P\,(B \geq k)\,. \end{aligned}$$

Um bom teste deverá ter P_5 e P_6 elevados, $P_1 - P_3$ reduzido; bem como $P\,(B < k)$ reduzido e $P\,(E_k)$ elevado.

No caso de um procedimento sequencial as medidas de desempenho que foram enunciadas para o caso de um único outlier necessitam também de ser generalizadas pois temos agora vários contaminantes.

Suponha-se que a hipótese alternativa indica dois valores discordantes superiores, numa amostra de n.

Os seguintes acontecimentos sob $\bar{H}$ são relevantes num procedimento interno (para um procedimento externo as definições D_0, D_1 e D_2 a seguir apresentadas devem ser adptadas).

Assim, temos:

$E_1 : x_{(n)}$ é um dos dois contaminantes
$E_2 : x_{(n-1)}$ é um dos dois contaminantes
$E = E_1 \frown E_2$: os dois contaminantes são os dois "outliers"

$D_0 : x_{(n)}$ não é considerado discordante (no primeiro teste)

$D_1 : x_{(n)}$ é considerado discordante (no primeiro teste) mas $x_{(n-1)}$ não é considerado discordante (no segundo teste)

$D_2 : x_{(n)}, x_{(n-1)}$ são ambos considerados discordantes (o que requer dois testes).

As medidas correspondentes a P_1, P_3 e P_5, são $P(D_1)$, $P(D_2 \frown E)$ e $P(D_2|E)$, respectivamente.

Outras medidas de interesse poderão ser $P(D_1|E)$ e $P(D_0|E)$ $(= 1 - P(D_1|E) - P(D_2|E))$.

8.6 Algumas Conclusões

No contexto das populações normais[8] acima introduzido, foram calculadas as diferentes medidas de desempenho considerando valores de δ entre 0.05 e 4 (correspondente a $\bar{H}_n$) para $n = 5; 10; 30; 100$. Os resultados são apresentados nos gráficos seguintes com o eixo δ em escala logarítmica.

Naturalmente os testes terão um melhor desempenho quanto maior for o desvio do parâmetro da distribuição da observação contaminante, ou seja quanto maior for o desvio de δ em relação a 1. Excepção para esta regra são os valores observados em P_7, pois esta é uma medida de não-desempenho, pelo que os resultados são o inverso dos outros. Os testes apresentam valores para as diferentes medidas de desempenho bastante consideráveis o que mostra que a detecção de outliers é viável para valores do parâmetro do contaminante não muito desviantes.

Um facto interessante é a diminuição das diferentes medidas de desempenho (excepto P_7 que é uma medida especial) à medida que aumenta o tamanho da amostra, contrariamente ao que sucede nos testes paramétricos de significância clássicos. Este resultado, já anteriormente referido por Braumann [21], para testes de discordância em populações exponenciais, explica-se pelas hipóteses alternativas.

Num teste de discordância, a hipótese alternativa refere-se à existência de um outlier, ou seja, um único elemento da amostra cujo valor do parâmetro da sua distribuição se afasta dos restantes. No caso de um teste clássico, quando a hipótese alternativa relativa ao valor de um parâmetro é verdadeira, todos os elementos da amostra vêm de uma distribuição com o valor desse parâmetro.

Deste modo no caso do teste clássico a potência de um teste de razão de verosimilhanças é geralmente uma função crescente do tamanho da

[8]Para um estudo geral e detalhado sobre este assunto pode consultar-se a recente tese de doutoramento de Palma [95].

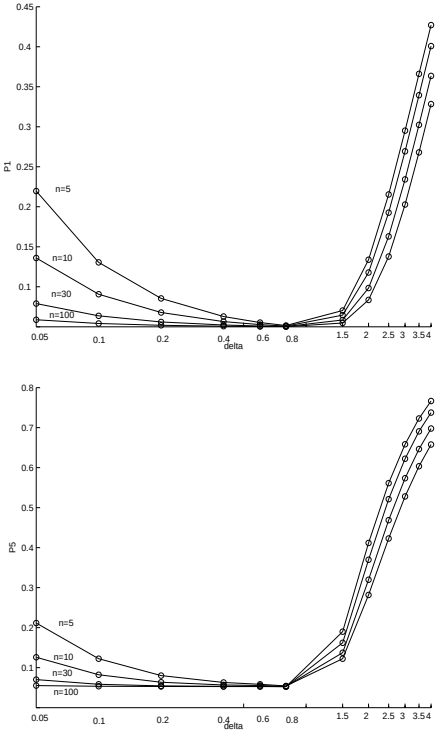

Figura 8.1: Desempenho de S_1 em Normais - P_1 e P_5.

amostra. É tanto mais fácil detectar um desvio do parâmetro em relação à hipótese nula quantos mais elementos da amostra contribuírem para o valor da estatística de teste (todos seguindo a hipótese alternativa).

No nosso caso, dado que a hipótese alternativa se refere à existência de um único elemento desviante, quanto maior a dimensão da amostra maior será o número de elementos e o mesmo para a proporção de elementos que têm um comportamento de acordo com a hipótese nula e, como tal, menor é a influência do único contaminante na estatística de teste. Este efeito traduz-se na diminuição da potência P_1do teste à medida que a dimensão da amostra aumenta, mas é igualmente visível nas outras medidas de desempenho, algumas ainda com maior intensidade.

Deste facto resulta que se a amostra é grande, dificilmente se detectará um "outlier", a menos que o valor de σ' se desvie significativamente de σ, isto é, se a distribuição do contaminante for bastante diferente.

Outra conclusão é que as estatísticas S_2 e S_3 são mais eficazes do que a corresponde estatística S_1, o que se compreende já que supõe-se na aplicação daquelas estatísticas que temos informação suplementar em relação ao comportamento dos parâmetros da população. A existência de informação suplementar traduz-se pois num melhor desempenho das estatísticas. Veja-se a figura 8.2 onde se pode comparar o desempenho das estatísticas S_2 e S_3 (na mesma linha a tracejado, com $\delta < 1$, valores para S_2 e $\delta > 1$ valores para S_3) com S_1.

Por último deve referir-se que o desempenho das estatísticas na situação em que $\sigma' < \sigma$ ($\delta < 1$) é pior do que na situação oposta $\sigma' > \sigma$ ($\delta > 1$), o que indica que é relativamente mais difícil detectar observações "outlier" que estão mais próximas do valor médio, isto é, que não são valores extremos.

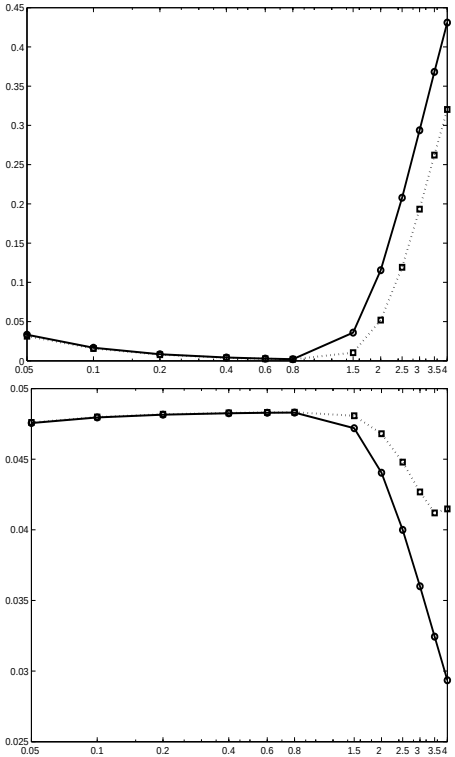

Figura 8.2: Desempenho de S_1, S_2 e S_3 - P_5 e P_7 ($n = 30$).

Neste breve percurso sobre a problemática dos "outliers" tivemos

oportunidade de focar alguns dos principais problemas que se colocam na sua detecção através da investigação no desempenho dos testes de discordância que, para esse fim, podem ser utilizados.

Salientamos a necessidade de alguma sistematização inclusive nas próprias noções envolvidas nas várias metodologias, desde logo a que se prende com a definição de outlier.

As dificuldades já enunciadas para a detecção de "outliers" é ainda agravada no momento da escolha da estatística de teste a adoptar face à grande variedade existente. A opção por uma ou outra estatística é uma questão fundamental pois uma observação poderá ser considerada "outlier" por um teste e não o ser por outro. Uma outra questão prende-se com a possibilidade de existência numa amostra de múltiplos "outliers" o que coloca dificuldades acrescidas na detecção.

Face à grande variedade de estatísticas disponíveis nos vários estudos para os testes de discordância existentes são necessárias medidas que permitam avaliar o desempenho e, simultaneamente, a sua comparação. Uma abordagem alternativa pode ser feita através dos contornos de sensibilidade que, para os múltiplos outliers, generalizam as curvas introduzidas por Tukey [134].

Neste assunto - sobre as medidas de desempenho - muito há ainda a fazer.

Capítulo 9

"Outliers" e Dimensão da Amostra

9.1 Introdução

A condição "outlier" em dados estatísticos é introduzida por alguma informação associada a uma observação e que é "divergente" em relação aos restantes. Este "afastamento" depende das diversas componentes a que está ligado o valor aberrante e, decerto, uma delas é a dimensão da amostra a que se encontra associado ou onde foi observado.

A selecção e tratamento de observações discordantes em sondagens e inquéritos é um tema recente e ainda pouco estudado[1] na teoria dos outliers.

Nas diversas questões sobre a amostragem em estatística, a dimensão da amostra é também um ponto fundamental.

Em análise de regressão, como já vimos no Exemplo IX de 1.2, várias abordagens podem ser utilizadas para elaborar um estudo "com" observações discordantes. Por exemplo, para métodos baseados em alterações na localização ou modelos onde possa acontecer uma inflação na variância, várias técnicas estão à disposição. De entre os métodos que incorporam alguma informação contida nesses dados contam-se "Trimming", "Winsorizing" ou "estimadores-M". Mas, os diagnósticos e a influência de "outliers" (ou grupos de outliers) trazem à liça (também) a dimensão da amostra. E principalmente nestes estudos onde o modelo depende de - e é condicionado por - todas as observações. Os resíduos

[1] Este tema apenas foi introduzido na 3ª edição de *Outliers in Statistical Data* de Barnett e Lewis. (Cf. os principais desenvolvimentos em [11], p. 440-7).

tornam-se então um elemento fundamental do estudo onde se usam todas as observações. Em circunstâncias apropriadas podemos usar métodos robustos, onde se pretende minimizar a influência dos "outliers". Assim estaremos a acomodar os valores discordantes. Sabemos que a variância dos resíduos não é constante e, quanto mais perto da média for o valor da variável regressora maior é a influência na variância estimada. Este efeito de alargamento[2], também depende da dimensão n da amostra.

Em modelação estatística recorre-se muitas vezes a aproximações que fazem intervir (por exemplo) as chamadas "leis dos grandes números". E, por consequência, em muitos estudos, sempre que se coloca a problemática da dimensão n, é habitual "desejar" n grande.

Num texto estatístico, mesmo de escrita "generalista", é comum surgir a conclusão: (...) para grandes amostras, isto é, para n maior do que 30 (...); ou então: (...) pelo teorema do limite central podemos garantir que (...) ou ainda, salientando dificuldades, (...) para pequenas amostras, para n menor do que 30 teremos (...). Como sabemos, não é simples esta separação dimensional das amostras.

Para este tema, sempre que se solicita um limite separador... o 30 é eleito.

Qual deverá ser a acepção desta perspectiva estatística numa teoria de "outliers"?

Numa primeira tese devemos admitir que: se, por exemplo, um modelo é apropriado para os dados em análise, os outliers são estatisticamente "mais esperados" há medida que n aumenta. O aumento da dimensão da amostra poderá produzir maior probabilidade de aparecimento de outliers. No entanto, o efeito de surpresa da observação outlier, como já sabemos, é um ponto fulcral na análise prática de observações discordantes. Também, esta possibilidade pode depender do modelo que se supõe. Em ambientes normais poderá existir um "comportamento" diferente de exponencialidade. E também, se os dados forem "aproximadamente normais" mas com caudas pesadas, outro mecanismo de geração de observações discordantes será criado. Assim, por exemplo, em amostras de distribuições Cauchy a "perturbação" na média e na variância amostral, criada por um outlier, para grandes valores de n, é maior[3] do que numa distribuição normal.

[2]Este efeito foi inicialmente introduzido em estudos de [18]. Uma abordagem geral desta problemática pode ver-se em [40]. Barnett e Lewis ([11] p. 321) propõem para este efeito a designação *ballooning*.

[3]Este temas foram inicialmente introduzidos por Neyman e Scott [93] e Green [54] provando que a família das log-normais e das distribuições gama têm maior possibilidade de fazer surgir outliers - são *outlier prone*. Main [90], em oposição, propõe que se introduza o conceito de *outlier resistant*. Por sua vez, Goldstein [53] examina

Nalguns livros de texto[4] pode ler-se: "*Any result of a series containing n... observations shall be rejected when the magnitude of its deviation from the mean of all measurements is such that the probability of occurrence of all deviations as large or larger is less than* $1/2n$".

O critério de Chauvenet, como sabemos, é baseado nesta mesma filosofia de afastamento ou de desvio em relação ao esperado. Desde há longa data muito utilizado na prática, também é um bom exemplo de reflexão sobre a influência da dimensão da amostra na selecção de outliers.

O primeiro "teste" objectivo para estudar k observações suspeitas aos olhos do experimentador foi desenvolvido por Peirce [98], em 1852, alguns anos antes do critério de Chauvenet. De acordo com aquele critério, devemos rejeitar uma obervação suspeita sempre que a "*probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations*". Peirce faz então introduzir no seu critério uma probabilidade de rejeição que depende de n. Assim, desde o início e mesmo nos mais simples critérios, vemos a sensibilidade dos estatísticos para a influência da dimensão da amostra.

9.2 Exemplos

Na secção 7.4, já introduzimos - no exemplo IX - alguma reflexão sobre a problemática da alteração de um estudo de outliers apenas pela variação da dimensão da amostra.

Vejamos outros exemplos.

Exemplo VIII (conclusão):

Para o conjunto de dados analisado em 1.2, e sobre os quais Peirce e Chauvenet consideraram que o valor -1.40 *deve ser rejeitado, sabemos também que ao nível de 5 por cento esse mínimo não deve*[5] *ser declarado outlier.*

Para os dados em referência vamos criar amostras modificadas - acrescentando valores iguais à média - utilizando o mesmo método de exemplos anteriores; desta vez para populações normais.

Admitindo um modelo de discordância com alternativa natural para o caso mais geral, a estatística de teste utilizada em (5.7) permite concluir que, logo para n=16 a média 0.018 - que foi acrescentada - transforma-se em candidato a "outlier". O mínimo deixa de ser o candidato apenas

algumas características bayesianas de uma observação gerada por uma distribuição resistente a outliers.

[4]Cf. [24], citado por Barnett e Lewis ([11], p.4).

[5]Cf. Barnett e Lewis ([11], p. 38).

por acrescentar o valor 0.018 aos dados iniciais. Além disso o valor 8.415E-18 atingido pela estatística de teste garante a decisão: a média é "outlier".

Se, em vez de 0.018 fosse usado 0.017 ou 0.019 - igualmente afastados da média - a simetria naquela estatística originava o mesmo valor 0.0004547 para a estatística de teste. Este valor não dá rejeição e portanto 0.017 ou 0.019 não seriam "outliers".

Exemplo XIII (conclusão):

Continuemos a estudar os dados exponenciais introduzidos no exemplo XIII de (5.2.4.3) e já analisados em 5.2

$$4, 7, 10, 17, 19, 25, 31, 34, 45, 52, 61,$$

$$64, 76, 87, 101, 116, 141, 181, 240, 446, 503.$$

Admitimos a presença de um "outlier" na amostra, suspeitámos do 4 e do 503 e concluímos ainda que, se o mínimo for 3 (e não 4!) então o 3 é candidato a "outlier". Para dois diferentes modelos de discordância, com base nos pontos críticos das respectivas tabelas podemos decidir que a amostra é homogénea e, portanto, não existe valor discordante.

n	$x_{(1)} = 4$	$x_{(n)} = 503$	$x_{(1)} = 3$	$x_{(n)} = 503$
21	1,71E-03	**1,45E-03**	**1,29E-03**	1,44E-03
22	1,63E-03	**1,41E-03**	**1,23E-03**	1,41E-03
23	1,56E-03	**1,37E-03**	**1,18E-03**	1,37E-03
24	1,49E-03	**1,34E-03**	**1,13E-03**	1,33E-03
25	1,43E-03	**1,30E-03**	**1,09E-03**	1,30E-03
30	1,20E-03	**1,15E-03**	**9,05E-04**	1,14E-03
32	1,12E-03	**1,09E-03**	**8,48E-04**	1,09E-03
34	1,05E-03	**1,04E-03**	**7,98E-04**	1,04E-03
35	**1,02E-03**	1,02E-03	**7,76E-04**	1,02E-03
36	**9,96E-04**	9,99E-04	**7,54E-04**	9,97E-04
37	**9,69E-04**	9,78E-04	**7,34E-04**	9,76E-04
38	**9,43E-04**	9,57E-04	**7,14E-04**	9,55E-04
39	**9,19E-04**	9,37E-04	**6,96E-04**	9,36E-04
40	**8,96E-04**	9,18E-04	**6,79E-04**	9,17E-04
45	**7,96E-04**	8,34E-04	**6,03E-04**	8,32E-04
50	**7,17E-04**	7,63E-04	**5,43E-04**	7,61E-04

A partir destes mesmos dados vamos fazer um breve estudo no sentido de verificar alguma influência da dimensão da amostra. Para isso,

"aumentemos" a dimensão n, juntando observações iguais ao valor da média quer para a amostra original (com mínimo 4) quer para a amostra "modificada" (com mínimo 3).

Na tabela acima, para cada valor da dimensão n da amostra apresentamos os correspondentes resultados da estatística de teste, para as duas amostras - a original e a modificada. Estes "novos dados" para estudo, permitem que não sejam alterados os principais indicadores na distribuição e no modelo de discordância. Os valores, salientados a negro carregado, indicam as observações onde a estatística do teste (5.7) para a discordância atinge o mínimo indicando por conseguinte, o correspondente candidato a "outlier". Não sendo suficientemente pequenos, os valores surpreendentes encontrados não confirmam os respectivos candidatos como "outliers". A tabela mostra a influência da dimensão da amostra na determinação do valor discordante. Podemos verificar que, para a amostra modificada, o mínimo 3 se mantém como candidato mesmo fazendo variar a dimensão. Parece mesmo que, à medida que n aumenta, mais se salienta esse facto. Na amostra original temos o máximo 503 como candidato a outlier até à dimensão 34. A partir desse valor surge o mínimo 4 como suspeito.

Na amostra modificada, o mínimo 3 é sempre suspeito, para os diversos valores de n.

Na amostra original, o mínimo 4 a princípio suspeito, perde essa condição.

De acordo com o exemplo anterior, deve-se registar que a influência, nas conclusões, se verifica para valores de n que estão um pouco acima das chamadas pequenas amostras onde o estudo de outliers deve ser mais eficaz. Para pequenas amostras a observação outlier tem uma conotação e "um peso" acrescido pelo facto de surgir discordante "entre poucos". Razão contrária leva à "diluição" de um valor discordante inserido numa grande amostra. Além disso, para se tornar suspeito "entre muitos" é exigido muito mais "sobre o acaso". Este é, também, um problema de robustez, que a par da selecção[6] é um tema igualmente importante no estudo de outliers em dados estatísticos.

O método GAN para as duas amostras que estamos a considerar permite um estudo comparado, analisando a presença de dois outliers e a respectiva influência com a dimensão da amostra.

[6]É vasta a bibliografia sobre esta temática. Para uma consulta de referências gerais, ainda podemos considerar actual o tratado de Barnett e Lewis [11]. Um historial da contribuição portuguesa foi recentemente publicado por Branco em [19].

Assim, na continuação do exemplo e com os mesmos dados, na tabela abaixo, podemos verificar que apenas os pares $(x_{(1)}, x_{(2)})$ ou $(x_{(n-1)}, x_{(n)})$ são considerados como candidatos a outliers utilizando o método GAN com um modelo de discordância exponencial com os parâmetros desconhecidos. Verificamos também os diferentes valores da dimensão da amostra onde se altera a condição de candidato a outlier (115 na amostra inicial e 65 na amostra modificada).

n	$(3,7)$	$(446,503)$	$(4,7)$
50	3,16E-06	**2,80E-06**	
55	2,61E-06	**2,43E-06**	
60	2,19E-06	**2,13E-06**	
62	2,05E-06	**2,03E-06**	
64	1,93E-06	1,93E-06	
65	**1,87E-06**	1,88E-06	
66	**1,81E-06**	1,84E-06	
70	**1,61E-06**	1,67E-06	
75	**1,40E-06**	1,49E-06	
-	-	-	-
105		**8,40E-07**	8,57E-07
110		**7,74E-07**	7,81E-07
112		**7,49E-07**	7,53E-07
114		**7,26E-07**	7,27E-07
115		7,15E-07	**7,14E-07**
116		7,04E-07	**7,02E-07**
118		6,83E-07	**6,79E-07**
120		6,62E-07	**6,56E-07**

Após a análise acima feita sobre um exemplo em populações exponenciais, vamos formular a mesma questão em ambiente normal.

Continuemos, utilizando de novo, o:

Exemplo III (conclusão):

Consideremos os dados já analisados em 1.2.

Numa primeira parte - exemplo II - admitimos os valores 2, 2.8 e 3.4. Uma análise preliminar destes dados não é conclusiva sobre a eventual discordância de alguma das observações. Aplicando um modelo de discordância onde admitimos os dois parâmetros de escala desconhecidos verificamos que a observação 2.8 é surpreendente.

Como nos exemplos acima, analisemos esse facto à medida que n aumenta e para tal construímos amostras acrescentando aquela mesma observação 2.8.

Apresentamos em seguida uma tabela onde registamos qual a observação discordante para cada valor da dimensão com a amostra artificialmente construída.

Verificamos que para uma dimensão superior a 8 passamos a ter o mínimo 2 como observação discordante e deixa de ser surpreendente a "observação do meio".

Constatamos ainda que esse valor 2 é "outlier" a partir da dimensão 15 (o respectivo valor da estatística de teste está a negro carregado).

n	2	2.8	3.4
3	3.36E-01	**6.68E-02**	3.69E-01
4	2.14E-01	**5.01E-02**	2.84E-01
5	1.33E-01	**4.00E-02**	2.21E-01
6	8.19E-02	**3.34E-02**	1.74E-01
8	3.05E-02	**2.50E-02**	1.09E-01
9	**1.85E-02**	2.22E-02	8.70E-02
10	**1.12E-02**	2.00E-02	6.92E-02
13	**2.46E-03**	1.54E-02	3.51E-02
14	**1.48E-03**	1.43E-02	2.80E-02
15	**8.93E-04**	1.33E-02	2.23E-02
20	**7.04E-05**	1.00E-02	7.26E-03
30	**4.31E-07**	6.67E-03	7.73E-04
40	**2.63E-09**	5.00E-03	8.27E-05
50	**1.60E-11**	4.00E-03	8.86E-06
75	**4.56E-17**	2.67E-03	3.34E-08
100	**1.30E-22**	2.00E-03	1.26E-10

Exemplo XIV (conclusão):

Como vimos, um exemplo "análogo ao anterior" pode construir-se através da amostra 4.9, 5.9, 7. Também aqui a observação "ao meio" surge como surpreendente.

n	4.9	5.9	7
3	3.59E-01	**2.24E-02**	3.48E-01
4	2.62E-01	**1.68E-02**	2.38E-01
5	1.92E-01	**1.35E-02**	1.62E-01
10	4.18E-02	**6.73E-03**	2.27E-02
13	1.69E-02	**5.17E-03**	6.96E-03
14	1.25E-02	4.80E-03	**4.68E-03**
18	3.73E-03	3.74E-03	**9.62E-04**
19	2.76E-03	3.54E-03	**6.48E-04**
20	2.04E-03	3.36E-03	**4.36E-04**
21	1.51E-03	3.20E-03	**2.93E-04**
22	1.12E-03	3.06E-03	**1.97E-04**
30	1.00E-04	2.24E-03	**8.30E-06**
40	4.92E-06	1.68E-03	**1.58E-07**
50	2.42E-07	1.35E-03	**2.99E-09**
75	1.30E-10	8.97E-04	**1.49E-13**
100	6.96E-14	6.73E-04	**7.37E-18**

Conclusões análogas às anteriores podem também neste exemplo ser estabelecidas. Os resultados estão apresentados no quadro acima.

Agora, é para n=13 que se altera a ordem do valor surpreendente e para uma amostra com dimensão superior a 18 vemos o valor 7 ser confirmado como um "outlier". Ao contrário do anterior que realçou o mínimo, neste exemplo a observação surpreendente "salta" para o máximo.

Para terminar estes exemplos de reflexão sobre a influência da dimensão da amostra na tomada de decisão sobre "outliers" retomemos o:

Exemplo X (conclusão)**:**

Consideremos, de novo, o caso H-H do exemplo X, já considerado nos capítulos 1 e 7. Um estudo análogo introduzindo a influência da dimensão da amostra pode fazer-se também para este exemplo de estatística forense.

Admitamos que o dado usado para construir a amostra modificada é, também neste caso, a média 273.33.

Utilizando as estatísticas para os testes de homogeneidade das amostras e para os diferentes valores da dimensão n podemos obter a tabela seguinte onde, a negro carregado, se indicam os valores suspeitos.

n	200	280	340
3	3.33E-01	**6.62E-02**	3.68E-01
4	2.13E-01	**5.00E-02**	2.83E-01
5	1.33E-01	**4.00E-02**	2.21E-01
6	8.18E-02	**3.33E-02**	1.74E-01
7	5.30E-02	**2.80E-02**	1.37E-01
8	3.06E-02	**2.50E-02**	1.09E-01
9	**1.84E-02**	2.20E-02	8.70E-02
10	**1.11E-02**	2.00E-02	6.92E-02
20	**7.03E-05**	1.00E-02	7.26E-03
30	**4.31E-07**	6.67E-03	7.73E-04

É claro que se, no exemplo anterior, tivéssemos acrescentado aos dados em estudo, o "valor médio conhecido 280" a análise que se faria iria sempre produzir a suspeita sobre 280.

Registe-se finalmente - e como seria de esperar - a coincidência de valores nas tabelas dos exemplos III e X.

9.3 Conclusão

No exemplo IX, que analisámos em 1.2 e em 7.4, temos oportunidade de verificar como a introdução de uma observação no estudo - que no presente contexto se pode entender como um pequeno aumento da dimensão da amostra - pode alterar totalmente uma análise de dados.

Os exemplos apresentados são elucidativos sobre a influência da dimensão no estudo de outliers numa amostra. Muitas outras situações podemos elaborar, inclusive a variação no próprio modelo admitindo outras hipóteses sobre o conhecimento ou não dos vários parâmetros envolvidos. Todas reforçam a subjectividade no estudo de outliers. Uma observação discordante num determinado modelo poderá deixar de o ser num outro onde pequena alteração foi feita. O estudo da presença de outliers numa amostra requer bastante experiência. Desde os princípios que os estatísticos se preocupam com os dados não representativos ou discordantes. Vários têm sido os mecanismos propostos para justificar a presença dessas observações que "reduzem" ou "distorcem" a informação contida numa amostra. Como causa principal é assumida a forma da distribuição. Também o modelo assumido pode ser responsável pela presença de observações discordantes. O conceito de "outlier" merece então ser visto em termos relativos. Neste estudo manifestou-se a dimensão da amostra também como um novo elemento dessa relatividade.

Da análise que fizemos também podemos concluir sobre a importância do estudo de "outliers" em amostras de média dimensão; sendo portanto

fundamental caracterizar o que se deve entender como tal pois, como se sabe, o valor de separação depende das distribuições em estudo.

Capítulo 10

Sobre o Estudo de "Outliers" Multivariados

10.1 Introdução e algumas notas gerais

Após uma abordagem univariada[1] para o estudo de outliers em dados estatísticos impõe-se que olhemos um pouco para as observações multivariadas e, principalmente, para as especificidades que, neste âmbito, as caracterizam. Aliás, numa era com (fácil) acesso à informação, é de observações multivariadas que, cada vez mais, se compõe a matéria prima para o trabalho dos estatísticos.

Para elaborar um estudo de outliers em amostras de uma só dimensão necessitamos principalmente de estatísticas ordinais.

Uma área da análise multidimensional onde também não se pode evitar relações de ordem é a identificação de observações discordantes. Pela impossibilidade de uma ordenação única dos dados multivariados, a selecção e a detecção de outliers numa perspectiva multidimensional fica, à partida, com essa grande limitação. No entanto, algumas subordens podem ser usadas no sentido de, neste domínio, permitir a selecção de observações discordantes. Em [4] podem consultar-se os principais resultados.

Um dos pioneiros, neste campo multivariado, foi Wilks [137] que considerou em detalhe um teste para um outlier envolvendo "estatísticas

[1]Embora com algumas incursões pelo domínio bivariado como fizemos nos exemplos que apresentámos sobre o estudo de outliers em dados estruturados e onde a problemática do estudo de outliers multivariados, embora superficialmente, já foi introduzida.

internas" baseadas em determinantes de somas de quadrados e produtos cruzados com e sem a observação a estudar.

Sobre a motivação dos estatísticos para a problemática dos outliers, do ponto de vista histórico, podemos referenciar [6]. É uma edição correspondente às actas da conferência "Looking at Multivariate Data" organizada pela Universidade de Sheffield em 1980 e onde se reuniram os principais especialistas desta área. Naquela obra publicaram um olhar sobre a estatística multivariada que em muitos campos nos dá a visão actual. Mas, no estudo dos outliers muito se evoluiu...

Numa avaliação inicial - por exemplo uma análise exploratória - de dados não estruturados, sabemos que a discordância de qualquer observação está associada ao seu afastamento em relação às restantes. E esta característica também exige que no nosso pensamento se ordenem os dados para então decidirmos qual está mais distante do que é admissível, para uma determinada amostra. Este é, obviamente, um discurso univariado. Para os dados multivariados podemos avançar um pouco admitindo que um outlier possa estar longe da "nuvem de pontos". Mas esta "muda" sempre que se faça uma rotação nos dados como podemos confirmar com o auxílio de um pacote estatístico com capacidades para esta função "plot 3-D".

A identificação de outliers em dados multivariados é vulgarmente baseada na distância de Mahalanobis. O uso de estimadores robustos para o vector dos valores médios ou da matriz de covariâncias também é uma preocupação neste estudo. Para uma perspectiva global pode-se consultar Becker ([17], p. 3-18) nas Actas do X Congresso Anual da SPE; uma boa e actualizada lista de referências.

Como já referimos no capítulo 2 e em particular na secção 2.9.3, podemos ter uma amostra de observações multivariadas, com vectores aleatórios compostos por medições para as quais todas e cada uma delas não são discordantes em relação à sua respectiva distribuição marginal mas existe uma (ou mais) que surpreende o experimentador por alguma característica que é revelada, por exemplo, numa análise[2] exploratória.

Portanto, para além da subjectividade, por diversas vezes já referida, existem particularidades, com dificuldades acrescidas, que apenas se verificam no estudo de outliers em dados multivariados.

[2]O exemplo mais "clássico" é o que associa a altura e o peso dos humanos. Um homem com cem quilos estará no limite dos "outliers no peso". Uma conclusão semelhante se pode alinhar em relação à altura de um outro indivíduo com um metro e meio. No entanto, é a associação - daqueles dois valores num indivíduo - que cria um dado surpreendente. Mas não só, como vimos, nos exemplos do capítulo 2.

A primeira é aquela que corresponde à necessidade de "reformular" a noção de "afastamento em relação aos restantes" para uma perspectiva multivariada. Como vimos, esta discordância para os dados univariados está sempre associada à ordenação dos mesmos. Ela é única e da maior utilidade para o estudo dos outliers. Mas, a ordenação é uma característica univariada e que se perde desde a dimensão dois. De entre muitos estudos sobre esta problemática, salientamos Kendall [78] que, há 40 anos, apresentou as diversas dificuldades no contexto da Classificação e Análise Discriminante.

Uma segunda dificuldade para um estudo multivariado junta-se em determinados dados estatísticos[3] - como no caso dos dados direccionais. Num exemplo muito simples, os valores 10 e 370 - coincidentes numa representação planar de coordenadas geográficas - podem corresponder ao mesmo ângulo de duas observações (ou de uma só em duas leituras distintas?) e que embora com valores diferentes sendo a mesma direcção aqueles dados não são discordantes.

Uma dificuldade acrescida no caso multidimensional surge em censos, em estudos de opinião ou em tabelas de contingência. Estas são três áreas onde se desenvolvem as investigações mais recentes[4] do estudo de outliers em dados estatísticos e que têm sido pouco exploradas, quiçá pela especificidade. O mecanismo de geração dos dados pode "enfrentar" o acaso com um "filtro" mais forte. Um valor surpreendente pode ser facilmente censurado por um inquiridor que deseja produzir dados de qualidade.

Com grandes volumes de dados multivariados é muito importante, do ponto de vista de prático, encontrar representações simples que permitam construir uma "sensibilidade" sobre o comportamento de cada observação em relação às restantes.

Os resíduos - entendidos na generalidade como afastamentos que tanto podem ser construídos numa análise em componentes principais como num modelo de regressão multivariada - podem usar-se para esse fim.

Além disso e também para observações multivariadas podem fazer-se estudos que envolvam a função influência[5] (ou função de influência?).

Todas estas questões se entroncam numa análise estatística robusta. Diversos estatísticos portugueses têm apresentado contribuições para

[3]De novo encontramos trabalhos de Collett [27] - percursor desta sensibilidade específica do estudo de outliers.

[4]Estas áreas são incluídas no capítulo 12 de [11]. É uma das novidades, na 3ª edição, de *Outliers in Statistical Data*.

[5]Estes estudos foram inicialmente desenvolvidos por Hampel (Cf. [59], [60], [62] e [65]).

esta área. Salientamos os estudos de Pires iniciados em [102]. Uma boa síntese - também com a contribuição portuguesa - foi recentemente feita por Branco, em [19].

O capítulo 7 de Barnett e Lewis [11] é uma indicação fundamental para conteúdo e métodos do estudo de outliers em dados multivariados. Entre muitas outras, o capítulo 3 de [132] é boa referência para o estudo de outliers em ambiente normal.

Para modelos estruturados pode usar-se [52] que no capítulo 6 apresenta e desenvolve técnicas internas e externas para a detecção de outliers multivariados. Igualmente, é um bom guia, para uma abordagem dos resíduos das componentes principais (*ib.* p. 294) que em 10.2.3, num estudo de caso, também usaremos.

Para a problemática da influência de outliers multivariados nas regras de decisão, por exemplo de uma análise discriminante, pode consultar-se o recente estudo de Pereira e Pires em [99], que apresenta uma nova regra de rejeição de outliers baseada em análise de clusters e utilizando a distância de Mahalanobis.

Para a recente área da estatística ambiental é muito importante o capítulo 3 de [9], onde é desenvolvida principalmente a questão da robustez em dados de um mundo que acaba de entrar num novo milénio com uma forte crise ambiental onde "amanhã é demasiado tarde".

Neste domínio, os dados aberrantes têm maior força?

10.2 "Outliers" e Componentes Principais

10.2.1 A Redução da Dimensionalidade como Objectivo

A redução da dimensionalidade é um dos grandes objectivos de uma Análise Estatística em Dados Multivariados. A escolha do número de "factores a reter", no contexto duma simplificação da complexidade dimensional através duma Análise em Componentes Principais (ACP), tem sido objecto de numerosos critérios, em que geralmente são necessárias opções de natureza subjectiva por parte do utilizador. Esta subjectividade também pode influenciar o estudo de outliers multivariados. Mas, deixemo-la de parte e abordemos a ACP com todas as variáveis e portanto na mesma dimensão que os dados originais.

Os principais instrumentos práticos, largamente utilizados, para a exploração de observações discordantes em "duas ou três dimensões" são os diagramas das variáveis originais e das transformadas correspondentes às componentes principais.

As curvas de Andrews, de fácil construção e com acesso em alguns pacotes estatísticos, são um poderoso instrumento exploratório dos dados.

Igualmente, as chamadas técnicas externas, como a análise discriminante ou a análise canónica podem ser usadas[6] com bons resultados.

Pela facilidade de construção[7] em qualquer pacote estatístico, consideramos de muita utilidade os gráficos de contornos de quantis. Estes permitem[8] identificar observações que estão no limite de algum comportamento probabilístico a que correspondem os diversos quantis.

A ACP é uma técnica estatística poderosa e de grande aplicação praticamente em todas as áreas do conhecimento. Como é sabido a ACP desenvolve-se na procura da maior parte da explicação para a variabilidade global dos dados, entendida como a soma das variâncias de todas as variáveis.

Na ACP as variáveis originais são transformadas em novas variáveis denominadas componentes principais, através de uma transformação e com o objectivo estratégico destas terem variâncias decrescentes.

Duas são as características mais relevantes das componentes principais e que as tornam mais importantes que as variáveis originais: a não correlação e a quantidade de informação retida. Por um lado, as variáveis originais podem estar fortemente correlacionadas e estas interdependências, não desejadas, são eliminadas nas componentes principais. Assim, por construção, as componentes principais são ortogonais entre si. Deste modo, cada componente traz uma informação estatística diferente das restantes. Além disso, cada componente maximiza aquela informação. As variáveis originais, comparadas entre si, têm a mesma importância, enquanto que as componentes têm relevância estatística decrescente. Destas características, podemos compreender como numa análise, as componentes principais:

- podem ser estudadas separadamente devido à ortogonalidade, permitindo interpretar o peso das variáveis originais na combinação das componentes principais mais importantes

- podem servir para visualizar o conjunto da amostra apenas pelo gráfico das duas primeiras, que detêm a maior parte da informação estatística.

Vários autores sugerem que se faça uma ACP preliminar aos dados em estudo e que se analisem as projecções das observações nas componentes principais de diferentes ordens. Como a última componente principal tem a menor variância - também sobre esta perspectiva - é sobre ela

[6]Consultar , por exemplo, Gnanadesikan [52].

[7]Veja-se [52], p. 232 e seguintes.

[8](*ib.* p. 252), com um bom exemplo para os famosos Dados Iris de Fisher.

que deve recair a maior importância para detecção de observações que se afastem e discordem das restantes.

Quantas componentes devemos reter? Esta é uma questão fundamental para um análise de dados multivariados e que, por consequência, se transmite para um estudo de outliers multidimensionais. Sobre este assunto, são várias as dificuldades que surgem, e portanto também por aqui é introduzida a subjectividade das conclusões.

Dadas as alternativas dessa selecção ser feita de um modo multivariado, como acima referimos, tem sido usada a ACP como uma alternativa para o estudo da discordância de observações numa amostra. Pretendemos estudar a ACP como metodologia para detecção de outliers. Assunto muito importante, é a dicotomia valor discordante - valor influente, mas que não é objecto deste estudo.

Como vimos, a presença de observações que são outliers influentes deve ser estudada através de métodos robustos. Algumas vezes, a pesquisa de outliers envolve, também, o estudo das diferenças entre valores ajustados e observados. No caso geral temos um vector multidimensional de resíduos. Muito mais importante do que numa só dimensão, devemos saber como interpretar essas observações diferenças, cuja presença pode ser devida a duas diferentes abordagens das questões multivariadas - uma análise da estrutura interna dos dados ou uma análise externa onde é suposta uma estrutura nas observações. Podemos englobar no primeiro caso, por exemplo, a análise em componentes principais e *multidimensional scaling* onde um objectivo é o estudo de dependências internas tendo em vista a redução da dimensionalidade. Impondo modelos de regressão multivariada ou numa análise multivariada da variância estaremos na segunda abordagem.

10.2.2 Resíduos nas componentes principais

Para o estudo de outliers multivariados sabe-se que as primeiras componentes principais são "sensíveis" à presença de observações discordantes que se liguem a dados com variâncias ou covariâncias inflacionadas. Por sua vez, as últimas permitem identificar perturbações nos resíduos.

No contexto acima referido, comecemos por analisar como as componentes condicionam e são decisivas para encontrar outliers multivariados. Uma ACP pode ser interpretada como o ajustamento aos dados de um conjunto de hiperplanos ortogonais minimizando a soma dos quadrados dos desvios ortogonais das observações a cada um daqueles planos. Consideremos o vector aleatório $\mathbf{x} = (X_1, ..., X_p)$ das observações num espaço de dimensão p. Admitamos $\mathbf{x}$, em geral, com valor médio e matriz de covariâncias.

O novo grupo de variáveis, as componentes principais $C_1, ..., C_p$ têm variâncias decrescentes e são não correlacionadas com $X_1, ..., X_p$. Supõe-se ainda que cada componente C_i é uma combinação linear das variáveis X_i pelo que $C_i = \mathbf{c}_i^t \ \mathbf{x}_i \ i = 1, ..., p$ onde $\mathbf{c}_i$ é um vector de constantes. Uma vez que, no cálculo, cada vector é arbitrário pode impor-se a condição de normalização $\mathbf{c}_i^t \mathbf{c}_i = 1$. Esta condição garante que a transformação das componentes principais é ortogonal e que, portanto, são mantidas as distâncias no espaço de dimensão p. Daqui infere-se que os resíduos podem ser importantes. Na abordagem proposta por Hotelling, as componentes principais devem ser determinadas pelo valor decrescente das respectivas variâncias. Devemos portanto calcular $\mathbf{c}_1$ de modo que C_1 tenha variância máxima e sujeita à condição de normalização acima referida. Este resultado é equivalente à metodologia inicialmente proposta por Pearson envolvendo os mínimos quadrados. Assim, e como acima referimos, podemos fazer uso de resíduos para posterior estudo de eventuais observações discordantes. Na metodologia dos multiplicadores de Lagrange e pela sua relação com a variância das componentes, principalmente as seleccionadas[9] para representar a maior variabilidade dos dados, os valores próprios $\lambda_i, (i = 1, 2, ..., p)$, da matriz factorizada, são bastante importantes no estudo da discordância nas observações. Para avaliar da sua importância, vários testes podem ser efectuados sobre os valores próprios. Os cálculos efectuados garantem ainda que as constantes determinadas para a primeira componente principal são as coordenadas do vector próprio $\mathbf{c}_1$ correspondente ao maior valor próprio. Para a segunda componente principal um estudo análogo conduz a relevar o segundo maior valor próprio, etc. A metodologia assim definida vai permitir construir todas as p componentes principais. Porque irrelevante para o estudo que pretendemos, não abordaremos o caso em que alguns dos valores próprios da matriz factorizada surgem iguais. Portanto, quer os valores próprios quer os vectores próprios da matriz factorizada (de covariâncias S ou de correlações R) devem ser utilizados para análise de valores discordantes numa amostra. Se representarmos por C a matriz (pxp) dos vectores próprios e porque os valores próprios são iguais às variâncias das componentes temos, para os traços das diferentes matrizes envolvidas nos cálculos

$$tr(S) \ (ou \ tr(R)) = tr(C) = \sum_{i=1}^{p} \ \lambda_i = \sum_{i=1}^{p} var(X_i).$$

Verificamos assim que, a transformação envolvendo as componentes principais, mantém a soma das variâncias das variáveis originais e das trans-

[9]Factorizando a matriz de covariâncias ou de correlações.

formadas. Este é um resultado importante que também fundamenta a aplicação da ACP na pesquisa de observações discordantes pois a transformação que lhe está associada mantém a variabilidade dos dados. Podemos portanto "medir" a importância de cada componente através da percentagem da variação total nos dados que, por ela, é explicada. Se pretendermos que cada componente tenha valor médio nulo devemos usar uma constante adequada que obviamente vai envolver a média amostral $\bar{x}$. Através de $\mathbf{c} = C^t(\mathbf{x} - \bar{\mathbf{x}})$ obtêm-se as componentes.

Para um dado indivíduo x_i tem-se agora o correspondente vector observado $\mathbf{z}_i = C^t(\mathbf{x}_i - \bar{\mathbf{x}})$ cujas coordenadas são vulgarmente chamados "scores". Estes valores podem ser usados para a pesquisa e selecção de eventuais outliers multivariados na amostra.

Se for X a matriz dos dados então $Z = C^t X$ é a chamada transformação das componentes principais e cada uma das colunas da matriz C permite a determinação dos "scores" enquanto cada linha de Z mostra os desvios das projecções da amostra original no centroide, em relação a uma determinada componente principal.

Como sabemos, uma questão importante na detecção de um outlier multivariado é o facto de essa observação poder não ser extremo em nenhuma das variáveis mas "tornar-se" discordante pela conjugação dessas variáveis na estrutura de dependência interna dos dados. Assim, perante aquela realidade estatística, o estudo de cada variável constituinte de uma observação, torna-se um instrumento exploratório com vista à identificação de outliers multivariados. É o que faremos.

Quando usamos a ACP com objectivos de ajustamento a um espaço de menor dimensão e a consequente detecção de outliers a nossa atenção deve, principalmente, deter-se nas projecções dos dados sobre as componentes que correspondem aos menores valores próprios (isto é, as últimas linhas da matriz Z).

Este facto na selecção e detecção de valores discordantes numa amostra multivariada torna bastante relevantes as "últimas" componentes principais. Na figura 10.1, para a dimensão 2, esquematizamos este raciocínio.

Verificamos que, em relação ao "corpo" dos dados em torno da primeira e principal componente, a discordância do ponto P pode ser avaliada pelo resíduo ortogonal PQ que é equivalente a OP' sendo P' a projecção de P sobre a segunda componente principal. Então, o "score" de P na segunda componente principal permite avaliar da discordância multivariada de P e, pela metodologia da ACP, é mais importante do que o correspondente na primeira componente. E assim sucessivamente...

Mais geralmente, podemos usar a projecção dos resíduos no espaço gerado pelas últimas componentes para estudar a discordância multivariada de cada ponto.

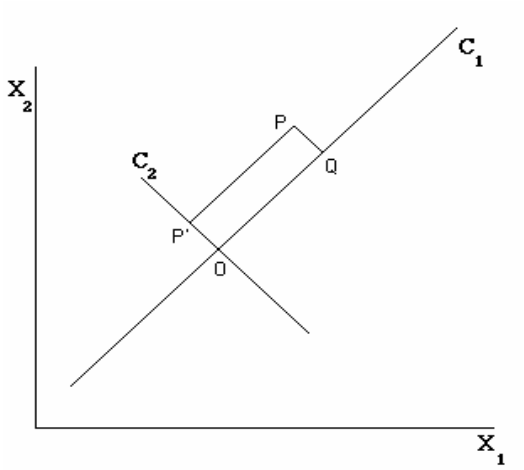

Figura 10.1: resíduos nas componentes principais

Para estudo destes resíduos Rao [106], discutido por Gnanadesikan e Kettenring [51], propõe que, para cada observação multivariada, seja calculada a estatística

$$d_{1i}^2 = \sum_{j=p-q+1}^{p} (\mathbf{c}_j^t(\mathbf{x}_i - \bar{\mathbf{x}}))^2$$

sendo um resíduo d_i^2 demasiado grande considerado como correspondente a uma observação que se afasta do espaço "gerado por todas as outras". Outras estatísticas têm entretanto surgido como alternativas ao estudo daqueles resíduos. Por exemplo, Hawkins (1980) sugere o uso dos valores próprios para dar igual peso às componentes e define

$$d_{2i}^2 = \sum_{j=p-q+1}^{p} (\mathbf{c}_j^t(\mathbf{x}_i - \bar{\mathbf{x}}))^2/\lambda_i$$

onde λ_i representa um valor próprio envolvido na componente principal do mesmo índice.

Note-se que, quando p=q, d_{2i}^2 é a distância de Mahalanobis entre a observação em estudo e a média geral $\bar{\mathbf{x}}$. Gnanadesikan e Kettenring [51] também consideram, com p=q, a estatística

$$d_{3i}^2 = \sum_{j=1}^{p} (\lambda_i \ \mathbf{c}_j^t (\mathbf{x}_i - \bar{\mathbf{x}}))^2$$

que realça observações que inflacionam as variâncias e as covariâncias e que afectam principalmente as primeiras componentes.

Hawkins [65], por sua vez, mostra que observações discordantes podem ser detectadas através da estatística

$$d_{4i} = max_{p-q+1 \leq j \leq p} \ \ \mathbf{c}_j^t (\mathbf{x}_i - \bar{\mathbf{x}}) / \sqrt{\lambda_j}$$

Em Gnanadesikan e Kettenring [51] e Hawkins [66] podem encontrar-se pontos críticos para as estatísticas acima referidas.

Numa fase exploratória, através do estudo das respectivas variâncias e também pela presença de outliers nas amostras univariadas correspondentes a cada uma das variáveis pode avaliar-se da sua contribuição para a discordância multivariada nos dados. Vamos seguir um possível trajecto para esse estudo aplicado a um modelo de discordância admitindo populações normais. Correspondendo a uma situação bastante geral, para elaborar um estudo, numa perspectiva não subjectiva na selecção de observações discordantes, podemos usar a estatística S_9 do capítulo 7

10.2.3 Um Estudo de Caso

O Estudo de Caso - muito em voga - é considerado um tipo de análise qualitativa "o irmão mais fraco dos métodos de investigação" e as pesquisas assim feitas têm sido consideradas desviadas das suas disciplinas, talvez porque as investigações que o utilizam possuem precisão, objectividade e rigor que facilmente se pode argumentar como insuficientes.

É também um recurso pedagógico ou uma maneira para se gerar "perspectivas" exploratórias.

Então, o método de Estudo de Caso, não deve ser usado com outros objectivos além da geração de ideias...

Posto isto, passemos ao exemplo seguinte:

O conjunto de dados que a seguir se apresenta corresponde a uma situação multivariada onde para cada um de 36 indivíduos foram registadas as oito variáveis indicadas (sendo as contínuas medidas em centímetros). Nesta amostra pretendemos fazer um estudo de observações discordantes. Para esse efeito, a par de uma análise em componentes principais bastante detalhada, faremos uma pesquisa de outliers univariados em cada uma das variáveis e assim poderemos concluir qual

delas pode ser mais responsabilizada pela eventual discordância na estrutura multivariada dos dados. Com vista à detecção de discordância nas componentes principais, complementando outros resultados, aplicaremos ainda as estatísticas d_{1i}^2 , d_{2i}^2 , d_{3i}^2 e d_{4i}.

Id	Sexo	Cabeça	Antebr	Pulso	Pé	Cintura	Peito	Altura
1	F	57	25	16	22	67	86	160
2	F	57	26	15	24.5	69	92	162
3	F	57	25	14	22	62	83	160
4	F	57	24	15	23	61	83	152.5
5	F	55	24	14	22	68	91	159
6	F	55	25	14.5	24	70	89	164
7	M	57	28	17	25	84	95	172
8	M	59	24	16	24	80	98	165
9	M	56	27	15	24	78	90	165
10	F	56	24	15.5	24	69	86	165
11	F	54	24	14	22	63	78	164
12	F	59	28	16	24.5	72	91	168
13	M	58	28	16	26	76	89	168
14	F	55	24	14	25	63	86	165
15	M	57	36	18	29	84	93	175
16	F	55	37	14	25	68	84	170
17	F	56	37	17	27	68	90	178
18	F	54	37	15	26.5	67.5	66	172
19	F	60	24	15	24	75	95	164
20	F	57	24	15	21	64	87	157
21	F	55	24	15	21	65	83	155
22	F	59	25	16	25	75	94	166
23	F	55	25	15	23.8	70	89	163
24	F	55	29	15.5	22	69	80	170
25	F	54	25	15.5	24.5	82	98	168
26	M	58	28	17	27	84	88	174
27	F	55	25	16	23	60	87	163
28	M	57	25	15	22	66	81	169
29	M	59	25	17	27	95	97	170
30	M	58	26	16	23	86	94	167
31	F	54	25	15	21	65	90	168
32	M	53	26	16	26	84	96	176
33	M	56	27	17	28	80	100	190
34	F	57	27	16	24	65	83	160
35	F	57	26	15	23.5	68	84	162
36	F	56	26	15	23	69	94	165

A partir da aplicação de um dos vários pacotes estatísticos podemos obter os resultados, os gráficos e as análises que a seguir apresentamos. Sendo instrumentos bastante disponíveis, podemos (e devemos!) avaliar da capacidade de cada um deles para a resolução do problema concreto para o qual pretendemos utilizá-los. Um estudo comparado, como primeira conclusão, permite afirmar a complementaridade daquelas aplicações estatísticas; enquanto uma é muito mais versátil quanto aos aspectos gráficos outra é muito mais potente em termos de cálculo ou mesmo na diversidade da metodologia e das várias soluções que nos propõe.

Os dados constam de medidas físicas efectuadas em 25 mulheres e em 11 homens. Pela análise da respectiva matriz vemos que nos homens a

maior correlação (0.84) se estabelece entre o pulso e o pé enquanto que nas mulheres aquele valor é apenas 0.31. Por outro lado, nas mulheres a maior correlação (0.71) aparece nas variáveis altura e antebraço sendo apenas 0.25 o valor correspondente para os homens. Na amostra global as variáveis mais correlacionadas (0.73) são altura e pé.

Continuemos o nosso estudo com uma breve análise em componentes principais feita em cada um dos sexos.

Para as mulheres e para os homens, usando as três primeiras componentes principais a partir da matriz das correlações, nas tabelas seguintes, apresentamos os "loadings" em cada uma das variáveis. Verificamos que estas componentes explicam 80% da variabilidade total nas mulheres e 86% nos homens.

Mulheres:

	Comp 1	Comp 2	Comp 3
Cabeça	0.05	0.69	0.56
Antebraço	0.76	-0.52	0.14
Pulso	0.52	0.36	0.52
Pé	0.85	-0.04	0.06
Cintura	0.55	0.58	-0.45
Peito	0.08	0.85	-0.33
Altura	0.88	-0.21	-0.21

Valores próprios: 2.66, 1.99 e 0.96
Variância explicada(%): 38, 28 e 14.

Homens:

	Comp 1	Comp 2	Comp 3
Cabeça	-0.08	0.76	0.49
Antebraço	0.59	-0.35	0.65
Pulso	0.93	0.13	0.26
Pé	0.93	-0.14	0.14
Cintura	0.66	0.60	-0.19
Peito	0.64	0.35	-0.59
Altura	0.67	-0.51	-0.31

Valores próprios: 3.37, 0.48 e 1.22.
Variância explicada(%): 48, 21 e 17.

A primeira componente, para as mulheres, mostra coeficientes positivos em todas as variáveis embora muito pequenos os correspondentes a Cabeça e Peito. Nos homens, quanto aos sinais, apenas Cabeça tem

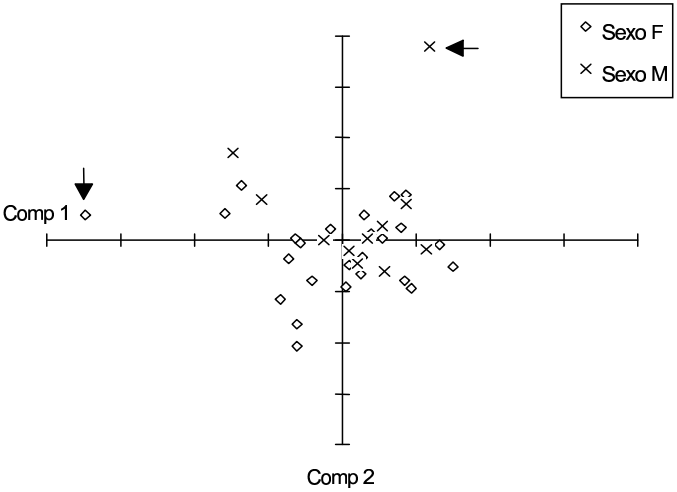

Figura 10.2: scores - componentes 1 e 2

coeficiente negativo igualmente pouco relevante e Peito passa a ser variável importante. Enquanto Altura, Antebraço e Pé são responsáveis pela maior contribuição nas mulheres, temos Pulso, Pé e Altura como variáveis mais correlacionadas com a primeira componente no caso dos homens. Na segunda e na terceira componentes existe uma concordância nos sinais dos coeficientes podendo considerar-se não relevantes as diferenças entre os valores observados que apenas invertem a ordem de importância nalguns casos.

Repetindo o estudo, desta vez para toda a amostra (com os dois sexos), podemos concluir que Pé, Pulso, Cintura e Altura são as variáveis mais importantes na construção da primeira componente principal. Antebraço, Peito e Cabeça, destacando-se das restantes, são as fundamentais para a segunda componente. Por sua vez, Cabeça é a única variável a considerar para a definição da terceira componente, que é responsável por 12% da variabilidade total. Relacionando as duas primeiras componentes, estabelece-se uma oposição entre Cabeça, Peito, Cintura, Pulso e Antebraço, Altura, Pé.

Porque, como vimos, são importantes para o estudo de outliers, registamos que as duas últimas componentes se repartem por três grupos de variáveis com diferente influência. Assim, Cabeça, Cintura, Antebraço, Peito e Pulso podem juntar-se, enquanto Altura e Pé estão isoladas e de

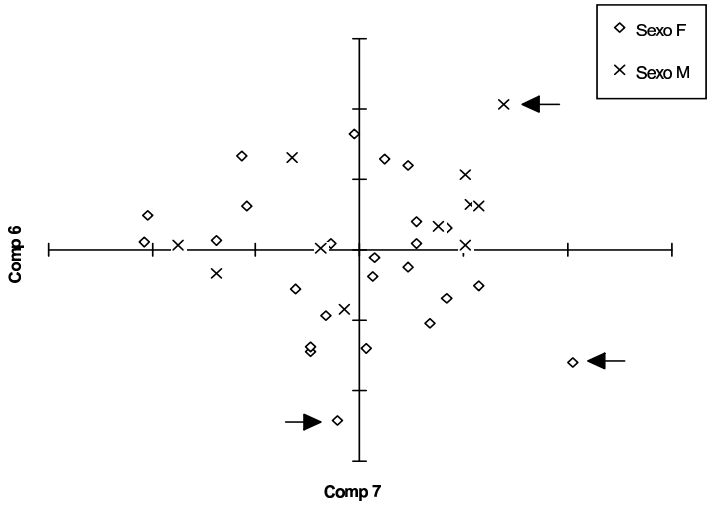

Figura 10.3: scores - componentes 7 e 8

algum modo, no contexto multivariado, com pesos opostos.

Depois de uma análise em torno das variáveis mais responsáveis pela estrutura multivariada dos dados, avancemos o estudo na direcção dos indivíduos com o objectivo primeiro de selecção de observações discordantes na amostra. Vamos, para esse fim, utilizar a matriz de covariâncias e analisar as duas primeiras e as duas últimas componentes principais. Os respectivos gráficos são apresentados nas figuras 10.2 e 10.3. Relativamente às duas últimas componentes salientam-se as observações 14 e 16 (a primeira mais do que a segunda) do sexo feminino e o indivíduo 15. Estes dados serão objecto da nossa atenção mais adiante. De facto as observações 14, 15 e 16 que se registam desde já como discordantes vão mais tarde confirmar-se como tal.

Em relação às duas primeiras componentes principais, que explicam 86% da variabilidade dos dados, verificamos a discordância do indivíduo 18 do sexo feminino e do 33 do sexo masculino. Esta última observação, que se destaca mais do que aquela, será mais adiante por outra metodologia de novo encontrada como discordante.

A observação 33 salienta-se também, embora ligeiramente, no gráfico correspondente às curvas de Andrews que a seguir se mostra. Numa análise aos valores de cada variável para este indivíduo verificamos que

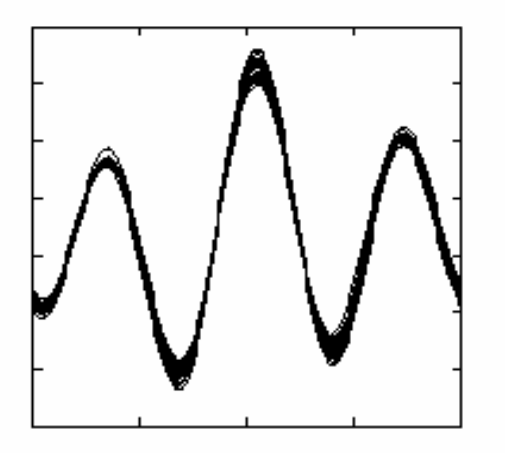

Figura 10.4: curvas de Andrews

tem a altura máxima e largamente acima da média. Esta pode ser uma explicação para a sua discordância neste gráfico. Por sua vez, como vimos acima, também a variável altura é importante na construção das duas últimas componentes principais que são fundamentais para a detecção de observações discordantes.

Procedamos agora a uma análise univariada em cada uma das variáveis com vista a um estudo de observações discordantes nesse contexto. Vamos para tal usar o modelo de discordância geral, acima referido, para populações normais com parâmetros desconhecidos e estudando a variação da estatística S_9 de 7.2.3.

Todas as variáveis são aceites numa distribuição normal. Por fornecerem valores bastante baixos para a estatística salientam-se a observação 12, muito próximo da média para a variável cintura, e a observação 33 nas variáveis altura e antebraço (esta igualmente junto da média). O indivíduo 18, já acima referenciado como discordante, também no contexto univariado surge com a variável peito, atingindo o mínimo, a atribuir um valor pequeno para a estatística de discordância. Porque é muito relevante para um estudo comparativo entre outliers a uma só e a mais do que uma dimensão registamos que as observações 14, 15 e 16, já encontradas como discordantes no domínio multivariado, não revelam qualquer suspeita no estudo univariado. Por sua vez e pelo contrário, o indivíduo 33 visto com alguma discordância a uma dimensão mantém

essa característica no estudo multivariado em várias análises.

Em resumo, no domínio univariado, surgem-nos indivíduos, como potenciais "outliers", que correspondem ao mínimo da amostra (observação 18 na variável peito), ao máximo (observação 33 na variável altura), ou estão demasiado próximas da média (observação 12 na variável cintura).

Para finalizar, vamos proceder a um estudo multivariado através das estatísticas d_{1i}^2 , d_{2i}^2 , d_{3i}^2 e d_{4i} . Para simplificação de notação vamos representar estas estatísticas por D_1, D_2, D_3 e D_4 respectivamente. Os quadros seguintes, apresentam os valores destas estatísticas quando se consideram as últimas q componentes principais. Para q=1 registamos apenas D_1 porque os valores das restantes estatísticas são iguais neste caso. A primeira tabela resulta de uma análise em componentes principais usando a matriz das covariâncias e a segunda foi obtida a partir das correlacções entre as sete variáveis.

q=1	q=2				q=3			
D_1	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
16	14	16	14	16	19	16	32	16
15	15	15	31	15	14	15	19	15
27	16	14	24	27	32	14	31	27
14	24	27	15	14	15	27	25	14

Nesta tabela, onde registamos as observações correspondentes aos quatro maiores valores para cada uma das estatísticas, pela sua maior frequência salientam-se, como discordantes, os indivíduos 14, 15 e 16. Embora não mantendo a ordem nas colunas são sempre essas observações que se evidenciam. Registe-se a informação, ligeiramente diferente, fornecida pela estatística D_3.

q=1	q=2				q=3			
D_1	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
31	33	33	33	33	33	33	33	33
14	24	24	4	31	29	29	29	29
24	31	31	24	14	30	30	30	31
30	4	28	28	24	4	31	4	14

Obtida a partir da matriz de correlações a tabela acima mostra, principalmente, a discordância das observações 24, 31 e 33. Note-se a repetitiva permanência da observação 33 no início de cada coluna correspondendo ao maior valor da respectiva estatística salientando assim a sua discordância multivariada. As observações que nesta última análise nos surgem como discordantes foram várias vezes estudadas como tal em vários contextos e salientam-se também por, em termos gráficos, aparecerem quase sempre nos respectivos "extremos". Este exemplo permitiu fazer uma abordagem univariada e simultaneamente um estudo bastante geral no campo multivariado com vista à detecção de observações discordantes e principalmente estabelecer conclusões numa análise comparativa entre esses dois domínios que facilita e eventualmente clarifica a própria noção de "outlier" multivariado.

Capítulo 11

Em Perspectiva

A Estatística é sempre associada à colheita e ao uso de dados de modo a apoiar a administração de um estado. O sistema de justiça é, na realidade, um dos pilares fundamentais de um estado moderno e é basilar na política da maior parte dos países. Os mais recentes avanços da teoria dos "outliers" têm surgido baseados na inferência estatística para interpretar dados de um ponto de vista legal. Os tribunais estão introduzindo novos desafios para os estatísticos que assim são solicitados a pronunciar-se em domínios de trabalho não tradicionais - por exemplo a correcta aplicação da legislação envolvendo os direitos de autor ou, com muito maior impacto, as evidências bioestatísticas ou genéticas em determinada prova. É o emergir da estatística forense; talvez o mais recente tema do estudo outlier.

Vimos que, a noção de outlier depende da área da estatística onde estamos a trabalhar. Pelo exposto, não é possível encontrar uma "definição geral". As séries temporais, por exemplo, exigem o desdobramento entre outliers aditivos e outliers inovadores. Os dados espaciais, por sua vez, requerem uma generalização dos poucos resultados existentes para dados circulares e onde a influência da dimensão nos transporta até aos outliers multivariados. Aqui, somos então confrontados com a dificuldade acrescida da ordenação dos dados que tinha sido fundamental para a pesquisa de outliers univariados.

O problema outlier também se pode envolver com a problemática geral do ensino da estatística - do ponto de vista conceptual e do ponto de vista prático. Esta questão coloca-se desde logo quando a maior parte da formação prática dos estudantes é feita sobre exercícios académicos. Estamos conscientes que o estatístico também se forma pela prática

profissional. No entanto, é importante a sensibilização para os problemas concretos do ponto de vista do experimentador; principalmente nas matérias de "formação final" que, por exemplo, envolvam a modelação estatística.

Embora com diferentes graus de dificuldade, muitos campos de investigação estão abertos para o estudo de outliers em dados estatísticos. A opção que neste livro fizemos - por uma abordagem generalista - tem a vantagem óbvia de tornar mais vasto o campo das possíveis aplicações mas, por outro lado, limitou o aprofundamento de alguns temas fundamentais. De entre estes - como pontualmente foi referido no texto - salientamos os estudos em séries temporais e em sondagens e censos, onde os primeiros desenvolvimentos são muito recentes. É claro que, podemos afirmar, a metodologia geral por nós introduzida também nestes domínios se pode aplicar; embora com as especificidades que daí sucederem.

A existência de um outlier é sempre relativa a um determinado modelo e uma observação pode ser discordante em relação a um modelo e não ser para outro.

O grande objectivo em qualquer estudo de outliers será sempre: O *que é um outlier e como tratar com essa observação.*

Definida uma teoria, tal como também referimos, é muito importante que se avalie o desempenho dos diversos testes de discordância. Este é também um domínio onde há muito trabalho para realizar.

O estudo de Beckman e Cook [16] - embora com mais de 20 anos - fez uma excelente síntese sobre o tratamento estatístico de outliers, quer do ponto de vista histórico quer das aplicações aos modelos padrão da estatística. Talvez seja o momento para uma actualização e, com a mesma intenção, fazer um novo ponto da situação. Nesse estudo, como já anteriormente referimos, ironicamente, Beckman e Cook concluíam que: "*Although much has been written, the notion of an outlier seems as vague today as it was 200 years ago*".

O que diremos hoje?

É claro que, desde aquela data, algo se avançou, mas muito há para fazer. Em 1974, Lindley predisse que o século XXI será bayesiano - sendo 2020 um ano crucial. Os métodos bayesianos são complicados especialmente para a teoria dos outliers onde como vimos, *a priori*, é (sempre) envolvida muita subjectividade. Haverá também aqui um grande tema de investigação?

Simbolicamente, como contrabalanço em 1998, Efron [41] prevê que "o velho Fisher terá um século XXI muito bom".

O mundo da estatística aplicada requer um compromisso entre os modos de pensar bayesiano e frequencista e, por agora, não há substituto para a síntese fisheriana. É interessante registar as ideias sobre as funções verosimilhança modificadas ou as pseudo-verosimilhanças, isto é, funções de parte ou todos os dados e parte ou todos os parâmetros que num sentido lato podem ser tratadas como verosimilhanças genuínas.

Como se cruzam todas estas questões com o estudo estatístico de outliers?

No estudo referido, onde fica este tema no triângulo estatístico de Efron ?

Este é um desafio científico para o futuro. Possivelmente, com dificuldade acrescida pelo desconhecimento do número de outliers em qualquer amostra.

Variados temas estão necessitados de maior avanço, como sejam: as causas (determinísticas e estatísticas) da presença de outliers e o problema da sua existência em modelos estruturados (univariados e multivariados); as diferenças entre outliers simples e outliers múltiplos.

Por sua vez, os diferentes objectivos que nos propomos ao estudar outliers numa amostra condicionam as conclusões. A síntese do trabalho efectuado será diversa se quisermos abordar apenas a detecção de outliers num conjunto de dados ou se pretendermos conjugá-la com modelos estatisticamente mais complexos, englobando por exemplo a presença de observações influentes. Aqui estaremos envolvendo problemas de robustez que muito se cruzam com o estudo de outliers mas que dele são distintos. Nesta perspectiva, não se fica muito longe da teoria dos valores extremos.

A teoria geral de outliers em dados estatísticos, em diversas direcções, avançou muito nos últimos 30 anos e, nela, uma grande parte dos desafios iniciais foi encontrando as contribuições que a tornaram uma área do saber já implantada como campo de investigação. Atingida essa fase deverá prosseguir com desenvolvimentos abrangentes nos domínios já explorados - o campo multivariado será um deles - à medida que também surgem novos temas; e, entre estes, o mais importante parece ser a avaliação do desempenho. De facto, a análise estatística de dados multivariados requer o nosso trabalho em duas vertentes principais - os testes e os modelos de discordância. E nesta temática é importante produzir novas ideias pois a estrutura complexa destes dados é inimiga da simplicidade científica exigida para obter o maior sucesso, principalmente nas aplicações.

No futuro, cada vez mais, os "outliers" continuarão a ocupar um lugar do centro na ciência estatística e nos métodos estatísticos, pois sempre uma observação discordante será um desafio para o analista e dela poderá depender o seu relatório final para a mais importante tomada de decisão. É de excelência que estamos a falar!

Mas, quando tudo está dito e feito, o principal problema no estudo de observações eventualmente suspeitas, continua a ser aquele que desafiou os primeiros investigadores - *O que é um "outlier" e como se deve trabalhar com essa observação?*

No final do segundo milénio, a revista Time organizou uma lista de personalidades de referência dos últimos mil anos. Os nomes foram ordenados através de uma votação. Como "personagem do milénio", no primeiro lugar, ficou S. Francisco de Assis seguido de Gutenberg, Cristóvão Colombo, Miguel Ângelo, Martinho Lutero, Galileu, Shakespeare, Thomas Jefferson, Mozart e, em décimo lugar, Einstein.

Um vencedor acumula valores que lhe conferem a distinção. Ora, com o objectivo de eleger a personalidade do milénio, os votantes terão elaborado os seus próprios critérios. Estes, emergindo de um conjunto de postulados permitiram gerar um primeiro lugar. S. Francisco é sempre indicado como referência e modelo de vida simples. Fala-se, muitas vezes, em "pobreza franciscana". O seu nome está também ligado à "ecologia" e à "paz". Quais terão sido, como se poderão descobrir, as variáveis mais importantes que fizeram eleger S. Francisco?

O conhecimento das componentes estatísticas que permitem encontrar (e definir) um valor discordante numa amostra é também um tema para a teoria dos outliers.

Em todos os modelos, qualquer que seja o critério de discordância, ficar em primeiro lugar é ser outlier!

Perante os exigentes temas sobre outliers em dados estatísticos acima descritos - e, parafraseando este "último outlier" - "ao menos comecemos a trabalhar, porque até agora pouco fizemos".

Referências

[1] Abramowitz, M. e Stegun, I. (1972) *Handbook of Mathematical Functions*.

[2] Anscombe, F. J. (1960). Rejection of Outliers. *Technometrics* **2**: 123-47.

[3] Bain, L. J. (1991). *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker, Inc. New York.

[4] Barnett, V. (1976).The Ordering of Multivariate Data (with discussion). *Journal of Royal Statistical Society* **A**: 318-354.

[5] Barnett, V. (1978). The Study of Outliers: Purpose and Model. *Applied Statistics* **27**: 242-50.

[6] Barnett, V. (Ed.) (1981). *Interpreting Multivariate Data*. Wiley.

[7] Barnett, V. (1983). Discussão do artigo "Outlier.....s" de Beckman e Cook. *Technometrics* **25**: 150-52.

[8] Barnett, V. (1988). Outliers and Order Statistics. *Commun. Statist. Theory and Meth.* **17**: 2109-18.

[9] Barnett, V. (2004). *Environmental Statistics. Methods and Applications*. Wiley.

[10] Barnett, V. e Lewis, T. (1967). A Study of Low-temperature Probabilities in the context of an Industrial Problem (with Discussion). *J. R. Statist. Soc.* **130**: 177-206.

[11] Barnett, V. e Lewis, T. (1994). *Outliers in Statistical Data*. 3ª Edição. Wiley.

[12] Basu, A. P. (1965). On some Tests of Hypotheses relating to the Exponential Distribution when some Outliers are present.
Jour. Amer. Stat. Assoc. **60**. 548-59.

[13] Basu, A. P. (1971). Bivariate Failure Rate.
Jour. Amer. Stat. Assoc. **66** : 103-5.

[14] Basu, A. P. (1998). Some Bivariate Distributions useful in Reliability Theory. *Frontiers in Probability and Statistics*. Ed. por S.P. Mukherjee *et al*. Narosa Publishing House: 35-40.

[15] Basu, A. P. e Singh, B. (1998). Order Statistics in Exponential Distribution. *Handbook of Statistics*, ed. por C.R. Rao and N. Balakrishnan, 1-23. North-Holland.

[16] Beckman, R.J. e Cook, R.D. (1983). Outlier.....s. *Technometrics* **25**: 119-63.

[17] Becker, C. (2003) Outliers and Robustness when Analysing Structured Data. *Literacia e Estatística* (Brito *et al* editores). *Actas do X Congresso Anual da Sociedade Portuguesa de Estatística*: 3-18. Edições SPE.

[18] Behnken, D. W. e Drapper, N. R. (1972). Residuals and their Variance Patterns. *Technometrics* **14**: 101-11.

[19] Branco, J. (2005). Estatística Robusta: Contribuição Portuguesa. *Memorial da Sociedade Portuguesa de Estatística* (Rosado, editor): 73-90. Edições SPE.

[20] Braumann, M. M. (1989). *Testes de Discordância para Outliers em Populações Normais e Gama*. Trabalho de síntese apresentado nas Provas de Aptidão Pedagógica e Capacidade Científica. Universidade de Évora.

[21] Braumann, M. M. (1994). *Sobre Testes de Detecção de "Outliers" em Populações Exponenciais*. Dissertação de Doutoramento. Universidade de Évora.

[22] Braumann, M. M. (1997). Medidas de Performance para Testes de Detecção de Outliers em Populações Exponenciais com Parâmetros Conhecidos. *Estatística: a diversidade na unidade* (Miranda *et al* editores). *Actas do X Congresso Anual da Sociedade Portuguesa de Estatística*: 183-97. Edições Salamandra.

[23] Braumann, M. M. (1999). Efeito do Uso de Hipóteses Gerais sobre as Medidas de Performance para Testes de Detecção de Outliers em Populações Exponenciais com Parâmetros Conhecidos. *Afirmar a Estatística - Um desafio para o século XXI* (Paulino *et al* editores). *Actas do VI Congresso Anual da Sociedade Portuguesa de Estatística*: 157-70.

[24] Calvin, M., Heidelberger, C., Reid, J. C., Tolbert, B. M. e Yankwich, P. F. (1949) *Isotopic Carbon: Techniques in its Measurement and Chemical Manipulation*. Wiley. Reeditado em 1960.

[25] Chauvenet, W. (1863). Method of Least Squares. Apêndice do *Manual of Spherical and Pratical Astronomy* **2**: 469-566. Reeditado em 1960. Dover.

[26] Cohen, A., Tiago de Oliveira, J. e Haim, M. (1972). Loccally Optimal Quasi-Linear Tests for Outliers. *Scientific Repport. The Technion*. Haifa. Israel.

[27] Collett, D. (1980). Outliers in Circular Data. *Applied Statistics* **29**: 50-7.

[28] Collett, D. e Lewis, T. (1976). The Subjective Nature of Outlier Rejection Procedures. *Applied Statistics* **25**: 228-37.

[29] Cook, R. D. e Weisberg, S. (1982) *Residuals and Influence in Regression*. Chapman and Hall.

[30] Costa, S. (2005). *Análise Estatística Multivariada na Segmentação de uma Companhia de Seguros*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

[31] Daniel, C. (1960). Locating Outliers in Factorial Experiments. *Technometrics* **1**: 149-56.

[32] Darling, D.A. (1952). On a Test for Homogeneity and Extremes Values. *Ann. Math. Stat* **23**: 450-6.

[33] David, H. A.(1981). *Order Statistics.* Wiley.

[34] David, H. A. e Paulson, A. S. (1965). The Performance of Several Tests for Outliers. *Biometrika*. **52**: 429-36.

[35] Dixon, W. J. (1950). Analysis of Extreme Values. *Ann. Math. Stat.* **21**: 488-506.

[36] Dixon, W. J. (1953). Processing Data for Outliers. *Biometrika* **9**: 74-89.

[37] Dixon, W. J. (1962). Rejection of Outliers. *Contributions to Order Statistics*. Sarhan and Greenberg (Ed.) [128].

[38] Dixon, W. J. (1964). Query 4: Rejection of Outlying Observations. *Technometrics* **6**: 238.

[39] Draper, N. R. (1983). Discussão do artigo "Outlier.....s" de Beckman e Cook. *Technometrics* **25**: 159-60.

[40] Draper, N. R. e Smith, H. (1998). *Applied Regression Analysis* (3ª edição). Wiley.

[41] Efron, B. (1998). R. A. Fisher in the 21st Century. *Statistical Science* **13**: 95-122.

[42] Epstein, B. (1960). Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential: part I. *Technometrics* **2**: 83-101.

[43] Epstein, B. (1960). Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential: part II. *Technometrics* **2**: 167-83.

[44] Ferguson, T. S. (1961). On the Rejection of Outliers. *Proceedings of the* 4^{th} *Berkeley Symposium* **1**: 253-87.

[45] Ferguson, T. S. (1961). Rules for the Rejection of Outliers. *Revue Inst. Int. Statistique* **29**: 29-43.

[46] Fieller, N. (1976). *Some Problems Related to the Rejection of Outlying Observations*. Tese de Doutoramento. Universidade de Sheffield.

[47] Figueira, M.M. (1995). *Identificação de Outliers: uma Aplicação ao Conjunto das Maiores Empresas com Actividade em Portugal*. Dissertação de Mestrado. Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão.

[48] Finney, D. J. (1974). Problems, Data and Inference: The Address of the President (with Proceedings). *J. R. Stat. Soc.* **137**: 1-23.

[49] Fisher, R. (1929). Tests of Significance in Harmonic Analysis. *Proc. Royal Soc. Edinburg, sect A* **125**: 54-9.

[50] Gander, W. e Gautschi W. (2000). "Adaptative Quadrature - revisited". *BIT*, **40**: 84-101.

[51] Gnanadesikan, R. e Kettenring, J.R. (1972). Robust Estimates, Residuals and Outlier Detection with Multiresponse data. *Biometrics* **28**: 81-124.

[52] Gnanadesikan, R. (1997). *Methods for Statistical Analysis of Multivariate Observations*, 2ª edição. Wiley.

[53] Goldstein, M. (1983). Outlier Resistant Distributions: Where Does the Probability Go? *Journal Royal Stat. Society* **45**: 355-7.

[54] Green, R. F. (1974). A Note on Outlier-Prone Families of Distributions. *Annals of Statistics* **2**: 1293-5.

[55] Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *Ann. Math. Stat.* **21**: 27-58.

[56] Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**: 1-21.

[57] Gumbel, E. J. (1960). Discussão dos artigos de Anscombe [2] e Daniel [31]. *Technometrics* **2**: 165-6.

[58] Guttman, I. e Smith, D. E. (1969). Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution. I: Estimation of the mean. *Technometrics* **11**, 527-50.

[59] Hampel, F. R. (1968). *Contributions to the Theory of Robustness*. Tese de Doutoramento. Universidade da California. Berkeley.

[60] Hampel, F. R. (1973). Robust Estimation: a Condensed Partial Survey. *Z. Wahr. Verw. Geb.* **27**: 87-104.

[61] Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *J. Americ. Stat. Assoc.* **69**: 383-93.

[62] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. e Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

[63] Harter, H. L. (1974-76). The Method of Least Squares and Some Alternatives. Parts I-VI. *Revue Institute International de Statistique* **42**: 147-174; 235-264; **43** 1-44; 125-190 e 269-278; **44**: 113-159.

[64] Harter, H. L. e Balakrishnan, N. (1998). Order Statistics: A Historical Prespective. *Handbook of Statistics*, vol. 16. Elsiever Science.

[65] Hawkins, D.M. (1974). The Detection of Errors in Multivariate Data using Principal Components. *Journal Americ. Statist. Association* **69**: 340-4.

[66] Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall. London.

[67] Hawkins, D. M. (1983). Discussão do artigo "Outlier.....s" de Beckman e Cook. *Technometrics* **25**: 155-6.

[68] Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley.

[69] Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (1992). *Análise Exploratória de Dados. Técnicas Robustas*. Edições Salamandra.

[70] Hogg, R. V. (1983). Discussão do artigo "Outlier.....s" de Beckman e Cook. *Technometrics* **25**: 158-9.

[71] Huber, P. J. (1972). Robust Statistics: a Review. (The 1972 Wald Lecture). *Ann. Math. Stat.* **43**: 1041-67.

[72] Joshi, P. C. (1972). Efficient Estimation of the Mean of an Exponential Distribution when an Outlier is Present. *Technometrics* **14**: 137-43.

[73] Kale, B. K. e Sinha, S. K (1971). Estimation of Expected Life in the Presence of an Outlier Observation. *Technometrics* **13**: 755-9.

[74] Kendall, M. G. (1966). Discrimination and Classification. *in* Krishnaiah [78] volume I: 165-84.

[75] Kimber, A. C. (1979). Tests for a Single Outlier in a Gamma Sample with unknown Shape and Scale Parameters. *Applied Statistics* **28**: 243-50.

[76] Kimber, A. C. e Stevens, H. J. (1981). The Null Distribution of a Test for Two Upper Outliers in an Exponential Sample. *Applied Statistics* **30**: 153-7.

[77] Kotz, S. e Johnson, N. L. (Ed.) (1989). *Encyclopedia of Statistical Sciences*. Wiley.

[78] Krishnaiah, P. R. (Ed.) (1966). *Multivariate Analysis*. Volume I. Academic Press.

[79] Lewis, T. e Fieller, N. R. J. (1979). A Recursive Algorithm for Null Distribution for Outliers: I. Gamma Samples. *Technometrics* **21**: 371-6.

[80] Mc Culloch, C. E. e Meeter, D. (1983). Discussão do artigo "Outlier.....s" de Beckman e Cook. *Technometrics* **25**: 152-5.

[81] Mc Millan, R. G. (1971). Tests for One or Two Outliers in Normal Samples with unknown Variance. *Technometrics* **13**: 87-100.

[82] Mc Millan, R. G. e David, H. A. (1971). Tests for One or Two Outliers in Normal Samples with known Variance. *Technometrics***13**: 75-85.

[83] Maddala, G. S. e Rao, C. R. (Ed.) (1997). *Handbook of Statistics*. Volume 15 - Robust Inference. Elsevier.

[84] Main, P. (1988). Asymptotic Behaviour of Reliability Functions. *Stat. and Prob. Letters* **7**: 259-63.

[85] Mardia, K. V. (1975). Statistics of Directional Data (with discussion). *J. Royal Stat. Society.* **37**: 349-93.

[86] Mardia, K. V. e Jupp P. E. (2000). *Directional Statistics*. Wiley.

[87] Martins, S. (2000). *Medidas de "Performance" em Modelos de Discordância Exponenciais*. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Ciências.

[88] Millard, S. P. e Nagaraj, K. N. (2000). *Environmental Statistics*. CRC Press.

[89] Muñoz-Garcia, J., Moreno-Rebollo, J. L. e Pascual-Acosta, A. (1990). Outliers: A Formal Approach. *International Statistical Review* **58**: 215-26.

[90] Murteira, B. J. (1988). *Estatística: Inferência e Decisão*. Imprensa Nacional-Casa da Moeda.

[91] Murteira, B. J. (1993). *Análise Exploratória de Dados. Estatística Descritiva*. Mc Graw-Hill.

[92] Murteira, B., Ribeiro, C. S., Silva, J. A. e Pimenta, C. (2002). *Introdução à Estatística*. Mc Graw Hill.

[93] Neyman, J. e Scott, E. L. (1971). Outlier Proneness of Phenomena and of Related Distribution. In Rustagi, J. (Eds.) (1971). *Optimising Method in Statistics*. Academica Press.

[94] Oliveira, P. (1988). *Tratamento Estatístico de "Outliers"*. Provas de Aptidão Pedagógica e Capacidade Científica. Universidade do Minho.

[95] Palma, J. (2006). *Medidas de Desempenho para Testes de Discordância em Populações Normais*. Tese de Doutoramento. Universidade de Lisboa.

[96] Passos, J. (1992). *Influência das Observações nos Coeficientes Estimados no Modelo de Regressão Múltipla*. Dissertação de Mestrado. Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão.

[97] Pearson, E. e Hartley, H. (1970). *Biometrika tables for statisticians*. Cambridge University Press.

[98] Peirce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal* **2**: 161-3.

[99] Pereira, C. M. e Pires, A. M. (2004). Método de Classificação com Rejeição por Indecisão e Observações Atípicas. *Estatística com Acaso e Necessidade* (Rodrigues *et al* editores). *Actas do XI Congresso Anual da Sociedade Portuguesa de Estatística*: 595-603. Edições SPE.

[100] Pestana, D., Turkman, M. A., Branco, J., Duarte, L. e Pires, A. (Eds.) (1994). *A Estatística e o Futuro e o Futuro da Estatística. Actas do I Congresso Anual da Sociedade Portuguesa de Estatística*. Edições Salamadra.

[101] Pestana, D. e Velosa, S. (2006). *Introdução à Probabilidade e à Estatística*. Volume I. 2ª edição. Fundação Calouste Gulbenkian.

[102] Pires, A. M. (1990). *Estimação Robusta e sua Aplicação a Componentes Principais*. Tese de Mestrado. Instituto Superior Técnico.

[103] Pires, A. M. e Branco, J. A. (1994). Estatística Robusta: Passado, Presente e Futuro. *A Estatística e o Futuro e o Futuro da Estatística* (Pestana *et al* editores). *Actas do I Congresso Anual da Sociedade Portuguesa de Estatística*: 531-49. Edições Salamandra.

[104] Prescott, P. (1979) Critical Values for a Sequencial Test for Many Outliers. *Applied Statistics,* **28**: 36-9.

[105] Prescott, P. (1983). Discussão do artigo "Outlier.....s" de Beckman e Cook. *Technometrics* **25**: 156-7.

[106] Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankya* **A26**: 329-58.

[107] Rao, C. R. e Székely, G. J. (2000). *Statistics for the 21st Century. Methodologies for Applications of the Future*. Marcel Dekker.

[108] Rao, C. R. Wegman, E. J. e Solka, J. L. (Ed.) (2005). *Handbook of Statistics*. Volume 24 - Data Mining and Data Visualization. Elsevier.

[109] Rosado, F. (1982). Análise Qualitativa de Densidades de Gram-Charlier e de Edgeworth como Modelos de Alternativas Inerentes para Outliers. *Nota 27/82 do CEAUL* - Centro de Estatística e Aplicações da Universidade de Lisboa.

[110] Rosado, F. (1982). Distribuições Assintóticas sobre Testes de Discordância para Outliers em Populações Exponenciais. *Nota 31/82 do CEAUL* - Centro de Estatística e Aplicações da Universidade de Lisboa.

[111] Rosado, F. (1982). Testes de Discordância para Outliers em Distribuições Exponenciais - Resultados Assintóticos. *Actas ds IX Jornadas Hispano - Lusas de Matemática*, vol. II: 623-6.

[112] Rosado, F. (1984). *Existência e Detecção de Outliers - Uma Abordagem Metodológica*. Tese de Doutoramento. Universidade de Lisboa.

[113] Rosado, F. (1984). The Null Distribution Function of Discordancy Tests for Outliers in Exponential Populations. *METRON, Rivista Internazionale di Statistica*, vol. XLII, nº 1-2: 51-7.

[114] Rosado, F. (1987). Outliers in Exponential Populations. *METRON, Rivista Internazionale di Statistica*, vol. XLV, n.1-2: 85-91.

[115] Rosado, F. (1987). Algumas Reflexões sobre a Condição Outlier. *Actas das XII Jornadas Luso-Espanholas de Matemática*, vol. III: 175-80.

[116] Rosado, F. (1990). Outliers, Inliers e Observações Influentes. *Actas das XV Jornadas Luso-Espanholas de Matemática*, vol. IV: 227-9.

[117] Rosado, F. (1996). Testes de Discordância para Outliers e sua Dependência da Dimensão da Amostra. *A Estatística a Decifrar o Mundo* (Vasconcelos *et al* editores). *Actas do IV Congresso Anual da Sociedade Portuguesa de Estatística*: 173-81. Edições Salamandra.

[118] Rosado, F. (2001). Outliers em Dados Estatísticos - o passado e o presente. E o futuro? *Um Olhar sobre a Estatística* (Oliveira *et al* editores). *Actas do VII Congresso Anual da Sociedade Portuguesa de Estatística*: 90-110. Edições SPE.

[119] Rosado, F. (2005). "Outliers" em Português. *Memorial da Sociedade Portuguesa de Estatística* (Rosado, editor): 139-51. Edições SPE.

[120] Rosado, F. (Ed.) (2005). *Memorial da Sociedade Portuguesa de Estatística*. Edições SPE.

[121] Rosado, F. e Alpiarça, I. (1994). Sobre Modelos de Discordância para Outliers em Populações Exponenciais e Gama. *Actas do II Congresso Anual da Sociedade Portuguesa de Estatística* (Lopes *et al* editores): 67-71.

[122] Rosado, F. e Braumann, M. M. (1990). Critical Values for a Lower Outlier in a Gamma Sample. *METRON, Rivista Internazionale di Statistica*, vol.XLVIII, n.1-4: 19-25.

[123] Rosado, F. e Mendes, Z. (1994). Sobre a Detecção de Observações Discordantes em Populações Normais. *A Estatística e o Futuro e o Futuro da Estatística*. (Pestana *et al* editores). *Actas do I Congresso Anual da Sociedade Portuguesa de Estatística*: 271-86. Edições Salamandra.

[124] Rosado, F. e Oliveira, I. (1996). Selecção Multivariada de Outliers - Uma Aplicação às Variedades de Castanheiro em Trás-os-Montes. *A Estatística a Decifrar o Mundo*. (Vasconcelos *et al* editores). *Actas do IV Congresso Anual da Sociedade Portuguesa de Estatística*: 163-73. Edições Salamandra.

[125] Rosado, F. e Palma, J. (2001). Measures of Performance for Discordancy Tests in Normal Populations. *Revista de Estatística.* Contributed Papers, vol. II, 2º Quad: 357-8.

[126] Rosado, F. e Palma, J. (2001). Sobre a Qualidade de Testes de Discordância em Populações Normais. *Nota CEAUL 12/2001* do Centro de Estatística e Aplicações da Universidade de Lisboa.

[127] Rosado, F. e Palma, J. (2003). Problemas e Limitações da Detecção de Outliers. *Nota CEAUL 2/2003* do Centro de Estatística e Aplicações da Universidade de Lisboa.

[128] Sarhan, A. E., e Greenberg, B. G. (Ed.) (1962). *Contributions to Order Statistics*. Wiley.

[129] Schwager, S. J. e Margolin, B. H. (1982). Detection of Multivariate Normal Outliers. *Annals of Statistics* **10**: 943-54.

[130] Shapiro, S. S. e Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**: 591-611.

[131] Shapiro, S. S. e Wilk, M. B. (1972). An Analysis of Variance Test for Exponential Distribution (Complete Samples). *Technometrics* **14**: 355-70.

[132] Srivastava, M. S. (2002). *Methods of Multivariate Observations*. Wiley.

[133] Tiejten, G. L., Moore, R. H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics***14**: 583-97.

[134] Tukey, J. W. (1977). *Exploratory Data Analysis.* Adison-Wesley Publishing Company.

[135] Veale, J. R. (1975). Improved Estimation of a Mean when One Observation may be Spurious. *Technometrics* **11**: 331-9.

[136] Velleman, P. F. e Hoaglin, D. C. (1981). *Applications Basics and Computing of Exploratory Data Analysis*. MA: Duxbury Press.

[137] Wilks, S. S. (1963). Multivariate Statistical Outliers. *Sankhia* **25**: 407-26.

Índice Remissivo

Agradecimentos

A Sociedade Portuguesa de Estatística agradece às seguintes entidades pelo valioso apoio concedido para a realização do XIV Congresso Anual:

British Council

Caixa Geral de Depósitos

Câmara Municipal da Covilhã

Câmara Municipal de Manteigas

CEAUL

Fonte da Fraga

Fundação Calouste Gulbenkian

Fundação para a Ciência e Tecnologia

Governo Civil de Castelo Branco

Hotel Tryp D.Maria

Instituto Nacional de Estatística

Livraria Escolar Editora

Lupim Design

Proengel

PSE

Região de Turismo da Serra da Estrela

Timberlake Consultores

Universidade da Beira Interior