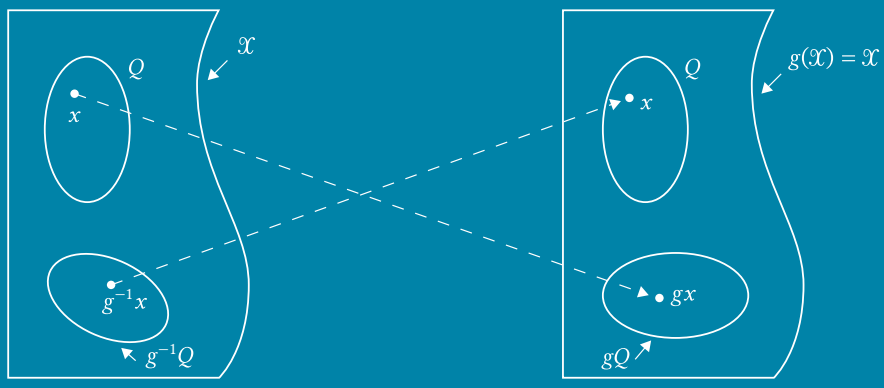
Bento José Ferreira Murteira

ESTATÍSTICA:
INFERÊNCIA E DECISÃO



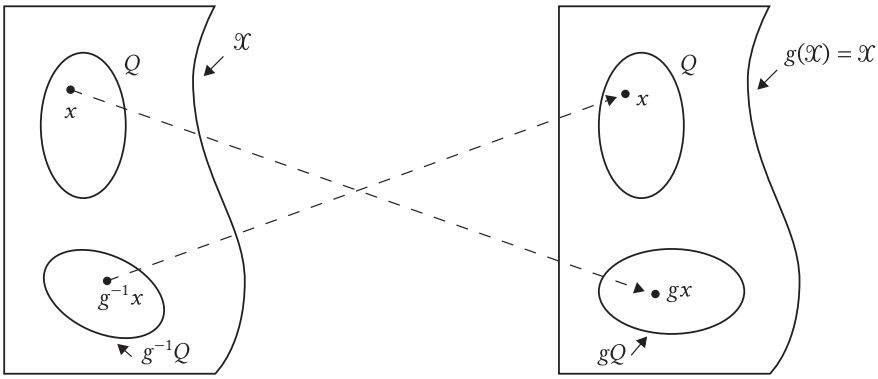
Edição comemorativa do centenário do Autor

SPE • CEMAPRE • CEAUL

ESTATÍSTICA:
INFERÊNCIA E DECISÃO

Bento José Ferreira Murteira

ESTATÍSTICA:
INFERÊNCIA E DECISÃO



Edição comemorativa do centenário do Autor

SPE • CEMAPRE • CEAUL

Estatística: Inferência e Decisão

Bento José Ferreira Murteira

Editor
Sociedade Portuguesa de Estatística
Bloco C6, Piso 4, sala 6.4.09
Campo Grande
1749-016 Lisboa

Capa, Design e Paginação
Typografia® | Atelier de Design de João Loureiro

Impressão
Instituto Nacional de Estatística

ISBN
978–972–8890–50–6

Depósito Legal
538273/24

ÍNDICE

A presente obra do Professor Bento Murteira, editada em 1988 pela Imprensa Nacional–Casa da Moeda, teve o patrocínio exclusivo do Instituto de Seguros de Portugal. O texto que se segue, é parte integrante da edição original, pelo que, por razões históricas, se considerou adequado manter nesta edição comemorativa do centenário.

O desenvolvimento económico do nosso País e a sua entrada em espaços mais vastos, como as Comunidades Europeias, têm conduzido, nas empresas ou noutras entidades, a uma aplicação mais frequente e generalizada de métodos científicos para a tomada de decisões.

Devem, por isso, ser apoiadas todas as iniciativas conducentes a uma maior divulgação das bases científicas dos métodos estatísticos a aplicar.

O Professor Bento Murteira vem-se dedicando, desde há largos anos, ao ensino e divulgação da Estatística, dos Métodos de Previsão e da Econometria.

Para além de numerosos artigos publicados, em revistas nacionais e estrangeiras, sobre estas matérias, publicou igualmente duas obras muito consultadas *Probabilidades e Estatística*, em dois volumes, e *Estatística Descritiva*.

O trabalho mais recente que ora se publica e que o Instituto de Seguros de Portugal tem o privilégio de patrocinar é precisamente sobre duas matérias centrais no campo da aplicação prática: a Inferência e a Decisão.

Na actividade seguradora, como aliás na maior parte das restantes actividades económicas, a tomada de decisões baseia-se, em regra, num conjunto de dados estatísticos que, devidamente tratados, permitem encontrar a melhor solução.

É, pois, com muito interesse que vemos a publicação desta obra. Pelo seu autor, tem a garantia do rigor e correcção a que nos habituou, e pela matéria abordada um interesse redobrado pela necessidade dum melhor conhecimento dos métodos de inferência estatística para a tomada de decisões.

Que todos saibam aproveitar os ensinamentos da presente obra, para cuja publicação tivemos a honra de contribuir.

Lisboa, 4 de Fevereiro de 1988
O Instituto de Seguros de Portugal

PREFÁCIO

O presente texto teve longa evolução. Começou por ser uma série de apontamentos em que se basearam algumas exposições feitas aos docentes de matemática do Instituto Superior de Economia no ano lectivo de 1974–75. Com sucessivos aperfeiçoamentos e extensões — efectuados em grande parte quando o Autor estava integrado no Centro de Estatística e Aplicações da Universidade de Lisboa, do INIC — serviu posteriormente de suporte às lições proferidas sobre decisão estatística no Curso de Mestrado em Métodos Matemáticos para Economia e Gestão de Empresas que regularmente se vem realizando no mesmo Instituto.

Foi grande o benefício retirado dos comentários e críticas feitos pelos participantes nos diferentes Cursos e de alguns colegas que tiveram a amabilidade de ler o manuscrito. Na impossibilidade de referir o rol dos que contribuíram para tornar o texto mais aceitável, citam-se, com uma palavra de reconhecimento extensível a todos, os colegas Maria Antónia do Amaral Turkman e Gustavo de Castro. Evidentemente, a responsabilidade por erros ou deficiências resta inteiramente com o Autor, aliás consciente das limitações que não conseguiu ultrapassar.

* * *

A publicação da Estatística: Inferência e Decisão tem como principal objectivo chamar a atenção para a existência de escolas ou correntes que afastando-se da Estatística Clássica não só abrem diferentes perspectivas e propõem métodos alternativos como permitem um melhor enquadramento e crítica dos métodos clássicos. A ênfase dada à lnferência Bayesiana e à Decisão Estatística, sobretudo a esta, visa, assim, contrabalançar o peso talvez excessivo que a Estatística Clássica tem no ensino.

Para manter a obra dentro de dimensões razoáveis consideram-se adquiridos os principais conceitos e métodos da Estatística Clássica, com particular relevo para estimação, intervalos e regiões de confiança, testes de significância e ensaios de hipóteses em problemas paramétricos (que são, pode dizer-se, os únicos tratados).

Esses conhecimentos correspondem, grosso modo, às matérias que podem estudar-se em Murteira (1980) ou, com maior desenvolvimento, em Mood, Graybill e Boes (1974) ou em Rohatgi (1976).

* * *

As referências contidas na obra não prestam justiça ao pesado débito que se tem para alguns autores. É agradável encargo mencionar Blackwell e Girshick (1954), Fraser (1957), Lehmann (1959 e 1983), Ferguson (1967), DeGroot (1970), Lindgren (1971), Berger (1980) e Barnett (1982), em que se inspiraram muitos dos resultados e exemplos apresentados. Infelizmente, Berger (1985) [segunda edição muito alargada de Berger (1980)], não saiu a tempo de lhe ser dada a atenção que merece: trata-se de leitura obrigatória para quem estuda Inferência Bayesiana e Decisão Estatística.

* * *

É também muito agradável manifestar apreço pelo acolhimento dispensado pela Imprensa Nacional-Casa da Moeda, E.P., sem o qual a publicação do texto não teria sido possível. Da solução encontrada não pode dissociar-se o apoio e encorajamento do colega J. Simões Lopes.

* * *

Finalmente, como prova de grande reconhecimento, dedica-se o presente trabalho a todos aqueles que pacientemente escutaram o Autor, estimulando-o com o seu interesse.

Lisboa, Outubro de 1987
Bento José Ferreira Murteira

PREFÁCIO À REEDIÇÃO DA OBRA, POR OCASIÃO DA CELEBRAÇÃO DO CENTENÁRIO DE BENTO MURTEIRA

É com grande entusiasmo que apresentamos a recuperação e reedição de *Estatística: Inferência e Decisão*, de Bento José Ferreira Murteira. Esta obra, que representa um marco na literatura estatística portuguesa, continua a ser de imensurável valor para a comunidade académica e científica. A reedição desta obra, no centenário do nascimento do autor, é uma iniciativa conjunta da Sociedade Portuguesa de Estatística (SPE) e do Instituto Superior de Economia e Gestão (ISEG), refletindo o compromisso de ambas as instituições em preservar o legado dos seus membros e em promover o desenvolvimento da Estatística em Portugal.

Bento Murteira, nascido a 17 de agosto de 1924 em Lisboa, foi um dos sócios fundadores da SPE e recebeu o Prémio Carreira SPE em 2013 pelo seu inestimável contributo ao ensino e à prática da Estatística. Licenciado em Finanças pelo ISCEF (atual ISEG), Bento Murteira doutorou-se em 1953 na Universidade Técnica de Lisboa com uma tese sobre processos auto-regressivos, uma área onde pioneiramente deixou a sua marca. O seu percurso académico e docente no ISEG, desde 1947 até à sua jubilação em 1994, como Professor Catedrático, foi fundamental na formação de várias gerações de estatísticos e na evolução da prática estatística em Portugal. A sua colaboração posterior com instituições como a Universidade Nova de Lisboa e a Universidade Autónoma de Lisboa continuou a enriquecer o panorama académico nacional até ao final da sua carreira.

Com uma vasta produção científica, Bento Murteira publicou numerosas obras de referência nas áreas de Estatística, Econometria e Investigação Operacional, algumas das quais em coautoria, tais como, *Probabilidades e Estatística*, *Introdução à Estatística*, *Análise Exploratória de Dados*, *Análise de Sucessões Cronológicas e Estatística Bayesiana*. Contudo, *Estatística: Inferência e Decisão* destaca-se pela profundidade da sua análise crítica das metodologias estatísticas e pelo enfoque

inovador nas questões de decisão estatística e seus fundamentos, sendo um texto que continua a alimentar debates na teoria e na prática da Estatística. Tal como Daniel Paulino referiu num dos seus escritos a respeito deste livro: «*Obra mais arrojada de Bento Murteira, pela profunda reflexão crítica sobre as metodologias estatísticas que constitui, foi no nosso meio um marco impulsionador de uma consciência crítica da teoria e prática estatística que, ainda hoje, deve ser leitura assídua e reflexiva de quem faz da Estatística a sua ocupação profissional*».

Esta reedição gratuita não apenas presta homenagem ao vasto legado de Bento Murteira, mas também visa colmatar uma lacuna significativa na literatura estatística em português de Portugal. Num contexto onde os materiais de estudo escritos em português são limitados, obras como esta são essenciais para fortalecer a formação académica e a investigação científica. A sua disponibilização gratuita demonstra o empenho da SPE e do ISEG em proporcionar recursos valiosos a todos os que, no presente e no futuro, pretendem aprofundar os seus conhecimentos em Probabilidade e Estatística.

Este gesto de tornar acessível uma obra desta relevância reforça a nossa missão de promover o avanço do conhecimento e de garantir que a sabedoria e o trabalho de figuras de destaque da Estatística como Bento Murteira permaneçam disponíveis para todos. Agradecemos a todos os que tornaram possível esta reedição e esperamos que a leitura de *Estatística: Inferência e Decisão* inspire e enriqueça novas gerações de estatísticos, tal como tem feito ao longo dos anos.

Lisboa, Outubro de 2024
Luís Meira Machado e Maria Antónia Amaral Turkman

A NOTE FROM PROFESSOR JAMES O. BERGER

Professor Bento Murteira was a towering figure in Portugal (and highly recognized in the wider world) in the sciences of statistics, forecasting, econometrics and decision making. This book, which he wrote in 1988, is a wonderful testament to his brilliance.

I have actually read most of the book; although I do not speak Portuguese, the important statistical ideas stand out, highlighted by the historical and philosophical discussions and the extensive examples. Elaborating on these latter topics:

1. Statistics has an astonishing history of over 250 years, weaving through a tapestry of the objective Bayesian statistics of Bayes and Laplace, that ruled the statistical landscape for 150 years; Fisherian and frequentist statistics (although they were highly oppositional) during the last 100 years; and then a huge Bayesian revival, occurring over the last 75 years. The book captures much of this history, although naturally focusing more on the years from 1920 on, since that is the most operationally important era. The scholarship in the book is amazing, essentially capturing all of the main Bayesian and frequentists ideas from 1920 to 1988.

2. I recently wrote an article "Learning Statistics Through Counterexamples"[1] for a volume in honor of Dev Basu who wrote a famous article with the same name[2]. I found that most of those examples were in the book. I am sent books all the time to review and I look for key things, such as these counterexamples, to see if the author is aware of the major issues and controversies in statistics; sadly, most authors do not understand at least some of these fundamentals. Professor Murteira did understand all of the fundamentals by the mid-80's (and probably long before).

The structure of the book is interesting. It starts with a massive Chapter 1, which more or less explains everything — frequentist, Bayesian and their

[1] Berger, J. (2024). Learning Statistics From Counterexamples. Sankhya A.

[2] Basu, D. (2011). Learning statistics from counter examples: ancillary statistics. *Selected Works of Debabrata Basu*, pages 391–397.

interactions — that most statisticians would need to know. The ensuing chapters delve into more technical details for those interested in serious frequentist methodology, such as minimaxity and invariance; and Bayesian methodology, such as the objective Bayesian and subjective Bayesian approaches. Professor Murteira is spot on in everything he says about either the frequentist or the Bayesian approaches. Even though these later chapters are more technical, there are still practical and philosophical gems sprinkled throughout.

I wish I had been able to interact with Professor Murteira; the depth of his understanding was phenomenal.

Durham, October, 2024
James Berger
Arts and Sciences Distinguished Professor Emeritus of
Statistics, Duke University

AGRADECIMENTOS

Comemorar o centenário do Nascimento do Professor Bento Murteira com uma reedição de um texto com a qualidade de *Estatística: Inferência e Decisão*, que se encontra esgotado há vários anos, é tarefa que, desde logo, nos entusiasmou. Como seria expectável, a sua realização contou com a colaboração e empenho de várias entidades e pessoas a quem não queremos deixar de agradecer.

Os primeiros agradecimentos dirigem-se, muito naturalmente, à Família do Professor Murteira e à Imprensa Nacional-Casa da Moeda (INCM) que autorizaram a reedição da obra. Neste âmbito um agradecimento especial é devido ao Dr Duarte Azinheira, administrador da INCM, e ao Dr Alípio Magalhães, nosso colega durante muitos anos no ISEG, que diligenciou os primeiros contactos com a INCM.

À Sociedade Portuguesa de Estatística (SPE), ao Centro de Matemática Aplicada à Previsão e Decisão Económica (CEMAPRE) e ao Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) um agradecimento muito especial, não só por terem apadrinhado o projeto desde a primeira hora, como pelo financiamento da edição do livro em formato aberto (ficheiro pdf a ser disponibilizado a toda a comunidade científica). Os financiamentos feitos pelos Centros de Investigação contaram com a generosa ajuda da Fundação para a Ciência e Tecnologia (FCT) no quadro dos projetos UIDB/05069/2020 e UIDB/00006/2020 a quem também se agradece.

Também se agradece ao Instituto Nacional de Estatística (INE) pela edição em papel dos 100 exemplares do livro que se destinam a ser distribuídos gratuitamente por várias entidades como forma de reforçar a divulgação desta obra de inegável relevo.

À Typografia — Atelier de Design de João Loureiro um agradecimento especial pelo cuidado posto na conceção da capa, design e paginação da obra em LaTeX.

Finalmente, um agradecimento muito particular ao professor James Berger por ter escrito uma breve nota introdutória a esta reedição do livro.

Maria Antónia Amaral Turkman
Alexandra Bugalho de Moura
João Andrade e Silva
Luís Meira Machado
Marília Antunes
Rui Paulo

CAPÍTULO 1

ENSAIO DE INFERÊNCIA ESTATÍSTICA COMPARADA

1.1 Inferência e decisão

Ao aprofundar um pouco o estudo da estatística, saindo naturalmente dos procedimentos clássicos, depara-se com grande número de correntes ou escolas. Sem falar na falta de unidade[1] dos chamados clássicos (Fisher para um lado, Neyman-Pearson para outro), o desfile é algo extenso: bayesianos (objectivos, subjectivos, ...), fiducialistas, verosimilhancistas, estruturalistas, ...

A diversidade não é inesperada! As conclusões ou informações retiradas dos dados estatísticos sobre parâmetros ou modelos enquadram-se na lógica indutiva e é bem sabido que a justificação da indução é um dos problemas mais controversos da filosofia.

Cada escola tem princípios e procedimentos próprios. Os princípios devem arrastar a validade das inferências correctas e nunca implicar a validade de qualquer inferência incorrecta. A respectiva análise conduz aos fundamentos da inferência estatística que Berger (1984) refere nos seguintes termos: «*Statistics needs a ‹foundation›, by which I mean a framework of analysis within which any statistical investigation can theoretically be planned, performed, and meaningfully evaluated. The words ‹any› and ‹theoretically› are key, in that the framework should apply to any situation but may only theoretically be implementable. Pratical difficulties or time limitations may prevent complete (or even partial) utilization of such framework, but the direction in which ‹truth› could be found would at least be known*».

[1] Alguns aspectos do confronto são abordados ao longo do presente capítulo.

Os fundamentos da inferência estatística constituem campo reservado à filosofia, onde, no entanto, vão fazer-se algumas breves e ligeiras incursões.

Correndo o risco de partir de uma simplificação demasiado drástica — com perigos que a seu tempo se apontam — consideram-se[2] apenas três correntes através de enquadramento que destaque as principais características e as ponha em confronto. Mais tarde dá-se maior ênfase à decisão estatística.

O ligeiro ensaio de inferência estatística comparada que vai seguir-se deve muito a Barnett (1982) e a Cox e Hinkley (1974), autores cuja consulta se recomenda para apuramento e desenvolvimento do estudo.

A distinção entre as três correntes pode esboçar-se segundo diversas vertentes. Consideram-se, por agora, duas: objectivos e informação relevante.

Ao nível dos objectivos os procedimentos estatísticos podem visar a inferência — descrição em termos probabilísticos do conjunto de fenómenos observados de modo a desenvolver o conhecimento científico — ou podem visar a decisão — prescrição de modos de acção prática no contexto de uma dada situação através do processamento de informação adequada.

Exemplo 1.1 — Os caudais numa secção de um rio podem observar-se estatisticamente para descrever a sua distribuição e melhorar o conhecimento do respectivo regime hidrológico. Alternativamente, a distribuição pode descrever-se com o propósito de informar a decisão sobre as características de um dique a construir para protecção contra inundações. □[3]

A importância da inferência foi assinalada por Fisher (1956) ao tratar de conceitos estatísticos: «*The concepts... have arisen in the study of numerical observations in the Natural Sciences; they are intended for use in inferences by which progress in the sciences is guided*». Jeffreys (1961) alinha claramente com Fisher: «*The fundamental problem of scientific progress, and a fundamental one of everyday life, is that of learning from experience. Knowledge obtained in this way is partly merely description of what we already observed, but part consists of making inferences from past experiences to predict future experience...*». E exemplifica: «*The Nautical Almanac's predictions of lhe position of the planets, an engineer's estimation of the output of a new dynamo, and an agricultural statistician's advice to a farmer about the utility of a fertilizer are all inferences from past experience*».

Um dos primeiros autores a pôr em causa a distinção entre inferência e decisão foi provavelmente Neyman por volta de 1937. Em obra mais recente (Neyman, 1950) pode ler-se: «*Claims are occasionally made that mathematical statistics and the theory of probability form the basis of some mental process described as ‹inductive*

[2] Tanto quanto possível sem tomar posição.

[3] O símbolo, □, indica que o exemplo terminou.

reasoning›. However, in spite of substancial literature on this subject the term ‹inductive reasoning› remains obscure and it is uncertain whether or not the term can be conveniently used to denote any clearly defined concept. On the other hand... there seems to be room for the term ‹inductive behavior›. This may be used to denote the adjustment of our behavior to limited amounts of observation».

Lehmann (1959) também considera irrelevante separar a inferência da decisão: «*Frequently it is a question of providing a convenient summary of the data or indicating what information is available.... This information will be used for guidance in various considerations but will not provide the sole basis for any specific decision. In such cases the emphasis is on inference rather than on the decision aspect of the problem, although formally it can still be considered a decision problem if the inferential statement itself is interpreted as the decision to be taken*».

Se Neyman e Lehmann não fazem distinção nítida entre inferência e decisão, argumentando que ao fazer uma inferência se está a decidir, Cox (1958) mostra desacordo e justifica a sua posição com dois pontos: por um lado, um dos principais problemas da inferência estatística é escolher quais os tipos de conclusões a que pode chegar-se, enquanto na decisão estatística as acções possíveis são previamente especificadas; por outro lado, as decisões estatísticas baseiam-se, quer, na informação usada para estabelecer inferências, quer, na avaliação das percas decorrentes de decisões erradas.

Lindley (1965), ao defender que a decisão é uma extensão da inferência, situa-se, assim parece, numa situação intermédia: «*The person making the inference need not have any particular decision problem in mind. The scientist in his laboratory does not consider the decisions that may subsequently have to be made concerning his discoveries. His task is to describe accurately what is known about... the question*».

As opiniões recolhidas mostram bem que a classificação dos procedimentos estatísticos segundo os dois objectivos — inferência ou decisão — não é tão pacífica como à primeira vista se podia esperar. Como não parece muito frutuoso alongar a discussão, espera-se que o desenvolvimento do estudo possa contribuir para lançar mais alguma luz sobre a matéria.

Ao nível da informação relevante há que posicionar as várias correntes relativamente a:

- informação a priori;
- informação obtida por amostragem;
- informação sobre as consequências das acções alternativas.

A informação a priori, anterior ou externa em relação à amostra ou à experiência, deriva normalmente dos conhecimentos acumulados no domínio em causa ou em domínios conexos, bem como da intuição ou sensibilidade do investigador.

A designação dá cobertura a um vasto espectro de elementos, objectivos ou subjectivos. A incorporação da informação a priori nos modelos estatísticos passa por uma quantificação nem sempre desprovida de dificuldades.

A informação obtida por amostragem ou experimentalmente pode assumir a forma de dados ou observações resultantes de inquéritos conduzidos sobre uma parte — amostra — de uma população, de experiências repetidas em condições constantes ou aproximadamente constantes ou, ainda, de «realizações» de uma situação real — caso, por exemplo, das sucessões cronológicas — em que o processo gerador tem algo de estruturalmente estável.

A avaliação das consequências potenciais das várias acções alternativas é o terceiro tipo de informação relevante. Trata-se, muito frequentemente, de informação de natureza económica quantificada em termos de custos e benefícios.

Exemplo 1.2 — No caso do Ex. 1.1, a informação a priori pode consistir nos conhecimentos acumulados sobre o comportamento do mesmo rio noutras secções ou de rios com regime semelhante; a amostra pode ser constituída pela observação ou cálculo dos caudais diários durante um certo período; finalmente, se o propósito é construir um dique de protecção e a respectiva altura é a variável em relação à qual se tem de tomar uma decisão, a avaliação das consequências exige o cálculo, para várias alturas num dado intervalo — acções alternativas — dos custos do dique e dos benefícios retirados da construção do mesmo. □

A natureza dos objectivos e o tipo de informação formalmente utilizado permitem distinguir entre:

- inferência clássica (ortodoxa ou «standard»);
- inferência bayesiana;
- decisão estatística.

A inferência clássica, em que pontificam nomes como os de R. A. Fisher, J. Neyman e E. S. Pearson, é o processo de inferência que em princípio emprega exclusivamente a informação obtida por amostragem[4].

A inferência bayesiana, com que aparecem associados F. P. Ramsey, de Finetti, H. Jeffreys, I. J. Good e D. V. Lindley, é o processo de inferência que através do teorema de Bayes — conhecido desde 1763 — se propõe combinar a informação a priori com a informação obtida por amostragem.

[4] Barnett (1982) ao referir que o não acolhimento formal pelos clássicos da informação a priori ou da informação sobre as consequências não resultou de uma falha involuntária, mas, sim, de uma escolha consciente, declara: «*They argued that such information could seldom be objective or detailed enough to form part of a formal theory of statistical inference. Instead they aimed at producing a theory which would be universal in application, free from subjective assessments, and based on information which would always exist in quantifiable form, namely sample data*».

A decisão estatística, em que A. Wald foi pioneiro, é o processo que formula regras de acção com base nos três tipos de informação referidos. Na linha inicialmente desenvolvida por Wald (1950) não se recorre necessariamente à informação a priori; verificou-se depois, com o impulso de L. J. Savage, que a introdução de critérios de raiz bayesiana facilitava a investigação de regras de decisão óptimas.

*

A arrumação do pensamento estatístico em três escolas, se, por um lado, tem o atractivo de oferecer um quadro de referência, tem, por outro lado, o grave inconveniente de ser demasiado simplista. De facto, não existem fronteiras perfeitamente definidas entre as três escolas e dentro de cada uma há modos de interpretação, conceitos e procedimentos que não são sempre formulados ou conduzidos da mesma maneira. Há, portanto, que ficar de sobreaviso em relação ao excessivo esquematismo e à existência de escolas que não se integram em nenhuma das apontadas [o estudo de outras correntes pode fazer-se em Barnett (1982)].

*

Distinção algo mais profunda entre inferência clássica, inferência bayesiana e decisão estatística carece de uma chamada aos princípios de interpretação das conclusões a que chegam e, correlativamente, às interpretações do conceito de probabilidade que aceitam e empregam.

Considere-se em primeiro lugar o posicionamento de cada corrente face ao conceito de probabilidade.

Como é bem sabido não há grandes divergências[5] no que respeita ao sistema formal — pensa-se, especialmente, na teoria desenvolvida a partir da axiomática de Kolmogorov — em que a probabilidade surge como uma função matemática e é, pode dizer-se, um termo indefinido. Desejando aplicar-se o sistema formal a problemas reais é indispensável introduzir uma ligação ou conexão entre esse ente abstracto e a interpretação do termo probabilidade quando este se refere a acontecimentos ou proposições com valor prático. E neste ponto, isto é, quando se visa a determinação das propriedades extramatemáticas da probabilidade, a controvérsia é enorme.

Para encurtar ideias referem-se apenas três interpretações do conceito de probabilidade: a frequencista, a lógica e a personalista. Acompanha-se aqui Savage (1954) quando diz a propósito: «*...the short taxonomy... is bound to infuriate any expert on the foundations of probability, but I trust it may do the less earned more good than harm*».

[5] Excepto, possivelmente, no que concerne a axiomatização da aditividade ou da aditividade-σ.

O ponto de vista frequencista (Venn, von Mises, Reichenbach, Salmon, etc.) foi adoptado de forma quase unânime pelos estatísticos durante a primeira metade do século e é ainda hoje considerado correcto pela maioria. Baseia-se, grosso modo, na estabilidade ou regularidade estatística das frequências relativas — empiricamente verificada — e sustenta que a probabilidade de um acontecimento pode ser medida observando a frequência relativa do mesmo numa sucessão «numerosa» de provas ou experiências «idênticas e independentes».

O ponto de vista lógico (Keynes, Jeffreys, Carnap, etc.) defende que a probabilidade representa uma relação lógica entre uma proposição ou conjunto de proposições — a evidência — e outra proposição — a hipótese — e mede o grau de implicação (grau de confirmação para Carnap) da hipótese pela evidência. O grau de implicação é único, racional, impessoal.

O ponto de vista personalista ou subjectivo (Ramsey, de Finetti, Savage, etc.) considera que a probabilidade representa ainda uma relação entre a evidência e a hipótese, não uma relação lógica, mas uma relação quase-lógica. Por outras palavras, a probabilidade mede o grau de credibilidade que uma dada pessoa na posse da evidência atribui à hipótese. Este ponto de vista pressupõe que a pessoa em questão tem uma atitude razoável — veja-se adiante a ideia de coerência — o que não impede que duas pessoas diferentes face à mesma evidência tenham diferentes graus de credibilidade relativamente à mesma proposição.

No quadro artificialmente simplificado das interpretações do conceito de probabilidade [há mais correntes e várias linhas dentro das três correntes indicadas; veja-se o excelente estudo de Fine (1973)] pode avançar-se preliminarmente o seguinte:

- na inferência clássica perfilha-se a interpretação ou ponto de vista frequencista[6];
- na inferência bayesiana perfilha-se a interpretação personalista (bayesianos subjectivos) e também, talvez em menor escala, a interpretação lógica (bayesianos objectivos — veja-se secção 1.6).

Mantendo mais à superfície o confronto entre clássicos e bayesianos subjectivos não é difícil de compreender a diferença de atitudes em relação ao conceito de probabilidade. De facto, enquanto a informação obtida por amostragem ou por experimentação é compatível com a interpretação frequencista, pois pode conceber-se

[6] Há que distinguir entre o puro frequencismo de Neyman-Pearson e a posição de Fisher. Apesar de ser considerado frequencista, por exemplo por Savage, Fisher foi um veemente crítico do frequencismo tradicional acabando por perfilhar um conceito de probabilidade que o mesmo Savage considera pouco claro. Outros autores, como Efron, parecem considerar Fisher mais perto dos bayesianos do que dos frequencistas, na medida em que se afasta destes quando usa o argumento fiducial e propõe distribuições de probabilidade (fiduciária) para parâmetros condicionadas pela particular amostra observada (sem, no entanto, admitir distribuições a priori).

a repetição indefinida do processo de recolha de amostras ou de realização de experiências e analisar as frequências dos resultados a que conduz, tal não sucede com a informação inicial ou a priori. Esta, incluindo juízos, ou experiências individuais da mais diversa índole, decorrentes em geral de situações não repetitivas, somente consegue formalizar-se ou introduzir-se no modelo através da interpretação em termos de graus de credibilidade. É, aliás, a ideia de Berger (1984) quando declara: «*The goal of statistics is to communicate evidence about uncertainties and the correct language of uncertainty is probability. Only subjective probability provides a broad enough framework to encompass the types of uncertainties encountered...*».

É a informação por amostragem compatível com a interpretação subjectiva? É! De facto a interpretação subjectiva não impede que a informação dada pela amostra — quer dizer, dada pela observação de frequências — seja aceite e empregue no campo bayesiano. Importa notar, porém, que só em 1931 quando de Finetti elaborou o conceito de acontecimentos permutáveis [«*exchangeable events*»; veja-se Kyburg e Smokler (1964)] se estabeleceu a «ponte» entre os métodos clássicos e a noção de probabilidade personalista, isto é, a via que permitiu aos bayesianos acolher os resultados frequencistas que não colidiam com os seus princípios.

A decisão estatística não está vinculada a nenhuma interpretação particular do conceito de probabilidade; nos casos em que é utilizada informação a priori — perspectiva de Savage — é em geral aceite a interpretação personalista.

Voltando aos princípios de interpretação das conclusões começa por referir-se o princípio de amostragem repetida segundo o qual os métodos estatísticos devem apreciar-se através do respectivo comportamento num número ilimitado de repetições — hipotéticas — efectuadas nas mesmas condições. Uma das faces do princípio reside, precisamente, na interpretação frequencista já mencionada, isto é, na utilização de frequências como medidas de incerteza; a outra face reside na avaliação dos procedimentos estatísticos em termos da frequência com que fornecem respostas correctas ou bons resultados.

Evidentemente, é no princípio de amostragem repetida que repousa a inferência clássica. Por exemplo, o estimador de um parâmetro pode considerar-se bom quando tem uma probabilidade elevada de conduzir a estimativas na vizinhança do verdadeiro valor do parâmetro; isto é, se for hipoteticamente repetido grande número de vezes o processo de estimação, a frequência relativa de estimativas que caiem nessa vizinhança é elevada (e não difere muito daquela probabilidade).

Aspecto importante da inferência bayesiana é o já assinalado propósito de descrever toda a incerteza por meio de probabilidades, inclusivamente a que resulta da contingência da informação a priori. Para o efeito baseia-se nas seguintes hipóteses [Cox e Hinkley (1974)]: (1) cada indivíduo tem relativamente a todo o acontecimento incerto uma atitude susceptível de ser medida por uma probabilidade;

(2) as probabilidades que dizem respeito ao mesmo indivíduo são comparáveis; (3) as probabilidades podem avaliar-se em função do comportamento do indivíduo ao apostar em certos jogos hipotéticos.

O princípio fundamental em que assenta a inferência bayesiana — princípio de coerência bayesiana — exige que as probabilidades acima referidas (probabilidades personalistas) assegurem a cada indivíduo um comportamento na realização de apostas que não conduza a um prejuízo certo, quer dizer, um indivíduo coerente nunca permite que seja feita contra si aquilo a que se chama «banca holandeza».

Por exemplo, se um indivíduo atribuir a um dado acontecimento grau de credibilidade igual a 1 /4, isso quer dizer que está disposto a apostar de modo a ganhar 1$ se o acontecimento se não realizar e a perder 3$ se o acontecimento se realizar. Se o mesmo indivíduo atribuir idêntico grau de credibilidade ao acontecimento contrário, isso quer dizer que está disposto a fazer uma aposta também de 1 para 3 mas, agora, em condições inversas. Logo, o «banqueiro» que faça duas apostas de 1$, uma no acontecimento, outra no contrário, inflige ao indivíduo incoerente um prejuízo certo de 2$.

Uma consequência extremamente importante do princípio de coerência bayesiana, princípio intimamente ligado ao caracter normativo da teoria personalista, é tornar possível aplicar ao sistema de graus de credibilidade de um indivíduo coerente, em dado contexto, as regras de cálculo de probabilidades, isto é, o conjunto de teoremas emanado dos axiomas de Kolmogorov[7].

O problema da operacionalidade do conceito de probabilidade personalista por meio de sistemas de apostas em jogos artificiais levanta alguns problemas. Em particular, é pertinente afirmar que os graus de credibilidade podem depender da incerteza e do montante das apostas. De Finetti ladeia a questão sugerindo que se considerem apostas de reduzido valor monetário — «miniapostas». Uma abordagem mais satisfatória tem de fazer-se à luz da teoria da utilidade (veja-se capítulo 2).

A decisão estatística assenta também num princípio de coerência, desenvolvido com certo paralelismo em relação ao princípio bayesiano, segundo o qual, nos problemas de decisão, relativamente ao decisor, cada combinação de uma

[7] Atenda-se no que dizem Edwards, Lindman e Savage (1963): «*A system of personal probabilities, or prices for contingent benefits, is inconsistent if a person who acts in accordance with it can be trapped into accepting a combination of bets that assures him of a loss no matter what happens. Necessary and sufficient conditions for consistency are the following, which are familiar as a basis for the whole mathematical theory of probability:*

$$O \leq P(A) \leq P(S) = 1, P(A \cup B) = P(A) + P(B),$$

where S is the tautological, or universal, event; A and B are any two incompatible, or nonintersecting, events; and A ∪ B is the event that either A or B is true, or the union of A and B».

acção com um particular valor do parâmetro — estado da natureza — produz uma consequência que tem associada uma dada utilidade. Portanto, a acção óptima — decisão coerente — é aquela que conduz à maximização da utilidade esperada. O assunto vai ser oportunamente retomado e desenvolvido.

1.2 Inferência Estatística e Investigação Científica

Antes de prosseguir convém preparar o terreno para a introdução das estruturas matemáticas ou modelos donde partem, quer a inferência clássica, quer a inferência bayesiana, quer ainda a decisão estatística.

Escusado seria frisar que a estatística não é uma teoria; é um instrumento ou alfaia (Castro, 1952) cujas aplicações mais relevantes se situam, naturalmente, no domínio da investigação científica.

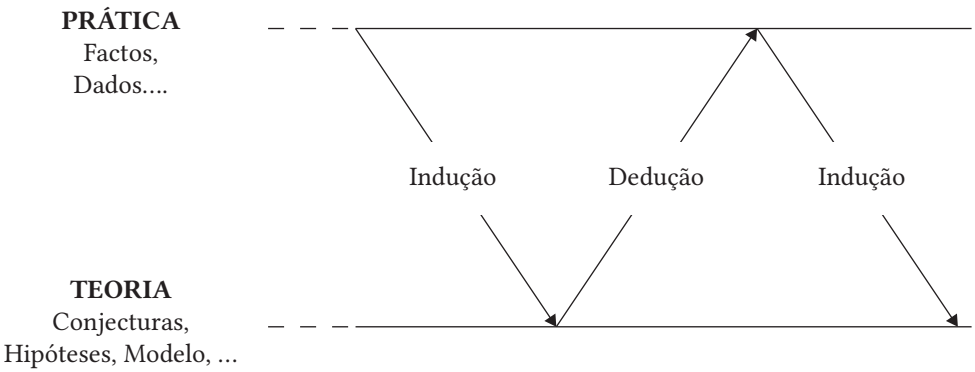
A investigação científica é um processo de aprendizagem essencialmente baseado na interacção teoria-prática e que, portanto, não se esgota com a mera especulação teórica ou com a estiolante acumulação de factos.

Box (1984) apresenta um exemplo típico da interacção teoria-prática que embora de interesse geral tem especial significado para os economistas: «*In 1927 Udny Yule was trying to understand what was wrong with William Beveridge's analysis of wheat price data. The fitting of sine waves of different frequencies by least squares had revealed significant oscillations at strange and inexplicable periods. Yule suggested that such series ought to be represented not by a deterministic function subject to error, but by a dynamic system (represented by a linear difference equation) responding to a series of random shocks — this model was likened to a pendulum being periodically hit by peas from a pea shooter*». Da ideia revolucionária de Yule nasceram os modelos mistos auto-regressivos e de médias móveis (modelos de Box-Jenkins) que ocupam hoje lugar central nos trabalhos de previsão a curto prazo.

A interacção teoria-prática encontra-se representada no esquema da página seguinte (Box, 1976).

Em dado ciclo, os factos conhecidos podem conduzir a ensaiar uma teoria. As deduções feitas a partir da teoria proposta podem revelar discrepâncias em relação a factos estabelecidos ou a dados especialmente recolhidos. Essas discrepâncias levam, plausivelmente, por indução, a uma teoria modificada ou a uma nova teoria. E assim por diante.

Um moderno ramo da estatística, a análise de dados (veja-se a exposição de Mallows e Tukey, 1982), referido de passagem por não se enquadrar no presente estudo, tem motivação na ideia de que actualmente o progresso da ciência exige que se atenda mais profundamente à aquisição, à qualidade e ao tratamento, análise e exploração dos dados.



Semelhante atitude está associada a uma diferente perspectivação do papel que cabe à estatística. Em vez de ser um auxiliar imparcial na avaliação da evidência que os dados fornecem em suporte de um dado modelo (ou a favor deste ou daquele valor particular dos parâmetros do mesmo modelo), hoje, por força da quantidade e da variedade dos dados acumulados e do poder de computação acessível, a estatística tem de empenhar-se muito mais na exploração, pesquisa e reconhecimento de padrões estáveis.

A posição da análise de dados é digna de nota. Por um lado, pede que na sequência, ... ⇒ dados ⇒ modelo ⇒ dados ⇒ ..., cada termo tenha a devida ponderação, isto é, adverte contra o perigo da teorização feita à margem das potencialidades que residem na exploração e tratamento dos dados; por outro lado, parece não pôr em dúvida a relevância dos modelos teóricos.

De facto, é lugar comum da metodologia das ciências que qualquer investigador que estuda um sistema — físico, biológico, económico, etc. — e é confrontado com uma dada informação factual, procura regra geral construir ou adaptar um modelo em termos do qual as características do sistema possam exprimir-se com simplicidade. O processo é dinâmico, quer dizer, o investigador bem sucedido vai elaborando modelos cada vez mais úteis ou mais válidos em função da atitude crítica tomada à luz da observação ou informação progressivamente acumulada.

Em tal contexto podem distinguir-se, em termos gerais, dois tipos de investigação estatística — dedutivo e indutivo — requeridos na relevante tarefa de apoio à investigação científica:

- o dedutivo, apropriado para realizar inferências condicionadas pela validade do modelo;
- o indutivo, necessário na fase de crítica do modelo («*diagnostic checking*»).

Mais desenvolvidamente tem-se: A inferência estatística associada com a incorporação de dados num modelo suposto ou considerado válido. Esta incorporação

designa-se por estimação; o modelo com os dados incorporados, permite prever o comportamento de novos factos ou observações.

A inferência estatística associada com a fase de crítica e que estabelece o confronto entre o que pode esperar-se se o modelo proposto for apropriado e os dados que na verdade são observados.

Assim, a análise dos resíduos (de uma equação de regressão, por exemplo), os ensaios de significância, os ensaios de ajustamento (por exemplo, o qui-quadrado), são técnicas de crítica; o método dos mínimos quadrados, o método da máxima verosimilhança, a regressão «*ridge*», são técnicas de estimação.

A separação crítica-estimação não deve entender-se de forma absoluta (a estimação pode fazer-se, quer na fase final do processo depois de concluir pela validade do modelo, quer nas fases intermédias de crítica que exigem normalmente estimações provisórias dos parâmetros) mas, na verdade, revela-se muito fecunda, nomeadamente na comparação da inferência clássica e da inferência bayesiana (veja-se secção 1.5).

1.3 Inferência clássica

Na secção anterior procurou chamar-se a atenção para o importante papel que os modelos representam na 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. Na inferência clássica[8] o modelo visa descrever a situação experimental ou o processo gerador dos dados ou observações. Para Birnbaum (1962): «*The adequacy of any such model is tipically supported, more or less adequately, by a complex informal synthesis of previous experimental evidence of various kinds and theoretical considerations concerning both subject-matter and experimental techniques*».

O modelo clássico compreende um espaço de resultados ou espaço da amostra, $\mathcal{X}$, cujos elementos, x, resultam da observação de uma variável ou vector aleatório[9], X. Cada elemento, $x \in \mathcal{X}$, representa o resultado da amostragem ou da experimentação; quer dizer, x corresponde genericamente aos dados estatísticos.

---

[8] E não só.

[9] Quando houver nítida vantagem em distinguir entre escalares e vectores, o símbolo destes escreve-se em normando: $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, etc. Sempre que a interpretação resulte clara da exposição omite-se o normando. Quando estiverem em causa operações com vectores (matrizes), $\mathbf{X}$ designa um vector coluna, isto é, $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$ é o vector coluna obtido por transposição do vector linha $(X_1, X_2, \ldots, X_N)$. Por exemplo, $\sum X_i^2 = \mathbf{X}'\mathbf{X}$, ou, com $\mathbf{B} = [b_{ij}]$, matriz $N \times N$, $\sum\sum b_{ij} X_i X_j = \mathbf{X}'\mathbf{B}\mathbf{X}$. De contrário não se tem a preocupação de distinguir entre vectores linha e vectores coluna.

A opção por um espaço da amostra — que ocorre ao nível do delineamento da experiência ou do plano de amostragem — é de tal modo fundamental no método clássico que este se designa também por teoria da amostragem.

As possíveis funções de probabilidade (caso discreto) ou funções densidade de probabilidade (caso contínuo) de X designam-se genericamente por $f(x\,|\,\theta)$, onde θ, escalar ou vector, é um parâmetro pertencente a um conjunto, Θ, o espaço do parâmetro. Assim,

$$\mathcal{F} = \{f(x\,|\,\theta) \,:\, \theta \in \Theta\}, \tag{1.1}$$

forma uma família de distribuições probabilísticas e constituí, por assim dizer, a parte nuclear do modelo.

O parâmetro, θ, é índice ou indicador dos membros da família, $\mathcal{F}$. Esta pode ser muito ampla a ponto de incluir, por exemplo, funções de densidade de todas as variáveis aleatórias (absolutamente) contínuas; no caso paramétrico propriamente dito, a expressão analítica de $f(x\,|\,\theta)$ é especificada a priori — supõe-se conhecida — e a atenção concentra-se na pesquisa do «verdadeiro» valor de θ, isto é, do valor particular que indexa a distribuição que descreve «apropriadamente» as condições em que se observa X.

Nas aplicações sabe-se, quase sempre, se X é variável aleatória discreta ou contínua, isto é, se $f(x\,|\,\theta)$ é função de probabilidade ou função de densidade de probabilidade. Para não estar sempre a distinguir entre caso discreto e caso contínuo exprimem-se genericamente os raciocínios em termos da função de densidade $f(x\,|\,\theta)$[10].

Para dar maior generalidade ao modelo — evitando também a dicotomia referida no parágrafo anterior — pode tomar-se em alternativa a (1.1) a família de medidas de probabilidade[11],

$$\mathcal{F} = \{P_\theta \,:\, \theta \in \Theta\}, \tag{1.2}$$

onde, $P_\theta(B) = P_\theta(X \in B)$ [B conjunto da álgebra-σ definida em $\mathfrak{X}$] ou a família de funções de distribuição,

$$\mathcal{F} = \{F(x\,|\,\theta) \,:\, \theta \in \Theta\}, \tag{1.3}$$

com, $F(x\,|\,\theta) = P_\theta(X \leqq x)$.

[10] O que é matematicamente correcto, pois no caso discreto a função de probabilidade é uma função de densidade em relação a uma medida de contagem e no caso contínuo está-se (correntemente) em presença de uma função de densidade em relação à medida à Lebesgue.

[11] Ao especificar uma medida de probabilidade, P, é necessário indicar, além do espaço da amostra, $\mathfrak{X}$, a classe de conjuntos de $\mathfrak{X}$, seja $\mathcal{A}$, para os quais a medida P é definida; em geral, $\mathcal{A}$, é uma álgebra-σ. Nos problemas aqui tratados $\mathfrak{X}$ é quase sempre um espaço euclideano e $\mathcal{A}$ o correspondente corpo de Borel. As referências a $\mathcal{A}$ são usualmente omitidas, usando-se por vezes a designação de «mensurável» para exprimir que B, $B \subset \mathfrak{X}$, pertence à classe $\mathcal{A}$, $B \in \mathcal{A}$.

Exemplo 1.3 — Considere-se ainda o caso do dique. Suponha-se que se projecta medir os caudais numa dada secção do rio em N momentos; em princípio tem-se $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, onde X_i, $i = 1, 2, \ldots, N$, é a variável aleatória que representa o caudal no momento i. A experiência consiste na observação de $\mathbf{X}$ e a sua concretização dá lugar a uma amostra, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, onde x_i, $i = 1, 2, \ldots, N$, é o caudal observado no momento i. O espaço da amostra é o subconjunto de $\mathbf{R}^N$,

$$\mathcal{X} = \{(x_1, x_2, \ldots, x_N) : x_i > 0, \quad i = 1, 2, \ldots, N\}.$$

Admitindo — hipótese muito forçada — que X_i são variáveis aleatórias independentes e identicamente distribuídas, simbolicamente I.I.D., com distribuição exponencial negativa, vem,

$$f(\mathbf{x} \mid \theta) = \Pi\, \theta \exp\{-\theta x_i\} = \theta^N \exp\{-\theta \Sigma x_i\}, x_i > 0, \theta > 0.$$

O espaço do parâmetro é,

$$\Theta = \{\theta : \theta > 0\},$$

e a família de funções densidade de probabilidade,

$$\mathcal{F} = \{\theta^N \exp\{-\theta \Sigma x_i\} : \theta > 0\}.$$

□

A inferência clássica procura responder a questões[12] como as seguintes (veja-se Cox e Hinkley (1974)):

a) Os dados, x, suportam ou são compatíveis com a família $\mathcal{F}$? É o trabalho de crítica de que se falou na secção anterior.
b) Supondo que os dados emanaram de uma das distribuições da família $\mathcal{F}$, isto é, admitindo a validade do modelo, que inferências podem fazer-se sobre o verdadeiro valor do parâmetro? É o trabalho de estimação de que igualmente se falou na secção anterior.

A escolha da família $\mathcal{F}$ é passo decisivo na inferência clássica que, insista-se, é dado através de uma complexa síntese da experiência passada e do conhecimento teórico sobre a natureza dos fenómenos em estudo[13].

Com, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, uma hipótese muito usada na construção dos modelos é a de que as variáveis aleatórias, X_i, são independentes mesmo quando as

[12] Questões que podem colocar-se igualmente quando se empregam métodos não clássicos.
[13] Como se adiantou no Prefácio, a presente obra aborda quase exclusivamente problemas paramétricos. Em problemas não paramétricos a falta de informação ou a procura de maior generalidade levam o investigador a trabalhar com famílias, $\mathcal{F}$, muito amplas, por exemplo, famílias constituídas por todas as distribuições absolutamente contínuas.

observações, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, não são obtidas por um claro processo de amostragem casual. No caso do Ex. 1.3, se os caudais não são independentes, o que é bem provável, o modelo teria de ser construído com base em,

$$f(x_1, x_2, \ldots, x_N \,|\, \theta),$$

função densidade conjunta dos caudais, não sendo lícito escrever,

$$f(x_1, x_2, \ldots, x_N \,|\, \theta) = \Pi\, f(x_i \,|\, \theta);$$

quer dizer, o modelo seria muito mais complexo na medida em que envolveria a estrutura de dependência — covariância, pelo menos — entre os caudais nos vários momentos considerados.

Outra hipótese a que se recorre com alguma frequência é a de que as variáveis, X_i, $i = 1, 2, \ldots, N$, possuem distribuição Normal com média, μ, e variância, σ^2. Simbolicamente, combinando esta hipótese com a anterior, é corrente a especificação,

$$X_i \text{ I.I.D.}, X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, N.$$

A normalidade, pelas vantagens teóricas e práticas que oferece, é não raras vezes pressuposta a partir de evidência directa insuficiente à luz dos critérios científicos usuais. De facto, mesmo depois de responder afirmativamente à questão *a)* atrás proposta, não é fácil estabelecer que determinada distribuição tem a forma apropriada para a variável aleatória objecto de observação. Em face da incerteza quanto ao modelo adequado, a hipótese da normalidade é em geral avançada com a seguinte justificação: 1) haver conhecimento de que a variável em questão não tem provavelmente distribuição muito diferente da Normal (teorema do limite central, etc.); 2) haver numerosos estudos sobre os efeitos da não normalidade nos procedimentos e métodos aplicáveis quando a distribuição Normal é válida (procedimentos robustos, etc.).

Hipóteses, mais aderentes ou menos aderentes à realidade, são sempre indispensáveis à construção de modelos, clássicos ou não. O que é preciso ter bem presente na resposta à questão *b)*, mais ainda quando houver tendência para encarar com ligeireza a fase de crítica, é que as inferências são sempre condicionadas pela validade do modelo (1.1) [ou (1.2) ou (1.3)].

Aliás, desde que os modelos se considerem indispensáveis, não pode escapar-se ao bem conhecido dilema: a análise estatística passa pela descrição sumária da informação contida num conjunto de dados por meio de um modelo probabilístico; porém, essa descrição, imprescindível para que a análise se desenvolva, pode, quando o modelo não é adequado, distorcer a informação ou mesmo excluir informação relevante.

As desagradáveis consequências do emprego de um modelo inadequado levam a dar grande atenção aos critérios de escolha. Cox e Hinkley (1974), apesar de reconhecerem a dificuldade em apontar regras precisas, destacam, entre outras, as seguintes (a regra [1] repisa um ponto já focado):

[1] O modelo deve, sempre que possível, estabelecer uma ligação com os conhecimentos teóricos sobre o sistema em questão e com o trabalho experimental anteriormente realizado.
[2] A forma do modelo deve ser tal que os respectivos parâmetros tenham uma interpretação clara.
[3] O modelo deve ser tão simples quanto possível; o conhecido princípio da parcimónia sugere que se adoptem modelos com um número de parâmetros tão reduzido quanto possível.
[4] O modelo deve ser acessível à aplicação de procedimentos estatísticos correntes ou que não careçam de uma teoria estatística muito elaborada.

Não são de excluir conflitos entre os princípios indicados. Em particular [3] pode contrariar [1] e [2]; trata-se, em regra, de casos onde se pode mostrar vantajoso uma análise com dois tipos de modelos, um vinculado à teoria dos fenómenos em causa, outro capaz de efectuar a descrição em termos mais económicos.

Exemplo 1.4 — As baterias produzidas por uma unidade industrial têm duração, X, com distribuição, por hipótese, log-normal. Uma amostra genérica de dimensão, N, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, com X_i I.I.D., $\log X_i \sim N(\mu, \sigma^2)$, tem função de densidade,

$$f(\mathbf{x} \mid \mu, \sigma^2) = (\Pi x_i)^{-1}(2\pi\sigma^2)^{-N/2} \exp\left[-\left(\frac{1}{2\sigma^2}\right)\Sigma(\log x_i - \mu)^2\right], \qquad (1.4)$$

com $\mathcal{X} = \{(x_1, x_2, \ldots, x_N) : x_i > 0\}$ e $\Theta = \{\boldsymbol{\theta} = (\mu, \sigma^2) : -\infty < \mu < \infty, \sigma > 0\}$.

As inferências que pretendam fazer-se sobre $\boldsymbol{\theta} = (\mu, \sigma^2)$, a partir da amostra, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, são condicionadas pela validade do modelo (1.4) e não é difícil conceber situações em que o modelo deixa de ser adequado. Por exemplo, se durante o período de observação um dos produtos químicos que entra no fabrico sofrer variações que afectam linearmente a média do logaritmo da duração das baterias, então, sendo z_i o valor da intensidade do mesmo produto na i-ésima bateria,

$$f(\mathbf{x} \mid \mathbf{z}, \mu_1, \mu_2, \sigma^2) = (\Pi x_i)^{-1}(2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\Sigma(\log x_i - \mu_1 - \mu_2 z_i)^2\right], \qquad (1.5)$$

é um modelo mais adequado para analisar o problema da duração das baterias, sendo também de notar que o vector parâmetro passa a ser, $\boldsymbol{\theta}^* = (\mu_1, \mu_2, \sigma^2)$ [Box e Tiao (1973)]. □

No caso clássico as inferências consistem, grosso modo, na estimação pontual, na estimação por intervalos ou regiões, nos ensaios de significância e nos ensaios de hipóteses. Apesar de se supor conhecido o tratamento elementar desses problemas convém recapitular rapidamente como intervém o princípio de amostragem repetida[14].

Considere-se, para exemplificação, a estimação pontual. Seja $\hat{\theta}(X)$ um estimador de θ; $\hat{\theta}(X)$ pode considerar-se em termos gerais um bom estimador quando, com algum ξ pequeno e positivo, e com base na respectiva distribuição por amostragem, se consegue provar que, por exemplo,

$$P_\theta(|\hat{\theta}(X) - \theta| < \xi) = 0{,}99, \tag{1.6}$$

para todo o $\theta \in \Theta$.

Quando a partir da amostra recolhida, x, se calcula a estimativa, $\hat{\theta}(x)$, pode apenas afirmar-se que:

(I) $|\hat{\theta}(x) - \theta| < \xi$,

ou

(II) um acontecimento improvável — de probabilidade 0,01 — ocorreu.

Obviamente, o investigador que emprega um tal procedimento não conhece θ; consequentemente não sabe se está na situação (I) ou na situação (II). Sendo adepto da inferência clássica socorre-se então do princípio de amostragem repetida e afirma: se o cálculo de uma estimativa for repetido um grande número de vezes a proporção de estimativas que verificam (I) não andará longe de 99%. Por outras palavras, o estimador $\hat{\theta}(X)$ é bom porque em 99% dos casos conduz a estimativas que diferem pouco do verdadeiro valor do parâmetro. O seu emprego encontra-se, portanto, justificado.

No princípio de amostragem repetida encontra-se nitidamente uma referência para o que se passa em todo o espaço da amostra; é o que sucede, nomeadamente, quando se concebe[15] o conjunto de estimativas, $\hat{\theta}(x)$, que podem ser geradas quando x percorre $\mathfrak{X}$. E é um facto que 99% dos elementos desse conjunto satisfazem (I).

Considerações análogas podem fazer-se em relação aos intervalos de confiança ou aos ensaios de hipóteses estatísticas. Por exemplo, quando no ensaio de uma hipótese simples, H_0, contra uma alternativa também simples, H_1, se diz que uma dada região crítica tem erro de 1.ª espécie com probabilidade 0,01 e erro de 2.ª espécie com probabilidade 0,05 pretende significar-se que numa longa repetição do

[14] Para uma análise mais profunda da perspectiva frequencista veja-se Berger (1985).

[15] Concepção associada com a ideia de distribuição por amostragem.

ensaio, a região crítica conduz à rejeição de H_0 quando verdadeira em 1% dos casos e à aceitação de H_1 quando falsa em 5% dos casos.

A probabilidade 0,99 em (1.6) [ou o par (0,01; 0,05) no exemplo do ensaio de hipóteses] refere-se à precisão inicial do estimador considerado. A precisão inicial também se designa por pre-experimental pelo óbvio motivo de se referir a uma situação anterior à observação dos dados.

A precisão inicial é, portanto, um importante critério de avaliação dos procedimentos clássicos e não se lhe dirigem críticas quando está em jogo a qualificação de um procedimento empregado de forma repetitiva. É o que acontece, por exemplo, com os ensaios de recepção por amostragem ou com o controlo estatístico da qualidade quando efectuados como rotina.

Do que os autores não clássicos discordam é do emprego da precisão inicial nas situações que ocorrem uma vez e não se repetem. É o que de facto exprime Berger (1980): «*When dealing with a one time situation, however, it is not clear what the relevance of initial precision is*». Consequentemente sugerem como critério a precisão final ou post-experimental que diz respeito, naturalmente, a uma situação posterior à observação dos dados. Citando ainda Berger (1980): «*In evaluating a one time 95% confidence interval for example, it is of little confort to know that in 95% of a long series of such experiments the interval would contain θ. What is desired is a feeling as to how likely it is that θ is contained in the particular interval being considered*».

Os exemplos seguintes ilustram a diferença entre precisão inicial e precisão final.

Exemplo 1.5 — A variável, X, assume dois valores, $\theta - 1$ e $\theta + 1$ ($-\infty < \theta < +\infty$), cada um dos quais com probabilidade igual a 1/2. A partir de uma amostra casual com dimensão, $N = 2$, (X_1, X_2), pretende construir-se um intervalo de confiança a 75%, de amplitude mínima, para θ. É fácil verificar que tal intervalo (aliás degenerado), seja $J(X_1, X_2)$, assume a forma,

$$J(x_1, x_2) = \begin{cases} (x_1 + x_2)/2 & \text{se } |\, x_1 - x_2 \,| = 2, \\ (x_1 + 1) & \text{se } |\, x_1 - x_2 \,| = 0, \end{cases}$$

[alternativamente pode tomar-se o ponto $x_1 - 1$ quando $|\, x_1 - x_2 \,| = 0$]; com efeito,

$$\begin{aligned} P[\theta \in J(X_1, X_2)] &= P[|\, X_1 - X_2 \,| = 2] P[\theta = (X_1 + X_2)/2 \,\big|\, |\, X_1 - X_2 \,| = 2] + \\ &\quad + P[|\, X_1 - X_2 \,| = 0] P[\theta = (X_1 + 1) \,\big|\, |\, X_1 - X_2 \,| = 0] \\ &= (1/2) + (1/2)(1/2) = 3/4. \end{aligned}$$

A precisão inicial é, portanto, 0,75. No entanto, se sair $|\, x_1 - x_2 \,| = 2$, está-se absolutamente certo que $\theta = (x_1 + x_2)/2$, isto é, a precisão final é 1; se sair $|\, x_1 - x_2 \,| = 0$ a incerteza reparte-se igualmente por $\theta = x_1 - 1$ ou $\theta = x_1 + 1$, isto é, a precisão final é 0,5. [Berger (1984)]. □

Exemplo 1.6 — Tomem-se, X_i I.I.D., $X_i \sim U(\theta - 1/2, \theta + 1/2)$, $i = 1, 2, \ldots, N$, onde $U(\alpha, \beta)$, $\alpha < \beta$, designa a distribuição uniforme no intervalo (α, β). Suponha-se N = 25; pode mostrar-se, considerando a distribuição da estatística, $[\min(X_i) + \max(X_i)]/2$, que,

$$P[\theta_1(\mathbf{X}) < \theta < \theta_2(\mathbf{X})] = 0{,}95,$$

quando,

$$\theta_1(\mathbf{X}) = \frac{1}{2}[\min(X_i) + \max(X_i)] - 0{,}056,$$

$$\theta_2(\mathbf{X}) = \frac{1}{2}[\min(X_i) + \max(X_i)] + 0{,}056.$$

Fica pois estabelecida, por processo tipicamente clássico, uma forma de proceder à construção de intervalos de confiança a 95% para θ, isto é, intervalos de confiança com uma precisão inicial igual a 0,95.

Suponha-se que para uma amostra concreta, particularmente «infeliz», se obteve $\min(x_i) = 3{,}1$ e $\max(x_i) = 3{,}2$. Sabe-se desde logo, sem margem para erro, que,

$$\max(x_i) < \theta + \frac{1}{2} \Rightarrow \theta > 2{,}7,$$

$$\min(x_i) > \theta - \frac{1}{2} \Rightarrow \theta < 3{,}6,$$

isto é, que θ pertence seguramente ao intervalo (2,7, 3,6).

Por outro lado, o intervalo de confiança obtido pela forma como acima se explicitou, ou seja (3,094, 3,206), parece extremamente enganador quando se apresenta associado com a precisão inicial de 0,95; de facto, em face da amostra concreta, é mais intuitivo e plausível atender à precisão final, calculada, por exemplo, pelo quociente,

$$(3{,}206 - 3{,}094)/(3{,}6 - 2{,}7) = 0{,}124.$$

A situação inversa ocorre quando se obtém uma amostra particularmente «feliz», com $\min(x_i) = 3{,}0$ e $\max(x_i) = 3{,}96$; neste caso, pode afirmar-se com certeza ser $3{,}46 < \theta < 3{,}5$, enquanto se tem como intervalo de confiança a 95%, (3,424, 3,536), resultado que sem dúvida obriga a meditar [Berger (1980)]. □

Os Exs. 1.5 e 1.6, através dos quais se pôs em confronto a precisão inicial e a precisão final, permitem ainda esboçar os motivos que levam os bayesianos a sugerir a incoerência dos intervalos de confiança.

Considere-se, primeiramente, o conceito de conjunto relevante («*recognizable*», na terminologia de Fisher). O conjunto do espaço da amostra, B, $B \subset \mathfrak{X}$,

diz-se relevante quando, para um intervalo (ou região) de confiança, $J(X)$, com probabilidade de cobertura,

$$P[\theta \in J(X)] = \alpha \text{ para todo o } \theta \in \Theta,$$

existe um número positivo, ξ, tal que,

$$P[\theta \in J(X) \,|\, X \in B) \leqq \alpha - \xi \text{ para todo o } \theta \in \Theta,$$

ou,

$$P[\theta \in J(X) \,|\, X \in B] \geqq \alpha + \xi \text{ para todo o } \theta \in \Theta.$$

Exemplo 1.5 — *Continuação.* Neste caso são conjuntos relevantes,

$$B_0 = \{(x_1, x_2) \,:\, |\, x_1 - x_2 \,| = 2\},$$
$$B_1 = \{(x_1, x_2) \,:\, |\, x_1 - x_2 \,| = 0\}.$$

Com efeito, para o intervalo de confiança a 75%, $J(X_1, X_2)$,

$$P[\theta \in J(X_1, X_2)] = 0{,}75,$$

mas,

$$P[\theta \in J(X_1, X_2) \,|\, (X_1, X_2) \in B_0] = 1 > 0{,}75,$$

e,

$$P[\theta \in J(X_1, X_2) \,|\, (X_1, X_2) \in B_1] = 0{,}5 < 0{,}75. \qquad \square$$

Exemplo 1.6 — *Continuação.* É fácil de ver que neste caso, com valores b_0 e b_1 convenientes, $0 < b_0 < b_1 < 1$, os conjuntos de $\mathfrak{X}$,

$$B_0 = \{\mathbf{x} \,:\, \max(x_i) - \min(x_i) < b_0\},$$
$$B_1 = \{\mathbf{x} \,:\, \max(x_i) - \min(x_i) > b_1\},$$

correspondem às amostras particularmente «infelizes» e «felizes» e constituem exemplos de conjuntos relevantes. □

Alguns autores sustentam que a existência de conjuntos relevantes, B, implica que o coeficiente de confiança (probabilidade de cobertura) não é uma medida apropriada da confiança que se tem na cobertura de θ por $J(X)$ quando se observa $X \in B$ (probabilidade condicionada de cobertura)[16]. A adjectivação de incoerência — bayesiana — parece então sair da seguinte análise. Retome-se o Ex. 1.5: se

[16] O problema da inferência condicionada volta a ser aflorado um pouco mais adiante. Há vasta literatura sobre o estabelecimento de condições em que existem ou não existem conjuntos relevantes [Berger e Wolpert (1984)].

Pedro usa $J(X_1, X_2)$ como intervalo de confiança a 75% está naturalmente disposto a apostar, ganhando 1$ se houver cobertura e perdendo 3$ se houver não cobertura ou, indiferentemente, ganhando 3$ se houver não cobertura e perdendo 1$ se houver cobertura. Na primeira situação diz-se que aposta na cobertura, na segunda diz-se que aposta na não cobertura. Agora, se Paulo entra no jogo com a possibilidade de optar pela aposta na cobertura/não cobertura depois de registar o resultado, (x_1, x_2), da observação, pode bater Pedro apostando na cobertura quando sai $|x_1 - x_2| = 2$ ou na não cobertura quando sai $|x_1 - x_2| = 0$. Pedro — utilizador de intervalos de confiança — sofre (em média) um prejuízo certo e é incoerente à luz da doutrina bayesiana.

Deve-se a Cox (1958) um exemplo particularmente instrutivo sobre o emprego indiscriminado da precisão inicial:

Exemplo 1.7 — A variável aleatória, $X \sim N(\theta, \sigma^2)$, é observada com o objectivo de ensaiar H_0: $\theta = 0$ contra H_1: $\theta = 10$. A observação de X pode fazer-se com um de dois instrumentos de medida: o primeiro, I_1, é pouco preciso ($\sigma = 10$); o segundo, I_2 , é preciso ($\sigma = 1$). O observador recebe I_1 com probabilidade p conhecida ($0 < p < 1$) e I_2 com probabilidade $1 - p$, ficando a saber qual dos instrumentos lhe é distribuído.

Para ensaiar H_0 contra H_1 podem propor-se, entre outros, os dois testes seguintes, ambos de dimensão α:

Teste 1:

- com I_1, rejeitar quando $X > K_0$, com[17] $\Phi(K_0/10) = 1 - \alpha$;
- com I_2 rejeitar quando $X > K_1$, com $\Phi(K_1) = 1 - \alpha$.

A dimensão é α, porquanto,

$$p \cdot P_{\theta=0,\,\sigma=10}(X > K_0) + (1 - p) \cdot P_{\theta=0,\sigma=1}(X > K_1) = p\alpha + (1 - p)\alpha = \alpha.$$

Teste 2:

- com I_1, rejeitar sempre;
- com I_2, rejeitar quando $X > K_2$, com (supondo $\alpha > p$),

$$\alpha = p + (1 - p) \cdot P_{\theta=0,\sigma=1}(X > K_2). \tag{1.7}$$

Em consequência de (1.7) tem-se também dimensão α.

[17] $\Phi(u)$ designa a função de distribuição da $N(0,1)$ embora mais adiante o símbolo Φ (sem argumento) seja empregue para designar classes de funções de decisão. A distinção resulta claramente de cada contexto.

Pode mostrar-se, para muitos valores de α e de p, que o Teste 2 é mais potente do que o Teste 1. Em tais casos, o Teste 2 é o recomendado pela doutrina clássica, quer dizer, é o melhor dos dois em termos de precisão inicial.

Se o observador se propõe efectuar uma longa série de ensaios não tem, muito provavelmente, motivo que o leve a recusar a recomendação clássica. Se está em causa uma única experiência, o observador fica naturalmente perplexo com a sugestão de ignorar o resultado da observação no caso de lhe ser distribuído I_1. De facto, é natural que não lhe interesse atender à precisão inicial mas à precisão que pode obter quando usa o instrumento que lhe é distribuído, seja ele qual for. □

O papel fulcral da precisão inicial nos procedimentos clássicos é, está bem de ver, consequência directa do princípio de amostragem repetida. Este princípio implica a transferência para os procedimentos estatísticos, mesmo quando haja intenção de os aplicar uma só vez, das propriedades que os mesmos revelam num conjunto ilimitado de repetições (necessariamente hipotéticas). É a atribuição das propriedades de um colectivo ou «conjunto de referência» às inferências específicas ou individuais que se torna de difícil aceitação para os opositores [Barnett (1982)]. Acresce ainda que não raras vezes pode haver ambiguidade no que toca ao conjunto de referência, isto é, à série de repetições relevante, sobretudo quando é possível imaginar várias formas de repetir a experiência ou observação e não é óbvio qual das formas é a mais apropriada para estabelecer inferências sobre o parâmetro[18].

Perante a crítica os clássicos reagem notando: *i)* os procedimentos clássicos têm fundamento próprio, nomeadamente o conceito frequencista de probabilidade, e nunca tiveram em mira a precisão final. Repare-se que a nota de incoerência (bayesiana), apresentada na forma como foi dirigida aos intervalos de confiança, é refutada nas mesmas linhas: tudo o que se reclama do nível de confiança é ser uma probabilidade não condicionada de cobertura susceptível de interpretação frequencista de acordo com a lei dos grandes números. Não é, portanto, justo pedir que os intervalos de confiança respondam a questões para as quais não foram orientados, mesmo que não se negue a surpresa que causa o aparecimento de conjuntos relevantes e das diferenças que se verificam em relação ao nível de confiança quando se tomam probabilidades condicionadas de cobertura; *ii)* o emprego da precisão inicial só muito raramente causa dificuldades, porquanto, nos problemas que

[18] Numa sucessão de provas de Bernoulli, em que θ é a probabilidade de um «sucesso», obtém-se $X = 15$ «sucessos» e $Y = 35$ «insucessos». Podem imaginar-se pelo menos três formas de repetir a experiência: 1) Repetir com $X+Y$ fixo em 50 (i.e., fixando o número de provas); 2) Repetir com X fixo em 15, o que implica ser Y o número (variável) de «insucessos» que antecedem a obtenção de 15 «sucessos»; 3) Repetir com Y fixo em 35, o que implica ser X o número (variável) de «sucessos» que antecedem a obtenção de 35 «insucessos». [Kalbfleisch (1985)]. A questão é adiante retomada.

envolvem as distribuições mais correntes — por exemplo, a Normal — verifica-se haver em regra coincidência entre a precisão inicial e a precisão final.

Convém prosseguir na apreciação da estatística clássica no quadro de alguns princípios gerais e de certo modo abstractos que têm sido propostos com o objectivo de estabelecer a forma como os dados devem afectar as inferências ou conclusões. Trata-se de princípios que dizem respeito aos aspectos dos dados e do modelo que devem considerar-se relevantes e que diferem dos princípios que dizem respeito à interpretação das conclusões, como, por exemplo, os de amostragem repetida e de coerência bayesiana.

Começa-se pelo princípio de verosimilhança[19] cujo enunciado carece da definição da função de verosimilhança.

A definição da função de verosimilhança é um ponto em que é importante manter a distinção entre o caso discreto e o caso contínuo, distinção que por vezes é descurada [Kempthorne e Folks (1971) é uma notável excepção].

Suponha-se que $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. com função de probabilidade $f(x \mid \theta)$, representa uma amostra casual. O acontecimento, $A = (X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$, tem probabilidade, $P_\theta(A) = \Pi f(x_i \mid \theta)$; fixando A e fazendo variar θ sobre Θ, tem-se a função de verosimilhança,

$$L(\theta) = L(\theta \mid A) = \Pi f(x_i \mid \theta),$$

que tem, portanto, domínio em Θ e que para cada $\theta \in \Theta$ indica a verosimilhança que lhe é atribuída quando se sabe que o acontecimento A se realizou, isto é, que foi observado $\mathbf{X} = \mathbf{x}$.

A verosimilhança não é uma probabilidade: por exemplo, não tem significado a adição de verosimilhanças. Somente a razão de verosimilhanças tem sentido: assim, $L(\theta)/L(\theta') = \Pi f(x_i \mid \theta)/\Pi f(x_i \mid \theta')$, mede o «peso da evidência» de θ contra θ' decorrente da realização do acontecimento A. Por isso a função de verosimilhança é definida a menos de um factor constante (i.e., independente de θ) positivo,

$$L(\theta) = K\Pi f(x_i \mid \theta), \quad \theta \in \Theta.$$

Se $\mathbf{X}$ não é uma amostra casual e possui componentes, $(X_1, X_2, \ldots, X_N)$, com função de probabilidade conjunta, $f(x_1, x_2, \ldots, x_N \mid \theta)$, tem-se como função de verosimilhança,

$$L(\theta) = K f(x_1, x_2, \ldots, x_N \mid \theta), \quad \theta \in \Theta.$$

[19] O termo verosimilhança foi introduzido por Fisher com o objectivo de evitar o teorema de Bayes; curiosamente, como vai ver-se, são os bayesianos e não os clássicos que respeitam o princípio de verosimilhança.

Suponha-se que $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. com função densidade de probabilidade $f(x \mid \theta)$, representa uma amostra casual. Como $f(x \mid \theta)$ é agora uma densidade e não uma probabilidade, para preservar a transposição,

$$\text{probabilidade} \Rightarrow \text{verosimilhança},$$

é necessário determinar um limite aproximado seguindo um raciocínio aliás bem conhecido.

Considere-se o acontecimento, $A^* = \bigcap \left(x_i - \frac{1}{2}\Delta_i < X_i \leqq x_i + \frac{1}{2}\Delta_i\right)$[20]; devido à independência tem-se a probabilidade,

$$P_\theta(A^*) = \Pi\, P_\theta \left(x_i - \frac{1}{2}\Delta_i < X_i \leqq x_i + \frac{1}{2}\Delta_i\right).$$

Se os Δ_i, $i = 1, 2, \ldots, N$, forem pequenos é em geral válida [veja-se Kempthorne e Folks (1971)] a aproximação,

$$\begin{aligned} P_\theta(A^*) &\approx \Pi f(x_i \mid \theta)\ \Delta_i \\ &\approx [\Pi \Delta_i]\ \Pi f(x_i \mid \theta). \end{aligned}$$

Como $\Pi\Delta_i$ pode ser absorvido na constante, pois não envolve θ, pode finalmente escrever-se a função de verosimilhança,

$$L(\theta) = K \Pi f(x_i \mid \theta), \quad \theta \in \Theta, \tag{1.8}$$

em termos aplicáveis no caso discreto e no caso contínuo desde que se tenha presente a distinção apontada e se não esqueça que a cada ponto $\mathbf{x} \in \mathcal{X}$ corresponde uma função de verosimilhança.

A função de verosimilhança goza de papel fundamental, quer na inferência clássica, quer na inferência bayesiana[21], como veículo portador da informação dada pela amostra acerca do parâmetro desconhecido. O princípio de verosimilhança sustenta que toda a informação está contida na função de verosimilhança.

Mais precisamente, no quadro do modelo $\mathcal{F}$ o princípio de verosimilhança fraco estabelece o seguinte: se as observações x e x' são tais que[22],

$$f(x \mid \theta) = K(x, x') f(x' \mid \theta) \quad \text{para todo o } \theta \in \Theta, \tag{1.9}$$

onde K não depende de θ, então são idênticas as inferências sobre θ feitas a partir de x ou a partir de x'.

[20] Recorde-se que devido à inevitável imprecisão dos instrumentos de medida as observações práticas de uma variável contínua formam um conjunto discreto, isto é, não se observam pontos mas sim intervalos.

[21] E noutras correntes, sobretudo, como era de esperar, na escola verosimilhancista.

[22] Repare-se que $f(x \mid \theta)$ corresponde, em (1.8), a $\Pi f(x_i \mid \theta)$.

O princípio de verosimilhança forte é introduzido em relação a dois modelos estatísticos ou sistemas aleatórios, o primeiro dando observações $x \in \mathcal{X}$, de uma variável (ou vector) aleatória, X, com função densidade de probabilidade (ou função de probabilidade), $f_X(x\,|\,\theta)$, e o segundo dando observações $y \in \mathcal{Y}$, de uma variável (ou vector) aleatória, Y, com função densidade (ou função de probabilidade), $f_Y(y\,|\,\theta)$, sendo o espaço do parâmetro, Θ, comum aos dois modelos. Se,

$$f_X(x\,|\,\theta) = K(x,y) f_Y(y\,|\,\theta) \quad \text{para todo o } \theta \in \Theta, \tag{1.10}$$

onde K não depende de θ, são idênticas as inferências sobre θ feitas a partir de x ou a partir de y.

Uma forma de olhar o princípio de verosimilhança[23] — que tem ilustres defensores, como G. A. Barnard, Allan Birnbaum, A. W. F. Edwards, etc. — é pensar que o mesmo estabelece que a observação particular ou amostra concreta, x, representa o único elemento do espaço da amostra, $\mathcal{X}$, relevante nas inferências sobre θ; consequentemente, os elementos de $\mathcal{X}$, que poderiam eventualmente ter sido obtidos, mas que não o foram, não fornecem qualquer informação sobre θ. Como imediatamente se depreende — releia-se a pág. 17 — os clássicos contam-se entre os opositores do princípio de verosimilhança na medida em que defendem que a avaliação da informação fornecida por $x \in \mathcal{X}$ passa necessariamente pela análise do que se passa com todos os pontos do espaço da amostra. Aliás, este é, normalmente, o conjunto de referência.

A definição do espaço da amostra está associada, logicamente, com o método de amostragem ou com o plano da experiência. Assim, o princípio de verosimilhança dispõe que o método de amostragem é irrelevante para as inferências; mais ainda, como o método de amostragem tem muito a ver com as chamadas regras de paragem, estas são também consideradas irrelevantes.

Impõem-se, nesta altura, a análise de alguns exemplos.

Exemplo 1.8 — Considerem-se duas experiências, E_1 e E_2, consistindo na observação de X_1 e X_2, respectivamente, com espaço de resultados, $\mathcal{X} = \{1, 2, 3\}$ e espaço do parâmetro, $\Theta = \{\theta_1, \theta_2\}$, comuns e funções de probabilidade,

E_1 :

X_1	$f_1(x_1\,\|\,\theta_1)$	$f_1(x_1\,\|\,\theta_2)$
1	0,90	0,09
2	0,05	0,055
3	0,05	0,855

E_2 :

X_2	$f_2(x_2\,\|\,\theta_1)$	$f_2(x_2\,\|\,\theta_2)$
1	0,26	0,026
2	0,73	0,803
3	0,01	0,171

[23] Forte quando referenciado sem qualificativo.

Se se observa, $X_1 = 1$, o princípio de verosimilhança estabelece que a informação acerca de θ deve depender da experiência unicamente através da relação entre, $(0{,}9, 0{,}09) = [f_1(1 \mid \theta_1),\ f_1(1 \mid \theta_2)]$. Além disso, como esses valores são proporcionais a, $(0{,}26, 0{,}026) = [f_2(1 \mid \theta_1),\ f_2(1 \mid \theta_2)]$, a observação de $X_2 = 1$ fornece a mesma informação que a observação de $X_1 = 1$. Ora, as razões de verosimilhanças continuam a ser iguais nas duas experiências quando se observam os valores 2 ou 3. Consequentemente, qualquer que seja a experiência realizada, o valor observado [seja 1, 2 ou 3], desde que seja o mesmo, conduz a idênticas conclusões sobre θ.

No entanto, as experiências E_1 e E_2 são muito diferentes numa perspectiva frequencista. Por exemplo, para ensaiar H_0: $\theta = \theta_1$ contra H_1: $\theta = \theta_2$, o teste com região de aceitação, $\{1\}$, e região de rejeição, $\{2, 3\}$, tem as seguintes características:

	E_1	E_2
Prob. Erro 1.ª espécie	0,10	0,74
Prob. Erro 2.ª espécie	0,09	0,026

e é o mais potente nos dois casos. A diferença é, porém, drástica.

Berger e Wolpert (1984), a quem se deve o exemplo, declaram: «*It is clear that experiment E_1 is more likely to provide useful information about θ, as reflected by the overall better error probabilities. The LP* [princípio de verosimilhança], *in no sense contradicts this. The LP applies only to the information about θ that is available from knowledge of the experiment and the observed x. Even though E_1 has a much better chance of yielding good information, the LP states that the conclusion, once x is at hand, should be the same, regardless of whether x came from E_1 or E_2*». □

Exemplo 1.9 — Considere-se uma sucessão de lançamentos independentes de uma moeda para a qual é θ a probabilidade de sair «coroa»; suponha-se que em dado momento se chega ao seguinte resultado ou amostra,

$$x = \{F, C, F, F, C, C, F, C, C, C\},$$

onde F designa «face» e C designa «coroa».

Este resultado pode ser obtido por diversos processos experimentais ou regras de paragem, nomeadamente:

- lançar a moeda com o número de lançamentos — no caso presente, 10 — fixado antecipadamente;
- lançar a moeda até que apareçam 6 «coroas»;
- lançar a moeda até que apareçam 3 «coroas» consecutivas;
- lançar a moeda até o lançador estar saturado, tendo a saturação sucedido com o 10.º lançamento.

Em qualquer dos casos a função de verosimilhança é proporcional a,

$$\theta^6(1-\theta)^4;$$

segundo o princípio de verosimilhança, toda a informação que x pode dar sobre θ encontra-se nesta expressão. Saber qual dos quatro processos experimentais foi adoptado ou saber qual foi a regra de paragem perfilhada nada vem acrescentar. Note-se que a possibilidade de o experimentador parar, por seu arbítrio, ao considerar o resultado x satisfatório, em nada altera o que acaba de dizer-se. □

Exemplo 1.9 — *Continuação.* Em relação à moeda do exemplo anterior pretende ensaiar-se a hipótese H_0: $\theta = 1/2$ contra a alternativa H_1: $\theta > 1/2$. São contemplados dois processos experimentais:

E_1: lançar a moeda 12 vezes;

E_2: lançar a moeda até que apareçam 3 «faces».

Admita-se que o resultado observado foi $y = 9$, valor particular da variável aleatória Y que designa o número de «coroas» (o número de «faces» foi portanto igual a 3).

Para um clássico o nível de significância do valor $y = 9$ difere nos dois casos.

No caso de E_1, $Y \sim B(12;\theta)$ [distribuição binomial], donde,

$$\alpha = P(Y \geqq 9 \,|\, \theta = 1/2) = \binom{12}{9}(1/2)^{12} + \binom{12}{10}(1/2)^{12} + \binom{12}{11}(1/2)^{12} + \binom{12}{12}(1/2)^{12}$$
$$= 0{,}075.$$

No caso de E_2, $Y \sim \mathrm{BN}(3; 1-\theta)$ [distribuição binomial negativa], donde,

$$\alpha' = P(Y \geqq 9 \,|\, \theta = 1/2) = \binom{11}{9}(1/2)^{12} + \binom{12}{10}(1/2)^{13} + \binom{13}{11}(1/2)^{14} + \cdots$$
$$= 0{,}0325.$$

Logo, se for adoptado um nível de significância de 5%, H_0 é rejeitada no caso de E_2 mas não o é no caso de E_1.

Para os adeptos do princípio de verosimilhança esta análise não é correcta, pois sustentam que as conclusões a tirar nos dois casos são idênticas visto que em qualquer deles a função de verosimilhança é proporcional a $\theta^9(1-\theta)^3$. De facto, as funções de verosimilhança são respectivamente,

$E_1)$ $L_1(\theta \,|\, y) = \binom{N}{y}\theta^y(1-\theta)^{N-y}$, ou seja, com $N = 12$, $y = 9$,
$L_1(\theta \,|\, 9) = \binom{12}{9}\theta^9(1-\theta)^3 = 220\,\theta^9(1-\theta)^3$;

$E_2)$ $L_2(\theta \,|\, y) = \binom{K+y-1}{y}\theta^y(1-\theta)^K$, ou seja, com $K = 3$, $y = 9$,
$L_2(\theta \,|\, 9) = \binom{11}{9}\theta^9(1-\theta)^3 = 55\,\theta^9(1-\theta)^3$

[Lindley (1976)]. □

O exemplo seguinte deve-se a Pratt (1962) e apresenta-se na língua original para não perder o aspecto pitoresco.

Exemplo 1.10 — «*An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate voltmeter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurement which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean. Later he visits the engineer's laboratory, and notices that the voltmeter used reads only as far as 100, so the population appears to be ‹censored›. This necessitates a new analysis, if the statistician is orthodox. However the engineer says he has another meter, equally accurate and reading to 1000 volts. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all. But the next day the engineer telephones and says, ‹I just discovered my high-range voltmeter was not working the day I did the experiment you analysed for me›. The statistician ascertains that the engineer would not have held the experiment until the meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. He says, ‹But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope›*».

Repare-se que no exemplo o que está em jogo é o espaço da amostra. Se o voltímetro não tiver limitações, tem-se, sendo N o número de observações,

$$\mathcal{X} = \{(x_1, x_2, ...x_N) : -\infty < x_i < +\infty, \quad i = 1, 2, \dots, N\};$$

se não medir voltagens superiores a 100, tem-se,

$$\mathcal{X}^* = \{(x_1, x_2, \dots, x_N) : -\infty < x_i \leqq 100, \quad i = 1, 2, \dots, N\}.$$

Sendo o modelo probabilístico a distribuição Normal (esqueça-se que as medidas obtidas no voltímetro são sempre positivas), no primeiro caso não há problema; no segundo caso, verifica-se a truncagem da distribuição no ponto 100. Do ponto de vista clássico a distinção é importante; em particular a estimação pontual ou por intervalos é diferente consoante a distribuição seja ou não truncada. Para quem aceite o princípio de verosimilhança somente conta a amostra de facto obtida, isto é, valores que poderiam ter sido obtidos ($x > 100$) mas que não o foram são irrelevantes para a inferência. □

Da exposição feita ressalta claramente o desrespeito do princípio de verosimilhança por parte dos métodos clássicos, residindo na obtenção de estimadores

pelo método da máxima verosimilhança talvez a única excepção. Pelo contrário, a propriedade de centragem ou não enviesamento,

$$E\left\{\hat{\theta}(X)\right\} = \int_{\mathfrak{X}} \hat{\theta}(x) f(x \mid \theta)\, dx = \theta, \quad \theta \in \Theta,$$

ou o emprego do erro quadrático médio,

$$E\left\{[\hat{\theta}(X) - \theta]^2\right\} = \int_{\mathfrak{X}} \left[\hat{\theta}(x) - \theta\right]^2 f(x \mid \theta)\, dx, \quad \theta \in \Theta,$$

são casos paradigmáticos de violação daquele princípio na medida em que envolvem a integração (no caso discreto, soma) sobre todo o espaço da amostra. Aliás a colisão também se verifica quando se procede a testes de significância ou a ensaios de hipóteses; nos primeiros, é corrente o emprego das abas da distribuição da variável ou estatística observada; nos segundos, costuma trabalhar-se com regiões críticas ou regiões de rejeição; ora, tanto as abas como as regiões críticas contém pontos que não foram observados. A propósito, a seguinte frase de Jeffreys tornou-se lendária: «*...a hypothesis which may be true may be rejected because it has not predicted observable results which have not ocurred*».

A sugestão de que os processos de amostragem e as regras de paragem são irrelevantes nas inferências não pode, como diz Gomes (1981): «*deixar de chocar qualquer pessoa cuja intuição estatística tenha sido desenvolvida no contexto da teoria de Neyman-Pearson*». Assim, sem no entanto conseguirem desacreditar seriamente o princípio de verosimilhança, tem aparecido regularmente contra-exemplos e os chamados paradoxos da regra de paragem.

Um dos paradoxos da regra de paragem foi rebatido por Barnard [na discussão de Savage (1962)]. Suponha-se que alguém quer provar a todo o custo a existência de percepção extra-sensorial e declara que vai prosseguir até obter uma proporção de sucessos significativa. Observa então Barnard: «*Knowing that this is what he is setting out to do would lead you to adopt a different test criterion. What you would look at would not be the ratio of sucesses obtained, but how long it took him to obtain it. ... And the reversing of the choice of test criteria would I think overcome the difficulty*».

Outra situação curiosa é a seguinte [veja-se Barnard e Godambe (1982)]: Considere-se uma sucessão de provas de Bernoulli em que θ representa a probabilidade de sucesso. Quando o verdadeiro valor de θ é θ_0 a lei do logaritmo iterado estabelece que realizando um número suficientemente grande de provas pode ter-se como resultado que a variável fulcral,

$$\left(\hat{\theta} - \theta_0\right)\sqrt{N}/\sqrt{[\theta_0(1-\theta_0]} \quad \left[\hat{\theta} \text{ est. max. ver.}\right],$$

excede qualquer número positivo previamente fixado. Aceitando a irrelevância da regra de paragem parece que é possível considerar o verdadeiro valor θ_0 como não plausível [referências adicionais podem colher-se em Joshi (1984)].

Os problemas suscitados pelo princípio de verosimilhança levaram Birnbaum (1962) a procurar obter a sua justificação lógica a partir de dois princípios aparentemente mais naturais e intuitivos: o princípio de suficiência e o princípio de condicionalidade.

O princípio de suficiência é bem conhecido[24]. No quadro do modelo $\mathcal{F}$ [veja-se (1.1)], se $T(X)$ é uma estatística suficiente para θ (em geral suficiente mínima) e se $T(x) = T(x')$, o princípio estabelece a identidade das inferências feitas a partir dos dados x ou dos dados x'. Por outras palavras, a evidência estatística fornecida por $T(x)$ é idêntica à fornecida pelos dados x. Recordando que uma condição necessária e suficiente para T ser suficiente para θ é verificar-se a decomposição,

$$f(x \mid \theta) = G[T(x), \theta]M(x), \quad [G \geqq 0, M \geqq 0],$$

para todo o $\theta \in \Theta$ e quase todo o x [excepto quando muito para um conjunto com medida à Lebesgue igual a zero], chega-se imediatamente à conclusão de que o princípio de suficiência implica o princípio de verosimilhança fraco.

O princípio de suficiência tem aceitação quase geral [veja-se, no entanto, Berger e Wolpert (1984)]. A existência de estatísticas suficientes mínimas de dimensão

[24] Está associado, como é natural, com o conceito de estatística suficiente. A estatística, $T(X)$, é suficiente para $\mathcal{F}$ — ou para θ quando não haja dúvida sobre o espaço Θ — quando a distribuição condicional de X dado $T = t$ não depende de θ [excepto, quando muito, para um conjunto de valores de T, seja A, tal que $P_\theta(T \in A) = 0$ para todo o $\theta \in \Theta$; veja-se em Fraser (1957) um tratamento rigoroso]. Uma famflia $\mathcal{F}$ admite em geral diversas estatísticas suficientes, nomeadamente a que consiste no próprio $X[\equiv (X_1, X_2, \ldots, X_N)$ no caso da amostragem casual], que é trivial e não conduz a qualquer redução dos dados. Como tal redução, sem perca de informação sobre θ, é o objectivo que Fisher teve em mente ao introduzir o conceito em 1921, interessa de facto escolher estatísticas suficientes que promovam a maior redução possível e que são designadas estatísticas suficientes mínimas. Estatística suficiente mínima é função de toda e qualquer estatística suficiente.
Por exemplo, se $(X_1, X_2, \ldots, X_N)$ é uma amostra casual, com $X_i \sim N(\theta, 1)$, é fácil de comprovar que,

$$\left(\sum_{i=1}^{i_0} X_i, \sum_{i=i_0+1}^{N} X_i\right), \quad 1 \leqq i_0 < N,$$

$$\left(\sum_{i=1}^{i_0} X_i, \sum_{i=i_0+1}^{i_1} X_i, \sum_{i=i_1+1}^{N} X_i\right), \quad 1 \leqq i_0 < i_1 < N,$$

$$\cdots$$

$$(X_1, X_2, \ldots, X_N)$$

são estatísticas suficientes para θ; no entanto, uma estatística suficiente mínima é, $T = \Sigma X_i$. Evidentemente, T é função das estatísticas acima. Se T' é função de T, $T' = \Psi(T)$, e a correspondência é biunívoca, T' é também estatística suficiente mínima.

fixa, independente da dimensão da amostra, permite uma apreciável redução dos dados sem perda de informação. Aliás, a fecundidade dos métodos clássicos depende não poucas vezes de tal existência que felizmente se verifica para uma vasta classe de modelos [por exemplo, modelos da família exponencial].

O princípio de condicionalidade diz respeito a experiências, E, definidas como mistura de experiências, E_m, $m = 1, 2, ...$, [seleccionadas com probabilidades, p_m, $\Sigma p_m = 1$, independentes de θ e das quais só uma é realizada] e estabelece que observar x por intermédio de E é equivalente a escolher m com probabilidade p_m e então observar x por intermédio de E_m.

Afirma o princípio de condicionalidade que as experiências componentes que não se realizaram são irrelevantes; por outras palavras, as conclusões a tirar da realização de E devem ser as conclusões a tirar da realização de E_m seleccionada pelo processo casual.

Exemplo 1.7 — *Continuação.* Trata-se de um caso típico de uma experiência, E, que é uma mistura de duas experiências, $E_1(\equiv I_1)$, escolhida com probalidade p e em que se observa $X \sim N(\theta, 100)$ [instrumento impreciso] e $E_2(\equiv I_2)$, escolhida com probabilidade $1 - p$ e em que se observa $X \sim N(\theta, 1)$ [instrumento preciso].

O princípio de condicionalidade estabelece em termos gerais que o procedimento correcto consiste em registar primeiro qual é o instrumento seleccionado, fazendo depois a observação com esse instrumento e desprezando completamente o que se poderia ter passado caso o outro instrumento tivesse sido seleccionado. Por exemplo, se o sorteio designou o segundo instrumento e a observação de $X \sim N(\theta, 1)$ deu $X = x$, só o valor de x é relevante nas inferências sobre θ; não interessa a observação que poderia ter sido feita se o primeiro instrumento tivesse sido designado.

Repare-se que o modo de actuar recomendado pelo princípio de condicionalidade vai contra a precisão inicial dos clássicos calculada antes de se proceder ao «sorteio» do instrumento. □

Considerando os dois princípios acabados de introduzir pode agora referir-se o resultado conhecido por Teorema de Birnbaum[25]: o princípio de suficiência mais o princípio de condicionalidade são equivalentes ao princípio de verosimilhança; quer dizer, os dois primeiros implicam o segundo e reciprocamente.

Esta proposição foi considerada de grande importância. Como o princípio de suficiência é largamente aceite e como se sabe que os frequencistas utilizam alguma forma de condicionalidade, parecia aberto o caminho para persuadir os clássicos a aderir em massa ao princípio de verosimilhança. No entanto, começaram a

[25] Birnbaum reformulou de maneira clara e concisa os resultados de 1962. O estudo de Birnbaum (1972) é particularmente gratificante mas não cabe no âmbito do presente trabalho.

aparecer objecções, quer por os raciocínios de Birnbaum se restringirem ao caso discreto (Joshi), quer por aspectos mais gerais (Durbin e Kalbfleish) — veja-se Joshi (1984) — que se não fizeram os adeptos do princípio de verosimilhança perder a fé, nem o teorema perder importância, explicam a não aceitação por parte de muitos estatísticos.

Independentemente de questões mais profundas que aqui não podem abordar-se, uma coisa é certa: os princípios de suficiência e de condicionalidade são incompatíveis com a teoria clássica quando tomados conjuntamente. Sendo algo pacífica a aceitação do princípio de suficiência qual é então a forma de condicionalidade que os frequencistas acolhem e empregam [com consequências por vezes ambíguas, Efron (1978)]?

A ideia de condicionalidade perfilhada pelos clássicos[26] conduz ao domínio da chamada inferência condicionada e tem íntima relação com o problema seguinte: é bem sabido que os procedimentos estatísticos são avaliados com base no comportamento que têm quando se procede à repetição indefinida do processo de amostragem ou experimentação; mas, sendo o estudo frequencista conduzido sobre um conjunto hipotético de repetições — conjunto de referência — sucede por vezes que é possível imaginar diferentes formas de repetir a amostragem ou experimentação, isto é, há vários conjuntos de referência susceptíveis de consideração [releia-se a pág. 21]. A opção por um conjunto de referência que não coincida com o espaço da amostra corresponde a uma inferência condicionada.

Exemplo 1.11 — Suponha-se X_i I.I.D., $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, N$; antes de iniciar a observação lança-se uma moeda «regular» para decidir se $N = 10$ ou se $N = 100$: $P(N = 10) = P(N = 100) = 1/2$. O objectivo é estimar μ a partir da observação de $X_1, X_2, \ldots, X_N$, supondo σ conhecido.

O estimador usual é $\overline{X} = \Sigma X_i/N$; como é bem sabido, a distribuição de $\overline{X}$ condicionada pelo valor determinado para N é, $\overline{X} \,|\, N \sim \mathrm{N}(\mu, \sigma^2/N)$.

O erro quadrático médio condicionado de $\overline{X}$ como estimador de μ calcula-se sem dificuldade,

$$E\left\{\left(\overline{X} - \mu\right)^2 \,|\, N\right\} = \begin{cases} \sigma^2/10 & \text{se } N = 10, \\ \sigma^2/100 & \text{se } N = 100. \end{cases} \tag{1.11}$$

O erro quadrático médio é então,

$$E\left\{\left(\overline{X} - \mu\right)^2\right\} = \frac{1}{2}(\sigma^2/10) + \frac{1}{2}(\sigma^2/100). \tag{1.12}$$

[26] Pode dizer-se que a condicionalidade perfilhada pelos clássicos é parcial enquanto (veja-se secção seguinte) a condicionalidade perfilhada pelos bayesianos é completa (para os bayesianos o conjunto de referência é $\{x\}$, i.e, contém apenas um elemento, a particular observação, x, efectuada).

Tanto o erro quadrático médio como o erro quadrático médio condicionado são conceitos frequencistas. O primeiro baseia-se na repetição da experiência que consiste em lançar repetidamente a moeda ao ar para escolher o valor de N, observando as N variáveis, $X_1, X_2, \ldots, X_N$, calculando depois $\overline{X}$ para estimar μ; o segundo baseia-se na repetição da experiência que consiste, com N fixo ($N = 10$ ou $N = 100$ consoante o resultado obtido com o lançamento da moeda que não volta a repetir-se), em observar as N variáveis, $X_1, X_2, \ldots, X_N$, calculando depois $\overline{X}$.

Que expressão usar, (1.11) ou (1.12), para medir a precisão de $\overline{X}$?

Na fase de planeamento da experiência, quando se deseja, por exemplo, optar entre a experiência mista já referida e uma outra experiência mista que consiste em escolher, também através do lançamento de uma moeda «regular», entre $N = 15$ e $N = 90$, procedendo em seguida, como anteriormente, à estimação de μ por meio de $\overline{X}$, parece defensável usar (1.12), comparando-a com a expressão análoga correspondente à segunda experiência.

Se o objectivo é fazer inferências sobre μ no quadro da primeira experiência, o erro quadrático médio, (1.12), apesar de nada ter de incorrecto, afigura-se irrelevante face aos valores de N e de $\overline{X}$ de facto observados: se $N = 10$, a expressão (1.12) é muito optimista; se $N = 100$, é muito pessimista. A inferência condicionada — emprego de (1.11) depois de conhecido N — é o procedimento clássico recomendável e não significa, repare-se, a aceitação do princípio de condicionalidade. □

A inferência condicionada é em geral apresentada em associação com o conceito de estatística ancilária ou subsidiária introduzido por Fisher em 1936. Uma primeira definição diz que $C(X)$ é estatística ancilária quando tem distribuição (marginal) independente do parâmetro em questão, seja θ. Esta definição é demasiado ampla; a seguinte é mais corrente: suponha-se que $(T, C) = [T(X), C(X)]$ é uma estatística suficiente mínima para θ, com $\dim\{(T, C)\} > \dim\{\theta\}$; se a distribuição marginal de C não depende de θ então C é uma estatística ancilária.

Uma estatística ancilária, C, não fornece directamente qualquer informação sobre θ pois, como se disse, a sua distribuição marginal é independente de θ. A informação primaria sobre θ é dada por T, fornecendo C apenas informação suplementar ou subsidiária[27].

[27] É interessante reflectir no seguinte [veja-se Lehmann (1983)]: uma estatística suficiente vai tanto mais longe na redução dos dados quanto menor for a «quantidade» de informação ancilária que contém. Assim, se, conforme se indicou, (T, C), é uma estatística suficiente mínima para θ e se $\dim\{(T, C)\}$ é maior do que $\dim\{\theta\}$, parece que se fica aquém da desejável redução dos dados. Um caso extremo passa-se, por exemplo, com a distribuição de Cauchy. Enquanto a estatística suficiente mínima tem a dimensão da amostra, pois é formada pelas N estatísticas de ordem, $[X_{(1)}, X_{(2)}, \ldots, X_{(N)}]$, o parâmetro tem dimensão um; em contrapartida a estatística suficiente mínima contém abundante informação subsidiária, por exemplo, as diferenças, $X_{(N)} - X_{(i)}$, $i = 1, 2, \ldots, N - 1$, são estatísticas ancilárias.

Designe, $f(t,c\,|\,\theta)$, a função de densidade conjunta de (T,C) e, $g(c)$, a função de densidade marginal de C; a relação,

$$f(t,c\,|\,\theta) = f(t\,|\,c,\theta)g(c), \quad \theta \in \Theta,$$

mostra que a informação sobre θ está contida no factor, $f(t\,|\,c,\theta)$. Assim, T é uma estatística condicionalmente suficiente para θ e as inferências, para os clássicos, devem fazer-se a partir de T depois de condicionar pelo valor observado para C, seja c. Por outras palavras, as conclusões, nomeadamente as que são baseadas no comportamento frequencista de T, são retiradas restringindo o espaço da amostra, $\mathcal{X}$, ao conjunto de pontos x tais que $C(x) = c$. Isto equivale a tomar como conjunto de referência $\{x : C(x) = c\} \subset \mathcal{X}$, e não $\mathcal{X}$ como acontece no problema não condicionado.

Exemplo 1.7 — *Continuação.* A experiência pode reformular-se nos seguintes termos: seja C uma variável aleatória que assume o valor $C = 0$ com probabilidade conhecida p e o valor $C = 1$ com probabilidade $1 - p$. Quando $C = 0$ observa-se $X \sim N(\theta, \sigma_0^2) \equiv N(\theta, 100)$; quando $C = 1$ observa-se $X \sim N(\theta, \sigma_1^2) \equiv N(\theta, 1)$. A função de verosimilhança para os dados (x, c) escreve-se,

$$f(x,c\,|\,\theta) = p^{1-c}(1-p)^c \left[1/\sigma_c\sqrt{2\pi}\right] \exp\{-(x-\theta)^2/2\sigma_c^2\},$$

devendo notar-se que a estatística suficiente mínima para θ é (X,C) e não somente X. Tem-se, $\dim\{(X,C)\} = 2 > \dim\{\theta\} = 1$; obviamente, C é uma estatística ancilária. É aceitavel para um clássico que as inferências sobre θ se façam a partir de X condicionalmente no valor obtido para C; assim, qualquer procedimento é avaliado em função de repetidas observações de X com C fixado no valor obtido, $C = 0$ *ou* $C = 1$. □

Exemplo 1.11 — *Continuação.* A estatística suficiente mínima para μ é $(\overline{X}, N)$ e não apenas $\overline{X}$. A dimensão da amostra, N, é uma estatística ancilária. As inferências sobre μ devem fazer-se a partir de $\overline{X}$ condicionalmente no valor obtido para N. O emprego de (1.11) tem suporte nesta ideia. □

As estatísticas suficientes completas (que são sempre suficientes mínimas embora a recíproca não seja verdadeira — T é estatística suficiente completa se é suficiente e se $E_\theta\{\varphi(T)\} = 0$ para todo o $\theta \in \Theta$ implica $P_\theta[\varphi(T) = 0] = 1$ para todo o $\theta \in \Theta$) são de facto as que permitem a redução de dados mais efectiva. Conforme o Teorema de Basu estabelece, se T é uma estatística suficiente completa, então, toda e qualquer estatística ancilária é independente de T, i.e., uma estatística suficiente completa não contém informação ancilária (em geral, $\dim\{T\} = \dim\{\theta\}$). O conceito de estatística suficiente completa volta a encontrar-se no capítulo 6 e pode estudar-se em Fraser (1957), Murteira (1980) ou Lehmann (1983).

Uma estatística ancilária pode muitas vezes interpretar-se como medida da precisão obtida nas inferências sobre θ feitas a partir da estatística condicionalmente suficiente, T. De facto, como os resultados da observação podem diferir grandemente no que diz respeito à quantidade de informação que fornecem sobre o parâmetro [recordem-se as amostras «felizes» e «infelizes» apontadas no Ex. 1.6], importa, nos problemas de inferência, ter em conta o peso informativo da amostra concreta recolhida. É isso que em regra se consegue através do valor obtido para a estatística ancilária.

Os Exs. 1.7 e 1.11 esclarecem bem a ideia que acaba de expor-se. No Ex. 1.7, quando a estatística ancilária, C, assume o valor 0, observa-se a variável $X \sim N(\theta, 100)$; quando assume o valor 1, observa-se $X \sim N(\theta, 1)$; no segundo caso a estimação de θ pode fazer-se com muito maior precisão. No Ex. 1.11 basta reparar na expressão (1.11); como é óbvio, neste caso, uma amostra de 100 permite maior precisão do que uma amostra de 10. O exemplo seguinte, retirado de Efron (1978), ilustra a mesma ideia.

Exemplo 1.12 — Têm-se duas amostras independentes, com a mesma dimensão, N: X_i I.I.D., $X_i \sim N(\mu, \sigma^2)$, Y_i I.I.D., $Y_i \sim N(\nu, \sigma^2)$, $i = 1, 2, \ldots, N$, e pretende estimar-se (μ, ν).

O estimador natural é, $(\overline{X}, \overline{Y})$; escreva-se,

$$\overline{X} \sim N(\mu, 1), \quad \overline{Y} \sim N(\nu, 1),$$

depois de tomar $\sigma^2/N = 1$ para facilitar a escrita sem perda de generalidade.

Quando se introduz a condição adicional de que o ponto (μ, ν) está sobre a circunferência de raio igual a 3 e centro na origem,

$$(\mu, \nu) = 3(\cos\theta, \operatorname{sen}\theta), -\pi < \theta \leqq \pi, \tag{1.13}$$

o problema passa a ser a estimação de θ.

Exprima-se $(\overline{X},\overline{Y})$ em coordenadas polares,

$$\hat{\theta} \equiv \operatorname{arctg}\left(\overline{Y}/\overline{X}\right), \quad R \equiv \sqrt{\left(\overline{X}^2 + \overline{Y}^2\right)}; \tag{1.14}$$

o estimador óbvio de θ é $\hat{\theta}$ (veja-se Fig. 1.1). Verifica-se que $\hat{\theta}$ é não enviesado, $E\{\hat{\theta}\} = \theta$, e tem erro quadrático médio (calculado por integração numérica),

$$E\left\{\left(\hat{\theta} - \theta\right)^2\right\} = 0{,}12. \tag{1.15}$$

O facto pouco óbvio para o qual Fisher chamou a atenção é o seguinte: R é uma estatística ancilária pois tem distribuição independente de θ [repare-se na simetria

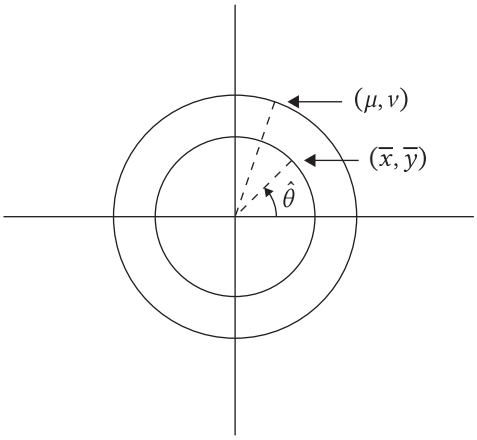


Fig. 1.1

circular da distribuição de $(\overline{X}, \overline{Y})$ em torno de (μ, ν)]; assim, a observação de R não dá qualquer informação directa sobre θ mas o seu valor determina a precisão de $\hat{\theta}$ como se indica no quadro abaixo [veja-se Efron (1978)]:

R	1,5	2	2,5	3	3,5	4	4,5	5
$E\left\{\left(\hat{\theta}-\theta\right)^2 \mid R\right\}$	0,26	0,18	0,14	0,12	0,10	0,09	0,08	0,07

Aceitando que a análise da precisão de $\hat{\theta}$ é conduzida condicionalmente no valor observado para R, a precisão dada por (1.15) é irrelevante. □

Não se pense que a inferência condicionada se apresenta sempre de forma clara. Na verdade.

1.º) não há qualquer método geral para a construção de estatísticas ancilárias;
2.º) em dado problema pode haver mais do que uma estatística ancilária e as inferências condicionadas dependem da que for usada como condicionante [Lindley (1971a)]. Basu sugere — veja-se Kiefer (1982) — que a dificuldade com a não unicidade pode residir na diferença entre uma experiência real ou «realizável» (como a observação fixando C ou N nos Exs. 1.7 e 1.11) e uma experiência conceitual ou «não realizável» (como a observação fixando R no Ex. 1.12) e propõe que se condicione no primeiro caso e não no segundo.

Outro aspecto interessante é referido por Kendall e Stuart (1967). Com $\theta = (\mu, \nu)$, μ e ν parâmetros escalares ou vectoriais, sendo μ o parâmetro de interesse e ν o parâmetro perturbador («*nuisance*» *parameter*), se (T, C) é estatística suficiente mínima para (μ, ν), suponha-se que C é ancilária para μ, i.e., tem distribuição

independente de μ. A factorização de $f(\mathbf{x} \mid \mu, \nu)$ pode assumir duas formas:

$$f(\mathbf{x} \mid \mu, \nu) = g_1(t \mid c, \mu, \nu) g_2(c \mid \nu) M(\mathbf{x}), \qquad [A]$$

ou,

$$f(\mathbf{x} \mid \mu, \nu) = g_1(t \mid c, \mu) g_2(c \mid \nu) M(\mathbf{x}). \qquad [B]$$

A forma [B] corresponde à situação em que C é suficiente para ν quando μ é dado; neste caso, é prática clássica corrente usar nas inferências sobre μ a estatística $T \mid C$ (condicionalmente) suficiente para μ. A forma [A] corresponde à situação em que não parece viável condicionar na estatística ancilária, C.

Exemplo 1.13 — Com $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D., $X_i \sim N(\mu, \sigma^2)$, pretende estimar-se μ com σ desconhecido. A situação é a seguinte: *(i)* com $\overline{X} = \Sigma X_i / N$, $S^2 = \Sigma (X_i - \overline{X})^2 / N$, a estatística $(\overline{X}, S^2)$ é suficiente mínima para (μ, σ^2); *(ii)* S^2 é estatística ancilária para μ porquanto, $NS^2/\sigma^2 \sim \chi^2(N-1)$ é independente de μ. No entanto, S^2 não é suficiente para o parâmetro perturbador σ^2, pois verifica-se sem dificuldade, trabalhando com a Normal,

$$f(\mathbf{x} \mid \mu, \sigma^2) = g_1(\overline{x} \mid s^2, \mu, \sigma^2) g_2(s^2 \mid \sigma^2) M(\mathbf{x}),$$

expressão que corresponde à forma [A]. As inferências sobre μ não se processam condicionando pelo valor obtido para S^2, mas, sim, empregando a distribuição t-Student de $\sqrt{N-1}(\overline{X} - \mu)/S$ [Kendall e Stuart (1967)][28]. □

Em resumo: se é verdade que os princípios de verosimilhança e de condicionalidade têm causado dificuldades aos seguidores de Neyman-Pearson, não é menos verdade que o princípio de verosimilhança tem limitações [Berger (1980)]:

Primeiro, refere-se exclusivamente às conclusões a tirar de uma experiência ou observação e não se aplica ao delineamento e planeamento de experiências em que o espaço da amostra tem necessariamente de ser considerado.

Segundo, aplica-se apenas às conclusões a tirar quando o modelo é válido ignorando portanto a fase de crítica. No caso do Ex. 1.6, em face da observação, $\min(x_i) = 3{,}1$, $\max(x_i) = 3{,}2$, numa amostra de 25 elementos, não é descabido pensar na inadequação do modelo uniforme, $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$.

Terceiro, é muito crítico de métodos de inferência, como o clássico, mas não ensina a utilizar a função de verosimilhança para efectuar inferências.

A matéria apresentada nas páginas anteriores permite já alguma reflexão sobre os métodos clássicos. Antes de fechar a secção não se resiste a citar a defesa do

[28] Como notam estes autores, o facto de no caso presente $\overline{X}$ e S^2 serem independentes é uma simplificação irrelevante para o argumento geral.

pensamento frequencista feita por Tiago de Oliveira (1981): «*E aqui está a diferença fundamental, o cerne da controvérsia: bayesianos, fiducialistas, verosimilhancistas, introduzem novos princípios (filosóficos) de inferência através das probabilidades a priori, da distribuição fiducial, da condicionalidade. A inferência clássica, agarrada à interpretação empirista das probabilidades, apenas tem de procurar critérios mais ajustados pois não perde o sentido do real. Nas outras inferências a escolha é a* gosto *num largo leque de princípios mas o real escapa-se*».

Abordam-se seguidamente alguns tópicos sobre a inferência bayesiana, o que não quer dizer que a inferência clássica não seja retomada aqui e ali, sobretudo para estabelecer o contraste.

1.4 Inferência Bayesiana

A semente para a abordagem bayesiana a problemas de inferência foi lançada por Richard Price quando em 1763 publicou a obra do Rev. Thomas Bayes intitulada «*An Essay Towards Solving a Problem in the Doctrine of Chances*» [veja-se a exposição de Turkman (1981)].

O Teorema de Bayes é uma proposição sobre probabilidades condicionadas indiscutível desde que se aceite, por exemplo, a axiomática de Kolmogorov. O que tem dado lugar a grande controvérsia é a sua aplicação a problemas de inferência estatística. Ocupa, de facto, lugar fulcral na inferência bayesiana.

Os métodos bayesianos passam por uma extensão do modelo clássico, extensão que tem raiz na seguinte divergência fundamental: o parâmetro θ, $\theta \in \Theta$, que no modelo clássico é um escalar ou vector desconhecido, mas fixo, passa a ser tomado como escalar ou vector aleatório (não observável); para os bayesianos, o que é desconhecido é incerto e toda a incerteza tem de ser quantificada em termos de probabilidade. Correlativamente admitem que a informação inicial ou a priori — anterior ou externa em relação à experiência mas demasiado importante para ser ignorada ou tratada *ad hoc* — pode traduzir-se formalmente[29] por uma distribuição de probabilidade (geralmente subjectiva) para θ, designada distribuição a priori. Assim, se θ é um parâmetro discreto, designando $h(\theta)$ a função de probabilidade a priori, tem-se que $h(\theta)$ exprime o grau de credibilidade[30] que o indivíduo que procede à analise atribui ao particular θ considerado; se θ é um parâmetro contínuo, designando $h(\theta)$ a função densidade de probabilidade a priori, tem-se que $h(\theta)d\theta$ exprime o grau de credibilidade que atribui ao intervalo $(\theta,\theta + d\theta)$. Note-se que a

[29] Em contraste pode talvez dizer-se que os clássicos atendem à informação a priori, quando muito. informalmente.

[30] A interpretação subjectiva da distribuição a priori não é a única perfilhada no campo bayesiano (veja-se a secção 1.6).

determinação e a interpretação da distribuição a priori se contam entre os pontos mais melindrosos da teoria bayesiana [veja-se adiante a secção 1.6][31].

A família $\mathcal{F}$ também faz parte do modelo bayesiano; quer dizer, a componente amostral ou experimental é comum aos modelos clássico e bayesiano, embora, eventualmente, com diferentes interpretações. Considerando um qualquer elemento de $\mathcal{F}$, seja $f(x\,|\,\theta)$, e a distribuição a priori, $h(\theta)$, o Teorema de Bayes para densidades conduz à relação,

$$h(\theta\,|\,x) = f(x\,|\,\theta)h(\theta)/\int_{\Theta} f(x\,|\,\theta)h(\theta)\,d\theta, \quad \theta \in \Theta, \tag{1.16}$$

onde $h(\theta\,|\,x)$ é a distribuição a posteriori de θ [quando Θ é discreto o integral do 2.º membro é substituído por um somatório]. Tendo em conta a observação, x, a atitude inicial em relação a θ, caracterizada por $h(\theta)$, é modificada, passando a nova atitude a traduzir-se por $h(\theta\,|\,x)$. No quadro bayesiano a distribuição a posteriori é elemento fundamental: serve de base a todas as inferências.

O denominador de (1.16) é a distribuição marginal de X,

$$f(x) = \int_{\Theta} f(x\,|\,\theta)h(\theta)\,d\theta, \quad x \in \mathcal{X}, \tag{1.17}$$

também chamada distribuição preditiva (vejam-se Exs. 1.14 e 1.15).

Como $f(x)$ não depende de θ, a relação (1.16) costuma escrever-se,

$$h(\theta\,|\,x) \propto f(x\,|\,\theta)h(\theta), \tag{1.18}$$

onde $\propto$ significa proporcional. Como a análise é condicionada pelo valor, x, observado, $f(x\,|\,\theta)$ é a função de verosimilhanca, isto é,

$$\{\text{dist. a posteriori} \propto \text{verosimilhança} \times \text{dist. a priori.}\}. \tag{1.19}$$

A função de verosimilhança tem importante papel na fórmula de Bayes pois representa o meio através do qual os dados, x, transformam o conhecimento a priori

[31] A discussão das distribuições a priori ilustra muito claramente aspectos importantes do confronto entre bayesianos e clássicos. Para os primeiros, como Berger (1984), «*The subjective choice of the model* [refere-se ao modelo experimental, $\mathcal{F}$] *is often a far more drastic use of prior information than is the use of prior distributions on parameters of the model*». Para os segundos, pelo menos na escrita de Lehmann (1983), há uma importante diferença na modelação de P_θ e na modelação de $h(\theta)$: «*Tipically, we have a number of observations from P_θ and can use these to check the assumption of the form of the distribution. Such a check of* Λ [h na notação aqui adoptada] *is not possible on the basis of one experiment because the value of θ under study represents only a single* observation *from this distribution*» [o sublinhado destina-se a destacar que o termo «observação» aqui empregado não significa que θ seja uma variável aleatória observável].

sobre θ; quer dizer, pode interpretar-se como expressão da informação sobre θ fornecida pelos dados, x.

O processo de revisão da distribuição a priori para obter a distribuição a posteriori esta esquematizado na Fig. 1.2. Nas «caixas» (1) e (2) indica-se que a distribuição a priori depende da informação inicial, em geral de tipo muito diverso: dados anteriormente recolhidos, conjecturas, análise subjectiva, observações acidentais, etc.; as «caixas» (3) e (4) exprimem que a informação adicional resultante da experiência entretanto realizada vai determinar a verosimilhança dos diferentes valores de θ; finalmente, nas «caixas» (5) e (6) refere-se que a combinação da distribuição a priori com a função de verosimilhança, operada pelo Teorema de Bayes, conduz à distribuição a posteriori.

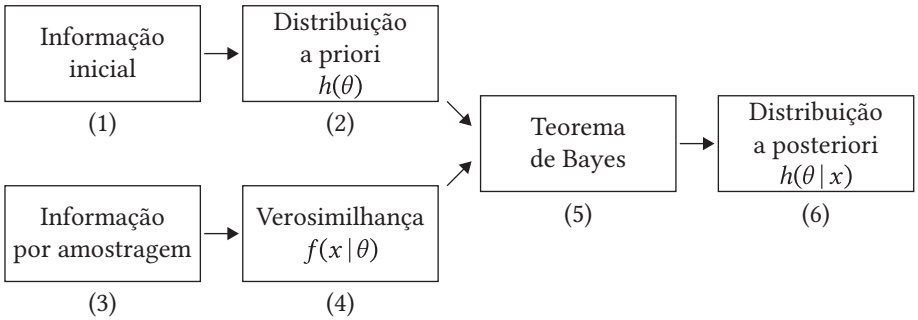


Fig. 1.2

Abra-se um parêntesis para apresentar alguns exemplos um tanto ou quanto formais. As dificuldades que cercam a determinação da distribuição a priori só adiante são afloradas (Secção 1.6).

Exemplo 1.14 — Uma experiência consiste na observação de uma variável aleatória, X, com distribuição binomial,

$$f(x\,|\,\theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}, \tag{1.20}$$

onde, $\mathcal{X} = \{0, 1, 2, \ldots, N\}, \Theta = \{\theta : 0 \leqq \theta \leqq 1\}$.

Se não há razão para distinguir entre os vários valores de θ «parece» natural — problema não pacífico retomado na pág. 72 — tomar uma distribuição a priori uniforme,

$$h(\theta) = \begin{cases} 1 & \text{para } \theta \in [0,1], \\ 0 & \text{para } \theta \notin [0,1]. \end{cases} \tag{1.21}$$

Por (1.17), a função de probabilidade preditiva de X é,

$$f(x) = \binom{N}{x} \int_0^1 \theta^x (1-\theta)^{N-x} d\theta = \binom{N}{x} B(x+1, N-x+1),$$

onde, $B(\alpha, \beta)$, é a função Beta definida pelo integral,

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du, \quad \alpha > 0, \ \beta > 0.$$

Notando que para α e β inteiros, $B(\alpha, \beta) = (\alpha-1)!(\beta-1)!/(\alpha+\beta-1)!$, tem-se,

$$f(x) = 1/(N+1), \quad X = 0, 1, 2, \ldots, N. \tag{1.22}$$

Como é intuitivo, se a distribuição a priori é (1.21), a distribuição preditiva associa uma probabilidade constante aos valores que é possível observar para X.

Supondo que se observa $X = x$, a distribuição a posteriori sai da aplicação de (1.16),

$$h(\theta \,|\, x) = (N+1)\binom{N}{x}\theta^x(1-\theta)^{N-x}, \quad 0 \leqq \theta \leqq 1, \tag{1.23}$$

que pode escrever-se,

$$h(\theta \,|\, x) = [B(x+1, N-x+1)]^{-1}\theta^x(1-\theta)^{N-x}, \quad 0 \leqq \theta \leqq 1. \tag{1.24}$$

Se a amostra for, por exemplo, constituída por N «sucessos», $X = N$, tem-se,

$$h(\theta \,|\, N) = (N+1)\theta^N, \quad 0 \leqq \theta \leqq 1.$$

Na Fig. 1.3 compara-se $h(\theta)$ com $h(\theta \,|\, N)$; a posição inicial é de indiferença, isto é, o reduzido conhecimento que se possui leva a considerar todos os valores de θ em pé de igualdade. Naturalmente, depois de obter N «sucessos» em N provas, os valores elevados de θ passam a ter muito mais credibilidade.

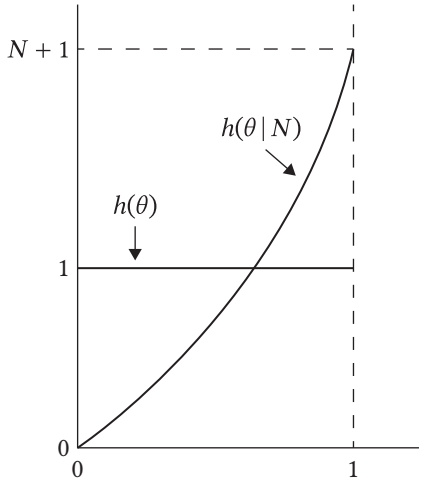


Fig. 1.3

Depois de observar $X \sim B(N;\theta)$ e obtido $X = x$, considere-se uma nova experiência do mesmo tipo em que se observa $Y \sim B(N;\theta)$, independente de X.

A expressão.

$$f(y\,|\,x) = \int_0^1 f(y\,|\,\theta)h(\theta\,|\,x)\,d\theta, \quad y = 0, 1, \ldots, N,$$

representa a função de probabilidade preditiva de Y condicionada por $X = x$. Fazendo a substituição de $f(y\,|\,\theta)$ e de $h(\theta\,|\,x)$ pelas respectivas expressões e integrando, obtém-se,

$$f(y\,|\,x) = \binom{N}{x}[B(x+y+1, 2N-x-y+1)][B(x+1, N-x+1)]^{-1}, \quad y = 0, 1, \ldots, N. \tag{1.25}$$

A função (1.25) permite atribuir a cada valor que é possível observar na segunda experiência, uma probabilidade condicionada pelo resultado da primeira experiência.

Suponha-se que saiu $X = N$ — na primeira experiência obtiveram-se N «sucessos» em N provas. Como caso particular de (1.25), vem,

$$f(y\,|\,N) = (N+1)\binom{N}{y}[B(N+y+1, N-y+1)] = \frac{(N+1)!\,(N+y)!}{(2N+1)!\,y!}, \quad y = 0, 1, \ldots, N. \tag{1.26}$$

Na Fig. 1.4 comparam-se, para $N = 5$, os valores saídos de (1.22) e de (1.26). Naturalmente, as predições dos valores da segunda experiência são muito diferentes das referentes à primeira; se na primeira experiência se obtiveram 5 «sucessos» em 5 provas, na segunda experiência predizem-se com mais elevada probabilidade, 3, 4 ou 5 «sucessos» (porquê, se as experiências são independentes?).

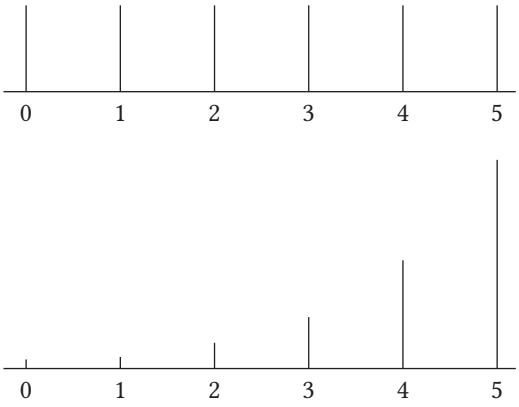


Fig. 1.4

□

Exemplo 1.15 — Ao investigar um novo medicamento pretende determinar-se, para um grupo de N pacientes, um limite inferior provável para o número de curas, seja $K \leqq N$, tal que,

$$P(X \geqq K) \geqq 0{,}95,$$

onde X é o número de curas.

Se a probabilidade de cura, θ, fosse conhecida, podia facilmente determinar-se K como sendo o maior inteiro para o qual se verifica,

$$\sum_{x=K}^{N} \binom{N}{x} \theta^x (1-\theta)^{N-x} \geqq 0{,}95.$$

Como θ é desconhecido há que ir por outro caminho. Se o investigador tem alguma ideia a priori sobre θ uma possível solução é dada pela distribuição preditiva.

Suponha-se que o investigador, dada a sua experiência, admite que θ não anda muito longe de 0,8 e que considera como distribuição a priori uma Beta com média 0,8 e variância 0,01. Dadas as relações que se verificam entre a média e a variância e os parâmetros da Beta (veja-se na pág. 76),

$$h(\theta) = [B(12,3)]^{-1}\theta^{11}(1-\theta)^2, \quad 0 \leqq \theta \leqq 1.$$

Como facilmente se calcula, a distribuição preditiva de X tem por expressão,

$$f(x) = \binom{N}{x} [B(12+x, 3+N-x)][B(12,3)]^{-1}, \quad x = 0, 1, \ldots, N.$$

Por exemplo, com $N = 10$, um simples cálculo numérico indica ser,

$$P(X \geqq 5) = 0{,}972, \quad P(X \geqq 6) = 0{,}925.$$

Há, portanto, uma elevada probabilidade de pelo menos 5 ou 6 curas (Zacks, 1981). □

Exemplo 1.16 — Considere-se a observação da variável aleatória $X \sim N(\theta, \sigma^2)$ com σ^2 conhecido e admita-se, a priori, que $\theta \sim N(\mu, \tau^2)$, com μ e τ^2 também conhecidos. A função de densidade conjunta do par (X, θ) é dada por,

$$\begin{aligned} h(x,\theta) &= f(x \mid \theta) h(\theta) \\ &= (2\pi\sigma\tau)^{-1} \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2}\right]\right\}. \end{aligned}$$

Introduzindo,

$$\rho = \frac{1}{\tau^2} + \frac{1}{\sigma^2},$$

e completando o quadrado do expoente, tem-se,

$$h(x,\theta) = (2\pi\sigma\tau)^{-1} \exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2\right\} \exp\left\{-\frac{(\mu - x)^2}{2(\tau^2 + \sigma^2)}\right\},$$

donde sai, por integração em relação a θ, a distribuição preditiva de X,

$$f(x) = (2\pi\rho)^{-1/2}(\sigma\tau)^{-1} \exp\left\{\frac{-(x-\mu)^2}{2(\tau^2+\sigma^2)}\right\}. \tag{1.27}$$

Por outro lado,

$$\begin{aligned} h(\theta \mid x) &= \frac{h(x,\theta)}{f(x)} \\ &= \left(\frac{\rho}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2\right\}. \end{aligned}$$

Quer dizer, $h(\theta \mid x)$ é ainda uma distribuição Normal com média,

$$\begin{aligned} \mu(x) &= \frac{1}{\rho}[(1/\tau^2)\mu + (1/\sigma^2)x] \\ &= \frac{(1/\tau^2)\mu + (1/\sigma^2)x}{(1/\tau^2) + (1/\sigma^2)} \\ &= x - \frac{\sigma^2}{\tau^2 + \sigma^2}(x - \mu), \end{aligned} \tag{1.28}$$

e variância,

$$\rho^{-1} = \frac{1}{(1/\tau^2) + (1/\sigma^2)}. \tag{1.29}$$

□

Exemplo 1.17 — Se em vez de X se observa uma amostra casual $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D., $X_i \sim N(\theta,\sigma^2)$, σ^2 conhecido, mantendo-se a distribuição a priori $h(\theta) \equiv N(\mu, \tau^2)$, é simples exercício, moldado no exemplo anterior, mostrar que $h(\theta \mid \mathbf{x})$ é ainda uma Normal com média,

$$\begin{aligned} \mu(\mathbf{x}) &= \frac{1}{\rho_N}[(1/\tau^2)\mu + (N/\sigma^2)\overline{x}] \\ &= \frac{(1/\tau^2)\mu + (N/\sigma^2)\overline{x}}{(1/\tau^2) + (N/\sigma^2)}, \end{aligned} \tag{1.30}$$

e vâriancia,

$$\rho_N^{-1} = \frac{1}{(1/\tau^2) + (N/\sigma^2)}. \tag{1.31}$$

Na expressão (1.30), $\mathbf{x}$ designa o vector amostra, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, e $\overline{x}$ a respectiva média, $\overline{x} = \Sigma x_i / N$.

Como é evidente, o presente exemplo generaliza o anterior; em particular, as expressões (1.28) e (1.29) são caso particular de (1.30) e (1.31), respectivamente, donde saiem quando se toma $N = 1$. Os comentários que vão fazer-se aplicam-se, portanto, aos dois exemplos.

Em primeiro lugar, note-se que sendo a distribuição a priori e a distribuição a posteriori da família Normal, a informação dada pela amostra (também proveniente de um universo Normal) reflecte-se exclusivamente sobre os parâmetros: comparem-se (1.30) com μ e (1.31) com τ^2.

Em segundo lugar, recorde-se que o inverso da variância também se designa por precisão; por exemplo, um método de observação tem tanto mais precisão quanto mais concentrados em torno do verdadeiro valor estejam os valores que produz, isto é, quanto menor for a variância das observações que produz. No caso presente podem considerar-se:

- precisão a priori $w' = 1/\tau^2$,
- precisão da amostragem (entenda-se a precisão da estatística suficiente, $\overline{x}$) $w'' = N/\sigma^2$,
- precisão a posteriori $w = \rho_N$.

Na nova notação tem-se a relação equivalente a (1.31),

$$w = w' + w'', \tag{1.32}$$

que estabelece que a precisão a posteriori é a soma da precisão a priori com a precisão da amostragem.

De (1.30) obtém-se,

$$\mu(\mathbf{x}) = \frac{w'\mu + w''\overline{x}}{w' + w''}, \tag{1.33}$$

isto é, a média da distribuição a posteriori é a média ponderada da média da distribuição a priori e da média da amostra. A ponderação da média a priori, $1/\tau^2$, é a precisão da credibilidade atribuída aos diferentes valores de θ, precisão que é grande se houver muita informação inicial sobre θ e pequena se a informação inicial for vaga ou difusa. A ponderação da média da amostra, N/σ^2, varia directamente com a dimensão da amostra, N, e com a precisão das observações individuais, $1/\sigma^2$.

Na Fig. 1.5 mostra-se graficamente a passagem de $h(\theta) \equiv N(\mu, \tau^2)$ para $h(\theta \mid \mathbf{x}) = N[\mu(\mathbf{x}), \rho_N^{-1}]$, pondo em destaque a mudança de localização e a redução na variância decorrentes da informação amostral.

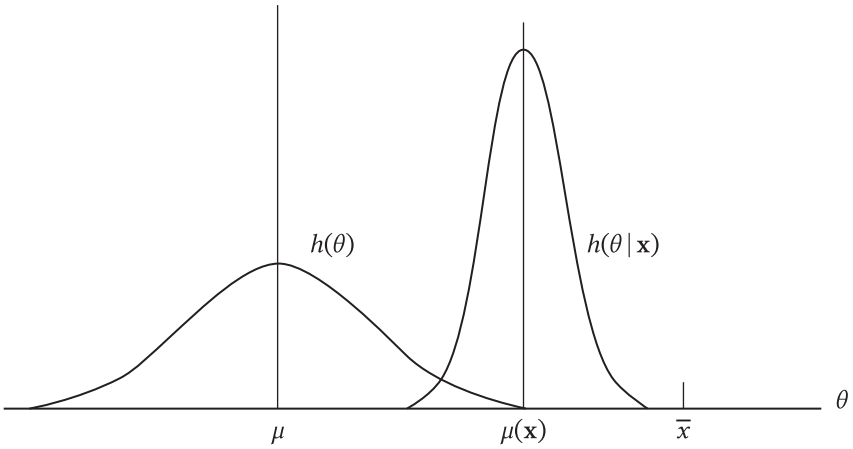


Fig. 1.5

Finalmente faça-se, $w' = N'/\sigma^2$, isto é, suponha-se que a precisão a priori é equivalente à precisão (da média) de uma amostra de N' observações. Nessa hipótese,

$$\mu(\mathbf{x}) = \frac{N'\mu + N\overline{x}}{N' + N},$$
$$\rho_N^{-1} = \frac{\sigma^2}{N' + N},$$

relações que permitem interpretar a distribuição a priori e a informação que fornece em termos de paridade com a informação dada pela amostra. De, $N'/\sigma^2 = 1/\tau^2$, sai $N' = \sigma^2/\tau^2$; quer dizer, a informação a priori equivale à informação dada por uma amostra (artificial) de dimensão σ^2/τ^2. □

Deixando, por agora, o formalismo bayesiano, retome-se o problema da inferência. Segundo os bayesianos a distribuição a posteriori incorpora, por via do Teorema de Bayes, toda a informação disponível sobre o parâmetro (informação inicial mais informação amostral). Daqui decorre que todos os procedimentos da inferência bayesiana são baseados exclusivamente em $h(\theta \mid x)$. Antes de passar a explorar as consequências desta perspectiva, repare-se na seguinte vantagem reclamada pelos bayesianos: as inferências baseadas na teoria da amostragem são mais restritas do que as bayesianas por usarem exclusivamente informação amostral. Os procedimentos bayesianos, porque recorrem também à informação a priori, quantificada pela distribuição a priori, produzem inferências mais informativas [recorde-se (1.32)] desde que essa quantificação seja correcta. Por outras palavras, os bayesianos sustentam que os seus métodos requerem menos informação

amostral para atingir o mesmo nível qualitativo dos métodos clássicos correspondentes. Ultrapassadas questões de princípio, este facto pode constituir motivação para optar pelos primeiros quando os dados da amostragem forem dispendiosos ou difíceis de obter, como sucede, por exemplo, nos problemas de fiabilidade.

Para estabelecer um primeiro contraste entre as inferências clássicas e bayesianas importa notar que as segundas são baseadas em probabilidades ou credibilidades associadas com diferentes valores do parâmetro, θ, que poderiam ter dado lugar à amostra, x, que de facto foi recolhida (veja-se Fig. 1.6). Pelo contrário, as primeiras são baseadas em probabilidades associadas com as diferentes amostras, x, que poderiam ocorrer para algum valor fixo, mas desconhecido, do parâmetro, θ (veja-se Fig. 1.7 e recorde-se o conceito de distribuição por amostragem).

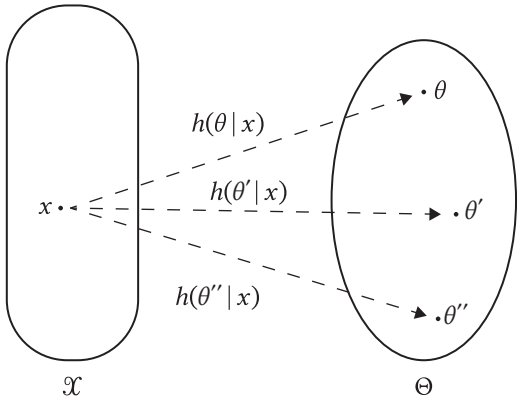


Fig. 1.6

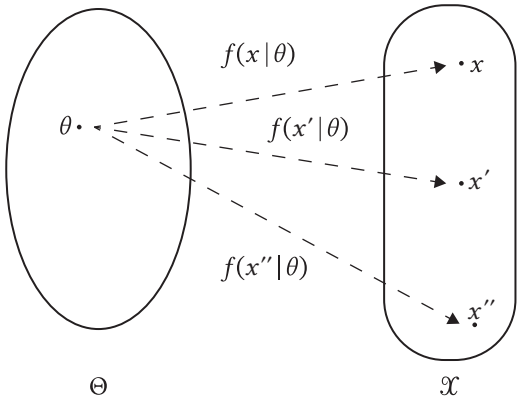


Fig. 1.7

Em muitos problemas paramétricos a variável ou vector experimental tem distribuição, $f(x \mid \theta)$, onde $\theta = (\mu, \nu)$, com ν parâmetro perturbador. Quando existem estatísticas suficientes para todos os parâmetros, a presença de parâmetros perturbadores não levanta, em regra, grandes problemas[32]. Se tal não sucede, parece haver manifesta vantagem na abordagem bayesiana. De facto, sendo,

$$h(\theta \mid x) \equiv h(\mu,\nu \mid x),$$

o parâmetro perturbador, ν, pode eliminar-se por integração sobre o respectivo domínio,

$$h(\mu \mid x) = \int h(\mu, \nu \mid x)\, d\nu,$$

dando lugar à distribuição marginal do parâmetro relevante, μ, com a qual pode avançar-se no trabalho inferencial.

A inferência bayesiana aceita, tal como a clássica, o princípio de suficiência.

Se $T(X)$ é estatística suficiente para θ, tem-se como condição necessária e suficiente a já referida factorização,

$$f(x \mid \theta) = G[T(x), \theta]M(x);$$

consequentemente,

$$h(\theta \mid x) = KG[T(x), \theta]h(\theta), \tag{1.34}$$

com,

$$K = \left\{ \int_{\Theta} G[T(x), \theta]\ h(\theta)\, d\theta \right\}^{-1},$$

independente de θ. Logo, se x e x' são duas amostras conducentes ao mesmo valor de T, conclui-se imediatamente que,

$$T(x) = T(x') \Rightarrow h(\theta \mid x) = h(\theta \mid x') \text{ para todo o } \theta \in \Theta; \tag{1.35}$$

quer dizer, x e x' conduzem às mesmas inferências bayesianas.

Equivalentemente,

$$h(\theta \mid x) = h[\theta \mid T(x)] \text{ para todo o } \theta \in \Theta, \tag{1.36}$$

porquanto, a verificar-se (1.35), $h(\theta \mid x)$ depende de x unicamente através de $T(x)$. A relação (1.36) pode empregar-se, aliás, para exprimir a suficiência bayesiana da

[32] Há mesmo assim casos — por exemplo, o problema de Behrens-Fisher — em que é difícil a eliminação dos parâmetros pertubadores.

estatística $T(X)$. Como foi dado verificar a suficiência fisheriana implica a suficiência bayesiana; a recíproca também é verdadeira [para uma demonstração rigorosa veja-se Zacks (1971)].

Ao contrário do que acontece na análise clássica, a eficácia da análise bayesiana não depende da existência de estatísticas suficientes mínimas de pequena dimensão. O motivo é simples: é quase tão fácil partir de $h(\theta \mid x)$ como partir de $h[\theta \mid T(x)]$. Semelhante propriedade resulta de a análise bayesiana se concentrar na distribuição do parâmetro condicionada por uma amostra concreta.

A inferência bayesiana aceita, ao invés da clássica, o princípio de verosimilhança, sustentando alguns autores que é a única via para a implementação deste princípio[33]. Se os dados x e x' conduzem às mesmas conclusões bayesianas, tem-se,

$$h(\theta \mid x) = h(\theta \mid x') \Rightarrow f(x \mid \theta) \propto f(x' \mid \theta),$$

e reciprocamente. Chega-se fatalmente a idêntico resultado recordando que na determinação de $h(\theta \mid x)$ entra exclusivamente a informação dada pela amostra particular, x, veiculada através da função de verosimilhança.

Entrando propriamente nos métodos inferenciais considere-se em primeiro lugar a estimação bayesiana. Um procedimento que imediatamente ocorre consiste em propor a estimação de θ por meio da moda da distribuição a posteriori. Assim, segundo o método da máxima verosimilhança generalizada, depois de observar, x, deve seleccionar-se o valor de θ com probabilidade ou credibilidade máxima.

Exemplo 1.18 — Quando se observa uma amostra casual, X_i I.I.D., $X_i \sim N(\theta, \sigma^2)$, $i = 1, 2, \ldots, N$, σ^2 conhecido, e θ tem distribuição a priori,

$$h(\theta) \equiv N(\mu, \tau^2),$$

obtém-se,

$$h(\theta \mid \mathbf{x}) \equiv N(\mu(\mathbf{x}), \rho_N^{-1}),$$

com $\mu(\mathbf{x})$ e ρ_N^{-1} dados por (1.30) e (1.31). A moda (= média = mediana) de $h(\theta \mid \mathbf{x})$ é, portanto, $\mu(\mathbf{x})$, que representa a estimativa da máxima verosimilhança generalizada de θ. □

Exemplo 1.19 — Considere-se,

$$f(x \mid \theta) = e^{-(x-\theta)}, \quad x \geqq \theta,$$

com,

$$h(\theta) = [\pi(1 + \theta^2)]^{-1}, \quad -\infty < \theta < +\infty.$$

[33] O que não merece o acordo dos verosimilhancistas puros que não aceitam o emprego de distribuições a priori nos processos inferenciais que propõem.

Consequentemente,

$$h(\theta \mid x) = \frac{e^{-(x-\theta)}}{\pi(1+\theta^2)f(x)}, \quad \theta \leqq x, \quad f(x) = \int_{-\infty}^{x} e^{-(x-\theta)}[\pi(1+\theta^2)]^{-1}d\theta.$$

Derivando, vem,

$$\begin{aligned}\frac{d}{d\theta}h(\theta \mid x) &= \frac{e^{-x}}{\pi f(x)}[e^{\theta}(1+\theta^2)^{-1} - 2\theta e^{\theta}(1+\theta^2)^{-2}] \\ &= \frac{e^{-x}}{\pi f(x)}\left[\frac{e^{\theta}(\theta-1)^2}{(1+\theta^2)^2}\right].\end{aligned}$$

Como esta derivada é sempre positiva, $h(\theta \mid x)$ é crescente para $\theta \leqq x$. O máximo é obtido com $\hat{\theta} = x$ que é, assim, a estimativa da máxima verosimilhança generalizada de θ [Berger (1980)]. □

Exemplo 1.20 — Observa-se $X \sim N(\theta, \sigma^2)$, com σ^2 conhecido e θ uma medida de uma quantidade positiva. A estimativa clássica de θ é x que se revela manifestamente inconveniente pois pode ser menor do que zero. Uma estimativa alternativa e mais coerente pode encontrar-se no quadro bayesiano. Tomando a distribuição a priori, $h(\theta)d\theta \propto d\theta$, $\theta > 0$ (consulte-se a secção seguinte no que diz respeito a distribuições a priori não informativas impróprias), vem,

$$h(\theta \mid x) = e^{-(\theta-x)^2/2\sigma^2}[f(x)]^{-1}, \quad \theta > 0,$$

onde,

$$f(x) = \int_0^{\infty} e^{-(\theta-x)^2/2\sigma^2}d\theta.$$

A estimativa da máxima verosimilhaça generalizada é $\hat{\theta} = x$ se $x > 0$ ou $\hat{\theta} = 0$ se $x \leqq 0$. Uma estimativa mais conveniente é a que consiste em tomar a média da distribuição a posteriori,

$$E\{\theta \mid x\} = \int_0^{\infty} \theta h(\theta \mid x)\, d\theta,$$

que depois de alguns cálculos — veja-se Berger (1980) — se verifica ter a expressão,

$$E\{\theta \mid x\} = x + \frac{(2\pi)^{-1/2}\sigma\, e^{-x^2/2\sigma^2}}{P[U > -(x/\sigma)]},$$

onde $U \sim N(0,1)$. $E\{\theta \mid x\}$ é maior do que zero por ser o valor médio de uma variável aleatória que só assume valores positivos. □

Como o exemplo anterior deixa antever, outras estimativas bayesianas eventualmente recomendadas são a média e a mediana de $h(\theta \mid x)$. Berger (1980) propõe que se calcule sempre a moda, a média e a mediana, procurando analisar a sua sensibilidade — robustez — face a modificações na distribuição a priori. No entanto, em princípio, a moda é o estimador bayesiano, posição muito distinta da dos clássicos, habituados a dar atenção a uma multiplicidade de estimadores.

Se a informação inicial é de tal modo vaga que leva a tomar uma distribuição a priori, $h(\theta) =$ constante [pelo menos em subdomínio de Θ relevante; veja-se secção 1.6], tem-se,

$$h(\theta \mid x) \dot{\propto} f(x \mid \theta), \quad \theta \in \Theta,$$

onde $\dot{\propto}$ significa aproximadamente proporcional. As estimativas da máxima verosimilhança dos clássicos e as estimativas da máxima verosimilhança generalizada dos bayesianos são nessa hipótese praticamente coincidentes. A respectiva interpretação é, porém, fundamentalmente diferente:

- no primeiro caso, a estimativa é o valor de θ que torna mais provável a amostra observada, x;
- no segundo caso, a estimativa é o valor de θ que, dado x, tem mais verosimilhança ou credibilidade.

Dois aspectos do método da máxima verosimilhança generalizada são dignos de citação:

[1] Com $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, suponha-se que o estimador da máxima verosimilhança generalizada é $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_k^*)$,

$$h(\theta^* \mid x) = \sup_\theta h(\theta \mid x);$$

se $\overline{\theta}_i$ é o estimador da máxima verosimilhança generalizada de θ_i determinado a partir da distribuição marginal de θ_i, $h(\theta_i \mid x)$, não se verifica em geral ser $\theta_i^* = \overline{\theta}_i$, embora para amostras grandes as diferenças possam ser pequenas[34].

[2] Contrariamente ao que sucede com a máxima verosimilhança, a máxima verosimilhança generalizada não fornece estimativas invariantes em relação a transformações paramétricas. Designando θ^* a estimativa da máxima verosimilhança generalizada de θ, se houver uma reparametrização, $\xi = \xi(\theta)$, não se segue que seja, $\xi^* = \xi(\theta^*)$. Assim, diferentes inferências podem resultar de diferentes parametrizações, facto que constitui um ponto fraco. A questão é por vezes iludida com a afirmação de que existe uma parametrização «natural» e que as inferências devem fazer-se a partir de tal parametrização [Barnett (1982)].

[34] Veja-se DeGroot (1970).

Além da estimação por pontos os clássicos procedem também, não raras vezes conjuntamente, à estimação por intervalos ou regiões. Na óptica bayesiana tem-se intervalos ou regiões de credibilidade e não intervalos ou regiões de confiança.

Observado, x,

$$R_\alpha(x) \subset \Theta,$$

é uma região de credibilidade a $(1-\alpha)100\%$ para o parâmetro θ quando se verifica,

$$\int_{R_\alpha(x)} h(\theta \mid x)\, d\theta = (1-\alpha), \quad 0 < \alpha < 1, \tag{1.37}$$

onde $(1-\alpha)$ é coeficiente de credibilidade ou nível de credibilidade consoante na relação acima se consiga manter o sinal de igualdade ou se consiga apenas uma região, $R_\alpha(x)$, satisfazendo,

$$P[\theta \in R_\alpha(x) \mid x] \geqq 1-\alpha.$$

Esta última situação ocorre quando o espaço do parâmetro é discreto e é paralela à que se depara aos clássicos quando a variável aleatória observada, X, é discreta.

Que diferenças podem encontrar-se entre as práticas seguidas pelas duas correntes?

Em relação a α a prática seguida coincide com a dos clássicos: toma-se α pequeno, 0,10, 0,05 ou 0,01, de modo a obter regiões de credibilidade a 90%, 95% ou 99%. Nos restantes aspectos o contraste é notável: «*Since the posterior distribution is an actual probability distribution on* Θ*, one can speak meaninfully (though usually subjectively) of the probability that* θ *is in* C [$R_\alpha(x)$ na nossa notação]. *This is in contrast to classical confidence regions, which can only be interpreted in terms of probability of coverage (the probability that the sample will be such that the resulting confidence region contains* θ*). Note that the classical probability of coverage is a measure of* initial precision *while the Bayesian coverage probability is a measure of* final precision. *After observing* x*, the Bayesian feels* θ *has probability at least* $1-\alpha$ *of being in* C». [Berger (1980), excepto o sublinhado].

A definição (1.37) possui um inconveniente: a região $R_\alpha(x)$ não é única, podendo muito bem suceder que valores de θ incluídos em $R_\alpha(x)$ tenham menor probabilidade ou credibilidade que valores de θ não incluídos em $R_\alpha(x)$. Assim, para proceder à escolha de uma particular região ao mesmo tempo que se minimiza o seu tamanho prefere-se trabalhar com regiões de credibilidade H.P.D. (com «*highest posterior density*»),

$$R^*_\alpha(x) = \{\theta \in \Theta : h(\theta \mid x) \geqq k(\alpha)\} \tag{1.38}$$

onde $k(\alpha)$ é o maior número real tal que,

$$P[\theta \in R^*_\alpha(x) \mid x] \geqq 1-\alpha. \tag{1.38'}$$

Quando $\theta \in \Theta \equiv \mathbf{R}$, o conceito correspondente é o de intervalos de credibilidade H.P.D.. No caso mais corrente em que $h(\theta \mid x)$ é unimodal podem sempre determinar-se, dada a amostra x, dois valores de θ, sejam $\theta_1(x)$ e $\theta_2(x)$, tais que,

$$\int_{-\infty}^{\theta_1} h(\theta \mid x)\, d\theta = \alpha_1, \quad \int_{\theta_2}^{\infty} h(\theta \mid x)\, d\theta = \alpha_2.$$

Assim, com $\alpha_1 + \alpha_2 = \alpha$, vem,

$$P[\theta_1(x) < \theta < \theta_2(x) \mid x] = \int_{\theta_1}^{\theta_2} h(\theta \mid x)\, d\theta = 1 - \alpha,$$

exigindo-se ainda que seja $h(\theta \mid x) > h(\theta' \mid x)$ para todo o $\theta \in [\theta_1(x), \theta_2(x)]$ e todo o $\theta' \notin [\theta_1(x), \theta_2(x)]$ de modo a que qualquer ponto excluído do intervalo não possua maior credibilidade que qualquer ponto do mesmo intervalo.

Exemplo 1.21 — Retome-se o Ex. 1.17 e considere-se, observado $\mathbf{x}$, a construção de um intervalo de credibilidade a 95% para θ. Tem-se imediatamente que,

$$P\left[\mu(\mathbf{x}) - 1{,}96\rho_N^{-1/2} < \theta < \mu(\mathbf{x}) + 1{,}96\rho_N^{-1/2} \mid \mathbf{x}\right] = 0{,}95;$$

quer dizer, verifica-se, com probabilidade ou credibilidade 0,95, que θ cai no intervalo,

$$\frac{[(1/\tau^2)\mu + (N/\sigma^2)\overline{x}]}{[(1/\tau^2) + (N/\sigma^2)]} \pm 1{,}96 \frac{1}{[(1/\tau^2) + (N/\sigma^2)]^{1/2}}.$$

É curioso notar que se a informação a priori for insignificante pode tomar-se $\tau^2 \to \infty$ (isto é, pode pensar-se que a precisão a priori tende para zero) sendo então praticamente coincidentes o intervalo de credibilidade bayesiano e o intervalo de confiança clássico,

$$\overline{x} \pm 1{,}96 \frac{\sigma}{\sqrt{N}}.$$

As interpretações, contudo, são muito diferentes. Enquanto para um bayesiano, $P[\overline{x} - 1{,}96\sigma/\sqrt{N} < \theta < \overline{x} + 1{,}96\sigma/\sqrt{N} \mid \mathbf{x}]$, tem uma interpretação directa como probabilidade, representando portanto a precisão final, para um clássico, antes da recolha da amostra, $P[\overline{X} - 1{,}96\sigma/\sqrt{N} < \theta < \overline{X} + 1{,}96\sigma/\sqrt{N}]$, é a probabilidade de cobertura de θ por um intervalo aleatório (precisão inicial); depois da recolha da amostra a única coisa que um clássico pode dizer é que das duas uma: (i) tem-se $\overline{x} - 1{,}96\sigma/\sqrt{N} < \theta < \overline{x} + 1{,}96\sigma/\sqrt{N}$ ou (ii) um acontecimento de probabilidade 0,05 ocorreu[35].

[35] Suponha-se $N = 100$, $\sigma = 10$ e que saiu uma amostra com $\overline{x} = 20$. O consequente intervalo de confiança a 95%, $18{,}04 < \theta < 21{,}96$, só faz sentido, no dizer dos bayesianos, quando interpretado em termos de credibilidade, pois a interpretação frequencista só é válida inicialmente, isto é, antes de conhecido o resultado da amostra.

Este exemplo tem uma grande virtude que é a de mostrar que, para algumas distribuições (no caso presente, a Normal), a precisão inicial coincide com a precisão final: em ambos os casos é 0,95. Segundo os bayesianos é esta feliz coincidência que tem permitido a sobrevivência dos procedimentos clássicos ... □

Exemplo 1.22 — Considere-se uma amostra casual, X_i I.I.D., $i = 1, 2, \ldots, N$, com função densidade de Cauchy,

$$f(x \mid \theta) = \{\pi[1 + (x - \theta)^2]\}^{-1}, \quad -\infty < x < +\infty, -\infty < \theta < +\infty,$$

e tome-se a distribuição a priori, $h(\theta)\, d\theta \propto d\theta$ [trata-se, mais uma vez, de uma distribuição imprópria; confronte-se com a secção seguinte]. Tem-se,

$$h(\theta \mid \mathbf{x}) = \frac{\Pi[1 + (x_i - \theta)^2]^{-1}}{\int_{-\infty}^{+\infty} \Pi[1 + (x_i - \theta)^2]^{-1}\, d\theta}.$$

Esta distribuição a posteriori, sem ser muito atraente, permite em computador obter regiões de credibilidade para θ. Pelo contrário, no âmbito dos procedimentos clássicos, não se afigura clara a forma de construir regiões de confiança para θ, pois a estatística suficiente mínima tem dimensão igual à da amostra e é constituída tipicamente pelas N estatísticas de ordem. Este exemplo serve assim, também, para enfatizar a dependência em que os procedimentos clássicos se encontram da existência de estatísticas suficientes mínimas com dimensão conveniente [Berger (1980)]. □

Exemplo 1.22 — *Continuação.* Com $N = 2$, vem,

$$f(\mathbf{x} \mid \theta) = 1/\pi^2[1 + (x_1 - \theta)^2][1 + (x_2 - \theta)^2];$$

mantendo a mesma distribuição a priori, tem-se,

$$h(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta),$$

função de que se encontra um exemplo na Fig. 1.8.

Quando $|x_2 - x_1| > 2$, $h(\theta \mid \mathbf{x})$ é uma distribuição bimodal[36], situação algo desconfortável, tanto para bayesianos como para clássicos, no que diz respeito

[36] Note que $(X_2 - X_1)$ é uma estatística ancilária, pois tem distribuição independente de θ:

$$X_2 - X_1 = (X_2 - \theta) + (X_1 - \theta),$$

e $Z = (X_i - \theta)$ tem função de densidade, $f(z) = 1/\pi(1 + z^2)$. Como tal não fornece qualquer informação primária sobre θ; no entanto, como se viu, condiciona a «configuração» da amostra, isto é, a forma da função de verosimilhança.

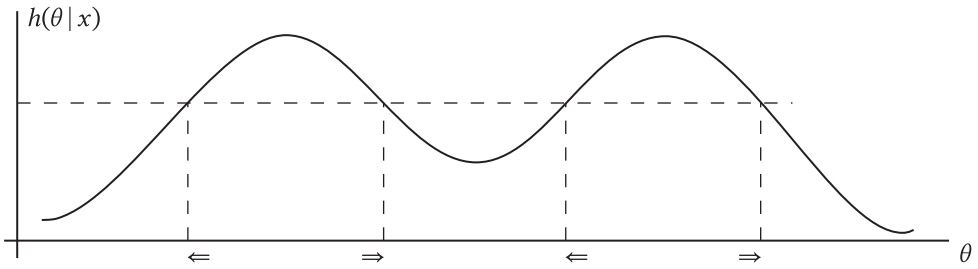


Fig. 1.8

à estimação pontual ou por intervalos. Como pode ver-se pela Fig. 1.8, que corresponde a essa situação, é-se conduzido a regiões de credibilidade representadas pela união de intervalos disjuntos e existem duas estimativas da máxima verosimilhança (generalizada ou não). [Barnett (1982)]. □

Considere-se uma dicotomia no espaço dos parâmetros,

$$\Theta = \Theta_0 \cup \Theta_1, \ \Theta_0 \cap \Theta_1 = \emptyset,$$

e suponha-se que o objectivo é ensaiar,

$$H_0\colon \theta \in \Theta_0 \text{ contra } H_1\colon \theta \in \Theta_1.$$

No campo clássico o ensaio[37] processa-se através da delimitação de uma região crítica no espaço da amostra,

$$W \subset \mathfrak{X},$$

que é avaliada em termos das probabilidades dos erros de 1.ª espécie e de 2.ª espécie, respectivamente, probabilidade de rejeitar H_0 quando é verdadeira,

$$P(X \in W \mid \theta), \quad \theta \in \Theta_0,$$

e probabilidade de aceitar H_0 quando é falsa,

$$P(X \notin W \mid \theta), \quad \theta \in \Theta_1.$$

[37] Na óptica de Newman-Pearson porquanto Fisher prefere os testes ou ensaios de significância aos ensaios de hipóteses. Tem-se aqui outro aspecto marcante da clivagem existente na escola clássica com Fisher em oposição a Neyman-Pearson a defender intransigentemente os testes de significância que considera mais informativos e possuem, segundo ele, a vantagem de dispensar a introdução de hipóteses alternativas. Quer dizer, Fisher ignora os erros de 2.ª espécie e volta ostensivamente as costas ao conceito de função potência de primordial importância na doutrina de Neyman-Pearson. (veja-se Capítulo 7).
Na análise comparativa entre a posição clássica e a bayesiana faz-se, de certo modo, tábua rasa desta clivagem.

Estas probabilidades, que dependem de θ, exprimem a precisão inicial (o problema do ensaio de hipóteses vai ser retomado no capítulo 7 no quadro da teoria da decisão estatística).

No campo bayesiano o problema é atacado de forma mais directa calculando as probabilidades a posteriori, $P(\Theta_0 \mid x)$ e $P(\Theta_1 \mid x)$, e decidindo depois de as comparar. Por exemplo, optando por H_0 se,

$$P(\Theta_0 \mid x) > P(\Theta_1 \mid x) = 1 - P(\Theta_0 \mid x),$$

isto é, se,

$$P(\Theta_0 \mid x) > \frac{1}{2}.$$

Estas probabilidades exprimem a precisão final do procedimento. A relação $P(\Theta_0 \mid x)/P(\Theta_1 \mid x)$, designa-se por rácio das «chances» a posteriori; quanto maior for o rácio mais plausível é H_0 em relação a H_1. Por vezes pode haver interesse em fazer a comparação com o rácio das «chances» a priori,

$$\frac{P(\Theta_0)}{P(\Theta_1)} = \frac{\int_{\Theta_0} h(\theta)\, d\theta}{\int_{\Theta_1} h(\theta)\, d\theta}.$$

No contexto do ensaio de hipóteses estatísticas pode talvez dizer-se que o resultado da inferência bayesiana não é a aceitação ou a rejeição mas sim a alteração da credibilidade que lhes é atribuída.

Exemplo 1.23 — Com X_i I.I.D., $X_i \sim N(\theta, \sigma^2)$, $\sigma^2 = 400$, $i = 1, 2, \ldots, N$, considere-se o ensaio da hipótese H_0: $\theta \leqq 100$ contra a alternativa H_1: $\theta > 100$.

A partir da informação dada por uma amostra concreta, a análise clássica é conduzida, grosso modo, nos termos seguintes. Determina-se o nível de significância, α, isto é, a probabilidade de obter um resultado tão bom ou pior do que o observado assumindo que a hipótese H_0 é verdadeira; quando menor for α menos plausível é H_0.

Suponha-se $N = 4$ e que se obteve uma amostra com média $\overline{x} = 106$. Dada a natureza unilateral direita de H_1, tem-se o nível de significância,

$$\begin{aligned}\alpha &= P(\overline{X} \geqq 106 \,|\, \theta = 100) = P[(\overline{X} - 100)/10 \geqq (106 - 100)/10] \\ &= P(U \geqq 0{,}6) = 1 - \Phi(0{,}6) = 0{,}274,\end{aligned}$$

onde $U \sim N(0, 1)$ e $\Phi(u)$ é a respectiva função de distribuição. A decisão a tomar face ao valor $\alpha = 0{,}274$ depende da atitude do investigador, embora para os limiares de significância «habituais» — 0,10, 0,05 e 0,01 — a hipótese H_0 não seja de rejeitar.

Na análise bayesiana as coisas passam-se de modo diferente. Suponha-se que a distribuição a priori é não informativa (caso $\tau^2 \to \infty$ do Ex. 1.17). Tem-se, $h(\theta \mid \mathbf{x}) \equiv N(\overline{x}, \sigma^2/N) \equiv N(106{,}100)$, donde,

$$\begin{aligned} P(H_0 \mid \mathbf{x}) &= P(\theta \leqq 100 \mid \overline{x} = 106) \\ &= P[(\theta - 106)/10 \leqq (100 - 106)/10] \\ &= P(U < -0.6) = \Phi(-0{,}6) = 0{,}274, \end{aligned}$$

e, consequentemente,

$$P(H_1 \mid \mathbf{x}) = 0{,}726.$$

Repare-se que o nível de significância é igual a $P(H_0 \mid \mathbf{x})$, facto que não surpreende porquanto se a distribuição a priori é difusa os dois procedimentos baseiam-se exclusivamente na informação amostral. No entanto, apesar de os resultados serem idênticos a interpretação é diferente. A diferença encontra-se ilustrada nas Figs. 1.9 e 1.10. A curva da Fig. 1.10, centrada no ponto 106, com domínio θ, representa a função de distribuição a posteriori; a zona a sombreado corresponde à probabilidade de ser $\theta \leqq 100$ quando $\overline{x} = 106$. A curva da Fig. 1.9, centrada no ponto 100, com domínio $\overline{x}$, representa a função de distribuição por amostragem da

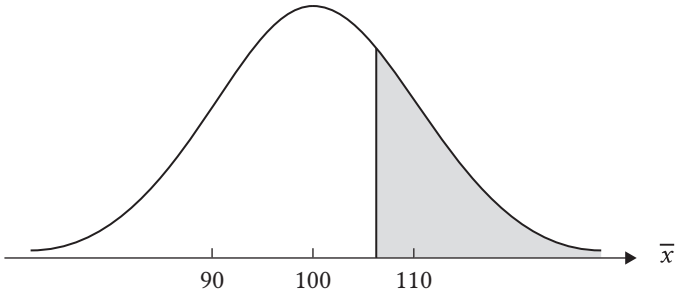


Fig. 1.9

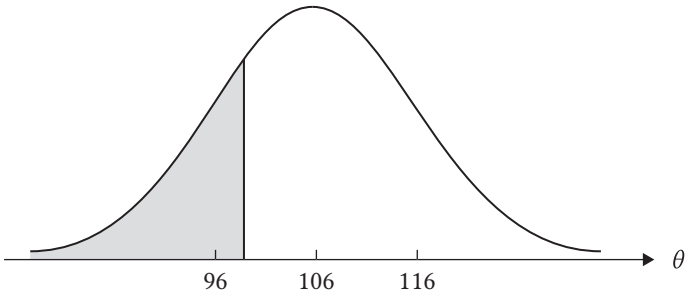


Fig. 1.10

média; a zona a sombreado corresponde à probabilidade de ser $\overline{X} \geqq 106$ quando $\theta = 100$. [Hays e Winkler (1970)]. □

O exemplo anterior tem um aspecto de muito interesse. Mostra a possibilidade de sair,

$$P(\overline{X} \geqq 106 \,|\, \theta = 100) = P(\theta \leqq 100 \,|\, \overline{x} = 106),$$

isto é, da área da aba da distribuição de $\overline{X}$ (= nível de significância) ser igual à probabilidade a posteriori da hipótese H_0. Esta conclusão tem servido de justificação a alguns autores para aproximarem clássicos e bayesianos. Os bayesianos, por sua vez, repetindo uma crítica semelhante feita aos intervalos de confiança, argumentam que os clássicos é que são forçados a interpretar o nível de significância em termos de graus de credibilidade pois, vincam ainda, tal não é viável em termos frequencistas.

Em qualquer caso parece importante inquirir até que ponto vai a «conciliação» entre clássicos e bayesianos na medida em que é traduzida pela igualdade entre a probabilidade a posteriori da hipótese unilateral [H_0: $\theta \leqq 0$ contra H_1: $\theta > 0$, sem perda de generalidade] e o nível de significância, isto é, pela igualdade,

$$P(\theta \leqq 0 \,|\, x) = P(X \geqq x \,|\, \theta = 0),$$

onde X é a variável ou estatística observada. De facto, o Ex. 1.23 representa um caso muito particular em que o modelo experimental é a Normal e a distribuição a priori é não informativa e a questão que se põe é a de saber se a «conciliação» se verifica com outros modelos experimentais e outras distribuições a priori. Casella e Berger (1987) obtiveram resposta afirmativa para um vasto conjunto de situações. Por exemplo, se X tem função densidade, $f(x - \theta)$, onde θ é um parâmetro de localização, $-\infty < \theta < +\infty$, e se $f(x)$ é simétrica em relação à origem [e possui razão de verosimilhança monótona — veja-se secção 7.3], então, no ensaio de $\theta \leqq 0$ contra $\theta > 0$, se sai $x > 0$ [se $x < 0$ toda a gente aceita a hipótese], tem-se,

$$\inf_h P(\theta \leqq 0 \,|\, x) = P(X \geqq x \,|\, \theta = 0),$$

onde o ínfimo é obtido sobre a classe das distribuições a priori unimodais simétricas em relação à origem.

Suponha-se θ escalar, $\Theta = \{\theta : -\infty < \theta < +\infty\}$. Quando o problema clássico de ensaiar uma hipótese simples, H_0: $\theta = \theta_0$, contra uma alternativa composta bilateral, H_1: $\theta \neq \theta_0$, é considerado no quadro bayesiano, depara-se com uma dificuldade: se a distribuição a priori for contínua, o que é de esperar dada a natureza de Θ, tem-se $P(H_0 \,|\, x) = 0$ qualquer que seja x, pois a priori $P(H_0) = 0$. Quer dizer, H_0 nunca poderá ser aceite.

Considerando que se trata essencialmente de um problema de formulação, há três saídas possíveis.

Primeiro, não é impensável adaptar um procedimento que tem alguma analogia com os testes de significância dos clássicos e que consiste em rejeitar a hipótese H_0: $\theta = \theta_0$ quando o valor θ_0 não pertence ao intervalo de credibilidade H.P.D. estabelecido a partir de $h(\theta \mid \mathbf{x})$ para o nível desejado.

Segundo, pode pensar-se que o que se pretende realmente ensaiar não é ser precisamente $\theta = \theta_0$ mas sim estar θ próximo de θ_0. Assim, desde que se tenha, H_0: $\theta_0 - \Delta < \theta < \theta_0 + \Delta$, $(\Delta > 0)$, já faz sentido falar em $P(H_0 \mid x)$.

Terceiro, desde que se insista na hipótese simples, H_0: $\theta = \theta_0$, pode pensar-se que o particular valor θ_0 tem uma ordem de importância diferente da dos outros valores de θ, sendo por essa razão que é objecto de atenção especial. Esta atitude pode reflectir-se na distribuição a priori atribuindo a θ_0 uma probabilidade a priori $h_0 > 0$ e distribuindo a restante probabilidade, $1 - h_0$, pelos valores $\theta \neq \theta_0$ (trata-se de uma distribuição a priori mista).

O exemplo seguinte ilustra a forma de proceder e põe em destaque uma curiosa situação conhecida por paradoxo de Lindley.

Exemplo 1.24 — Com $X \sim N(\theta, \sigma^2)$, σ^2 conhecido, pretende ensaiar-se H_0: $\theta = \theta_0$ contra H_1: $\theta \neq \theta_0$. A probabilidade a priori de θ_0 é h_0; a densidade a priori de $\theta \neq \theta_0$ é $(1 - h_0)h_1(\theta)$, com $h_1(\theta) \equiv N(\mu, \tau^2)$. Designando a densidade de $N(\theta, \sigma^2)$ por $f(x \mid \theta)$, vem,

$$h(\theta \mid x) = \begin{cases} \dfrac{f(x \mid \theta_0)h_0}{f(x)} & \text{se } \theta = \theta_0, \\ \dfrac{f(x \mid \theta)(1 - h_0)h_1(\theta)}{f(x)} & \text{se } \theta \neq \theta_0, \end{cases}$$

onde,

$$\begin{aligned} f(x) &= f(x \mid \theta_0)h_0 + (1 - h_0) \int_{\theta \neq \theta_0} f(x \mid \theta)h_1(\theta)\, d\theta \\ &= f(x \mid \theta_0)h_0 + (1 - h_0)g(x), \end{aligned}$$

com $g(x)$ densidade de $N(\mu, \tau^2 + \sigma^2)$ [veja-se (1.27) e repare-se que a exclusão de θ_0 do integral do 2.º membro é irrelevante].

Definindo $\Theta_0 = \{\theta_0\}$ e $\Theta_1 = \{\theta : \theta \neq \theta_0\}$, o rácio das «chances» a posteriori tem por expressão,

$$\begin{aligned}\frac{P(\Theta_0 \mid x)}{P(\Theta_1 \mid x)} &= \frac{h(\theta_0 \mid x)}{1 - h(\theta_0 \mid x)} \\ &= \frac{f(x \mid \theta_0) h_0}{(1 - h_0) g(x)} \\ &= \frac{h_0 \sigma^{-1} \exp\{-(x - \theta_0)^2 / 2\sigma^2\}}{(1 - h_0)(\tau^2 + \sigma^2)^{-1/2} \exp\{-(x - \mu)^2 / 2(\tau^2 + \sigma^2)\}} \\ &= \frac{h_0}{1 - h_0} \left(1 + \frac{\tau^2}{\sigma^2}\right)^{1/2} \exp\{-A\} \exp\{B\},\end{aligned}$$

onde,

$$A = \frac{\{x - [\theta_0 + (\theta_0 - \mu)\sigma^2/\tau^2]\}^2}{2\sigma^2 \left(1 + \frac{\sigma^2}{\tau^2}\right)},$$

$$B = \frac{(\theta_0 - \mu)^2}{2\tau^2}.$$

A desigualdade seguinte,

$$\begin{aligned}&\left| \frac{(x - \theta_0)}{\sigma} + \frac{(\mu - \theta_0)\sigma}{\tau^2} \right| \\ &< \left(1 + \frac{\sigma^2}{\tau^2}\right)^{1/2} \left\{ \log \left[\left(\frac{h_0}{1 - h_0} \right)^2 \left(1 + \frac{\tau^2}{\sigma^2}\right) \right] + \frac{(\theta_0 - \mu)^2}{\tau^2} \right\}^{1/2}\end{aligned} \tag{1.39}$$

é condição necessária e suficiente para que $P(\Theta_0 \mid x) > P(\Theta_1 \mid x)$, isto é, para a posteriori Θ_0 ter mais credibilidade do que Θ_1.

Quando $\sigma \to 0$ — precisão experimental grande — a desigualdade acima pode representar-se aproximadamente por,

$$\left| \frac{x - \theta_0}{\sigma} \right| < (\log \sigma^{-2})^{1/2}.$$

Para comparar o resultado bayesiano (1.39) e o clássico tome-se um caso particular em que,

$$h_0 = 1/2, \quad \mu = \theta_0, \quad \tau = 1, \quad \sigma^2 = \exp(-25).$$

Fazendo a substituição conclui-se que Θ_0 tem maior credibilidade do que Θ_1, quando, aproximadamente,

$$\left| \frac{x - \theta_0}{\sigma} \right| < 5.$$

A divergência em relação ao resultado clássico é flagrante, porquanto, grosso modo, a aceitação de H_0 processa-se quando,

$$\left| \frac{x - \theta_0}{\sigma} \right| < 2.$$

É esta divergência substancial que se conhece como paradoxo de Lindley [Berger (1980)]. □

No caso do ensaio de hipóteses bilaterais a conciliação entre bayesianos e frequencistas, nos termos quantitativos indicados para o caso unilateral, não é possível. Mostra-se, de facto, que a resposta bayesiana difere radicalmente da resposta frequencista [consulte-se Berger (1985)].

Exemplo 1.24 — *Continuação.* O distanciamento entre as perspectivas clássica e bayesiana no ensaio de hipóteses bilaterais [$\theta = \theta_0$ contra $\theta \neq \theta_0$] posto em destaque no exemplo pode ser apreendido de modo mais esclarecedor [veja-se Berger e Sellke (1987) para maior desenvolvimento]. Com efeito, no quadro do modelo experimental e da distribuição a priori considerados e mantendo as premissas.

$$\mu = \theta_0, h_0 = 1/2 \text{ e } \tau = \sigma,$$

pode calcular-se sem dificuldade de maior a expressão de $h(\theta_0 \mid x) = P(\theta = \theta_0 \mid x) = P(H_0 \mid x)$,

$$P(H_0 \mid x) = \left(1 + (1+N)^{-1/2} \exp\left\{\frac{z^2}{2[1+(1/N)]}\right\}\right)^{-1},$$

de que se apresentam alguns valores no quadro abaixo para os níveis de significância mais correntes e para $N = 10$, 50 e 100 (Note-se que $z = \sqrt{N}\,|\,\overline{x} - \theta_0\,|\,/\sigma$).

Nível de Significância	z	$P(H_0 \mid x)$		
		$N = 10$	$N = 50$	$N = 100$
0,10	1,645	0,47	0,65	0,72
0,05	1,960	0,37	0,52	0,60
0,01	2,576	0,14	0,22	0,27

Como se verifica é grande o conflito, para um dado x conducente aos valores de z que constam do quadro, entre nível de significância e probabilidade a posteriori da hipótese. Por exemplo, com $N = 50$ se sai $z = 1{,}96$ um clássico pode rejeitar a hipótese ao nível de significância de 5% quando para um bayesiano, que tenha adoptado a distribuição a priori indicada, a probabilidade a posteriori da hipótese é 0,52.

Mantendo o modelo experimental $[N(\theta, \sigma^2)]$ pode perguntar-se o que sucede quando a distribuição a priori em vez da componente $h_1(\theta) \equiv N(\mu, \tau^2)$ passa a assumir uma forma tão desfavorável a H_0 quanto possível (mantendo-se $h_0 = 1/2$).

Demonstra-se que nesse caso,

$$P(H_0 \mid x) = \left[1 + \exp\left\{\frac{z^2}{2}\right\}\right]^{-1},$$

expressão que dá ainda valores substancialmente superiores ao correspondente nível de significância como pode ver-se pelo quadro abaixo:

Nível de Significância	0,1 ($z = 1{,}645$)	0,05 ($z = 1{,}960$)	0,01 ($z = 2{,}576$)
$P(H_0 \mid x)$	0,205	0,127	0,035

[Berger e Sellke (1987)]. □

Há problemas de inferência que não são dos tipos referidos (estimação, etc.) e que possuem a particularidade de não serem facilmente atacados fora do quadro bayesiano [consulte-se Lindley (1982)]. Trata-se de questões em que observada uma amostra, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, pretende inferir-se sobre o valor de uma nova observação de X, seja X_{N+1}. A respectiva distribuição preditiva,

$$f(x_{N+1} \mid \mathbf{x}) = \int_\Theta f(x_{N+1} \mid \theta) h(\theta \mid \mathbf{x})\, d\theta,$$

permite em geral dar-lhes resposta. Como a expressão acima mostra, a plausibilidade dos diferentes valores de θ, obtida através da distribuição a posteriori, $h(\theta \mid \mathbf{x})$, entra como factor de ponderação da função de densidade, $f(x_{N+1} \mid \theta)$, e induz a plausibilidade dos resultados da nova experiência.

Exemplo 1.25 — Num estudo de fiabilidade observa-se o tempo de vida de uma amostra casual de N componentes, X_i, $i = 1, 2, \ldots, N$. Supõe-se que cada X_i tem função de densidade de probabilidade, $f(x_i \mid \theta) = \theta \exp\{-\theta x_i\}$, $x_i \geqq 0$, $\theta > 0$. Durante o estudo há r componentes que falham nos tempos $x_1, x_2, \ldots, x_r$, e as restantes são retiradas nos tempos $x_{r+1}, \ldots, x_N$, sem terem falhado (diz-se, em casos como este, que a amostra é «censurada»). Notando que $P(X > x) = \exp\{-\theta x\}$ e recordando a definição (1.8), tem-se — com $K = 1$ — a função de verosimilhança,

$$f(\mathbf{x} \mid \theta) = \theta^r \exp\left\{-\theta \sum_{i=1}^{r} x_i\right\} \exp\left\{-\theta \sum_{i=r+1}^{N} x_i\right\} = \theta^r \exp\{-\theta t\},$$

com,

$$t = \sum_{i=1}^{N} x_i.$$

Considerando [confronte-se com o conceito de distribuição conjugada introduzido na secção 1.6] a distribuição a priori da família Gama[38],

$$h(\theta) = \alpha^{\beta+1} \exp\{-\alpha\theta\}\theta^{\beta}/\Gamma(\beta+1),$$

obtém-se, depois de algum cálculo,

$$h(\theta \mid \mathbf{x}) = (\alpha+t)^{\beta+r+1} \exp\{-\theta(\alpha+t)\}\theta^{\beta+r}/\Gamma(\beta+r+1).$$

A função de densidade preditiva,

$$\begin{aligned} f(x_{N+1} \mid \mathbf{x}) &= \int_0^{\infty} \exp\{-\theta x_{N+1}\} h(\theta \mid \mathbf{x})\, d\theta \\ &= (\beta+r)[(\alpha+t)^{\beta+r+1}/(\alpha+t+x_{N+1})^{\beta+r+2}], \end{aligned}$$

permite calcular, por exemplo, a probabilidade da $(N+1)$-ésima componente durar mais do que o tempo x_0,

$$\begin{aligned} P(X_{N+1} > x_0 \mid \mathbf{x}) &= \int_{x_0}^{\infty} f(x_{N+1} \mid \mathbf{x})\, dx_{N+1} \\ &= \left(\frac{\alpha+t}{\alpha+t+x_0}\right)^{\beta+r+1}, \end{aligned}$$

[Lindley (1982)]. □

1.5 O Ecumenismo de Box[39]

Face ao antagonismo entre clássicos e bayesianos não pode ficar-se indiferente à tentativa de «conciliação» de Box (1983, 1984) através de uma teoria dualista da inferência estatística. Box sustenta, sem ser provavelmente o primeiro, que as doutrinas bayesiana e clássica são mais complementares do que concorrentes. Recordando a construção e emprego de modelos (releia-se a secção 1.2) como procedimento corrente na investigação, a tese nasce do reconhecimento de que a controvérsia se centra na realização de inferências condicionadas pela adequação do modelo, deixando em segundo plano o facto de o modelo especificado nunca ser verdadeiro. Por outras palavras, o choque dá-se sobretudo ao nível da estimação, ficando a crítica um tanto na penumbra.

Para Box o conflito resulta de pensar que há dois candidatos para um só lugar e não dois lugares para dois candidatos. Tal posição vem na esteira de Dempser

[38] Admitindo que tal distribuição traduz a posição inicial do analista em relação aos valores do parâmetro.

[39] A leitura desta secção pode omitir-se sem prejudicar a sequência do estudo.

(1971): «*I do not believe that either the Bayesian approach or the sampling distribution approach to unity is a total error, but I do find that subtle issues are involved which compromise parts of both schools, so that a mixed viewpoint becomes desirable. Specifically, one must reckon with the weakness of sampling distribution methods for estimation and of Bayesian methods for significance testing*».

O «ecumenismo» de Box traduz-se por uma divisão de trabalho: aos frequencistas o que é crítica (o modelo é adequado?); aos bayesianos o que é estimação (se o modelo é adequado, estimem-se os parâmetros!).

Em face da divisão proposta parece relevante abordar as duas questões seguintes:

— os métodos frequencistas não são considerados bons para a estimação, porquê?
— os métodos bayesianos não são considerados bons para testar o ajustamento ou a significância, porquê?

A resposta não tem, obviamente, aceitação geral. O que se diz prejudicar os clássicos no trabalho de estimação é a tónica que dão à precisão inicial (pre-observação) deixando de fora a precisão final (post-observação). Os bayesianos, pelo contrário, na medida em que fazem intervir a distribuição a posteriori no trabalho de estimação podem controlar a precisão final desde que o modelo seja «correcto».

Por outro lado, se não há grandes ataques à capacidade crítica dos métodos clássicos, no campo bayesiano as coisas parece que não se passam do mesmo modo. O melhor a fazer é voltar a citar Dempster (1971): «*Under the sampling distribution approach, the theory of what constitutes a good test depends on the formulation of precise alternative hypothesis, but actually carrying out a test requires only sampling distributions specified by a null hypothesis. In contrast, a Bayesian test rests on a posteriori distribution over a collection of null and alternative hypothesis, so that a precondition for carrying out the test is a precise specification of alternative hypothesis. This requirement is both a strength and weakness of Bayesian testing. On the positive side, Bayesian testing can claim to impeccable logical foundation* when alternative hypothesis are not in dispute. *On the negative side, Bayesian testing cannot be a discursive exploratory tool aimed in part at discovering plausible alternative hypothesis, since it operates within a closed system*».

A frase «*it operates within a closed system*» serve para dar ênfase ao contraste da posição bayesiana com o dinamismo ou «*open loop*»:

$$\ldots \Rightarrow \text{modelo} \Rightarrow \text{dados} \Rightarrow \text{modelo} \Rightarrow \ldots,$$

característico da investigação estatística. Quer dizer, na praxis bayesiana tem-se um sistema de hipóteses ou modelos, $H_1, H_2, \ldots, H_m$, com credibilidades iniciais,

$P(H_1), P(H_2), \ldots, P(H_m)$; obtida a informação amostral ou experimental, x, as credibilidades são transformadas em obediência ao Teorema de Bayes, $P(H_1 \mid x)$, $P(H_2 \mid x), \ldots, P(H_m \mid x)$, mas o sistema mantém-se. Ao assinalar que os métodos bayesianos se dirigem à comparação de modelos e não se ocupam da adequação ou crítica de um só modelo, sustenta-se que na tarefa de crítica não há necessidade de especificar modelos alternativos, pois a experiência quotidiana mostra que pode muito bem considerar-se uma situação surpreendente ou pouco usual sem ser obrigado a apontar, imediatamente, uma explicação alternativa.

A sugestão de Box para a referida divisão de trabalho processa-se então do modo seguinte:

1) A distribuição a posteriori, $h(\theta \mid x)$, permite — sem necessidade de recorrer a qualquer outro instrumento — realizar toda e qualquer estimação do parâmetro θ se o modelo for válido; no entanto, não fornece qualquer informação sobre a adequação do modelo.

2) A distribuição preditiva, $f(x)$ [veja-se (1.17)], permite testar a adequação do modelo. Assim, quando se observa $X = x$, e a probabilidade,

$$P[f(X) < f(x)]^{40},$$

é muito pequena, tem de concluir-se que é muito improvável que o valor, x, tenha sido gerado pelo modelo[41] e este deve ser posto em causa; repare-se na Fig. 1.11. Esta actuação crítica é nitidamente frequencista, pois, $[f(X) < f(x)]$, corresponde a um teste de significância clássico. O desempenho da tarefa exige porém dos clássicos uma posição de compromisso em relação ao emprego de distribuições a priori [note-se a sua presença em (1.17)].

Exemplo 1.26 — Para ilustração do ponto 2) suponha-se que se propõe para a variável, X, a distribuição de Poisson,

$$f(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \theta > 0, \quad x = 0, 1, 2, \ldots,$$

e que se propõe para θ a distribuição a priori,

$$h(\theta) = \frac{e^{-\theta}\theta^{m-1}}{\Gamma(m)}, \quad \theta > 0, \quad m > 0.$$

[40] Alternativamente a crítica pode conduzir-se com funções critério, $g_m(x)$, convenientes, calculando,

$$P\{f[g_m(X)] < f[g_m(x)]\}, \quad m = 1, 2, \ldots,$$

e apreciando estas probabilidades nos termos frequencistas em que se aprecia a probabilidade, $P[f(X) < f(x)]$. Em qualquer caso vejam-se os comentários de Berger (1985).

[41] Quer dizer, $h(\theta)$ e/ou $f(x \mid \theta)$, não «previram» bem o que ocorreu.

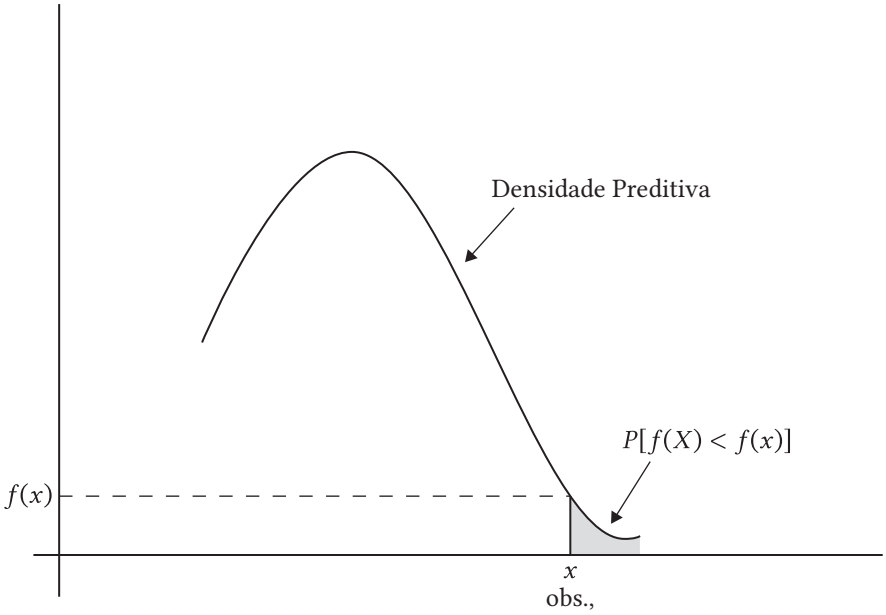


Fig. 1.11. (a) — Modelo questionável

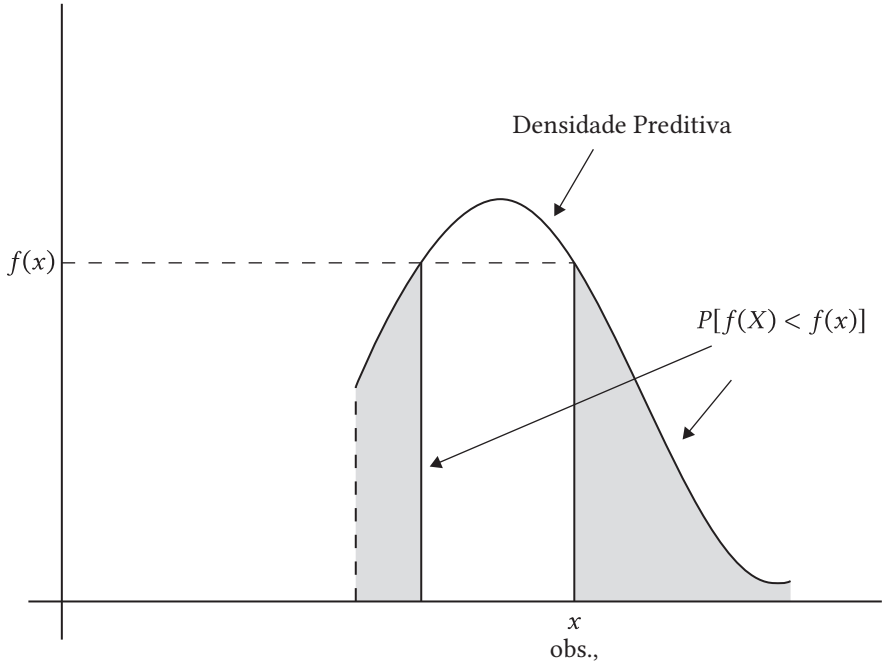


Fig. 1.11. (b) — Modelo aceitável

A distribuição a posteriori tem a forma,

$$h(\theta \mid x) \propto e^{-2\theta}\theta^{m+x-1},$$

sendo a distribuição preditiva,

$$\begin{aligned} f(x) &= [x!\Gamma(m)]^{-1} \int_0^\infty e^{-2\theta}\theta^{m+x-1}d\theta \\ &= [\Gamma(m+x)/\Gamma(x+1)\Gamma(m)][1/2^{m+x}]. \end{aligned}$$

Observado $X = x$, um dos possíveis testes críticos do modelo [$h(\theta)$ e $f(x \mid \theta)$][42] consiste em calcular,

$$\alpha = P[f(X) < f(x)],$$

concluindo pela inadequação do mesmo quando $\alpha < 0{,}05$, 0,01 ou 0,001, de acordo com a prática corrente para testar a significância. Um simples cálculo (que se deixa como exercício) mostra que o modelo é de rejeitar ao nível de significância de 5% quando, por exemplo, com $m = 1$, sair $x \geqq 5$, ou, com $m = 5$, sair $x \geqq 12$. □

1.6 Distribuições a priori

Salvo acidentais referências a casos de informação inicial vaga ou difusa, em que a correspondente distribuição a priori é suposta assumir formas extremamente achatadas ou com variância tendendo para infinito, nada se disse ainda sobre a determinação e interpretação da distribuição a priori.

Trata-se de problema central, e muito controverso, da análise dos procedimentos bayesianos de que passa a fazer-se ligeira abordagem[43].

Há três tipos de interpretação possível para a distribuição a priori de um parâmetro [Cox e Hinkley (1974)]:

[1] interpretação frequencista;
[2] interpretação como representação normativa e objectiva da credibilidade racional;
[3] interpretação como medida subjectiva da credibilidade de um dado indivíduo (investigador, decisor, etc.).

[1] *Interpretação frequencista*
Pode ocorrer o valor do parâmetro ser gerado por um mecanismo aleatório estável; nesse caso, há acordo praticamente unânime em considerar θ uma variável ou vector aleatório[44].

[42] Ou só de $h(\theta)$ se o modelo experimental, $f(x \mid \theta)$, não estiver em causa.
[43] Para um estudo muito mais desenvolvido recomenda-se Berger (1985).
[44] No tipo [3] o parâmetro θ é considerado aleatório num sentido subjectivo.

O exemplo mais citado vem da produção em série e da sucessiva constituição de lotes de uma dada dimensão. Se θ designa a proporção de peças defeituosas em cada lote e se o processo de produção for fisicamente estável, pode perfeitamente admitir-se que θ é uma variável aleatória. A respectiva distribuição a priori pode pois interpretar-se em termos frequencistas; a sua estimação pode fazer-se apreciando o comportamento de θ sobre um conjunto numeroso de lotes.

[2] *Distribuições a priori não informativas*
São muito limitadas as situações em que a interpretação frequencista faz sentido. Assim, não desejando desistir-se da aplicação dos métodos bayesianos — por reconhecer a sua simplicidade e fecundidade — há que ter uma diferente perspectiva das distribuições a priori.

A interpretação em termos da credibilidade racional, extensivamente tratada por Jeffreys (1961), tem-se dirigido quase exclusivamente aos casos em que a informação inicial é inexistente, isto é, casos em que há ignorância a priori. A distribuição a priori diz-se então não informativa.

Na medida em que a informação amostral ou experimental é a única reconhecida, esta abordagem tem pontos de contacto com a abordagem clássica. A diferença reside na insistência da primeira em fazer reflectir a informação da amostra numa distribuição a posteriori, de modo a garantir a utilização do formalismo bayesiano e a interpretar as conclusões obtidas nos termos peculiares que o caracterizam.

O problema a resolver é, portanto, o de especificar objectivamente uma distribuição a priori para um parâmetro acerca do qual nada se sabe.

Se o parâmetro pode assumir um número finito de valores, por exemplo $\Theta = \{1, 2, \dots, m\}$, parece natural tomar uma distribuição uniforme, $h(\theta) = 1/m$, $\theta = 1, 2, \dots, m$, para traduzir ignorância a priori. De contrário estavam a considerar-se uns valores mais credíveis do que outros e o estado não seria de ignorância.

A proposta do parágrafo anterior vem na linha do princípio da razão insuficiente de Bayes-Laplace e julga-se pacífica. As dificuldades surgem quando pretende alargar-se o princípio a situações em que o espaço do parâmetro é contínuo.

Quando Θ é um intervalo finito, $\Theta = \{\theta : \theta_0 \leqq \theta \leqq \theta_1\}$, $\theta_0 < \theta_1$, parece natural propor, $h(\theta) = 1/(\theta_1 - \theta_0)$, $\theta_0 \leqq \theta \leqq \theta_1$; a proposta não é, porém, pacífica. Há sempre algo de arbitrário na parametrização dos modelos probabilísticos, quer dizer, em vez de escolher θ para parâmetro pode optar-se por $\lambda = \Lambda(\theta)$, onde Λ é uma função suposta não linear; o estado de ignorância refere-se tanto a θ como a λ, mas a distribuição uniforme para θ é incompatível com a distribuição uniforme para λ. A questão é adiante retomada.

Quando Θ é um intervalo infinito, a sugestão de Jeffreys é a seguinte:

(a) $\Theta = (-\infty, +\infty)$, θ parâmetro de localização, deve tomar-se[45],

$$h(\theta)\, d\theta \propto d\theta. \tag{1.40}$$

(b) $\Theta = (0, \infty)$, θ parâmetro de escala, deve tomar-se,

$$h(\theta)\, d\theta \propto \frac{1}{\theta} d\theta. \tag{1.41}$$

Note-se que tanto (1.40) como (1.41) são distribuições impróprias, porquanto,

$$\int_{-\infty}^{+\infty} d\theta = \infty, \quad \int_0^{\infty} \frac{1}{\theta} d\theta = \infty.$$

Semelhante conclusão tem dado origem a grande controvérsia.

Os seguidores de Jeffreys, que são numerosos, não se impressionam e declaram que tudo está bem quando o formalismo bayesiano conduz a uma distribuição a posteriori própria, $h(\theta \mid x)$, mesmo quando a distribuição a priori, $h(\theta)$, é imprópria. E acrescentam: se é certo que $h(\theta \mid x)$ não pode ser rigorosamente considerada como a distribuição de θ condicionada por x, não faltam argumentos heurísticos para suportar a sua interpretação como tal. Por exemplo, considerando uma sucessão de distribuições a priori próprias, $h_\tau(\theta)$, que tendem para $h(\theta)$ imprópria quando $\tau \to \infty$ [veja-se o Ex. 1.20].

Box e Tiao utilizam (1.40) e (1.41) para representar o comportamento local das distribuições a priori na região em que a função de verosimilhança tem valor apreciável, mas recusam a sua extensão a todo o domínio do parâmetro. E justificam-se: «*By supposing that to a sufficient approximation the prior follows the form (1.2.17) or (1.2.18)* [no caso presente (1.40) e (1.41) respectivamente] *only over the range of appreciable likelihood and that it suitably tails to zero outside that range we ensure that the priors actually used are proper. Thus, by employing the distribution in a way that makes pratical sense we are relieved of a theoretical difficulty*» [Box e Tiao (1973)].

Neyman [veja-se Barnett (1982)] conta-se entre os que não aceitam o emprego de distribuições a priori impróprias e vai mesmo ao ponto de considerar como «saída fácil» da análise bayesiana a ideia de representar a ignorância a priori de uma maneira formal. Cornfield tentou de princípio defender as distribuições a priori impróprias, chamando-lhes «funções iniciais», mas acabou por considerar o seu uso pouco satisfatório.

A favor ou contra, interessa estudar rapidamente a justificação dada por Jeffreys a (1.40) e (1.41).

[45] «Distribuição» uniforme em $(-\infty, +\infty)$.

No caso (a), sendo θ um parâmetro de localização, a distribuição a priori deve ser invariante para transformações da forma $\lambda = \alpha + \beta\theta$; é o que de facto sucede com (1.40) pois $d\lambda \propto d\theta$. Chega-se à mesma conclusão estabelecendo que, para um parâmetro de localização, todos os intervalos $(a, a + \Delta)$ devem ter a mesma probabilidade a priori para qualquer Δ dado e para todo o a.

No caso (b), sendo θ um parâmetro de escala, a distribuição a priori deve ser invariante para transformações da forma $\lambda = \theta^n$; é o que sucede com (1.41) pois $d\lambda = n\theta^{n-1}d\theta$, donde $d\lambda/\lambda \propto d\theta/\theta$. Chega-se à mesma conclusão estabelecendo que, para um parâmetro de escala, todos os intervalos $(a, \kappa a)$, devem ter a mesma probabilidade a priori para qualquer $\kappa > 1$ dado e para todo o $a > 0$.

A invariância em relação a transformações do tipo $\lambda = \theta^n$ é importante porquanto a escala pode ser parametrizada, por exemplo, através do desvio padrão, σ, através da variância, σ^2, através da precisão, $\eta = 1/\sigma^2$, etc. Adoptando (1.41) tem-se que as distribuições a priori são da mesma forma e coerentes entre si: $d\sigma/\sigma \propto d\sigma^2/\sigma^2 \propto d\eta/\eta$.

Quando $\theta = (\mu, \sigma)$, onde μ é um parâmetro de localização e σ um parâmetro de escala, nomeadamente numa distribuição Normal, $N(\mu, \sigma^2)$, a sugestão de Jeffreys consiste em considerar μ e σ independentes e em combinar (1.40) e (1.41):

$$h(\theta) \equiv h(\mu,\sigma) \propto (1/\sigma)\, d\mu d\sigma. \tag{1.42}$$

As propriedades de invariância que conduziram a (1.40) e (1.41) foram generalizadas por Jeffreys.

Seja, com $X \sim f(x \mid \theta)$, $J(\theta)$ a quantidade de informação de Fisher,

$$\begin{aligned} J(\theta) &= \int_{\mathcal{X}} \left(\frac{\partial \log f(x \mid \theta)}{\partial \theta} \right)^2 f(x \mid \theta)\, dx \\ &= -\int_{\mathcal{X}} \left(\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} \right) f(x \mid \theta)\, dx, \end{aligned} \tag{1.43}$$

bem conhecida da desigualdade de Cramér-Rao. Alternativamente,

$$J(\theta) = E\left\{ \left[\frac{\partial \log f(X \mid \theta)}{\partial \theta} \right]^2 \right\} = -E\left\{ \frac{\partial^2 \log f(X \mid \theta)}{\partial \theta^2} \right\}. \tag{1.44}$$

Introduzindo como distribuição a priori,

$$h(\theta) \propto [J(\theta)]^{1/2}, \tag{1.45}$$

pode mostrar-se que se o investigador resolve parametrizar o modelo em termos de $\lambda = \Lambda(\theta)$, onde Λ é uma função biunívoca e derivável, e toma como distribuição a priori,

$$h(\lambda) \propto [J(\lambda)]^{1/2}, \tag{1.46}$$

então,

$$[J(\theta)]^{1/2}d\theta = [J(\lambda)]^{1/2}d\lambda; \tag{1.47}$$

quer dizer, as distribuições a posteriori são análogas e as inferências delas decorrentes equivalem-se.

A demonstração de (1.47) é fácil. Escrevendo, para simplificar, $f \equiv f(x\,|\,\theta)$, tem-se,

$$\frac{\partial \log f}{\partial \theta} = \frac{\partial \log f}{\partial \lambda}\frac{d\lambda}{d\theta},$$

donde sai,

$$E\left\{\left[\frac{\partial \log f}{\partial \theta}\right]^2\right\} = E\left\{\left[\frac{\partial \log f}{\partial \lambda}\right]^2\right\}\left(\frac{d\lambda}{d\theta}\right)^2,$$

e finalmente (1.47).

Exemplo 1.27 — Seja $X \sim N(\mu, \sigma^2)$. Suponha-se σ conhecido; vem,

$$\log f = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2\sigma^2},$$

donde,

$$\frac{\partial \log f}{\partial \mu} = \frac{x-\mu}{\sigma^2},$$

e,

$$J(\mu) = E\left\{\left[\frac{\partial \log f}{\partial \mu}\right]^2\right\} = E\left\{\frac{(X-\mu)^2}{\sigma^4}\right\} = \text{ constante},$$

É-se, portanto, reconduzido a (1.40), $h(\mu)d\mu \propto d\mu$, por (1.45).

Suponha-se μ conhecido; vem,

$$\frac{\partial \log f}{\partial \sigma} = -(1/\sigma) + (x-\mu)^2/\sigma^3,$$

$$\frac{\partial^2 \log f}{\partial \sigma^2} = (1/\sigma^2) - 3(x-\mu)^2/\sigma^4,$$

donde,

$$J(\sigma) = 2/\sigma^2,$$

que conduz precisamente a (1.41) depois de notar (1.45).

Com μ e σ desconhecidos o parâmetro é bidimensional, (μ, σ), e em vez da quantidade de informação tem-se a matriz de informação, $\mathbf{J}(\mu, \sigma)$; a distribuição a priori é então construída com o respectivo determinante,

$$h(\mu, \sigma) \propto |\,\mathbf{J}(\mu, \sigma)\,|^{1/2}, \tag{1.48}$$

onde,

$$\mathbf{J}(\mu,\sigma) = \begin{bmatrix} -E\left\{\dfrac{\partial^2 \log f}{\partial \mu^2}\right\} & -E\left\{\dfrac{\partial^2 \log f}{\partial \mu \partial \sigma}\right\} \\ -E\left\{\dfrac{\partial^2 \log f}{\partial \sigma \partial \mu}\right\} & -E\left\{\dfrac{\partial^2 \log f}{\partial \sigma^2}\right\} \end{bmatrix}.$$

O cálculo do determinante e da respectiva raiz quadrada conduz à distribuição a priori,

$$h(\mu,\sigma)\, d\mu d\sigma \propto (1/\sigma^2)\, d\mu d\sigma, \tag{1.49}$$

que Jeffreys rejeita em favor de (1.42) por sustentar que a priori μ e σ devem considerar-se independentes. □

Zellner (1971) trata o problema da ignorância a priori no contexto da teoria da informação. Utilizando a medida de Shannon pode dizer-se que a informação dada pela observação de $X \sim f(x\,|\,\theta)$, com um dado θ, é,

$$I_f(\theta) = \int_{\mathfrak{X}} f(x\,|\,\theta)[\log f(x\,|\,\theta)]\, dx;$$

considerando todos os possíveis valores de θ ponderados pela respectiva distribuição a priori tem-se a informação média fornecida por X,

$$\bar{I}_f = \int_{\Theta} I_f(\theta) h(\theta)\, d\theta.$$

Por sua vez, a informação a priori contida em $h(\theta)$ é medida por,

$$I_h = \int_{\Theta} h(\theta)[\log h(\theta)]\, d\theta;$$

a diferença,

$$G = \bar{I}_f - I_h,$$

representa o ganho em informação conseguido com os dados, em média. Uma distribuição a priori com «informação mínima» é então aquela que, para um dado $f(x\,|\,\theta)$, maximiza o ganho em informação.

A exemplificação com a distribuição Normal conduz precisamente a (1.40), (1.41) e (1.42), consoante os casos. Nestes casos pode dizer-se que a distribuição a priori não informativa é a que fornece informação mínima.

Jeffreys e Zellner são dois nomes de uma longa lista de proponentes de técnicas para produzir distribuições a priori não informativas[46]. De facto tem de aceitar-se

[46] Veja-se Berger (1985) para maior desenvolvimento e, nomeadamente, para a aplicação no contexto do conceito de entropia.

que em geral existe uma larga classe de distribuições a priori não informativas logicamente plausíveis [uma situação em que a escolha parece ser natural é abordado na secção 5.5].

Exemplo 1.14 — *Continuação.* Ao introduzir este exemplo disse-se que não era pacífico tomar a distribuição uniforme no intervalo $[0, 1]$ como distribuição a priori não informativa para o parâmetro, θ, da binomial. Na verdade conhecem-se, ou melhor, foram propostas, quatro distribuições a priori não informativas plausíveis:

$$h_1(\theta) = 1 \qquad \text{[já referida]},$$
$$h_2(\theta) = \theta^{-1}(1-\theta)^{-1},$$
$$h_3(\theta) \propto [\theta(1-\theta)]^{-1/2},$$
$$h_4(\theta) \propto \theta^{\theta}(1-\theta)^{(1-\theta)}.$$

Das quatro, apenas $h_2(\theta)$ é imprópria. Para mais pormenores veja-se Berger (1985). □

O embaraço na escolha de uma distribuição a priori não informativa pode ser mitigado considerando que tal escolha (dentro da classe das plausíveis) raramente afecta de forma significativa a solução do problema envolvido. Se tal não sucede deve haver motivo para empregar uma distribuição a priori subjectiva do tipo adiante considerado.

Outro aspecto embaraçoso decorre de as distribuições a priori não informativas serem quase sempre obtidas por métodos — veja-se (1.45) e (1.48) — que dependem da estrutura experimental [Berger (1985)], isto é, envolvem a integração sobre o espaço da amostra, não estando portanto o seu emprego em consonância com o princípio de verosimilhança.

Confunde-se, por vezes, as distribuições a priori não informativas com as distribuições que traduzem um conhecimento a priori vago ou difuso, talvez porque na prática as conclusões a que se chega nos dois casos são muito semelhantes. No entanto há uma diferença importante; as segundas referem-se a situações em que não é razoável sustentar que existe ignorância a priori mas em que a informação dada pela amostra é de tal forma dominante que a distribuição a posteriori é praticamente proporcional à função de verosimilhança.

O efeito do «esmagamento» da informação a priori pela informação amostral é descrito por Savage como o princípio de medição precisa [se no Ex. 1.15 se tomar σ muito pequeno e τ muito grande, tem-se um caso de medição precisa]. O princípio de medição precisa ou estimação estável é referido um pouco adiante.

[3] *Distribuições subjectivas*
Ao estudar a construção das distribuições a priori subjectivas convém adaptar uma perspectiva mais ampla aplicável a qualquer variável aleatória, represente ou não um parâmetro.

Começa por tratar-se o caso em que Θ é finito, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$. O problema consiste em avaliar,

$$h(\theta_i) = \text{probabilidade subjectiva de ser } \theta = \theta_i;$$

isto é, consiste em quantificar a incerteza que um dado indivíduo (investigador, decisor, etc.) associa com os diferentes valores de θ. Obviamente, $h(\theta_i) \geqq 0$, $\Sigma h(\theta_i) = 1$.

Uma probabilidade subjectiva ou personalista representa o grau em que um indivíduo coerente acredita num dado valor de um parâmetro ou de uma variável aleatória, na realização de um dado acontecimento ou na veracidade de uma dada proposição. A coerência significa, como já foi dito, que esse indivíduo ao avaliar ou declarar probabilidades subjectivas o faz sem introduzir qualquer contradição fundamental entre essas probabilidades.

Evidentemente, a maneira mais fácil de obter probabilidades subjectivas consiste em perguntar directamente ao indivíduo; é o chamado método directo. Um indivíduo com algumas luzes sobre probabilidades pode em certos casos exprimir, por meio de um número compreendido entre 0 e 1, o grau de credibilidade que atribui a cada valor da variável aleatória ou a cada um dos acontecimentos de um dado conjunto.

Por vezes a tarefa de avaliação é facilitada se a atitude do indivíduo é revelada em termos de vantagens relativas ou «chances» relativas. Se $\Theta = \{\theta_1, \theta_2, \theta_3\}$, por exemplo, o indivíduo pode declarar que a vantagem relativa de θ_1 em relação a θ_2 é de 3 para 1 e que a vantagem relativa de θ_2 em relação a θ_3 é de 2 para 1. Assim, $h(\theta_1)/h(\theta_2) = 3$, $h(\theta_2)/h(\theta_3) = 2$; como, admitindo coerência, $h(\theta_1) + h(\theta_2) + h(\theta_3) = 1$, vem, por substituição,

$$3h(\theta_2) + h(\theta_2) + (1/2)h(\theta_2) = 1,$$

donde, $h(\theta_2) = 2/9$, e, consequentemente, $h(\theta_1) = 2/3$ e $h(\theta_3) = 1/9$.

A avaliação pelos métodos indirectos visa inferir as probabilidades a partir do comportamento do indivíduo. Quer dizer, em lugar do inquérito directo, observa-se o comportamento em certas situações que envolvem lotarias, apostas, prémios de seguro, etc. e procuram tirar-se conclusões.

Um método indirecto muito simples emprega como padrão de referência uma urna equivalente. Suponha-se, $\Theta = \{0, 1, 2, 3, 4, 5, 6\}$, e considere-se $\theta = 3$. Imagine-se uma urna com 1000 bolas, brancas e pretas, fisicamente idênticas, em que a

proporção de brancas e pretas pode ser controlada substituindo brancas por pretas ou pretas por brancas. Em dada altura em que o indivíduo conhece a composição da urna é-lhe pedido para optar por uma das seguintes apostas:

Aposta A: O indivíduo ganha 1000$ se sair $\theta = 3$ e não ganha nada se for $\theta \neq 3$.

Aposta B: O indivíduo ganha 1000$ se a bola extraída ao acaso da urna for branca e não ganha nada se for preta.

Se a urna contém apenas bolas pretas é óbvio que o indivíduo prefere certamente a aposta A. Agora, se as bolas pretas forem sendo substituídas uma a uma por bolas brancas e se após cada substituição o indivíduo é inquirido para declarar a sua preferência, é natural que em dada fase o indivíduo fique indiferente entre as duas apostas. Isso, é claro, antes de chegar ao ponto em que todas as bolas pretas foram substituídas por bolas brancas caso em que a aposta B é certamente preferida à aposta A.

Se, por exemplo, a situação de indiferença é atingida quando há na urna 750 bolas pretas e 250 bolas brancas, entende-se que o indivíduo atribui ao valor $\theta = 3$ uma credibilidade ou probabilidade subjectiva, $h(3) = 250/1000 = 0{,}25$. Repetindo o exercício com os outros valores de θ pode eventualmente chegar-se a uma distribuição subjectiva sobre Θ. Quando não se tem $\Sigma h(\theta) = 1$ é porque existe incoerência e o processo tem de ser repetido. Outra forma de testar a coerência é pedir ao indivíduo para em face das duas novas apostas,

Aposta A*: o indivíduo ganha 1000$ se for $\theta \leqq 3$ e não ganha nada se $\theta > 3$;

Aposta B*: o indivíduo ganha 1000$ se uma bola extraída ao acaso da urna for branca e não ganha nada se for preta,

declarar o número de bolas brancas no ponto de indiferença. Se esse número não for igual a $1000[h(0) + h(1) + h(2) + h(3)]$, o indivíduo não está a ser coerente e há que retomar o exercício.

Repare-se que a operacionalidade deste método, e de outros que a seguir se apresentam, depende da possibilidade de observar o valor de θ.

No método das lotarias supõe-se que o indivíduo é confrontado com a seguinte aposta ou lotaria.

Lotaria: o indivíduo ganha 1000$ se for $\theta = 5$ e não ganha nada se for $\theta \neq 5$,

e é chamado a declarar qual a quantia certa, seja Q$, tal que lhe é indiferente participar na lotaria ou receber imediatamente Q$.

A hipótese em que se baseia a aplicação do método é a de que Q$ designa o valor monetário esperado da lotaria, isto é,

$$\text{Q\$} = p \cdot 1000\$ + (1 - p) \cdot 0\$,$$

donde,

$$p = \frac{\text{Q\$}}{1000\$},$$

valor que é tomado como probabilidade subjectiva de obter o prémio de 1000$ ou seja, $p = h(5)$. E analogamente para outros valores de θ.

Uma das principais críticas feitas aos métodos indirectos descritos advém do emprego de terminologia derivada dos «jogos de azar» e sobretudo da possibilidade de o comportamento do indivíduo ser determinado, não apenas pelos seus graus de credibilidade, mas também pela sua atitude em relação ao risco, isto é, maior ou menor propensão a arriscar ou jogar. Alguns aspectos da questão são retomados a propósito da Teoria da Utilidade.

Considere-se agora o caso em que Θ é infinito contínuo,

$$\Theta = \{\theta : \theta' \leqq \theta \leqq \theta''\},$$

onde pode ser $\theta' = -\infty$ e/ou $\theta'' = +\infty$. Neste caso a distribuição a priori pode ser caracterizada por uma função de densidade de probabilidade, $h(\theta)$, ou por uma função de distribuição, $H(\theta)$.

O método do histograma é bastante intuitivo. Consiste em subdividir Θ em intervalos, determinar a credibilidade atribuída a cada intervalo, ajustando finalmente uma função de densidade ao histograma assim construido. Trata-se de um processo grosseiro que pode dar bom resultado quando não há grande exigência em relação à distribuição a priori. Os dois maiores inconvenientes são: (1) conduzir, por vezes, a distribuições a priori dificilmente manuseáveis; (2) conduzir a distribuições a priori sem abas.

No método das alturas relativas começa por pedir-se ao indivíduo que indique o valor de θ que considera mais credível ou provável. Suponha-se que o indivíduo indica como valor modal, θ_0. Em seguida inquire-se quais são, no entender do indivíduo, os valores abaixo e acima de θ_0, com credibilidade igual a metade da credibilidade de θ_0. Admita-se que,

$$h^*(\theta_1) = h^*(\theta_2) = (1/2)h^*(\theta_0),$$

$\theta_1 < \theta_0 < \theta_2$, onde se escreveu h^* em vez de h para indicar que não se trata ainda de uma função de densidade. Depois pode obter-se do indivíduo a indicação de que,

$$h^*(\theta_3) = h^*(\theta_4) = (1/4)h^*(\theta_0),$$

$\theta_3 < \theta_1 < \theta_0 < \theta_2 < \theta_4$, e, finalmente, depois de mais algumas inquirições do mesmo estilo, pede-se a indicação de θ' e de θ'' tais que,

$$h^*(\theta') = h^*(\theta'') \simeq 0.$$

Considerando provisoriamente $h^*(\theta_0) = 1$, pode traçar-se uma curva suave passando pelos pontos, $(\theta', 0)$, $(\theta_3, 0{,}25)$, $(\theta_1, 0{,}5)$, $(\theta_0, 1)$, $(\theta_2, 0{,}5)$, $(\theta_4, 0{,}25)$ e $(\theta'', 0)$. Ajustando a escala das ordenadas de modo a que a área total venha igual a 1, tem-se uma estimativa de $h(\theta)$.

Quando Θ não é intervalo finito, em particular quando $\Theta = (-\infty, +\infty)$, o método das alturas relativas, tal como o método do histograma, estabelece o comportamento aproximado de $h(\theta)$ num intervalo finito, ficando por definir o que se passa nas abas. Ao ajustar a escala importa portanto não esquecer a credibilidade de cada uma das abas.

A proposta de uma dada forma funcional é talvez o método mais usado. Se o indivíduo tem alguma ideia sobre a forma da distribuição a priori, por exemplo $\theta \sim N(\mu, \tau^2)$, para especificar completamente a distribuição pode tentar obter-se do indivíduo uma opinião sobre a média, μ, e a variância, τ^2. Em geral é mais fácil dar uma opinião sobre a média do que sobre a variância (ou precisão). Nesse caso pode inquirir-se do indivíduo quais são os valores equidistantes da média, sejam $\mu \pm \Delta$, que no seu entender compreendem 50% da credibilidade total ou probabilidade total (em vez de 50% pode fazer-se o ensaio com 90% ou 95%). Se a resposta for $\Delta \simeq 10$, tem-se como é sabido,

$$\mu \pm 10 \simeq \mu \pm 0{,}675\tau,$$

donde a estimativa $\tau \simeq 15$.

No Ex. 1.15 foi referido um caso em que o investigador considera como razoável uma distribuição a priori Beta, $h(\theta) = [B(\alpha, \beta)]^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, com média 0,8 e variância 0,01. Como, neste caso,

$$E\{\theta\} = \frac{\alpha}{\alpha+\beta}, \quad V\{\theta\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

tem-se $\alpha = 12$ e $\beta = 3$.

O método dos quantis está voltado para a determinação da função de distribuição. Recorde-se que com $0 < \alpha < 1$, θ_α é o quantil de ordem α se $H(\theta_\alpha) = \alpha$, isto é, se o acontecimento $\theta \leqq \theta_\alpha$ tem credibilidade α. Suponha-se, para concretizar, que θ designa a inflação no próximo ano e que se trata de inquirir um dado economista. O método processa-se, por exemplo, nas seguintes fases:

(1) Começa por pedir-se ao economista para indicar um valor de θ, seja $\theta_{0,50}$, tal que $H(\theta_{0,50}) = 0{,}5$, isto é, um valor tal que para o economista é tão provável uma inflação inferior a $\theta_{0,50}$ como uma inflação superior a $\theta_{0,50}$. Suponha-se que indica $\theta_{0,50} = 10\%$; então, 10% é a mediana da distribuição e $H(10\%) = 0{,}5$.
(2) Pede-se depois para indicar dois valores de θ, sejam $\theta_{0,25}$ e $\theta_{0,75}$, tais que $H(\theta_{0,25}) = 0{,}25$ e $H(\theta_{0,75}) = 0{,}75$. O 1.º quartil, $\theta_{0,25}$, é um valor tal que para o economista é tão provável que a inflação esteja no intervalo $(\theta', \theta_{0,25})$ como no intervalo $(\theta_{0,25}, \theta_{0,50})$. O 3.º quartil, $\theta_{0,75}$, é tal que para o economista é tão provável que a inflação esteja no intervalo $(\theta_{0,50}, \theta_{0,75})$ como no intervalo $(\theta_{0,75}, \theta'')$. Suponha-se que indica, $\theta_{0,25} = 7\%$ e $\theta_{0,75} = 12\%$; então, $H(7\%) = 0{,}25$ e $H(12\%) = 0{,}75$.
(3) Finalmente (pode ainda haver fases intermédias) pede-se para avaliar os limites θ' e θ'', entre os quais, na sua opinião, deve a inflação situar-se quase certamente. Se indica 5% e 15%, tem-se $H(5\%) \simeq 0$ (ou $= 0{,}01$), $H(15\%) \simeq 1$ (ou $= 0{,}99$).

Realizadas as três fases referidas fica-se com 5 pontos da imagem da função de distribuição a priori. Ajustando uma curva suave a esses pontos fica-se com uma ideia aproximada de $H(\theta)$.

A análise dos métodos descritos permite ter uma ideia das contigências que cercam a determinação da distribuição a priori subjectiva. E repare-se que até agora se considerou θ escalar. Se θ é um vector, essa determinação complica-se muito quando as respectivas componentes não possam considerar-se independentes. Se tal sucede o método mais acessível é o da forma funcional, em que se procura estimar os parâmetros (quanto mais parcimónia, melhor). Quando a dimensão de θ é pequena, por exemplo, com $\theta = (\theta_1, \theta_2)$, é recomendável começar por determinar uma distribuição marginal, seja $h(\theta_1)$, passando depois à distribuição condicionada, $h(\theta_2 \mid \theta_1)$, e obtendo finalmente a distribuição conjunta por multiplicação.

Face às dificuldades que cercam a determinação das distribuições a priori compreendem-se melhor as críticas feitas aos procedimentos bayesianos, sobretudo naqueles casos em que as inferências finais estão muito condicionadas pela validade das mesmas distribuições, isto é, em que a distribuição a posteriori é particularmente sensível a variações dentro da classe de distribuições a priori consideradas plausíveis.

Situação em que tal inconveniente se encontra grandemente esbatido é a que foi referida por medição precisa ou estimação estável. A natureza incisiva da informação experimental é posta por Edwards, Lindman e Savage (1963) nos termos seguintes: «*If observations are precise, in a certain sense, relative to the prior distribution on which they bear, then the form and properties of the prior distribution have*

negligible influence on the posterior distribution. From a pratical point of view, then, the untrammeled subjectivity of opinion about a parameter ceases to apply as soon as much data become available». Os mesmos autores estudaram as condições em que a distribuição a posteriori construída a partir da distribuição a priori uniforme — própria ou não — era uma boa aproximação da distribuição a posteriori construída a partir de uma distribuição a priori traduzindo adequadamente a atitude do analista, mas de difícil eliciação.

Seja, $h(\theta \mid x)$, a distribuição a posteriori correspondente à «exacta» distribuição a priori, $h(\theta)$, e $h^*(\theta \mid x)$ a distribuição a posteriori aproximada obtida de uma distribuição a priori uniforme (própria ou imprópria),

$$h^*(\theta \mid x) = f(x \mid \theta) / \int_\Theta f(x \mid \theta)\, d\theta, \quad 0 < \int_\Theta f(x \mid \theta)\, d\theta < \infty;$$

o teorema de Edwards, Lindman e Savage (1963) desenvolve-se nas seguintes linhas gerais:

Considere-se um subconjunto do espaço dos parâmetros, $\Omega \subset \Theta$, tal que,

$$\int_\Omega h^*(\theta \mid x)\, d\theta \geqq 1 - \alpha, \quad 0 \leqq \alpha < 1; \qquad [A]$$

na prática deve ter-se α muito pequeno, por exemplo, $\alpha = 10^{-4}$, e, assim, o conjunto Ω indica uma parte de Θ especialmente favorecida pelos dados.

Suponha-se, sendo m um número real positivo, que,

$$m < h(\theta) < (1 + \beta)m, \quad \theta \in \Omega, \qquad [B]$$

$\beta \geqq 0$ (na prática, 0,01 ou 0,05 são valores convenientes); a hipótese corresponde a admitir que no conjunto Ω a distribuição a priori tem uma evolução quase constante.

Suponha-se ainda que,

$$h(\theta) \leqq (1 + \gamma)m, \quad \theta \in \overline{\Omega}, \qquad [C]$$

onde $\overline{\Omega}$ é o complementar de Ω, hipótese que corresponde a admitir que a distribuição a priori não assume valores excepcionalmente elevados fora de Ω; valores de $\gamma = 100\alpha$ ou mais elevados podem ter de ser tolerados.

Finalmente defina-se ε pela relação,

$$\varepsilon = \max\left\{\frac{\alpha + \beta}{1 - \alpha}, \frac{\alpha + \beta + \alpha\gamma}{1 + \alpha + \beta + \alpha\gamma}\right\} + \frac{\alpha(2 - \alpha + \gamma)}{1 - \alpha};$$

então,

$$\int_\Theta |\, h(\theta \mid x) - h^*(\theta \mid x)\,|\, d\theta \leqq \varepsilon,$$

i.e., a «distância» entre $h(\theta \mid x)$ e $h^*(\theta \mid x)$ é muito pequena.

Na Fig. 1.12 ilustra-se o teorema.

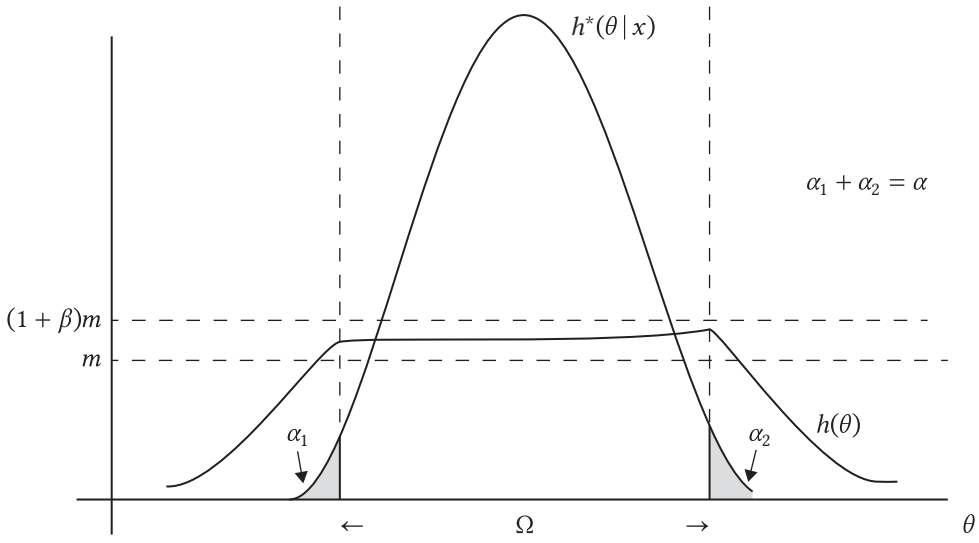


Fig. 1.12

Preocupação semelhante[47] embora desligada das restrições da estimação estável, levou alguns bayesianos subjectivos [veja-se Berger (1984)] a propor o conceito de procedimento bayesiano robusto. O ponto de partida de tal concepção é o pressuposto de que as distribuições a priori não podem nunca ser quantificadas exactamente, porquanto só quando se dispõe de tempo infinito se pode fazer uma quantificação ou eliciação sem erros. Sendo assim deve trabalhar-se com procedimentos bayesianos que se revelem satisfatórios para todas as distribuições a priori consideradas plausíveis ao fim do esforço de determinação que foi possível empreender. O atributo robusto aplica-se, portanto, aos procedimentos que fornecem bons resultados qualquer que seja a distribuição a priori dentro da classe das tidas como plausíveis [o problema da robustez volta a referir-se na secção 3.6].

As dificuldades não se esgotam com o problema da sensibilidade. Um outro problema que se depara é o seguinte: a aplicação de (1.16), fulcral para os procedimentos bayesianos, depende, muitas vezes, da possibilidade de calcular o integral (1.17), tarefa que pode não ser pacífica[48]. Para evitar bloqueamentos ou o recurso a morosos cálculos Raiffa e Schlaifer (1961) introduziram o conceito de família conjugada de distribuições a priori.

[47] I.e., com a sensibilidade das distribuições a posteriori.

[48] Por exemplo, se $X \sim N(\theta, \sigma^2)$ e se θ tem distribuição de Cauchy, $h(\theta \mid x)$ só pode obter-se por cálculo numérico.

Em termos genéricos o objectivo é permitir uma passagem «suave» às distribuições a posteriori e, ao mesmo tempo, ter alguma percepção do peso relativo da informação a priori e da informação amostral. A designação conjugada entende-se relativamente à família, $\mathcal{F} = \{f(x\,|\,\theta) : \theta \in \Theta\}$, especificada quando se elege a variável ou vector aleatório a observar.

Concretamente pedem-se para uma família conjugada as propriedades seguintes [Hays e Winkler (1970)]: (1) fácil tratamento matemático; (2) variedade; (3) interpretação acessível.

A propriedade (1) está na base da introdução do próprio conceito e resulta, aliás, de exigir que a distribuição a posteriori seja ainda membro da família conjugada, com vantagens patentes no exemplo seguinte:

Exemplo 1.28 — Se $X \sim B(N;\theta)$, tem-se $f(x\,|\,\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}$, $\Theta = [0,1]$, $x = 0, 1, 2, \ldots, N$. Quando a distribuição a priori é uma Beta,

$$h(\theta) = [B(\alpha,\beta)]^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 \leqq \theta \leqq 1,$$

também a distribuição a posteriori é uma Beta,

$$\begin{aligned} h(\theta\,|\,x) &\propto f(x\,|\,\theta)h(\theta) \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1} \qquad 0 \leqq \theta \leqq 1, \\ &= [B(\alpha+x, \beta+N-x)]^{-1}\theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}. \end{aligned}$$

A passagem à distribuição a posteriori foi extremamente fácil porquanto, como adiante se mostra, a família Beta é conjugada da família Binomial.

Por outro lado, se a seguir a X se resolve observar Y, X e Y independentes, $Y \sim B(N';\theta)$, vem,

$$\begin{aligned} h(\theta\,|\,x,y) &\propto f(y\,|\,\theta)h(\theta\,|\,x) \\ &\propto \theta^{\alpha+x+y-1}(1-\theta)^{\beta+N+N'-x-y-1}, \end{aligned} \qquad 0 \leqq \theta \leqq 1,$$

isto é, $h(\theta\,|\,x,y)$ continua a ser uma Beta. □

Propriedade (2): dado que as distribuições a priori devem reflectir a informação inicial possuída pelo indivíduo antes de realizar a experiência, se a família conjugada é muito limitada ou pouco flexível pode acontecer que nenhum dos seus membros seja capaz de traduzir a atitude do indivíduo em relação aos diferentes valores de θ. Assim, convém que haja variedade, isto é, membros da família com diferentes médias, diferentes variâncias, diferentes assimetrias, etc. A família Beta é exemplo típico de família variada no intervalo $[0,1]$.

Propriedade (3): se possível, a família conjugada deve permitir uma interpretação do processo de passagem da distribuição a priori para a distribuição a posteriori. Esta propriedade nem sempre se consegue concretizar; existem, porém casos,

em que se estabelece uma certa equivalência entre a informação a priori e a informação obtida por amostragem de modo que a informação a posteriori aparece em resultado da combinação das duas primeiras.

Exemplo 1.28 — *Continuação.* A observação de $X = x$ significa que em N provas se obtiveram x «sucessos» e $N - x$ «insucessos». Analisando os parâmetros da distribuição a priori e da distribuição a posteriori pode pensar-se que a informação a priori equivale a α «sucessos» e β «insucessos» em $\alpha + \beta$ provas (é uma pseudo-amostra) e que a informação a posteriori equivale a $\alpha + x$ «sucessos» e $\beta + N - x$ «insucessos» em $\alpha + \beta + N$ provas. Em esquema:

	«Sucessos»	«Insucessos»	Provas
Informação a priori	α	β	$\alpha + \beta$
Informação amostral	x	$N - x$	N
Informação a posteriori	$\alpha + x$	$\beta + N - x$	$\alpha + \beta + N$

. □

Dada a família, $\mathcal{F} = \{f(x \mid \theta) : \theta \in \Theta)$, importa investigar em que condições existe uma família conjugada, seja $\mathcal{H}(\mathcal{F}) = \{h(\theta; \gamma) : \gamma \in \Gamma\}$, onde cada $h(\theta; \gamma) \in \mathcal{H}(\mathcal{F})$ é indexado por um parâmetro, $\gamma \in \Gamma$, escalar ou vector e onde se $h(\theta) \in \mathcal{H}(\mathcal{F})$ [i.e., $h(\theta) \equiv h(\theta; \gamma_0)$, $\gamma_0 \in \Gamma$] também $h(\theta \mid x) \in \mathcal{H}(\mathcal{F})$ [i.e., $h(\theta \mid x) \equiv h(\theta; \gamma_1)$, $\gamma_1 \in \Gamma$].

Se uma família de distribuições, $\mathcal{H}_0 = \{h(\theta; \gamma) : \gamma \in \Gamma]$, é tal que para todo o par $\gamma_0, \gamma_1 \in \Gamma$ existe $\gamma_2 \in \Gamma$, verificando,

$$h(\theta; \gamma_2) \propto h(\theta; \gamma_0) h(\theta; \gamma_1), \tag{1.50}$$

a família diz-se fechada em relação à multiplicação.

Considerando a relação, $h(\theta \mid x) \propto f(x \mid \theta) h(\theta)$ e recordando a análise feita no Ex. 1.28 pode estabelecer-se a condição em que existe uma família conjugada de $\mathcal{F}$: para qualquer $x \in \mathcal{X}$, $f(x \mid \theta)$ considerada função de θ no domínio Θ deve ser proporcional a um dos membros de uma família de distribuições fechada em relação à multiplicação. A família obtida nessa condição diz-se conjugada natural[49] de $\mathcal{F}$.

O seguinte teorema, que se refere ao importante caso em que $\mathbf{X}$ é uma amostra casual de dimensão, N, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, mostra que o conceito de família conjugada natural tem estreitas ligações com o conceito de suficiência.

Teorema 1.1 — Se existe estatística suficiente de dimensão fixa para $\mathcal{F}$, então existe família conjugada natural de $\mathcal{F}$.

[49] No presente texto foca-se apenas o conceito de conjugada natural. No entanto, o conceito de família conjugada é mais geral; por exemplo, como é evidente, a família de todas as distribuições é conjugada de toda e qualquer família $\mathcal{F}$!

Dem. Suponha-se, sem perda de generalidade, que T é estatística suficiente para $\mathcal{F}$ com dimensão fixa $k = 1$. Com $\mathbf{x} = (x_1, x_2, \dots, x_N)$, vem pelo teorema da factorização,

$$f_N(\mathbf{x} \mid \theta) = G_N[T_N(\mathbf{x}),\theta]M(\mathbf{x}),$$

onde se destaca a dimensão da amostra. Fazendo $T_N(\mathbf{x}) = t$ e admitindo,

$$\int_\Theta G_N(t,\theta)\, d\theta < \infty,$$

então,

$$h(\theta; t,N) = G_N(t,\theta) / \int_\Theta G_N(t,\theta)\, d\theta,$$

é uma função de densidade sobre Θ e, além disso,

$$h(\theta; t,N) \propto G_N(t,\theta).$$

Considere-se a família de distribuições,

$$\mathcal{H}_0 = \{h(\theta; t,N) \,:\, t \in \mathbf{R}, \quad N = 1, 2, \dots,\}, \tag{1.51}$$

admitindo que $T_N(\mathbf{x})$ assume valores em $\mathbf{R}$ para todo o N. A família $\mathcal{H}_0$ tem as seguintes propriedades: (I) por construção, $f(\mathbf{x} \mid \theta) \propto h(\theta; t,N)$; (II) é fechada em relação à multiplicação.

Para provar a propriedade (II) tomem-se duas amostras independentes; seja, $\mathbf{x} = (x_1, x_2, \dots, x_N)$, $\mathbf{x}' = (x'_1, x'_2, \dots, x'_N)$, $T_N(\mathbf{x}) = t$, $T_{N'}(\mathbf{x}') = t'$, $\mathbf{x}'' = (\mathbf{x},\mathbf{x}')$, $T_{N+N'}(\mathbf{x}'') = t''$.

Como as amostras são independentes,

$$f_{N+N''}(\mathbf{x}'' \mid \theta) = f_N(\mathbf{x} \mid \theta) f_{N'}(\mathbf{x}' \mid \theta); \tag{1.52}$$

em face da suficiência de $T(\mathbf{X})$, de que T_N, $T_{N'}$ e $T_{N+N'}$ são versões para as dimensões amostrais em índice,

$$\begin{aligned}
&f_N(\mathbf{x} \mid \theta) \propto G_N[T_N(\mathbf{x}), \theta] \propto h(\theta; t,N),\\
&f_{N'}(\mathbf{x}' \mid \theta) \propto G_{N'}[T_{N'}(\mathbf{x}'), \theta] \propto h(\theta; t',N'),\\
&f_{N+N'}(\mathbf{x}'' \mid \theta) \propto G_{N+N'}[T_{N+N'}(\mathbf{x}''), \theta] \propto h(\theta; t'',N + N').
\end{aligned}$$

Finalmente, de (1.52),

$$h(\theta; t'',N + N') \propto h(\theta; t,N) h(\theta; t',N'),$$

como era necessário provar. □□ [50]

[50] O símbolo □□ indica que a demonstração foi concluída ou que terminou o enunciado do teorema.

A família conjugada, $\mathcal{H}_0$, obtida no quadro do teorema anterior — veja-se (1.51) — peca por demasiado restrita; na prática é mais conveniente trabalhar com uma família mais ampla, admitindo, designadamente, que o parâmetro correspondente a N pode assumir valores não inteiros. Reconsidere-se o Ex. 1.28: a família $\mathcal{H}_0$ é, nesse caso,

$$\mathcal{H}_0 = \{[B(x+1, n-x+1)]^{-1}\theta^x(1-\theta)^{N-x}, \quad N = 1, 2, \ldots, 0 \leqq x \leqq N\};$$

é, no entanto, preferível empregar a família completa das Betas,

$$\mathcal{H} = \{[B(\alpha,\beta)]^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}; \alpha > 0, \beta > 0\},$$

como família conjugada (de futuro omite-se a designação natural) da Binomial. Evidentemente, $\mathcal{H}_0$ é uma subfamília de $\mathcal{H}$.

A arbitrariedade na parametrização do modelo $\mathcal{F}$, que como já foi dito suscita problemas quando pretende caracterizar-se uma posição de partida não informativa, implica a não unicidade da família conjugada de um processo amostral. Por exemplo, quando se recolhem amostras de um processo Normal (suponha-se a média conhecida para facilitar) parametrizado pela variância, σ^2, a família conjugada é a Gama Inversa; se o processo é parametrizado pela precisão, $\eta = 1/\sigma^2$, a família conjugada é a Gama.

As famílias conjugadas podem definir-se de forma mais geral [veja-se Dickey (1982)]; no entanto, a existência de uma estatística suficiente de dimensão fixa conduz à família conjugada natural.

Há duas famílias de distribuições que admitem estatísticas suficientes de dimensão fixa: a exponencial (ou regular) e a não regular.

A variável aleatória, X, tem distribuição da família exponencial uniparamétrica[51] quando,

$$f(x\,|\,\theta) = C(\theta)\exp\{Q(\theta)R(x)\}H(x),$$

onde $C(\theta)$ e $Q(\theta)$ são funções reais em Θ e $R(x)$ e $H(x)$ são funções reais em $\mathbf{R}$ e o conjunto, $\{x : f(x\,|\,\theta > 0\}$, não depende de θ. Para uma amostra casual de

[51] A família exponencial uniparamétrica contém subfamílias bem conhecidas: a Binomial, a Poisson, a Normal (variância dada), etc.

Diz-se que a variável alealória, X, tem distribuição da família exponencial k-paramétrica quando,

$$f(x\,|\,\theta) = C(\theta)\exp\left\{\sum_{j=1}^{k} Q_j(\theta)\; R_j(x)\right\}H(x),$$

onde $C(\theta)$ e $Q_j(\theta)$ são funções reais em Θ e $R_j(x)$ e $H(x)$ são funções reais em $\mathbf{R}$ e o conjunto, $\{x : f(x\,|\,\theta) > 0\}$, não depende de θ.

A Normal pertence à família exponencial biparamétrica.

Na distribuição exponencial k-paramétrica dim $\{\theta\} \leqq k$, podendo θ ser escalar.

dimensão, N, tem-se,

$$f(\mathbf{x}\,|\,\theta) = [C(\theta)]^N \exp\{Q(\theta)\Sigma R(x_i)\}\Pi H(x_i).$$

Como facilmente se verifica, tomando uma distribuição a priori da forma,

$$h(\theta) \propto [C(\theta)]^\alpha \exp\{Q(\theta)\beta\},$$

a distribuição a posteriori tem ainda a mesma forma,

$$h(\theta\,|\,\mathbf{x}) \propto [C(\theta)]^{\alpha+N} \exp\{Q(\theta)[\beta + \Sigma R(x_i)]\},$$

A construção da família conjugada é portanto imediata. Os hiper-parâmetros da distribuição a priori são α (correspondente à dimensão da pseudo-amostra inicial) e β (correspondente ao valor da estatística suficiente mínima, $\Sigma R(x_i)$, assumido na pseudo-amostra inicial).

Nas expressões acima considerou-se o caso uniparamétrico; a adaptação ao caso k-paramétrico é imediata.

A família não regular é composta por distribuições do tipo,

$$f(x\,|\,\theta) = C(\theta)H(x), \quad Q_1(\theta) < x < Q_2(\theta),$$

com,

$$[C(\theta)]^{-1} = \int_{Q_1(\theta)}^{Q_2(\theta)} H(x)\,dx.$$

Qualquer que seja a dimensão da amostra, $\mathbf{T} = [\min(X_i), \max(X_i)]$, é estatística suficiente mínima de dimensão dois. Um pouco adiante refere-se a família conjugada da distribuição uniforme, sem dúvida a subfamília mais simples da família não regular; para estudar casos mais gerais pode consultar-se o trabalho de DeGroot (1970).

Seja $\mathcal{H}(\mathcal{F}) = \{h(\theta;\gamma) : \gamma \in \Gamma\}$ a família conjugada de $\mathcal{F}$. Quando,

$$h(\theta) \equiv h(\theta;\gamma_0) \text{ e } h(\theta\,|\,x) \equiv h(\theta;\gamma_1), \quad \gamma_0, \gamma_1 \in \Gamma,$$

o mecanismo bayesiano que consiste na passagem da distribuição a priori, $h(\theta)$, para a distribuição a posteriori, $h(\theta\,|\,x)$, pode representar-se simbolicamente,

$$\gamma_0 \overset{\mathcal{F}}{\longrightarrow} \gamma_1, \tag{1.53}$$

destacando o papel representado pela experimentação, isto é, pela família $\mathcal{F}$. A (1.53) pode dar-se a forma alternativa,

$$\gamma_0 \xrightarrow{\mathcal{F}} \gamma_0 + (\gamma_1 - \gamma_0), \tag{1.54}$$

onde $\gamma_1 - \gamma_0$ exprime a variação sofrida por γ_0 em resultado da informação decorrente do processo de amostragem. Note-se que (1.54) permite avaliar em certo sentido o peso da informação a priori relativamente à informação experimental.

Exemplo 1.28 — *Continuação.* No caso do Ex. 1.28 tem-se,

$$\mathcal{F} = \{\binom{N}{x}\theta^x(1-\theta)^{N-x} \,:\, \theta \in [0,1]\}.$$

Como é de prever e adiante se confirma, a família conjugada é a Beta,

$$\mathcal{H}(\mathcal{F}) = \{h(\theta;\gamma) \,:\, \gamma \in \Gamma\},$$

com,

$$h(\theta;\alpha,\beta) = [\beta(\alpha,\beta)]^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

e, $\gamma = (\alpha,\beta)$, $\Gamma = \{(\alpha,\beta) \,:\, \alpha > 0, \beta > 0\}$.

Suponha-se $h(\theta) \equiv h(\theta;\alpha_0,\beta_0)$, isto é, $\gamma_0 = (\alpha_0,\beta_0)$; da análise feita no Ex. 1.28, vem, $h(\theta \mid x) \equiv h(\theta;\alpha_0 + x, \beta_0 + N - x)$. Logo,

$$(\alpha_0,\beta_0) \xrightarrow{\mathcal{F}} (\alpha_0 + x, \beta_0 + N - x),$$

ou ainda,

$$(\alpha_0,\beta_0) \xrightarrow{\mathcal{F}} (\alpha_0,\beta_0) + (x, N - x),$$

como se representa na Fig. 1.13.

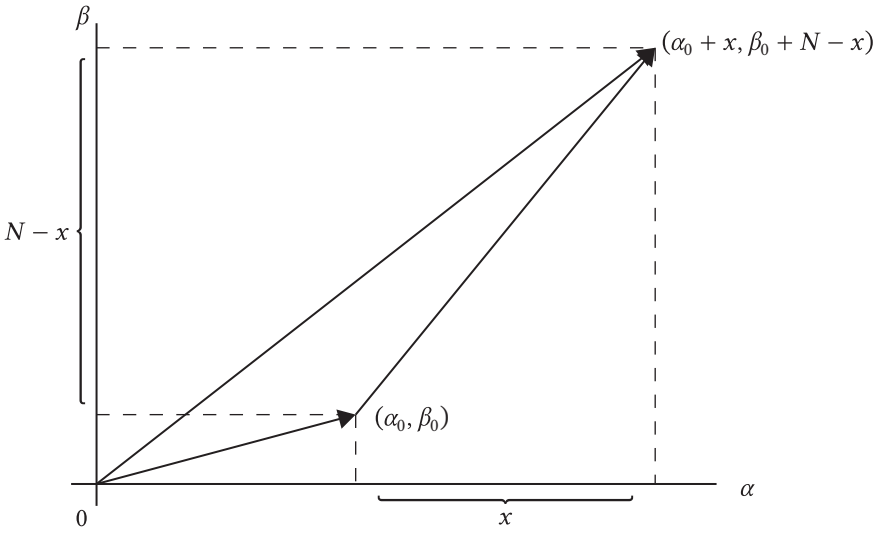


Fig. 1.13

□

Vão indicar-se seguidamente alguns casos importantes de famílias conjugadas; paralelamente apresentam-se as expressões de $E\{\theta \mid x\}$ e de $V\{\theta \mid x\}$ que são de uso frequente na teoria da decisão.

Distribuição de Bernoulli. Com X_i I.I.D., $X_i \sim B(1;\theta)$ — distribuição de Bernoulli — $i = 1, 2, \ldots, N$, vem,

$$f(\mathbf{x} \mid \theta) = \theta^{\Sigma x_i}(1-\theta)^{N-\Sigma x_i}, \quad 0 \leqq \theta \leqq 1.$$

e sabe-se que $T = \Sigma X_i$, é estatística suficiente para θ. Assim,

$$f(\mathbf{x} \mid \theta) \propto \theta^t(1-\theta)^{N-t},$$

e conclui-se imediatamente que a família conjugada é a Beta, pois, como é de fácil verificação, é fechada em relação à multiplicação. Assim, se,

$$h(\theta) \equiv h(\theta;\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

vem,

$$h(\theta \mid \mathbf{x}) \propto h(\theta;\alpha+t,\beta+N-t) \propto \theta^{\alpha+t-1}(1-\theta)^{\beta+N-t-1}.$$

Repondo os x_i e calculando a constante, sai,

$$h(\theta \mid \mathbf{x}) = [B(\alpha+\Sigma x_i, \beta+N-\Sigma x_i)]^{-1}\theta^{\alpha+\Sigma x_i-1}(1-\theta)^{\beta+N-\Sigma x_i-1}, \tag{1.55}$$

distribuição que tem média,

$$E\{\theta \mid \mathbf{x}\} = \frac{\alpha+\Sigma x_i}{\alpha+\beta+N}, \tag{1.56}$$

e variância,

$$V\{\theta \mid \mathbf{x}\} = \frac{(\alpha+\Sigma x_i)(\beta+N-\Sigma x_i)}{(\alpha+\beta+N)^2(\alpha+\beta+N-1)}. \tag{1.57}$$

Distribuição Binomial. Obtém-se imediatamente do caso anterior fazendo $X = \Sigma X_i$. A família conjugada é ainda a Beta (veja-se Ex. 1.28).

Distribuição de Poisson. Com X_i I.I.D., X_i com distribuição de Poisson, $i = 1, 2, \ldots, N$,

$$f(\mathbf{x} \mid \theta) = \frac{e^{-N\theta}\theta^{\Sigma x_i}}{\Pi x_i!}.$$

Sendo $T = \Sigma X_i$ estatística suficiente para θ,

$$f(\mathbf{x} \mid \theta) \propto e^{-N\theta}\theta^t,$$

e a família conjugada é a Gama por ser fechada em relação à multiplicação. Assim,

$$h(\theta) \equiv h(\theta;\alpha,\beta) \propto e^{-\alpha\theta}\theta^{\beta-1}$$

e,

$$h(\theta \mid \mathbf{x}) \equiv h(\theta; \alpha + N, \beta + t) \propto e^{-(\alpha+N)\theta}\theta^{\beta+t-1};$$

repondo os x_i e calculando a constante, sai,

$$h(\theta \mid \mathbf{x}) = \left(\frac{(\alpha+N)^{\beta+\Sigma x_i}}{\Gamma(\beta+\Sigma x_i)}\right) e^{-(\alpha+N)\theta}\theta^{\beta+\Sigma x_i-1}, \tag{1.58}$$

com média e variância,

$$E\{\theta \mid \mathbf{x}\} = \frac{\beta+\Sigma x_i}{\alpha+N}, \tag{1.59}$$

$$V\{\theta \mid \mathbf{x}\} = \frac{\beta+\Sigma x_i}{(\alpha+N)^2}. \tag{1.60}$$

Distribuição Exponencial Negativa. Quando X_i I.I.D., X_i com distribuição exponencial negativa,

$$f(\mathbf{x} \mid \theta) = \theta^N e^{-\theta\Sigma x_i}, \quad \theta > 0, x_i > 0,$$

a família conjugada é ainda a Gama. De facto, $T = \Sigma X_i$ é estatística suficiente para θ e,

$$f(\mathbf{x} \mid \theta) \propto e^{-t\theta}\theta^N.$$

Tomando,

$$h(\theta) \equiv h(\theta; \alpha, \beta) \propto e^{-\alpha\theta}\theta^{\beta-1},$$

vem,

$$h(\theta \mid \mathbf{x}) \equiv h(\theta; \alpha + t, \beta + N) \propto e^{-(\alpha+t)\theta}\theta^{\beta+N-1},$$

donde,

$$h(\theta \mid \mathbf{x}) = \left(\frac{(\alpha+\Sigma x_i)^{\beta+N}}{\Gamma(\beta+N)}\right) e^{-(\alpha+\Sigma x_i)\theta}\theta^{\beta+N-1}, \tag{1.61}$$

$$E\{\theta \mid \mathbf{x}\} = \frac{\beta+N}{\alpha+\Sigma x_i}, \tag{1.62}$$

$$V\{\theta \mid \mathbf{x}\} = \frac{\beta+N}{(\alpha+\Sigma x_i)^2}. \tag{1.63}$$

Distribuição Uniforme. Quando X_i I.I.D., $X_i \sim U(0, \theta)$, tem-se,

$$f(\mathbf{x} \mid \theta) = \theta^{-N}, \quad \theta > 0, 0 < x_i < \theta,$$

e $T = \max(X_i)$ é estatística suficiente para θ. Tome-se a priori uma distribuição de Pareto,

$$h(\theta) \equiv h(\theta; \theta_0, \alpha) = \alpha(\theta_0^\alpha/\theta^{\alpha+1}), \quad \theta > \theta_0, \alpha > 0,$$

ou,

$$h(\theta;\theta_0,\alpha) \propto \theta^{-(\alpha+1)}.$$

Vem,

$$h(\theta \mid \mathbf{x}) \equiv h(\theta;\theta_0',\alpha+N) \propto \theta^{-(\alpha+N+1)}, \quad \theta > \theta_0',$$

onde $\theta_0' = \max\{\theta_0, \max(x_i)\}$. Escrevendo a constante,

$$h(\theta \mid \mathbf{x}) = \frac{(\alpha+N){\theta_0'}^{\alpha+N}}{\theta^{\alpha+N+1}}, \quad \theta > \theta_0', \tag{1.64}$$

$$E\{\theta \mid \mathbf{x}\} = \frac{(\alpha+N)\theta_0'}{\alpha+N-1}, \tag{1.65}$$

$$V\{\theta \mid \mathbf{x}\} = \frac{(\alpha+N){\theta_0'}^2}{(\alpha+N-1)^2(\alpha+N-2)}. \tag{1.66}$$

Distribuição Normal (Variância dada). Quando X_i I.I.D., $X_i \sim N(\theta,\sigma^2)$, facilmente se obtém,

$$f(\mathbf{x} \mid \theta) \propto \exp\{-N(\overline{x}-\theta)^2/2\sigma^2\}.$$

Sabe-se que $T = \overline{X}$ é estatística suficiente para θ. Tomando para distribuição a priori uma Normal, $h(\theta) \equiv h(\theta;\mu,\tau^2) \equiv N(\mu,\tau^2)$, a análise desenvolvida nos Ex. 1.15 e 1.16 mostra que a distribuição a posteriori é ainda Normal, $h(\theta \mid \mathbf{x}) \equiv h(\theta;\mu',\tau'^2)$, com μ' dado por (1.30) e τ'^2 dado por (1.31). A família conjugada da Normal (Variância dada) é portanto a própria Normal.

Distribuição Normal (Média dada). Tem-se, parametrizando em termos da precisão, $\eta = 1/\sigma^2$,

$$f(\mathbf{x} \mid \eta) \propto \eta^{N/2} \exp\{-[\Sigma(x_i-\theta)^2/2]\eta\}.$$

Se a distribuição a priori for uma Gama com parâmetros $\alpha_0 > 0$, $\beta_0 > 0$,

$$h(\eta) \equiv h(\eta;\alpha_0,\beta_0) \propto \exp\{-\alpha_0\eta\}\eta^{\beta_0-1},$$

vem, a posteriori,

$$h(\eta \mid \mathbf{x}) \equiv h(\eta;\alpha_1,\beta_1) \propto \exp\{-[\alpha_0 + \Sigma(x_i-\theta)^2/2]\eta\}\eta^{\beta_0+(N/2)-1},$$

que se trata de uma Gama com parâmetros,

$$\alpha_1 = \alpha_0 + \Sigma(x_i-\theta)^2/2,$$
$$\beta_1 = \beta_0 + \frac{N}{2}.$$

Portanto, a conjugada da Normal (Média dada) é a Gama.

Se a parametrização for feita em termos do desvio padrão (ou da variância) a conjugada é a Gama Inversa [veja-se Raiffa e Schlaifer (1961)].

Distribuição Normal. Com $X_i \sim N(\theta, \eta^{-1})$, toma-se a priori, $h(\theta, \eta) = h(\theta \mid \eta)h(\eta)$, com $h(\theta \mid \eta)$ Normal e $h(\eta)$ Gama, caso em que a distribuição conjunta se diz Normal-Gama. Mostra-se [De Groot (1970)] que a distribuição a posteriori é ainda uma Normal-Gama.

1.7 Decisão Estatística

Apesar de o homem ser chamado diariamente a tomar decisões, só muito recentemente os problemas que estas suscitam começaram a ser tratados segundo uma óptica científica. Para evitar qualquer mal entendido importa especificar em termos precisos no que consiste a atitude científica relativamente à tomada de decisões.

Primeiro, a teoria que vai desenvolver-se diz respeito à decisão individual e não de grupo. A diferença entre decisão individual e de grupo é funcional e não socio-biológica: qualquer decisor, seja indivíduo ou organização, do qual se pode pensar que tem propósitos ou objectivos unitários, deve tratar-se como um indivíduo.

Segundo, a teoria que vai estudar-se não pretende substituir o decisor — normalmente um indivíduo dotado de personalidade — mas, sim, fornecer um conjunto de regras que auxiliem o decisor. Como essas regras devem ser fundamentadas na lógica e na razão é evidente que têm de ser despidas de quaisquer factores emocionais.

Terceiro, a teoria que vai abordar-se é do tipo normativo — indica qual deve ser o comportamento dos decisores coerentes — e não do tipo descritivo — não analisa o comportamento concreto dos seres humanos quando tomam decisões nas mais variadas circunstâncias.

Em termos muito 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. A necessidade de escolher significa: 1.º) há que proceder a uma afectação irrevogável de recursos; 2.º) existem várias afectações possíveis.

A primeira tarefa que o decisor deve efectuar consiste no arrolamento ou identificação de todas as acções possíveis. Os cursos de acção alternativos que se abrem ao decisor devem fazer parte de um conjunto construido atendendo às duas propriedades seguintes: exaustividade — pode ser erro grave ignorar ou não detectar possibilidades — e exclusividade — convém evitar duplicações e afastar qualquer comportamento que leve a escolher mais do que uma acção.

O conjunto de acções possíveis designa-se por A; cada $a \in A$ representa uma acção diferente. O decisor tem forçosamente de escolher um $a \in A$ (exaustividade)

e não pode escolher mais do que um $a \in A$ (exclusividade). Quando A é finito, escreve-se $A = \{a_1, a_2, \ldots, a_m\}$.

A aparente simplicidade com que se descreve o arrolamento das acções alternativas não deve criar ilusões; de facto, no estudo de problemas concretos, ao procurar definir todos os caminhos depara-se, em regra, com enormes dificuldades de ordem técnica. Importa também notar que «nada fazer» é quase sempre uma das acções possíveis.

As decisões não são tomadas no «vazio»; um decisor quando considera qualquer uma das acções atende normalmente aos diferentes «cenários» ou «estados» em que se pode apresentar o «meio ambiente» em que está inserido. A descrição do meio nos seus aspectos alternativos pode revestir diversos graus de complexidade; por exemplo, nos problemas empresariais envolve, entre outros, aspectos sociais, económicos, políticos e tecnológicos. Casos há, todavia, em que se consegue identificar um conjunto de factores, de sistemas ou de circunstâncias particularmente relevantes.

A designação «estados da natureza» encontra-se consagrada e supõe-se corresponder à variedade de circunstâncias, situações ou acontecimentos susceptíveis de caracterizar o meio envolvente em que o decisor se move. Supõe-se, em geral, que é viável definir — de forma exaustiva e exclusiva —o conjunto de estados da natureza.

O conjunto de estados da natureza designa-se por Θ; cada $\theta \in \Theta$ representa um estado diferente; concretiza-se sempre um $\theta \in \Theta$ (exaustividade), e nunca pode concretizar-se mais do que um $\theta \in \Theta$ (exclusividade). É útil, por vezes, pensar que o conjunto Θ representa as alternativas que se oferecem à «natureza». A definição de Θ pode também revestir-se de grandes dificuldades de ordem técnica. Quando Θ é finito, escreve-se $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$.

É lícito pensar que a escolha exigida em qualquer problema de decisão depende de uma avaliação das acções em função das consequências que arrastam e das preferências do decisor por essas consequências. Somente interessam para o presente estudo as escolhas feitas em condições de incerteza, isto é, sem o decisor conhecer antecipadamente o «verdadeiro» estado da natureza.

O conjunto das consequências pode associar-se com o produto cartesiano, $\Theta \times A$; cada par ordenado, (θ, a), designa uma consequência, isto é, o resultado da interacção de uma acção, $a \in A$, e de um estado da natureza, $\theta \in \Theta$.

Exemplo 1.29 — Considere-se um produtor que acaba de vender uma peça a um cliente e tem de decidir se procede ou não à inspecção antes de efectuar a respectiva remessa.

As acções são duas: a_1 — inspeccionar, a_2 — não inspeccionar. Admita-se que os estados da natureza são também dois: θ_1 — a peça está em boas condições, θ_2 — a peça está em más condições. Se o produtor resolve inspeccionar e a peça está em

boas condições (acção a_1 combinada com o estado θ_1) a consequência, c_{11}, traduz-se pela satisfação do cliente e pelos custos decorrentes do trabalho de inspecção. Apreciando as outras três combinações obtém-se o quadro seguinte:

	θ_1: peça em boas condições	θ_2: peça em más condições
a_1: inspeccionar	c_{11}: cliente satisfeito mais custo da inspecção	c_{21}: procurar outra peça em boas condições mais custo da inspecção
a_2: não inspeccionar	c_{12}: cliente satisfeito	c_{22}: reclamação do cliente e substituição por outra peça em boas condições

Ao considerar o quadro não se esqueça que o problema é o produtor decidir se inspecciona ou se vende sem inspeccionar sem saber em que reais condições se encontra a peça. [Lindley (1971b)]. □

Do exemplo anterior, que já deixa antever algumas dificuldades criadas pela natureza das consequências, pode passar-se para casos bem mais complexos envolvendo a invalidez ou perda de vidas humanas ou, com muito menor gravidade, traduzindo situações de falência ou quebra de «*goodwill*».

Para não bloquear o desenvolvimento da teoria da decisão há que de algum modo abrir caminho à avaliação ou quantificação das consequências. Podem, nesse contexto, seguir-se dois métodos: (1) pressupor a existência de uma função perca, $L(\theta, a)$, com domínio em $\Theta \times A$, e contra-domínio em $\mathbf{R}$ (os ganhos são percas negativas), traduzindo em «escudos» o resultado para o decisor de tomar a acção a quando o estado é θ; (2) investigar as condições em que existe uma função utilidade, $U(\theta, a)$, com domínio e contra-domínio como indicado para a função perca e tal que, designadamente, $U(\theta, a) > U(\theta', a')$ quando e só quando (θ, a) é preferível a (θ', a') do ponto de vista do decisor.

A existência de função utilidade e a sua relação com a função perca são questões a tratar no capítulo 2. O pressuposto (1) vai, no entanto, predominar daqui por diante.

Quando os conjuntos A e Θ são finitos, a função perca, se existir, é equivalente à matriz $m \times n$, $[L(\theta_j, a_i)]$, designada matriz perca.

Exemplo 1.30 — Um supermercado é abastecido diariamente com certo tipo de artigo perecível de procura aleatória. Como estabelecer uma política de aprovisionamento que atenda a que no fim do período os artigos não vendidos perdem

grande parte do seu valor e que a procura não satisfeita pode desviar os clientes para outros pontos de venda?

Os estados são os diferentes níveis de procura, $\Theta = \{0, 1, 2, \ldots, n\}$, e as acções os níveis de aprovisionamento, $A = \{0, 1, 2, \ldots, n\}$ (supõe-se que o decisor conhece a procura máxima, n). Admita-se,

$$L(\theta,a) = \begin{cases} 2(\theta - a) & \text{se} \quad \theta \geqq a, \\ 4(a - \theta) & \text{se} \quad \theta < a, \end{cases}$$

isto é, que a perca resultante de cada unidade de procura não satisfeita é de 2 e a de cada unidade não vendida é de 4. A matriz perca tem o seguinte aspecto:

$\theta \rightarrow$	0	1	2	...	n
0	0	2	4	...	$2n$
1	4	0	2	...	$2(n-1)$
$a \rightarrow 2$	8	4	0	...	$2(n-2)$
...	...	...	...	...	...
n	$4n$	$4(n-1)$	$4(n-2)$	...	0

□

Exemplo 1.31 — O ensaio de drogas para detectar compostos com determinadas propriedades biológicas é feito em laboratórios especialmente equipados. Em regra, duas acções são possíveis, $A = \{a_1, a_2\}$, onde a_1 consiste em rejeitar o composto quando se julga que tem pequeno ou nulo efeito e a_2 consiste em aceitar o composto, embora provisoriamente, para o submeter a testes mais intensos. Por outro lado, $\Theta = \{\theta_1, \theta_2\}$, em que θ_1 corresponde à situação em que o composto tem realmente pequeno ou nulo efeito e θ_2 à situação em que o efeito é de facto significativo. Com L_{12} e L_{21} números reais positivos, a matriz perca pode assumir, no caso presente, a forma,

	θ_1	θ_2
a_1	0	L_{12}
a_2	L_{21}	0

As consequências menos desejáveis são: (θ_2, a_1) — as que resultam de abandonar o composto quando na verdade pode ser útil; (θ_1, a_2) — as que resultam de continuar os ensaios com um composto desprovido de potencialidades. As decisões correctas conduzem a consequências que por hipótese não implicam qualquer perca [Dunnett (1972)]. □

Na teoria da decisão costuma distinguir-se entre decisão sem dados e decisão com dados.

A decisão sem dados implica a escolha de uma acção em face da função perca ou da matriz perca sem recorrer a experiências, observações ou dados estatísticos que possam eventualmente reduzir a incerteza em relação aos estados da natureza. A expressão sem dados dirige-se essencialmente a novas experiências ou observações e não quer significar que a informação anteriormente acumulada é posta de parte ou ignorada.

A decisão com dados ou decisão estatística implica que o conjunto das acções terminais (conjunto A) deve combinar-se com as opções decorrentes da possibilidade que se oferece ao decisor de realizar experiências. A colheita de informação estatística tem um nítido objectivo: melhorar o conhecimento sobre os estados para decidir, se possível, em posição mais vantajosa. A realização de experiências é vista como espécie de «espionagem» exercida sobre a natureza e pode revestir várias modalidades; a mais corrente nos problemas que aqui interessam é a amostragem casual.

Existem fundamentalmente três tipos de amostragem casual:

- amostragem de dimensão fixa, em que o número de elementos da amostra é predeterminado;
- amostragem dupla ou múltipla, em que se prevê a recolha de duas ou mais amostras, sendo a dimensão de cada amostra e o número máximo de amostras, seja K, fixado antecipadamente; a tiragem da κ-ésima amostra, $2 \leqq \kappa \leqq K$, depende dos resultados observados nas primeiras $\kappa - 1$ amostras e da apreciação dos mesmos segundo critérios também definidos previamente;
- amostragem progressiva ou «sequencial», em que o número de elementos a observar não é fixado antecipadamente; as observações são feitas uma a uma e a realização da κ-ésima observação depende do resultado das primeiras $\kappa - 1$ e da apreciação das mesmas segundo regras de paragem definidas inicialmente.

A possibilidade de experimentação torna necessário introduzir o conjunto, $\mathbf{E}$, das várias experiências admissíveis entre as quais o decisor tem de optar. No caso mais simples, em que o tipo de experiência se encontra bem definido, por exemplo amostragem casual de dimensão fixa, pode representar-se, $\mathbf{E} = \{e_0, e_1, e_2, \dots, e_N, \dots\}$, onde e_N corresponde a optar por uma amostra de dimensão N; evidentemente, e_0 corresponde a não fazer qualquer experiência (decisão sem dados). A escolha de um elemento $e \in \mathbf{E}$ equivale ao planeamento de uma experiência ou inquérito cujo custo deve logicamente fazer parte da economia do problema. Para o efeito pode passar a operar-se com função perca, $L(\theta, a, e)$, definida no produto cartesiano $\Theta \times A \times \mathbf{E}$.

Para limitar o presente estudo ignora-se o custo da experimentação e considera-se que esta consiste sempre na amostragem casual com dimensão fixa[52]. Assim, em termos gerais, o planeamento de uma experiência tal como aqui é entendido tem dois aspectos:

- especificação da variável ou vector aleatório a observar;
- fixação da dimensão da amostra, isto é, do número de observações.

Devem destacar-se neste contexto os seguintes pontos: 1) nas situações reais as possibilidades de experimentação podem ser muito limitadas; 2) a dimensão da amostra deve ser fixada por compromisso entre o custo da experiência e o custo das decisões incorrectas; acontece porém muitas vezes que a dimensão óptima depende do próprio estado e só por acaso pode ser fixada antecipadamente — são casos como esses que tornam vantajoso recorrer à amostragem múltipla ou à sequencial; 3) as considerações de ordem estatística associadas ao planeamento de experiências não são menos importantes que as que presidem à análise dos dados produzidos.

A observação de uma variável ou vector aleatório, X, só é relevante para um dado problema de decisão quando o respectivo comportamento probabilístico depende do estado θ, $\theta \in \Theta$. Assim, quando θ percorre o conjunto Θ — e não é por acaso que se adapta o mesmo símbolo para o conjunto de estados e o espaço do parâmetro — obtém-se uma família de funções densidade de probabilidade ou de probabilidade,

$$\mathcal{F} = \{f(x \mid \theta) \,:\, \theta \in \Theta\},$$

e reencontra-se (1.1). Quando isso for conveniente pode trabalhar-se com a família de medidas de probabilidade — (1.2) — ou de funções de distribuição — (1.3).

Feita a distinção entre decisão sem dados e decisão com dados importa considerar alguns princípios de decisão. Advirta-se desde já que o confronto entre

[52] Quer dizer, no presente texto não se trata da chamada análise preposteriori que diz respeito ao planeamento ou escolha da experiência ou modalidade de inquérito; a designação deve-se aos bayesianos e tem justificação no facto de envolver problemas anteriores à recolha dos dados. O objectivo da análise preposteriori é minimizar o custo global que resulta de agregar o custo (ou perda) da decisão e o custo de conduzir e analisar a experiência ou inquérito.
Em geral pode escrever-se, $L(\theta, a, e) \equiv L(\theta, a, s, N)$, onde s é o processo de recolha e N o número de observações; uma hipótese simples consiste em admitir que,

$$L(\theta, a, s, N) = L(\theta, a) + C(s, N),$$

ou,

$$L(\theta, a, N) = L(\theta, a) + C(N),$$

quando $s \equiv$ observar uma amostra de dimensão fixa. No segundo caso a análise preposteriori tem por objectivo determinar a dimensão óptima da amostra [para a análise preposteriori Raiffa e Schlaifer (1961) é uma boa referência].

bayesianos e frequencistas se transporta para o campo da teoria da decisão com curiosos exemplos de coexistência.

Como as decisões são feitas em condições de incerteza, a perca em que se acaba de incorrer, $L(\theta, a)$, nunca é conhecida, pelo menos antecipadamente. Um procedimento que se afigura coerente é o que consiste em minimizar a perca esperada [ou maximizar a utilidade esperada — veja-se capítulo 2]. É preciso no entanto ter em atenção as diferentes versões de perca esperada que são propostas no domínio da decisão.

Para um decisor bayesiano, no momento da decisão, em que conhece o resultado da experiência, $X = x$, a única incerteza é a que diz respeito a θ. A «espionagem» que consiste na observação de $X = x$ permite-lhe alterar o sistema de credibilidade dos estados da natureza através da passagem da distribuição a priori, $h(\theta)$, para a distribuição a posteriori, $h(\theta \mid x)$. Consequentemente interessa-lhe considerar a perca esperada bayesiana ou risco a posteriori,

$$r_x(a) = \int_\Theta L(\theta, a) h(\theta \mid x)\, d\theta,$$

associado com cada $a \in A$, escolhendo, se existir, a acção a_h que minimiza $r_x(a)$. [Note-se que na decisão sem dados a perca esperada bayesiana assume a forma, $r(a) = \int_\Theta L(\theta, a) h(\theta)\, d\theta$]. Quando $r_x(a_h) \leqq r_x(a)$ para todo o $a \in A$, diz-se que a_h é a acção Bayes contra h; o decisor que escolhe a_h diz-se que segue o princípio Bayes condicional.

Para um decisor frequencista, muito mais do que para um bayesiano, o conceito de função de decisão é fundamental quando está em causa a instituição de princípios de decisão.

Em presença do resultado da experiência, isto é, depois de observado um $x \in \mathcal{X}$, a escolha de uma acção, $a \in A$, deve atender a esse resultado, pelo menos em princípio. Aqui se insere o importante conceito de função de decisão, δ, aplicação do espaço da amostra sobre o conjunto de acções, $\delta : \mathcal{X} \to A$. O campo de escolha do decisor desvia-se do conjunto de acções para o conjunto de funções de decisão, seja D. Ao estabelecer o modo como procede à interpretação dos resultados da experiência pode pensar-se que o decisor procede na verdade à escolha de uma função de decisão, isto é, de um elemento $\delta \in D$.

Exemplo 1.32 — Para ensaiar um medicamento no tratamento do cancro administra-se o mesmo a um grupo de cobaias e estuda-se a natureza dos efeitos. Em certas experiências regista-se o número de curas conseguido. Noutras experiências procede-se a uma graduação dos efeitos depois de implantar células cancerosas; uma parte das cobaias é tratada com o medicamento em questão e as restantes são mantidas em condições completamente idênticas excepto que não recebem aplicações

do tratamento. Decorrido um certo período de tempo os tumores são removidos e pesados. O problema da decisão estatística consiste em interpretar os resultados da experiência com o propósito, por exemplo, de aceitar ou rejeitar o medicamento como eficaz.

No primeiro tipo de ensaio o espaço de resultados é o número possível de curas nos N animais da amostra, $\mathcal{X} = \{0, 1, 2, \dots, N\}$. Logicamente, quanto maior for o número de curas, x, $x \in \mathcal{X}$, mais inclinado está o investigador a aceitar as propriedades do medicamento — acção a_1; quanto menor for x maior é a propensão para não reconhecer essas propriedades — acção a_2 . Seja ν um inteiro, $0 \leqq \nu \leqq N + 1$; uma classe de funções de decisão que decorre das anteriores considerações é a seguinte:

$$\delta_\nu(x) = \begin{cases} a_1 & \text{se} \quad x \geqq \nu, \\ a_2 & \text{se} \quad x < \nu. \end{cases}$$

Fazendo ν percorrer a sucessão, $0, 1, 2, \dots, N+1$, obtém-se $N+2$ funções de decisão algumas das quais fazem pouco sentido. Por exemplo, $\delta_0(x) \equiv a_1$, implica a aceitação do medicamento independente do resultado do ensaio; $\delta_{N+1}(x) \equiv a_2$, a rejeição qualquer que seja o número de curas. Algumas questões podem então formular-se. Como fixar o número crítico de curas, ν? Que outras funções de decisão podem propor-se?

No segundo tipo de ensaio, designe N_1 o número de cobaias sujeitas a tratamento e N_2 o número de cobaias piloto. Sejam,

$$x_{11}, x_{12}, \dots, x_{1N_1} \quad \text{e} \quad x_{21}, x_{22}, \dots, x_{2N_2},$$

os pesos dos respectivos tumores. O espaço dos resultados é conjunto de $\mathbf{R}^{N_1+N_2}$. Os pesos médios em cada uma das amostras são, respectivamente,

$$\overline{x}_1 = \frac{\Sigma x_{1i}}{N_1} \quad \text{e} \quad \overline{x}_2 = \frac{\Sigma_{2i}}{N_2}.$$

Com ω número real, uma classe de funções de decisão que parece de sugerir é a seguinte:

$$\delta_\omega(\mathbf{x}) = \begin{cases} a_1 & \text{se} \quad \overline{x}_2 - \overline{x}_1 \geqq \omega, \\ a_2 & \text{se} \quad \overline{x}_2 - \overline{x}_1 < \omega, \end{cases}$$

onde $\mathbf{x} = (x_{11}, x_{12}, \dots, x_{1N_1}, x_{21}, \dots, X_{2N_2})$. Se ω é grande, o medicamento só é aceite quando se verifica uma elevada diferença para mais entre os pesos médios dos tumores das cobaias não tratadas e tratadas; à medida que se tomam valores de ω mais pequenos o ensaio vai-se tornando menos exigente. Quando este tipo de função de decisão é aplicado — e é muitas vezes — há que fixar a diferença crítica, ω, de forma adequada[53]. □

[53] Adaptado de Dunnett (1972).

Em tudo o que se segue trabalha-se sempre com funções de decisão mensuráveis. Quer dizer, quando em qualquer função de decisão, $\delta \in D$, se toma como argumento a variável aleatória, X, obtém-se uma outra variável aleatória, $\delta(X)$.

Quando emprega $\delta \in D$ e observa $x \in \mathfrak{X}$, o decisor incorre numa perca igual a $L[\theta, \delta(x)]$. Quando o argumento é a variável aleatória, X, a perca, $L[\theta, \delta(X)]$ é também uma variável aleatória (desde que se suponha que, para cada $\theta \in \Theta$, $L[\theta, \delta(x)]$ é função mensurável de x). Assim, para um frequencista tem importante papel o valor esperado de $L[\theta, \delta(X)]$,

$$R(\theta, \delta) = E_\theta\{L[\theta, \delta(X)]\} = \int_{\mathfrak{X}} L[\theta, \delta(x)] f(x \mid \theta)\, dx, \tag{1.67}$$

que se designa por função risco ou perca esperada. A função, $R(\theta, \delta)$, definida em $\Theta \times D$, exprime a perca média sofrida pelo decisor quando emprega a função de decisão δ e o estado é θ.

A função risco usa-se, muitas vezes, para comparar funções de decisão: se com $\delta_1, \delta_2 \in D$, se tem $R(\theta, \delta_1) \leqq R(\theta, \delta_2)$ para todo o $\theta \in \Theta$, verificando-se a desigualdade estrita pelo menos para um θ, diz-se que δ_1 domina estritamente δ_2[54]. Uma função de decisão que não é dominada estritamente por qualquer outra diz-se admissível; caso contrário diz-se inadmissível. O conceito de admissibilidade é estudado mais a fundo no capítulo 4.

Parece intuitivo que as funções de decisão inadmissíveis não devem ser usadas, a menos que tenham atractivos de outra ordem, por exemplo, facilidade de cálculo. No entanto convém não perder de vista que a função risco é obtida — leia-se (1.67) — por integração sobre o espaço da amostra, $\mathfrak{X}$, isto é, considerando todos os possíveis valores de x e não exclusivamente o particular x observado. Trata-se de uma medida de precisão inicial que pode não ser uma boa medida de precisão final. É por esse facto que autores como Berger (1980) dizem: «*Using expected loss to investigate the performance of a decision rule seems on the one hand reasonable, but on the other hand appears to be rather arbitrary*».

Apesar do conceito de função risco não ser compatível com o princípio de verosimilhança, a sua aplicação encontra-se muito generalizada, naturalmente no campo clássico e mesmo, até certo ponto, nas fileiras bayesianas.

Por uma parte, é perfeitamente lógico usar $R(\theta, \delta)$ nas situações repetitivas (por exemplo, inspecção por amostragem) ou no planeamento de experiências (por exemplo, determinação da dimensão da amostra), pois faz sentido considerar o que se passa quando, virtualmente, x percorre $\mathfrak{X}$.

[54] Ou que δ_1 é R-melhor do que δ_2. Apesar da exposição dos próximos parágrafos na prática diz-se apenas que δ_1 é melhor do que δ_2

Por outra parte, quando não existe ou não se aceita uma distribuição a priori, como sucede com os clássicos, $R(\theta, \delta)$, parece não ter concorrentes como instrumento orientador na escolha de boas funções de decisão. Com os bayesianos, se não está em causa a distribuição a priori como elemento do modelo, já parecem ser inescapáveis as dificuldades levantadas em torno da sua correcta eliciação, pelo menos em condições de assegurar o sucesso de um puro procedimento bayesiano. A desculpa de alguns bayesianos para usar a função risco é então a falta de tempo (!). E afirmam: se o tempo disponível para investigação fosse infinito poderia sempre determinar-se exactamente a distribuição a priori e consequentemente a distribuição a posteriori e evitar a violação do princípio de verosimilhança. Porquê? Porque em tal caso não seria necessário recorrer à função risco por haver como alternativa a função risco a posteriori. No entanto, como nunca se dispõe de tempo infinito, a mor das vezes tem de determinar-se a distribuição a priori aproximadamente e — reconhecem — a melhor forma de avaliar a «bondade» de tal aproximação é ainda através da função risco[55].

Para algumas funções de decisão e para certos valores do parâmetro θ, o integral (1.67) pode ser $\pm\infty$ ou pode mesmo não existir, caso em que a variável aleatória, $L[\theta, \delta(X)]$, não tem média. Para ladear essa dificuldade toma-se para D a classe de funções de decisão tais que (1.67) existe e é finito para todo o $\theta \in \Theta$.

Com D finito, $D = \{\delta_1(x), \delta_2(x), \ldots, \delta_r(x)\}$, e Θ finito, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, a função risco, $R(\theta, \delta)$, corresponde à matriz $r \times n$,$[R(\theta_j, \delta_k)]$, designada matriz risco ou perca esperada.

A comparação das matrizes $[L(\theta_j, a_i)]$ e $[R(\theta_j,\delta_k)]$ — muito embora referentes aos casos particulares $A \in \Theta$ ou $D \in \Theta$ finitos — permite afirmar que os problemas de decisão sem dados têm uma certa analogia formal com os problemas de decisão com dados.

Exemplo 1.33 — O clássico exemplo do barómetro, apesar da sua extrema simplicidade, é sempre instrutivo. Determinado indivíduo, que tem conveniência em sair de casa em dia de tempo incerto, contempla três acções:

a_1 — ficar em casa;
a_2 — sair sem guarda-chuva;
a_3 — sair com guarda-chuva.

Os estados da natureza são dois: θ_1 — chuva; θ_2 — bom tempo.

[55] Essa forma de avaliação pode consistir em comparar as funções risco, $R(\theta, \delta_{h'})$, $R(\theta, \delta_{h''})$, ..., onde $\delta_{h'}, \delta_{h''}, \ldots$, são funções de decisão Bayes contra distribuições a priori, $h', h'', \ldots$, propostas como aproximações da verdadeira distribuição a priori. O conceito de função de decisão Bayes é estudado na secção 3.4; a análise da robustez em termos de risco — secção 3.6 — é particularmente instrutiva.

A matriz perca é a seguinte,

	θ_1 [chuva]	θ_2 [b. tempo]
a_1 [ficar em casa]	4	4
a_2 [sair sem guarda-chuva]	5	0
a_3 [sair com guarda-chuva]	2	5

Antes de tomar uma decisão o indivíduo pode efectuar uma «experiência»: observar um barómetro. Suponha-se, para facilitar, que a leitura do barómetro — observação da variável aleatória X — pode fornecer apenas dois valores: x_1 (indicação de chuva) e x_2 (indicação de bom tempo). Tem-se, portanto, $\mathcal{X} = \{x_1, x_2\}$. O barómetro oferece uma relativa orientação, mas, como todas as experiências (não determinísticas), não é infalível. Admita-se que o seu comportamento, para cada estado, é dado pelas seguintes funções de probabilidade:

$$P(X = x_1 \,|\, \theta_1) = f(x_1 \,|\, \theta_1) = 0{,}8; \quad P(X = x_2 \,|\, \theta_1) = f(x_2 \,|\, \theta_1) = 0{,}2;$$
$$P(X = x_1 \,|\, \theta_2) = f(x_1 \,|\, \theta_2) = 0{,}1; \quad P(X = x_2 \,|\, \theta_2) = f(x_2 \,|\, \theta_2) = 0{,}9;$$

que exprimem:

- se o estado é chuvoso, o barómetro acerta em 80% dos casos;
- se o estado é bom tempo, o barómetro prevê com correcção em 90% dos casos.

Com $\mathcal{X} = \{x_1, x_2\}$ e $A = \{a_1, a_2, a_3\}$ o número total de funções de decisão é $3^2 = 9$, conforme passa a enumerar-se,

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
x_1	a_1	a_2	a_3	a_1	a_2	a_1	a_3	a_2	a_3
x_2	a_1	a_2	a_3	a_2	a_1	a_3	a_1	a_3	a_2

O cálculo da matriz risco processa-se sem qualquer dificuldade. Por exemplo,

$$\begin{aligned} R(\theta_1, \delta_4) &= L[\theta_1, \delta_4(x_1)] f(x_1 \,|\, \theta_1) + L[\theta_1, \delta_4(x_2)] f(x_2 \,|\, \theta_1) \\ &= L(\theta_1, a_1) f(x_1 \,|\, \theta_1) + L(\theta_1, a_2) f(x_2 \,|\, \theta_1) \\ &= 4(0{,}8) + 5(0{,}2) = 4{,}2; \end{aligned}$$

$$\begin{aligned} R(\theta_2, \delta_4) &= L[\theta_2, \delta_4(x_1)] f(x_1 \,|\, \theta_2) + L[\theta_2, \delta_4(x_2)] f(x_2 \,|\, \theta_2) \\ &= L(\theta_2, a_1) f(x_1 \,|\, \theta_2) + L(\theta_2, a_2) f(x_2 \,|\, \theta_2) \\ &= 4(0{,}1) + 0(0{,}9) = 0{,}4. \end{aligned}$$

Procedendo do mesmo modo com os restantes elementos $R(\theta_j, \delta_k)$, obtém-se,

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
θ_1	4	5	2	4,2	4,8	3,6	2,4	4,4	2,6
θ_2	4	0	5	0,4	3,6	4,9	4,1	4,5	0,5

Note-se que $A \subset D$ pois as funções de decisão $\delta_1(x) \equiv a_1$, $\delta_2(x) \equiv a_2$, $\delta_3(x) \equiv a_3$, desprezam a informação dada pela experiência — decisor de ideias fixas — e são equivalentes respectivamente a a_1, a_2 e a_3.

Avançar com a escolha de um $a \in \{a_1, a_2, a_3\}$, no caso de não haver barómetro, ou de um $\delta \in \{\delta_1, \delta_2, \ldots, \delta_9\}$, no caso de haver barómetro, seria uma antecipação relativamente ao que vai expor-se adiante. O exemplo deixa, no entanto, patente, a analogia formal entre os dois problemas [Lindgren (1976)]. □

A função risco, sendo embora um importante instrumento, não permite ir muito longe na selecção de funções de decisão pela simples razão de que há muitas funções de decisão admissíveis [ou se se quiser, como se explica melhor no capítulo 3, não há funções de decisão com risco uniformemente mínimo].

Restringindo por agora a atenção a critérios globais [veja-se a secção 3.2] podem introduzir-se dois princípios de decisão. O primeiro puramente frequencista, o segundo híbrido.

O princípio minimax admite que a natureza é o mais hostil possível e recomenda a escolha da função de decisão, se existir, que minimize o máximo risco. Isto é, recomenda que se escolha $\tilde{\delta} \in D$ tal que[56],

$$\sup_\theta R(\theta, \tilde{\delta}) = \inf_\delta \sup_\theta R(\theta, \delta).$$

O princípio risco Bayes exige que seja proposta para θ uma distribuição a priori ou função de ponderação, $h(\theta)$ [que para um bayesiano pode representar um estado de espírito e para um frequencista um simples instrumento matemático ou método razoável para obter funções de decisão]. Em resultado do compromisso a função risco ou perca esperada passa a escrever-se,

$$R(h, \delta) = \int_\Theta R(\theta, \delta) h(\theta)\, d\theta, \tag{1.68}$$

ou ainda,

$$R(h, \delta) = \int_\Theta \left\{ \int_{\mathfrak{X}} L[\theta, \delta(x)] f(x \mid \theta)\, dx \right\} h(\theta)\, d\theta. \tag{1.69}$$

[56] $\sup_\theta R(\theta, \delta)$ é o supremo de $R(\theta, \delta)$ para $\theta \in \Theta$. Etc. Veja-se nota do pé da página 146.

Se existir uma função de decisão, $\delta_h \in D$, tal que,

$$R(h, \delta_h) \leqq R(h, \delta) \text{ para todo o } \delta \in D,$$

essa função de decisão é óptima à luz do princípio e designa-se por função de decisão Bayes contra h; $R(h, \delta_h)$ representa o risco Bayes de h.

Os princípios Bayes condicional, minimax e risco Bayes são adiante retomados e desenvolvidos.

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 realizado por von Neumann e Morgenstern. O grande mérito de Wald [além de inúmeros resultados sobre admissibilidade, problemas de existência de soluções minimax e Bayes, classes completas, etc. — veja-se capítulo 4 e Wald (1950)], foi fornecer uma estrutura unificada no quadro da qual pode estudar-se a grande maioria dos problemas de estatística. Em termos gerais, os procedimentos clássicos são casos particulares de decisão estatística. Na estimação, tem-se $A \equiv \Theta$ e uma acção $a \in A$ corresponde a propor uma estimativa de θ; as funções de decisão são estimadores [veja-se Ex. 3.4 e o capítulo 6]; no ensaio da hipótese H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$, tem-se, $A = \{a_0 \equiv$ aceitação de $H_0, a_1 \equiv$ rejeição de $H_0\}$; as funções de decisão generalizam o conceito de região crítica [veja-se Ex. 1.32 e o capítulo 7].

O impulso de Wald fez-se também sentir no domínio do planeamento de experiências, nomeadamente através da concepção de novos processos de amostragem (amostragem sequencial), a da multidecisão (veja-se capítulo 8). São de Wald (1950) estas palavras: «*Until about ten years ago, the available statistical theories, except scattered results, were restricted in two important respects: (1) experimentation was assumed to be carried out in a single stage; (2) decision problems were restricted to problems of testing hypothesis, and that of point or interval estimation. The general theory, as given in this book, is freed from both of these restrictions. It allows for multi-stage experimentation and includes the general multi-decision problem*».

O enfoque introduzido pela chamada Escola de Savage inclui um apelo mais profundo às ideias bayesianas. Não é que Wald — que pode considerar-se um frequencista — não tenha considerado distribuições a priori e soluções Bayes; fê-lo, contudo, sem se envolver em questões de princípio e encarando os procedimentos bayesianos puramente como instrumento matemático.

CAPÍTULO 2

TEORIA DA UTILIDADE

2.1 Generalidades

As consequências reflectem, já foi dito, a interacção entre o decisor e a «natureza»; a representação por um par simbólico, (θ, a), $(\theta, a) \in \Theta \times A$, é imagem simplificada imprescindível ao tratamento matemático e não deve iludir quanto à complexidade que na prática podem revestir. De facto, uma descrição completa das consequências, prospectivas ou historiais futuros pode ser difícil ou até impossível, mesmo quando se omitem pormenores pouco relevantes.

A teoria da utilidade é um dos caminhos que pode seguir-se na avaliação das consequências potenciais das várias acções alternativas. Propõe um modelo para o comportamento dos indivíduos quando enfrentam situações de risco ou incerteza, isto é, define com precisão o que deve entender-se por comportamento consistente e coerente em tais situações. Demonstra, concomitantemente, que um indivíduo consistente e coerente atribui, implicitamente, valores numéricos ou utilidades às consequências e opta pela acção conducente à maximização da utilidade. A atribuição de valores numéricos às consequências corresponde, ainda segundo a teoria, ao estabelecimento de uma ordem de preferências entre as diferentes opções caindo então a escolha na preferida.

Assim, a teoria da utilidade é normativa: prescreve a forma como os indivíduos devem reagir e fornece um modelo de comportamento «racional». Não pretende descrever a forma como os indivíduos reagem na vida real, mas tem o propósito de tornar «racionais» os que não o são depois de os convencer das bases lógicas da teoria.

O modelo proposto pela teoria da utilidade não tem aceitação geral, o que não surpreende quando se assinala que tem natureza subjectiva. Não podendo entrar-se na discussão filosófica do problema, resta enfatizar que se trata de um dos vários modelos possíveis para enquadramento da tomada de decisões em face do risco e da incerteza. Trata-se de um modelo passivo de contestação como qualquer outro, embora em muitos casos concretos, sobretudo de natureza económica, os aspectos subjectivos sejam diluídos. Quando assim sucede, atenua-se também a controvérsia acerca da sua validade.

2.2 Função utilidade

Para desenvolver a exposição designa-se por $\mathbf{C}$ o conjunto de consequências ou prospectivas e por c o elemento genérico de $\mathbf{C}$. No caso do Ex. 1.29, tem-se, $\mathbf{C} = \{c_{11}, c_{12}, c_{21}, c_{22}\}$.

O decisor nunca tem de posicionar-se face a uma consequência certa ou determinada por um motivo muito simples: a incapacidade de prever exactamente o estado da natureza que vinga. Assim, tem de admitir-se que o carácter contingente ou aleatório dos estados coloca o decisor na situação em que tem de escolher ou optar por distribuições de probabilidade sobre $\mathbf{C}$ ou lotarias.

Uma lotaria ou consequência mista é um jogo em que os prémios são consequências puras, $c_1, c_2, \ldots \in \mathbf{C}$, que saiem com probabilidades conhecidas, p_1, $p_2, \ldots, p_i \geqq 0, \Sigma p_i = 1$.

No caso do Ex. 1.29, se for possível admitir que os estados da natureza, θ_1 (peça em boas condições) e θ_2 (peça em más condições) tem probabilidade, respectivamente, 9/10 e 1/10, conclui-se que optar entre as acções a_1 e a_2 equivale a optar entre as lotarias,

	LOTARIA [1]:			LOTARIA [2]:	
	prem.	prob.		prem.	prob.
a_1	c_{11}	9/10	a_2	c_{12}	9/10
	c_{21}	1/10		c_{22}	1/10.

Este exemplo procura mostrar que o conceito de lotaria não é tão excêntrico como a primeira impressão podia levar a concluir. Exemplo mais pertinente é o de um investigador que emprega métodos estatísticos para realizar inferências sobre determinada parâmetro. A consequência (ou recompensa) dos procedimentos que contempla é a quantidade de informação recolhida. Evidentemente, o investigador pode debruçar-se sobre vários esquemas de experimentação ou de amostragem; no entanto, a quantidade de informação obtida é sempre aleatória (qualquer que seja o esquema há sempre amostras mais ou menos «felizes» ou «infelizes»).

Em qualquer hipótese a escolha é sempre entre distribuições de probabilidade ou «lotarias» em que os prémios são diferentes quantidades de informação.

Mais adiante faz-se referência a lotarias em que as probabilidades dos prémios não são conhecidas, mas, sim, imperfeitamente conhecidas ou subjectivas. Por exemplo, um investidor quando decide em relação a um projecto considera como prémios os «*cash flows*» prospectivos gerados pelo negócio. No entanto, a incerteza em relação ao complexo sistema de factores (estados) que condicionam os resultados do projecto (concorrência interna e externa, intervenção dos poderes públicos, preços das matérias, etc.), não permite probabilizar que não seja subjectivamente as diferentes consequências da decisão.

Entre os dois tipos de lotarias existe a diferença que se verifica, por exemplo, entre o Totoloto (ou a Lotaria da Santa Casa) e o Totobola. Os anglo-saxões falam em «*roulette lotteries*» e «*horse lotteries*».

Quando $\mathbf{C}$ tem k elementos distintos a lotaria genérica pode representar-se,

$$c^* = [c_1, c_2, \dots, c_k]_{(p_1, p_2, \dots, p_k)}, \quad p_i \geqq 0, \Sigma p_i = 1. \tag{2.1}$$

O conjunto de lotarias sobre $\mathbf{C}$ designa-se por $\mathbf{C}^*$.

Cada $c_i \in \mathbf{C}$ é uma lotaria particular — lotaria degenerada — correspondente ao caso em que $p_i = 1$ e $p_j = 0, j \neq i$.

Se uma lotaria tem apenas dois prémios escreve-se, simplesmente,

$$[c_1, c_2]_p, \quad 0 < p < 1, \tag{2.2}$$

uma vez que $p_1 = p$, $p_2 = 1 - p$ e $p_j = 0$, $j = 3, 4, \dots, k$.

Se $\mathbf{C}$ tem infinidade numerável de elementos, tem-se,

$$c^* = [c_1, c_2, \dots]_{(p_1, p_2, \dots)}, \quad p_i \geqq 0, \Sigma p_i = 1; \tag{2.3}$$

se $\mathbf{C}$ é conjunto não numerável pode em geral associar-se $\mathbf{C}^*$ à família de distribuições de probabilidade ou medidas de probabilidade sobre $\mathbf{C}$ [introduzindo uma álgebra-σ, Γ, de subconjuntos de $\mathbf{C}$ de modo a operar com um espaço de probabilidade, $(\mathbf{C}, \Gamma, \mathbf{C}^*)$]. A análise seguinte processa-se a nível que não carece de referência para o caso não numerável.

O conceito de comportamento a que se convenciona chamar «racional» pede à partida que o decisor tenha propósitos ou objectivos suficientemente claros para, em presença de duas lotarias quaisquer, c_1^* e c_2^*, pertencentes a $\mathbf{C}^*$, poder sempre afirmar qual delas prefere ou se lhe são indiferentes. Por outras palavras, o comportamento «racional» exige que o decisor tenha um padrão de preferências.

Um padrão de preferências é uma relação binária de suporte $\mathbf{C}^*$, simbolicamente representada por «$\succcurlyeq$»„ satisfazendo os seguintes axiomas:

{A1} — Para quaisquer, $c_1^*, c_2^* \in \mathbf{C}^*$, tem-se,

$$c_1^* \succcurlyeq c_2^* \quad \text{ou} \quad c_2^* \succcurlyeq c_1^*,$$

podendo verificar-se ambas as relações (propriedade conectiva).

{A2} — Se $c_1^*, c_2^*, c_3^* \in \mathbf{C}^*$ e se $c_1 \succcurlyeq c_2^*$ e $c_2^* \succcurlyeq c_3^*$, então, $c_1^* \succcurlyeq c_3^*$ (propriedade transitiva).

A relação, $c_1^* \succcurlyeq c_2^*$, indica que para o decisor a lotaria c_1^*; é pelo menos tão preferível como c_2^*.

Se $c_1^* \succcurlyeq c_2^*$ sem que se verifique $c_2^* \succcurlyeq c_1^*$ diz-se que c_1^* é preferível a c_2^* e escreve-se, $c_1^* \succ c_2^*$.

Quando $c_1^* \succcurlyeq c_2^*$ e $c_2^* \succcurlyeq c_1^*$, c_1^* e c_2^* são indiferentes, simbolicamente, $c_1^* \approx c_2^*$.

A proposição seguinte demonstra-se sem dificuldade,

Teorema 2.1 — «$\approx$» representa uma relação de equivalência. Para qualquer par de lotarias, $c_1^*, c_2^* \in \mathbf{C}^*$, verifica-se uma e uma só das seguintes relações:

$$c_1^* \succ c_2^*, \quad c_2^* \succ c_1^* \quad \text{ou} \quad c_1^* \approx c_2^*.$$

Se $c_1^* \succ c_2^*$ e $c_2^* \succcurlyeq c_3^*$, tem-se, $c_1^* \succ c_3^*$. □□

Função real, $U(c^*)$, com domínio em $\mathbf{C}^*$, diz-se função utilidade quando cumpre as seguintes condições:

{U1} — Com, $c^* = [c_1, c_2]_p$,

$$U(c^*) = U(c_1)p + U(c_2)(1-p); \tag{2.4}$$

quer dizer, a utilidade de tal lotaria é o valor esperado da utilidade dos prémios.

{U2} — Se $U(c_1^*) \geqq U(c_2^*)$ então $c_1^* \succcurlyeq c_2^*$ e reciprocamente; isto é, quando c_1^* é pelo menos tão preferível como c_2^* a utilidade de c_1^* não é inferior à utilidade de c_2^*.

Com c^* dado por (2.1), obtém-se de {U1}, por indução finita,

$$U(c^*) = \sum_{i=1}^{k} U(c_i)p_i. \tag{2.5}$$

Enquanto o padrão de preferências estabelece a ordenação das consequências ou lotarias, a função utilidade permite ir mais longe e introduzir uma escala numérica para as consequências ou lotarias, compatível com aquela ordenação, desde que, é claro, digam respeito ao mesmo decisor («racional»).

A função utilidade permite quantificar as consequências; a sua existência é, compreensivelmente (recordem-se as considerações feitas na secção 2.1), o problema central da teoria da utilidade.

Um decisor «racional», como tal detentor de padrão de preferências de acordo com as normas dos axiomas {A1} e {A2}, não usufrui necessariamente de função utilidade compatível. A afirmação não é inesperada, ponderando o que acima ficou dito acerca do maior alcance da função utilidade; de facto, para assegurar a existência desta torna-se necessário introduzir dois axiomas adicionais[1]:

{A3} — Se $c_1^*, c_2^* \in \mathbf{C}^*$, $c_1^* \succcurlyeq c_2^*$, tem-se para qualquer $0 < p < 1$ e para qualquer $c^* \in \mathbf{C}^*$,

$$[c_1^*, c^*]_p \succcurlyeq [c_2^*, c^*]_p.$$

{A4} — Dados, $c_1^*, c_2^*, c_3^* \in \mathbf{C}^*$, tais que, $c_1^* \succ c_2^* \succ c_3^*$, existem lotarias,

$$[c_1^*, c_3^*]_p \quad \text{e} \quad [c_1^*, c_3^*]q, \quad 0 < p, q < 1,$$

verificando,

$$c_1^* \succ [c_1^*, c_3^*]_p \succ c_2^* \succ [c_1^*, c_3^*]_q \succ c_3^*.$$

O sistema de axiomas {A1} – {A4} merece alguns comentários.

Têm sido concebidas consequências ou lotarias de tal forma enredadas que muitos indivíduos ficam perplexos ou completamente hesitantes quando se lhes pede para declararem a sua preferência. Outras opções podem ser de tal modo repugnantes ou terríveis que os indivíduos negam manifestar qualquer preferência. A possibilidade de tal indefinição é rejeitada por {A1}.

---

[1] Alguns autores introduzem ainda outro axioma (axioma das lotarias compostas): para qualquer lotaria composta (com prémios que são já lotarias), por exemplo, para simplificar, $c^* = [c_1^*, c_2^*]_p$, onde,

$$c_1^* = [c_1, c_2, \dots, c_k]_{(p_1, p_2, \dots, p_k)} \quad \text{e} \quad c_2^* = [c_1, c_2, \dots, c_k]_{(q_1, q_2, \dots, q_k)},$$

tem-se, $c^* \approx c^{**}$, com,

$$c^{**} = [c_1, c_2, \dots, c_k]_{(r_1, r_2, \dots, r_k)}, r_i = pp_i + (1-p)q_i, \quad i = 1, 2, \dots, k.$$

Para não introduzir este axioma adoptou-se o ponto de vista de DeGroot (1970): «*The allusion to lotteries is very important here because it properly conveys the following notion: When two probability distributions on R* [na presente notação **C**] *are compared, only the probability of receiving the various rewards are relevant, and the statistician need not take in consideration the particular events of lhe lottery that generated these probabilities*». Por outras palavras, supõe-se que no conceito de lotaria está implícito que o que interessa de facto são os prémios finais e as respectivas probabilidades, sendo imaterial que estas sejam obtidas por meio de uma tabela de números aleatórios ou, por exemplo, tirando primeiro uma carta de um bom baralho, rodando em seguida uma colorida tômbola e finalmente usando uma roleta construída com excelentes materiais. Em resumo, na presente análise o «gosto pelo jogo» está excluído.

A transitividade das preferências pode ser aceite com a simplicidade com que o faz Lindley (1971a): «*...transitivity is a modest requirement whose violation has peculiar and obviously undesirable properties*». Na prática têm sido observados experimentalmente casos em que a transitividade não se verifica e, por isso, a apreciação de Luce e Raiffa (1957) é menos pacífica: «*No matter how intransitivities arise, we must recognize they exist, and we can take little confort in the thought that they are an anathema to most of what constitutes theory in the behavioral sciences today. We may say that we are only concerned with behavior which is transitive, adding hopefully that we believe this need not always be a vacuous study. Or we may contend that the transitive description is often a ‹close› aproximation to reality. Or we may limit our interest to ‹normative› or ‹idealized› behavior in the hope that such studies will have a methatheoretic impact on more realistic studies. In order to get on, we shall be flexible and accept all of these as possible defences, and to them add the traditional mathematician's hedge: transitive relations are far more mathematically tractable than intransitive ones*».

A aceitação de {A3} parece menos discutível, pois, se numa lotaria com dois prémios, $[c_1^*, c^*]$, um destes, seja c_1^*, é substituído por um pior, seja $c_2^*, c_1^* \succ c_2^*$, a lotaria inicial é naturalmente preferível à nova lotaria.

O axioma {A4} volta a provocar controvérsia. Considerem-se três prémios ou lotarias degeneradas,

$$c_1 \equiv 1000\$, \quad c_2 \equiv 10\$ \quad \text{e} \quad c_3 \equiv \text{«morte»};$$

certamente, $c_1 \succ c_2 \succ c_3$. Logo, segundo o axioma em discussão, existe uma probabilidade, $0 < p < 1$, tal que a obtenção de um prémio de 1000\$ com probabilidade p e a obtenção da «morte» com probabilidade $1 - p$ é preferível a um prémio certo de 10\$.

$$[1000\$,\ \text{«morte»}]_p \succ 10\$.$$

A maioria das pessoas não aceita este pressuposto; outras podem aceitar se p for suficientemente grande (por exemplo, $p = 1 - 10^{-10}$) e tornar o prémio «morte» praticamente impossível.

Considerando três prémios arbitrários, tais que $c_1 \succ c_2 \succ c_3$, conclui-se que {A4} pressupõe a não existência:

1) de um c_3 tão terrível que iniba a lotaria que o combina com outro prémio c_1 de ser preferível ao prémio intermédio c_2,

ou,

2) de um c_1 tão maravilhoso que iniba a lotaria que o combina com outro prémio c_3 de ser preferida pelo prémio intermédio c_2.

Felizmente, na prática, em geral, nem a pior das consequências é terrivelmente má, nem a melhor das consequências é maravilhosamente boa.

A demonstração da existência de função utilidade com base nos quatro axiomas discutidos é feita na secção seguinte. Fica assim disponível uma solução — mesmo que contestada ou contestável — para o problema da quantificação das consequências[2] e feita, de certo modo, a consagração do princípio de maximização da utilidade de grande importância para a teoria da decisão.

Para chamar mais uma vez a atenção para o carácter normativo da teoria da utilidade, registando a frequência com que na prática as pessoas podem não a respeitar, fecha-se a presente secção com o conhecido exemplo do economista francês Allais [veja-se Savage (1954) e a discussão de Raiffa (1968)].

Considerem-se duas situações em cada uma das quais se comparam duas lotarias (prémios em milhares de contos):

Situação 1

Lotaria c_1^* ≡ prémio 5 com probabilidade 1,00;

Lotaria c_2^* ≡ prémio 25 com probabilidade 0,10,
prémio 5 com probabilidade 0,89 e
prémio 0 com probabilidade 0,01;

Situação 2

Lotaria c_3^* ≡ prémio 5 com probabilidade 0,11 e
prémio 0 com probabilidade 0,89;

Lotaria c_4^* ≡ prémio 25 com probabilidade 0,10 e
prémio 0 com probabilidade 0,90.

Allais notou que muitas pessoas [Raiffa fala que obteve esse resultado em centenas de experiências] preferem c_1^* a c_2^* e c_4^* a c_3^*. A declaração de que, $c_1^* \succ c_2^*$, é justificada nos seguintes termos: receber uma soma elevada (5 mil contos) de certeza é preferível a correr o risco de nada ganhar muito embora adquirindo a possibilidade de ganhar uma soma muito elevada (25 mil contos); a declaração de que, $c_4^* \succ c_3^*$, é justificada nos seguintes termos: a «chance» de ganhar é quase a mesma em ambas as lotarias, logo a que tem o prémio mais elevado é preferível.

A atitude que consiste em considerar,

$$c_1^* \succ c_2^* \quad \text{e} \quad c_4^* \succ c_3^*,$$

[2] Desde que interpretadas como lotarias em que os prémios saiem com probabilidades conhecidas ou objectivas.

é incompatível com a axiomática introduzida, porquanto, sendo U qualquer função utilidade decorrente dos quatro pressupostos, tem-se,

$$U(5) > 0{,}1U(25) + 0{,}89U(5) + 0{,}01U(0),$$
$$0{,}1U(25) + 0{,}9U(0) > 0{,}11U(5) + 0{,}89U(0),$$

o que é uma impossibilidade.

A propósito do exemplo de Allais, Raiffa (1968) tece considerações que seguem, mais ou menos, nas seguintes linhas: Numa óptica «descritiva» verifica-se que a maioria das pessoas não se comportam de acordo com os princípios da teoria da utilidade. Mas ninguém exige que a maioria das pessoas se comporte como «deve». Na verdade, a razão primeira para adoptar uma teoria «normativa» para o comportamento na escolha é a observação de que, quando o processo de decisão é deixado inteiramente a cargo do julgamento, sem ajuda formal, as opções são tomadas muitas vezes de modo internamente contraditório, e isso indica que o decisor talvez pudesse agir melhor do que age. Se todo o mundo se comportasse de acordo com a teoria normativa não havia motivo para falar desta e bastava simplesmente dizer ao decisor: faça o que lhe vier à cabeça.

Na secção seguinte, para evitar a proliferação de asteriscos, designam-se por c os elementos de $\mathbf{C}^*$, isto é, as lotarias degeneradas e não degeneradas, sendo clara do contexto a situação contemplada.

2.3 Existência de função utilidade

O sistema de axiomas {A1}–{A4} permite estabelecer as seguintes proposições [Lindgren (1971)]:

Teorema 2.2 — Se $c_1 \succ c_2$, então, $c_1 \succ [c_1, c_2]_p \succ c_2$, para todo o $p \in (0,1)$.

Dem. Consequência imediata de {A3} pois, para todo o $p \in (0,1)$,

$$c_1 \approx [c_1, c_1]_p \succ [c_1, c_2]_p \succ [c_2, c_2]_p \approx c_2.$$

□□

Teorema 2.3 — Dados c_1 e c_2, com $c_1 \succ c_2$, e c tal que $c_1 \succ c \succ c_2$, existe um único número $r \in (0{,}1)$ que faz $[c_1, c_2]_r \approx c$.

Dem. Por {A4} fica assegurada a existência de $p, q \in (0{,}1)$, conducentes à relação,

$$[c_1, c_2]_p \succ c \succ [c_1, c_2]_q,$$

sendo, como facilmente se verifica, $p > q$.

Seja,

$$\Pi = \{p \,:\, p \in (0,1) \text{ e } [c_1, c_2]_p \succ c\};$$

como o conjunto Π é limitado, existe o ínfimo, $\inf \Pi$. Tome-se $\inf \Pi = r$; tem-se, $[c_1, c_2]_r \approx c$, como passa a mostrar-se.

Se a indiferença não se verifica, pelo Teorema 2.1, $[c_1, c_2]_r \succ c$ ou $c \succ [c_1, c_2]_r$. No primeiro caso, vem $[c_1, c_2]_r \succ c \succ c_2$ e por {A4} existe $r' \in (0,1)$ tal que,

$$[c_1, c_2]_r \succ [[c_1, c_2]_r, c_2]_{r'} \succ c,$$

onde, na posição intermédia, se tem uma lotaria onde o prémio $[c_1, c_2]_r$ — que é já de si uma lotaria — sai com probabilidade r' e o prémio c_2 com probabilidade $1 - r'$. Por ser,

$$[[c_1, c_2]_r, c_2]_{r'} \approx [c_1, c_2]_{rr'},$$

a comparação desta relação com a anterior permite concluir ser $rr' \in \Pi$, com $rr' < r$, o que contradiz $r = \inf \Pi$. No segundo caso, $c_1 \succ c \succ [c_1, c_2]_r$, e por {A4} existe $r'' \in (0,1)$ tal que,

$$c \succ [c_1, [c_1, c_2]_r]_{r''} \succ [c_1, c_2]_r;$$

ora, como,

$$[c_1, [c_1, c_2]_r]_{r''} \approx [c_1, c_2]_{r'' + (1-r'')r},$$

deve ser $r'' + (1 - r'')r > r$ e ainda $p > r'' + (1 - r'')r$ para todo o $p \in \Pi$, pois,

$$[c_1, c_2]_p \succ c \succ [c_1, c_2]_{r'' + (1-r'')r}.$$

Logo, $p > r$ para todo o $p \in \Pi$ o que também contradiz $r = \inf \Pi$. □□

O Teorema 2.3 tem uma consequência notável: permite codificar todas as lotarias, c, intermédias a duas lotarias arbitrárias, c_1 e c_2, tomadas para referência.

Com, $c_1 \succ c \succ c_2$, convencionando, $U(c_2) = 0$ e $U(c_1) = 1$, faça-se $U(c) = r$ com $r \in (0,1)$ tal que $[c_1, c_2]_r \approx c$; é fácil de ver que a função assim definida é função utilidade: verifica {U1} e {U2}. Quanto a {U2} a conclusão é imediata:

$$c \approx [c_1, c_2]_r \succ c' \approx [c_1, c_2]_s,$$

implica,

$$U(c) = r \geqq U(c') = s,$$

e reciprocamente, como se conclui revendo a parte inicial da demonstração do Teorema 2.3. Quanto a {U1}: com c e c' lotarias intermédias, isto é, verificando,

$c_1 \succ c \succ c_2$ e $c_1 \succ c' \succ c_2$, tome-se a combinação, $[c, c']_p$; suponha-se ser $c \approx [c_1, c_2]_r$ e $c' \approx [c_1, c_2]_s$; então, vem $U(c) = r$, $U(c') = s$ e,

$$[c, c']_p \approx [c_1, c_2]_{rp + s(1-p)},$$

donde,

$$U\{[c, c']_p\} = rp + s(1-p) = U(c)p + U(c')(1-p).$$

A função instituída atribui univocamente uma utilidade numérica a cada lotaria intermédia. No entanto, qualquer função linear de U,

$$U^*(c) = \alpha U(c) + \beta, \alpha \text{ e } \beta \text{ números reais}, \alpha > 0, \tag{2.6}$$

satisfaz {U1} e {U2} e serve o mesmo propósito.

Teorema 2.4 — Se $U(c)$ é função utilidade para lotarias intermédias c, $c_1 \succ c \succ c_2$, também $U^*(c)$ definida por (2.6) é função utilidade para as mesmas lotarias.

Dem. Com $c \approx [c_1, c_2]_r$, vem,

$$U^*(c) = \alpha U(c) + \beta = \alpha U\{[c_1, c_2]_r\} + \beta,$$

donde,

$$\begin{aligned} U^*(c) &= \alpha[U(c_1)r + U(c_2)(1-r)] + \beta \\ &= [\alpha U(c_1) + \beta]r + [\alpha U(c_2) + \beta](1-r) \\ &= U^*(c_1)r + U^*(c_2)(1-r), \end{aligned}$$

e U^* satisfaz {U1}.

Considerando duas lotarias intermédias, c e c', $c \succcurlyeq c'$, tem-se,

$$U(c) \geqq U(c') \Rightarrow \frac{U^*(c) - \beta}{\alpha} \geqq \frac{U^*(c') - \beta}{\alpha},$$

e, finalmente, $U^*(c) \geqq U^*(c')$: U^* satisfaz {U2}. Note-se que a utilidade das lotarias de referência passa a ser, $U^*(c_1) = \alpha + \beta$ e $U^*(c_2) = \beta$. □□

A construção da função utilidade, U, a partir do Teorema 2.3 deixa em aberto dois problemas: 1) atribuição de utilidades a lotarias não intermédias; 2) esclarecimento da dependência em que U fica das lotarias de referência. São esses problemas que resolve o,

Teorema 2.5 — Existe uma função utilidade, com domínio em $\mathbf{C}^*$, única a menos de uma transformação linear, com $U(c)$ finito para todo o $c \in \mathbf{C}^*$.

Dem. Para alargar o domínio da função utilidade a todo o conjunto $\mathbf{C}^*$, seja U' a função procurada e tome-se,

$$U'(c_1) = 1, U'(c_2) = 0, U'(c) = U(c) \quad \text{para} \quad c_1 \succcurlyeq c \succcurlyeq c_2,$$

onde U é a função construída através da aplicação do Teorema 2.3.

Considere-se uma lotaria c^* tal que,

$$[a]\ c_1 \succ c_2 > c^* \quad \text{ou} \quad [b]\ c^* \succ c_1 \succ c_2.$$

No caso $[a]$, existe $s \in (0,1)$ que faz $c_2 \approx [c_1, c^*]_s$; tomando,

$$U'(c_2) = U'(c_1)s + U'(c^*)(1-s),$$

vem,

$$U'(c^*) = -s/(1-s).$$

No caso $[b]$, existe $t \in (0,1)$ que faz $c_1 \approx [c^*, c_2]_t$; tomando,

$$U'(c_1) = U'(c^*)t + U'(c_2)(1-t),$$

vem,

$$U'(c^*) = 1/t.$$

A função U' fica assim definida para todo e qualquer $c \in \mathbf{C}^*$. Resta mostrar que é função utilidade.

Sejam c e c^* duas lotarias arbitrárias e fixem-se as lotarias de referência, c_1 e c_2, para as quais $U'(c_1) = 1$ e $U'(c_2) = 0$. Tomem-se duas outras lotarias, c_1^0 e c_2^0, em relação às quais c, c^*, c_1 e c_2 sejam intermédias. O Teorema 2.3 pode aplicar-se a partir das lotarias, c_1^0 e c_2^0, para instituir uma função utilidade U^0 para as lotarias c tais que $c_1^0 \succcurlyeq c \succcurlyeq c_2^0$.

Suponha-se que c^* se encontra no caso $[a]$. Tem-se,

$$c_2 \approx [c_1, c^*]_s \text{ com } s = \frac{U'(c^*)}{U'(c^*) - 1},$$

donde, por U^0 satisfazer {U1} e {U2},

$$U^0(c_2) = \{U'(c^*)/[U'(c^*) - 1]\}U^0(c_1) + \{-1/[U'(c^*) - 1]\}U^0(c^*),$$

e ainda,

$$U^0(c^*) = [U^0(c_1) - U^0(c_2)]U'(c^*) + U^0(c_2). \tag{2.7}$$

Se c^* está no caso $[b]$,

$$c_1 \approx [c^*, c_2]_t \text{ com } t = 1/U'(c^*),$$

donde, tal como acima,

$$U^0(c_1) = [1/U'(c^*)]U^0(c^*) + \{1 - [1/U'(c^*)]\}U^0(c_2),$$

e ainda,

$$U^0(c^*) = [U^0(c_1) - U^0(c_2)]U'(c^*) + U^0(c_2). \tag{2.8}$$

Assim, como se deduz de (2.7) e (2.8), quer seja $c^* \succ c \succ c_2$, quer seja $c_1 \succ c_2 \succ c^*$,

$$U^0(c^*) = \alpha_0 U'(c^*) + \beta_0,\ \alpha_0 > 0, \tag{2.9}$$

ou, equivalentemente,

$$U'(c^*) = \alpha_1 U^0(c^*) + \beta_1,\ \alpha_1 > 0. \tag{2.10}$$

Além disso, se $c_1 \succcurlyeq c^* \succcurlyeq c_2$, $c^* \approx [c_1, c_2]_r$, $U'(c^*) = r$,

$$\begin{aligned} U^0(c^*) &= U^0(c_1)r + U^0(c_2)(1-r) \\ &= [U^0(c_1) - U^0(c_2)]r + U^0(c_2) \\ &= [U^0(c_1) - U^0(c_2)]U'(c^*) + U^0(c_2), \end{aligned}$$

tal como em (2.7) e (2.8) e, portanto, (2.9).

A relação (2.10) permite agora mostrar que U' é função utilidade. De facto, se $c \succcurlyeq c^*$, tem-se $U^0(c) \geqq U^0(c^*)$ e de (2.10) sai $U'(c) \geqq U'(c^*)$. Por outro lado, ainda de (2.10),

$$\begin{aligned} U'\{[c,c^*]_p\} &= \alpha_1 U^0\{[c,c^*]_p\} + \beta_1 \\ &= \alpha_1[U^0(c)p + U^0(c^*)(1-p)] + \beta_1 \\ &= U'(c)p + U'(c^*)(1-p). \end{aligned}$$

Logo, U' satisfaz {U1} e {U2}.

A função utilidade ampliada, U', depende das lotarias de referência, c_1 e c_2, para as quais $U'(c_1) = 1$ e $U'(c_2) = 0$. Partindo do novo par de referência, c_1^0 e c_2^0, obteve-se outra função utilidade, U^0, relacionada com a primeira pela expressão linear (2.10). O novo par tem contudo a particularidade de ser $c_1^0 \succ c_1 \succ c_2 \succ c_2^0$; no entanto, por meio de raciocínios do mesmo tipo, pode mostrar-se que seja qual for o par de referência escolhido em alternativa a c_1 e c_2 — par inicial — se obtém sempre uma função utilidade que é transformação linear da primeira. Por isso se diz que a função utilidade é única a menos de uma transformação linear. Por outras palavras, a arbitrariedade na escolha do par de referência, c_1 e c_2, acompanhada da convenção, $U(c_1) = 1$ e $U(c_2) = 0$, corresponde afinal à arbitrariedade na fixação da origem e escala na medição da utilidade. □□

O axioma seguinte é mais forte do que {A3} e permite ir um pouco mais longe no desenvolvimento da teoria:

{A3*} — Se $c_j \succcurlyeq c'_j$ para $j = 1, 2, ...$, então,

$$[c_1, c_2, ...]_{(p_1,p_2,...)} \succcurlyeq [c'_1, c'_2, ...]_{(p_1,p_2...)},$$

para qualquer sucessão infinita, $p_1, p_2, ...$, $(p_j > 0, \Sigma p_j = 1)$. Se adicionalmente, $c_j \succ c'_j$, para algum j para o qual $p_j > 0$, então,

$$[c_1, c_2, ...,]_{(p_1,p_2,...)} \succ [c'_1, c'_2, ...]_{(p_1,p_2...)}.$$

Com o sistema de axiomas, {A1}, {A2}, {A3*}, {A4}, podem introduzir-se dois teoremas adicionais:

Teorema 2.6 — A função utilidade é limitada: $|U(c)| \leqq M$ para todo o $c \in \mathbf{C}^*$ e para alguma constante positiva, M.

Dem. Parta-se da hipótese de que a função utilidade não é limitada. Considere-se a sucessão de lotarias, $c_1, c_2, c_3, ...$, com,

$$U(c_1) = 1,\ U(c_j) \geqq 2^{j-1}, j = 2, 3, ...,$$

o que é sempre possível em consequência da hipótese. Como é óbvio, $c_j \succ c_1$ para $j = 2, 3, ...$, e pelo axioma {A3*},

$$\begin{aligned} c'_j &\approx [c_{j+1}, c_{j+2}, ...]_{(p_{j+1}/q_j, p_{j+2}/q_j, ...)} \\ &\succ [c_1, c_1, ...]_{(p_{j+1}/q_j, p_{j+2}/q_j, ...)} \approx c_1 \end{aligned}$$

com $q_j = 1 - p_1 - \cdots - p_j$ [$(p_1, p_2, ...)$ vector de probabilidades].

Assim, $U(c'_j) > U(c_1) = 1$. Ora,

$$[c_1, c_2, ...]_{(p_1,p_2,...)} \approx [c_1, c_2, ..., c_j, c'_j]_{(p_1,p_2,...,p_j,q_j)},$$

onde $q_j = 1 - p_1 - p_2 - \cdots - p_j \to 0$ quando $j \to \infty$. Designe c a lotaria do primeiro membro; vem,

$$U(c) = U(c_1)p_1 + U(c_2)/p_2 + \cdots + U(c_j)p_j + U(c'_j)q_j.$$

Fazendo $p_j = 2^{-j}$, sai,

$$\begin{aligned} U(c) &\geqq (1/2) \cdot 1 + (1/2^2) \cdot 2 + \cdots + (1/2^j) \cdot 2^{j-1} + (1/2^j)U(c'_j) \\ &\geqq (j/2) + (1/2^j), \end{aligned}$$

o que mostra que $U(c)$ não é número finito pois com j arbitrário excede sempre $j/2$. Entra-se pois em contradição com um dos resultados do Teorema 2.5 ao supor a função utilidade não limitada, ficando assim provado o teorema em questão. □□

Teorema 2.7 — A utilidade de uma lotaria, $[c_1, c_2, \ldots]_{(p_1, p_2, \ldots)}$, é o valor esperado da utilidade dos prémios.

Dem. Tem-se, do teorema anterior,

$$U(c) = U(c_1)p_1 + U(c_2)p_2 + \cdots + U(c_j)p_j + U(c'_j)q_j;$$

mas,

$$U(c) - U(c_1)p_1 - U(c_2)p_2 - \cdots - U(c_j)p_j = U(c'_j)q_j \to 0,$$

quando $j \to \infty$, porquanto, neste caso, $q_j \to 0$ e $U(c)$ é função limitada.

Portanto,

$$U(c) = \sum_{j=1}^{\infty} U(c_j)p_j, \tag{2.11}$$

expressão que generaliza (2.5). □□

2.4 Utilidade de valores monetários

Em muitos problemas de decisão, nomeadamente quando envolvem aspectos económicos e financeiros, as consequências têm expressão monetária, isto é, representam-se por custos ou benefícios em «escudos». O conjunto das consequências, **C**, finito, infinito numerável ou contínuo, tem elementos numéricos que são quantias em moeda, positivas quando se trata de benefícios ou lucros, negativas quando se trata de custos ou percas[3]. O elemento genérico de **C** passa a designar-se por m para fazer a distinção relativamente ao caso geral abordado nas secções anteriores.

Quando o argumento da função utilidade é uma quantia em moeda variável, m, $m \in \Delta$, com Δ intervalo que se supõe corresponder a **C** ou conter **C**, uma propriedade que parece razoável introduzir desde logo é a monotonicidade:

$$m_2 > m_1 \Leftrightarrow U(m_2) > U(m_1) \quad [\text{ou } U(m_2) \geqq U(m_1)]. \tag{2.12}$$

Tratando-se de dinheiro, uma quantia tem tanto mais utilidade quanto mais elevada for[4].

[3] Há aqui uma troca de sinal em relação ao que se passa com a função perca.

[4] Não deve perder-se de vista que a quantia, m, representa quando positiva (negativa) um acréscimo (decréscimo) em relação ao «status quo», isto é, em relação a um dado activo ou riqueza do decisor, seja M. Em alguns problemas, para comparar o «status quo» com o ganho de m ou a perca de m unidades monetárias, confronta-se, $U(M)$ com $U(M+m)$ e $U(M-m)$. Em tudo o que segue, como a função utilidade é definida a menos de uma transformação linear, toma-se $U(M)$ para zero da escala das utilidades. Tal convenção não possui inconvenientes desde que não sejam envolvidas situações em que houve alteração da riqueza do decisor.

Se $m^* \in \mathbf{C}^*$ é uma lotaria com prémios monetários,

$$m^* = [m_1, m_2, \dots, m_k]_{(p_1, p_2, \dots, p_k)},$$

pode associar-se-lhe uma variável aleatória, seja $\tilde{m}$, assumindo valores, m_1, $m_2, \dots, m_k$, com probabilidades, $p_1, p_2, \dots, p_k$. Assim,

$$U(m^*) = \Sigma U(m_j) p_j = E\{U(\tilde{m})\}, \tag{2.13}$$

quer dizer, a utilidade da lotaria m^* é o valor esperado da variável aleatória $U(\tilde{m})$. Por outro lado,

$$E\{\tilde{m}\} = \Sigma m_j p_j, \tag{2.14}$$

estabelece o importante conceito de valor monetário esperado.

Alguns indivíduos estabelecem equivalência entre a utilidade de uma quantia e essa mesma quantia: tomam $U(m) = m$. Consequentemente,

$$U(m^*) = \Sigma U(m_j) p_j = \Sigma m_j p_j = E\{\tilde{m}\}; \tag{2.15}$$

quer dizer, para eles, a utilidade de m^* é igual ao valor monetário esperado.

Se um indivíduo com função utilidade $U(m) = m$ for colocado perante um conjunto de lotarias, $m_1^*, m_2^*, m_3^*, \dots$, entre as quais tem de optar, a menos que não actue «racionalmente», deve escolher aquela que tem valor monetário esperado máximo.

Não é difícil mostrar através de exemplos que muitos indivíduos não possuem função utilidade do tipo indicado.

Exemplo 2.1 — Considerem-se três lotarias em relação às quais é obrigatório exercer uma opção,

$$\begin{aligned} m_1^* &= [100\,002\$, -100\,000\$]_{1/2}, \\ m_2^* &= [\ 10\,000\$, \ -10\,000\$]_{1/2}, \\ m_3^* &= [\ \ 1\,000\$, \ \ -1\,002\$]_{1/2}. \end{aligned}$$

Qualquer das lotarias tem apenas dois prémios e em todas elas os prémios têm a mesma probabilidade, 1/2. Na hipótese do decisor possuir função utilidade, $U(m) = m$, tem-se,

$$\begin{aligned} U(m_1^*) &= U(100\,002\$)(1/2) + U(-100\,000\$)(1/2), \\ &= 100\,002\$(1/2) + (-100\,000\$)(1/2), \\ &= 1\$; \end{aligned}$$

$$\begin{aligned} U(m_2^*) &= U(10\,000\$)(1/2) + U(-10\,000\$)(1/2), \\ &= 10\,000\$(1/2) + (-10\,000\$)(1/2), \\ &= 0\$; \\ U(m_3^*) &= U(1\,000\$)(1/2) + U(-1\,002\$)(1/2), \\ &= 1\,000\$(1/2) + (-1\,002\$)(1/2), \\ &= -1\$, \end{aligned}$$

donde a seguinte ordem de preferências, $m_1^* \succ m_2^* \succ m_3^*$.

Certamente que muitos indivíduos discordam desta ordenação feita com base no valor monetário esperado, podendo até afirmar-se sem receio de errar que muitos preferem a ordenação inversa, $m_3^* \succ m_2^* \succ m_1^*$. Quer dizer, a menos que estejam actuando «irracionalmente», os discordantes não possuem função utilidade, $U(m) = m$. □

Exemplo 2.2 — Se todas as pessoas tivessem função utilidade, $U(m) = m$, não havia companhias de seguros!

Para investigar esta proposição considere-se uma empresa com instalações industriais avaliadas em $10^8\$$ e que tem de optar entre fazer ou não fazer o seguro contra o risco de incêndio e explosão. O prémio de seguro é de 0,5%, sejam $5 \cdot 10^5\$$ por ano, e sabe-se, pela experiência adquirida com fábricas do mesmo tipo, ser aproximadamente 0,001 a probabilidade de haver um sinistro no período de um ano [por hipótese, um sinistro acarreta a perda total da fábrica].

A opção é entre um pagamento certo de $5 \cdot 10^5\$$ — lotaria degenerada — e uma lotaria,

$$m^* = [0\$, -10^8\$]_{0,999},$$

em que se recebe um prémio nulo com probabilidade 0,999 (não havendo sinistro) ou se tem um prejuízo de $10^8\$$ com probabilidade 0,001 (havendo sinistro).

Admita-se que o empresário (ou órgão de gestão) tem função utilidade $U(m) = m$. Como,

$$U(-5 \cdot 10^5\$) = -5 \cdot 10^5\$ < U(m^*) = U(-10^8\$)(0{,}001) = -10^5\$,$$

a decisão «racional» é não fazer o seguro.

Os prémios cobrados pelas seguradoras são, em regra, superiores ao produto do montante seguro pela probabilidade de sinistro. Se todas as pessoas decidissem na base da maximização do valor monetário esperado não haveria, provavelmente, actividade seguradora. Isso não sucede porque a «aversão ao risco» que muitas pessoas têm se exprime por uma função utilidade diferente da acima considerada.

Na prática o empresário prefere naturalmente ter uma perca certa de $5 \cdot 10^5\$$ [que procura repercurtir sobre os preços] a correr o risco de perder $10^8\$$ e ficar

arruinado, mesmo que a probabilidade de tal ocorrer seja relativamente pequena. Quer dizer, o empresário que faz o seguro tem função utilidade, U_0, tal que,

$$U_0(-5 \cdot 10^5\$) > U_0(0\$)(0{,}999) + U_0(-10^8\$)(0{,}001),$$

ou, supondo $U_0(0\$) = 0$,

$$U_0(-5 \cdot 10^5\$) > U(-10^8\$)(0{,}001).$$

Por seu lado, a companhia seguradora que prefere participar na lotaria,

$$[-10^8\$ + 5 \cdot 10^5\$, 5 \cdot 10^5\$]_{0{,}001},$$

a não aceitar o seguro, tem função utilidade, U_1, tal que,

$$U_1(-10^8\$ + 5 \cdot 10^5\$)(0{,}001) + U_1(5 \cdot 10^5\$)(0{,}999) > U_1(0\$).$$

As seguradoras subscrevem muitas apólices em cada ramo e possuem avultados recursos (não falando nas possibilidades de resseguro). Não é, portanto, descabido supor $U_1(m) = m$, caso em que a desigualdade anterior se verifica.

Em geral, com U_0 função utilidade do segurado, U_1 função utilidade do segurador, M verba sujeita a risco de perca total, Π prémio de seguro e p probabilidade de sinistro, a operação é contratada quando,

$$\begin{aligned} U_0(-\Pi) &> U_0(0)(1-p) + U_0(-M)p, \\ U_1(0) &< U_1(\Pi)(1-p) + U_1(-M+\Pi)p. \end{aligned}$$

□

Exemplo 2.3 — O caso clássico de crítica à função utilidade, $U(m) = m$, é conhecido por Paradoxo de S. Petersburgo. Uma moeda, perfeita por hipótese, é lançada até sair face; se a primeira face é obtida no n-ésimo lançamento, $n = 1, 2, 3, \ldots$, a pessoa que participa no jogo recebe 2^n escudos. Quanto estará disposta a pagar para participar?

A probabilidade de a face sair pela primeira vez no n-ésimo lançamento é $1/2^n$. Consequentemente, o valor monetário esperado,

$$\Sigma 2^n(1/2^n) = 1 + 1 + 1 + 1 + \cdots = \infty;$$

se a pessoa tem função utilidade, $U(m) = m$, deve estar disposta a pagar uma entrada arbitrariamente grande! Na realidade, as pessoas que têm sido questionadas oferecem-se para pagar entradas muito limitadas que variam — ou deviam

variar — consoante as respectivas funções utilidade. Por exemplo, se uma pessoa tem função utilidade,

$$U(m) = \begin{cases} m & \text{para} \quad m \leqq 2^{20} \\ 2^{20} & \text{para} \quad m > 2^{20}, \end{cases}$$

tem-se,

$$U(m^*) = \sum_{n=1}^{\infty} U(2^n)(1/2^n) = \sum_{n=1}^{20} 2^n(1/2^n) + 2^{20} \sum_{n=21}^{\infty} (1/2^n) = 21,$$

onde, $m^* = [2, 2^2, 2^3, ...]_{(1/2,\, 1/2^2,\, 1/2^3, ...)}$, é a lotaria correspondente ao jogo. A entrada que essa pessoa aceita pagar é de 21 escudos! □

O Teorema 2.5 ensina que as funções utilidade,

$$U(m) = m \quad \text{e} \quad U(m) = \alpha m + \beta,\ \alpha > 0,$$

traduzem o mesmo padrão de preferências e levam, portanto, a idêntico comportamento quando se consideram lotarias arbitrárias. Os exemplos anteriores — e também um pouco de instrospecção — mostram que as funções utilidade lineares são caso muito particular. Importa, pois, efectuar a sua comparação com outras formas de função utilidade, tarefa em que desempenha papel de relevo o conceito de equivalente certo.

Designe, m^*, uma lotaria qualquer, $m^* \in \mathbf{C}^*$; o respectivo equivalente certo é a quantia certa, m_c (lotaria degenerada em que se recebe essa quantia com probabilidade igual a um), que tem a mesma utilidade que m^*,

$$U(m_c) = U(m^*), \tag{2.16}$$

ou, em termos da variável aleatória, $\tilde{m}$, equivalente à lotaria, m^*,

$$U(m_c) = E\{U(\tilde{m})\} \quad \Rightarrow \quad m_c = U^{-1}[E\{U(\tilde{m}\}], \tag{2.17}$$

quando U é função monótona.

De acordo com esta definição é indiferente para o decisor receber a quantia certa m_c ou participar na lotaria m^*.

Recapitulando alguns resultados anteriores tem-se:

Teorema 2.8 — Se um indivíduo tem função utilidade linear, para qualquer lotaria a utilidade do valor monetário esperado é igual à utilidade da lotaria,

$$U(E\{\tilde{m}\}) = E\{U(\tilde{m})\}, \tag{2.18}$$

e o valor monetário esperado é igual ao equivalente certo,

$$E\{\tilde{m}\} = m_c. \tag{2.19}$$

Dem. Deixa-se como exercício. □□

Além das funções utilidade lineares merecem especial referência as funções utilidade convexas e as funções utilidade côncavas. Como vai ver-se, estas funções traduzem atitudes opostas em relação ao risco: gosto pelo risco, as primeiras, aversão ao risco, as segundas. As funções utilidade lineares exprimem uma atitude intermédia ou de neutralidade.

Função[5], $U(m)$, real de variável real, definida no intervalo, Δ, arbitrário — aberto, semi-aberto ou fechado, finito ou infinito — diz-se convexa nesse intervalo se para quaisquer, $m_1, m_2 \in \Delta$ e qualquer $\lambda \in (0,1)$, se tem,

$$U[\lambda m_1 + (1-\lambda)m_2] \leqq \lambda U(m_1) + (1-\lambda)U(m_2). \tag{2.20}$$

Verificando-se em (2.20) desigualdade estrita a função diz-se estritamente convexa. Na Fig. 2.1 representam-se as imagens da função nos dois casos.

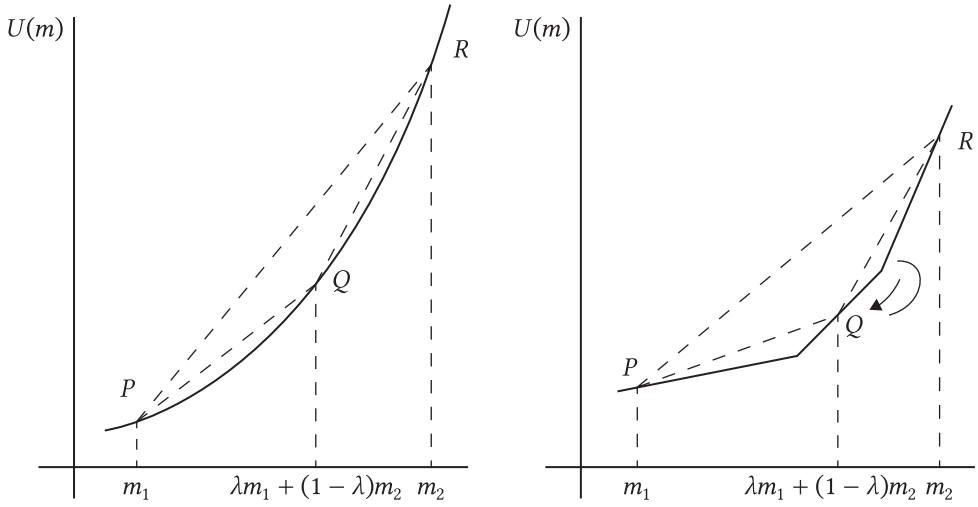


Fig. 2.1

Geometricamente (2.20) significa que sendo P, Q e R três pontos da imagem da função $U(m)$, localizados tal como se indica na Fig. 2.1, Q está abaixo da corda PR se a função é estritamente convexa ou está abaixo ou sobre a corda se a função é convexa. Evidentemente, no segundo caso, Q está sobre a corda se a função for linear entre P e R.

As funções convexas gozam de importantes propriedades.

[5] Apesar de se usar a notação, $U(m)$, de função utilidade, o conceito de convexidade é relevante para qualquer função nas condições indicadas.

Diz-se que U (convexa ou não) tem suporte no ponto $m_0 \in \Delta$ se existe uma função afim,

$$\Psi(m) = U(m_0) + v(m - m_0),$$

tal que $\Psi(m) \leqq U(m)$ para todo o $m \in \Delta$. A imagem da função suporte denomina-se recta ou linha de suporte de U em m_0 (veja-se Fig. 2.2).

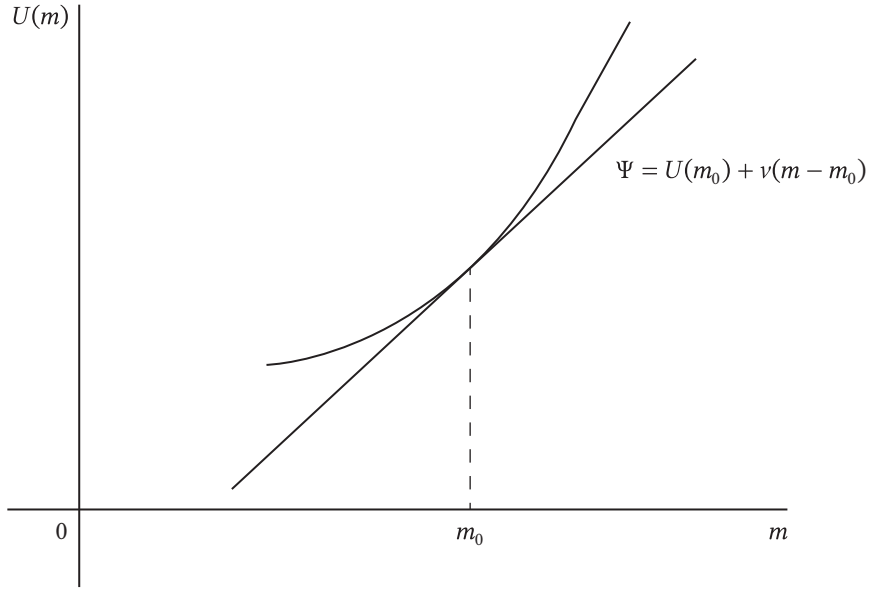


Fig. 2.2

Teorema 2.9 — A função U é convexa se e somente se existe pelo menos uma recta de suporte em todo o $m \in \Delta_0$ [interior de Δ].

Dem. Veja-se Roberts e Varberg (1973). □□

Teorema 2.10 — Se U é convexa no intervalo Δ, então, com $i = 1, 2, \ldots, k$, $m_i \in \Delta$, $\lambda_i > 0$, $\Sigma\lambda_i = 1$, verifica-se,

$$U(\Sigma\lambda_i m_i) \leqq \Sigma\lambda_i U(m_i) \qquad \text{[Desigualdade de Jensen]}. \tag{2.21}$$

Dem. Pelo teorema anterior tem-se, para todo o $m_0 \in \Delta_0$, um número real v tal que,

$$U(m) \geqq U(m_0) + v(m - m_0), \tag{2.22}$$

qualquer que seja $m \in \Delta$ (veja-se Fig. 2.2). Em particular,

$$U(m_i) \geqq U(m_0) + v(m_i - m_0), \quad i = 1, 2, \ldots, k;$$

tomando $m_0 = \Sigma\lambda_i m_i$, multiplicando ambos os membros por λ_i, somando e simplificando, obtém-se (2.21). □□

Repare-se que a desigualdade de Jensen podia demonstrar-se por indução finita a partir de (2.20); a dedução acima tem a vantagem de abrir caminho para o importante,

Teorema 2.11 — Se U é função convexa no intervalo Δ e se $\tilde{m}$ é variável aleatória tal que $P(\tilde{m} \in \Delta) = 1$, então, existindo os valores esperados $E\{\tilde{m}\}$ e $E\{U(\tilde{m})\}$, verifica-se,

$$U(E\{\tilde{m}\}) \leqq E\{U(\tilde{m})\}. \tag{2.23}$$

Em particular, se U é estritamente convexa e $\tilde{m}$ variável aleatória não degenerada,

$$U(E\{\tilde{m}\}) < E\{U(\tilde{m})\}. \tag{2.24}$$

Dem. Faça-se em (2.22) $m_0 = E\{\tilde{m}\}$ e substitua-se m por $\tilde{m}$,

$$U(\tilde{m}) \geqq U(E\{\tilde{m}\}) + v(\tilde{m} - E\{\tilde{m}\});$$

tomando o valor esperado de ambos os membros e notando $E\{[\tilde{m} - E\{\tilde{m}\}]\} = 0$, obtém-se (2.23). Se U é estritamente convexa a igualdade em (2.22) só pode verificar-se para $m = m_0$; como se supõe que $\tilde{m}$ é não degenerada, $\tilde{m}$ tem de assumir pelo menos dois valores com probabilidade positiva. Logo tem-se (2.24). □□

Pode avançar-se agora na caracterização do gosto pelo risco.

Um indivíduo tem gosto pelo risco quando prefere qualquer lotaria não degenerada ao respectivo valor monetário esperado. Isto é, sendo m^* uma lotaria não degenerada e $\tilde{m}$ a variável aleatória com ela associada, o gosto pelo risco traduz-se por,

$$U(E\{\tilde{m}\}) < E\{U(\tilde{m})\}. \tag{2.25}$$

Uma forma alternativa de enunciado consiste em dizer que o valor monetário esperado é sempre inferior ao equivalente certo,

$$E\{\tilde{m}\} < m_c. \tag{2.26}$$

Tem-se, como é evidente,

Teorema 2.12 — Um indivíduo tem gosto pelo risco se e somente se a sua função utilidade é estritamente convexa. □□

A principal conclusão a tirar da caracterização precedente é a seguinte: decisor com gosto pelo risco — função utilidade estritamente convexa — está sempre disposto a pagar pelo menos o valor monetário esperado para adquirir o direito a participar na lotaria considerada; no máximo está disposto a pagar o valor monetário esperado, $E\{\tilde{m}\}$, acrescido da diferença, $m_c - E\{\tilde{m}\}$.

Exemplo 2.4 — Considere-se a lotaria, $m^* = [100\$, -100\$]_{1/2}$, e suponha-se que o decisor tem função utilidade estritamente convexa num intervalo, $\Delta = (m_1, m_2)$, $m_1 < -100\$$ e $m_2 > 100\$$. A variável aleatória, $\tilde{m}$, associada à lotaria tem valor esperado, $E\{\tilde{m}\} = 0\$$. Por (2.24),

$$\begin{aligned} U(m^*) = E\{U(m)\} = U(-100\$)(1/2) + \\ + U(100\$)(1/2) > U(E\{\tilde{m}\}) = U(0\$); \end{aligned} \tag{2.27}$$

o decisor, colocado perante a situação em que tem de optar pela participação ou não participação na lotaria, prefere participar porquanto a lotaria tem para ele maior utilidade do que a acção neutra (ganho = 0\$). O decisor está mesmo interessado em pagar alguma coisa para participar apesar do valor monetário ser zero. É aliás a posição dos indivíduos cujo gosto pelo risco os leva a tomar parte em jogos com entrada superior ao ganho esperado — jogos não equitativos — como são os praticados em todos os casinos.

Quanto está o decisor disposto a pagar para se habilitar na lotaria m^*?

A resposta é dada através do cálculo do equivalente certo m_c, $U(m_c) = U(m^*)$.

Considerando a Fig. 2.3 e a expressão (2.27) conclui-se imediatamente que $U(m^*) = OA$. Fazendo, $U(m_c) = OA$, determina-se m_c, quantia máxima que o decisor está pronto a entregar para adquirir o direito de jogar na lotaria. □

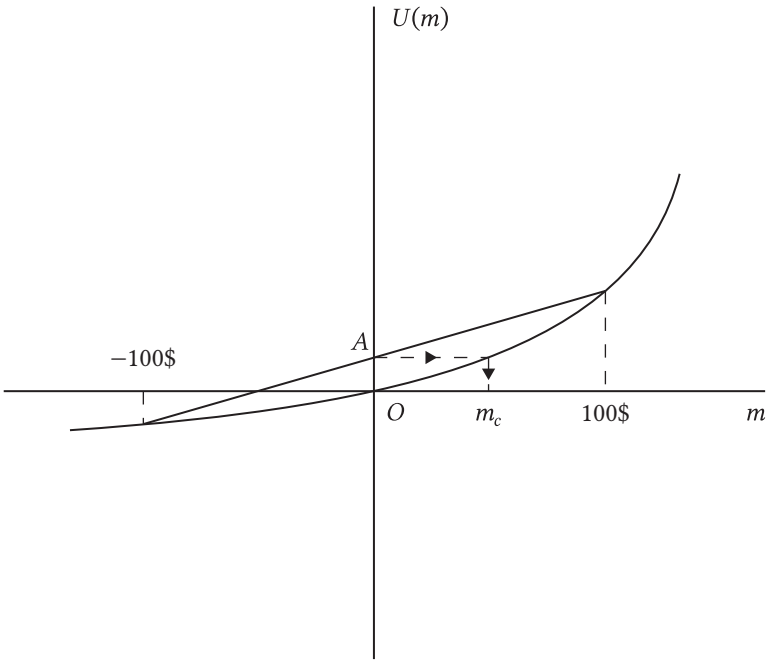


Fig. 2.3

As funções utilidade estritamente côncavas exprimem aversão ao risco, atitude oposta à caracterizada pelas funções utilidade estritamente convexas, e estão de acordo com a lei da utilidade marginal decrescente bem conhecida dos economistas.

Função, $U(m)$, real de variável real, definida no intervalo, Δ, arbitrário, diz-se côncava (estritamente côncava) nesse intervalo se $-U(m)$ for convexa (estritamente convexa), isto é, se para quaisquer valores, $m_1, m_2 \in \Delta$ e $\lambda \in (0,1)$, se tem,

$$U[\lambda m_1 + (1-\lambda)m_2] \geqq \lambda U(m_1) + (1-\lambda)U(m_2), \tag{2.28}$$

(com $>$ no caso estrito).

Um indivíduo tem aversão ao risco quando prefere o valor monetário esperado de qualquer lotaria não degenerada à própria lotaria, isto é, quando,

$$U(E\{\tilde{m}\}) > E\{U(\tilde{m})\}, \tag{2.29}$$

ou, equivalentemente,

$$E\{\tilde{m}\} > m_c. \tag{2.30}$$

Evidentemente,

Teorema 2.13 — Um indivíduo tem aversão ao risco se e somente se a sua função utilidade é estritamente côncava. □□

Exemplo 2.5 — Seja ainda a lotaria,

$$m^* = [100\$, -100\$]_{1/2}$$

e o caso em que o decisor tem função utilidade estritamente côncava. Tem-se

$$E\{\tilde{m}\} = 0\$$$

e, por (2.29),

$$\begin{aligned} U(m^*) = E\{U(\tilde{m})\} = U(-100\$)(1/2) + \\ + U(100\$)(1/2) < U(E\{\tilde{m}\}) = U(0\$). \end{aligned} \tag{2.31}$$

O decisor, se tiver de exercer uma opção, prefere não participar na lotaria, pois, para ele, um ganho de 0$ tem mais utilidade do que a lotaria. Pode mesmo dizer-se que, se o tentarem forçar a entrar no jogo, está na disposição de pagar alguma coisa para ser dispensado de o fazer. É a posição das pessoas que fazem seguros.

Na Fig. 2.4 mostra-se como se calcula a importância máxima que o decisor está inclinado a pagar para não entrar no jogo. Trata-se do equivalente certo, m_c,

$$U(m_c) = U(m^*) = OA,$$

com $U(m^*)$ dado pelo primeiro membro de (2.30). O sinal negativo de m_c indica que o decisor é indiferente entre o prejuízo de $|m_c|$ e a participação na lotaria. □

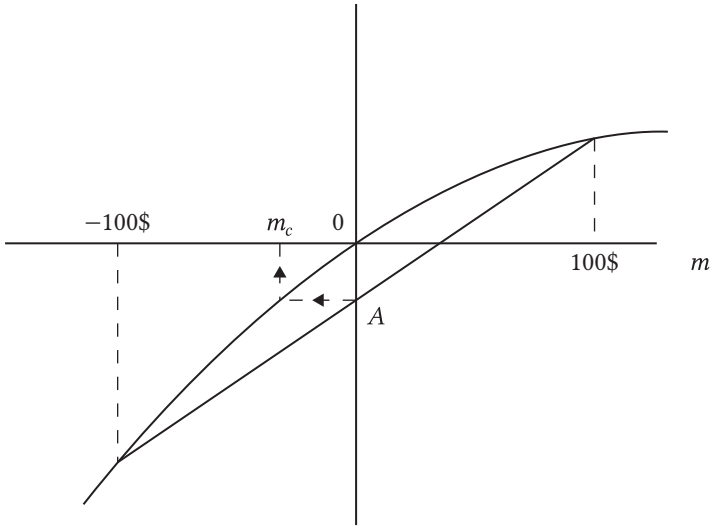


Fig. 2.4

Nada obriga um indivíduo a possuir função utilidade estritamente convexa, linear ou estritamente côncava em todo o intervalo de valores monetários considerado. A função pode ser estritamente côncava num dado troço, linear noutro e ainda estritamente convexa noutro. O indivíduo pode ser prudente quanto estão em jogo quantitativos de uma dada ordem de grandeza e audacioso quando os quantitativos são de ordem diferente. Por outro lado a função utilidade pode apresentar pontos de descontinuidade: na Fig. 2.5 indica-se o que pode acontecer com a função de utilidade de um indivíduo que necessita desesperadamente de uma quantia Q.

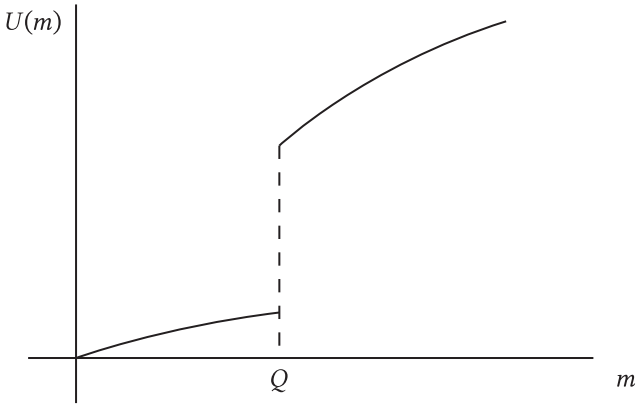


Fig. 2.5

Para finalizar o resumido estudo da função utilidade de valores monetários refere-se, de passagem, o processo que pode usar-se para determinar empiricamente a função utilidade, $U(m)$, de um dado indivíduo. Embora o problema seja mais psicológico do que matemático o Teorema 2.3 contém alguns ensinamentos pertinentes.

Suponha-se, para facilitar a exposição, $0 \leqq m \leqq M$. O processo compreende os seguintes passos[6]:

1.º) Convenciona-se, sem perda de generalidade,

$$U(0) = 0 \text{ e } U(M) = 1.$$

2.º) Pede-se ao indivíduo para indicar a quantia certa, m_1, tal que $m_1 \approx [0, M]_{1/2}$, quer dizer, tal que para o indivíduo é indiferente receber m_1 ou participar na lotaria $[0, M]_{1/2}$. Indicado m_1, vem,

$$U(m_1) = U(0)(1/2) + U(M)(1/2) = 0{,}5.$$

3.º) Pede-se para indicar a quantia certa, m_2, tal que $m_2 \approx [0, m_1]_{1/2}$; indicado m_2, tem-se,

$$U(m_2) = U(0)(1/2) + U(m_1)(1/2) = 0{,}25.$$

4.º) Pede-se para indicar a quantia certa, m_3, $m_3 \approx [m_1, M]_{1/2}$; indicado m_3, vem,

$$U(m_3) = U(m_1)(1/2) + U(M)(1/2) = 0{,}75.$$

5.º) Pede-se para indicar a quantia certa, m_4, $m_4 \approx [m_2, m_3]_{1/2}$; indicado m_4, tem-se,

$$U(m_4) = U(m_2)(1/2) + U(m_3)(1/2) = 0{,}5.$$

Nos passos 1.º) a 4.º) ficam determinados cinco pontos da função utilidade do indivíduo,

$$(0{,}0), (m_2, 0{,}25), (m_1, 0{,}5), (m_3, 0{,}75), (M, 1).$$

Note-se que o 5.º) passo deve fornecer $m_4 = m_1$, porquanto se tem,

$$U(m_4) = 0{,}5 = U(m_1),$$

e U é função monótona. Trata-se de um teste de coerência que quando não satisfeito implica afastamento do indivíduo em relação ao comportamento «racional» que assegura a existência de função utilidade. Assim, se for obtido $m_4 \neq m_1$, repetem-se os passos 2.º) a 5.º) o número de vezes necessário para que haja convergência. Ajustando uma curva suave aos pontos finalmente estimados fica-se com uma ideia da imagem da função utilidade do indivíduo inquirido.

[6] O número de passos pode, evidentemente, alargar-se.

2.5 Utilidade e probabilidades subjectivas

Toda a análise do presente capítulo foi conduzida sobre lotarias ou jogos em que determinados prémios, $c_1, c_2, \ldots, c_i \in \mathbf{C}$, saiem com probabilidades, $p_1, p_2, \ldots$ ($p_i \geqq 0, \Sigma p_i = 1$), conhecidas ou objectivas [o termo «probabilidade objectiva» deve escrever-se entre aspas pois não é inteiramente pacífica a respectiva interpretação; uma possibilidade é recorrer ao conceito frequencista mas Fine (1973) refere outras]. A teoria da utilidade exposta nas secções anteriores serve para orientar a escolha nos problemas de decisão em face do risco mas não tem alcance ou, pelo menos, não é suficiente para tratar problemas de decisão em face da incerteza, designadamente aqueles em que as probabilidades dos estados são imperfeitamente conhecidas ou subjectivas.

Para que a teoria da utilidade tenha amplitude suficiente para cobrir problemas gerais de decisão, é necessário introduzir um novo tipo de lotaria cujos prémios saiem com probabilidades imperfeitamente conhecidas.

Para estabelecer o confronto entre os dois tipos de lotarias pode pensar-se no seguinte exemplo: nas lotarias tratadas na secção anterior os prémios são atribuídos em função dos resultados obtidos com uma roleta; nas lotarias que passam a tratar-se os prémios são atribuídos em função dos resultados obtidos numa corrida de cavalos. São as já referidas «*roulette lotteries*» e «*horse lotteries*».

Suponha-se que o número de estados da natureza é finito, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, e designe b_0 a lotaria,

$$b_0 = [c_1, c_2, \ldots, c_n].$$

em que o prémio $c_j \in \mathbf{C}$, $j = 1, 2, \ldots, n$, é atribuído quando se verifica ser θ_j o estado[7]. Admitindo a mistura de probabilidades objectivas-subjectivas tem-se a lotaria mais geral,

$$b = [c_1^*, c_2^*, \ldots, c_n^*],$$

em que os prémios são lotarias do tipo roleta. As lotarias b passam a ser referência e o respectivo conjunto designa-se por $\mathbf{B}$.

A doutrina exposta permite estabelecer que se o decisor tem um padrão de preferências,«$\succcurlyeq$» em $\mathbf{C}^*$, satisfazendo o sistema de axiomas, {A1}–{A4}, existe, para ele, função utilidade, U, em $\mathbf{C}^*$. O aspecto interessante que Anscombe e Aumann (1963) investigaram é o seguinte: introduzido um padrão de preferências, «$\succcurlyeq_\beta$», em $\mathbf{B}$, verificando um sistema de axiomas moldado no anterior, em que condições [1] fica assegurada a existência de função utilidade, U_β, em $\mathbf{B}$; mais ainda, como as probabilidades dos estados são imperfeitamente conhecidas, em que condições [2] o decisor actua como se tivesse atribuído probabilidades aos vários estados da

[7] Ou se verifica que venceu o «cavalo» n.º j.

natureza, optando depois entre as lotarias de **B** de forma a maximizar a utilidade esperada. Se estas condições forem identificadas, as probabilidades atribuídas aos estados da natureza podem considerar-se probabilidades subjectivas ou personalistas, conclusão de óbvia importância, pelo menos na óptica bayesiana.

Seja então «$\succcurlyeq_\beta$» uma relação binária com suporte em **B**, conjunto de lotarias do segundo tipo, satisfazendo um sistema de axiomas homólogo de {A1}–{A4}. Introduzam-se os pressupostos adicionais:

{B1} — Para qualquer i, $i = 1, 2, \ldots, n$, $c_i^* \succcurlyeq c_i^{**}$, com $c_i^*, c_i^{**} \in \mathbf{C}^*$, implica,

$$[c_1^*, \ldots, c_i^*, \ldots, c_n^*] \succcurlyeq_\beta [c_1^*, \ldots, c_i^{**}, \ldots, c_n^*].$$

{B2} — $c^* \succcurlyeq c^{**}$, com $c^*, c^{**} \in \mathbf{C}^*$, implica,

$$[c^*, \ldots, c^*, \ldots, c^*] \succcurlyeq_\beta [c^{**}, \ldots, c^{**}, \ldots, c^{**}].$$

{B3} — Com $b_i = [c_{i1}^*, c_{i2}^*, \ldots, c_{in}^*]$, $b_i \in \mathbf{B}$, $c_{ij}^* \in \mathbf{C}^*$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, tem-se,

$$[b_1, b_2, \ldots, b_k]_{(p_1, p_2, \ldots, p_k)} \approx_\beta [[c_{11}^*, c_{21}^*, \ldots, c_{k1}^*]_{(p_1, p_2, \ldots, p_k)}, \ldots, [c_{1n}^*, c_{2n}^*, \ldots, c_{kn}^*]_{(p_1, p_2, \ldots, p_k)}],$$

onde $(p_1, p_2, \ldots, p_k)$ é um vector arbitrário de probabilidades.

Os axiomas {B1}–{B3} exprimem a relação entre os padrões de preferência «$\succcurlyeq$» em $\mathbf{C}^*$ e «$\succcurlyeq_\beta$» em **B**. Os dois primeiros axiomas mostram como se processa a transposição do primeiro padrão para o segundo; o terceiro axioma estabelece que a casualização operada com o vector de probabilidades conhecidas, $(p_1, p_2, \ldots, p_k)$, pode permutar-se com a casualização operada pela «natureza» ao fixar o estado. Quer dizer, se numa lotaria «combinada» os prémios dependem do resultado da roleta e do resultado da corrida de cavalos, a ordem por que são conhecidos ou obtidos os dois resultados [roleta ⇒ cavalos ou cavalos ⇒ roleta] é completamente indiferente para o decisor.

Para exemplificar {B3} considerem-se dois estados, $\{\theta_1, \theta_2\}$, e duas lotarias de cavalos,

$$b_1 = [c_{11}^*, c_{12}^*], b_2 = [c_{21}^*, c_{22}^*],$$

cujos prémios são as lotarias de roleta, $c_{11}^*, c_{12}^*, c_{21}^*, c_{22}^*$. O axioma estabelece a equivalência das lotarias representadas no esquema da Fig. 2.6.

Posto isto, se «$\succcurlyeq$» e «$\succcurlyeq_\beta$» são padrões de preferência em $\mathbf{C}^*$ e **B**, respectivamente, verificando o sistema {A1}–{A4} e relacionados pelo sistema {B1}–{B3}, demonstra-se que existe uma função utilidade, U_β, com domínio em **B**, tal que, para todo o $b \in \mathbf{B}$,

$$U_\beta(b) = U(c_1^*)h_1 + U(c_2^*)h_2 + \cdots + U(c_n^*)h_n, \tag{2.32}$$

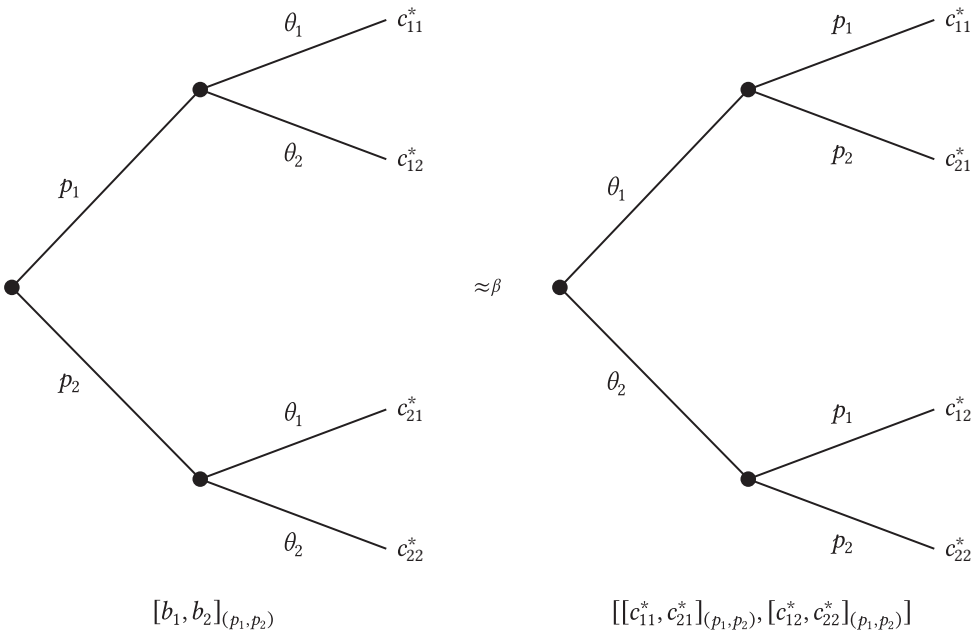


Fig. 2.6

onde U é função utilidade com domínio em $\mathbf{C}^*$ e $(h_1, h_2, \ldots, h_n)$ é um vector de probabilidades.

O decisor actua, assim, como se conhecesse as probabilidades dos estados e essas probabilidades podem considerar-se probabilidades subjectivas ou personalistas atribuídas aos diferentes estados.

Em resumo [Ferguson (1967)],

Teorema 2.14 — Se os padrões de preferência, «$\succcurlyeq$» em $\mathbf{C}^*$ e «$\succcurlyeq_\beta$» em $\mathbf{B}$, satisfazem, {A1}–{A4} e {B1}–{B3}, existem funções utilidade, U em $\mathbf{C}^*$ e U_β em $\mathbf{B}$ e existe um vector de probabilidades, $(h_1, h_2, \ldots, h_n)$, verificando (2.32). □□

Embora a exposição tenha sido conduzida considerando um espaço de estados finito, o estudo de Anscombe e Aumann (1963) [que Ferguson apura] pode desenvolver-se sobre um espaço mensurável qualquer.

O tratamento simultâneo da utilidade e da probabilidade subjectiva é extremamente importante. De facto, quando se supõe — veja-se secção 1.1 — que as probabilidades são medidas em função do comportamento do indivíduo ao apostar em certos jogos hipotéticos pode cair-se em dificuldades. A fonte de preocupação é o argumento de que esse comportamento depende do montante das apostas e não só da credibilidade atribuída aos eventos relevantes. Precisamente por isso

alguns autores pensam em termos de apostas monetárias «suficientemente pequenas»; Savage (1954) chega a considerar apostas monetárias «infinitesimais». O que importa, na verdade, é a utilidade do montante das apostas; assim, a teoria da utilidade e as probabilidades subjectivas aparecem interligadas.

A primeira formulação da teoria da utilidade em situações não determinísticas é geralmente atribuída a von Neumann e Morgenstern (1944), autores seguidos de perto nas secções 2.2 e 2.3 através da clara exposição de Lindgren (1971). Essa prioridade só faz sentido numa óptica puramente matemática, pois Ramsey (1926), um tanto informalmente, já havia provado que devia existir uma função utilidade a qual, por sua vez, arrastava um conceito de probabilidade que permitia quantificar a credibilidade.

O grande mérito de Ramsey foi deduzir primeiro, a partir de um conjunto de pressupostos, a existência de função utilidade na qual depois se baseou para avaliar as probabilidades subjectivas. Um dos pressupostos é a existência de uma proposição «eticamente neutra», seja π, com grau de credibilidade 1/2, por exemplo uma proposição como «esta moeda, quando lançada, cai com a ‹face› voltada para cima». Garantida a existência de função utilidade tem-se, por exemplo, para a lotaria, $c^* \equiv [c_1 \,|\, \pi, c_2 \,|\, \overline{\pi}]$ — a notação foi ligeiramente modificada e significa saída do prémio c_1 quando π é verdadeira e saída do prémio c_2 quando $\overline{\pi}$ é verdadeira — a utilidade,

$$U(c^*) = U(c_1)\frac{1}{2} + U(c_2)\frac{1}{2}.$$

Dados dois conjuntos de estados, $\alpha, \beta \subset \Theta$, se o decisor manifesta indiferença perante as lotarias,

$$\left[c_1 \,|\, \beta, c_2 \,|\, \overline{\beta}\right] \quad \text{e} \quad \left[c_3 \,|\, \alpha \cap \beta, c_4 \,|\, \overline{\alpha} \cap \beta, c_2 \,|\, \overline{\beta}\right],$$

a credibilidade de α, dado β, é dada pela expressão,

$$\frac{U(c_1) - U(c_4)}{U(c_3) - U(c_4)},$$

e verifica-se que segue as leis de probabilidade usuais.

A formulação de Anscombe e Aumann tem um aspecto que alguns autores consideram pouco satisfatório [por exemplo, Fine (1973)]: mistura probabilidades objectivas (as das «*roulette lotteries*») com probabilidades subjectivas (as das «*horse lotteries*») o que tem como consequência que as avaliações subjectivas são tomadas em relação a uma escala objectiva de probabilidade preexistente.

A construção axiomática de Pratt, Raiffa e Schlaifer (1964) é tida como uma das mais rigorosas se bem que limitada ao caso finito. Estes autores introduzem a ideia de uma «experiência canónica» que corresponde efectivamente a postular a existência de variáveis aleatórias, x e y, independentes e com distribuição uniforme no

intervalo $[0, 1]$. Este postulado tem algo a ver com a proposição «eticamente neutra» de Ramsey. O que qualquer dos pressupostos[8] mostra essencialmente é, como diz Lindley (1971a), a necessidade de: « *...a standard to which, other statements of uncertainty can be compared*».

O tratamento de DeGroot (1970) é diferente e tem a particularidade de introduzir um sistema de axiomas para as probabilidades subjectivas e, posteriormente, um sistema de axiomas para as utilidades. Mas, note-se, do primeiro sistema faz parte o pressuposto de que existe uma variável aleatória com distribuição uniforme no intervalo $[0, 1]$. A estruturação de von Neumann e Morgenstern conduz a uma função utilidade limitada; DeGroot, na sua análise, mostra que esse resultado é consequência das restrições impostas ao conjunto de lotarias e que um sistema mais amplo de axiomas, admitindo uma mais vasta classe de lotarias, leva a uma função utilidade não limitada, facto que na prática pode ter vantagens.

Outras construções, nomeadamente as devidas a Savage e a Krantz e Luce, podem estudar-se em Fine (1973). Fine afirma que a do primeiro, retomada e desenvolvida pelos segundos, é baseada unicamente numa noção primitiva de preferência e faz aparecer as probabilidades subjectivas independentemente de quaisquer outras concepções de probabilidade.

2.6 Função utilidade e função perca

Segundo o princípio de coerência em que se baseia a decisão estatística — veja-se secção 1.1 — a optimização sobre o conjunto das funções de decisão deve ter como critério a maximização da utilidade esperada. As condições em que existe função utilidade são, por isso, importantes, pelo menos em teoria.

Quando se consegue determinar correctamente a função utilidade, $U(\theta, a)$, a função perca sai imediatamente da relação,

$$L(\theta, a) = -U(\theta, a), \tag{2.33}$$

que tem o atractivo de permitir a correspondência entre a maximização da utilidade esperada e a minimização da perca esperada.

Autores há que sustentam que o que é importante é a perca relativamente à melhor acção, posição que os leva a empregar de preferência a função pesar,

$$\begin{aligned} L_0(\theta, a) &= \sup_a U(\theta, a) - U(\theta, a) \\ &= L(\theta, a) - \inf_a L(\theta, a). \end{aligned} \tag{2.34}$$

[8] E a presença de probabilidades objectivas na formulação Anscombe-Aumann.

Quando a^* é a melhor acção para o estado θ, tem-se $L_0(\theta, a^*) = 0$, propriedade de inegável interesse. Os economistas chamam a $L_0(\theta, a)$ função perca de oportunidade ou custo de oportunidade.

Quando se não aceita a teoria da utilidade ou se não consegue construir uma função utilidade satisfatória, fica-se impedido de aplicar a teoria da decisão estatística? A resposta é negativa; na própria elaboração de Wald em 1950, a função perca — que também designa por função ponderação — é um dado do problema de decisão e o objectivo é a minimização da perca esperada. Esta posição de partida é tomada com plena consciência dos obstáculos que cercam a determinação prática da função perca e que levam com frequência a adoptar formas expeditas ou puramente convencionais. Por outro lado, a marginalização da teoria da utilidade deixa a minimização da perca esperada sem o «perfume» da coerência.

Se o desenvolvimento da teoria da decisão estatística se processa pressupondo que a função perca é dada, nada existe nessa teoria que justifique tomar a função perca com sinal trocado para representar a utilidade!

Nos problemas de decisão estatística, θ, o estado da natureza, é muitas vezes o parâmetro de uma distribuição, facto que pode dar origem a dificuldades na interpretação da função utilidade. Berger (1980) cita a propósito o seguinte exemplo: uma empresa realiza um estudo de mercado antes de lançar um novo produto e procura estimar a proporção do universo, seja θ, que se diz interessada em adquirir o mesmo; na verdade a empresa tem resultados que dependem da proporção que vai de facto comprar o produto, proporção que é função de outros factores aleatórios tais como a conjuntura económica. Assim, as consequências da tomada da acção a quando o estado é θ, não são caracterizadas pelo par (θ, a), mas sim por uma variável ou resultado aleatório, R; nesse caso parece legítimo definir,

$$U(\theta, a) = E_{\theta,a}\{U(R)\},$$

onde a expressão do segundo membro é o valor esperado de $U(R)$ calculado para o par (θ, a), que evidentemente condiciona o comportamento de R.

O comentário que as considerações acima imediatamente sugerem é o seguinte: por que motivo não se consideram estados da natureza que compreendam todos os factores relevantes, desde as intenções de compra às condições económicas envolventes? Em teoria seria esta a atitude lógica; na prática revela-se muitas vezes conveniente separar os factores em relação aos quais se pode obter informação estatística — caso da proporção de compradores potenciais — dos factores em que se não pode ir além de vagas conjecturas — tidas em conta ao esboçar a distribuição de R.

CAPÍTULO 3

FUNÇÕES DE DECISÃO MINIMAX E BAYES

3.1 Acções puras. Acções mistas

Suponham-se finitos, por agora, o conjunto de acções, $A = \{a_1, a_2, \dots, a_m\}$, e o conjunto de estados, $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$. A cada acção $a_i \in A$ pode fazer-se corresponder um ponto ou vector de $\mathbf{R}^n$, chamado vector perca, que passa a designar-se por,

$$\boldsymbol{\ell}(a_i) = [L(\theta_1, a_i), L(\theta_2, a_i), \dots, L(\theta_n, a_i)], \quad i = 1, 2, \dots, m. \tag{3.1}$$

As componentes de $\boldsymbol{\ell}(a_i)$ representam as percas correspondentes aos vários estados.

Podem conceber-se situações em que o decisor exerce a sua opção por meio de um processo ou mecanismo aleatório. É o que sucede quando lança uma moeda ao ar para decidir a ida ao teatro ou ao cinema. A casualização conduz ao conceito de acção mista: escolha entre as acções puras, a_i, com probabilidades, ζ_i, $i = 1, 2, \dots, m$, $\zeta_i \geqq 0$, $\Sigma\zeta_i = 1$. Optar por uma acção mista equivale a indicar um vector de probabilidade, $\zeta \in S_m$, onde S_m é o simplex de $\mathbf{R}^m$,

$$S_m = \{\zeta = (\zeta_1, \zeta_2, \dots, \zeta_m) \,:\, \zeta_i \geqq 0, \Sigma\zeta_i = 1\}.$$

Cada acção mista é definida por uma distribuição de probabilidade sobre o espaço de acções puras. As acções puras são casos particulares de acções mistas: correspondem aos vértices do simplex, $(1, 0, \dots, 0)$, $(0, 1, \dots, 0)$, ..., $(0, 0, \dots, 1)$.

Designe a^* a acção mista genérica obtida com o vector de probabilidade genérico, $\zeta \in S_m$. A perca associada com a^* quando o estado é θ_j é definida pelo valor

esperado,

$$L(\theta_j, a^*) = \sum_{i=1}^{m} L(\theta_j, a_i)\zeta_i, \quad j = 1, 2, \ldots, n. \tag{3.2}$$

Com,

$$\boldsymbol{\ell}(a^*) = [L(\theta_1, a^*), L(\theta_2, a^*), \ldots, L(\theta_n, a^*)],$$

tem-se,

$$\boldsymbol{\ell}(a^*) = \sum_{i=1}^{m} \boldsymbol{\ell}(a_i)\zeta_i. \tag{3.3}$$

O espaço de acções mistas, obtido quando ζ percorre S_m, designa-se por A^*; como é óbvio, $A \subset A^*$.

Considerem-se os conjuntos de $\mathbf{R}^n$,

$$S = \{\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n) \,:\, \ell_j = L(\theta_j, a_i),\ a_i \in A\}, \tag{3.4}$$

$$S^* = \{\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n) \,:\, \ell_j = L(\theta_j, a^*),\ a^* \in A^*\}. \tag{3.5}$$

A S chama-se conjunto perca e a S^* conjunto perca casualizada. Claramente, $S \subset S^*$; por outro lado, como se conclui de (3.3), os pontos de S^* são obtidos por combinação linear convexa dos m pontos de S. Portanto, S^* é o ambiente convexo de S, isto é, o conjunto convexo mínimo que contém S ou ainda a intersecção de todos os conjuntos convexos que contém S. Sendo S finito, S^* é: [1] para $n = 2$, um polígono convexo, excepto quando os pontos $\boldsymbol{\ell}(a_i)$ são colineares caso em que se reduz a um segmento de recta; [2] para $n = 3$, um poliedro convexo, com excepções óbvias; [3] para $n > 3$, um hiperpoliedro convexo.

Exemplo 3.1 — Dada a matriz perca,

	a_1	a_2	a_3	a_4	a_5
θ_1	2	4	3	6	4
θ_2	3	1	3	2	5

na Fig. 3.1 representam-se os cinco pontos do conjunto S,

$$\begin{aligned} S &= \{\boldsymbol{\ell}(a_1), \boldsymbol{\ell}(a_2), \boldsymbol{\ell}(a_3), \boldsymbol{\ell}(a_4), \boldsymbol{\ell}(a_5)\}, \\ &= \{(2,3), (4,1), (3,3), (6,2), (4,5)\}, \end{aligned}$$

e o quadrilátero correspondente ao conjunto S^*,

$$S^* = \left\{2(\zeta_1 + 4\zeta_2 + 3\zeta_3 + 6\zeta_4 + 4\zeta_5, 3\zeta_1 + \zeta_2 + 3\zeta_3 + 2\zeta_4 + 5\zeta_5) \,:\, \zeta_i \geqq 0, \sum_{i=1}^{5} \zeta_i = 1\right\}.$$

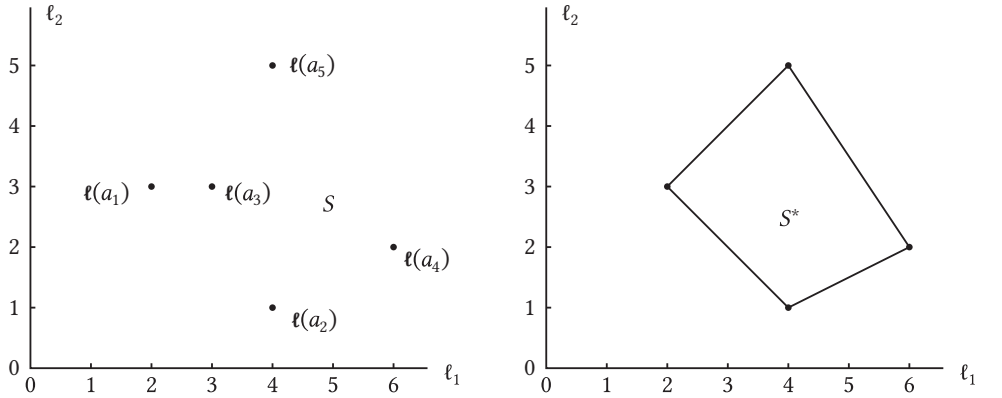


Fig. 3.1

□

Exemplo 3.2 — Para mostrar a vantagem da introdução de acções mistas, considere-se a matriz perca,

	a_1	a_2	a_3
θ_1	4	1	3
θ_2	1	4	3

Tem-se,

$$S = \{(4,1), (1,4), (3,3)\},$$
$$S^* = \{(4\zeta_1 + \zeta_2 + 3\zeta_3,\ \zeta_1 + 4\zeta_2 + 3\zeta_3) : \zeta_i \geqq 0,\ \zeta_1 + \zeta_2 + \zeta_3 = 1\}.$$

Pergunta-se: há lugar para a escolha de a_3? À primeira vista parece que sim, pois a_3 é preferível a a_1 quando $\theta = \theta_1$ e preferível a a_2 quando $\theta = \theta_2$. Se o decisor resolve optar por a_1 com probabilidade $1/2$ e por a_2 com probabilidade $1/2$, isto é, escolher a acção mista a^* associada com o vector de probabilidades, $(1/2,\ 1/2)$, a perca esperada tem por valor,

$$\text{para } \theta = \theta_1 : \ell_1(a^*) = 4(1/2) + 1(1/2) = 5/2 < 3;$$
$$\text{para } \theta = \theta_2 : \ell_2(a^*) = 1(1/2) + 4(1/2) = 5/2 < 3.$$

Deste modo, qualquer que seja o estado, a acção mista a^* é sempre preferível a a_3 (veja-se Fig. 3.2). Por outras palavras, no conjunto $A = \{a_1, a_2, a_3\}$, todas as acções são admissíveis; no conjunto A^*, a acção a_3 é inadmissível por ser dominada estritamente pela acção mista,

$$a^* = < a_1 > \frac{1}{2} + < a_2 > \frac{1}{2}.$$

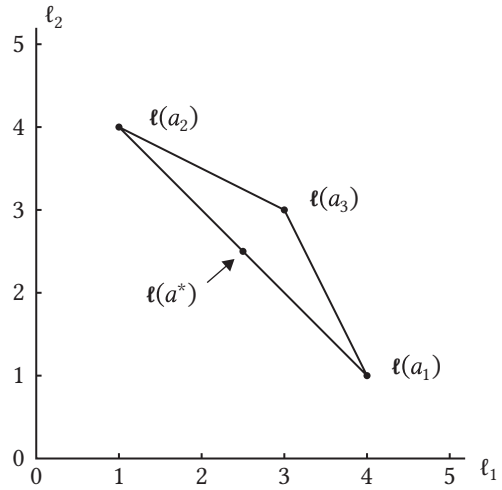


Fig. 3.2

□

Quando A é infinito numerável ou contínuo, uma acção mista, $a^* \in A^*$, tem associada uma distribuição de probabilidade sobre A, seja ζ. Generalizando (3.2), tem-se o valor esperado,

$$L(\theta, a^*) = E\{L(\theta, Z)\}, \tag{3.2'}$$

onde Z é uma variável aleatória assumindo valores em A com distribuição ζ. Em tudo o que segue supõe-se $L(\theta, a^*)$ finito para todo o $a^* \in A^*$.

3.2 Funções de decisão mistas e aleatórias

Nos problemas de decisão com dados representa importante papel a função risco, $R(\theta, \delta) = E_\theta\{L[\theta, \delta(X)]\}$, $\theta \in \Theta$, $\delta \in D$, definida em (1.67), pois corresponde à perca esperada que muitas vezes o decisor procura minimizar (pelo menos o decisor frequencista).

Se $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$, o conjunto de $\mathbf{R}^n$,

$$S_D = \{\boldsymbol{\ell} = (\ell_1, \ell_2, \dots, \ell_n) \,:\, \ell_j = R(\theta_j, \delta), \quad \delta \in D\} \equiv \{\boldsymbol{\ell}(\delta) \,:\, \delta \in D\}, \tag{3.6}$$

designa-se por conjunto risco. Ao conjunto D pertencem sempre as funções de decisão $\delta(x) \equiv a$, para cada $a \in A$; logo, $S \subset S_D$, onde S é o conjunto perca (3.4).

A relação, $S \subset S_D$, mostra que o conjunto risco sofre uma ampliação relativamente ao conjunto perca por força da experimentação. Os custos da experimen-

tação são normalmente compensados pelo alargamento do espaço de manobra do decisor e pelas potencialidades de redução do risco.

Na secção anterior foi apreciado como a introdução de acções mistas permitia já uma certa ampliação de S, conduzindo a S^*, com $S \subset S^*$. Na decisão com dados existe também a possibilidade de usar funções de decisão mistas por combinação ou mistura de funções de decisão puras, isto é, elementos $\delta \in D$[1]. Em geral, δ^*, representa uma função de decisão mista e D^* o respectivo conjunto; δ^* equivale a uma distribuição de probabilidade sobre D. Naturalmente, $D \subset D^*$, pois as distribuições de probabilidade degeneradas, que atribuem probabilidade 1 a cada $\delta \in D$, são casos particulares em que sai $\delta^* \equiv \delta$.

Quando $\delta^* \in D^*$ resulta da combinação de número finito ou infinidade numerável de funções de decisão puras, $\delta_i \in D$, $i = 1, 2, \ldots, \delta^*$ está associada com uma distribuição discreta ou vector de probabilidade, seja ζ. A função risco assume a nova forma,

$$R(\theta, \delta^*) = \sum_i R(\theta, \delta_i)\zeta_i. \tag{3.7}$$

No caso geral, em que δ^* está associada com uma distribuição, seja ζ,

$$R(\theta, \delta^*) = E\{R(\theta, Z)\}, \tag{3.8}$$

onde Z é um elemento aleatório assumindo valores em D e tendo distribuição ζ, Quando um tratamento rigoroso é necessário há que instituir D em espaço de probabilidade.

Empregando funções de decisão mistas passa a ter-se o conjunto risco casualizado,

$$S_D^* = \{\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n) : \ell_j = R(\theta_j, \delta^*), \quad \delta^* \in D^*\} \equiv \{\boldsymbol{\ell}(\delta^*) : \delta^* \in D^*\}. \tag{3.9}$$

Tem-se, $S_D \subset S_D^*$, em virtude de $D \subset D^*$. Se D é conjunto finito, resulta imediatamente de (3.7) que S_D^* é o ambiente convexo de S_D. Com D qualquer, demonstra-se [veja-se Ferguson (1967)] que S_D^* continua sendo o ambiente convexo de S_D.

[1] A casualização ou mistura de acções ou de funções de decisão é um procedimento importado da Teoria dos Jogos que faz confusão a muita gente. Na Teoria dos Jogos a casualização de estratégias é feita com o propósito de manter o adversário na ignorância da escolha. Mas porquê casualizar quando não se tem pela frente um adversário inteligente disposto a infligir a maior perca possível? A ideia, contra a qual se insurge Fisher (1956), é dar maior escopo à teoria da decisão tornando possível a exposição de certas matérias nomeadamente as que se prendem com a teoria minimax adiante aflorada (veja-se também, no Capítulo 7, início da secção 7.2, a vantagem da casualização no ensaio de hipóteses). No entanto, como diz Berger (1985), a casualização raramente é proposta para usos práticos. Aliás importa não confundir a casualização introduzida aqui com a casualização utilizada no planeamento de experiências cuja fecundidade ninguém põe em dúvida.

Exemplo 3.3 — Para apreciar as vantagens da experimentação e da casualização de forma muito esquemática, retome-se o Ex. 1.33. Na decisão sem dados a matriz perca é,

	a_1	a_2	a_3
θ_1	4	5	2
θ_2	4	0	5

A possibilidade de examinar o barómetro conduz à matriz risco,

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
θ_1	4	5	2	4,2	4,8	3,6	2,4	4,4	2,6
θ_2	4	0	5	0,4	3,6	4,9	4,1	4,5	0,5

As funções de decisão puras, $D = \{\delta_1, \dots, \delta_9\}$, foram indicadas na pág. 100. A matriz risco contém nas três primeiras colunas as funções de decisão, $\delta_1 \equiv a_1$, $\delta_2 \equiv a_2$, $\delta_3 \equiv a_3$, que ignoram o resultado da leitura do barómetro.

Na Fig. 3.3 inscrevem-se os pontos, $[R(\theta_1, \delta_i), R(\theta_2, \delta_i)]$, para $i = 1, 2, \dots, 9$. A região limitada a tracejado corresponde ao conjunto S^* e a limitada a traço grosso corresponde a S_D^*. Como é evidente, $S^* \subset S_D^*$, i.e., na decisão com dados o campo de escolha do decisor é em geral mais alargado.

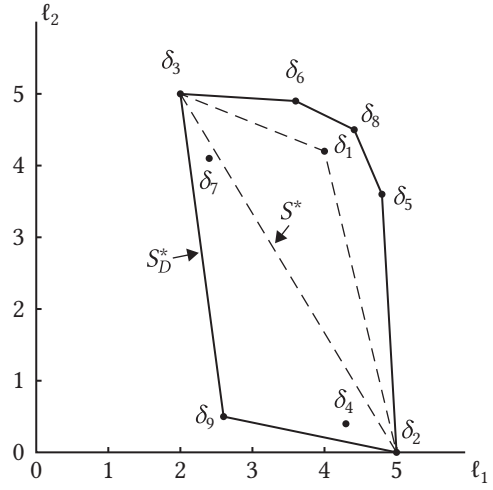


Fig. 3.3

□

Quando Θ não é finito deixa de poder trabalhar-se com os conjuntos de risco, S_D ou S_D^*[2]. Para ter uma ideia da bondade das funções de decisão pode analisar-se o comportamento da função risco, $R(\theta, \delta)$ ou $R(\theta, \delta^*)$, para alguns elementos de D ou de D^*, quando θ percorre Θ.

Exemplo 3.4 — A característica de um dado universo, seja X, tem distribuição exponencial negativa,

$$f(x \mid \theta) = (1/\theta) \exp\{-x/\theta\}, \quad x > 0,\ \theta > 0.$$

Pretende estimar-se o parâmetro θ. O espaço de estados ou do parâmetro é, $\Theta = (0, \infty)$; uma acção, a, interpreta-se no presente contexto como a proposta do número, a, como estimativa de θ. Assim, para o espaço de acções toma-se $A = (0, \infty)$, donde, $A \equiv \Theta$, como sucede em geral nos problemas de estimação.

A experiência consiste, por hipótese, na observação de N variáveis aleatórias, X_i I.I.D. a X. Uma função de decisão, $\delta \in D$, faz corresponder uma estimativa $a \in A$ a cada ponto $\mathbf{x} \in \mathcal{X}$, onde,

$$\mathcal{X} = \{(x_1, x_2, \ldots, x_N) : x_i > 0, \quad i = 1, 2, \ldots, N\}.$$

Por outras palavras, função de decisão é estimador.

Considerem-se as três funções de decisão,

$$\begin{aligned}
\delta_1(\mathbf{x}) &= \sum x_i/N = \overline{x}, \\
\delta_2(\mathbf{x}) &= \overline{x} + 1, \\
\delta_3(\mathbf{x}) &= x_N \text{ (última observação)},
\end{aligned}$$

que não esgotam, como é evidente, o conjunto D. A respectiva função risco pode calcular-se desde que seja dada a função perca. Suponha-se,

$$L(\theta, a) = (a - \theta)^2;$$

vem,

$$\begin{aligned}
R(\theta, \delta_1) &= E_\theta\{[\delta_1(\mathbf{X}) - \theta]^2\} = E_\theta\{(\overline{X} - \theta)^2\} = \theta^2/N; \\
R(\theta, \delta_2) &= E_\theta\{\delta_2(\mathbf{X}) - \theta]^2\} = (\theta^2/N) + 1; \\
R(\theta, \delta_3) &= E_\theta\{[\delta_3(\mathbf{X}) - \theta]^2\} = \theta^2,
\end{aligned}$$

expressões cujo cálculo se deixa como exercício. As funções risco apresentam-se na Fig. 3.4. Entre as três funções de decisão consideradas, a melhor é δ_1; δ_1 domina estritamente δ_2 e δ_3 que são, portanto, inadmissíveis.

[2] Que dão um panorama geral das consequências das possíveis escolhas e permitem substituir conjuntos provavelmente complicados (D ou D^*) por simples conjuntos de $\mathbf{R}^n$.

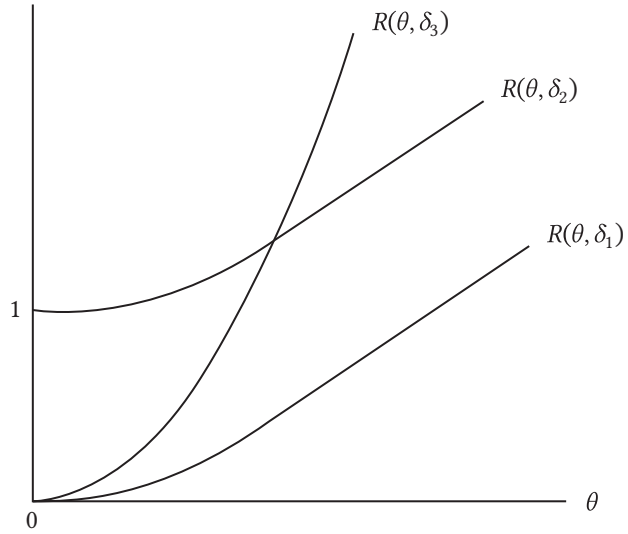


Fig. 3.4

□

Revela-se muitas vezes conveniente recorrer a um processo de casualização de funções de decisão diferente do atrás apresentado. Aliás, alguns autores [Berger (1980), por exemplo] empregam exclusivamente funções de decisão obtidas pelo processo de casualização que passa a descrever-se.

Considere-se o produto cartesiano, $\mathcal{X} \times A$ e seja Φ a classe de funções ϕ com domínio em $\mathcal{X} \times A$, tais que, para cada $x \in \mathcal{X}$, $\phi(a \mid x)$ é uma distribuição de probabilidade sobre A; ϕ é função de decisão aleatória e Φ a classe de funções de decisão aleatórias. Tomar $\phi \in \Phi$ equivale a adoptar a seguinte regra: realizada a experiência e observado, $x \in \mathcal{X}$, o decisor escolhe um $a \in A$ de acordo com a distribuição $\phi(a \mid x)$.

Para precisar a ideia de função de decisão aleatória, suponha-se A finito, $A = \{a_1, a_2, \ldots, a_m\}$; neste caso, para cada $x \in \mathcal{X}$, $\phi(a_i \mid x)$, $i = 1, 2, \ldots, m$, representa a probabilidade com que é tomada a acção a_i condicionada pela observação de x. Assim, qualquer que seja $\phi \in \Phi$, tem-se,

$$\begin{aligned} &\phi \equiv [\phi(a_1 \mid x), \phi(a_2 \mid x), \ldots, \phi(a_m \mid x)], \\ &\phi(a_i \mid x) \geqq 0, i = 1, 2, \ldots, m, \sum_i \phi(a_i \mid x) = 1. \end{aligned} \tag{3.10}$$

A respectiva esquematização é feita na Fig. 3.5.

De modo geral, função de decisão aleatória, $\phi \in \Phi$, é função que a cada $x \in \mathcal{X}$ faz corresponder uma acção mista, isto é, um elemento $a^* \in A^*$; esta ideia é parti-

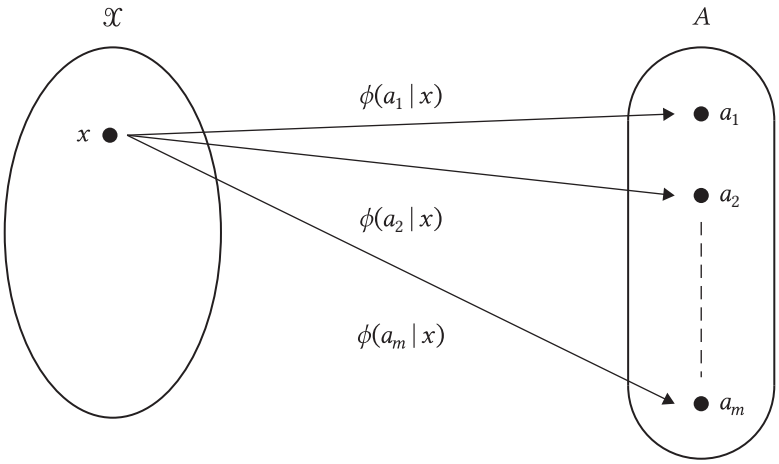


Fig. 3.5

cularmente clara no caso A finito. Função de decisão pura, $\delta \in D$, é função que a cada $x \in \mathcal{X}$ faz corresponder uma acção pura, isto é, um elemento $a \in A$.

Quando se recorre a uma função de decisão mista, a casualização sobre D que a mesma implica pode fazer-se indiferentemente antes ou depois de ser conhecido o resultado da observação, x. Quando se emprega uma função de decisão aleatória, a casualização sobre A que a mesma implica só pode fazer-se, em geral, depois de conhecido x. Parece assim intuitivo que as primeiras não são mais gerais do que as segundas, pois a casualização sobre D depois de conhecido x é perfeitamente equivalente a uma casualização sobre A.

Antes de referir de passagem o problema da equivalência entre os dois métodos de casualização convém introduzir a função risco, $\hat{R}(\theta, \phi)$, com $\phi \in \Phi$, fazendo a distinção relativamente a $R(\theta, \delta)$ e $R(\theta, \delta^*)$.

Por definição de perca esperada,

$$\hat{R}(\theta, \phi) = E_\theta\{L_X(\theta, \phi)\}, \tag{3.11}$$

onde,

$$L_x(\theta, \phi) = E\{L(\theta, Z)\}, \tag{3.12}$$

com Z elemento aleatório assumindo valores em A segundo a distribuição $\phi(a \mid x)$.

Por exemplo, com $A = \{a_1, a_2, \ldots, a_m\}$, tem-se,

$$L_x(\theta, \phi) = \sum_{i=1}^{m} L(\theta, a_i)\phi(a_i \mid x),$$

donde,

$$\hat{R}(\theta, \phi) = \int_{\mathfrak{X}} \left[\sum_i L(\theta, a_i) \phi(a_i \mid x) \right] f(x \mid \theta)\, dx,$$

no caso contínuo, ou,

$$\hat{R}(\theta, \phi) = \sum_x \left[\sum_i L(\theta, a_i) \phi(a_i \mid x) \right] f(x \mid \theta),$$

no caso discreto.

Se $A = \mathbf{R}$,

$$L_x(\theta, \phi) = \int_A L(\theta, a) \phi(a \mid x)\, da,$$

entendendo, $\phi(a \mid x)$, como densidade sobre A.

Em certas condições, demonstra-se — o tratamento mais geral deve-se a Wald e Wolfowitz (1951) — que os dois processos de casualização são equivalentes, quer dizer, para cada $\delta^* \in D^*$ existe um $\phi \in \Phi$ tal que para todo o $\theta \in \Theta$, $R(\theta, \delta^*) = \hat{R}(\theta, \phi)$ e reciprocamente.

O emprego de funções de decisão casualizadas — mistas ou aleatórias — vai ser reduzido ao mínimo, limitando-se aqueles casos em que do mesmo resultam claras vantagens.

Equacionado um problema de decisão com dados, a questão fundamental consiste na escolha de uma função de decisão[3] $\delta^* \in D^*$. Logicamente, a primeira ideia que ocorre é procurar um δ^* tal que, $R(\theta, \delta^*) \leqq R(\theta, \delta')$, para qualquer $\delta' \in D^*$ e todo o $\theta \in \Theta$. A desilusão não se faz esperar; função de decisão com risco mínimo para todo o $\theta \in \Theta$ (risco uniformemente mínimo) existe raramente e só em situações triviais. Normalmente, ocorre o seguinte: uma função de decisão que é óptima para um dado estado θ', deixa de o ser quando $\theta \neq \theta'$.

Exemplo 3.4 — *Continuação.* Considere-se a função de decisão,

$$\delta_0(\mathbf{x}) \equiv \theta_0,$$

com função risco,

$$R(\theta, \delta_0) = (\theta_0 - \theta)^2,$$

representada na Fig. 3.6 juntamente com $R(\theta, \delta_1)$. Se o estado é θ_0, o melhor que o decisor tem a fazer é utilizar δ_0; no entanto, se $\theta = \theta' \neq \theta_0$, $\delta'(\mathbf{x}) \equiv \theta'$ passa a ser melhor do que δ_0. Como é de verificação imediata, não existe função de decisão óptima para todos os estados.

[3] Ou $\phi \in \Phi$.

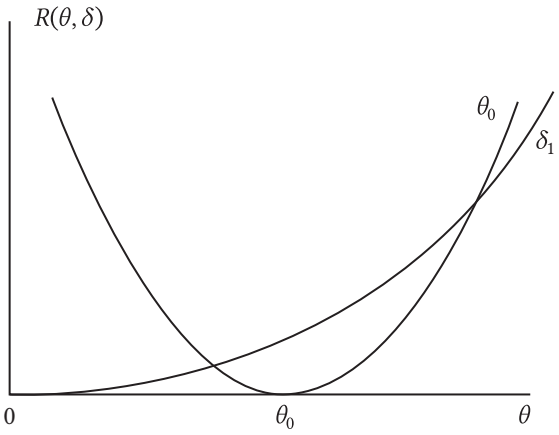


Fig. 3.6

□

São conhecidas duas vias para sair do «impasse» a que leva a não existência de função de decisão uniformemente melhor do que todas as outras, i.e., com risco uniformemente mínimo.

A primeira via consiste em considerar o conjunto D^* demasiado amplo, contendo funções de decisão, como δ_0 no exemplo anterior, muito desequilibradas na medida em que minimizam o risco para um só estado e são insensíveis ao que se passa com os outros estados. Segundo esta óptica limita-se ou restringe-se o estudo a um subconjunto de D^* cujos elementos possuam certas características de «imparcialidade» ou simetria e procura-se optimizar dentro desse subconjunto. A restrição de D^* pode fazer-se recorrendo aos conceitos de invariância e de não enviesamento que adiante se estudam. Advirta-se desde já que a escolha em classes restritas pode levar a incoerências.

A segunda via consiste em introduzir princípios ou critérios de escolha — problema «filosófico» — determinando em D^* a função (ou funções) óptima à luz desses princípios — problema «técnico». Enquadram-se nesta orientação as soluções minimax e Bayes que passam a estudar-se[4].

Há uma diferença entre as duas vias que importa destacar. Enquanto depois de introduzir os conceitos de não enviesamento ou de invariância se insiste em procurar uma função de decisão uniformemente melhor do que todas as outras da classe restrita, isto é, procurar $\delta \in D_o^*$ (classe restrita de D^*), tal que,

$$R(\theta, \delta) \leqq R(\theta, \delta^*) \text{ para todo o } \theta \in \Theta \text{ e todo o } \delta^* \in D_o^*,$$

[4] Aliás já referidas na secção 1.7.

no caso dos critérios minimax ou Bayes procura-se optimizar num sentido global, isto é, procura-se $\delta \in D^*$ (classe universal) que minimize medidas globais de risco, como

$$\sup_{\theta \in \Theta} R(\theta, \delta) \qquad \text{[critério minimax]},$$

ou,

$$\int_{\Theta} R(\theta, \delta) h(\theta)\, d\theta \qquad \text{[critério risco Bayes]}.$$

Evidentemente, em relação à classe universal, a optimização no sentido global conduz sempre a propriedade mais fraca do que a traduzida por risco uniformemente mínimo.

3.3 Funções de decisão minimax

O princípio minimax é pessimista por excelência: recomenda ao decisor que conte com uma «natureza» radicalizada na sua hostilidade. O decisor deve preparar-se para o pior, isto é, considerar em relação a cada função de decisão, $\delta^* \in D^*$, a máxima perca esperada ou risco em que pode incorrer[5],

$$\sup_\theta R(\theta, \delta^*), \quad \delta^* \in D^*, \tag{3.13}$$

e proceder então à escolha da função de decisão que minimiza a máxima perca esperada ou risco.

Função de decisão, $\tilde{\delta}^* \subset D^*$, é minimax quando verifica,

$$\sup_\theta R(\theta, \tilde{\delta}^*) = \inf_{\delta^*} \sup_\theta R(\theta, \delta^*); \tag{3.14}$$

se existir $\tilde{\delta}^*$ com esta propriedade o decisor que a emprega pode estar seguro que não tem risco superior ao segundo membro de (3.14), designado risco minimax.

É simples exercício verificar que $\tilde{\delta}^*$ é minimax se e somente se,

$$R(\theta', \tilde{\delta}^*) \leqq \sup_\theta R(\theta, \delta^*), \tag{3.15}$$

para todo $\theta' \in \Theta$ e todo o $\delta^* \in D^*$.

O princípio minimax teve origem na teoria dos jogos estratégicos e fornece soluções aceitáveis quando se defrontam dois oponentes inteligentes e o ganho de um é a perca do outro — jogos com soma nula. Nos jogos estatísticos, em que os adversários são o decisor e a «natureza», parece atitude extremamente pessimista admitir que a natureza actua sistematicamente contra o decisor, procurando infligir-lhe o maior prejuízo possível. Tal atitude, além de não encontrar suporte na

[5] Escreve-se $\sup_\xi$ ou $\inf_\xi$ em vez de $\sup_{\xi \in \Xi}$ ou $\inf_{\xi \in \Xi}$ sempre que não haja dúvida sobre o conjunto Ξ.

teoria da utilidade, conduz por vezes a decisões muito pouco razoáveis. Apesar das suas limitações, o princípio minimax continua a ser aplicado na decisão estatística, não sendo de excluir a hipótese de que seja o mais recomendável em determinado tipo de problemas.

Exemplo 3.3 — *Continuação.* O critério minimax pode servir de orientação na escolha de uma acção na decisão sem dados que é, aliás, caso muito particular da decisão com dados.

Considere-se a matriz perca,

	a_1	a_2	a_3
θ_1	4	5	2
θ_2	4	0	5
$\max_\theta L(\theta, a_i)$	**4**	5	5

Como imediatamente se reconhece, a acção pura minimax, $\tilde{a}$, é $\tilde{a} = a_1$, porquanto,

$$\max_\theta L(\theta, a_1) = \min_a \max_\theta L(\theta, a) = 4,$$

isto é,

$$\max_\theta L(\theta, a_1) = 4 < \max_\theta L(\theta, a_2) = \max_\theta L(\theta, a_3) = 5.$$

Considere-se a matriz risco quando $D = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_9\}$;

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
θ_1	4	5	2	4,2	4,8	3,6	2,4	4,4	2,6
θ_2	4	0	5	0,4	3,6	4,9	4,1	4,5	0,5
$\max_\theta L(\theta, \delta_i)$	4	5	5	4,2	4,8	4,9	4,1	4,5	**2,6**

A função de decisão pura minimax $\tilde{\delta} = \delta_9$,

$$\max_\theta L(\theta, \delta_9) = \min_\delta \max_\theta R(\theta, \delta) = 2{,}6;$$

ao empregar δ_9 o decisor está seguro que a perca esperada ou risco nunca sai superior a 2,6 (risco minimax).

Adiante mostram-se as vantagens que podem colher-se da casualização no quadro do mesmo exemplo. □

Exemplo 3.5 — Com ξ número real, arbitrariamente pequeno e positivo, considere-se a matriz perca em baixo. A aplicação do critério minimax conduz à escolha de a_3, sugestão que ninguém de bom senso aceita.

	a_1	a_2	a_3
θ_1	ξ	ξ	$100-\xi$
θ_2	ξ	ξ	$100-\xi$
θ_3	100	100	$100-\xi$

□

No caso Θ finito, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, a interpretação geométrica das soluções minimax é feita recorrendo ao conceito de ortante inferior. Para qualquer $\mathbf{z} \in \mathbf{R}^n$, o ortante (quadrante se $n = 2$, octante se $n = 3$) inferior em $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ é o conjunto,

$$Q_{\mathbf{z}} = \{\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n) \,:\, \ell_j \leqq z_j, \quad j = 1, 2, \ldots, n\}. \tag{3.16}$$

Com $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n)$ ponto de S_D^*, o máximo risco em que o decisor pode incorrer é $\max_j \ell_j$. Assim, no que respeita à solução minimax, os pontos de S_D^* que conduzem ao mesmo valor para $\max_j \ell_j$, são equivalentes.

Se $\max_j \ell_j = c$, são pontos equivalentes os que pertencem à intersecção de S_D^* com a fronteira do ortante em $\mathbf{c} = (c, c, \ldots, c)$,

$$Q_{\mathbf{c}} = \{\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n) \,:\, \ell_j \leqq c, j = 1, 2, \ldots, n\}.$$

Portanto, para obter a solução minimax procura-se o ínfimo dos valores c, seja c_0, para os quais a intersecção é não vazia,

$$Q_{\mathbf{c}_0} \cap S_D^* \neq \emptyset;$$

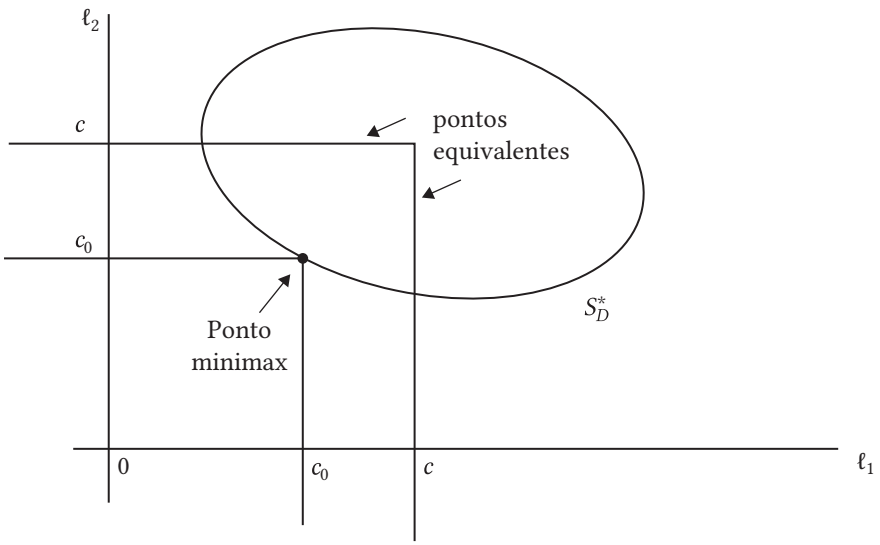


Fig. 3.7

então, se existir c_0 e se houver $\tilde{\delta}^* \in D^*$ tal que[6],

$$\boldsymbol{\ell}(\tilde{\delta}^*) \in Q_{c_0} \cap S_D^*, \tag{3.17}$$

$\tilde{\delta}^*$ é função de decisão mista minimax (veja-se a Fig. 3.7).

A Fig. 3.8 exemplifica três casos de interpretação geométrica das soluções minimax.

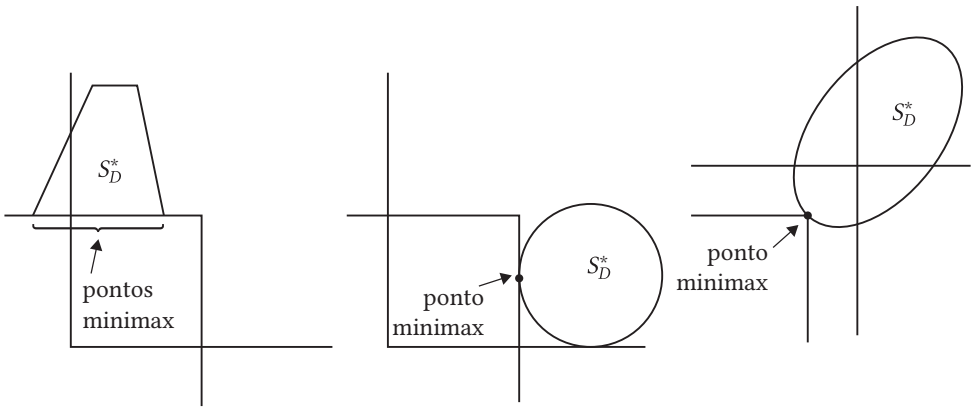


Fig. 3.8

Exemplo 3.3 — *Continuação.* Na Fig. 3.9 procede-se a uma ampliação da Fig. 3.3, retendo apenas a localização dos pontos de interesse para o estudo que segue: $\boldsymbol{\ell}(\delta_9), \boldsymbol{\ell}(\delta_2) \equiv \boldsymbol{\ell}(a_2)$ e $\boldsymbol{\ell}(\delta_3) \equiv \boldsymbol{\ell}(a_3)$.

Analisando a Fig. 3.9 rapidamente se conclui que a acção mista minimax, $\tilde{a}^*$, é dada por combinação de a_2 e a_3, com probabilidades que é simples exercício mostrar serem 3/8 e 5/8 respectivamente. A perca minimax é,

$$L(\theta_1, \tilde{a}^*) = L(\theta_2, \tilde{a}^*) = L(\theta_1, a_2)(3/8) + L(\theta_1, a_3)(5/8) = 5(3/8) + 2(5/8) = 3{,}125.$$

Quando se emprega a acção pura minimax a respectiva perca é 4. Do ponto de vista minimax revela-se, portanto, vantajoso casualizar sobre o espaço de acções. A função de decisão mista minimax, $\tilde{\delta}^*$, resulta da combinação de δ_3 e δ_9, com probabilidades, 0,41 e 0,59, respectivamente; o risco minimax tem por valor,

$$\begin{aligned} R(\theta_1, \tilde{\delta}^*) = R(\theta_2, \tilde{\delta}^*) &= R(\theta_1, \delta_3)(0{,}41) + R(\theta_1, \delta_9)(0{,}59) \\ &= 2{,}354. \end{aligned}$$

Note-se a melhoria conseguida em relação à função de decisão pura minimax, δ_9, cujo risco minimax é igual a 2,6. Do ponto de vista minimax — e como seria de

[6] Recorde-se: $\boldsymbol{\ell}(\delta^*) = [\ell_1(\delta^*), \ell_2(\delta^*), \ldots, \ell_n(\delta^*)] = [R(\theta_1, \delta^*), R(\theta_2, \delta^*), \ldots, R(\theta_n, \delta^*)]$.

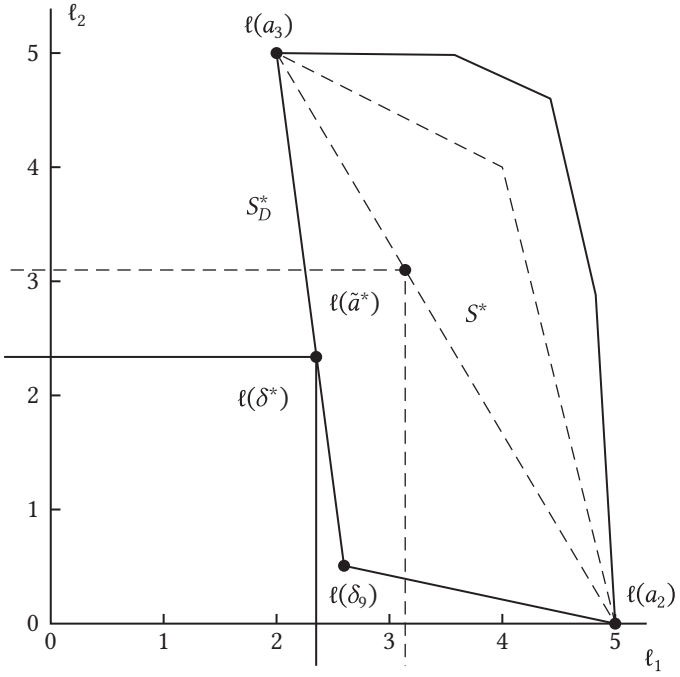


Fig. 3.9

esperar dado o que se verificou com as acções puras — é vantajoso casualizar sobre o espaço de funções de decisão. □

Exemplo 3.6 — Uma empresa de transportes aéreos tem oportunidade de adquirir um grupo de aviões usados e as acções possíveis são:

a_1 — comprar o grupo de aviões;
a_2 — não comprar o grupo de aviões.

Sabe-se que 1000 horas de voo representam o ponto crítico na exploração de aviões do tipo dos daquele grupo: se um avião voa 1000 horas ou mais a empresa realiza um resultado positivo r; se voa menos realiza um resultado negativo s. O estado da natureza é o número de aviões em condições de voar 1000 horas ou mais: $\Theta = \{0, 1, 2, \ldots, 10\}$. Antes de tomar uma decisão a empresa pode experimentar um dos aviões; tem-se $\mathcal{X} = \{x_1, x_2\}$, onde x_1, indica teste satisfatório e x_2 indica teste não satisfatório. A distribuição de X, no caso presente função de probabilidade, depende de θ,

$$f(x_1 \,|\, \theta) = \frac{\theta}{10}; \quad f(x_2 \,|\, \theta) = 1 - \frac{\theta}{10}.$$

As quatro funções de decisão encontram-se no quadro abaixo:

	δ_1	δ_2	δ_3	δ_4
x_1	a_1	a_2	a_1	a_2
x_2	a_1	a_2	a_2	a_1

A função perca é dada por,

$$L(\theta, a_1) = -r\theta + s(10 - \theta); \quad L(\theta, a_2) = 0.$$

Para as funções risco, vem,

$$\begin{aligned}
R(\theta, \delta_1) &= L(\theta, a_1) f(x_1 \,|\, \theta) + L(\theta, a_1) f(x_2 \,|\, \theta) = -r\theta + s(10 - \theta); \\
R(\theta, \delta_2) &= L(\theta, a_2) f(x_1 \,|\, \theta) + L(\theta, a_2) f(x_2 \,|\, \theta) = 0; \\
R(\theta, \delta_3) &= L(\theta, a_1) f(x_1 \,|\, \theta) + L(\theta, a_2) f(x_2 \,|\, \theta) \\
&= [-r\theta + s(10 - \theta)] \left[\frac{\theta}{10}\right]; \\
R(\theta, \delta_4) &= L(\theta, a_2) f(x_1 \,|\, \theta) + L(\theta, a_1) f(x_2 \,|\, \theta) \\
&= [-r\theta + s(10 - \theta] \left[1 - \frac{\theta}{10}\right].
\end{aligned}$$

A representação gráfica, no caso particular $r = 2$ e $s = 3$, faz-se na Fig 3.10, devendo notar-se que só os valores inteiros de θ, $0 \leqq \theta \leqq 10$, são relevantes.

Como pode ver-se não existe nenhuma função de decisão que seja a melhor para todos os estados; δ_3, não sendo a melhor para nenhum dos estados, é bastante razoável em relação ao conjunto dos estados.

Por outro lado, notando que,

$$\begin{aligned}
\max_\theta R(\theta, \delta_1) &= 30, \\
\max_\theta R(\theta, \delta_2) &= \ \ 0, \\
\max_\theta R(\theta, \delta_3) &= \ \ 4{,}5, \\
\max_\theta R(\theta, \delta_4) &= 30,
\end{aligned}$$

conclui-se que a função de decisão pura minimax é δ_2, facto que põe em relevo a atitude de extrema prudência ou conservadora recomendada pelo critério minimax [Silvey (1970)].

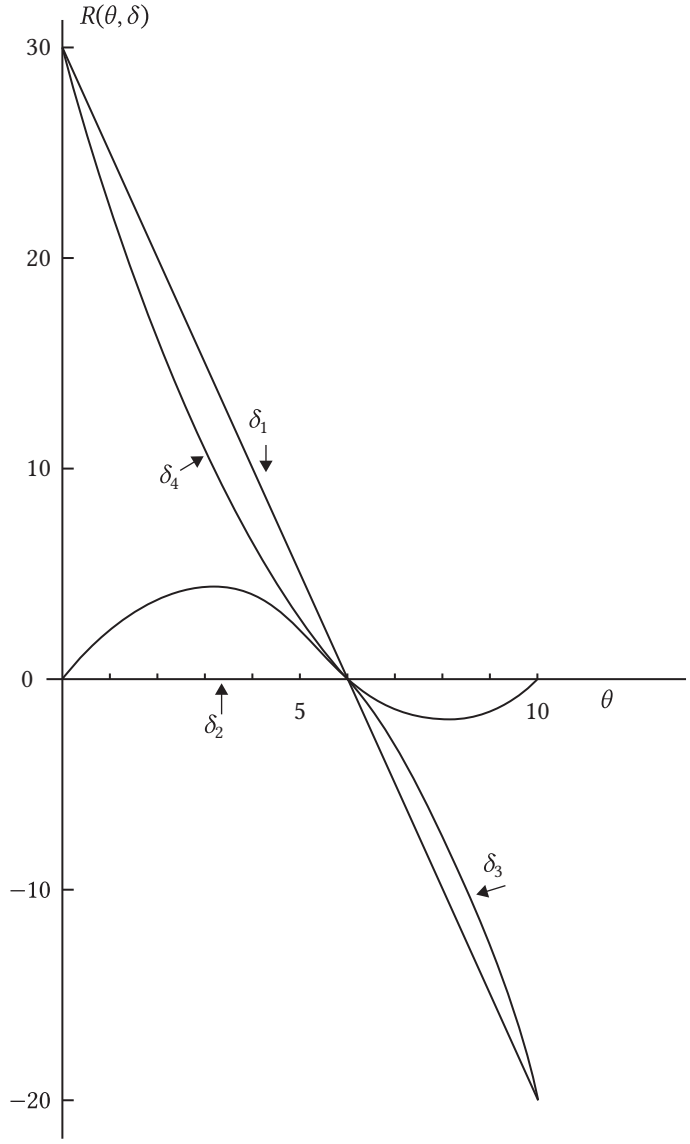


Fig. 3.10

□

Quando não existe função de decisão minimax — os problemas de existência são melhor abordados no capítulo seguinte — o decisor pode recorrer a uma função de decisão minimax-ξ, com ξ número real arbitrariamente pequeno e positivo; $\tilde{\delta}^{**} \in D^*$ é minimax-ξ se,

$$\sup_\theta R(\theta, \tilde{\delta}^{**}) \leqq \inf_{\delta^*} \sup_\theta R(\theta, \delta^*) + \xi. \tag{3.18}$$

Usando $\tilde{\delta}^{**}$ o decisor pode estar seguro de que a sua perca esperada não excede o risco minimax em mais do que ξ. Na Fig. 3.11, ainda no caso Θ finito, indica-se o conjunto de pontos de S_D^* correspondentes a funções de decisão minimax-ξ.

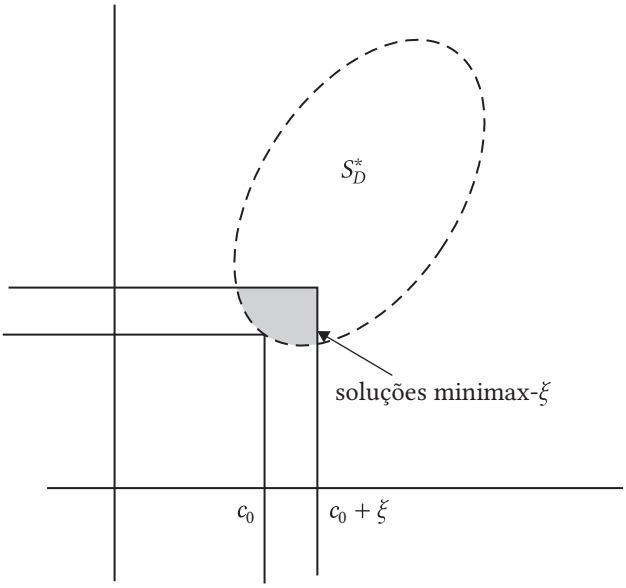


Fig. 3.11

Se, por qualquer motivo, o estado θ fosse conhecido antecipadamente, o decisor, actuando racionalmente, escolheria $\delta' \in D^*$ tal que,

$$R(\theta, \delta') = \inf_{\delta^*} R(\theta, \delta^*).$$

Esta perca mínima não pode ser evitada em nenhuma circunstância. Na prática o decisor não conhece o estado e tem de escolher uma função de decisão de D^*. Se opta por $\delta^* \neq \delta'$, quando vem a saber que o estado é θ sente pesar por não ter escolhido δ'. A função pesar ou perca de oportunidade é então definida, para todo o par, (θ, δ^*), pela diferença,

$$R_0(\theta, \delta^*) = R(\theta, \delta^*) - \inf_{\delta^*} R(\theta, \delta^*). \tag{3.19}$$

Em alguns problemas é indiferente empregar a função risco ou a função pesar; casos há, todavia, em que a segunda é preferível, quer pela importância que os custos de oportunidade têm nos estudos económicos, quer pela possibilidade que dá de atenuar as críticas ao critério minimax decorrentes de situações como a esquematizada no Ex. 3.5.

Uma função de decisão, $\tilde{\delta}_0 \in D^*$, tal que,

$$\sup_\theta R_0(\theta, \tilde{\delta}_0) = \inf_{\delta^*} \sup_\theta R_0(\theta, \delta^*), \tag{3.20}$$

diz-se função de decisão pesar minimax.

Exemplo 3.5 — *Continuação.* Como está em causa um problema de decisão sem dados, a função pesar tem por expressão,

$$L_0(\theta, a) = L(\theta, a) - \min_a L(\theta, a);$$

no caso presente tem-se a matriz pesar,

	a_1	a_2	a_3
θ_1	0	0	$100 - 2\xi$
θ_2	0	0	$100 - 2\xi$
θ_3	ξ	ξ	0
$\max_\theta L_0(\theta, a)$	ξ	ξ	$100 - 2\xi$

concluindo-se imediatamente que a_1 e a_2 são acções puras pesar minimax. A falta de senso na escolha de a_3 fica eliminada. □

Os dois casos seguintes, apontados por Lindley (1971b), mostram incoerências nas aplicações minimax e pesar minimax.

Exemplo 3.7 — Para ilustrar como a aplicação do critério pesar minimax pode revestir-se de incoerência, considerem-se as seguintes:

Matriz Perca

	a_1	a_2
θ_1	-8	-2
θ_2	0	-4

Matriz Pesar

	a_1	a_2
θ_1	0	6
θ_2	4	0

Evidentemente, a acção pesar minimax é a_1. Suponha-se que entretanto se identifica uma nova acção possível, a_3, cuja introdução leva à matriz de perca abaixo e obriga a recalcular a correspondente matriz pesar,

Matriz Perca

	a_1	a_2	a_3
θ_1	-8	-2	-1
θ_2	0	-4	-7

Matriz Pesar

	a_1	a_2	a_3
θ_1	0	6	7
θ_2	7	3	0

Verifica-se, agora, que a acção pesar minimax passa a ser a_2. Quer dizer, o efeito de introduzir uma nova acção, que não vem a ser recomendada, é mudar a preferência do critério de a_1 para a_2! A situação é um pouco comparável aquela em que o decisor, tendo optado por ficar em casa em vez de ir ao teatro, resolve alterar a decisão e ir ao teatro depois de saber que há um bom concerto algures... . □

Exemplo 3.8 — Considere-se a seguinte matriz perca,

	a_1	a_2
θ_1	-10	-1
θ_2	0	-1

A solução minimax é obviamente a_2. Para a matriz perca,

	a_2	a_3
θ_1	-1	0
θ_2	-1	-10

a solução minimax continua a ser a_2. Reunam-se os dois problemas, isto é, forme-se a matriz perca,

	a_1	a_2	a_3
θ_1	-10	-1	0
θ_2	0	-1	-10

A solução minimax passa a ser a acção mista que combina a_1 e a_3 com probabilidade 1/2. Não deixa de ser matéria para reflexão o facto de o critério minimax escolher a_2 quando compara a_1 e a_2, e escolher também a_2 quando compara a_2 e a_3, enquanto a opção entre a_1 e a_3 por meio do lançamento de uma moeda é preferível a a_2. □

Exemplo 3.9 — Berger (1980) apresenta uma situação em que o axioma {A3} da teoria da utilidade é nitidamente violado pelo critério minimax.

No lado esquerdo da Fig. 3.12 comparam-se as funções risco de duas funções de decisão, δ_1 e δ_2; δ_2 é claramente a preferida pelo critério minimax e naturalmente por qualquer outro critério de decisão, desde que não se considere θ concentrado na vizinhança de 1,5. Suponha-se, no entanto, que para $1 < \theta < 2$ é introduzida uma perca adicional de 1; o risco, R^*, passa a caracterizar-se pelas funções representadas no lado direito e pelo critério minimax δ_1 passa a ser preferível a δ_2, o que é ridículo. A questão que se põe é esta: porquê modificar a ordem de preferência de δ_1 e δ_2 quando se tem em conta uma perca inevitável a que ambas as funções de decisão estão igualmente sujeitas?

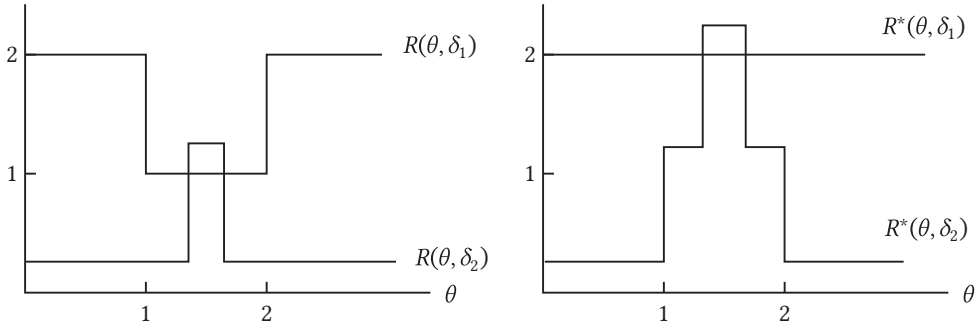


Fig. 3.12

□

O problema da existência e determinação[7] das soluções minimax — em condições pelo menos mais gerais do que as apresentadas para exemplificação — vai ser retomado no capítulo seguinte. Termina-se, por agora, com a seguinte citação: «*The last bastion of defence for the minimax principle is the argument that someone might really want to act conservatively, no matter how silly the result seems to be. The obvious counter-argument has already been mentioned, namely that conservatism will naturally be built into the utility function, and hence into the loss and risk function. Further attempts at conservatism would be overkill*». [Berger (1980)].

3.4 Funções de decisão Bayes

Enquanto o critério minimax conduz à mesma solução independentemente de quem o aplica, o critério Bayes depende da atitude do decisor, dos seus conhecimentos, informações, palpites, etc., e baseia-se na convicção de que mesmo sendo desconhecido o estado da natureza, há sempre alguma informação disponível que importa não desprezar.

O critério Bayes pressupõe que toda a informação inicial, seja de que tipo for, é susceptível de traduzir-se por uma distribuição a priori[8] sobre os estados da natureza a qual deve ser incorporada no problema de decisão. Se Θ é finito, a premissa equivale muitas vezes a admitir que o decisor atribui probabilidades subjectivas ou graus de credibilidade aos vários estados; em geral, a distribuição a priori é uma

[7] Como diz Berger (1985): « *...the minimax principie can be devilishly hard to implement*».

[8] Lehmann (1983) considera que a introdução de uma distribuição a priori ou sistema de ponderação para os estados da natureza pode entender-se como: (i) instrumento matemático, sem conotações filosóficas (é, de certo modo, a posição de Wald); (ii) forma de utilizar a experiência acumulada no passado; (iii) descrição de um «*state of mind*» (seguindo de Finetti e Savage ou Jeffreys); (iv) método geral que gera funções de decisão razoáveis nos problemas de estimação e de ensaio de hipóteses.

distribuição de probabilidade sobre Θ, seja h. O conjunto de todas as distribuições a priori consideradas designa-se por $\mathcal{H}$ [podia ser Θ^*]; evidentemente, quando $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$, tem-se $\mathcal{H} \equiv S_n$.

A função risco, $R(\theta, \delta^*)$, definida por (3.8) — veja-se também (1.67) — tem domínio em $\Theta \times D^*$; para alargar o domínio a $\mathcal{H} \times D^*$ escreve-se,

$$R(h, \delta^*) = E\{R(\theta, \delta^*)\} = \int_\Theta R(\theta, \delta^*) h(\theta)\, d\theta. \tag{3.21}$$

Em tudo o que segue supõem-se verificadas as condições que asseguram a existência da função risco. A versão alargada, $R(h, \delta^*)$, também se designa por risco Bayes; alguns autores preferem usar diferente notação, $r(h, \delta^*)$.

Dada a distribuição a priori, $h \in \mathcal{H}$, $\delta_h^* \in D^*$, diz-se função de decisão Bayes contra h, se,

$$R(h, \delta_h^*) = \inf_{\delta^*} R(h, \delta^*); \tag{3.22}$$

$R(h, \delta_h^*) = r(h)$ é o risco Bayes de h.

A solução Bayes[9] é inevitavelmente óptima à luz dos princípios de coerência porquanto minimiza, sem restrições, a perca esperada (maximiza a utilidade esperada quando se estabelece uma correcta relação entre perca e utilidade). As soluções Bayes apresentam em relação às soluções tipo minimax a vantagem de tornar desnecessária a casualização sobre o espaço das funções de decisão; quer dizer, no contexto da análise bayesiana pode dispensar-se D^* e trabalhar com D.

Suponha-se D finito, $D = \{\delta_1, \delta_2, \dots, \delta_r\}$, e seja δ^* função de decisão mista arbitrária associada com o vector de probabilidades $(\zeta_1, \zeta_2, \dots, \zeta_r)$. Tem-se,

$$R(h, \delta^*) = \sum_{i=1}^{r} R(h, \delta_i)\zeta_i \geqq \sum_{i=1}^{r} [\min_\delta R(h, \delta)]\zeta_i = \min_\delta R(h, \delta),$$

isto é,

$$\min_{\delta^*} R(h, \delta^*) \geqq \min_\delta R(h, \delta);$$

por outro lado, como $D \subset D^*$,

$$\min_{\delta^*} R(h, \delta^*) \leqq \min_\delta R(h, \delta);$$

consequentemente,

$$\min_{\delta^*} R(h, \delta^*) = \min_\delta R(h, \delta).$$

Assim, se $\delta_h^* \in D^*$ é Bayes contra h, existe $\delta_h \in D$, também Bayes contra h.

[9] Sobretudo quando obtida na forma extensiva adiante estudada. Os problemas de existência de soluções Bayes são abordados no capítulo 4. A análise do presente capítulo é meramente formal, como aliás se procedeu em relação às soluções minimax.

A demonstração no caso geral apresenta-se apenas nas suas linhas gerais. Suponha-se δ_h^* Bayes contra h e seja Z um elemento aleatório assumindo valores em D segundo a lei de probabilidade, ζ, que caracteriza δ_h^*. Por (3.8),

$$R(\theta, \delta_h^*) = \int_D R(\theta, z)\, d\zeta(z);$$

considerando (3.21) e admitindo que é legítimo permutar os integrais que figuram na expressão, vem,

$$\begin{aligned} R(h, \delta_h^*) &= \int_\Theta \left[\int_D R(\theta, z)\, d\zeta(z) \right] h(\theta)\, d\theta \\ &= \int_D R(h, z)\, d\zeta(z). \end{aligned}$$

Como δ_h^* é Bayes contra h deve verificar-se, $R(h, \delta_h^*) \leqq R(h, z)$ para todo o $z \in D$, o que implica $R(h, \delta_h^*) = R(h, Z)$ com probabilidade 1. Quer dizer, para todo o δ assumido por Z deve ter-se $R(h, \delta_h^*) = R(h,\delta)$ com probabilidade 1 e, assim, todos esses δ são Bayes contra h. A principal dificuldade da demonstração rigorosa [Ferguson (1967)] reside na permutabilidade dos referidos integrais.

Quando Θ é finito a interpretação geométrica das soluções Bayes pode fazer-se no espaço $\mathbf{R}^n$ a exemplo do que se passa com as soluções minimax. Seja $\mathbf{h} = (h_1, h_2, \ldots, h_n)$, $\mathbf{h} \in S_n$, a distribuição a priori; o risco de δ^* em relação a $\mathbf{h}$ escreve-se,

$$\begin{aligned} R(\mathbf{h}, \delta^*) &= \sum_{j=1}^{n} R(\theta_j, \delta^*) h_j \\ &= \boldsymbol{\ell}(\delta^*) \cdot \mathbf{h}, \end{aligned}$$

onde a última expressão é o produto interno dos vectores $\boldsymbol{\ell}(\delta^*)$ e $\mathbf{h}$.

Com $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n) \subset \mathbf{R}^n$, considere-se o hiperplano,

$$\Delta(\mathbf{h}, b) = \{\boldsymbol{\ell} : \boldsymbol{\ell} \cdot \mathbf{h} = b\}, \tag{3.23}$$

com b número real. Se para algum b a intersecção de $\Delta(\mathbf{h}, b)$ com S_D^* [ou com S_D] for não vazia, os pontos da intersecção correspondem a funções de decisão possuindo o mesmo risco em relação a $\mathbf{h}$, isto é, funções de decisão equivalentes do ponto de vista Bayes. O hiperplano definido por (3.23) tem as seguintes propriedades: (I) é perpendicular ao vector $\mathbf{h}$; (II) a intersecção com a bissectriz, $\ell_1 = \ell_2 = \cdots = \ell_n$ é o ponto $(b, b, \ldots, b)$.

Se existir o ínfimo dos valores b para os quais $\Delta(\mathbf{h}, b)$ tem intersecção não vazia com S_D^*, seja b_0, qualquer $\delta_h^* \in D^*$ tal que,

$$\boldsymbol{\ell}(\delta_h^*) \in \Delta(\mathbf{h}, b_0) \cap S_D^*, \tag{3.24}$$

é Bayes contra $\mathbf{h}$.

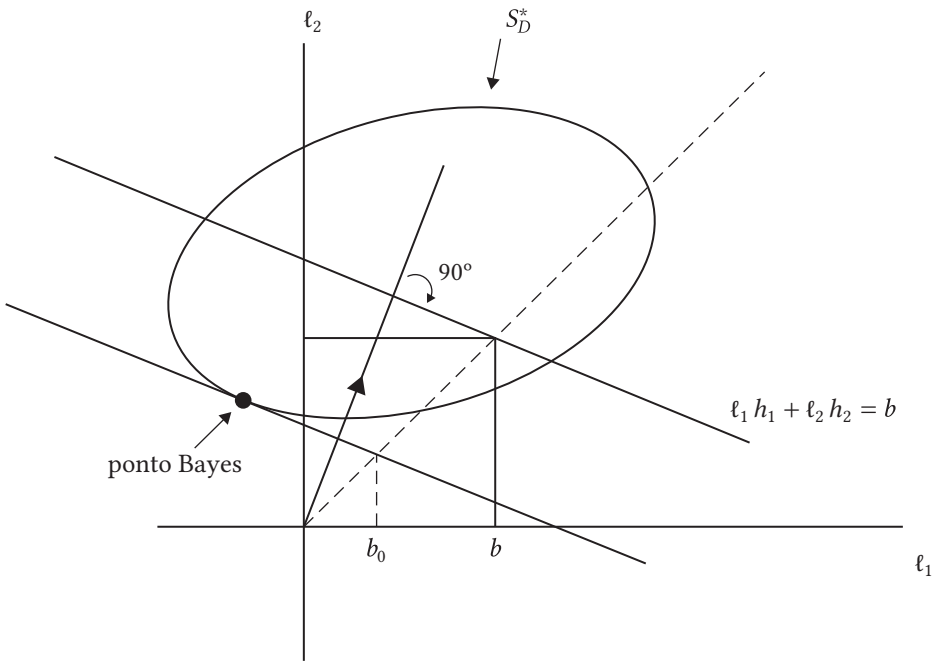


Fig. 3.13

Na Fig.3.13 faz-se a representação geométrica quando $n = 2$. As Figs. 3.14 e 3.15 referem-se ao caso particular em que S_D é finito e S_D^* é o respectivo ambiente convexo; S_D^* tem número finito de pontos extremos — é polígono convexo — cada um dos quais corresponde a uma função de decisão pura (elemento de D). Como se verifica, o ínfimo de b é atingido pelo menos num ponto extremo; quando o é em dois pontos extremos, há uma infinidade de pontos Bayes que podem obter-se por combinação linear convexa dos mesmos (Fig. 3.14). Em qualquer hipótese nada se ganha com a casualização.

A determinação das funções de decisão Bayes pode fazer-se através de duas formas de análise, inspiradas, também, na Teoria dos Jogos. A análise na forma normal equivale à adopção do princípio risco Bayes; a análise na forma extensiva equivale à adopção do princípio Bayes condicional.

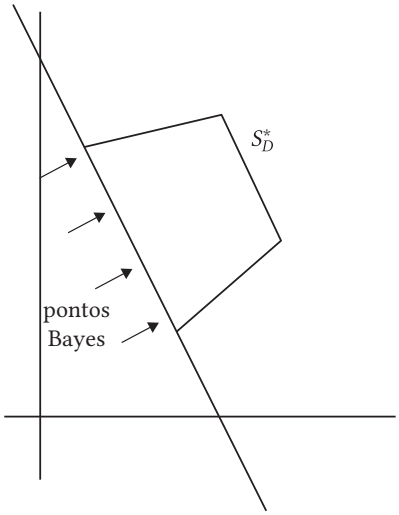


Fig. 3.14

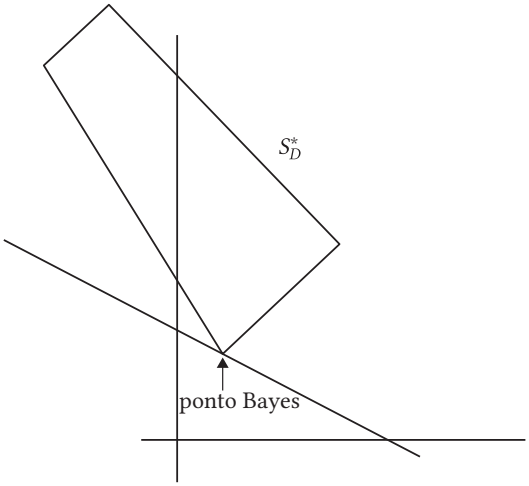


Fig. 3.15

Análise na forma normal. Procura-se uma função de decisão, $\delta_h \in D$, que minimize a função risco, $R(h, \delta_h) = \inf_\delta R(h, \delta)$, na sequência de (3.22) e das considerações que levam a dispensar a casualização por irrelevante.

Se D é finito, $D = \{\delta_1, \delta_2, \ldots, \delta_r\}$, a obtenção da função de decisão Bayes contra h é fácil: calcula-se, $R(h, \delta_i)$, $i = 1, 2, \ldots, r$, e procura-se $\delta_h \equiv \delta_j$, tal que,

$$R(h, \delta_j) \leqq R(h, \delta_i), \quad i = 1, 2, \ldots, r.$$

Exemplo 3.6 — *Continuação.* Suponha-se que o decisor atribui, a priori, a mesma credibilidade aos 11 estados. Tem-se,

$$h(\theta) = 1/11, \quad \theta = 0, 1, 2, \ldots, 10,$$

e, assim,

$$R(h, \delta_1) = (1/11) \sum_{\theta=0}^{10} [-r\theta + s(10 - \theta)] = -5r + 5s;$$

$$R(h, \delta_2) = 0;$$

$$R(h, \delta_3) = (1/11) \sum_{\theta=0}^{10} [-r\theta + s(10 - \theta)](\theta/10) = -3{,}5r + 1{,}5s;$$

$$R(h, \delta_4) = (1/11) \sum_{\theta=0}^{10} [-r\theta + s(10 - \theta)] \left[1 - \frac{\theta}{10}\right] = -1{,}5r + 3{,}5s.$$

No caso particular, $r = 2$ e $s = 3$, vem,

$$R(h, \delta_1) = 5,\ R(h, \delta_2) = 0,\ R(h, \delta_3) = -2{,}5,\ R(h, \delta_4) = 7{,}5,$$

donde se conclui que δ_3 é Bayes contra a distribuição a priori indicada. □

Quando D não é finito, basta recordar (1.69),

$$R(h, \delta) = \int_\Theta \left\{ \int_{\mathcal{X}} L[\theta, \delta(x)] f(x \mid \theta)\, dx \right\} h(\theta)\, d\theta,$$

para concluir que, em geral, a minimização de $R(h, \delta)$ para algum $\delta \in D$, é problema de cálculo das variações de difícil solução.

Análise na forma extensiva. Na expressão (1.69), acima repetida, tome-se, atendendo a (1.16) e (1.17), $f(x \mid \theta)h(\theta) = h(\theta \mid x)f(x)$, e faça-se a respectiva substituição, admitindo legítimo inverter a ordem de integração (o que se verifica se $L(\theta, a) \geqq -K > -\infty$, dado que as densidades correspondem a medidas finitas); vem,

$$R(h, \delta) = \int_{\mathcal{X}} \left\{ \int_\Theta L[\theta, \delta(x)] h(\theta \mid x)\, d\theta \right\} f(x)\, dx. \tag{3.25}$$

por ser $f(x) \geqq 0$, função de decisão, $\delta_h \in D$, que minimiza (3.25) é função que a cada $x \in \mathcal{X}$ faz corresponder o valor $\delta_h(x) \in A$ que minimiza a expressão,

$$\int_\Theta L[\theta, \delta(x)] h(\theta \mid x)\, d\theta = r_x(\delta), \tag{3.26}$$

já designada por risco a posteriori. Note-se que a função r_x tem domínio em A, pois, com x dado, $\delta(x) \in A$; por isso, em alternativa a (3.26), pode ser mais conveniente escrever,

$$r_x(a) = \int_\Theta L(\theta, a) h(\theta \mid x)\, d\theta. \tag{3.27}$$

Vale a pena comparar os dois métodos retomando um exemplo de extrema simplicidade:

Exemplo 3.1 — *Continuação.* Suponha-se que o decisor, depois de espreitar o céu e olhar as núvens, atribui à chuva credibilidade 0,4 e ao bom tempo credibilidade 0,6; tem-se, $h(\theta_1) = 0{,}4$ e $h(\theta_2) = 0{,}6$. Reconsiderando a matriz perca, pode calcular-se,

$$R(h, a_i) = L(\theta_1, a_i)h(\theta_1) + L(\theta_2, a_i)h(\theta_2),$$

isto é,

$$R(h, a_i) = L(\theta_1, a_i)(0{,}4) + L(\theta_2, a_i)(0{,}6), \quad i = 1, 2, 3.$$

	a_1	a_2	a_3	
θ_1	4	5	2	$h(\theta_1) = 0{,}4$
θ_2	4	0	5	$h(\theta_2) = 0{,}6$
$R(h, a_i)$	4	2	3,8	

Se o decisor não dispõe de barómetro — decisão sem dados — o critério Bayes recomenda a tomada da acção a_2; quer dizer, $a_h \equiv a_2$, é a acção pura Bayes contra a distribuição a priori, h, considerada. Se o decisor dispõe de barómetro, procede a cálculos formalmente análogos mas, agora, no quadro da matriz risco.

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	
θ_1	4	5	2	4,2	4,8	3,6	2,4	4,4	2,6	$h(\theta_1) = 0{,}4$
θ_2	4	0	5	0,4	3,6	4,9	4,1	4,5	0,5	$h(\theta_2) = 0{,}6$
$R(h, \delta_i)$	4	2	3,8	1,92	4,14	4,38	3,42	4,46	1,34	

A função de decisão pura Bayes contra h é manifestamente, $\delta_h \equiv \delta_9$, porquanto, $R(h, \delta_9) \leqq R(h, \delta_i)$, $i = 1, 2, \ldots, 9$ [coincide, acidentalmente, com a função de decisão minimax]; tal função de decisão foi obtida através da análise na forma normal.

Continuando a dispor de barómetro, suponha-se que o decisor observou que o mesmo indicava chuva, isto é, observou $X = x_1$. De posse desta informação pode dispensar a construção da matriz risco (e dispensar o arrolamento das funções de decisão) e optar por uma reavaliação das credibilidades que atribui aos estados por meio do Teorema de Bayes. Assim,

$$\begin{aligned} h(\theta_1 \mid x_1) &= f(x_1 \mid \theta_1)h(\theta_1)/[f(x_1 \mid \theta_1)h(\theta_1) + f(x_1 \mid \theta_2)h(\theta_2)] \\ &= 16/19; \end{aligned}$$

$$\begin{aligned} h(\theta_2 \mid x_1) &= f(x_1 \mid \theta_2)h(\theta_2)/[f(x_1 \mid \theta_1)h(\theta_1) + f(x_1 \mid \theta_2)h(\theta_2)] \\ &= 3/19. \end{aligned}$$

Se observou bom tempo — $X = x_2$ — tem-se,

$$\begin{aligned} h(\theta_1 \mid x_2) &= f(x_2 \mid \theta_1)h(\theta_1)/[f(x_2 \mid \theta_1)h(\theta_1) + f(x_2 \mid \theta_2)h(\theta_2)] \\ &= 4/31; \end{aligned}$$

$$\begin{aligned} h(\theta_2 \mid x_2) &= f(x_2 \mid \theta_2)h(\theta_2)/[f(x_2 \mid \theta_1)h(\theta_1) + f(x_2 \mid \theta_2)h(\theta_2)] \\ &= 27/31. \end{aligned}$$

Reconsiderando a matriz perca, calcula-se facilmente o risco a posteriori das várias acções:

— se $X = x_1$:

$$\begin{aligned} r_{x_1}(a_1) &= L(\theta_1, a_1)h(\theta_1 \mid x_1) + L(\theta_2, a_1)h(\theta_2 \mid x_1) \\ &= 4(16/19) + 4(3/19) = 4; \\ r_{x_1}(a_2) &= L(\theta_1, a_2)h(\theta_1 \mid x_1) + L(\theta_2, a_2)h(\theta_2 \mid x_1) \\ &= 5(16/19) + 0(3/19) = 80/19; \\ r_{x_1}(a_3) &= L(\theta_1, a_3)h(\theta_1 \mid x_1) + L(\theta_2, a_3)h(\theta_2 \mid x_1) \\ &= 2(16/19) + 5(3/19) = 47/19; \end{aligned}$$

— se $X = x_2$:

$$\begin{aligned} r_{x_2}(a_1) &= L(\theta_1, a_1)h(\theta_1 \mid x_2) + L(\theta_2, a_1)h(\theta_2 \mid x_2) \\ &= 4(4/31) + 4(27/31) = 4; \\ r_{x_2}(a_2) &= L(\theta_1, a_2)h(\theta_1 \mid x_2) + L(\theta_2, a_2)h(\theta_2 \mid x_2) \\ &= 5(4/31) + 0(27/31) = 20/31; \\ r_{x_2}(a_3) &= L(\theta_1, a_3)h(\theta_1 \mid x_2) + L(\theta_2, a_3)h(\theta_2 \mid x_2) \\ &= 2(4/31) + 5(27/31) = 143/31; \end{aligned}$$

Comparando os riscos a posteriori imediatamente se conclui que a solução Bayes obtida através da forma extensiva consiste na seguinte regra:

$$\begin{aligned} &\text{— se } X = x_1 \quad \text{escolher } a_3; \\ &\text{— se } X = x_2 \quad \text{escolher } a_2, \end{aligned}$$

que equivale precisamente à solução obtida pela forma normal, δ_9. Finalmente, notando a distribuição preditiva,

$$\begin{aligned} f(x_1) &= f(x_1 \mid \theta_1)h(\theta_1) + f(x_1 \mid \theta_2)h(\theta_2) = 0{,}38; \\ f(x_2) &= f(x_2 \mid \theta_1)h(\theta_1) + f(x_2 \mid \theta_2)h(\theta_2) = 0{,}62, \end{aligned}$$

tem-se,

$$\begin{aligned} r_{x_1}(a_3)f(x_1) + r_{x_2}(a_2)f(x_2) &= (47/19)0{,}38 + (20/31)0{,}62 \\ &= 1{,}34, \end{aligned}$$

que é o valor de $R(h, \delta_9)$ — veja-se (3.28) abaixo. □

Como imediatamente se reconhece — vejam-se, aliás, os exemplos seguintes — a análise na forma extensiva é de mais fácil aplicação do que a análise na forma

normal. Para os bayesianos, que recusam ou aplicam relutantemente a análise na forma normal, é a análise na forma extensiva que acolhem decisivamente. O motivo é claro: a análise na forma extensiva não contempla a integração sobre o espaço da amostra — releia-se (3.26) — porque faz intervir a distribuição a posteriori; consequentemente, concentra-se no risco a posteriori que consideram a única medida aceitável de precisão final. A clivagem não tem efeitos práticos pois são quase sempre idênticos os resultados obtidos com o risco a priori e com o risco a posteriori. Por outras palavras, os princípios Bayes condicional e risco Bayes conduzem praticamente aos mesmos resultados[10].

O risco a priori é a já conhecida função risco ou risco Bayes,

$$R(h,\delta) = \int_{\mathfrak{X}} r_x(\delta) f(x)\, dx; \tag{3.28}$$

quando a função perca não é limitada pode surgir uma dificuldade: a função de decisão obtida minimizando $r_x(\delta)$ para cada x pode ter risco a priori infinito. Para vincar a diferença diz-se então tratar-se de função de decisão Bayes formal.

Dado $X = x$, tem-se por (1.18), $h(\theta \,|\, x) \propto f(x \,|\, \theta) h(\theta)$; conclui-se, portanto, ser a minimização de (3.26) equivalente à minimização de,

$$\int_{\Theta} L[\theta, \delta(x)] f(x \,|\, \theta) h(\theta)\, d\theta, \tag{3.29}$$

procedimento que permite dispensar o cálculo de $h(\theta \,|\, x)$.

Exemplo 3.10 — Considere-se uma variável aleatória, X, com distribuição uniforme no intervalo $(0, |\,\theta\,|^{-1})$, com $1 \leqq |\,\theta\,| < \infty$. A função densidade é, portanto, $f(x \,|\, \theta) = |\,\theta\,|$, $0 < x < |\,\theta\,|^{-1}$.

Suponha-se que a distribuição a priori é,

$$h(\theta) = \begin{cases} (1/2)|\,\theta\,|^{-2} & \text{se } 1 \leqq |\,\theta\,| < \infty, \\ 0 & \text{outros valores de } \theta. \end{cases}$$

Observado $X = x$, tem-se a distribuição a posteriori,

$$h(\theta \,|\, x) = \begin{cases} \dfrac{(1/2)|\,\theta\,|^{-1}}{-\log x} & \text{se } x \leqq |\,\theta\,|^{-1} \leqq 1, \\ 0 & \text{outros valores de } \theta. \end{cases}$$

Se a função perca é $L(\theta, a) = (\theta - a)^2$, a acção Bayes correspondente, $[a = \delta_h(x)]$, deve minimizar,

$$r_x(a) = -(1/2) \int_{\Omega} \frac{(\theta - a)^2 |\,\theta\,|^{-1}}{\log x}\, d\theta,$$

[10] Fala-se, por isso, de forma unificada, em funções de decisão Bayes.

onde $\Omega = \{\theta : x \leqq |\theta|^{-1} \leqq 1\}$. O único valor de a que minimiza $r_x(a)$ é, $a = E\{\theta \mid x\}$, isto é, $a = 0$, dado que a distribuição a posteriori é simétrica em relação a $\theta = 0$. A função de decisão Bayes é, portanto, $\delta_h(x) \equiv 0$.

O risco a posteriori associado com $a = \delta_h(x) = 0$, é,

$$r_x(0) = -(1/2)\int_\Omega \frac{|\theta|}{\log x}\, d\theta = \frac{-(1/2)(1-x^2)}{x^2(\log x)}.$$

O risco a priori é, contudo,

$$R(h, \delta_h) = R(h, 0) = \int_0^1 r_x(0) f(x)\, dx = \infty,$$

pois $f(x) = -\log x$, $0 < x < 1$. A função de decisão $\delta_h(x) \equiv 0$, é Bayes formal. Zacks (1971). □

Exemplo 3.11 — No quadro do Ex. 1.14 procure-se a função de decisão Bayes contra a distribuição a priori (1.21) quando está em causa a estimação de θ — parâmetro da distribuição Binomial — com função perca quadrática,

$$L(\theta, a) = (\theta - a)^2.$$

Tome-se, $A = \Theta = [0, 1]$ e escreva-se (1.23) na forma,

$$h(\theta \mid x) = \frac{\theta^x (1-\theta)^{N-X}}{B(x+1, N-x+1)};$$

de (3.27) tem-se,

$$r_x(a) = \int_0^1 \frac{(\theta - a)^2 \theta^x (1-\theta)^{N-x}}{B(x+1, N-x+1)}\, d\theta.$$

Derivando $r_x(a)$ em relação ao respectivo argumento e igualando a zero, vem,

$$\int_0^1 (\theta - a)\theta^x (1-\theta)^{N-x} d\theta = 0,$$

isto é,

$$\int_0^1 \theta^{x+1}(1-\theta)^{N-x} d\theta = a \int_0^1 \theta^x (1-\theta)^{N-x} d\theta,$$

donde,

$$a_h = \delta_h(x) = \frac{B(x+2, N-x+1)}{B(x+1, N-x+1)}$$

$$= \frac{x+1}{N+2}.$$

A função de decisão, $\delta_h(x) = (x+1)/(N+2)$, é Bayes contra a distribuição uniforme no intervalo [0,1]; também se diz que $\delta_h(X) = (X+1)/(N+2)$ é um estimador bayesiano do parâmetro θ, no caso presente a probabilidade de um «sucesso» [compare-se com o estimador da máxima verosimilhança, $\hat{\delta}(X) = X/N$].

Para determinar o valor do risco a posteriori correspondente a δ_h, tem-se,

$$r_x(\delta_h) = [B(x+1, N-x+1)]^{-1} \int_0^1 \left[\theta - \frac{x+1}{N+2}\right]^2 \theta^x (1-\theta)^{N-x} d\theta,$$

ou seja, depois de alguns cálculos directos com a função Beta,

$$r_x(\delta_h) = \frac{(x+1)(N-x+1)}{(N+2)^2(N+3)}.$$

Por sua vez, o risco a priori, $R(h, \delta_h)$, obtém-se de (3.28), notando ser $f(x) = 1/(N+1)$, $x = 0, 1, \ldots, N$ [recorde-se (1.22)],

$$\begin{aligned} R(h, \delta_h) &= \sum_{x=0}^{N} r_x(\delta_h) f(x) = \frac{1}{N+1} \sum_{x=0}^{N} \frac{(x+1)(N-x+1)}{(N+2)^2(N+3)} \\ &= 1/6(N+2). \end{aligned}$$

[Silvey (1970)]. □

Exemplo 3.11 — *Continuação.* Altere-se a distribuição a priori considerando uma Beta de parâmetros α e β,

$$h(\theta) = [B(\alpha, \beta)]^{-1} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \quad 0 \leqq \theta \leqq 1.$$

Face à distribuição a posteriori [veja-se Ex. 1.28], a função de decisão Bayes passa a ser,

$$\delta_h(x) = E\{\theta \mid x\} = \frac{\alpha + x}{\alpha + \beta + N}.$$

É interessante comparar o estimador Bayes com o estimador usual, $\hat{\delta}(X) = X/N$. Antes de fazer qualquer observação (decisão sem dados), o estimador bayesiano é o valor esperado da distribuição a priori, $\alpha/(\alpha+\beta)$. Depois de observar, X, o estimador bayesiano, $\delta_h(X)$, assume a forma,

$$\frac{\alpha + X}{\alpha + \beta + N} = \left(\frac{\alpha + \beta}{\alpha + \beta + N}\right) \frac{\alpha}{\alpha + \beta} + \left(\frac{N}{\alpha + \beta + N}\right) \frac{X}{N},$$

que é precisamente a média ponderada de $\alpha/(\alpha+\beta)$ e de X/N: o estimador a priori é combinado com o estimador «clássico» (baseado na informação amostral) para dar o estimador a posteriori.

Quando $\alpha \to \infty$ e $\beta \to \infty$, mantendo-se fixo o rácio, β/α, $\delta_h(x) \to \alpha/(\alpha+\beta)$; quer dizer, a informação a priori é de tal forma preponderante que determina praticamente o estimador Bayes. O que se passa é que a distribuição, $B(\alpha, \beta)$, concentra toda a massa no ponto, $\alpha/(\alpha+\beta)$, sucedendo que a respectiva variância, $[\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)] \to 0$.

Se α e β se mantêm fixos e $N \to \infty$, $\delta_h(x) \to x/N$: quer dizer, a informação amostral é de tal forma preponderante que determina praticamente o estimador Bayes.

Um tratamento mais completo do exemplo pode ver-se em Lehmann (1983). □

Exemplo 3.12 — Considere-se o problema de estimação de θ em que se tem $A = \Theta = (0, \infty)$, função perca quadrática, $L(\theta, a) = K(\theta - a)^2$, distribuição a priori, $h(\theta) = \theta e^{-\theta}$, $\theta > 0$. A experiência consiste em observar uma variável aleatória $X \sim U(0, \theta)$. Tem-se,

$$\begin{aligned} f(x) &= \int_x^\infty (1/\theta)\theta e^{-\theta} d\theta = \int_x^\infty e^{-\theta} d\theta \\ &= e^{-x}, \quad x > 0, \end{aligned}$$

donde,

$$h(\theta \mid x) = f(x \mid \theta) h(\theta) / f(x) = e^{x-\theta}, \quad \theta > x.$$

Atendendo a (3.27),

$$r_x(a) = K \int_x^\infty (\theta - a)^2 e^{x-\theta} d\theta = K e^x \int_x^\infty (\theta - a)^2 e^{-\theta} d\theta,$$

expressão que é mínima para,

$$a_h = \delta_h(x) = \int_x^\infty \theta e^{-\theta} d\theta \Big/ \int_x^\infty e^{-\theta} d\theta = (x+1),$$

isto é, $\delta_h(x) = x + 1$ é a função de decisão Bayes contra h. Como imediatamente se reconhece, $\delta_h(x)$, é a média da distribuição a posteriori,

$$\delta_h(x) = x + 1 = \int_x^\infty \theta h(\theta \mid x)\, d\theta = \int_x^\infty \theta e^{x-\theta} d\theta,$$

como a expressão de $r_x(a)$ deixava antever [Ferguson (1967)]. □

Exemplo 3.13 — Um fabricante de fazendas mede a qualidade das peças que produz pelo número médio de defeitos por metro. Em relação a cada peça, partindo do princípio que o fabricante não pode inspeccionar totalmente cada uma, o número

médio de defeitos por metro representa o estado da natureza, θ; logo, $\Theta = (0, \infty)$. É prática corrente classificar as peças do modo seguinte:

$$\begin{aligned} 0 \leqq \theta < r &- 1.^{\text{a}} \text{ qualidade}, \\ r \leqq \theta < s &- 2.^{\text{a}} \text{ qualidade}, \\ \theta \geqq s &- 3.^{\text{a}} \text{ qualidade ou «refugo»}. \end{aligned}$$

Tem-se, $A = \{a_1, a_2, a_3\}$, onde a_i representa a venda de uma peça como sendo de i-ésima qualidade. A matriz perca tem a seguinte estrutura,

	a_1	a_2	a_3
$0 \leqq \theta < r$	-500	-300	-100
$r \leqq \theta < s$	-100	-300	-100
$\theta \geqq s$	300	100	-100

Esta estrutura resulta das condições de venda: (i) os preços são 500 para peças de 1.ª qualidade, 300 para peças de 2.ª qualidade e 100 para peças de 3.ª qualidade; (ii) o comprador, que em princípio acaba sempre por saber a real qualidade de cada peça, concorda em pagar o preço pedido se a qualidade for igual ou superior à declarada pelo fabricante no acto da venda; se a qualidade for inferior à especificada, o comprador paga o preço justo mas é indemnizado com uma quantia igual à diferença entre o preço facturado e o preço justo.

A matriz perca de oportunidade ou pesar facilita os cálculos e não altera, como facilmente se mostra, as soluções Bayes:

	a_1	a_2	a_3
$0 \leqq \theta < r$	0	200	400
$r \leqq \theta < s$	200	0	200
$\theta \geqq s$	400	200	0

O fabricante, com base na experiência pessoal, tem a seguinte distribuição a priori,

$$h(\theta) = e^{-\theta}, \ \theta > 0;$$

supõe-se que o número de defeitos por metro é uma variável aleatória com distribuição de Poisson com média θ. A experiência consiste em inspeccionar N metros de cada peça, registando o número de defeitos em cada metro. Assim, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, a respectiva amostra casual, tem função de probabilidade,

$$f(\mathbf{x} \,|\, \theta) = \frac{e^{-N\theta}\theta^{\Sigma x_i}}{\Pi x_i}.$$

A expressão dos riscos a posteriori é,

$$r_{\mathbf{x}}(a_1) = \left[200\int_r^s e^{-(N+1)\theta}\theta^z d\theta + 400\int_s^\infty e^{-(N+1)\theta}\theta^z d\theta\right]/\Delta(\mathbf{x}),$$

$$r_{\mathbf{x}}(a_2) = \left[200\int_0^r e^{-(N+1)\theta}\theta^z d\theta + 200\int_s^\infty e^{-(N+1)\theta}\theta^z d\theta\right]/\Delta(\mathbf{x}),$$

$$r_{\mathbf{x}}(a_3) = \left[400\int_0^r e^{-(N+1)\theta}\theta^z d\theta + 200\int_r^s e^{-(N+1)\theta}\theta^z d\theta\right]/\Delta(\mathbf{x}),$$

onde,

$$z = \sum_{i=1}^{N} x_i \quad \text{e} \quad \Delta(\mathbf{x}) = \int_0^\infty e^{-(N+1)\theta}\theta^z d\theta.$$

Seja, $H(\theta\,|\,\mathbf{x}) = H(\theta\,|\,z)$, a função de distribuição a posteriori de θ,

$$H(\theta\,|\,z) = \int_0^\theta e^{-(N+1)\xi}\xi^z d\xi \Big/ \int_0^\infty e^{-(N+1)\xi}\xi^z d\xi.$$

A função de decisão Bayes contra h pode então escrever-se,

(1) $\delta_h(\mathbf{x}) = a_1$, quando $r_{\mathbf{x}}(a_1) \leqq r_{\mathbf{x}}(a_2)$ e $r_{\mathbf{x}}(a_1) \leqq r_{\mathbf{x}}(a_3)$, isto é, quando z é tal que,

$$1 - H(r\,|\,z) \leqq H(r\,|\,z) \quad \text{e} \quad 1 - H(s\,|,z) \leqq H(r\,|\,z);$$

(2) $\delta_h(\mathbf{x}) = a_2$ quando $r_{\mathbf{x}}(a_2) \leqq r_{\mathbf{x}}(a_1)$ e $r_{\mathbf{x}}(a_2) \leqq r_{\mathbf{x}}(a_3)$, isto é, quando,

$$H(r\,|\,z) \leqq 1 - H(r\,|\,z) \quad \text{e} \quad 1 - H(s\,|\,z) \leqq H(s\,|\,z);$$

(3) $\delta_h(\mathbf{x}) = a_3$ quando $r_{\mathbf{x}}(a_3) \leqq r_{\mathbf{x}}(a_1)$ e $r_{\mathbf{x}}(a_3) \leqq r_{\mathbf{x}}(a_2)$, isto é quando,

$$H(r\,|\,z) \leqq 1 - H(s\,|\,z) \quad \text{e} \quad H(s\,|\,z) \leqq 1 - H(s\,|\,z).$$

Considerando que $r < s$ e que $H(\theta\,|\,z)$ é função crescente de θ, tem-se,

(1) z tal que $H(r\,|\,z) \geqq 1/2$ implica $\delta_h(\mathbf{x}) = a_1$;

(2) z tal que $H(r\,|\,z) \leqq 1/2$ e $H(s\,|\,z) \geqq 1/2$ implica $\delta_h(\mathbf{x}) = a_2$;

(3) z tal que $H(s\,|\,z) \leqq 1/2$ implica $\delta_h(\mathbf{x}) = a_3$.

Sejam z_1 e z_2 valores de z para os quais se verifica,

$$H(r\,|\,z_1) = 1/2 \quad \text{e} \quad H(s\,|\,z_2) = 1/2;$$

é fácil de ver que $z_1 < z_2$ e que,

(1) $z < z_1$ ou seja, $\Sigma x_i < z_1$, implica $\delta_h(\mathbf{x}) = a_1$;

(2) $z_1 < z < z_2$ ou seja, $z_1 < \Sigma x_i < z_2$, implica $\delta_h(\mathbf{x}) = a_2$;

(3) $z > z_2$ ou seja, $\Sigma x_i > z_2$, implica $\delta_h(\mathbf{x}) = a_3$.

Se $z = z_1$ as acções a_1 e a_2 são indiferentes e preferíveis a a_3; se $z = z_2$ as acções a_2 e a_3 são indiferentes e preferíveis a a_1 [Blackwell e Girshick (1954)]. □

Exemplo 3.14 — Determinados indivíduos podem ser provenientes de um universo U_1 ou de um universo U_2. Deseja-se proceder à classificação dos indivíduos com base na observação de k características: $(X_1, X_2, \ldots, X_k)$.

A matriz perca é a seguinte,

Origem	Classificação em U_1 a_1	em U_2 a_2
$U_1 \equiv \theta_1$	0	L
$U_2 \equiv \theta_2$	L'	0

Sejam $f(\mathbf{x}\,|\,\theta_1)$ e $f(\mathbf{x}\,|\,\theta_2)$ as funções de densidade das características para cada uma das origens e admita-se que os dois universos estão misturados nas proporções ζ e $1-\zeta$. A distribuição a priori, logicamente,

$$\mathbf{h} = (\zeta, 1-\zeta),$$

tem neste caso interpretação frequencista.

Conclui-se imediatamente que sendo $\delta_h(\mathbf{x})$ função de decisão Bayes contra $\mathbf{h}$,
(1) $\delta_h(\mathbf{x}) = a_1$, quando $r_{\mathbf{x}}(a_1) \leqq r_{\mathbf{x}}(a_2)$, isto é, quando,

$$L' h(\theta_2\,|\,\mathbf{x}) \leqq L h(\theta_1\,|\,\mathbf{x});$$

como,

$$h(\theta_1\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,\theta_1)\zeta}{f(\mathbf{x}\,|\,\theta_1)\zeta + f(\mathbf{x}\,|\,\theta_2)(1-\zeta)},$$

$$h(\theta_2\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,\theta_2)(1-\zeta)}{f(\mathbf{x}\,|\,\theta_1)\zeta + f(\mathbf{x}\,|\,\theta_2)(1-\zeta)},$$

tem-se a desigualdade equivalente,

$$\frac{f(\mathbf{x}\,|\,\theta_1)}{f(\mathbf{x}\,|\,\theta_2)} \geqq \frac{L'(1-\zeta)}{L\zeta}. \tag{3.30}$$

(2) $\delta_h(\mathbf{x}) = a_2$ quando $r_{\mathbf{x}}(a_2) \leqq r_{\mathbf{x}}(a_1)$, isto é, quando,

$$\frac{f(\mathbf{x}\,|\,\theta_1)}{f(\mathbf{x}\,|\,\theta_2)} \leqq \frac{L'(1-\zeta)}{L\,\zeta}. \tag{3.31}$$

Evidentemente, se se verifica a igualdade em (3.30) e (3.31) — acontecimento que tem probabilidade zero — as acções são indiferentes.

Situação de interesse é aquela em que as k características têm distribuição Normal k-dimensional,

$$f(\mathbf{x}\,|\,\boldsymbol{\theta}_i) = \frac{|\,\sigma^{rs}\,|^{1/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2}\sum_{r=1}^{k}\sum_{s=1}^{k}\sigma^{rs}(x_r - \theta_{ir})(x_s - \theta_{is})\right\}, \quad i = 1, 2,$$

com matriz de covariâncias, $[\sigma_{rs}]$, idêntica e médias, $\boldsymbol{\theta}_1 = (\theta_{11}, \theta_{12}, \ldots, \theta_{1k})$, $\boldsymbol{\theta}_2 = (\theta_{21}, \theta_{22}, \ldots, \theta_{2k})$, diferentes. Note-se que na expressão acima $|\,\sigma^{rs}\,|$ é o determinante da matriz $[\sigma^{rs}]$ inversa de $[\sigma_{rs}]$.

De (3.30) sai, depois de aplicar logaritmos,

(1) $\delta_h(\mathbf{x}) = a_1$ se $\mathbf{x}$ é tal que,

$$\Lambda(\mathbf{x}) = \sum_{r=1}^{k}\lambda_r x_r \leqq \frac{1}{2}[\Lambda(\boldsymbol{\theta}_1) + \Lambda(\boldsymbol{\theta}_2)] + \log\frac{L\,\zeta}{L'(1-\zeta)}, \tag{3.32}$$

onde,

$$\lambda_r = \sum_{s=1}^{k}\sigma^{rs}(\theta_{2s} - \theta_{1s});$$

(2) $\delta_h(\mathbf{x}) = a_2$ se $\mathbf{x}$ conduz à desigualdade inversa de (3.32) [Blackwell e Girshick (1954)]. □

Exista ou não função de decisão Bayes, o decisor pode estar interessado em utilizar uma função de decisão Bayes-ξ contra h — ξ número real arbitrariamente pequeno e positivo — isto é, pode usar $\delta' \in D^*$ verificando,

$$R(h, \delta') \leqq \inf_{\delta^*} R(h, \delta^*) + \xi. \tag{3.33}$$

Uma função de decisão é quase-Bayes ou Bayes por extensão se for Bayes-ξ para todo o $\xi > 0$, isto é, se para todo $\xi > 0$ existir $h_\xi \in \mathcal{H}$ contra o qual é Bayes-ξ. No caso Θ finito, a Fig. 3.16 indica o conjunto de pontos de S_D^* que correspondem a funções de decisão Bayes-ξ contra um dado h.

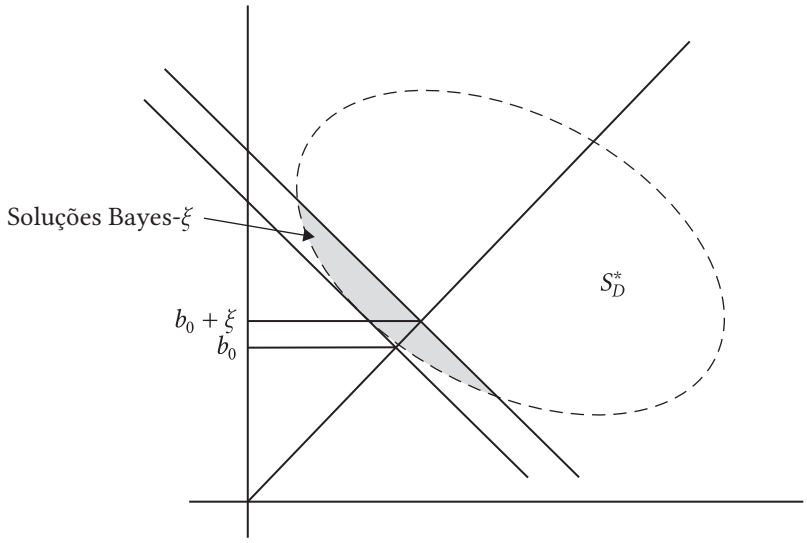


Fig. 3.16

Exemplo 3.15 — A partir de uma amostra casual, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i com distribuição $N(\theta, 1)$, pretende estimar-se θ. Tem-se, $A = \Theta = (-\infty, +\infty)$,

$$f(\mathbf{x} \mid \theta) = (2\pi)^{-N/2} \exp\{-(1/2)\Sigma(x_i - \theta)^2\}.$$

Como é sabido, a função de decisão obtida pelo clássico método da máxima verosimilhança é $\hat{\delta}(\mathbf{x}) = \overline{x}$.

Suponha-se $h(\theta) = h_\tau(\theta) \equiv N(0, \tau^2)$; os cálculos feitos nos Exs. 1.16 e 1.17 permitem obter,

$$h_\tau(\theta \mid \mathbf{x}) = \left(\frac{1 + N\tau^2)}{2\pi\tau^2}\right)^{1/2} \exp\left\{-\frac{1 + N\tau^2}{2\tau^2}\left[\theta - \frac{N\tau^2\overline{x}}{1 + N\tau^2}\right]^2\right\}.$$

Com função perca quadrática, a função de decisão Bayes contra $h_\tau(\theta)$ é,

$$\begin{aligned}\delta_\tau(\mathbf{x}) = E\{\theta \mid \mathbf{x}\} &= \int_{-\infty}^{+\infty} \theta h_\tau(\theta \mid x)\, d\theta \\ &= \frac{N\theta^2\overline{x}}{1 + N\tau^2},\end{aligned}$$

vindo para valor do risco a priori,

$$R(h_\tau, \delta_\tau) = E\left\{\left[\theta - \frac{N\tau^2\overline{x}}{1 + N\tau^2}\right]^2\right\} = \frac{\tau^2}{1 + N\tau^2}.$$

Por outro lado,

$$R(h_\tau, \hat{\delta}) = \int_{-\infty}^{+\infty} \left\{ \int_{\mathfrak{X}} (\theta - \overline{x})^2 f(\mathbf{x} \,|\, \theta)\, d\mathbf{x} \right\} h_\tau(\theta)\, d\theta = \frac{1}{N};$$

assim, qualquer que seja $\xi > 0$,

$$R(h_\tau, \hat{\delta}) \leqq \inf_{\delta^*} R(h_\tau, \delta^*) + \xi = \frac{\tau^2}{1 + N\tau^2} + \xi,$$

desde que se tome τ^2 suficientemente grande. Logo, para todo o $\xi > 0$ existe uma distribuição a priori, $h_\tau(\theta)$, contra a qual $\hat{\delta}$ é Bayes-ξ: basta tomar $\tau^2 \geqq (1 - \xi N)/\xi N^2$. Conclui-se, portanto, que o estimador da máxima verosimilhança é função de decisão quase-Bayes [Ferguson (1967)]. □

Na secção 1.5 foi referido o emprego que por vezes se faz de distribuições a priori impróprias para caracterizar situações iniciais de completa ignorância. As funções de decisão Bayes contra distribuições a priori impróprias dizem-se Bayes generalizadas; são funções de decisão que por vezes interessa estudar.

Exemplo 3.15 — *Continuação.* Considere-se a estimação de θ com distribuição a priori não informativa (imprópria), $h(\theta)\, d\theta \propto d\theta$ $(-\infty < \theta < +\infty)$. Tem-se,

$$\begin{aligned} h(\theta \,|\, \mathbf{x}) &\propto \exp\{-(1/2)\Sigma(x_i - \theta)^2\} \\ &\propto \exp\{-(N/2)(\theta - \overline{x})^2\}, \end{aligned}$$

donde, $E\{\theta \,|\, \mathbf{x}\} = \overline{x}$. Como a função perca é quadrática, $\delta(\mathbf{x}) = \overline{x}$, é função de decisão Bayes generalizada. Como era de esperar chega-se ao mesmo resultado tomando, $h_\tau(\theta) \equiv N(0, \tau^2)$, e fazendo $\tau^2 \to \infty$ (precisão inicial $\to 0$) no estimador Bayes contra $h_\tau(\theta)$,

$$\delta_\tau(\mathbf{x}) = \frac{N\tau^2 \overline{x}}{1 + N\tau^2}.$$

□

3.5 Decisão Bayes para funções perca particulares

Há certas formas de função perca que surgem amiúde em problemas de decisão, justificando-se o seu tratamento em termos gerais — sem especificar a distribuição a priori e a natureza da experiência — a partir da distribuição a posteriori, $h(\theta \,|\, x)$.

[1] *Função perca quadrática.* Com $A = (-\infty, +\infty)$, $\Theta = (-\infty, +\infty)$ e função perca quadrática,

$$L(\theta, a) = K(\theta - a)^2, \tag{3.34}$$

o risco a posteriori escreve-se,

$$\begin{aligned} r_x(a) &= K \int_{-\infty}^{+\infty} (\theta - a)^2 h(\theta \mid x)\, d\theta \\ &= K \cdot E\{(\theta - a)^2 \mid x\} \\ &= K[E\{\theta^2 \mid x\} - 2aE\{\theta \mid x\} + a^2]. \end{aligned}$$

Derivando em ordem a a, sai,

$$dr_x(a)/da = K[2a - 2E\{\theta \mid x\}];$$

resolvendo a equação, $dr_x(a)/da = 0$, tem-se finalmente,

$$\delta_h(x) = E\{\theta \mid x\}. \tag{3.35}$$

A função,

$$L(\theta, a) = W(\theta)(\theta - a)^2, \tag{3.36}$$

é mais geral do que (3.34). Nas linhas do raciocínio anterior tem-se,

$$\delta_h(x) = \frac{E\{\theta W(\theta) \mid x\}}{E\{W(\theta) \mid x\}}. \tag{3.37}$$

[2] *Função perca erro absoluto.* Com $A = (-\infty, +\infty)$, $\Theta = (-\infty, +\infty)$ e função perca erro absoluto,

$$L(\theta, a) = K \mid \theta - a \mid, \tag{3.38}$$

seja ν a mediana de $h(\theta \mid x)$ e considere-se $a > \nu$. Tem-se, com $K = 1$ sem perder generalidade,

$$L(\theta, \nu) - L(\theta, a) = \begin{cases} \nu - a & \text{se } \theta \leqq \nu, \\ 2\theta - (\nu + a) & \text{se } \nu < \theta < a, \\ a - \nu & \text{se } \theta \geqq a, \end{cases}$$

donde se conclui,

$$L(\theta, \nu) - L(\theta, a) \leqq (\nu - a)\, I_{(-\infty, \nu]}(\theta) + (a - \nu)\, I_{(\nu, \infty)}(\theta),$$

onde, $I_{(u,v)}(\theta)$ é a função indicatriz do intervalo (u, v). Por ser,

$$\begin{aligned} P(\theta \leqq \nu \mid x) &\geqq 1/2, \\ P(\theta > \nu \mid x) &\leqq 1/2, \end{aligned}$$

conclui-se que,

$$\begin{aligned} E\{[L(\theta,v)-L(\theta,a)]\,|\,x\} &\leqq (v-a)\,P(\theta \leqq v\,|\,x)+(a-v)\,P(\theta>v\,|\,x) \\ &\leqq (v-a)\,(1/2)+(a-v)\,(1/2)=0, \end{aligned}$$

isto é, que $r_x(v) \leqq r_x(a)$. Um raciocínio semelhante feito para $a < v$, permite concluir ser,

$$\delta_h(x) = \text{mediana de } h(\theta\,|\,x). \tag{3.39}$$

[3] *Função perca linear por troços.* Com $A=(-\infty,+\infty)$, $\Theta=(-\infty,+\infty)$ e função perca,

$$L(\theta,a)=\begin{cases} K_0(a-\theta) & \text{se } \theta \leqq a, K_0>0, \\ K_1(\theta-a) & \text{se } \theta>a, K_1>0, \end{cases}$$

o risco a posteriori assume a forma,

$$r_x(a)=K_0\int_{-\infty}^{a}(a-\theta)h(\theta\,|\,x)\,d\theta+K_1\int_{a}^{+\infty}(\theta-a)h(\theta\,|\,x)\,d\theta. \tag{3.40}$$

Derivando em ordem a a, obtém-se,

$$dr_x(a)/da = K_0\int_{-\infty}^{a}h(\theta\,|\,x)\,d\theta-K_1\int_{a}^{+\infty}h(\theta\,|\,x)\,d\theta,$$

ou seja,

$$dr_x(a)/da = K_0H(a\,|\,x)-K_1[1-H(a\,|\,x)],$$

fazendo intervir a função de distribuição a posteriori, $H(\theta\,|\,x)$. Resolvida a equação, $dr_x(a)/da=0$, tem-se,

$$\delta_h(x)=a_h \text{ tal que } H(a_h\,|\,x)=\frac{K_1}{K_0+K_1}. \tag{3.41}$$

[4] *Função perca linear por troços: caso de duas acções.*
Se $A=\{a_1,a_2\}$, $\Theta=(-\infty,+\infty)$ e a função perca é do tipo,

$$\begin{aligned} L(\theta,a_1)&=\begin{cases} 0 & \text{para } \theta \leqq \theta_0, \\ K_1(\theta-\theta_0) & \text{para } \theta>\theta_0, K_1>0, \end{cases} \\ L(\theta,a_2)&=\begin{cases} K_2(\theta_0-\theta) & \text{para } \theta \leqq \theta_0, K_2>0, \\ 0 & \text{para } \theta>\theta_0, \end{cases} \end{aligned} \tag{3.42}$$

o risco a posteriori vem dado pela expressão,

$$r_x(a_1)=K_1\int_{\theta_0}^{\infty}(\theta-\theta_0)h(\theta\,|\,x)\,d\theta; \quad r_x(a_2)=K_2\int_{-\infty}^{\theta_0}(\theta_0-\theta)h(\theta\,|\,x)\,d\theta. \tag{3.43}$$

Se $r_x(a_1) < r_x(a_2)$, $\delta_h(x) = a_1$; se $r_x(a_2) > r_x(a_1)$, $\delta_h(x) = a_2$; verificando-se a igualdade, qualquer das acções é Bayes, quer dizer, para esse valor de x é indiferente fazer $\delta_h(x) = a_1$ ou $= a_2$.

Sendo $K_1 = K_2 = K$, a análise pode simplificar-se significativamente. Nessa hipótese, $r_x(a_1) < r_x(a_2)$ implica,

$$\int_{\theta_0}^{+\infty} \theta h(\theta \mid x)\, d\theta - \theta_0[1 - H(\theta_0 \mid x)] < \theta_0 H(\theta_0 \mid x) - \int_{-\infty}^{\theta_0} \theta h(\theta \mid x)\, d\theta,$$

isto é,

$$r_x(a_1) < r_x(a_2) \Leftrightarrow E\{\theta \mid x\} < \theta_0. \tag{3.44}$$

Como é óbvio,

$$r_x(a_2) < r_x(a_1) \Leftrightarrow E\{\theta \mid x\} > \theta_0. \tag{3.45}$$

Se $E\{\theta \mid x\} = \theta_0$, as acções são indiferentes.

[5] *Função perca* «0–K_i». Na expressão (3.42) substitua-se $K_1(\theta - \theta_0)$ pela constante K_1 e $K_2(\theta_0 - \theta)$ pela constante K_2. De (3.43) obtém-se imediatamente,

$$r_x(a_1) \lesseqgtr r_x(a_2) \Leftrightarrow K_1 P(\theta > \theta_0 \mid x) \lesseqgtr K_2 P(\theta \leqq \theta_0 \mid x). \tag{3.46}$$

[6] *Função perca* «0–1». Caso ainda mais particular, com $K_1 = K_2 = 1$:

$$r_x(a_1) \lesseqgtr r_x(a_2) \Leftrightarrow P(\theta > \theta_0 \mid x) \lesseqgtr P(\theta \leqq \theta_0 \mid x). \tag{3.47}$$

3.6 Robustez das soluções Bayes

A importância dos modelos na investigação científica e, em particular, na investigação estatística (inferencial ou decisional), ficou bem vincada na secção 1.2.

O modelo utilizado na decisão estatística tem várias componentes das quais se destacam a distribuição a priori, $h(\theta)$, a família de distribuições da variável ou vector associado à experiência, $\mathcal{F} = \{f(x \mid \theta) : \theta \in \Theta\}$, e a função perca, $L(\theta, a)$. Recordando o que diz Box (1979) — «*All models are wrong but some are useful*». — convém ter bem presente que os pressupostos ou hipóteses introduzidos na modelação daquelas componentes não são exactamente verdadeiros e resultam, quase sempre, de um esforço de racionalização ou simplificação para viabilizar o tratamento matemático. Semelhante esforço, aliás indispensável nos mais variados ramos da matemática aplicada, encontra justificação [Huber (1981)] num vago princípio de estabilidade ou continuidade: um pequeno erro no modelo matemático deve causar um pequeno erro nas conclusões finais.

Os exemplos de não verificação do princípio de estabilidade são deveras abundantes e levaram, no domínio da estatística, à introdução do conceito de robustez:

um procedimento considera-se robusto quando é insensível a pequenas variações nos pressupostos ou hipóteses em que se baseia.

A robustez em relação à família $\mathcal{F}$ [em relação a $f(x\,|\,\theta)$] não é problema específico da decisão estatística pois constitui preocupação comum a qualquer sistema de inferência estatística. O respectivo estudo centra-se na especificação de classes de densidades ou de distribuições e na pesquisa de procedimentos que se comportem bem para todas as densidades da classe. Semelhante estudo não pode aqui fazer-se mas o interesse do assunto justifica a apresentação de um exemplo.

Exemplo 3.16 — Considere-se a estimação do parâmetro de localização das seguintes famílias de densidades: Normal, Mistura da Normal e da Uniforme, «Student»-*t* [4] (4 graus de liberdade), Laplace e Cauchy. As funções de decisão consideradas foram: $\delta_1(\mathbf{X}) = \overline{X}$ e $\delta_2(\mathbf{X}) = M_e$ (mediana da amostra); a comparação é feita através das respectivas variâncias assintóticas e do respectivo rácio, $V\{\overline{X}\}/V\{M_e\}$, que mede a eficiência da mediana em relação à média.

Famílias	Variâncias $V\{\overline{X}\}$	Assintóticas $V\{M_e\}$	Eficiência $V\{\overline{X}\}/V\{M_e\}$
Normal	σ^2/N	$\pi\sigma^2/2N$	0,6366
$0{,}9N + 0{,}1U(-3{,}3)^*$	$1{,}2\ \sigma^2/N$	$1{,}77\ \sigma^2/N$	0,6776
$0{,}5N + 0{,}5U(-3{,}3)^*$	$2\ \sigma^2/N$	$3{,}1258\ \sigma^2/N$	0,6398
«Student»-*t* [4]	$2\ \sigma^2/N$	$16\ \sigma^2/9N$	1,125
Laplace	$2\ \sigma^2/N$	σ^2/N	2,0
Cauchy	—	$\sigma^2\pi^2/4$	∞

* A família mistura tem função de densidade,

$$\alpha N + (1-\alpha)U(-3{,}3) \equiv f(x) = \alpha\left(\sqrt{2\pi}\right)^{-1}\exp\{-x^2/2\} + \\ + \frac{1-\alpha}{6}I_{(-3,3)}(x), \quad -\infty < x < +\infty.$$

Os valores do quadro mostram que $\overline{X}$ é um bom estimador para a família Normal ou para a família mistura Normal/Uniforme [$\alpha = 0{,}9$ e 0,5], mas deixa de ser superior à mediana quando se consideram outras famílias, tornando-se completamente ineficiente para a família Cauchy. A justificação do fenómeno encontra-se na extrema sensibilidade da média a flutuações nos valores extremos da amostra, donde resulta um comportamento muito desfavorável quando se aplica a uma família com abas que associam densidades relativamente elevadas a valores da variável muito desviados do ponto central, como é o caso da Cauchy. Tem pois de concluir-se que a média não é um estimador robusto [Zacks (1981)] . □

A robustez em relação à função perca já é peculiar dos problemas decisionais mas não pode ser também objecto de estudo desenvolvido dada a escassez de resultados conhecidos. Seguindo Berger (1980) de perto — como de resto vai fazer-se em toda a presente secção — podem avançar-se os seguintes comentários.

As funções de decisão são em geral robustas em relação às formas alternativas como a função perca penaliza os grandes erros; por exemplo, no caso da estimação, se a função perca é da forma, $L(\theta - a)$, uma função de decisão é usualmente robusta em relação à forma de L para valores grandes de $(\theta - a)$ porque estes aparecem normalmente associados com densidades de probabilidade ou probabilidades evanescentes. Sendo assim, há que ter o maior cuidado na definição ou determinação de L para valores pequenos de $(\theta - a)$. Por outro lado, a presença de um factor de ponderação, $W(\theta)$, na função perca, por exemplo, $L(\theta, a) = W(\theta)(\theta - a)^2$, é fonte de dificuldades; no entanto, como pode ser absorvido na distribuição a priori — repare-se em (3.29) — o problema pode ser transferido para as linhas que seguem.

Surge, assim, em plano de destaque o estudo da robustez em relação à distribuição a priori o qual envolve, sobretudo, a análise da sensibilidade das funções de decisão Bayes contra h quando h se faz percorrer uma classe $\mathcal{H}_0 \subset \mathcal{H}$. Dadas as contingências e dificuldades com que se depara na descrição da informação a priori por meio de uma adequada distribuição, a robustez das funções de decisão Bayes parece conceito extremamente relevante, embora, infelizmente, se trate de campo muito pouco explorado.

Em tudo o que segue, $H(\theta)$, é a função de distribuição correspondente à densidade, $h(\theta)$: $H(\theta) = \int_{-\infty}^{\theta} h(\xi)\, d\xi$ ou $= \sum_{\xi \leqq \theta} h(\xi)$.

O processo natural para investigar a robustez das soluções Bayes em relação à distribuição a priori consiste, portanto, em introduzir uma classe de distribuições a priori plausíveis e em inquirir como a escolha de h na classe afecta as referidas soluções. Nesse contexto, Berger (1980), refere as seguintes quatro classes:

$\mathcal{H}_1 = \{h : |H(\theta) - H_0(\theta)| \leqq \eta$ para todo o $\theta \in \Theta\}$;

$\mathcal{H}_2 = \{h : |\zeta(\alpha_i) - \zeta_0(\alpha_i)| \leqq, i = 1, 2, \ldots, k$, onde $\zeta(\alpha_i)$ e $\zeta_0(\alpha_i)$ são os quantis de ordem-α_i de h e h_0 respectivamente$\}$;

$\mathcal{H}_3 = \{h : |\nu_i - \nu_i^0| \leqq \eta_i, i = 1, 2, \ldots, K$, onde ν_i e ν_i^0 são os momentos de ordem i em relação à origem de h e h_0 respectivamente$\}$;

$\mathcal{H}_4 = \{h$: h é uma função de densidade de uma dada forma funcional, $h(\theta \,|\, \gamma_1, \ldots, \gamma_\kappa)$, onde $\gamma_i \in \Gamma_i, i = 1, 2, \ldots, \kappa\}$.

Em todas as classes a ideia predominante é a de que h_0 é a distribuição a priori proposta pelo decisor sendo a classe constituída por distribuições próximas de h_0, variando nos quatro casos o critério de proximidade.

As classes $\mathcal{H}_1$ e $\mathcal{H}_2$ são as mais atractivas por envolverem probabilidades ou parâmetros de mais fácil determinação subjectiva; no entanto, como o seu trata-

mento é difícil, recorre-se às classes $\mathcal{H}_3$ e $\mathcal{H}_4$, menos sugestivas mas de mais fácil manuseamento.

Robustez a posteriori. Uma primeira via para investigar a robustez, aliás a preferida numa óptica bayesiana, atende ao risco a posteriori, $r_x(a)$. Suponha-se que a proposta[11] distribuição a priori é h_0 e que a solução Bayes contra h_0 é a_0: estude-se a diferença,

$$\left| \int_\Theta L(\theta, a_0) h_0(\theta \mid x)\, d\theta - \inf_a \int_\Theta L(\theta, a) h(\theta \mid x)\, d\theta \right|,$$

para todo o $h \in \mathcal{H}_0$; se a diferença for pequena ($\leqq \xi$) a_0 diz-se robusta a posteriori (robusta a posteriori-ξ) em relação a $\mathcal{H}_0$.

Quando numa dada situação o princípio da medição precisa é verificado (veja-se pág. 77), a informação a priori é submersa pela informação amostral; consequentemente, a distribuição a priori é muito achatada em contraste com o pontiagudo da função de verosimilhança e por mais variações que se façam em torno daquela — dentro de limites razoáveis — mantém-se a preponderância desta e a distribuição a posteriori pouco ou nada varia. Claro que em tal situação — aliás de pouco interesse na perspectiva decisional — as acções Bayes são robustas a posteriori em relação à classe em que se incluem essas variações. Casos mais interessantes são os ilustrados nos exemplos seguintes:

Exemplo 3.17 — Observa-se $X \sim N(\theta, 1)$ e pretende ensaiar-se a hipótese H_0: $\theta \leqq 0$ contra H_1: $\theta > 0$ com função perca «0–1». Tem-se $A = \{a_0, a_1\}$, onde a_0 significa aceitação de H_0 e a_1 aceitação de H_1 [H_0 e H_1 designam hipóteses e nada têm a ver com a função de distribuição a priori]. Supõe-se que a distribuição a priori é $h_0 \equiv N(1, 4)$ mas admite-se que a verdadeira distribuição a priori possa ser um elemento qualquer da classe $\mathcal{H}_1$ [veja-se página anterior]. Considerando (3.47) conclui-se que para qualquer $h_0 \in \mathcal{H}_1$, a_0 é a acção Bayes contra h_0 se,

$$P(\theta \leqq 0 \mid x) = \int_{-\infty}^{0} h_0(\theta \mid x)\, d\theta > P(\theta > 0 \mid x) = \int_0^{\infty} h_0(\theta \mid x)\, d\theta,$$

e que a_1 é a acção Bayes contra h_0 se a desigualdade sai invertida.

Se, ao realizar a observação, sai $x = -10$, a informação amostral pesa decisivamente a favor de H_0 e a tal ponto que para todo o h razoavelmente próximo de h_0, ter-se-á,

$$P(\theta \leqq 0 \mid x = -10) > P(\theta > 0 \mid x = -10).$$

Assim, observado $x = -10$, a_0 é a acção Bayes contra qualquer $h \in \mathcal{H}_1$. A acção Bayes é, portanto, dado $x = -10$, extremamente robusta a posteriori; infelizmente,

[11] Presumida como «verdadeira».

tivesse saído x próximo de zero a desigualdade acima seria de difícil verificação para um qualquer $h \in \mathcal{H}_1$ e nada se concluia.

A lição a colher é a seguinte: em geral, a robustez de uma acção Bayes depende drasticamente da amostra ou valor observado [Berger (1980)]. □

Exemplo 3.18 — Com $X \sim N(\theta, 1)$, $\Theta = (-\infty, +\infty)$, e função perca quadrática pretende-se um estimador bayesiano para θ. Tomando para distribuição a priori, $h_N \equiv N(0, 2{,}19)$, obtém-se o estimador [veja-se (1.28)],

$$\delta_N(x) = x - \frac{x}{3{,}19};$$

se a distribuição a priori é a Cauchy, $h_C \equiv C(0{,}1)$ $[h(\theta) = 1/\pi(1+\theta^2)]$, a função de decisão Bayes é, conforme Berger (1980),

$$\delta_C(x) \cong x - \frac{2x}{2+x^2}.$$

Repare-se que as distribuições $N(0, 2{,}19)$ e $C(0, 1)$ possuem ambas mediana igual a zero, 1.º quartil igual a -1 e 3.º quartil igual a 1. Com um pouco de boa vontade pode considerar-se que as duas distribuições formam uma classe do tipo $\mathcal{H}_2$ [com $\alpha_1 = 0{,}25$, $\alpha_2 = 0{,}5$ e $\alpha_3 = 0{,}75$] ou melhor, que cobrem uma razoável parcela da variação em $\mathcal{H}_2$, premissa justificada pelo desigual comportamento das abas.

Suponha-se $x = 10$; vem $\delta_N(10) = 6{,}87$ e $\delta_C(10) = 9{,}80$, valores substancialmente diferentes que mostram a reduzida robustez a posteriori, no caso presente, das funções de decisão Bayes em relação a $\mathcal{H}_2$. Para x próximo de zero a diferença atenua-se grandemente, não sendo então descabido afirmar que os procedimentos são robustos.

A lição adicional a colher é a seguinte: dada a dificuldade em operar com classes de densidades, resta o expediente de seleccionar alguns elementos típicos da classe e confrontar as soluções Bayes contra esses elementos [Berger (1980)]. □

Robustez em termos de risco. Se uma acção Bayes é robusta a posteriori, a análise pode considerar-se concluída; caso contrário, ou havendo dúvida, a robustez pode investigar-se por outras vias.

O caminho mais simples, e o melhor em certo sentido, consiste em comparar as funções risco, $R(\theta, \delta)$, correspondentes às funções de decisão Bayes contra as distribuições (ou algumas distribuições) de uma dada classe $\mathcal{H}_0$.

Exemplo 3.18 — *Continuação.* Dadas as distribuições a priori, h_N e h_C, e as respectivas funções de decisão Bayes, δ_N e δ_C, as correspondentes funções risco apresentam-se na Fig. 3.17.

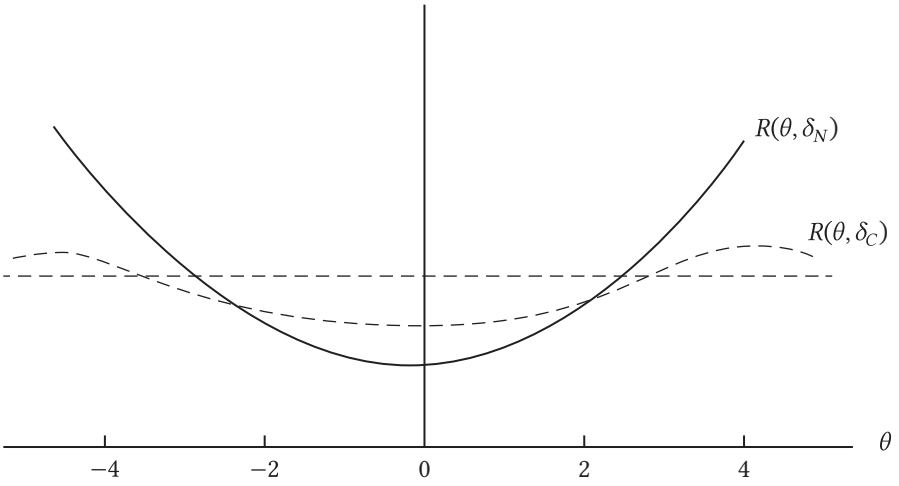


Fig. 3.17

Ambos os procedimentos são bons quando θ está próximo de zero (e as distribuições a priori parecem correctas). No entanto, quando $|\theta|$ aumenta sucessivamente, δ_N é muito pior do que δ_C; havendo incerteza quanto à natureza das abas da verdadeira distribuição a priori, δ_C é claramente preferível; havendo a certeza de que θ está na proximidade de zero, δ_N pode ser preferível [Berger (1980)]. □

A análise feita no quadro do exemplo anterior é pouco rigorosa. Para efectuar uma análise formal em termos de risco tem de considerar-se a função $R(h, \delta^*)$.

Considere-se uma classe $\mathcal{H}_0$ de distribuições a priori; a expressão,

$$R_{\mathcal{H}_0}(\delta^*) = \sup_{h \in \mathcal{H}_0} R(h, \delta^*),$$

representa a perca esperada máxima que o decisor pode sofrer quando utiliza $\delta^* \in D^*$ e h percorre $\mathcal{H}_0$. Na linha de ideias já anteriormente afloradas, uma função de decisão, $\delta^*_{\mathcal{H}_0}$, tal que,

$$\sup_{h \in \mathcal{H}_0} R(h, \delta^*_{\mathcal{H}_0}) = \inf_{\delta^* \in D^*} \sup_{h \in \mathcal{H}_0} R(h, \delta^*) = R_{\mathcal{H}_0},$$

diz-se minimax-$\mathcal{H}_0$. Evidentemente, $R_{\mathcal{H}_0}(\delta^*) \geqq R_{\mathcal{H}_0}$ [= risco minimax-$\mathcal{H}_0$]; daqui a ideia de apreciar a robustez em termos de risco de δ^* através de $R_{\mathcal{H}_0}(\delta^*)$. Se $R_{\mathcal{H}_0}(\delta^*)$ está próximo de $R_{\mathcal{H}_0}$, pode concluir-se que o decisor não tem vantagem em optar por uma função de decisão diferente de δ^*; com efeito, ao abandonar δ^* pouco pode reduzir a perca esperada em que incorre quando dentro de $\mathcal{H}_0$ é tomada a mais desfavorável distribuição a priori. Repare-se que se $\mathcal{H}_0$ coincide com a classe de todas as distribuições a priori, $\mathcal{H}$, está-se a adoptar o princípio

minimax estudado na secção 3.3; se $\mathcal{H}_0$ contém apenas um elemento está-se a adoptar o princípio Bayes estudado na secção 3.4.

Exemplo 3.19 — Com $X \sim N(\theta, 1)$ pretende estimar-se θ com função perca quadrática [como a função perca é convexa não há necessidade de casualizar conforme teorema demonstrado no capítulo seguinte]. Suponha-se que a distribuição a priori, h, tem média, μ, e variância, τ^2, sendo desconhecidas quaisquer outras propriedades. Parece então razoável tomar a classe,

$$\mathcal{H}_0 = \{h : \text{média} = \mu \text{ e variância} = \tau^2\}.$$

Considere-se a função de decisão,

$$\delta_N(x) = \frac{\tau^2}{1+\tau^2}x + \frac{1}{1+\tau^2}\mu.$$

Tem-se,

$$R(\theta, \delta_N) = E\left\{\left[\frac{\tau^2}{1+\tau^2}X + \frac{1}{1+\tau^2}\mu - \theta\right]^2\right\},$$

ou, depois de alguns cálculos,

$$R(\theta, \delta_N) = \frac{\tau^4}{(1+\tau^2)^2} + \frac{(\mu-\theta)^2}{(1+\tau^2)^2}.$$

Consequentemente, para qualquer $h \in \mathcal{H}_0$,

$$R(h, \delta_N) = E\{R(\theta, \delta_N)\} = \frac{\tau^2}{1+\tau^2}. \tag{3.48}$$

No entanto, como δ_N é Bayes contra $N(\mu, \tau^2) \equiv h_N$ (que pertence à classe $\mathcal{H}_0$),

$$R(h_N, \delta^*) \geqq R(h_N, \delta_N) = \frac{\tau^2}{1+\tau^2},$$

o que implica,

$$\sup_{h \in \mathcal{H}_0} R(h, \delta^*) \geqq R(h_N, \delta^*) \geqq \frac{\tau^2}{1+\tau^2}.$$

Combinando com (3.48) tem-se, qualquer que seja δ^*,

$$\sup_{h \in \mathcal{H}_0} R(h, \delta^*) \geqq \frac{\tau^2}{1+\tau^2} = \sup_{h \in \mathcal{H}_0} R(h, \delta_N),$$

e conclui-se que δ_N é minimax-$\mathcal{H}_0$ [Berger (1980)]. □

Exemplo 3.18 — *Continuação.* Tomem-se, $h_C \equiv C(0,1)$ e $h_N \equiv N(0,2{,}19)$ como elementos típicos (mas substancialmente diferentes) de uma classe de distribuições a priori, seja $\mathcal{H}_0$. Seja δ_N a função de decisão Bayes contra h_N e δ_C a função de decisão Bayes contra h_C. O cálculo numérico mostra que $R(h_N, \delta_C)$ é significativamente inferior a um, enquanto $R(h_C, \delta_N) = \infty$. Citando Berger (1980): «*While certainly not a very exhaustive investigation of* Γ*-minimaxility* [na notação aqui adoptada $\Gamma \equiv \mathcal{H}_0$], *this at least indicates that* δ_N *is very bad from a* Γ*-minimax viewpoint, while* δ_C *appears reasonable. Of course, the fact that* $R(h_C, \delta_N) = \infty$ *is due to the use of an unbounded loss function, but, even for bounded losses,* δ_N *will be a good deal worse than* δ_C». □

Procurou chamar-se a atenção, abordando alguns aspectos do conceito de robustez, para o cuidado a ter na especificação da distribuição a priori, sobretudo quanto ao melindre do que se passa nas respectivas abas. Repare-se, a terminar, que os bayesianos puros, se aceitam a análise em termos do risco a posteriori, não avalizam a análise em termos do risco a priori [$R(\theta, \delta^*)$ ou $R(h, \delta^*)$ envolvem a integração sobre o espaço de resultados, $\mathcal{X}$] e, portanto, também não avalizam o critério minimax-$\mathcal{H}_0$. Curiosamente, o risco a priori exprime as propriedades «clássicas» (precisão inicial) de procedimentos bayesianos e estes, por vezes, têm de conformar-se com o seu emprego na falta de saídas alternativas, nomeadamente quando não é adequada a análise de sensibilidade da distribuição a posteriori em relação a uma classe de distribuições a priori.

CAPÍTULO 4

TEOREMAS FUNDAMENTAIS

4.1 Admissibilidade: classes completas

O problema fundamental da teoria da decisão é o da escolha de uma função de decisão dentro da classe D^* [ou D se a casualização é irrelevante]. Uma das vias para facilitar a solução foi referida na secção 3.2: consiste em restringir o conjunto D^* através da eliminação dos elementos que não possuem determinadas propriedades. Semelhante via traduz-se pela imposição de requisitos adicionais (invariância, não enviesamento, etc.) e pode violar o princípio de coerência. É possível, porém, operar uma redução inicial de D^* sem impôr condições adicionais para além da aceitação da função risco: trata-se de eliminar pura e simplesmente as funções de decisão que são claramente inferiores em termos de risco. Para clarificar este ponto retomam-se os conceitos de dominação e de admissibilidade.

A função de decisão, $\delta \in D^*$, domina, $\delta' \in D^*$, quando,

$$R(\theta, \delta) \leqq R(\theta, \delta') \text{ para todo o } \theta \in \Theta, \tag{4.1}$$

isto é, quando a perca esperada ou risco correspondente a δ nunca é superior ao risco correspondente a δ' qualquer que seja o estado; δ domina estritamente δ' quando se verifica (4.1) e,

$$R(\theta, \delta) < R(\theta, \delta') \text{ para algum } \theta \in \Theta. \tag{4.2}$$

Alternativamente diz-se que δ' é dominada ou dominada estritamente por δ.

Função de decisão que não é dominada estritamente por nenhuma outra diz-se admissível; $\mathcal{A}$, $\mathcal{A} \in D^*$, designa a classe das funções de decisão admissíveis.

Como se verifica por (4.1) e (4.2) a admissibilidade é um conceito condicional: depende da função risco e, em última análise, da função perca.

A dominação estrita estabelece em D^* uma ordem parcial. Assim, pode afirmar-se que as funções de decisão admissíveis são «óptimas» num sentido muito fraco. Quer dizer, a classe $\mathcal{A}$ pode ser muito ampla e conter até elementos pouco razoáveis.

Exemplo 4.1 — Com $X \sim N(\theta, 1)$ — mais uma vez — pretende estimar-se θ com função perca quadrática. Considere-se o conjunto de funções de decisão, $\delta_c(x) = cx$; tem-se,

$$\begin{aligned} R(\theta, \delta_c) &= E\{(\theta - cX)^2\} \\ &= E\{[c(\theta - X) + (1-c)\theta]^2\} \\ &= c^2 E\{(\theta - X)^2\} + 2c(1-c)\theta E\{\theta - X\} + (1-c)^2\theta^2 \\ &= c^2 + (1-c)^2\theta^2. \end{aligned}$$

Como para $c > 1$,

$$R(\theta, \delta_1) = 1 < c^2 + (1-c)^2\theta^2 = R(\theta, \delta_c),$$

as funções de decisão δ_c, $c > 1$, são inadmissíveis. Por outro lado, para $0 \leqq c \leqq 1$ as funções de decisão não são comparáveis; em particular, tem-se que $\delta_0(x) \equiv 0$ é admissível! A ilação a tirar é a seguinte: a admissibilidade, embora considerada desejável pela maioria dos autores, não conduz necessariamente a funções de decisão razoáveis [Berger (1980)]. □

A admissibilidade pode ser de difícil investigação. Por vezes pode justificar-se o emprego de uma função de decisão inadmissível desde que as conhecidas como admissíveis sejam de cálculo mais complexo e não introduzam melhoria apreciável em termos de risco.

Com Θ finito, seja $\Theta = \{\theta_1, \theta_2\}$, é fácil interpretar geometricamente os conceitos apresentados. Na Fig. 4.1, onde os pontos $[R(\theta_1, \delta), R(\theta_2, \delta)]$ são referenciados por δ, a função de decisão δ^* domina estritamente δ_1 e δ_2, é dominada estritamente por δ_3 e δ_4, e não é comparável com δ_5 e δ_6.

Com Θ infinito, seja $\Theta = (\theta', \theta'')$, a situação não é tão clara. Na Fig. 4.2, a função de decisão δ_1 é dominada estritamente por δ_2 e δ_3, mas estas não são comparáveis.

Uma classe de funções de decisão diz-se completa e representa-se por $\mathfrak{C}$, $\mathfrak{C} \subset D^*$, se para qualquer função de decisão não pertencente à classe, $\delta' \notin \mathfrak{C}$, existe uma função de decisão pertencente à classe que domina estritamente δ'. Uma classe diz-se essencialmente completa se na definição anterior a relação de dominação estrita é substituída pela de dominação.

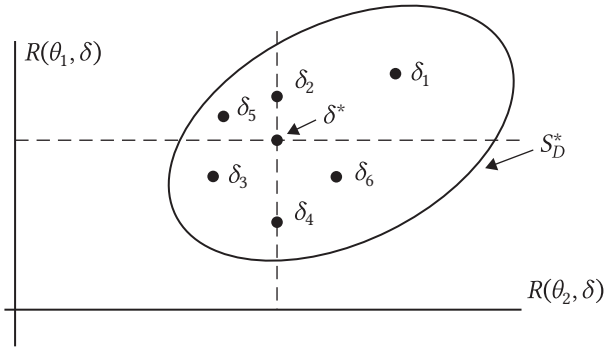


Fig. 4.1

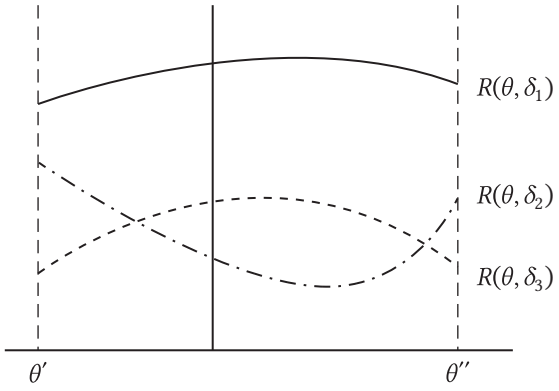


Fig. 4.2

Usando uma classe completa opera-se, eventualmente, uma redução de D^*, com a certeza de que as funções de decisão eliminadas não podem ser óptimas. Aliás demonstra-se facilmente por absurdo que,

Teorema 4.1 — Uma classe completa contém todas as funções de decisão admissíveis. □□

Duas funções de decisão, δ e δ', dizem-se equivalentes quando,

$$R(\theta, \delta) = R(\theta, \delta') \text{ para todo o } \theta \in \Theta;$$

é simples exercício mostrar que,

Teorema 4.2 — Se uma função de decisão admissível, δ, não pertence a classe essencialmente completa $\mathcal{C}$, existe $\delta' \in \mathcal{C}$ equivalente a δ. □□

Podem existir muitas classes completas entre as quais se conta a classe universal, D^*. Ao investigar uma classe completa tão reduzida quanto possível é-se conduzido ao conceito de classe completa mínima: classe completa sem subclasse própria completa.

Teorema 4.3 — Se a classe de funções de decisão admissíveis, $\mathcal{A}$, é completa, então é necessariamente classe completa mínima. Se existir classe completa mínima essa classe é $\mathcal{A}$.

Dem. Pelo Teorema 4.1, $\mathcal{A} \subset \mathcal{C}$, para toda a classe completa $\mathcal{C}$. Portanto, se $\mathcal{A}$ for classe completa é completa mínima. Se $\mathcal{A}$ não for completa, dada uma classe completa qualquer, $\mathcal{C}$, a diferença, $\mathcal{C} - \mathcal{A}$, não é conjunto vazio. Seja $\delta' \in \mathcal{C} - \mathcal{A}$ e considere-se, $\mathcal{C}_1 = \mathcal{C} - \{\delta'\}$. Como δ' é inadmissível existe δ'' que domina estritamente δ' e pode pertencer a $\mathcal{C}$, logo também a $\mathcal{C}_1$; se não pertencer a $\mathcal{C}$, existe δ''' em $\mathcal{C}$ (por ser classe completa) que domina δ'' e, consequentemente, δ'. Obviamente, em qualquer hipótese, $\mathcal{C}_1$ também é classe completa e $\mathcal{C}_1 \subset \mathcal{C}$. Repetindo o raciocínio com $\mathcal{C}_1 - \mathcal{A} \neq \emptyset$ obtém-se uma nova classe completa, $\mathcal{C}_2$, $\mathcal{C}_2 \subset \mathcal{C}_1$, e assim por diante. Se para uma classe completa arbitrária existe sempre subclasse própria completa isso quer dizer que não existe classe completa mínima. Logo, se existir classe completa mínima, $\mathcal{A}$ é completa e coincide com essa classe completa mínima. □□

4.2 Classes essencialmente completas notáveis

Na secção 3.4 mostrou-se que segundo o ponto de vista Bayes não há qualquer vantagem no emprego de funções de decisão casualizadas. Abandonando a perspectiva exclusivamente bayesiana e introduzindo algumas restrições sobre o espaço de acções e sobre a função perca podem estabelecer-se, com alguma generalidade, condições em que a casualização é irrelevante.

Teorema 4.4 — Se o espaço de acções, A, é conjunto convexo de $\mathbf{R}$ [de $\mathbf{R}^k$], se para todo o $\theta \in \Theta$, $L(\theta, a)$, é função convexa de a e se para algum $\theta' \in \Theta$ existem $\xi > 0$ e c tais que,

$$L(\theta', a) \geqq \xi |\, a \,| + c,$$

então, para toda a acção mista, $a^* \in A^*$, existe uma acção pura, $a_0 \in A$, verificando,

$$L(\theta, a_0) \leqq L(\theta, a^*) \text{ para todo o } \theta \in \Theta.$$

Dem. A cada $a^* \in A^*$ corresponde uma distribuição de probabilidade sobre A; seja ζ a distribuição associada com a^* e seja Z uma variável aleatória assumindo

valores em A com distribuição ζ. Como,

$$\xi E\{|\, Z\,|\} + c \leqq E\{L(\theta', Z)\} = L(\theta', a^*),$$

tem-se que $E\{Z\}$ é finito [do conjunto A^* excluem-se, por definição, os elementos a^* para os quais $L(\theta, a^*)$ não é finita]. Aplicando a relação (2.23) à função convexa de a, $L(\theta, a)$, vem, para todo o $\theta \in \Theta$,

$$L[\theta, E\{Z\}] \leqq E\{L(\theta, Z)\};$$

como A é convexo[1], $E\{Z\} \in A$. Tome-se $a_0 = E\{Z\}$; assim, para todo o $a^* \in A^*$ existe $a_0 \in A$, verificando,

$$L(\theta, a_0) \leqq L(\theta, a^*) \text{ para todo o } \theta \in \Theta,$$

e a demonstração fica completa. □□

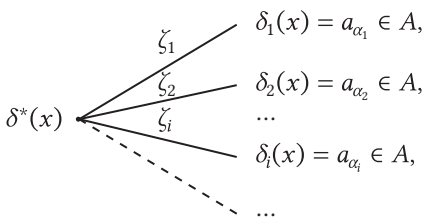
Recorde-se que função de decisão pura, $\delta \in D$, é aplicação,

$$\delta : \mathfrak{X} \to A,$$

isto é, faz corresponder a cada observação, $x \in \mathfrak{X}$, uma acção, $a \in A$; função de decisão mista, $\delta^* \in D^*$, é aplicação,

$$\delta^* : \mathfrak{X} \to A^*,$$

isto é, faz corresponder a cada observação, $x \in \mathfrak{X}$, uma acção mista, $a^* \in A^*$. Com efeito, suponha-se que $\delta^* \in D^*$ resulta da mistura de $\delta_i \in D$, $i = 1, 2, \ldots$, com probabilidades, ζ_i, $i = 1, 2, \ldots$, $\zeta_i \geqq 0$, $\Sigma\zeta_i = 1$. Dado $x \in \mathfrak{X}$, tem-se o seguinte esquema,



[1] Se A é conjunto convexo, $A \subset \mathbf{R}^k$, e para a variável aleatória, Z, $P(Z \in A) = 1$ — como se admite no presente caso — demonstra-se [Ferguson (1967)] que $E\{Z\} \in A$.

Assim, δ^* faz corresponder ao dado x a acção mista que equivale a tomar as acções, puras,

$$a_{\alpha_1}, a_{\alpha_2}, \ldots, a_{\alpha_i}, \ldots$$

com probabilidades,

$$\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots$$

A característica das funções de decisão mistas que acaba de referir-se vem na linha da conclusão a que se havia chegado — veja-se pág. 143 — de que a casualização que implicam tanto pode ser feita antes como depois de obter x. O que interessa de facto destacar é que para cada $x \in \mathcal{X}$, $\delta(x) \in A$, enquanto, $\delta^*(x) \in A^*$. Logo, verificadas as condições em que é válido o teorema anterior, tem-se, para todo o $x \in \mathcal{X}$, dada uma função de decisão mista, seja $\delta^* \in D^*$, que existe sempre uma função de decisão pura, $\delta_0 \in D$, satisfazendo,

$$L[\theta, \delta_0(x)] \leqq L[\theta, \delta^*(x)],$$

para todo o $\theta \in \Theta$. Assim, na medida em que existam valores esperados,

$$R(\delta, \theta_0) = E\{L[\theta, \delta_0(X)]\} \leqq R(\theta, \delta^*) = E\{L[\theta, \delta^*(X)]\},$$

propriedade que permite reformular o Teorema 4.4.

Teorema 4.5 — Verificadas as condições do teorema anterior, a classe das funções de decisão puras, D, é essencialmente completa. □□

Exemplo 4.2 — Sendo $X \sim B(N; \theta)$, analise-se a estimação do parâmetro θ, com função perca, $L(\theta, a) = (\theta - a)^2$; tem-se A [convexo] $= \Theta = [0, 1]$ e espaço de resultados, $\mathcal{X} = \{0, 1, 2, \ldots, N\}$. Considerem-se as funções de decisão puras,

$$\delta_1(x) = x/N \quad \text{e} \quad \delta_2(x) = 1/2,$$

e a função de decisão mista, δ^*, obtida combinando δ_1 e δ_2 com igual probabilidade. Do Teorema 4.4 conclui-se que a função de decisão pura,

$$\begin{aligned} \delta_0(x) &= E\{Z\} \\ &= \delta_1(x) \cdot \frac{1}{2} + \delta_2(x) \cdot \frac{1}{2} \\ &= \frac{2x + N}{4N}, \end{aligned}$$

é pelo menos tão boa como δ^*. Na relação acima há que reconhecer que para cada x a variável aleatória Z assume o valor $\delta_1(x) \in A$ com probabilidade 1/2 e o valor $\delta_2(x) \in A$ com probabilidade 1/2.

No que diz respeito às funções risco um cálculo simples mostra ser $R(\theta, \delta_0) = R(\theta, \delta^*)/2$; quer dizer, δ_0 é mesmo melhor do que δ^*, propriedade que decorre da convexidade estrita da função perca. □

Como é legítimo conjecturar a partir dos Teoremas 4.4 e 4.5, as funções perca convexas permitem introduzir simplificações em muitos problemas de decisão. Assim, entre as percas mais usadas contam-se,

$$L(\theta, a) = (\theta - a)^2 \text{ ou } L(\theta, a) = \lambda(\theta)(\theta - a)^2 \quad [\lambda(\theta) > 0],$$

que para cada $\theta \in \Theta$ são funções estritamente convexas de a, e

$$L(\theta, a) = |\theta - a| \text{ ou } L(\theta, a) = \lambda(\theta)|\theta - a| \quad [\lambda(\theta) > 0],$$

que para cada $\theta \in \Theta$ são funções convexas de a. Aliás o segundo grupo tem a particularidade de corresponder às funções convexas pares que apresentam o mais lento crescimento para variações no erro absoluto $|\theta - a|$. Em relação ao primeiro grupo traduzem situações em que há menor sensibilidade ou se penalizam menos os erros grandes.

Quando se emprega uma função de decisão, $\delta \in D^*$, valores grandes da perca, $L[\theta, \delta(x)]$, estão em geral associados com valores grandes de $|x|$, isto é, com o comportamento das abas da distribuição de X (maior ou menor possibilidade de «*outliers*»)[2]. Para traduzir uma sensibilidade intermédia entre erro quadrático e erro absoluto, Huber [veja-se Lehmann (1983)] propõe funções perca da forma,

$$L(\theta, a) = \begin{cases} (\theta - a)^2 & \text{se } |\theta - a| \leqq K, \\ 2K|\theta - a| - K^2 & \text{se } |\theta - a| \geqq K, \end{cases}$$

que são convexas mas não estritamente convexas. Uma forma alternativa, que resulta da interpolação entre aqueles dois tipos, $L(\theta, a) = |\theta - a|^q$, $1 < q < 2$, é estritamente convexa. Não obstante serem funções convexas, os tipos intermédios são de tratamento mais difícil e conduzem a soluções de interpretação menos directa: por esse facto são aqui raramente considerados.

Os próximos resultados dizem respeito ao acolhimento do princípio de suficiência no âmbito da teoria da decisão estatística. O tratamento da questão faz-se de modo mais conveniente retomando as funções de decisão aleatórias oportunamente introduzidas [veja-se pág. 142].

[2] Confronte-se com as considerações feitas na secção 3.6 sobre a função perca na óptica da robustez bayesiana — análise condicionada por $X = x$. Estando em jogo o risco (frequencista) $R(\theta, \delta)$, a presente análise é particularmente relevante.

Uma função de decisão aleatória, $\phi \in \Phi$, diz-se baseada na estatística suficiente, $T(X)$, quando somente depende de x através de $T(x)$,

$$\phi(a \,|\, x) = \phi[a \,|\, T(x)], \tag{4.3}$$

ou, de outro modo, quando,

$$T(x) = T(x') \;\Rightarrow\; \phi(a \,|\, x) = \phi(a \,|\, x'). \tag{4.4}$$

Dada uma amostra casual, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, com espaço de resultados, $\mathfrak{X} = \mathbf{R}^N$, uma estatística suficiente, $T(\mathbf{X})$ — e de uma maneira geral qualquer estatística — estabelece em $\mathbf{R}^N$ uma partição, seja Π. Os conjuntos da partição, sendo $\mathfrak{J}$ o domínio de T, podem designar-se,

$$\Pi_t = \{\mathbf{x} = (x_1, x_2, \ldots, x_N) \,:\, T(\mathbf{x}) = t, \quad t \in \mathfrak{J}\},$$
$$\Pi_t \cap \Pi_{t'} = \emptyset,\; t \neq t',\; \bigcup_{t \in \mathfrak{J}} \Pi_t = \mathbf{R}^N.$$

Se ϕ é baseada na estatística suficiente T então ϕ é constante [a mesma distribuição sobre A] sobre cada conjunto da partição,

$$\phi(a \,|\, \mathbf{x}) = \phi(a \,|\, t) \text{ para todo o } \mathbf{x} \in \Pi_t. \tag{4.5}$$

Exemplo 4.3 — Considere-se a amostra casual, $\mathbf{X} = (X_1, X_2, X_3)$, $X_i \sim B(1;\theta)$, $i = 1, 2, 3$, caso em que $T = X_1 + X_2 + X_3$ é estatística suficiente para θ. Suponha-se, $A = \{a_1, a_2\}$; uma função de decisão aleatória, $\phi \in \Phi$, representa-se por,

$$\phi \equiv [\phi(a_1 \,|\, \mathbf{x}),\; \phi(a_2 \,|\, \mathbf{x})],\; \phi(a_i \,|\, \mathbf{x}) \geqq 0,\; \phi(a_1 \,|\, \mathbf{x}) + \phi(a_2 \,|\, \mathbf{x}) = 1,$$

ou seja, em virtude da última relação,

$$\phi \equiv [\phi(a_1 \,|\, \mathbf{x}),\; 1 - \phi(a_1 \,|\, \mathbf{x})], \quad 0 \leqq \phi(a_1 \,|\, \mathbf{x}) \leqq 1,$$

onde $\phi(a_1 \,|\, \mathbf{x})$ é a probabilidade com que a acção a_1 é tomada quando se emprega $\phi \in \Phi$ e a observação é $\mathbf{X} = \mathbf{x}$. Assim, o estudo de ϕ pode fazer-se considerando simplesmente $\phi(a_1 \,|\, \mathbf{x})$.

No quadro abaixo indica-se a partição de $\mathfrak{X}$ operada pela estatística suficiente, T, e compara-se uma função de decisão aleatória genérica com uma função de decisão aleatória baseada em T, seja $\phi' \in \Phi_0$, onde $\Phi_0 \subset \Phi$ é a classe das funções de decisão baseadas em T.

Π	$\mathbf{x}$	$\phi \in \Phi$	$\phi' \in \Phi_0$
Π_0	(0,0,0)	$\phi[a_1 \mid (0,0,0)]$	$\phi'[a_1 \mid (0,0,0)]$
Π_1	(1,0,0)	$\phi[a_1 \mid (1,0,0)]$	$\phi'[a_1 \mid (1,0,0)] =$
	(0,1,0)	$\phi[a_1 \mid (0,1,0)]$	$= \phi'[a_1 \mid (0,1,0)] =$
	(0,0,1)	$\phi[a_1 \mid (0,0,1)]$	$\phi'[a_1 \mid (0,0,1)]$
Π_2	(1,1,0)	$\phi[a_1 \mid (1,1,0)]$	$\phi'[a_1 \mid (1,1,0)] =$
	(1,0,1)	$\phi[a_1 \mid (1,0,1)]$	$= \phi'[a_1 \mid (1,0,1)] =$
	(0,1,1)	$\phi[a_1 \mid (0,1,1)]$	$\phi'[a_1 \mid (0,1,1)]$
Π_3	(1,1,1)	$\phi[a_1 \mid (1,1,1)]$	$\phi'[a_1 \mid (1,1,1)]$

□

O teorema seguinte mostra que no caso de haver uma estatística suficiente não se retira qualquer vantagem do emprego de funções de decisão que não sejam baseadas em tal estatística[3].

Teorema 4.6 — Se T é estatística suficiente para θ e se $\Phi_0 \subset \Phi$ representa a classe de funções de decisão aleatórias baseadas em T, então Φ_0 é essencialmente completa.

Dem. É necessário provar que dado qualquer $\phi \in \Phi$ existe um $\phi' \in \Phi_0$, tal que,

$$\hat{R}(\theta, \phi') \leqq \hat{R}(\theta, \phi) \text{ para todo o } \theta \in \Theta,$$

onde $\hat{R}(\theta, \phi)$ é a função risco definida por (3.11) e (3.12).

Seja, $\Pi = \{\Pi_t : t \in \mathfrak{J}\}$, a partição instituída em $\mathfrak{X}$ por T. Se ϕ' é por hipótese baseada em T então ϕ' é constante sobre cada conjunto Π_t; com $\phi \in \Phi$ arbitrária, tome-se, para todo o $x \in \Pi_t$,

$$\phi'(a \mid x) = \phi'(a \mid t) = E\{\phi(a \mid X) \mid T = t\}. \tag{4.6}$$

Isto é, para cada $t \in \mathfrak{J}$, $\phi'(a \mid t)$ calcula-se como média condicionada de $\phi(a \mid X)$ quando X percorre Π_t,

$$E\{\phi(a \mid X) \mid T = t\} = \int_{\Pi_t} \phi(a \mid x) f(x \mid t)\, dx, \tag{4.7}$$

onde $f(x \mid t)$ é a densidade de X condicionada por $T = t$ que é independente de θ por T ser suficiente para θ.

[3] O princípio de suficiência encontra assim pleno acolhimento na teoria da decisão estatística.

Por outro lado [o exemplo seguinte permite acompanhar melhor a demonstração] tem-se,

$$E\{\phi'(a\,|\,X)\} = \int_{\mathfrak{X}} \phi'(a\,|\,x) f(x\,|\,\theta)\,dx$$
$$= \int_{\mathfrak{J}} \left\{ \int_{\Pi_t} \phi'(a\,|\,x) f(x\,|\,t)\,dx \right\} g(t\,|\,\theta)\,dt,$$

onde $g(t\,|\,\theta)$ é a densidade de T; mas, por (4.6),

$$E\{\phi'(a\,|\,X)\} = \int_{\mathfrak{J}} \phi'(a\,|\,t) g(t\,|\,\theta)\,dt = E\{\phi'(a\,|\,T)\}.$$

Por outro lado,

$$E\{\phi(a\,|\,X)\} = \int_{\mathfrak{X}} \phi(a\,|\,x) f(x\,|\,\theta)\,dx$$
$$= \int_{\mathfrak{J}} \left\{ \int_{\Pi_t} \phi(a\,|\,x) f(x\,|\,t)\,dx \right\} g(t\,|\,\theta)\,dt,$$

ou ainda, por (4.6) e (4.7),

$$E\{\phi(a\,|\,X)\} = \int_{\mathfrak{J}} \phi'(a\,|\,t) g(t\,|\,\theta)\,dt = E\{\phi'(a\,|\,T)\}.$$

Em resumo,

$$E\{\phi'(a\,|\,X)\} = E\{\phi(a\,|\,X)\}. \tag{4.8}$$

Por (3.11) e (3.12),

$$\hat{R}(\theta, \phi) = \int_{\mathfrak{X}} \left\{ \int_A L(\theta, a) \phi(a\,|\,x)\,da \right\} f(x\,|\,\theta)\,dx,$$

entendendo $\phi(a\,|\,x)$ como densidade sobre A. Admitindo válida a permuta dos integrais,

$$\hat{R}(\theta, \phi) = \int_A L(\theta, a) \left\{ \int_{\mathfrak{X}} \phi(a\,|\,x) f(x\,|\,\theta)\,dx \right\} da.$$

Logo, por (4.8), tem-se finalmente,

$$\hat{R}(\theta, \phi') = \hat{R}(\theta, \phi), \quad \theta \in \Theta,$$

e o teorema fica demonstrado. □□

Exemplo 4.3 — *Continuação.* Supondo que $\phi \in \Phi$ é arbitrária, proceda-se à construção de $\phi' \in \Phi_0$ segundo (4.6). Tem-se,

(i) $\phi'[a_1\,|\,(0,0,0)] = \phi'(a_1\,|\,t=0) = \phi[a_1\,|\,(0,0,0)]$;
(ii) $\phi'[a_1\,|\,(1,0,0)] = \phi'[a_1\,|\,(0,1,0)] = \phi'[a_1\,|\,(0,0,1)] = \phi'(a_1\,|\,t=1)$,

com,

$$\phi'(a_1\,|\,t=1) = \phi[a_1\,|\,(1,0,0)]\cdot P(\mathbf{X}=(1,0,0)\,|\,T=1) + \\ +\ \phi[a_1\,|\,(0,1,0)]\cdot P(\mathbf{X}=(0,1,0)\,|\,T=1) + \phi[a_1\,|\,(0,0,1)]\cdot P(\mathbf{X}=(0,0,1)\,|\,T=1),$$

donde,

$$\phi'(a_1\,|\,t=1) = \{\phi[a_1\,|\,(1,0,0)] + \phi[a_1\,|\,(0,1,0)] + \phi[a_1\,|\,(0,0,1)]\}(1/3);$$

(iii) $\phi'[a_1\,|\,(1,1,0)] = \phi'[a_1\,|\,(1,0,1)] = \phi'[a_1\,|\,(0,1,1)] = \phi'(a_1\,|\,t=2)$,

com,

$$\phi'(a_1\,|\,t=2) = \phi[a_1\,|\,(1,1,0)]\cdot P(\mathbf{X}=(1,1,0)\,|\,T=2) + \\ +\ \phi[a_1\,|\,(1,0,1)]\cdot P(\mathbf{X}=(1,0,1)\,|\,T=2) + \phi[a_1\,|\,(0,1,1)]\cdot P(\mathbf{X}=(0,1,1)\,|\,T=2),$$

donde,

$$\phi'(a_1\,|\,t=2) = \{\phi[a_1\,|\,(1,1,0)] + \phi[a_1\,|\,(1,0,1)] + \phi[a_1\,|\,(0,1,1)]\}(1/3);$$

(iv) $\phi'[a_1\,|\,(1,1,1)] = \phi'(a_1\,|\,t=3) = \phi[a_1\,|\,(1,1,1)]$. □

Exemplo 4.3 — *Continuação.* Não obstante o pesado formalismo, vai ilustrar-se a relação (4.8). Como $A=\{a_1,a_2\}$, basta tomar a_1. Tem-se,

$$\begin{aligned} E\{\phi(a_1\,|\,\mathbf{X})\} &= \sum_{\mathbf{x}\in\mathfrak{X}} \phi[a_1\,|\,(x_1,x_2,x_3)]\ \theta^{\Sigma x_i}(1-\theta)^{3-\Sigma x_i} \\ &= \phi[a_1\,|\,(0,0,0)](1-\theta)^3 + \{\phi[a_1\,|\,(1,0,0)] + \\ &\quad + \phi[a_1\,|\,(0,1,0)] + \phi[a_1\,|\,(0,0,1)]\}\theta(1-\theta)^2 + \\ &\quad + \{\phi[a_1\,|\,(1,1,0)] + \phi[a_1\,|\,(1,0,1)] + \\ &\quad + \phi[a_1\,|\,(0,1,1)]\}\theta^2(1-\theta) + \phi[a_1\,|\,(1,1,1)]\theta^3 \\ &= \sum_{t=0}^{3} \phi'(a_1\,|\,t)\,\binom{3}{t}\,\theta^t(1-\theta)^{3-t} \\ &= E\{\phi'(a_1\,|\,T)\} = E\{\phi'(a_1\,|\,\mathbf{X})\}. \end{aligned}$$

□

Segundo o Teorema 4.6, o conhecimento de uma estatística suficiente, T, permite restringir a escolha da função de decisão à classe das baseadas em T. De facto, qualquer que seja $\phi\in\Phi$ existe sempre $\phi'\in\Phi_0$, ϕ' dado por (4.6), com risco idêntico a ϕ.

Por outro lado, desde que se verifiquem as condições do Teorema 4.4, a classe das funções de decisão puras, D, é também essencialmente completa. Retendo essas condições é interessante investigar se D_0, $D_0 \subset D$,

$$D_0 = \{\delta \in D : \delta(x) = \delta[T(x)]\},$$

subclasse das funções de decisão puras baseadas na estatística suficiente, T, é ainda essencialmente completa. A verificar-se esta propriedade, além da redução de dados operada com a estatística suficiente, pode dispensar-se a casualização. De facto, tem-se,

Teorema 4.7 — Verificadas as condições do Teorema 4.4, a classe das funções de decisão puras baseadas numa estatística suficiente, D_0, é essencialmente completa.

Dem. Em termos gerais, pouco rigorosos, é a seguinte: qualquer que seja $\phi \in \Phi$, existe, pelo Teorema 4.6, $\phi' \in \Phi_0$, pelo menos tão boa como ϕ; a ϕ' pode fazer-se corresponder, de acordo com o Teorema 4.4, a função de decisão pura, seja δ, definida pela relação,

$$\delta(x) = \int_A a\phi'(a \mid x)\, da \quad \text{para todo o } x \in \mathfrak{X},$$

que é pelo menos tão boa como ϕ'. Finalmente, dado que ϕ' é baseada na estatística suficiente, T, $\phi'(a \mid x) = \phi'[a \mid T(x)]$, também, $\delta(x) = \delta[T(x)]$, isto é, também δ é baseada em T e a demonstração fica completa. □□

4.3 Admissibilidade das funções de decisão Bayes

As proposições seguintes focam alguns aspectos da relação que existe entre funções de decisão Bayes e funções de decisão admissíveis. Da sua análise pode concluir-se que as regras Bayes são virtualmente admissíveis, propriedade que aliada à optimização formal em termos da teoria da utilidade — no quadro da teoria da utilidade as funções de decisão Bayes são inevitavelmente as melhores — representa forte argumento em favor do respectivo emprego.

Teorema 4.8 — Se para uma distribuição a priori, h, δ_h é a única função de decisão [a menos de uma relação de equivalência] Bayes contra h, então, δ_h é admissível.

Dem. Por hipótese,

$$R(h, \delta_h) = \inf_{\delta^*} R(h, \delta^*),$$

ou ainda,

$$R(h, \delta_h) < R(h, \delta^*),$$

para todo o $\delta^* \in D^*$ não equivalente a δ_h. Considerando o valor esperado em relação a h, vem,

$$E\{R(\theta, \delta_h) - R(\theta, \delta^*)\} < 0,$$

para todo o $\delta^* \in D^*$ não equivalente a δ_h. A desigualdade anterior implica a impossibilidade de verificar-se, $R(\theta, \delta_h) \geqq R(\theta, \delta^*)$, para todo o $\theta \in \Theta$, ficando assim provada a admissibilidade de δ_h. □□

O teorema anterior mostra o interesse de investigar em que condições uma função de decisão é Bayes única. A seguinte proposição pode estudar-se em Lehmann (1983) e descreve uma condição suficiente:

Lema 4.1 — Com função perca quadrática (ou estritamente convexa), a função de decisão, δ_H, Bayes contra H (função de distribuição a priori), é única (a menos de conjuntos de $\mathfrak{X}$ com probabilidade zero qualquer que seja o elemento de $\mathcal{F}$) se: (i) $R(H, \delta_H)$ é finito; (ii) para qualquer conjunto B (boreliano) $\subset \mathfrak{X}$,

$$P(X \in B) = \int P_\theta(X \in B)\, dH(\theta) = 0 \;\Rightarrow\; P_\theta(X \in B) = 0 \text{ para todo o } \theta \in \Theta. \quad \square\square$$

Exemplo 4.4 — Com $x \sim B(N; \theta)$ considere-se o problema de estimar θ, com $\Theta = [0,1]$. Suponha-se que a distribuição a priori é discreta e atribui probabilidade $1/2$ aos pontos $\theta = 0$ e $\theta = 1$. Para qualquer $x \in \mathfrak{X}$, onde, $\mathfrak{X} = \{0, 1, 2, \ldots, N\}$, tem-se,

$$P(X = x) = f(x \mid 0)(1/2) + f(x \mid 1)(1/2),$$

com $f(x \mid \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$. Agora, para qualquer $x \neq 0$, N, $0 < x < N$, vem, $P(X = x) = 0$, mas, $P_\theta(X = x) = f(x \mid \theta) > 0$, para $0 < \theta < 1$.

Dado que,

$$h(0 \mid x) = \begin{cases} 1 & \text{se } x = 0, \\ 0 & \text{se } x \neq 0, \end{cases}$$

$$h(\theta \mid x) = \quad 0 \quad \text{qualquer } x,\ 0 < \theta < 1,$$

$$h(1 \mid x) = \begin{cases} 0 & \text{se } x \neq N, \\ 1 & \text{se } x = N, \end{cases}$$

tem-se, com função perca quadrática,

$$r_x(a) = a^2 h(0 \mid x) + (1-a)^2 h(1 \mid x),$$

isto é,

$$r_x(a) = \begin{cases} a^2 & \text{se } x = 0, \\ 0 & \text{se } x \neq 0, N, \\ (1-a)^2 & \text{se } x = N. \end{cases}$$

Assim, qualquer função de decisão tal que $\delta(0) = 0$ e $\delta(N) = 1$ — e há uma infinidade — é Bayes contra a distribuição a priori considerada. A não unicidade advém da não verificação da condição (ii) do Lema 4.1 [adaptado de Lehmann (1983)] . □

Teorema 4.9 — Com $\Theta = \mathbf{R}$, se $R(\theta, \delta^*)$ é função contínua de θ para todo o $\delta^* \in D^*$, se δ_h é Bayes contra h e $R(h, \delta_h)$ é finito e se h tem $\mathbf{R}$ como suporte[4], então δ_h é admissível.

Dem. Suponha-se que δ_h não é admissível; existe então $\delta^* \in D^*$ tal que $R(\theta, \delta^*) \leqq R(\theta, \delta_h)$ para todo o $\theta \in \Theta$ e $R(\theta_0, \delta^*) < R(\theta_0, \delta_h)$ para algum $\theta_0 \in \Theta$. Em consequência: por um lado, existe um número real $\xi > 0$, tal que,

$$R(\theta_0, \delta^*) \leqq R(\theta_0, \delta_h) - \xi;$$

por outro lado, pela continuidade de $R(\theta, \delta^*)$ em relação a θ, existe uma vizinhança de θ_0, seja $v(\theta_0, \xi)$, tal que,

$$R(\theta, \delta^*) \leqq R(\theta, \delta_h) - \xi \text{ para todo o } \theta \in v(\theta_0, \xi).$$

Finalmente,

$$\begin{aligned} R(h, \delta^*) &= \int_v R(\theta, \delta^*) h(\theta)\, d\theta + \int_{v^C} R(\theta, \delta^*) h(\theta)\, d\theta \\ &\leqq \int_v [R(\theta, \delta_h) - \xi] h(\theta)\, d\theta + \int_{v^C} R(\theta, \delta_h) h(\theta)\, d\theta \\ &= R(h, \delta_h) - \xi P(\theta \in v), \end{aligned}$$

onde $v \equiv v(\theta_0, \xi)$ e $v^C \equiv \Theta - v(\theta_0, \xi)$. Mas, $P(\theta \in v) > 0$, por h ter $\mathbf{R}$ por suporte; logo, $R(h, \delta^*) < R(h, \delta_h)$, o que contradiz a hipótese de que δ_h é Bayes contra h. Portanto, δ_h, é admissível. □□

Dada a dificuldade em apresentar resultados muito gerais, grande parte das proposições do presente capítulo exige que o espaço dos estados seja finito, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Tem-se, então, que a família das distribuições a priori $\mathcal{H} \equiv S_n$, isto é,

$$\mathbf{h} \in \mathcal{H} \;\Rightarrow\; \mathbf{h} = (h_1, h_2, \ldots, h_n),\; h_j \geqq 0, \quad \sum h_j = 1;$$

por outro lado, $\mathcal{H}^+$, designa o subconjunto de $\mathcal{H}$ constituído por vectores de probabilidade com componentes todas positivas,

$$\mathbf{h} \in \mathcal{H}^+ \;\Rightarrow\; \mathbf{h} = (h_1, h_2, \ldots, h_n),\; h_j > 0, \quad \sum h_j = 1.$$

[4] O suporte de uma distribuição sobre Θ é o conjunto de pontos $\theta' \in \Theta$ tais que, para todo $\xi > 0$, $P(\theta' - \xi < \theta < \theta' + \xi) > 0$. No caso presente o suporte é o próprio $\Theta[= \mathbf{R}]$.

Seja $\mathcal{B}$ conjunto das funções de decisão que são Bayes contra algum $\mathbf{h} \in \mathcal{H}$. Se $\delta \in \mathcal{B}$ existe $\mathbf{h} \in \mathcal{H}$ tal que,

$$R(\mathbf{h}, \delta) = \sum R(\theta_j, \delta)h_j \leqq R(\mathbf{h}, \delta^*) = \sum R(\theta_j, \delta^*)h_j \tag{4.9}$$

para todo o $\delta^* \in D^*$.

Analogamente $\mathcal{B}^+$ designa o conjunto de funções de decisão que são Bayes contra algum $\mathbf{h} \in \mathcal{H}^+$. Evidentemente, se $\delta \in \mathcal{B}^+$, verificam-se as desigualdades (4.9) mas nenhum dos h_j pode ser igual a zero. O teorema seguinte pode considerar-se a versão do anterior para o caso Θ finito.

Teorema 4.10 — $\mathcal{B}^+ \subset \mathcal{A}$, isto é, as funções de decisão Bayes contra algum $\mathbf{h} \in \mathcal{H}^+$ são admissíveis.

Dem. Se $\delta \in \mathcal{B}^+$ existe $\mathbf{h} \in \mathcal{H}^+$ tal que,

$$\sum R(\theta_j, \delta)h_j \leqq \sum R(\theta_j, \delta^*)h_j \text{ para todo o } \delta^* \in D^*,$$

ficando desde logo afastada qualquer hipótese de haver δ' dominando estritamente δ. Portanto, $\delta \in \mathcal{A}$. □□

Considere-se o conjunto S_D^* representado na Fig. 4.3. As funções de decisão, δ', δ'', δ''', são todas Bayes contra $\mathbf{h} = (1,0)$, tal como sucede com a infinidade de funções de decisão a que correspondem pontos no mesmo lado do rectângulo.

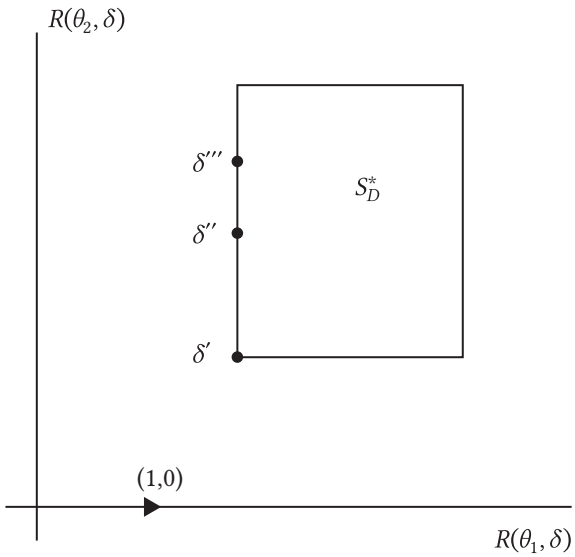


Fig. 4.3

No entanto, só δ' é admissível. Este caso, muito simples, mostra como a admissibilidade das soluções Bayes falha quando $\mathbf{h}$ não tem todas as componentes positivas e serve de contra-exemplo relativamente ao Teorema 4.8 na medida em que apresenta uma situação em que a função de decisão Bayes não é única.

As funções de decisão Bayes formais [veja-se pág. 164] não são necessariamente admissíveis.

Exemplo 4.5 — Suponha-se $X \sim N(\theta, 1)$, $\theta \sim N(0, 1)$ e

$$L(\theta, a) = \exp\{3\theta^2/4\}\,(\theta - a)^2.$$

Considerando o Ex. 1.15 conclui-se que $h(\theta \mid x)$ é uma $N(x/2, 1/2)$. Recorrendo a (3.37) mostra-se que a função de decisão Bayes formal é $\delta_h(x) = 2x$ [o risco a priori, $R(h, \delta_h)$, é infinito]. No entanto, não é difícil calcular, ser,

$$R(\theta, \delta_h) = \exp\{3\theta^2/4\}\,(4 + \theta^2) > \exp\{3\theta^2/4\}\,(1) = R(\theta, \delta_1),$$

onde $\delta_1(x) = x$. Em conclusão, δ_h é seriamente inadmissível [Berger (1980)]. □

O exemplo anterior é, por assim dizer, dilemático. Por um lado, a análise post--experimental, conduz a uma função de decisão que, em sintonia com a teoria da utilidade, resulta da minimização do risco a posteriori e é, em princípio, impecável. Por outro lado, a análise pre-experimental, leva a concluir que tal função de decisão é grosseiramente inadmissível e deve ser descartada. Uma possível via para ultrapassar a dificuldade consiste em excluir da análise os casos em que a função perca (ou a função utilidade) não é limitada (recorde-se o Teorema 2.6).

Tal como se passa com as funções de decisão Bayes formais, as Bayes generalizadas não são necessariamente admissíveis, embora a verificação da sua admissibilidade ou inadmissibilidade possa ser bastante difícil.

Quando $\delta \in D^*$ é Bayes generalizada, a função perca é positiva e a função risco $R(h, \delta) < \infty$, não é em regra árduo provar a admissibilidade. Porém, o que quase sempre acontece é que $R(h, \delta) = \infty$, caso em que, mesmo funções Bayes generalizadas «naturais» ou «clássicas» podem ser inadmissíveis.

Exemplo 4.6 — Considere-se $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim N_k(\boldsymbol{\theta}, I_k)$, onde $\boldsymbol{\theta}$ é o vector das médias, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, I_k é a matriz identidade $k \times k$ e N_k designa a distribuição Normal k-dimensional. Pretende-se estimar $\boldsymbol{\theta}$ com função perca,

$$L(\boldsymbol{\theta}, \mathbf{a}) = \sum_{i=1}^{k} (\theta_i - a_i)^2.$$

Como $\boldsymbol{\theta}$ é um parâmetro de localização, a distribuição a priori não informativa, $h(\boldsymbol{\theta}) = 1$, $\boldsymbol{\theta} \in \mathbf{R}^k$, é a indicada quando não existe informação a priori. Pode então

mostrar-se que a distribuição a posteriori de $\boldsymbol{\theta}$ é ainda Normal k-dimensional, $N_k(\mathbf{x}, I_k)$. A função de decisão Bayes generalizada [$h(\boldsymbol{\theta}) = 1$ é imprópria] vem $\delta_h(\mathbf{x}) = \mathbf{x}$ [tratando-se de uma amostra casual, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, o estimador Bayes generalizado era o vector das médias]. O que é algo surpreendente é que $\delta_h(\mathbf{x})$ é admissível para $k \leqq 2$ e inadmissível para $k \geqq 3$. De facto, James e Stein demonstraram que, para $k \geqq 3$,

$$\delta^{JS}(\mathbf{x}) = \left\{1 - \left[(k-2)/\sum x_i^2\right]\right\} \mathbf{x}$$

domina estritamente δ_h, isto é, $R(\boldsymbol{\theta}, \delta^{JS}) < R(\boldsymbol{\theta}, \delta)$, para todo o $\boldsymbol{\theta} \in \Theta$. Note-se que δ_h é o estimador «óptimo» do ponto de vista clássico [Berger (1980)]. Na secção 6.6 retoma-se o problema dos estimadores de James-Stein. □

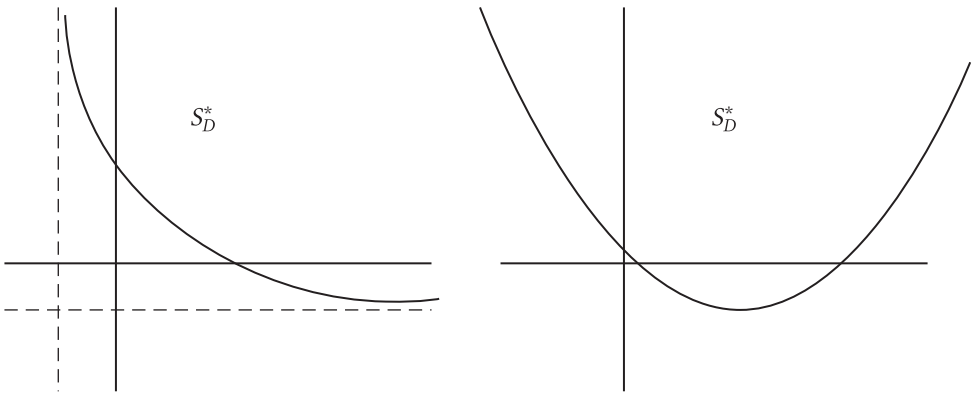
4.4 Existência de soluções Bayes. Teorema da classe completa

Para atacar o problema da existência de soluções Bayes é necessário apresentar alguns conceitos, aliás igualmente relevantes no estudo da existência de soluções minimax.

O conjunto risco, S_D^* [e, em geral, um conjunto de $\mathbf{R}^n$], diz-se limitado inferiormente se existe um número real finito, M, tal que

$$\ell_j \geqq -M, j = 1, 2, \ldots, n, \text{ para todo o ponto } \boldsymbol{\ell} \in S_D^*.$$

Na Fig. 4.4 ilustram-se as duas situações.



(a) conjunto limitado inferiormente

(b) conjunto não limitado inferiormente

Fig. 4.4

Com $\mathbf{z} \in \mathbf{R}^n$, seja $Q_{\mathbf{z}}$ o ortante inferior em $\mathbf{z}$ [recorde-se (3.16)]; como imediatamente se reconhece, $Q_{\mathbf{z}} \cap S_D^*$, quando não for conjunto vazio, compreende os pontos de risco correspondentes a funções de decisão que dominam a função de decisão a que corresponde $\mathbf{z}$ (note-se a Fig. 4.5).

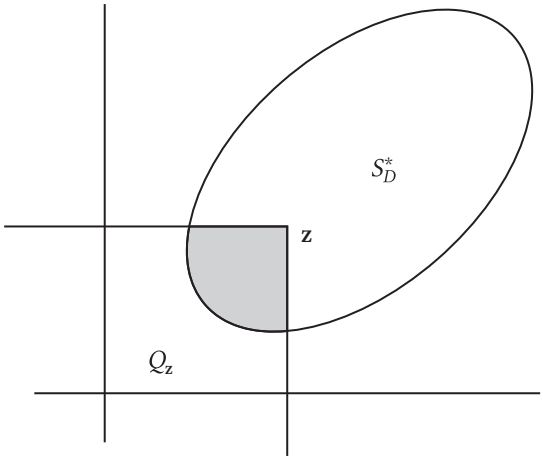


Fig. 4.5

Designe $\overline{S}_D^*$ o fecho de S_D^*; um ponto $\mathbf{z}$ pertence à fronteira inferior de S_D^* [ou, em geral, de um conjunto convexo de $\mathbf{R}^n$], se,

$$Q_{\mathbf{z}} \cap \overline{S}_D^* = \{\mathbf{z}\},$$

onde $\{\mathbf{z}\}$ é o conjunto que tem $\mathbf{z}$ como único elemento.

A fronteira inferior de S_D^* representa-se por $\Lambda(S_D^*)$; diz-se que S_D^* é fechado inferiormente quando,

$$\Lambda(S_D^*) \subset S_D^*. \tag{4.10}$$

Ao longo do presente capítulo vai ter-se uma ideia da importância que têm os problemas de decisão em que S_D^* é limitado inferiormente e fechado inferiormente. Por esse facto interessa fazer referência à forma de investigar a verificação destas duas propriedades.

Quando é válido admitir ser,

$$L(\theta, a) \geqq -K > -\infty,$$

é fácil verificar que $R(\theta, \delta^*) \geqq -K$ e que, consequentemente, S_D^* é limitado inferiormente.

A verificação de que S_D^* é fechado inferiormente é consideravelmente mais difícil. O caminho mais acessível ainda é ensaiar a demonstração de que S_D^* é fechado e, portanto, obviamente fechado inferiormente.

O primeiro dos lemas seguintes permite mostrar que o conjunto risco, S_D^*, pode «herdar» certas propriedades do conjunto perca, S [veja-se 3.4].

Lema 4.2 — Se o conjunto perca, S, é limitado e fechado, o conjunto risco, S_D^* é limitado e fechado. □□

Lema 4.3 — Seja, A, o espaço de acções: o conjunto perca, S, é limitado e fechado quando se verifica qualquer das condições:

1) A é finito;
2) A é conjunto limitado e fechado de $\mathbf{R}^k$ e $L(\theta_i, a)$ é função contínua de a para $i = 1, 2, \ldots, n$. □□

Referências a estes resultados podem ver-se em Blackwell e Girshick (1954) e Berger (1980). Felizmente, na grande maioria dos problemas de decisão estatística o conjunto risco é fechado (vejam-se os Exs. 7.3, 7.4 e 7.5).

A importância da fronteira inferior relativamente à admissibilidade é estabelecida no,

Teorema 4.11 — Se $\boldsymbol{\ell}(\delta) \in \Lambda(S_D^*)$ então $\delta \in \mathcal{A}$, isto é, funções de decisão a que correspondem pontos na fronteira inferior do conjunto risco casualizado são admissíveis.

Dem. Se $\boldsymbol{\ell}(\delta) \in \Lambda(S_D^*)$, a existência de δ' com $\boldsymbol{\ell}(\delta') \in S_D^*$, $\ell_j(\delta') \leqq \ell_j(\delta)$ para $j = 1, 2, \ldots, n$ e desigualdade estrita para algum j conduz a uma contradição. De facto, tem-se nesse caso,

$$Q_{\boldsymbol{\ell}(\delta')} \subset Q_{\boldsymbol{\ell}(\delta)},$$

o que implica,

$$[Q_{\boldsymbol{\ell}(\delta')} \cap S_D^*] \subset [Q_{\boldsymbol{\ell}(\delta)} \cap S_D^*] = \{\boldsymbol{\ell}(\delta)\}$$

e $\boldsymbol{\ell}(\delta') \neq \boldsymbol{\ell}(\delta)$ não pode pertencer a S_D^*. □□

E, agora, a primeira proposição sobre a existência de soluções ou funções de decisão Bayes:

Teorema 4.12 — Se S_D^* é limitado inferiormente e fechado inferiormente, existe, para todo o $\mathbf{h} \in \mathcal{H}^+$, uma função de decisão Bayes contra $\mathbf{h}$.

Dem. Para qualquer $\mathbf{h} \in \mathcal{H}^+$, seja $\mathbf{B}$ o conjunto de números reais,

$$\mathbf{B} = \{b : b = \Sigma \ell_j h_j \text{ para algum } \boldsymbol{\ell} \in S_D^*\}.$$

Na Fig. 4.6 indicam-se, com $n = 2$, as rectas correspondentes a quatro valores de b e o vector $\mathbf{h} = (h_1, h_2)$. Como S_D^* é limitado inferiormente, então o conjunto $\mathbf{B}$ é

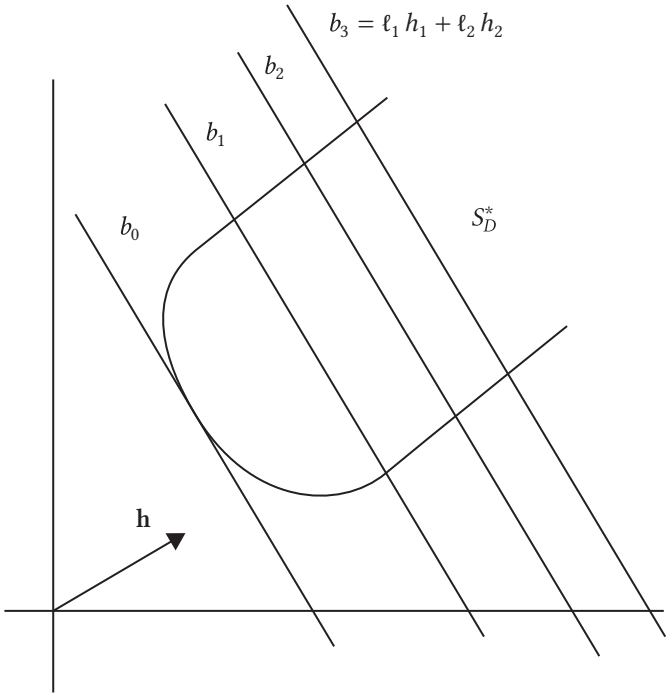


Fig. 4.6

minorado e tem ínfimo. Seja b_0 o ínfimo de **B**. Em qualquer sucessão de pontos, $\boldsymbol{\ell}^\nu$, $\nu = 1, 2, \ldots$, $\boldsymbol{\ell}^\nu \in S_D^*$, tal que,

$$\sum \ell_j^\nu h_j \to b_0,$$

o facto de ser $h_j > 0$ implica que para $j = 1, 2, \ldots, n$, as sucessões de números, ℓ_j^ν, $\nu = 1, 2, \ldots$, são limitadas superiormente. Existe, portanto,

$$\lim_{\nu \to \infty} \boldsymbol{\ell}^\nu = \boldsymbol{\ell}^0,$$

e $\sum \ell_j^0 h_j = b_0$, Ora, $\boldsymbol{\ell}^0 \in \Lambda(S_D^*)$: por uma parte, $\boldsymbol{\ell}^0 \in \overline{S}_D^*$, por ser ponto limite e $\{\boldsymbol{\ell}^0\} \subset Q_{\boldsymbol{\ell}^0} \cap \overline{S}_D^*$, por outra parte, $Q_{\boldsymbol{\ell}^0} \cap \overline{S}_D^* \cap \{\boldsymbol{\ell}^0\}$, porque se $\boldsymbol{\ell}' \in Q_{\boldsymbol{\ell}^0}$ e é diferente de $\boldsymbol{\ell}^0$, $\Sigma \ell_j' h_j < b_0$ e se $\boldsymbol{\ell}' \in \overline{S}_D^*$ haverá pontos $\boldsymbol{\ell} \in S_D^*$ para os quais $\Sigma \ell_j h_j < b_0$ contradizendo a hipótese de b_0 ser o ínfimo de **B**. Portanto, $\{\boldsymbol{\ell}^0\} = Q_{\boldsymbol{\ell}^0} \cap \overline{S}_D^*$ e $\boldsymbol{\ell}^0 \in \Lambda(S_D^*)$[5].

Como, por hipótese, S_D^* é fechado inferiormente, $\boldsymbol{\ell}^0 \in S_D^*$ e o mínimo de $\Sigma \ell_j h_j$ é obtido num ponto de S_D^*; $\boldsymbol{\ell}^0$ é, assim, ponto que corresponde a função de decisão Bayes contra **h** e a demonstração fica completa. □□

[5] Acaba de demonstrar-se que, em geral, conjunto não vazio, convexo e limitado inferiormente, tem fronteira inferior não vazia.

Exemplo 4.7 — Para mostrar que o Teorema 4.12 não pode alargar-se a toda e qualquer distribuição a priori, $h \in \mathcal{H}$, considere-se, com $n = 2$, o conjunto risco,

$$S_D^* = \{(\ell_1, \ell_2) \,:\, \ell_1\ell_2 \geqq 1,\ \ell_1 > 0\},$$

representado na Fig. 4.7. O conjunto S_D^* é convexo, limitado inferiormente e fechado inferiormente. Tome-se $\mathbf{h} = (1{,}0) \notin \mathcal{H}^+$; vem,

$$\ell_1 h_1 + \ell_2 h_2 = \ell_1,$$

e o risco mínimo é zero; contudo o mínimo não é atingido para qualquer $\boldsymbol{\ell} = (\ell_1, \ell_2) \in S_D^*$ o que mostra não existir solução Bayes contra $\mathbf{h}$.

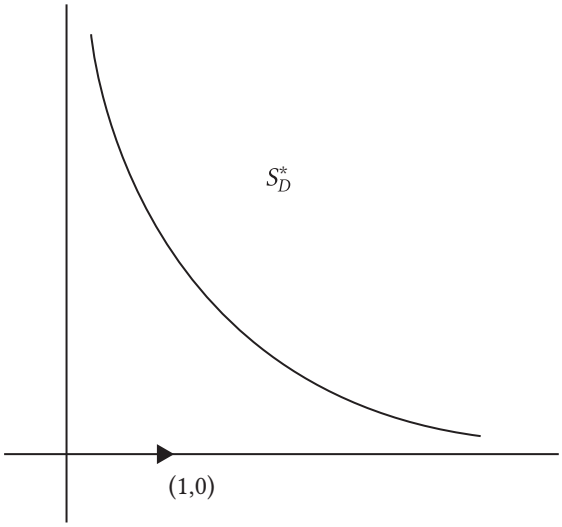


Fig. 4.7

□

Analisando a demonstração do teorema anterior verifica-se facilmente que se a hipótese «S_D^* limitado inferiormente» for substituída por «S_D^* limitado», existe função de decisão Bayes contra todo o $\mathbf{h} \in \mathcal{H}$.

Nas condições do teorema anterior tem-se ainda,

Teorema 4.13 — Se S_D^* é limitado inferiormente e fechado inferiormente, a classe de funções de decisão,

$$\mathfrak{C}_0 = \{\delta \in D^* \,:\, \boldsymbol{\ell}(\delta) \in \Lambda(S_D^*)\},$$

é completa mínima.

Dem. Mostra-se, primeiro, que $\mathcal{C}_0$ é completa. Seja $\delta \in D^*$, $\delta \in \mathcal{C}_0$; tem-se, $\boldsymbol{\ell}(\delta) \in S_D^*$ e $\boldsymbol{\ell}(\delta) \in \Lambda(S_D^*)$. Designe-se,

$$S^\Omega = Q_{\boldsymbol{\ell}(\delta)} \cap \overline{S}_D^*;$$

S^Ω é conjunto não vazio, convexo [por ser intersecção de conjuntos convexos ($\overline{S}_D^*$ é convexo por ser fecho de conjunto convexo)] e limitado inferiormente. Da primeira parte do teorema anterior pode concluir-se que todo o conjunto não vazio, convexo e limitado inferiormente, tem fronteira inferior não vazia; logo, $\Lambda(S^\Omega) \neq \emptyset$. Tome-se, $\boldsymbol{\ell} \in \Lambda(S^\Omega)$, isto é, $\boldsymbol{\ell}$ tal que, $\{\boldsymbol{\ell}\} = Q_{\boldsymbol{\ell}} \cap \overline{S}^\Omega$; verifica-se ser $\boldsymbol{\ell} \in Q_{\boldsymbol{\ell}(\delta)}$, pois,

$$\boldsymbol{\ell} \in \overline{S}^\Omega = \overline{Q_{\boldsymbol{\ell}(\delta)} \cap \overline{S}_D^*} \subset \overline{Q}_{\boldsymbol{\ell}(\delta)} = Q_{\boldsymbol{\ell}(\delta)}.$$

Finalmente,

$$\begin{aligned} \{\boldsymbol{\ell}\} = Q_{\boldsymbol{\ell}} \cap \overline{S}^\Omega &= Q_{\boldsymbol{\ell}} \cap \overline{[Q_{\boldsymbol{\ell}(\delta)} \cap \overline{S}_D^*]} \\ &= Q_{\boldsymbol{\ell}} \cap [Q_{\boldsymbol{\ell}(\delta)} \cap \overline{S}_D^*] = Q_{\boldsymbol{\ell}} \cap \overline{S}_D^*. \end{aligned}$$

Assim, como S_D^* é fechado inferiormente, existe $\delta' \in D^*$ tal que $\boldsymbol{\ell}(\delta') = \boldsymbol{\ell}$, dominando estritamente δ, pois $\boldsymbol{\ell} \in [Q_{\boldsymbol{\ell}(\delta)} - \{\boldsymbol{\ell}(\delta)\}]$. A classe $\mathcal{C}_0$ é pois completa.

Pelo Teorema 4.11, qualquer $\delta \in \mathcal{C}_0$ é admissível. Consequentemente nenhuma subclasse própria de $\mathcal{C}_0$ pode ser completa uma vez que essa subclasse não conteria algumas funções de decisão admissíveis contradizendo $\mathcal{A} \in \mathcal{C}$ para toda a classe completa $\mathcal{C}$. Portanto, $\mathcal{C}_0$ é classe completa mínima. □□

Na Fig 4.8 ilustram-se alguns passos do teorema que acaba de demonstrar-se.

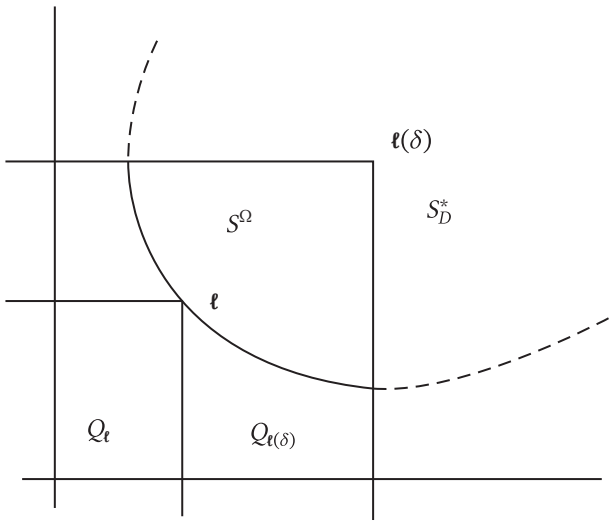


Fig. 4.8

Dos Teoremas 4.3 e 4.13 sai o seguinte,

Teorema 4.14 — A classe,

$$\mathcal{C}_0 = \{\delta \in D^* \,:\, \boldsymbol{\ell}(\delta) \in \Lambda(S_D^*)\},$$

coincide com a classe $\mathcal{A}$. □□

Para prosseguir na apresentação de alguns teoremas fundamentais da teoria da decisão estatística é necessário introduzir o teorema do hiperplano de separação [a demonstração pode ver-se em Blackwell e Girshick (1954) ou em Ferguson (1967)]:

Teorema 4.15 — Se U e V são conjuntos convexos e disjuntos de $\mathbf{R}^n$ existe um hiperplano de separação, isto é, um hiperplano $\mathbf{k} \cdot \boldsymbol{\ell} = c$ tal que $\mathbf{k} \cdot \boldsymbol{\ell} \leqq c$ quando $\boldsymbol{\ell} \in U$ e $\mathbf{k} \cdot \boldsymbol{\ell} \geqq c$ quando $\boldsymbol{\ell} \in V$. □□

Teorema 4.16 — Tem-se $\mathcal{A} \subset \mathcal{B}$.

Dem. Se $\delta \in \mathcal{A}$, $\{\boldsymbol{\ell}(\delta)\} = Q_{\boldsymbol{\ell}(\delta)} \cap S_D^*$; os conjuntos $Q_{\boldsymbol{\ell}(\delta)} - \{\boldsymbol{\ell}(\delta)\}$ e S_D^* são convexos e disjuntos. Pelo teorema anterior existe um hiperplano de separação, isto é, existe em $\mathbf{R}^n$ um vector não nulo, seja $\mathbf{h}$, tal que,

$$\sum z_j h_j \leqq \sum w_j h_j \text{ para todo o } \mathbf{z} \in Q_{\boldsymbol{\ell}(\delta)} - \{\boldsymbol{\ell}(\delta)\} \text{ e } \mathbf{w} \in S_D^*.$$

Assim, para todo o $j = 1, 2, \ldots, n$, tem de ser $h_j \geqq 0$, pois, caso contrário, pode tomar-se um z_j negativo e suficientemente grande em valor absoluto de modo a sair $\sum z_j h_j > \sum w_j h_j$. Normalizando $\mathbf{h}$ para que seja $\sum h_j = 1$, vem, $\mathbf{h} \in \mathcal{H}$ e, como $\boldsymbol{\ell}(\delta)$ pertence ao hiperplano de separação,

$$\boldsymbol{\ell}(\delta) \cdot \mathbf{h} \leqq \mathbf{w} \cdot \mathbf{h},$$

para todo o $\mathbf{w} \in S_D^*$. Logo, δ é Bayes contra $\mathbf{h}$ e $\delta \in \mathcal{B}$. □□

Se a subclasse de uma classe de funções de decisão é completa é evidente que a própria classe também é completa. Como, pelo Teorema 4.16, $\mathcal{A} \subset \mathcal{B}$, e, pelo Teorema 4.13, $\mathcal{A}$ é classe completa (mínima), tem-se que $\mathcal{B}$ é classe completa. Fica assim demonstrado o Teorema da Classe Completa,

Teorema 4.17 — Se S_D^* é limitado inferiormente e fechado inferiormente, a classe $\mathcal{B}$ é completa e as funções de decisão admissíveis formam classe completa mínima. □□

Exemplo 4.8 — Para reforçar a ideia de que o Teorema 4.16 se refere exclusivamente ao caso Θ finito considere-se a matriz risco,

	δ_1	δ_2	δ_3	δ_4	δ_5	...
θ_1	1/2	0	0	0	0	...
θ_2	1/2	1	0	0	0	...
θ_3	1/2	1	1	0	0	...
θ_4	1/2	1	1	1	0	...
θ_5	1/2	1	1	1	1	...
...	.	.	.	.	.	...

Note-se que δ_1 é a única função de decisão admissível. Por outro lado, δ_2 é dominada estritamente por δ_3, δ_3 por δ_4, e assim por diante. No entanto, δ_1 não é Bayes contra nenhum $\mathbf{h}$, pois, $R(\mathbf{h}, \delta_1) = 1/2$, e qualquer que seja $\mathbf{h}$ é sempre possível encontrar um δ_i verificando,

$$R(\mathbf{h}, \delta_i) < R(\mathbf{h}, \delta_1) = 1/2.$$

Para tanto basta tomar i suficientemente grande para ser $\sum_{j=i}^{\infty} h_j < 1/2$. □

Exemplo 4.9 — Seja $\Theta = \{\theta_1, \theta_2\}$ e S_D^* tal como representado na Fig. 4.9. A classe $\mathcal{B}^+$ é formada pelas funções de decisão que correspondem aos pontos dos segmentos $\overline{bc}$ e $\overline{cd}$, o mesmo sucedendo com $\mathcal{A}$. Logo, tem-se, $\mathcal{B}^+ = \mathcal{A}$. Os pontos do segmento $\overline{ab}$ correspondem a funções de decisão que são Bayes contra (1,0) e os pontos $\overline{de}$ a Bayes contra (0,1). Portanto, a classe $\mathcal{B}$ compreende as funções de decisão a que correspondem pontos dos segmentos $\overline{ab}$, $\overline{bc}$, $\overline{cd}$ e $\overline{de}$. Evidentemente, $\mathcal{B}^+ = \mathcal{A} \subset \mathcal{B}$.

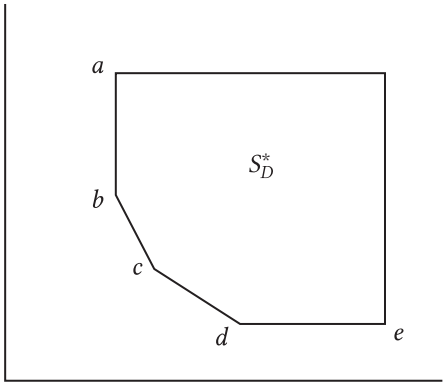


Fig. 4.9

□

Exemplo 4.10 — Suponha-se S_D^* como representado na Fig. 4.10. O arco bc, sem os pontos b e c corresponde a $\mathcal{B}^+$. O arco bc corresponde a $\mathcal{A}$. O conjunto formado pelo arco bc e pelo segmento $\overline{cd}$ corresponde, a $\mathcal{B}$. Portanto, $\mathcal{B}^+ \subset \mathcal{A} \subset \mathcal{B}$.

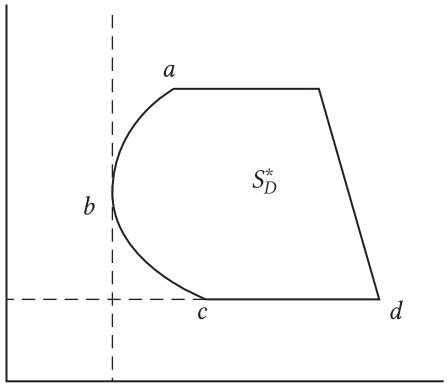


Fig. 4.10

□

Exemplo 4.11 — Seja $\boldsymbol{\ell} = (\ell_1, \ell_2)$ e considerem-se os conjuntos,

$$P = \{\boldsymbol{\ell} : (\ell_1 - 8)^2 + (\ell_2 - 8)^2 \leqq 100\}, \quad Q = \{\boldsymbol{\ell} : 0 \leqq \ell_1 \leqq 4,\ 0 \leqq \ell_2 \leqq 4\},$$
$$R = \{\boldsymbol{\ell} : 2 \leqq \ell_1 \leqq 3,\ \ell_2 = 0\}, \quad S = \{\boldsymbol{\ell} : \ell_1 = 0,\ \ell_2 = 2\}.$$

Tome-se,

$$S_D^* = P \cap Q - (R \cup S),$$

conjunto que se representa na Fig. 4.11.

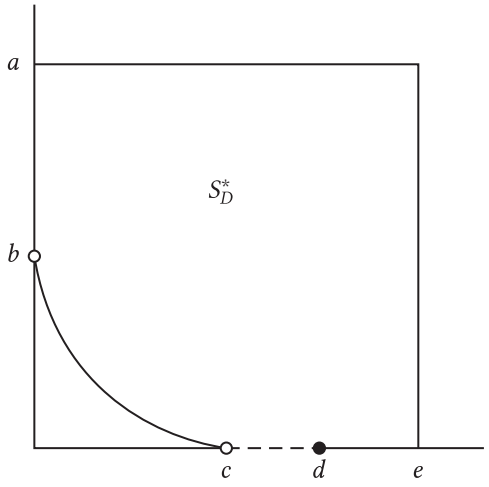


Fig. 4.11

Observe-se que S_D^* não é fechado pois não contém os pontos b e c e o segmento $\overline{cd}$ embora contenha o ponto d. O arco bc, com exclusão dos pontos b e c, corresponde a $\mathcal{B}^+$. A classe $\mathcal{A}$ compreende $\mathcal{B}^+$ mais a função de decisão correspondente ao ponto d. A classe $\mathcal{B}$ compreende os seguintes pontos: segmento $\overline{ab}$ (com exclusão do ponto b), arco bc (com exclusão dos pontos b e c) e segmento $\overline{de}$. A classe $\mathcal{A}$ não é completa visto que nenhum dos pontos do segmento $\overline{ab}$ (sem o ponto b) é dominado estritamente por qualquer ponto de $\mathcal{A}$. Por outro lado, não existe classe completa mínima porque $\mathcal{A}$ não é classe completa [Blackwell e Girshick (1954)]. □

A presente secção é muito incompleta; em particular parte-se da hipótese restritiva de que Θ é finito. Infelizmente, o tratamento de casos mais gerais não cabe no âmbito das presentes notas. A título complementar referem-se apenas duas proposições.

O primeiro resultado, devido a Brown [veja-se a citação em Berger (1980) visto referir-se a trabalho não publicado], exige a continuidade das funções risco.

Teorema 4.18 — Verificando-se as condições,

i) A e Θ são conjuntos limitados e fechados de espaços euclideanos;
ii) $L(\theta, a)$ é função contínua de a para todo o $\theta \in \Theta$;
iii) todas as funções de decisão possuem função risco contínua, as funções de decisão Bayes formam uma classe completa. □□

Esta proposição mostra a importância que tem investigar em que circunstâncias as funções risco são contínuas. Quando é possível admitir que a função perca é limitada [além de hipóteses menos restritivas como sejam $\Theta \subset \mathbf{R}^m$, $L(\theta, a)$ contínua em θ para todo o $a \in A$ e $f(x \mid \theta)$ contínua em θ para todo o $x \in \mathcal{X}$] sucede que todas as funções de decisão possuem função risco contínua. Como muitas vezes tal não é possível, sobretudo quando Θ não é limitado [por exemplo, quando $\Theta \equiv \mathbf{R}$ e se trabalha com função perca quadrática], há que restringir a densidade, $f(x \mid \theta)$, para garantir a continuidade desejada.

O segundo resultado, devido a Wald (1950), exige certas propriedades topológicas do espaço de funções de decisão, D^*, e do espaço do parâmetro, Θ.

Um espaço de funções de decisão, D^*, diz-se fracamente compacto em relação à função risco, $R(\theta, \delta)$, se para cada sucessão de funções de decisão,

$$\{\delta_\nu\}, \quad \nu = 1, 2, 3, \ldots, \delta_\nu \in D^*,$$

existe $\delta' \in D^*$ tal que,

$$\lim_{\nu \to \infty} R(\theta, \delta_\nu) \geqq R(\theta, \delta') \text{ para todo o } \theta \in \Theta.$$

Por outro lado, o espaço de estados, Θ, diz-se fracamente semi-separável no sentido de Wald se e somente se existe uma sucessão de pontos, $\{\theta_\nu\}$, $\nu = 1, 2, 3, ..., \theta_\nu \in \Theta$, tal que,

$$\overline{\lim_{\nu \to \infty}} R(\delta_\nu, \delta) \geqq R(\theta, \delta) \text{ para todo o } \theta \in \Theta \text{ e todo o } \delta \in D^*.$$

Tem-se, então,

Teorema 4.19 — Se para cada $\delta \in D^*$, $R(\theta, \delta)$ é função limitada de θ, se D^* é fracamente compacto e se Θ é fracamente semi-separável, então, a classe $\mathcal{C}$ é completa e a classe de funções de decisão Bayes, próprias e generalizadas, é essencialmente completa. □□

4.5 Soluções minimax: existência e cálculo

O princípio minimax foi introduzido na teoria dos jogos estratégicos. A respectiva abordagem torna-se mais típica e, possivelmente, mais intuitiva, quando se emprega a linguagem própria desta teoria.

Considere-se um jogo de duas pessoas onde intervém a «natureza» — jogador J_1 — e o «decisor» — jogador J_2. Suponha-se que se trata de jogo com soma nula, isto é, o que um jogador ganha é igual ao que o outro perde. A função risco, $R(h, \delta^*)$, pode interpretar-se como representando pagamento (= perca esperada) a fazer pelo decisor à natureza quando o decisor emprega a estratégia (= função de decisão) $\delta^* \in D^*$ e a natureza emprega a estratégia (= distribuição a priori) $h \in \mathcal{H}$. Nota-se que a incerteza do jogo resulta de qualquer dos jogadores fazer a sua escolha na ignorância da escolha feita pelo outro.

A quantidade,

$$\inf_{\delta^*} \sup_h R(h, \delta^*) = \overline{V}, \tag{4.11}$$

chama-se valor superior do jogo; como já foi dito, se,

$$\sup_h R(h, \tilde{\delta}^*) = \inf_{\delta^*} \sup_h R(h, \delta^*), \tag{4.12}$$

diz-se que $\tilde{\delta}^*$ é uma estratégia ou função de decisão minimax (de J_2).

A quantidade,

$$\sup_h \inf_{\delta^*} R(h, \delta^*) = \underline{V}, \tag{4.13}$$

chama-se valor inferior do jogo; se,

$$\inf_{\delta^*} R(\tilde{h}, \delta^*) = \sup_h \inf_{\delta^*} R(h, \delta^*), \tag{4.14}$$

diz-se que $\tilde{h}$ é uma estratégia maximin (de J_1) ou distribuição a priori mais desfavorável.

Tem-se sempre,

$$\underline{V} \leqq \overline{V}, \tag{4.15}$$

porquanto com quaisquer h_0 e δ_0,

$$\inf_{\delta^*} R(h_0, \delta^*) \leqq \sup_h R(h, \delta_0),$$

e ao tomar $\sup_{h_0}$ no primeiro membro e $\inf_{\delta_0}$ no segundo a desigualdade mantém-se.

Quando, $\underline{V} = \overline{V}$, diz-se que o jogo tem valor $V = \underline{V} = \overline{V}$. O problema central da teoria dos jogos de duas pessoas com soma nula é investigar em que condições um jogo tem valor e em que condições existem estratégias maximin e minimax. No caso Θ finito o teorema seguinte é uma resposta,

Teorema 4.20 — Se S_D^* é limitado inferiormente, o jogo tem valor e existe uma distribuição a priori mais desfavorável, $\tilde{h}$. Se adicionalmente S_D^* é fechado inferiormente, existe uma função de decisão minimax, $\tilde{\delta}$, que é admissível e Bayes contra $\tilde{h}$.

Dem. (*a*) Começa por mostrar-se que $\underline{V} \geqq \overline{V}$, isto é, que o jogo tem valor. Seja $Q_{\mathbf{c}}$ o ortante inferior no ponto $\mathbf{c} = (c, c, \ldots, c)$; como S_D^* é limitado inferiormente, existe,

$$c_0 = \inf\{c : Q_{\mathbf{c}} \cap S_D^* \neq \emptyset\},$$

como se indica na Fig. 4.12 para um caso particular.

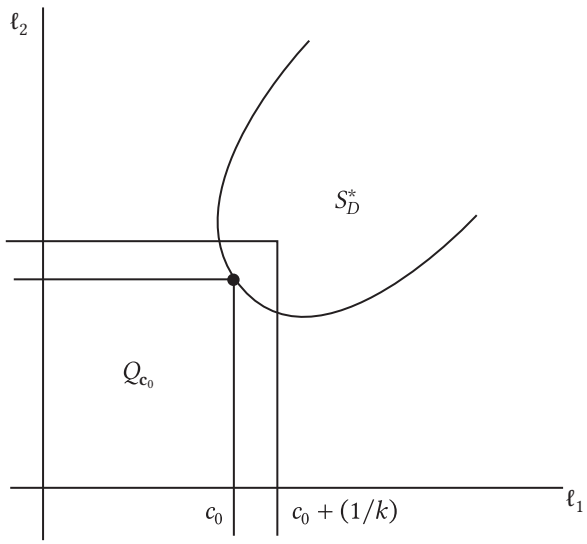


Fig. 4.12

Nestas condições, qualquer que seja $k > 0$, existe uma função de decisão, δ_k, tal que,

$$R(\theta_j, \delta_k) \leqq c_0 + \frac{1}{k}, \tag{4.16}$$

para $j = 1, 2, \ldots, n$. Portanto,

$$R(h, \delta_k) \leqq c_0 + \frac{1}{k},$$

para todo o $h \in \mathcal{H}$ e,

$$\sup_h R(h, \delta_k) \leqq c_0 + \frac{1}{k},$$

o que implica $\overline{V} \leqq c_0$, quando se toma o ínfimo em D^*.

Designe, $Q^o_{c_0}$, o interior de Q_{c_0}; note-se que $Q^o_{c_0}$ e S^*_D são conjuntos convexos e disjuntos. Logo, pelo Teorema 4.15, deve existir um hiperplano de separação, isto é, um hiperplano, $\sum \ell_j \eta_j = b$, tal que, $\sum z_j \eta_j \leqq b$ quando $\mathbf{z} \in Q^o_{c_0}$ e $\sum w_j \eta_j \geqq b$ quando $\mathbf{w} \in S^*_D$. O vector, $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)$, não pode ter qualquer das componentes negativa. Por exemplo, admitindo que se tem $\eta_j < 0$, tomando $\mathbf{z} \in Q^o_{c_0}$ e fazendo $z_j \to -\infty$ com as outras componentes fixas, conclui-se que a expressão $\sum z_j \eta_j \to \infty$, o que é uma contradição.

O vector, seja $\tilde{h}$, obtido por normalização de $\boldsymbol{\eta}$ de forma a sair $\sum \tilde{h}_j = 1$, pode tomar-se como distribuição a priori. Dado que $\sum z_j \tilde{h}_j \leqq b'$ $[b' = b/\sum \eta_j]$ fazendo $\mathbf{z} \to (c_0, c_0, \ldots, c_0)$, tem-se $c_0 \leqq b'$ e para toda a função de decisão $\delta \in D^*$,

$$R(\tilde{h}, \delta) = \sum_{j=1}^{n} R(\theta_j, \delta)\tilde{h}_j \geqq b' \geqq c_0; \tag{4.17}$$

logo,

$$\underline{V} = \sup_h \inf_{\delta^*} R(h, \delta^*) \geqq \inf_{\delta^*} R(\tilde{h}, \delta^*) \geqq c_0 \geqq \overline{V}.$$

Como, (4.15), $\underline{V} \leqq \overline{V}$, conclui-se,

$$\inf_{\delta^*} R(\tilde{h}, \delta^*) = \sup_h \inf_{\delta^*} R(h, \delta^*);$$

assim, o jogo tem valor e $\tilde{h}$ é uma distribuição a priori mais desfavorável.

(*b*) Suponha-se adicionalmente que S^*_D é fechado inferiormente. Reconsidere-se a função de decisão δ_k nas condições (4.16) e tome-se $\boldsymbol{\ell}(\delta_k)$, na notação habitual, $\boldsymbol{\ell}(\delta_k) = [R(\theta_1, \delta_k), R(\theta_2, \delta_k), \ldots, R(\theta_n, \delta_k)]$. Porque os vectores da sucessão, $\{\boldsymbol{\ell}(\delta_k)\}$, têm componentes limitadas superiormente [por (4.16)] e inferiormente [por hipótese], pelo Teorema de Bolzano-Weierstrass existe um ponto limite, $\boldsymbol{\ell}^o$. Evidentemente, $\boldsymbol{\ell}^o \in \overline{S}^*_D$, e, assim, $Q_{\boldsymbol{\ell}^o} \cap \overline{S}^*_D \neq \emptyset$. Por o conjunto, $Q_{\boldsymbol{\ell}^o} \cap \overline{S}^*_D$, ser limitado inferiormente,

vem $\Lambda(Q_{\ell^o} \cap S_D^*) \neq \emptyset$, conforme se assinalou ao demonstrar o Teorema 4.12. Tome-se o ponto $\tilde{\boldsymbol{\ell}} \in \Lambda(Q_{\ell^o} \cap \overline{S}_D^*)$; como, por definição,

$$Q_{\tilde{\ell}} \cap [Q_{\ell^o} \cap \overline{S}_D^*] = \{\tilde{\boldsymbol{\ell}}\},$$

também $\tilde{\boldsymbol{\ell}} \in Q_{\ell^o}$ e, por ser necessariamente $Q_{\tilde{\ell}} \subset Q_{\ell^o}$, $Q_{\tilde{\ell}} \cap \overline{S}_D^* = \{\tilde{\boldsymbol{\ell}}\}$, tem-se $\tilde{\boldsymbol{\ell}} \in \Lambda(S_D^*)$. Como S_D^* é fechado inferiormente por hipótese, também $\tilde{\boldsymbol{\ell}} \in S_D^*$. Qualquer $\tilde{\delta} \in D^*$, tal que, $[R(\theta_1, \tilde{\delta}), R(\theta_2, \tilde{\delta}), \ldots, R(\theta_n, \tilde{\delta})] = \tilde{\boldsymbol{\ell}}$, é admissível [Teorema 4.11] e satisfaz, $R(h, \delta) \leqq c_0$, porque se tem $R(\theta_j, \delta) \leqq c_0$, para $j = 1, 2, \ldots, n$ [$\tilde{\delta}$ obtém-se da sucessão (ou de uma subsucessão) dos δ_k quando $k \to \infty$]. Além disso, de (4.17), substituindo $\tilde{\delta}$ no primeiro membro sai $R(\tilde{h}, \tilde{\delta}) \geqq c_0$, e logo vem $R(\tilde{h}, \tilde{\delta}) = c_0 = \inf_{\delta^*} R(\tilde{h}, \delta^*)$, o que mostra que $\tilde{\delta}$ é Bayes contra $\tilde{h}$ e obviamente minimax. □□

O caso seguinte [Ferguson (1967)] mostra que se Θ não é finito a proposição demonstrada — Teorema Minimax — não é necessariamente verdadeira.

Exemplo 4.12 — Suponha-se, $\Theta = D = \{1, 2, 3, \ldots,\}$, e,

$$R(\theta, \delta) = \begin{cases} +1 & \text{se} \quad \delta < \theta, \\ 0 & \text{se} \quad \delta = \theta, \\ -1 & \text{se} \quad \delta > \theta. \end{cases}$$

Tem-se, para qualquer $h \in \mathcal{H}$,

$$R(h, \delta) = \left[\sum_{j>\delta} h_j\right] - \left[\sum_{j<\delta} h_j\right],$$

donde, com h fixo,

$$\inf_\delta R(h, \delta) = \lim_{\delta \to \infty} R(h, \delta) = -1,$$

e,

$$\underline{V} = \sup_h \inf_\delta R(h, \delta) = -1.$$

De modo análogo se mostrava que,

$$\overline{V} = \inf_\delta \sup_h R(h, \delta) = +1.$$

Como $\underline{V} < \overline{V}$ o jogo não tem valor e não existem estratégias minimax ou maximin. □

O Teorema Minimax pode alargar-se ao caso Θ infinito. Uma das formas de o fazer é através do conceito de semi-continuidade inferior. Uma função real, $\psi(z)$, é semi-continua inferiormente, se para todo o número real, c, o conjunto, $\{z : \psi(z) > c\}$, é aberto.

Teorema 4.21 — Seja $\mathcal{C} \subset D^*$ uma classe essencialmente completa de funções de decisão. Se existe em $\mathcal{C}$ uma topologia tal que (i) $\mathcal{C}$ é compacto e (ii) $R(\theta, \delta)$ é semi--continua inferiormente em $\delta \in \mathcal{C}$ para todo o $\theta \in \Theta$, então, $\underline{V} = \overline{V}$ e existe uma função de decisão minimax. □□

A demonstração pode ver-se em Ferguson (1967). É interessante notar que a partir das hipóteses deste teorema pode demonstrar-se que a classe das funções de decisão quase-Bayes ou Bayes por extensão é essencialmente completa. Como é evidente, as proposições libertas do condicionalismo Θ finito revestem-se de maior complexidade.

Quando Θ e D são finitos [situação formalmente análoga aquela em que Θ e A são finitos e o pagamento ou perca esperada é dado pela expressão, $R(h, a^*) = E\{L(\theta, a^*)\}$, com $L(\theta, a^*)$ definido por (3.2)] fala-se em jogo rectangular ou finito. Sendo S_D conjunto finito, o respectivo ambiente convexo, S_D^* é um hiperpoliedro convexo, portanto, conjunto limitado e fechado. A proposição seguinte é simples corolário do Teorema 4.20,

Corolário 4.1 — Todo o jogo finito tem valor e estratégias maximin e minimax. □□

Quando Θ e D são finitos a determinação do valor do jogo, da estratégia maximin [há uma pelo menos e havendo mais do que uma há uma infinidade] e da estratégia minimax [idem] é problema de matemática de solução conhecida, nomeadamente através da relação com a programação linear, com cálculo numérico facilmente realizado em computador quando o cardinal dos conjuntos o exija [vejam-se Blackwell e Girshick (1954) e Gass (1969)]. O problema não é aqui aprofundado, já por ser algo extenso, já por não ser especialmente relevante para a decisão estatística. Passa, por isso, a estudar-se com maior ênfase a determinação das funções de decisão minimax em contexto mais geral.

Quando a classe de funções de decisão, D^*, é já de si reduzida, ou pode reduzir--se por eliminação das inadmissíveis, torna-se às vezes possível utilizar um método directo de cálculo a partir da definição. Um exemplo vai encontrar-se a seguir ao Teorema 7.4.

Os métodos mais fáceis para determinar funções de decisão minimax têm como ponto de partida as funções de decisão Bayes. Um desses métodos baseia-se na ideia de que é possível ao decisor conhecer ou conjecturar a distribuição a priori mais desfavorável [a estratégia mais temível da «natureza» ou maximin]. Isso sucede, nomeadamente, quando $\mathcal{H}$ tem estrutura pouco complicada e D^*, pelo contrário, tem estrutura intrincada a ponto de tornar difícil a pesquisa directa da solução minimax. Uma vez vislumbrada a distribuição mais desfavorável, $\tilde{h}$, se for possível determinar a função de decisão, $\tilde{\delta}$, Bayes contra $\tilde{h}$, então, $\tilde{\delta}$ é função de decisão minimax. Poderoso auxiliar no contexto é o seguinte,

Teorema 4.22 — Se $\tilde{\delta}$ é Bayes contra $\tilde{h}$ e, para todo o $\theta \in \Theta$,

$$R(\theta, \tilde{\delta}) \leqq R(\tilde{h}, \tilde{\delta}), \tag{4.18}$$

então $\tilde{\delta}$ é minimax e $\tilde{h}$ é a distribuição a priori mais desfavorável.

Dem. De (4.18) e do facto de $\tilde{\delta}$ ser Bayes contra $\tilde{h}$, resulta, para todo $h \in \mathcal{H}$,

$$R(h, \tilde{\delta}) \leqq R(\tilde{h}, \tilde{\delta}) = \inf_{\delta^*} R(\tilde{h}, \delta^*),$$

donde,

$$\sup_h R(h, \tilde{\delta}) \leqq R(\tilde{h}, \tilde{\delta}) = \inf_{\delta^*} R(\tilde{h}, \delta^*),$$

e, ainda,

$$\inf_{\delta^*} \sup_h R(h, \delta^*) \leqq \sup_h R(h, \tilde{\delta}) \leqq \sup_h \inf_{\delta^*} R(h, \delta^*).$$

Por outro lado — veja-se (4.15) — verifica-se sempre,

$$\sup_h \inf_{\delta^*} R(h, \delta^*) \leqq \inf_{\delta^*} \sup_h R(h, \delta^*);$$

logo,

$$\sup_h R(h, \tilde{\delta}) = \inf_{\delta^*} \sup_h R(h, \delta^*),$$

e a demonstração fica completa. □□

Exemplo 4.13 — Apesar de esquemático o problema tem algum interesse [Ferguson (1967)]. Seja, $\Theta = \{1/3, 2/3\}$, $A = (-\infty, +\infty)$, $L(\theta, a) = (\theta - a)^2$. A experiência consiste em observar a variável aleatória, X, assumindo dois valores, x_1 e x_2, com $P(X = x_1) = f(x_1 \,|\, \theta) = \theta$ e $P(X = x_2) = f(x_2 \,|\, \theta) = 1 - \theta$.

O conjunto de funções de decisão puras, D, pode identificar-se com $\mathbf{R}^2$,

$$D = \{\mathbf{z} = (z_1, z_2) \,:\, z_1 = \delta(x_1), z_2 = \delta(x_2)\};$$

um palpite razoável sobre a distribuição a priori mais desfavorável é o que aponta para $\tilde{\mathbf{h}} = (\tilde{h}_1, \tilde{h}_2) = (1/2, 1/2)$. O risco associado com cada $\mathbf{z} \in D$ é,

$$\begin{aligned} R(\tilde{h}, \mathbf{z}) &= R(\theta_1, \mathbf{z})\tilde{h}_1 + R(\theta_2, \mathbf{z})\tilde{h}_2 = R(1/3, \mathbf{z})(1/2) + R(2/3, \mathbf{z})(1/2) \\ &= \left[\frac{1}{3}\left(\frac{1}{3} - z_1\right)^2 + \frac{2}{3}\left(\frac{1}{3} - z_2\right)^2\right]\frac{1}{2} + \left[\frac{2}{3}\left(\frac{2}{3} - z_1\right)^2 + \frac{1}{3}\left(\frac{2}{3} - z_2\right)^2\right]\frac{1}{2}; \end{aligned}$$

esta expressão é mínima para $z_1 = 5/9$ e $z_2 = 4/9$, quer dizer, a função de decisão Bayes contra $\tilde{h} = (1/2, 1/2)$ é $\tilde{\mathbf{z}} = (5/9, 4/9)$. Voltando à expressão de $R(\tilde{h}, \tilde{\mathbf{z}})$, tem-se,

$$R(\tilde{h}, \tilde{\mathbf{z}}) = R(1/3, \tilde{\mathbf{z}}) = R(2/3, \tilde{\mathbf{z}}) = 2/81;$$

a condição (4.18) é satisfeita; $\tilde{\mathbf{z}}$ é minimax e $\tilde{h}$ maximin. □

Diz-se que δ^* é função de decisão igualadora se $R(\theta, \delta^*) = \rho$ (constante) para todo o $\theta \in \Theta$. Olhando a Fig. 4.13, traçada para Θ finito, $n = 2$, compreende-se que na situação apresentada (mas, não, por exemplo, quando se está nos dois primeiros casos da Fig. 3.8) uma função de decisão igualadora para ser minimax carece de propriedades adicionais.

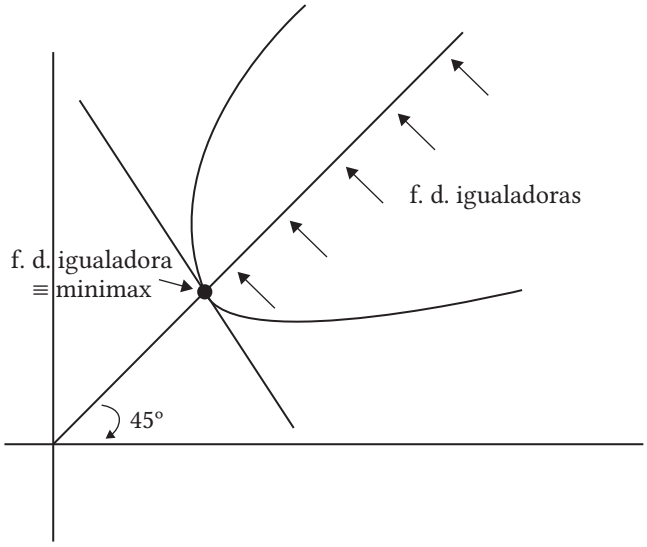


Fig. 4.13

Como é intuitivo uma propriedade adicional que parece garantir que uma função de decisão igualadora seja minimax é a admissibilidade. O teorema seguinte, além de estabelecer uma condição de admissibilidade, mostra que essa conjectura é verdadeira,

Teorema 4.23 — Se $\tilde{\delta}$ é a única função de decisão minimax, $\tilde{\delta}$ é admissível. Se $\tilde{\delta}$ é função de decisão igualadora admissível, $\tilde{\delta}$ é minimax.

Dem. 1) Para provar a primeira proposição suponha-se que $\tilde{\delta}$, única função de decisão minimax, não é admissível. Existe, então, $\delta \in D^*$, tal que,

$$R(\theta, \delta) \leqq R(\theta, \tilde{\delta}) \text{ para todo o } \theta \in \Theta,$$

e desigualdade estrita para algum $\theta \in \Theta$. Consequentemente,

$$\sup_\theta R(\theta, \delta) \leqq \sup_\theta R(\theta, \tilde{\delta}),$$

e δ seria também minimax o que é uma contradição.

2) Para provar a segunda proposição, suponha-se que $\tilde{\delta}$ nas condições indicadas não é minimax. Existe então uma função de decisão, $\delta' \in D^*$, tal que,

$$\sup_\theta R(\theta, \delta') < \sup_\theta R(\theta, \tilde{\delta}).$$

No entanto, como $R(\theta, \tilde{\delta}) = \rho$ (constante) para todo o $\theta \in \Theta$, esta desigualdade implica,

$$R(\theta, \delta') < R(\theta, \tilde{\delta}) \text{ para todo o } \theta \in \Theta.$$

Assim, δ' domina estritamente $\tilde{\delta}$ o que é uma contradição. □□

Não é preciso ser tão exigente para garantir que uma função de decisão igualadora seja minimax. Basta pedir que seja igualadora e Bayes.

Teorema 4.24 — Se para $\tilde{\delta} \in D^*$, $R(\theta, \tilde{\delta}) = \rho$, constante, para todo o $\theta \in \Theta$, e se existe $\tilde{h} \in \mathcal{H}$ contra o qual $\tilde{\delta}$ é Bayes, então $\tilde{\delta}$ é minimax e $\tilde{h}$ é maximin.

Dem. Sendo $R(\theta, \tilde{\delta}) = \rho$ para $\theta \in \Theta$ e $\tilde{\delta}$ Bayes contra $\tilde{h}$, vem,

$$\rho = \int_\Theta R(\theta, \tilde{\delta}) \tilde{h}(\theta)\, d\theta = R(\tilde{h}, \tilde{\delta}) = \inf_{\delta^*} R(\tilde{h}, \delta^*);$$

por outro lado,

$$\inf_{\delta^*} R(\tilde{h}, \delta^*) \leqq \sup_h \inf_{\delta^*} R(h, \delta^*) \leqq \inf_{\delta^*} \sup_h R(h, \delta^*) \leqq \sup_h R(h, \tilde{\delta}) = \rho,$$

e a demonstração fica completa. □□

Mais ainda, para que função de decisão igualadora seja minimax é suficiente pedir que seja também quase-Bayes ou Bayes por extensão [vejam-se Ferguson (1967) e Berger (1985) e comparem-se os Exs. 3.15 e 4.16].

Exemplo 4.14 — Suponha-se que a variável aleatória, X, tem função de probabilidade Binomial,

$$f(x \mid \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}, x = 0, 1, \ldots, N, \quad 0 < \theta < 1.$$

Considere-se a estimação de θ, $A = \Theta = (0, 1)$, com função perca quadrática,

$$L(\theta, a) = (\theta - a)^2,$$

e procure-se a função de decisão minimax.

Para aplicar o Teorema 4.24 há que investigar a existência de função de decisão com risco constante.

Ensaiando,

$$\delta(x) = \alpha x + \beta,$$

tem-se,

$$\begin{aligned} R(\theta, \delta) &= E\{(\theta - \alpha X - \beta)^2\} \\ &= [(\alpha N - 1)^2 - \alpha^2 N]\theta^2 + [\alpha^2 N + 2\beta(\alpha N - 1)]\theta + \beta^2, \end{aligned}$$

notando que,

$$\begin{aligned} E\{X\} &= N\theta, \\ E\{X^2\} &= N\theta(1-\theta) + N^2\theta^2. \end{aligned}$$

O risco, $R(\theta, \delta)$, é constante quando se verificam as duas relações seguintes:

$$\begin{aligned} (\alpha N - 1)^2 - \alpha^2 N &= 0, \\ \alpha^2 N + 2\beta(\alpha N - 1) &= 0, \end{aligned}$$

isto é, se $\alpha = \alpha_0$ e $\beta = \beta_0$, com,

$$\alpha_0 = [\sqrt{N}(1 + \sqrt{N})]^{-1}, \beta_0 = [2(1 + \sqrt{N})]^{-1};$$

de facto, fazendo a respectiva substituição, vem, $R(\theta, \delta_0) = [2(1 + \sqrt{N})]^{-2}$, onde, $\delta_0(x) = \alpha_0 x + \beta_0$. Será δ_0 Bayes contra alguma distribuição a priori?

Tome-se,

$$h(\theta) = \left[\frac{1}{B(\kappa, \omega)}\right] \theta^{\kappa - 1}(1-\theta)^{\omega - 1}, \quad 0 < \theta < 1,$$

donde sai, como facilmente se verifica,

$$h(\theta \mid x) = \left[\frac{1}{B(\kappa + x, \omega + N - x)}\right] \theta^{\kappa + x - 1}(1-\theta)^{\omega + N - x - 1}, \quad 0 < \theta < 1.$$

O risco a posteriori,

$$r_x(\delta) = \int_0^1 (\theta - \delta)^2 h(\theta \mid x)\, d\theta,$$

é minimizado para

$$\delta_h(x) = E\{\theta \mid x\} = \frac{\kappa + x}{\kappa + \omega + N}.$$

Igualando,

$$(\kappa + \omega + N)^{-1} = \alpha_0, \quad \left(\frac{\kappa}{\kappa + \omega + N}\right)^{-1} = \beta_0,$$

obtém-se, $\kappa = \sqrt{N}/2$ e $\omega = \sqrt{N}/2$. Logo, δ_0 é Bayes contra a distribuição a priori

$$h_0(\theta) = \left[\frac{1}{B(\sqrt{N}/2, \sqrt{N}/2)}\right]\theta^{(\sqrt{N}/2)-1}(1-\theta)^{(\sqrt{N}/2)-1}, \quad 0 < \theta < 1,$$

e é, por isso, minimax. A distribuição $h_0(\theta)$ tem a pecularidade de ter parâmetros que dependem do número de provas (dimensão da amostra) o que não é bem aceite pelos bayesianos. □

Exemplo 4.14 — *Continuação.* É interessante comparar o risco da função de decisão minimax do exemplo anterior, $\tilde{\delta}(x) = \delta_0(x)$, com o risco da função de decisão ou estimador da máxima verosimilhança, $\hat{\delta}(x) = x/N$. Tem-se,

$$R(\theta, \hat{\delta}) = \frac{\theta(1-\theta)}{N} \leqq R(\theta, \tilde{\delta}) = [2(1+\sqrt{N})]^{-2},$$

se e somente se,

$$\left|\theta - \frac{1}{2}\right| \geqq \frac{\left(1+2\sqrt{N}\right)^{1/2}}{2(1+\sqrt{N})};$$

assim, $R(\theta, \tilde{\delta}) < R(\theta, \hat{\delta})$, num intervalo $\left(\frac{1}{2} - \Gamma_N, \frac{1}{2} + \Gamma_N\right)$ onde $\Gamma_N \to 0$ quando $N \to \infty$. Além disso,

$$\frac{\sup_\theta R(\theta, \hat{\delta})}{\sup_\theta R(\theta, \tilde{\delta})} = \frac{1/4N}{[2(1+\sqrt{N})]^{-2}} = \frac{N + 2\sqrt{N} + 1}{N} \to 1,$$

quando $N \to \infty$. Portanto, para N pequeno, o estimador minimax é preferível; para N grande, justifica-se o emprego do estimador da máxima verosimilhança pela maior simplicidade de cálculo. □

Exemplo 4.15 — Hodges e Lehmann (1950) — a quem se deve também o Ex. 4.14 — provaram por processo análogo que o estimador,

$$\tilde{\delta}(x) = \left\{\frac{M}{N + [N(M-N)/(M-1)]^{1/2}}\right\}x + \frac{M}{2}\left\{1 - \frac{N}{N + [N(M-N)/(M-1)]^{1/2}}\right\},$$

é minimax para o parâmetro Ξ de uma distribuição hipergeométrica,

$$f(x \mid M, N, \Xi) = \binom{\Xi}{x}\binom{M-\Xi}{N-x} \Big/ \binom{M}{N}, \; x = 0, 1, \ldots, \; N, \Xi = 0, 1, \ldots, M.$$

Este resultado tem sido aplicado em problemas de controlo estatístico da qualidade e em inquéritos por amostragem. □

Os Teoremas 4.22 e 4.24 traduzem uma estratégia muito clara na pesquisa de soluções minimax, que consiste em descobrir ou conjecturar uma distribuição a priori — própria — mais desfavorável, e só são válidos quando tal distribuição existe, [veja-se, no entanto, Berger (1985)].

Sucede, por vezes, que a conjectura aponta para uma distribuição a priori mais desfavorável imprópria. Por exemplo, na estimação da média, θ, de uma população Normal com variância conhecida, a distribuição a priori mais desfavorável parece ser aquela que torna todos os valores de θ igualmente difíceis de estimar (uma distribuição a priori concentrada num dado intervalo finito de $\Theta = \mathbf{R}$ facilitava o trabalho de estimação). No entanto, a distribuição uniforme em $\mathbf{R}$ é imprópria [veja-se (1.40)].

Segundo Lehmann (1983) há duas saídas possíveis.

[1] Se a distribuição a priori mais desfavorável é imprópria pode acontecer que a correspondente distribuição a posteriori seja própria. Nesse caso, com função perca quadrática, pode tomar-se $\tilde{\delta}(x) = E\{\theta \mid x\}$ [ou $E\{\tau(\theta) \mid x\}$ quando estiver em jogo o estimador minimax de $\tau(\theta)$], na esperança de que seja o desejado estimador minimax.

[2] Preferivelmente, pode introduzir-se uma sucessão de distribuições a priori próprias, seja $\{h_\nu(\theta)\}$, $\nu = 1, 2, \ldots$, tal que quando $\nu \to \infty$, $h_\nu(\theta)$ tende para a distribuição a priori mais desfavorável imprópria [por exemplo, no caso da distribuição uniforme sobre $\mathbf{R}$, pode fazer-se, $h_\nu(\theta) = 1/2\nu$, para $-\nu \leqq \theta \leqq \nu$, $\nu \to \infty$, ou $h_\tau(\theta) \equiv N(0, \tau^2)$, $\tau^2 \to \infty$; vejam-se Exs. 1.21 e 3.15].

A ideia é substituir o conceito de distribuição a priori mais desfavorável pelo conceito mais geral de sucessão de distribuições a priori mais desfavorável. Considere-se uma sucessão de distribuições a priori, $\{h_\nu(\theta)\}$, e seja δ_ν a função de decisão Bayes contra h_ν e r_ν o respectivo risco Bayes, $r_\nu = R(h_\nu, \delta_\nu)$; suponha-se que $r_\nu \to r$ quando $\nu \to \infty$; a mesma sucessão diz-se mais desfavorável quando sendo h uma distribuição a priori (própria) arbitrária, se verifica $r(h) = R(h, \delta_h) \leqq r$.

Demonstra-se então o seguinte:

Teorema 4.25 — Se $\{h_\nu\}$ é uma sucessão de distribuições a priori com risco Bayes, $r_\nu \to r$ e se $\tilde{\delta}$ é uma função de decisão verificando,

$$\sup_\theta R(\theta, \tilde{\delta}) = r, \tag{4.19}$$

então: (i) $\tilde{\delta}$ é minimax; (ii) a sucessão, $\{h_\nu\}$, é mais desfavorável.

Dem. (i) Suponha-se que δ é uma outra função de decisão qualquer. Tem-se,

$$\sup_\theta R(\theta, \delta) \geqq \int_\Theta R(\theta, \delta) h_\nu(\theta)\, d\theta \geqq r_\nu,$$

para todo o ν. Consequentemente,

$$\sup_\theta R(\theta, \delta) \geqq \sup_\theta R(\theta, \tilde{\delta}), \tag{4.20}$$

e $\tilde{\delta}$ é minimax.

(ii) Se h é uma distribuição a priori qualquer,

$$r(h) = \int_\Theta R(\theta, \delta_h) h(\theta)\, d\theta \leqq \int_\Theta R(\theta, \tilde{\delta}) h(\theta)\, d\theta \leqq \sup_\theta R(\theta, \tilde{\delta}) = r,$$

e a demonstração fica completa. □□

Do teorema anterior obtém-se imediatamente o seguinte,

Teorema 4.26 — Se a função de decisão, $\tilde{\delta}$, tem função risco, $R(\theta, \tilde{\delta}) = \rho$, constante, para todo o $\theta \in \Theta$, e se existe uma sucessão de distribuições a priori, $\{h_\nu\}$, tal que, sendo $\{\delta_\nu\}$ a respectiva sucessão de soluções Bayes,

$$\lim_{\nu \to \infty} R(h_\nu, \delta_\nu) = \rho,$$

então $\tilde{\delta}$ é minimax. □□

Exemplo 4.16 — Retome-se o Ex. 3.15 onde se trata da estimação da média da Normal. Quando a função perca é quadrática, será minimax o estimador da máxima verosimilhança, $\hat{\delta}(\mathbf{x}) = \overline{x}$?

Evidentemente, $\hat{\delta}$, tem risco constante,

$$R(\theta, \hat{\delta}) = E\{(\theta - \overline{X})^2\} = 1/N; \tag{4.21}$$

pelo Teorema 4.22 para ser minimax é bastante que seja Bayes contra alguma distribuição a priori. Se $\hat{\delta}$ for Bayes contra algum $h \in \mathcal{H}$, deve ter-se,

$$\delta(\mathbf{x}) = \overline{x} = E\{\theta \,|\, \mathbf{x}\} = E\{\theta \,|\, \overline{x}\}, \tag{4.22}$$

por se tratar de função perca quadrática (e a última relação por a média da amostra ser suficiente para θ)[6].

A relação (4.22) é absurda. Calcule-se, $E\{\theta\overline{X}\}$, valor esperado em relação à distribuição conjunta das variáveis aleatórias, θ e $\overline{X}$, condicionando primeiro em relação a $\overline{X}$ no primeiro cálculo, e condicionando primeiro em relação a θ no segundo cálculo. Tem-se,

$$E\{\theta\overline{X}\} = E\{E\{\theta\overline{X} \,|\, \overline{X}\}\} = E\{\overline{X} E\{\theta \,|\, \overline{X}\}\} = E\{\overline{X}^2\}, \tag{4.23}$$

[6] O Teorema 6.20 trata este problema com maior generalidade.

onde (4.22) interveio na obtenção do último termo; por outro lado,

$$E\{\theta\overline{X}\} = E\{E\{\theta\overline{X}\,|\,\theta\}\} = E\{\theta E\{\overline{X}\,|\,\theta\}\} = E\{\theta^2\}, \tag{4.24}$$

onde na obtenção do último termo se notou que $E\{\overline{X}\,|\,\theta\} = \theta$. Agora, (4.23) e (4.24) implicam,

$$R(h,\hat{\delta}) = E\{(\theta - \overline{X})^2\} = E\{\theta^2\} - 2E\{\theta\overline{X}\} + E\{\overline{X}^2\} = 0,$$

com h qualquer [compatível com a existência dos valores esperados envolvidos] o que é impossível, porquanto,

$$R(h,\hat{\delta}) = E\{(\theta - \overline{X})^2\} = E\{E\{(\theta - \overline{X})^2\,|\,\theta\}\} = \frac{1}{N}.$$

Se $\hat{\delta}$ não é Bayes contra nenhuma distribuição a priori a doutrina do Teorema 4.24 não permite chegar a qualquer conclusão: trata-se de condição suficiente mas não necessária. Em contra-partida o Teorema 4.26 é completamente esclarecedor. Tomem-se as distribuições a priori, h_τ, e as funções de decisão, δ_τ, contra h_τ, definidas no Ex. 3.15; formem-se as sucessões, $\{h_\tau\}$ e $\{\delta_\tau\}$ para alguma sucessão de valores $\tau \to \infty$. Tem-se,

$$\lim_{\tau\to\infty} R(h_\tau,\delta_\tau) = \lim_{\tau\to\infty} \frac{\tau^2}{1+N\tau^2} = \frac{1}{N},$$

donde se conclui que $\hat{\delta}$ é minimax. □

Exemplo 4.17 — Ainda relativamente ao problema tratado no Ex. 3.15 altere-se a função perca para,

$$L(\theta,a) = \begin{cases} 0 & \text{se } |\,\theta - a\,| < \Delta, \\ 1 & \text{se } |\,\theta - a\,| \geqq \Delta. \end{cases}$$

Mostra-se que a função de decisão,

$$\delta_\tau(\mathbf{x}) = \overline{x}\left(\frac{N\tau^2}{1+N\tau^2}\right),$$

é Bayes contra h_τ que corresponde a $\theta \sim N(0,\tau^2)$ — recorde-se que X_i I.I.D. $X_i \sim N(\theta,1)$, $i = 1,2,\ldots,N$. Com efeito, com,

$$h(\theta\,|\,\mathbf{x}) = N\left(\frac{\overline{x}N\tau^2}{1+N\tau^2},\ \frac{\tau^2}{1+N\tau^2}\right)$$

o risco a posteriori,

$$\begin{aligned} r_{\mathbf{x}}(\delta) &= E\{L(\theta,\delta)\,|\,\mathbf{x}\} = \int_{-\infty}^{+\infty} L(\theta,\delta)h(\theta\,|\,\mathbf{x})\,d\theta \\ &= 1 - P(\delta - \Delta < \theta < \delta + \Delta) \\ &= 1 - \left[\Phi\left\{\frac{\delta + \Delta - [\overline{x}N\tau^2/(1+N\tau^2)]}{[\tau^2/(1+N\tau^2)]^{1/2}}\right\} - \Phi\left\{\frac{\delta - \Delta - [\overline{x}N\tau^2/(1+N\tau^2)]}{[\tau^2/(1+N\tau^2)]^{1/2}}\right\}\right], \end{aligned}$$

onde $\Phi(u) \equiv N(0,1)$. Para minimizar $r_{\mathbf{x}}(\delta)$ em δ há que maximizar a expressão que no 2.º membro é subtraída da unidade.

Considere-se a expressão geral,

$$\Phi(\Omega + \Lambda - \Gamma) - \Phi(\Omega - \Lambda - \Gamma),$$

e calcule-se a derivada em relação a Ω; vem, sendo $\phi(u)$ a densidade de $\Phi(u)$,

$$\phi(\Omega + \Lambda - \Gamma) - \phi(\Omega - \Lambda - \Gamma) = 0,$$

donde, dada a forma e a simetria de $\phi(u)$, sai $\Omega = \Gamma$. Repondo as variáveis iniciais fica provado o que se pretendia.

Retome-se a função de decisão da máxima verosimilhança, $\hat{\delta}(\mathbf{x}) = \overline{x}$; vai mostrar-se que é minimax.

Tem-se,

$$R(\theta, \hat{\delta}) = P(|\,\theta - \overline{X}\,| \geqq \Delta) = 2[1 - \Phi(\Delta\sqrt{N})] = \rho,$$

para todo o $\theta \in (-\infty, +\infty)$.

Por outro lado, não é exercício muito difícil calcular,

$$R(\theta, \delta_\tau) = 2 - \left[\Phi\left\{\sqrt{N}\frac{\Delta + \theta(1+N\tau^2)^{-1}}{[N\tau^2/(1+N\tau^2)]}\right\} + \Phi\left\{\sqrt{N}\frac{\Delta - \theta(1+N\tau^2)^{-1}}{[N\tau^2/(1+N\tau^2)]}\right\}\right]; \quad (4.25)$$

o correspondente risco a priori é,

$$R(h_\tau, \delta_\tau) = \int_{-\infty}^{+\infty} R(\theta, \delta_\tau) h_\tau(\theta)\, d\theta.$$

Como a função integranda é uniformemente limitada por 2 — repare-se em (4.25) — pelo teorema da convergência limitada de Lebesgue,

$$\begin{aligned}\lim_{\tau \to \infty} R(h_\tau, \delta_\tau) &= \int_{-\infty}^{+\infty} \lim_{\tau \to \infty} R(\theta, \delta_\tau) h_\tau(\theta)\, d\theta \\ &= 2[1 - \Phi(\Delta\sqrt{N})] = \rho;\end{aligned}$$

logo, $\hat{\delta}(\mathbf{x}) = \overline{x}$ é minimax pelo Teorema 4.26 [Zacks (1971)]. □

Exemplo 4.18 — [Os pormenores podem ver-se em Zacks (1971)]. Considerem-se duas amostras casuais de dimensão, N, $\mathbf{X}$ e $\mathbf{Y}$, com X_i I.I.D., Y_i I.I.D., $X_i \sim N(\theta, \sigma^2)$, $Y_i \sim N(\theta, \kappa\sigma^2)$, sendo desconhecidos a média comum, θ, e as variâncias. O objectivo é estimar a média comum.

Se a função perca é $L_1(\theta, a) = (\theta - a)^2/\sigma^2$, mostra-se que o estimador minimax é a média da primeira amostra, $\overline{x}$; trata-se de um estimador minimax que

ninguém estará interessado em empregar e que resulta da parcialidade da função perca. Se, $L_2(\theta, a) = (\theta - a)^2/\sigma^2 \cdot \max\{1, \kappa\}$, a função de decisão minimax é $(\overline{x} + \overline{y})/2$.

As conclusões indicadas são obtidas com o emprego do Teorema 4.26 e mostram, mais uma vez, o condicionalismo introduzido ao adaptar uma dada função perca. □

Se a função perca é convexa [e se verificam as restantes condições do Teorema 4.4] o Teorema 4.5 mostra a desnecessidade de proceder a qualquer casualização, quer dizer, no presente contexto, se existe uma função de decisão mista minimax existe com certeza uma função de decisão pura minimax [vejam-se os Exs. 4.14, 4.15 e 4.16] e a solução minimax pode procurar-se em D.

Se a função perca não é convexa retiram-se vantagens da casualização como se ilustra no exemplo seguinte [veja-se também o Ex. 3.3].

Exemplo 4.14 — *Continuação.* Considere-se a função perca,

$$L(\theta, a) = \begin{cases} 0 & \text{se } |\theta - a| \leqq \alpha, \\ 1 & \text{se } |\theta - a| > \alpha. \end{cases} \qquad \alpha < 1/2(N+1)$$

Qualquer função de decisão pura, $\delta(x)$, pode assumir quando muito $N+1$ valores diferentes porquanto, $\mathfrak{X} = \{0, 1, 2, \ldots, N\}$. Portanto, para tal δ,

$$\sup_\theta R(\theta, \delta) = 1 \quad \text{para} \quad 0 \leqq \theta \leqq 1.$$

Por outro lado, o estimador independente de X, $\delta_0 = Z$, onde Z é uma variável aleatória com distribuição uniforme no intervalo $(0, 1)$, tem função risco,

$$R(\theta, \delta_0) = 1 - P(|Z - \theta| \leqq \alpha),$$

e,

$$\sup_\theta R(\theta, \delta_0) = 1 - \alpha < 1.$$

Nenhuma função de decisão pura é minimax pois, δ_0, para todos os efeitos um estimador casualizado, verifica, $\sup_\theta R(\theta, \delta_0) < \sup_\theta R(\theta, \delta)$ — Lehmann (1983). □

4.6 Complementos sobre admissibilidade

A Admissibilidade das funções de decisão é estabelecida — veja-se (4.1) e (4.2) — com base no conceito de função risco. Como semelhante conceito é de natureza pre-experimental, põe-se a questão de investigar se não há qualquer conflito entre a admissibilidade ou inadmissibilidade e a posição bayesiana.

Em relação às funções de decisão que são Bayes contra distribuições a priori próprias parece não haver qualquer incompatibilidade pois, como já se afirmou,

tais funções são virtualmente admissíveis e, seguramente, as funções de decisão admissíveis são Bayes.

O confronto agudiza-se, porém, quando passam a considerar-se funções de decisão Bayes generalizadas, pois como também já foi assinalado, se a função risco, $R(h, \delta)$, é infinita, podem as mesmas ser inadmissíveis [veja-se Ex. 4.6].

Para aprofundar a questão convém pensar que uma distribuição a priori imprópria, seja $h(\theta)$, representa, em princípio, uma aproximação de uma verdadeira distribuição a priori, seja $h_0(\theta)$. O que importa então saber é o seguinte: se a função de decisão Bayes generalizada, seja δ_h, é dominada estritamente por uma função de decisão, seja δ', e, portanto, inadmissível, pode, numa óptica bayesiana, considerar-se δ_h «melhor» do que δ'?

Uma sugestão óbvia é comparar os riscos a posteriori, com $h_0(\theta \mid x)$ correspondente a $h_0(\theta)$,

$$r_x^0(\delta_h) = \int_\Theta L[\theta, \delta_h(x)] h_0(\theta \mid x)\, d\theta,$$

e,

$$r_x^0(\delta') = \int_\Theta L[\theta, \delta'(x)] h_0(\theta \mid x)\, d\theta.$$

Naturalmente, se,

$$r_x^0(\delta') < r_x^0(\delta_h) \quad \text{para todo o } x, \tag{4.26}$$

a preferência deve recair em δ'. No entanto o que é de esperar que se verifique é ter-se a desigualdade (4.26) para uns valores de x e a desigualdade contrária para outros valores de x. Em qualquer caso, como se desconhece $h_0(\theta)$, o dilema mantém-se,

Se δ' domina estritamente δ_h, suponha-se que,

$$R(\theta, \delta') < R(\theta, \delta_h),$$

sobre um subconjunto de valores de Θ, seja Θ', a que $h_0(\theta)$ atribui probabilidade positiva. Consequentemente,

$$R(h_0, \delta') < R(h_0, \delta_h),$$

e, por (3.25),

$$\int r_x^0(\delta') f_0(x)\, dx < \int r_x^0(\delta_h) f_0(x)\, dx,$$

com,

$$f_0(x) = \int_\Theta f(x \mid \theta) h_0(\theta)\, d\theta.$$

Assim, como se verifica «em média» uma relação do tipo de (4.26), pode orientar-se a preferência para δ' ao mesmo tempo que se obtém de certo modo uma justificação bayesiana para o conceito de admissibilidade.

Hill (1974) reformulou a justificação bayesiana da admissibilidade em termos de coerência.

Pedro e Paulo entram num jogo em que cada um se apresenta com uma função de decisão, δ_1 e δ_2, respectivamente. Seguidamente, Paulo escolhe à sua vontade uma distribuição a priori, seja $h(\theta)$, procedendo-se imediatamente à escolha casual de um valor de θ, seja θ_1, de acordo com essa distribuição, e à escolha casual de um valor da variável aleatória, X, de acordo com a distribuição experimental, $f(x\,|\,\theta_1)$. O processo é repetido n vezes dando lugar a n pares independentes, $(\theta_1, x_1), (\theta_2, x_2), \ldots, (\theta_n, x_n)$. Após a selecção de cada par, (θ_i, x_i), $i = 1, 2, \ldots, n$,

- Pedro paga a Paulo $L[\theta_i, \delta_1(x_i)]$,
- Paulo paga a Pedro $L[\theta_i, \delta_2(x_i)]$.

Se,

$$\lim_{n\to\infty} P\left[\sum_{i=1}^{n}\{L[\theta_i, \delta_2(x_i)] - L[\theta_i, \delta_1(x_i)]\} > 0\right] = 1, \tag{4.27}$$

onde a probabilidade indicada é calculada com a distribuição conjunta dos n pares $[= \Pi f(x_i\,|\,\theta_i)h(\theta_i)]$, qualquer que seja a distribuição a priori escolhida por Paulo, considera-se que este é incoerente ao optar por δ_2.

Se a função risco é limitada, pode mostrar-se que a condição necessária e suficiente para se verificar (4.27) é ter-se, $R(\theta, \delta_1) < R(\theta, \delta_2)$, para todo o $\theta \in \Theta$. Esta proposição mostra que o conceito de incoerência de Hill é simplesmente uma condição ligeiramente mais forte que a inadmissibilidade.

O tratamento do problema da admissibilidade, para além dos resultados simples apresentados, exige conhecimentos de matemática acima do nível pressuposto na redacção do presente texto. Assim, pouco mais se pode fazer do que citar a caracterização da admissibilidade feita por Stein através de uma condição suficiente e as condições necessárias e suficientes de admissibilidade estabelecidas pelo mesmo autor e generalizadas por LeCam. Sobre o assunto, e para obter mais referências, podem consultar-se Zacks (1971) e Berger (1985). O segundo autor enuncia as condições necessárias e suficientes de forma simplificada e indica algumas consequências das mesmas.

CAPÍTULO 5

INVARIÂNCIA

5.1 Conceito de invariância

A determinação de funções de decisão óptimas à luz do critério minimax e do critério Bayes foi oportunamente abordada. Muitas vezes, porém, não existe óptimo ou a optimização é difícil de levar a bom termo. A constatação deste facto conduz a outra via de ataque ao problema da decisão, aliás já mencionada, que consiste em restringir a escolha a procedimentos ou subclasses de funções de decisão possuindo certas características de simetria.

Se um problema de decisão é simétrico ou invariante para determinadas operações, parece razoável recorrer preferentemente a regras ou funções de decisão que sejam simétricas ou invariantes para as mesmas operações.

Para introduzir o conceito de invariância o melhor a fazer é analisar os dois exemplos seguintes [Berger (1980)]:

Exemplo 5.1 — O tempo de vida de uma dada componente electrónica, X, tem função de densidade,

$$f(x\,|\,\theta) = (1/\theta)\exp\{-x/\theta\},\ x > 0,\ \theta > 0.$$

Pretende estimar-se θ a partir da observação de X, com função perca,

$$L(\theta, a) = [1 - (a/\theta)]^2;$$

suponha-se que X é medido em segundos e que é proposta uma função de decisão, $\delta(x)$.

Considere-se o problema de decisão que se põe quando o tempo de vida é medido em minutos e não em segundos. A variável observada vem então, $Y = X/60$; definindo, $\eta = \theta/60$, a função de densidade de Y vem dada por,

$$f(y\,|\,\eta) = (1/\eta)\exp\{-y/\eta\},\ y > 0,\ \eta > 0.$$

Se uma acção no novo problema, seja a', é também expressa em minutos, $a' = a/60$, é claro que,

$$\begin{aligned} L(\theta, a) = [1 - (a/\theta)]^2 &= \left[1 - \frac{a/60}{\theta/60}\right]^2 \\ &= [1 - (a'/\eta)]^2 = L(\eta, a'). \end{aligned}$$

Assim, a estrutura formal do problema em termos de minutos [a classe de densidades, o espaço do parâmetro, a função perca] é exactamente idêntica à estrutura formal do problema em termos de segundos. Logo parece natural usar a mesma função de decisão nas duas formulações; designando por $\delta'(y)$ a função de decisão proposta no problema transformado, isso quer dizer que,

$$\delta'(y) = \delta(y) = \delta(x/60). \tag{5.1}$$

Segundo outra óptica parece natural exigir que seja tomada a mesma decisão independentemente da unidade escolhida para medir o tempo; isso quer dizer que,

$$\delta'(y) = \delta(x)/60. \tag{5.2}$$

Combinando (5.1) e (5.2) obtém-se a relação,

$$\delta(x) = 60\,\delta'(y) = 60\,\delta(y) = 60\,\delta(x/60). \tag{5.3}$$

O raciocínio anterior pode seguir-se para qualquer transformação da forma, $Y = cX$, $c > 0$. Consequentemente,

$$\delta(x) = c^{-1}\delta(cx),\ c > 0, \tag{5.4}$$

ou, fazendo $c = 1/x$,

$$\delta(x) = x\delta(1). \tag{5.5}$$

Em semelhante problema parece intuitivo que as funções de decisão coerentes com a análise feita são apenas as do tipo $\delta(x) = kx$, $k > 0$; tais funções de decisão dizem-se invariantes e o princípio de invariância recomenda que somente devem usar-se as mesmas. □

Exemplo 5.1 — *Continuação.* Suponha-se que considerações teóricas implicam ser $\theta > 120$ segundos, caso em que o espaço do parâmetro é $\Theta = (120, \infty)$. Se o tempo é medido em minutos o espaço do parâmetro, η, é $\mathbf{H} = (2, \infty)$. A estrutura formal do problema original e do transformado deixam de ser idênticas e não há motivo para recomendar (5.1); pode continuar a aceitar-se (5.2), relação que por si só pouco permite avançar. □

Os argumentos apresentados intuitivamente no Ex. 5.1 traduzem, de facto, dois importantes princípios:

Princípio de invariância: se dois problemas de decisão possuem a mesma estrutura formal [em termos de $\mathcal{X}$, Θ, $f(x \mid \theta)$ e L], então deve usar-se a mesma função de decisão nos dois. A relação (5.1) decorre deste princípio.

Princípio de invariância racional: em qualquer problema de decisão as acções não devem estar na dependência das unidades de medida ou de quaisquer factores arbitrariamente fixados. A relação (5.2) decorre deste princípio.

O segundo princípio é pacífico. O primeiro princípio tem um aspecto bastante controverso: a não inclusão da distribuição a priori — juntamente com $\mathcal{X}$, Θ, f e L — na lista dos elementos do problema de decisão em termos dos quais se caracteriza a invariância. Quer dizer, pode aduzir-se que a aplicação do princípio de invariância só é sensata quando a distribuição a priori for também invariante; na maioria dos casos práticos não se atende a tal recomendação.

No caso do Ex. 5.1, a preocupação com a invariância da distribuição a priori, obrigaria a sustentar: 1.º) $h'(\eta) = h(\eta)$, tal como em (5.1); 2.º) $h'(\eta) = (1/c)h(\eta/c)$, face à transformação, $\eta = c\theta$. Considerando as duas relações em conjunto, a distribuição a priori invariante é da forma,

$$h(\theta) \propto 1/\theta,$$

ou seja, dado que θ é um parâmetro de escala, da forma não informativa [e imprópria — veja-se (1.41)].

O estudo de outros problemas e transformações parece sugerir que o princípio de invariância só é ‹adequado› quando não existe informação a priori[1], isto é, que o recurso à invariância corresponde à análise bayesiana com uma distribuição a priori não informativa.

Sobre a questão é útil citar Berger (1980): «*If invariance is equivalent to the use of noninformative priors, a natural question is — why should we bother with ...invariance? There are three basic reasons. First, people who don't like to talk about noninformative priors are welcome to do the same thing using invariance. Second, it is not strictly true that the two approaches always correspond, although when they*

[1] Distribuições a priori próprias invariantes raramente existem.

don't, invariance is probably suspect. Third, and most importantly ... invariance suggests one particular noninformative prior for use, namely that which is called the right invariant Haar measure on the group of transformations»[2].

5.2 Grupos de transformações. Problemas invariantes

Em termos matemáticos a invariância exprime-se geralmente em relação a grupos de transformações.

Seja $\mathcal{Y}$ um conjunto arbitrário e considere-se uma família, $G = \{g\}$, de transformações ou aplicações bijectivas (correspondências biunívocas),

$$g : \mathcal{Y} \to \mathcal{Y}.$$

Para qualquer $y \in \mathcal{Y}$ o transformado de y por $g \in G$ representa-se por $g(y)$ ou simplesmente por gy. Por definição, $g(\mathcal{Y}) = \mathcal{Y}$.

O produto de duas transformações ou aplicação composta define-se do modo habitual: com $g, g' \in G$, o produto designa-se por $g \circ g'$, e,

$$(g \circ g')(y) = g(g'(y)) \text{ para cada } y \in \mathcal{Y}. \tag{5.6}$$

Dado $g \in G$ a transformação inversa designa-se por g^{-1}; a transformação idêntica ou identidade designa-se por e e define-se por,

$$e(y) = y \text{ para cada } y \in \mathcal{Y}. \tag{5.7}$$

Tem-se,

$$e \circ g = g \circ e = g \text{ para todo o } g \in G, \tag{5.8}$$

e ainda, por definição de inversa,

$$g \circ g^{-1} = g^{-1} \circ g = e \text{ para todo o } g \in G. \tag{5.9}$$

A família G constitui um grupo de transformações quando:

{GT1} $e \in G$;
{GT2} se $g \in G$ também $g^{-1} \in G$;
{GT3} se $g \in G$, $g' \in G$ então $g \circ g' \in G$.

Considere-se a variável ou vector aleatório, X, e seja $\mathcal{X}$ o respectivo espaço de resultados ou domínio. Em tudo o que se segue, G, representa um grupo de transformações bijectivas, $\mathcal{X} \to \mathcal{X}$. Para qualquer, $g \in G$, designa-se por $g(X)$ ou

[2] As medidas Haar invariantes são referidas na secção 5.5.

simplesmente por gX a variável aleatória que assume o valor $g(x)$ quando X assume o valor x [supõe-se, portanto, que as transformações ou funções, $g \in G$, são mensuráveis].

No caso presente, pelo menos inicialmente, é mais conveniente trabalhar com a família de distribuições de X na forma (1.2). A família, $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$, de medidas de probabilidade sobre $\mathcal{X}$, diz-se invariante em relação ao grupo de transformações, G, se para todo o $g \in G$ e para todo o $\theta \in \Theta$ existe um único $\theta' \in \Theta$ tal que,

$$X \sim P_\theta \quad \Rightarrow \quad gX \sim P_{\theta'}.$$

Sendo θ' unicamente determinado por g e por θ, escreve-se $\theta' = \overline{g}\theta$, onde $\overline{g}$ é uma aplicação, $\overline{g} : \Theta \to \Theta$, induzida por g. Assim, G, grupo de transformações em $\mathcal{X}$, induz uma família de aplicações em Θ, seja,

$$\overline{G} = \{\overline{g} = \Gamma(g) : g \in G\}.$$

Teorema 5.1 — Se a família, $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$, é invariante em relação a G, então $\overline{G}$ é um grupo de transformações.

Dem. Se X tem distribuição P_θ, $g(X)$ tem distribuição $P_{\overline{g}\theta}$ e $g'(g(X)) = (g' \circ g)(X)$ tem distribuição $P_{\overline{g}'(\overline{g}\theta)} = P_{(\overline{g}' \circ \overline{g})\theta}$. Da unicidade inerente ao conceito de invariância de $\mathcal{F}$ sai,

$$\overline{g'} \circ \overline{g} = \overline{g' \circ g}, \quad \overline{g' \circ g} \in \overline{G},$$

e $\overline{G}$ é fechado em relação ao produto de aplicações. Tomando, em particular, $g' = g^{-1}$, vem,

$$\overline{g^{-1}} \circ \overline{g} = \overline{g^{-1} \circ g} = \overline{e},$$

onde $\overline{e}$ é a transformação identidade de $\overline{G}$. Logo, $\overline{e} \in \overline{G}$ e, mais,

$$(\overline{g})^{-1} = \overline{g^{-1}}, (\overline{g})^{-1} \in \overline{G},$$

e $\overline{G}$ é um grupo. □□

A aplicação, $G \to \overline{G}$, atrás simbolicamente representada por $\overline{g} = \Gamma(g)$, estabelece um morfismo entre os sistemas algébricos, $[G, \circ]$ e $[\overline{G}, \circ]$; porém, vai ver-se que se trata de aplicação sobrejectiva mas não, em regra, bijectiva; está-se em presença de um homomorfismo e não, em geral, de um isomorfismo.

Exemplo 5.2 — Suponha-se $\mathcal{X} = \mathbf{R}$ e considere-se a família de transformações, $G = \{g_c : c \in \mathbf{R}\}$, onde $g_c(x) = x + c$. Como facilmente se verifica, $g_{c_1} \circ g_{c_2} = g_{(c_1+c_2)} \in G$, $g_c^{-1} = g_{(-c)} \in G$, e $e = g_o \in G$. Trata-se, portanto, de um grupo de transformações conhecido por grupo aditivo ou grupo localização (em $\mathbf{R}$). □

Exemplo 5.3 — Seja $\mathcal{X} = (0, \infty)$ e considere-se a família de transformações, $G = \{g_c : c > 0\}$, onde $g_c(x) = cx$. Tem-se,

$$g_{c_2} \circ g_{c_1}(x) = g_{c_2}(g_{c_1}(x)) = g_{c_2}(c_1 x) = c_2 c_1 x = g_{c_2 c_1}(x),$$

donde, $g_{c_2} \circ g_{c_1} = g_{c_2 c_1} \in G$. De modo análogo se mostra que $g_c^{-1} = g_{c^{-1}} \in G$ e $e = g_1 \in G$. G é consequentemente um grupo [multiplicativo ou escala]. □

Exemplo 5.4 — Suponha-se $\mathcal{X} = \mathbf{R}$ e que $X \sim N(\mu, \sigma^2)$:

$$dP_{\boldsymbol{\theta}}(x) = (\sigma\sqrt{2\pi})^{-1} \exp\{-(x-\mu)^2/2\sigma^2\}\, dx,$$

e,

$$\Theta = \{\boldsymbol{\theta} = (\mu, \sigma^2) : -\infty < \mu < +\infty,\ 0 < \sigma^2 < +\infty\}.$$

A família de transformações,

$$G = \{g_{\alpha,\beta} : 0 < \alpha < \infty,\ -\infty < \beta < +\infty\},$$

com,

$$g_{\alpha,\beta}(x) = \alpha x + \beta,$$

é um grupo.

Com efeito, tanto a transformação identidade, $e = g_{1,0}$, como a transformação inversa,

$$g_{\alpha,\beta}^{-1} = g_{1/\alpha, -\beta/\alpha},$$

como o produto de transformações,

$$g_{\alpha',\beta'} \circ g_{\alpha,\beta} = g_{\alpha'\alpha, \alpha'\beta+\beta'},$$

pertencem a G.

Por outro lado, a família $\mathcal{F}$ é invariante em relação a G: para todo o $g_{\alpha,\beta} \in G$ e para todo o $\boldsymbol{\theta} \in \Theta$ existe um único $\boldsymbol{\theta}' \in \Theta$ tal que quando a distribuição de X é $P_{\boldsymbol{\theta}}$ a distribuição de $g_{\alpha,\beta}(X)$ é dada por $P_{\boldsymbol{\theta}'}$, com,

$$\boldsymbol{\theta}' = (\alpha\mu + \beta, \alpha^2\sigma^2);$$

isto é,

$$\overline{g}_{\alpha,\beta}(\boldsymbol{\theta}) \equiv \overline{g}_{\alpha,\beta}(\mu, \sigma^2) = (\alpha\mu + \beta, \alpha^2\sigma^2).$$

Pelo Teorema 5.1,

$$\overline{G} = \{\overline{g}_{\alpha,\beta} : 0 < \alpha < \infty,\ -\infty < \beta < +\infty\},$$

é também um grupo, no caso presente imagem isomórfica de G. □

Exemplo 5.4 — *Continuação.* Tome-se $\mu = 0$ e $\sigma = \theta$. Vem,

$$dP_\theta(x) = (\theta\sqrt{2\pi})^{-1} \exp\{-x^2/2\theta^2\}\, dx,$$
$$\Theta = \{\theta : 0 < \theta < \infty\}.$$

Considerem-se as duas transformações, $e(x) = x$ e $g(x) = -x$; $G = \{e, g\}$, forma um grupo — note-se que $g^{-1} \equiv g$ — em relação ao qual $\mathcal{F}$ é invariante. No entanto, como a distribuição de X é idêntica à distribuição de $-X$, $\overline{g}(\theta) = \theta$ para todo o $\theta \in \Theta$ e o grupo induzido, $\overline{G} = \{\overline{e}\}$, é composto apenas pela transformação identidade e é uma imagem homomórfica de G. □

Como gX tem distribuição $P_{\overline{g}\theta}$ quando X tem distribuição P_θ, para qualquer conjunto Q mensurável, $Q \subset \mathcal{X}$,

$$P_\theta(Q) = P_{\overline{g}\theta}(gQ), \tag{5.10}$$

onde gQ é a imagem de Q por g (veja-se Fig. 5.1).

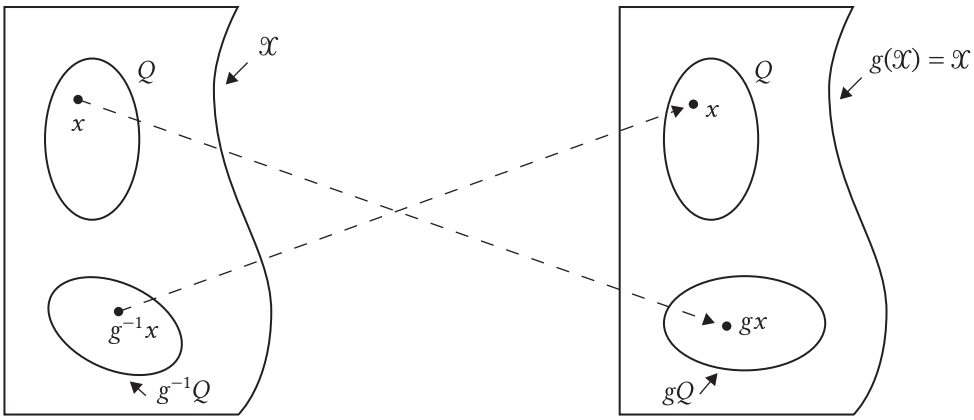


Fig. 5.1

A relação (5.10) pode escrever-se,

$$P_\theta(X \in Q) = P_{\overline{g}\theta}(X \in gQ). \tag{5.11}$$

Substituindo, Q por $g^{-1}Q$ e, consequentemente, gQ por Q, obtém-se a relação equivalente,

$$P_\theta(g^{-1}Q) = P_{\overline{g}\theta}(Q), \tag{5.12}$$

ou,

$$P_\theta(X \in g^{-1}Q) = P_{\overline{g}\theta}(X \in Q), \tag{5.13}$$

ou ainda,

$$P_\theta(gX \in Q) = P_{\overline{g}\theta}(X \in Q), \tag{5.14}$$

onde o índice θ na medida de probabilidade do primeiro membro se refere à distribuição de X e não à de gX.

Se $\psi(x)$ é função mensurável, $\psi(gX)$ e $\psi(X)$ são variáveis aleatórias e, em consequência de (5.14),

$$E_\theta\{\psi(gX)\} = E_{\overline{g}\theta}\{\psi(X)\}, \tag{5.15}$$

relação de que adiante se faz largo emprego e onde se indica em índice a expressão do parâmetro em relação a qual se calcula o valor esperado das funções de X.

Exemplo 5.5 — No quadro do Ex. 5.4 seja $Q = (-\infty, x_0)$, $-\infty < x_0 < +\infty$. Como ilustração de (5.11), vem,

$$P_{(\mu,\sigma^2)}[X \in (-\infty, x_0)] = P_{(\alpha\mu+\beta,\,\alpha^2\sigma^2)}[X \in (-\infty, \alpha x_0 + \beta)] = \Phi[(x_0 - \mu)/\sigma];$$

exemplificando (5.13),

$$P_{(\mu,\sigma^2)}\left[X \in \left(-\infty, \frac{x_0 - \beta}{\alpha}\right)\right] = P_{(\alpha\mu+\beta,\,\alpha^2\sigma^2)}[X \in (-\infty, x_0)] = \Phi[(x_0 - \alpha\mu - \beta)/\alpha\sigma];$$

exemplificando (5.14),

$$P_{(\mu,\sigma^2)}[\alpha X + \beta \in (-\infty, x_0)] = P_{(\alpha\mu+\beta,\,\alpha^2\sigma^2)}[X \in (-\infty, x_0)] = \Phi[(x_0 - \alpha\mu - \beta)/\alpha\sigma].$$

Seja $\psi(x) = x^2$; exemplificando (5.15), tem-se,

$$E_{(\mu,\sigma^2)}\{(\alpha X + \beta)^2\} = E_{(\alpha\mu+\beta,\,\alpha^2\sigma^2)}\{X^2\} = \alpha^2\sigma^2 + (\alpha\mu + \beta)^2.$$

□

Um problema de decisão, caracterizado pelo espaço de acções, A, pelo espaço de estados, Θ, e pela função perca, $L(\theta, a)$, em que a experiência consiste na observação da variável aleatória, X, com espaço de resultados, $\mathcal{X}$, e distribuição na família, $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$, diz-se invariante em relação ao grupo de transformações, G, se:

{I1} $\mathcal{F}$ é invariante em relação a G;

{I2} $L(\theta, a)$ é invariante em relação a G, isto é, para todo o $g \in G$ e todo o $a \in A$, existe um único $a' \in A$ tal que,

$$L(\theta, a) = L(\overline{g}\theta, a') \text{ para } \theta \in \Theta,$$

onde $\overline{g}$ é a transformação induzida por g.

A acção, $a' \in A$, unicamente determinada por g e por a, designa-se por $\tilde{g}(a)$ ou simplesmente por $\tilde{g}a$. A condição {I2} pode então tomar a forma alternativa,

$$L(\theta, a) = L(\overline{g}\theta, \tilde{g}a) \text{ para todo o } \theta \in \Theta \text{ e todo o } g \in G.$$

Note-se que o requisito de a' ser único não é de modo nenhum restritivo, pois se para $a'' \in A$, $L(\overline{g}\theta, a') = L(\overline{g}\theta, a'')$, para todo o $\theta \in \Theta$, tem-se $L(\theta, a') = L(\theta, a'')$ para todo o $\theta \in \Theta$ e a'' pode eliminar-se de A.

Teorema 5.2 — Se um problema de decisão é invariante em relação ao grupo de transformações, G, então, $\tilde{G} = \{\tilde{g} = \Lambda(g) : g \in G\}$, é um grupo de transformações.

Dem. É muito semelhante à do teorema anterior. Note-se,

$$L(\theta, a) = L(\overline{g}\theta, \tilde{g}a) = L[\overline{g}'(\overline{g}\theta), \tilde{g}'(\tilde{g}a)] = L[(\overline{g}' \circ \overline{g})\theta, (\tilde{g}' \circ \tilde{g})a],$$

donde,

$$L(\theta, a) = L[(\overline{g' \circ g})\theta, (\tilde{g}' \circ \tilde{g})a],$$

e pela unicidade,

$$\tilde{g}' \circ \tilde{g} = \widetilde{g' \circ g}, \widetilde{g' \circ g} \in \tilde{G}; \tag{5.16}$$

o resto da demonstração é imediato [Ferguson (1967)]. □□

Exemplo 5.6 — Suponha-se $\mathcal{X} = \mathbf{R}$ e $X \sim N(\theta, 1)$. Assim,

$$dP_\theta(x) = (\sqrt{2\pi})^{-1} \exp\{-(x-\theta)^2/2\}\, dx.$$

O problema é a estimação de θ, $A = \Theta = \mathbf{R}$; seja, $L(\theta, a) = (\theta - a)^2$ e considere-se o grupo de transformações,

$$G = \{g_c : -\infty < c < +\infty\}, g_c x = x + c.$$

A distribuição de $g_c X = X + c$ é ainda Normal com média $\theta + c$ e variância igual a um. Portanto, $\mathcal{F}$ é invariante em relação a G e,

$$\overline{G} = \{\overline{g}_c : -\infty < c < +\infty\}, \overline{g}_c\theta = \theta + c.$$

Além disso, a relação,

$$L(\theta, a) = L(\overline{g}_c\theta, \tilde{g}_c a),$$

é verificada para todo o $\theta \in \Theta$ e todo o $g_c \in G$, logo para todo o $\overline{g}_c \in \overline{G}$, desde que $\tilde{g}_c a = a + c$. Em conclusão, o problema de decisão é invariante em relação a G. □

Exemplo 5.7 — A variável aleatória, X, tem distribuição Binomial,

$$f(x\,|\,\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x},$$

num problema de estimação de $\theta - A = \Theta = (0,1)$ — com função perca, $L(\theta, a) = W(\theta - a)$, função par de $\theta - a$.

Seja o grupo de transformações,

$$G = \{e, g\}, \text{ onde } gx = N - x, g^{-1} \equiv g;$$

a distribuição de gX é,

$$P_\theta(gX = x) = \binom{N}{N-x}\theta^{N-x}(1-\theta)^x = f(x\,|\,1-\theta).$$

Portanto, $\mathcal{F} = \{f(x\,|\,\theta) : \theta \in (0,1)\}$, é invariante em relação a G, tendo-se, $\overline{g}\theta = 1 - \theta$. Com, $\tilde{g}a = 1 - a$,

$$W(\overline{g}\theta - \tilde{g}a) = W(\theta - a),$$

e conclui-se que o problema é invariante em relação a G. □

Exemplo 5.8 — Com $X \sim N(\theta, \sigma^2)$, θ e σ^2 desconhecidos, pretende ensaiar-se a hipótese H_0: $\theta \leqq 0$ contra a alternativa H_1: $\theta > 0$, com função perca «0–1». Tem-se, $A = \{a_0, a_1\}$, onde a_i designa a aceitação de $H_i\,(i = 0, 1)$. Considere-se o grupo de transformações, $G = \{g_c : c > 0\}$, $g_c x = cx$; como facilmente se deduz, a distribuição de $g_c X$ é a $N(c\theta, c^2\sigma^2)$ o que mostra ser $\mathcal{F}$ invariante em relação a G com $\overline{g}_c(\theta, \sigma^2) = (c\theta, c^2\sigma^2)$. Por outro lado, se $\theta \leqq 0$ também $c\theta \leqq 0$ e se $\theta > 0$ também $c\theta > 0$, isto é, designando $\Theta_0 = \{\theta : \theta \leqq 0\}$, e $\Theta_1 = \{\theta : \theta > 0\}$, tem-se $\overline{g}_c(\Theta_0) = \Theta_0$ e $\overline{g}_c(\Theta_1) = \Theta_1$. Consequentemente, definindo, como bem se compreende, $\tilde{g}_c(a_i) = \tilde{e}(a_i) = a_i$, $i = 0, 1$, o problema é invariante em relação a G. Repare-se que $\tilde{G} = \{\tilde{e}\}$, isto é, contém apenas a transformação identidade, facto que é muito corrente nos problemas de ensaio de hipóteses. □

Exemplo 5.8 — *Continuação.* Se o grupo de transformações for,

$$G = \{g_c : c < 0\},\ g_c x = cx,$$

o problema continua a ser invariante com,

$$\overline{g}_c(\theta, \sigma^2) = (c\theta, c^2\sigma^2),$$

e,

$$\tilde{g}_c(a_i) = a_{1-i}.$$

A troca das acções é de certo lógica, porquanto, agora, $\overline{g}_c(\Theta_0) = \Theta_1$, $\overline{g}_c(\Theta_1) = \Theta_0$. □

5.3 Funções de decisão invariantes[3]

Dado um problema de decisão invariante em relação ao grupo de transformações, G, uma função de decisão pura diz-se invariante em relação a G, se, com $\delta \in D$, se tem,

$$\delta(gx) = \tilde{g}\delta(x), \tag{5.17}$$

para todo o $x \in \mathcal{X}$ e para todo o $g \in G$. Na Fig. 5.2 esquematiza-se o comportamento de uma função de decisão invariante. O conceito resulta das seguintes considerações: a invariância traduz-se pela equivalência da estrutura formal dos problemas em que se observa X ou se observa gX; sejam, δ e δ' as funções de decisão aplicadas, respectivamente, em cada caso; tomando $Y = gX$, para simplificar, pelo princípio de invariância deve ter-se, $\delta(y) = \delta'(y)$; pelo princípio de invariância racional deve ter-se $\delta'(y) = \tilde{g}\delta(x)$; da combinação sai (5.17).

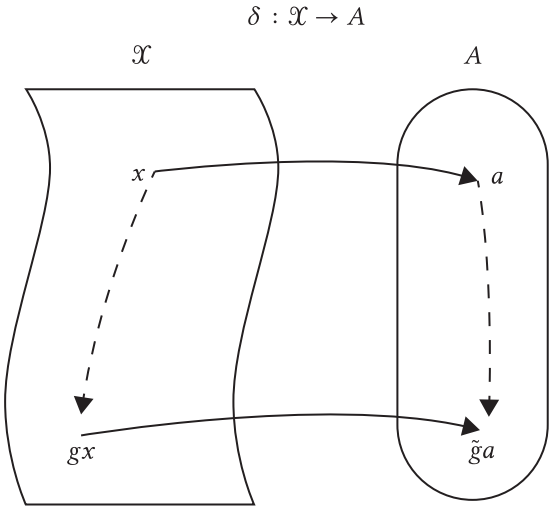


Fig. 5.2

Uma função de decisão mista, $\delta^* \in D^*$, diz-se invariante em relação a G quando resulta da mistura de funções de decisão puras invariantes.

Uma função de decisão aleatória, $\phi(a \mid x)$, $\phi \in \Phi$, é invariante em relação a G se, para todo o $x \in \mathcal{X}$ e todo o $g \in G$,

$$\phi(a \mid gx) = \tilde{g}\phi(a \mid x). \tag{5.18}$$

onde $\tilde{g}\phi$ deve entender-se como a distribuição de $\tilde{g}Z$ quando Z tem distribuição ϕ, sendo Z uma variável aleatória assumindo valores em A.

[3] Alguns autores preferem a designação de «equivariantes» reservando, com mais propriedade, a de «invariantes» para as funções que satisfazem a condição [1] da secção 5.4.

Para melhor compreender a definição acima, considere-se um subconjunto (mensurável) A_0, do espaço de acções A, $A_0 \subset A$; quando se emprega a função de decisão aleatória invariante definida por (5.18), seja ϕ, tem-se, mantendo a mesma fixa,

$$P(Z \in A_0 \,|\, gx) = P(\tilde{g}Z \in A_0 \,|\, x),$$

ou,

$$P(Z \in A_0 \,|\, gx) = P(Z \in \tilde{g}^{-1}A_0 \,|\, x), \tag{5.18′}$$

ou ainda, tomando $A_0 = \tilde{g}A_1$,

$$P(Z \in \tilde{g}A_1 \,|\, gx) = P(Z \in A_1 \,|\, x), \tag{5.18′′}$$

onde se reconhecem expressões do tipo (5.11), (5.13) e (5.14) [a parametrização da distribuição, $\phi(a \,|\, x)$, é aqui feita em termos de $x \in \mathcal{X}$; como sempre, Z, é uma variável aleatória com distribuição ϕ e assumindo valores em A]; veja-se Fig. 5.2-a.

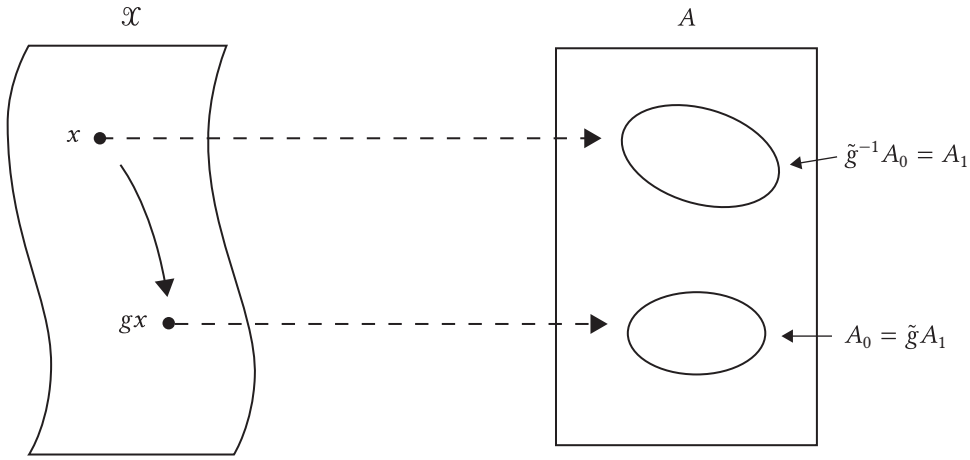


Fig. 5.2-a

Para introduzir o importante conceito de órbita[4], importa instituir, em $\mathcal{X}$, a relação de equivalência — simbolicamente $\cong$ — definida do modo seguinte,

- com $x, x' \in \mathcal{X}$, $x' \cong x$, se e somente se existe $g \in G$, tal que $x' = gx$.

A classe de equivalência de x,

$$C(x) = \{x' \in \mathcal{X} \,:\, x' \cong x\},$$

[4] Vão definir-se órbitas no espaço de resultados, órbitas no espaço do parâmetro e órbitas no espaço de acções. As mais importantes são sem dúvida as definidas no espaço do parâmetro.

designa-se por órbita de x relativamente a G. Por outras palavras, a órbita de x é o conjunto de todos os transformados de x por meio de g quando g percorre o grupo G. A partição de $\mathfrak{X}$ que tem as órbitas por elementos é induzida por G.

De modo absolutamente análogo se introduzem órbitas e partições em Θ induzidas por $\overline{G}$ e em A induzidas por $\tilde{G}$.

Teorema 5.3 — Se um problema de decisão é invariante em relação a G, toda a função de decisão pura, $\delta \in D$, invariante em relação a G, tem função risco constante em cada órbita de Θ.

Dem. Seja $\delta \in D$ invariante em relação a G; tem-se, sucessivamente[5],

$$\begin{aligned} R(\theta,\delta) &= E_\theta\{L[\theta,\delta(X)]\} \\ &= E_\theta\{L[\overline{g}\theta,\tilde{g}\delta(X)]\} && \text{(invariância da função perca)} \\ &= E_\theta\{L[\overline{g}\theta,\delta(gX)]\} && \text{(invariância de } \delta) \\ &= E_{\overline{g}_\theta}\{\overline{g}\theta,\delta(X)\}, && \text{(invariância de } \mathcal{F} - (5.15)) \end{aligned}$$

e, finalmente,

$$R(\theta,\delta) = R(\overline{g}\theta,\delta),$$

para todo o $\overline{g} \in \overline{G}$. □□

Teorema 5.4 — Se um problema de decisão é invariante em relação a G, toda a função de decisão mista, $\delta^* \in D^*$, invariante em relação a G, tem função risco constante em cada órbita de Θ.

Dem. Consequência imediata do teorema anterior e da definição de invariância para as funções de decisão mistas. □□

Com $\phi \in \Phi$ e $g \in G$, defina-se $\phi_g \in \Phi$ do modo seguinte,

$$\phi_g(a \mid x) = \tilde{g}^{-1}\phi(a \mid gx), \tag{5.19}$$

onde, $\tilde{g}^{-1}(a \mid gx)$, representa a distribuição de $\tilde{g}^{-1}Z$ quando Z tem distribuição $\phi(a \mid gx)$, com Z variável aleatória assumindo valores em A. Desta definição resulta que ϕ é invariante se e somente se, $\phi(a \mid x) = \phi_g(a \mid x)$ para todo o $g \in G$. Com efeito, de (5.19) tem-se,

$$\tilde{g}\phi(a \mid x) = \tilde{g}\phi_g(a \mid x) = \phi(a \mid gx),$$

como em (5.18).

[5] Repare-se que $E_\theta\{\psi(X,\cdot)\}$, designa o valor esperado da expressão em X relativamente à distribuição P_θ:

$$E_\theta\{\Psi(X,\cdot)\} = \int_{\mathfrak{X}} \Psi(x,\cdot)\, dP_\theta(x).$$

Teorema 5.5 — Se um problema de decisão é invariante em relação a G, tem-se para todo o $\phi \in \Phi$ e todo o $g \in G$,

$$\hat{R}(\theta, \phi_g) = \hat{R}(\overline{g}\theta, \phi). \tag{5.20}$$

Dem. Por definição, $\hat{R}(\theta, \phi_g) = E_\theta\{L(\theta, \phi_g)\}$, onde, por (3.12), $L(\theta, \phi_g) = E\{L(\theta, Z)\}$, com Z variável aleatória assumindo valores em A com distribuição ϕ_g. A invariância da função perca implica, $L(\theta, \phi) = L(\overline{g}\theta, \tilde{g}\phi)$; portanto, por (5.19),

$$\begin{aligned} L[\theta, \phi_g(a \,|\, x)] &= L[\theta, \tilde{g}^{-1}\phi(a \,|\, gx)] \\ &= L[\overline{g}\theta, \phi(a \,|\, gx)]. \end{aligned}$$

Em consequência,

$$\begin{aligned} \hat{R}(\theta, \phi_g) &= E_\theta\{L[\overline{g}\theta, \phi(a \,|\, gX)]\} \\ &= E_{\overline{g}\theta}\{L[\overline{g}\theta, \phi(a \,|\, X)]\}, \end{aligned}$$

sendo a segunda relação justificada pela invariância de $\mathcal{F}$ — veja-se (5.15).

Finalmente,

$$\hat{R}(\theta, \phi_g) = \hat{R}(\overline{g}\theta, \phi),$$

como pretendia demonstrar-se. □□

Teorema 5.6 — Se um problema de decisão é invariante em relação a G, e $\phi \in \Phi$ é invariante em relação a G, então a função risco é constante em cada órbita de Θ.

Dem. Consequência imediata do teorema anterior e de ser $\phi = \phi_g$ para todo o $g \in G$ quando ϕ é invariante. □□

Um grupo de transformações, $\overline{G}$, em Θ, diz-se transitivo quando Θ consiste numa única órbita, isto é, quando quaisquer que sejam $\theta, \theta' \in \Theta$, existe sempre um $\overline{g} \in \overline{G}$, tal que $\theta' = \overline{g}\theta$.

Os Teoremas 5.3, 5.4 e 5.6 ganham especial importância quando $\overline{G}$ é um grupo transitivo, porquanto havendo uma única órbita as funções de decisão invariantes possuem risco constante sobre Θ. Assim, aquela que tiver risco mínimo é a melhor função de decisão invariante ou função de decisão IRM (invariante com risco mínimo). Quando $\overline{G}$ não é transitivo, função de decisão invariante óptima, é função de decisão invariante com risco uniformemente mínimo ou IRUM. Como não podia deixar de ser no presente texto contempla-se sobretudo o primeiro caso [para maior generalidade veja-se Lehmann (1983)].

Apresentam-se seguidamente alguns exemplos para ilustrar a doutrina que tem vindo a ser exposta. A determinação de funções de decisão invariantes óptimas é retomada de forma mais extensa nos dois capítulos seguintes.

Exemplo 5.9 — Retomando o Ex. 5.7, onde,

$$gx = N - x,$$
$$\overline{g}\theta = 1 - \theta,$$
$$\tilde{g}a = 1 - a,$$

conclui-se, por (5.17), que qualquer função de decisão pura invariante deve satisfazer,

$$\delta(N - x) = 1 - \delta(x).$$

Por outro lado repare-se que as órbitas em Θ são formadas pelos pares de pontos, $\{\theta, 1 - \theta\}$, para $0 \leqq \theta < 1/2$ e pelo ponto $\theta = 1/2$.

A propriedade da invariância não tem no caso da Binomial consequências de interesse. Como se refere na secção 5.7, a classe das funções de decisão invariantes é demasiado ampla. □

Exemplo 5.10 — O presente exemplo e o seguinte envolvem casos em que o grupo de transformações em Θ, $\overline{G}$, não é transitivo. O seu interesse reside sobretudo na ilustração que fazem das órbitas instituídas no espaço de resultados e no espaço do parâmetro.

Considere-se, com $i = 1, 2, \ldots, N$, X_i I.I.D., $X_i \sim N(\mu, \sigma^2)$,

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = (\sigma\sqrt{2\pi})^{-N} \exp\{-\Sigma(x_i - \mu)^2/2\sigma^2\},$$
$$\Theta = \{\boldsymbol{\theta} = (\mu, \sigma^2), -\infty < \mu < +\infty, \sigma > 0\}.$$

O problema consiste na estimação de $\mu - A = \mathbf{R}$ — com função perca,

$$L(\boldsymbol{\theta}, a) = W(|\,\mu - a\,|/\sigma),$$

onde W é qualquer função crescente com $|\,\mu - a\,|$.

Seja G_1 o grupo de transformações com elementos, g_α, $\alpha \in \mathbf{R}$, $\alpha \neq 0$,

$$g_\alpha \mathbf{x} = \alpha \mathbf{x}, \quad \mathbf{x} \in \mathbf{R}^N;$$

G_1 deixa o problema invariante, sendo,

$$\overline{g}_\alpha \boldsymbol{\theta} = (\alpha\mu, \alpha^2\sigma^2), \quad \tilde{g}_\alpha a = \alpha a.$$

Na Fig. 5.3 indica-se a configuração das órbitas em $\mathcal{X}$ (supondo $N = 2$) e em Θ,

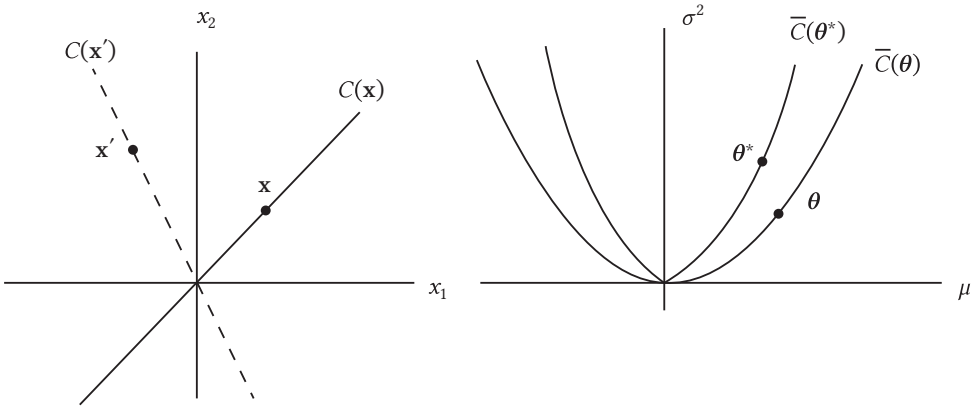


Fig. 5.3

Uma função de decisão, $\delta \in D$, para ser invariante em relação a G_1 deve satisfazer,

$$\delta(g_\alpha \mathbf{x}) = \tilde{g}_\alpha \delta(\mathbf{x}) \rightarrow \delta(\alpha \mathbf{x}) = \alpha \delta(\mathbf{x}).$$

É fácil de ver que $R(\boldsymbol{\theta}, \delta) = R(\overline{g}_\alpha \boldsymbol{\theta}, \delta)$ para todo o $\boldsymbol{\theta} \in \Theta$ e todo o $\overline{g}_\alpha \in \overline{G}_1$. Com efeito,

$$R(\overline{g}_\alpha \boldsymbol{\theta}, \delta) = E_{\overline{g}_\alpha \boldsymbol{\theta}}\{W[|\alpha\mu - \delta(\mathbf{X})|/|\alpha|\sigma]\},$$

donde, por (5.15),

$$\begin{aligned} R(\overline{g}_\alpha \boldsymbol{\theta}, \delta) &= E_{\boldsymbol{\theta}}\{W[|\alpha\mu - \delta(g_\alpha \mathbf{X})|/|\alpha|\sigma] \\ &= E_{\boldsymbol{\theta}}\{W[|\alpha\mu - \alpha\delta(\mathbf{X})|/|\alpha|\sigma] \\ &= R(\boldsymbol{\theta}, \delta). \end{aligned}$$

□

Exemplo 5.10 — *Continuação.* Considere-se o grupo de transformações, G_2, com elementos g_β, $\beta \in \mathbf{R}$,

$$g_\beta \mathbf{x} = \mathbf{x} + \beta \mathbf{1}, \quad \mathbf{x} \in \mathbf{R}^N, \quad \mathbf{1} = (1, 1, \ldots, 1);$$

G_2 deixa o problema invariante, com,

$$\overline{g}_\beta \boldsymbol{\theta} = (\mu + \beta, \sigma^2), \tilde{g}_\beta a = a + \beta.$$

Na Fig. 5.4 indica-se a configuração das órbitas em $\mathcal{X}$ e Θ.

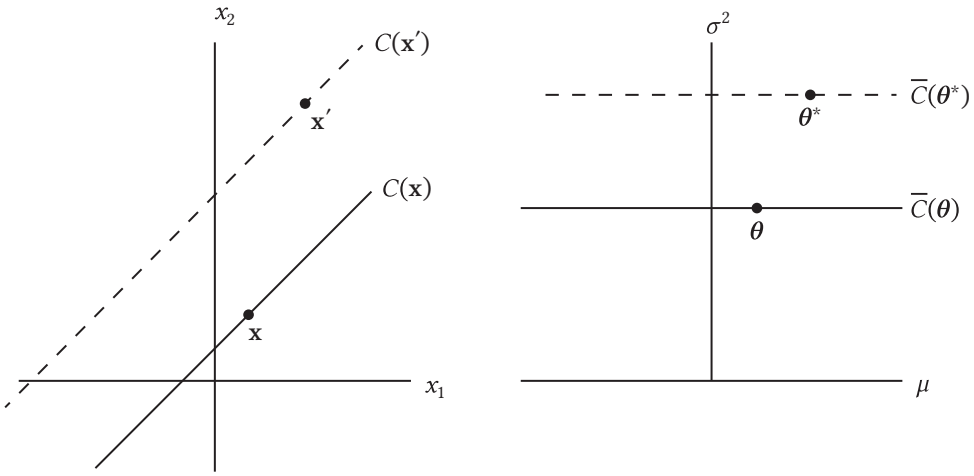


Fig. 5.4

Uma função de decisão, $\delta \in D$, para ser invariante em relação a G_2 deve verificar,

$$\delta(g_\beta \mathbf{x}) = \tilde{g}_\beta \delta(\mathbf{x}) \to \delta(\mathbf{x} + \beta \mathbf{1}) = \delta(\mathbf{x}) + \beta.$$

Tem-se, $R(\boldsymbol{\theta}, \delta) = R(\overline{g}_\beta \boldsymbol{\theta}, \delta)$ para todo o $\boldsymbol{\theta} \in \Theta$ e todo o $\overline{g}_\beta \in \overline{G}_2$, pois,

$$\begin{aligned} R(\overline{g}_\beta \boldsymbol{\theta}, \delta) &= E_{\overline{g}_\beta \boldsymbol{\theta}}\{W[|\, \mu + \beta - \delta(\mathbf{X})\,|\,/\sigma]\} \\ &= E_{\boldsymbol{\theta}}\{W[|\, \mu + \beta - \delta(g_\beta \mathbf{X})\,|/\sigma]\}, \end{aligned}$$

ou ainda,

$$\begin{aligned} R(\overline{g}_\beta \boldsymbol{\theta}, \delta) &= E_{\boldsymbol{\theta}}\{W[|\, \mu + \beta - \delta(\mathbf{X}) - \beta\,|\,/\sigma]\} \\ &= E_{\boldsymbol{\theta}}\{W[|\, \mu - \delta(\mathbf{X})\,|\,/\sigma]\} = R(\boldsymbol{\theta}, \delta). \end{aligned}$$

□

Exemplo 5.11 — No quadro do Ex. 5.1 considere-se o grupo de transformações, G, com elementos $g_c x = cx$, $c > 0$. Como é imediato, $Y = g_c X = cX$, tem função de densidade,

$$\begin{aligned} c^{-1} f(y/c\,|\,\theta) &= (c\theta)^{-1} \exp\{-y/c\theta\} \\ &= f(y\,|\,c\theta). \end{aligned}$$

Consequentemente, $\overline{g}_c \theta = c\theta$ é a transformação induzida em Θ, e a classe $\mathcal{F}$ das funções de densidade exponenciais é invariante em relação a G.

Facilmente se conclui também que $\tilde{g}_c a = ca$, porquanto,

$$L(\theta, a) = [1 - (a/\theta)]^2 = \left[1 - \frac{\tilde{g}_c a}{\overline{g}_c \theta}\right]^2$$
$$= L(\overline{g}_c \theta, \tilde{g}_c a);$$

segue-se, enfim, que o problema de decisão é invariante em relação a G.

As funções de decisão invariantes devem verificar,

$$\delta(g_c x) = \tilde{g}_c \delta(x),$$

isto é, devem ser do tipo,

$$\delta(x) = kx,$$

como aliás já se havia concluído [veja-se (5.4) e (5.5)].

O grupo de transformações em Θ, $\overline{G} = \{\overline{g}_c : c > 0\}$, é transitivo, isto é, quaisquer que sejam $\theta, \theta' \in \Theta$, é sempre possível encontrar um $c > 0$ tal que $\theta' = c\theta$. Assim, Θ consiste numa única órbita; com $\delta_k(x) = kx$, deve ser $R(\theta, \delta_k)$ constante sobre Θ para cada $k > 0$; com efeito,

$$R(\theta, \delta_k) = E_\theta\{[1 - (kX/\theta)]^2\}$$
$$= \int_0^\infty \theta^{-1}[1 - (kx/\theta)]^2 \exp\{-x/\theta\}\, dx$$
$$= 1 - 2k + 2k^2.$$

Designando por D_I a classe das funções de decisão invariantes, seja $D_I = \{\delta_k : \delta_k(x) = kx, k > 0\}$, a melhor função de decisão dentro da classe é a obtida com o valor de k que minimiza $1 - 2k + 2k^2$, isto é, $\delta_{1/2}(x) = x/2$ [Berger (1980)]. □

Exemplo 5.12 — Seja, $\mathbf{X} = (X_1, X_2)$, X_1 e X_2 variáveis aleatórias independentes com distribuição de Poisson e médias θ_1 e θ_2 respectivamente. Tem-se,

$$\Theta = \{(\theta_1, \theta_2) : \theta_1 > 0,\ \theta_2 > 0\},$$
$$\mathcal{X} = \{(x_1, x_2) : x_i = 0, 1, 2, \ldots, \quad i = 1, 2\}.$$

Suponha-se $A = \{0, 1\}$ e,

$$L[(\theta_1, \theta_2), 0] = \begin{cases} 1 & \text{se} \quad \theta_1 < \theta_2, \\ 0 & \text{se} \quad \theta_1 \geqq \theta_2, \end{cases} \qquad L[(\theta_1, \theta_2), 1] = \begin{cases} 1 & \text{se} \quad \theta_1 > \theta_2, \\ 0 & \text{se} \quad \theta_1 \leqq \theta_2. \end{cases}$$

Nestes termos, a acção $a = 0$ é correcta quando $\theta_1 > \theta_2$ e $a = 1$ é correcta se $\theta_1 < \theta_2$; são ambas correctas se $\theta_1 = \theta_2$.

Tem-se que o grupo de transformações, $G = \{e, g\}$, onde,

$$g(x_1, x_2) = (x_2, x_1),$$

deixa o problema invariante, com,

$$\overline{g}(\theta_1, \theta_2) = (\theta_2, \theta_1), \ \tilde{g}a = 1 - a.$$

A função de decisão pura, $\delta \in D$, é invariante se,

$$\delta[g(x_1, x_2)] = \tilde{g}\delta[(x_1, x_2)],$$

isto é, se,

$$\delta[(x_2, x_1)] = 1 - \delta[(x_1, x_2)].$$

Para $x_1 = x_2$ vem $\delta[(x_1, x_2)] = 1/2$ o que é impossível por ser $A = \{0, 1\}$; não existe função de decisão pura invariante nem, consequentemente, função de decisão mista invariante.

Defina-se à função de decisão aleatória, $\phi \in \Phi$, do modo seguinte:

$$\begin{aligned}
&\phi[0\,|\,(x_1, x_2)] = 1, \quad \phi[1\,|\,(x_1, x_2)] = 0 \quad \text{para } x_1 > x_2; \\
&\phi[0\,|\,(x_1, x_2)] = \frac{1}{2}, \quad \phi[1\,|\,(x_1, x_2)] = \frac{1}{2} \quad \text{para } x_1 = x_2; \\
&\phi[0\,|\,(x_1, x_2)] = 0, \quad \phi[1\,|\,(x_1, x_2)] = 1 \quad \text{para } x_1 < x_2.
\end{aligned} \tag{A.1}$$

Tem-se que ϕ é invariante uma vez que satisfaz (5.18). Note-se que a expressão geral das funções de decisão aleatórias invariantes é, tomando $0 \leqq \gamma \leqq 1$,

$$\begin{aligned}
&\phi_\gamma[0\,|\,(x_1, x_2)] = \gamma, \quad \phi_\gamma[1\,|\,(x_1, x_2)] = 1 - \gamma \quad \text{para } x_1 > x_2; \\
&\phi_\gamma[0\,|\,(x_1, x_2)] = \frac{1}{2}, \quad \phi_\gamma[1\,|\,(x_1, x_2)] = \frac{1}{2} \quad \text{para } x_1 = x_2; \\
&\phi_\gamma[0\,|\,(x_1, x_2)] = 1 - \gamma, \quad \phi_\gamma[1\,|\,(x_1, x_2)] = \gamma \quad \text{para } x_1 < x_2;
\end{aligned} \tag{A.2}$$

a correspondente função risco com $\boldsymbol{\theta} = (\theta_1, \theta_2)$,

$$R(\boldsymbol{\theta}, \phi_\gamma) = \begin{cases}
\gamma P_{\boldsymbol{\theta}}(X_1 > X_2) + \frac{1}{2} \cdot P_{\boldsymbol{\theta}}(X_1 = X_2) + (1 - \gamma) P_{\boldsymbol{\theta}}(X_1 < X_2) & \text{se } \theta_1 < \theta_2, \\
0, & \text{se } \theta_1 = \theta_2, \\
(1 - \gamma) P_{\boldsymbol{\theta}}(X_1 > X_2) + \frac{1}{2} \cdot P_{\boldsymbol{\theta}}(X_1 = X_2) + \gamma P_{\boldsymbol{\theta}}(X_1 < X_2) & \text{se } \theta_1 > \theta_2.
\end{cases}$$

Dado que,

$$\begin{aligned}
&P_{\boldsymbol{\theta}}(X_1 > X_2) < P_{\boldsymbol{\theta}}(X_1 < X_2) \quad \text{para } \theta_1 < \theta_2, \\
&P_{\boldsymbol{\theta}}(X_1 > X_2) > P_{\boldsymbol{\theta}}(X_1 < X_2) \quad \text{para } \theta_1 > \theta_2,
\end{aligned}$$

verifica-se que ϕ definido por [A.1], isto é, saído de [A.2] fazendo $\gamma = 1$, é a função de decisão aleatória invariante óptima [Ferguson (1967)]. □

Exemplo 5.13 — Suponha-se que $\mathcal{F}$ é a família das distribuições absolutamente contínuas (com função de densidade) em $\mathbf{R}$. Trata-se de um problema não paramétrico [Ferguson (1967)] em que $\mathcal{F}$ pode identificar-se com Θ. Seja $A = \mathbf{R}$ e,

$$L(F, a) = \left[F(a) - \frac{1}{2}\right]^2, \quad F \in \mathcal{F},\ a \in A,$$

a função perca. A experiência consiste em observar, X_i I.I.D., $i = 1, 2, \ldots, N$, com função de distribuição F, isto é, $P(X_i \leqq y) = F(y), i = 1, 2, \ldots, N$; o problema, como se verifica pela estrutura da função perca, é estimar a mediana da «verdadeira» função de distribuição.

O grupo de transformações G, com,

$$g_\psi(x_1, x_2, \ldots, x_N) = [\psi(x_1), \psi(x_2), \ldots, \psi(x_N)],$$

onde ψ é uma função contínua e crescente, aplicando $\mathbf{R}$ sobre $\mathbf{R}$, deixa o problema invariante, com,

$$\overline{g}_\psi F(y) = F[\psi^{-1}(y)], y \in \mathbf{R},$$
$$\tilde{g}_\psi(a) = \psi(a), \quad a \in A.$$

Se em vez de $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ se observa o vector das estatísticas de ordem, $\mathbf{Y} = [X_{(1)}, X_{(2)}, \ldots, X_{(N)}]$, com função de distribuição,

$$F_{\mathbf{Y}}(x_1, x_2, \ldots, x_N) = N! \prod_i F(x_i),$$

$x_1 \leqq x_2 \leqq \ldots \leqq x_N$, não se alteram $\mathcal{F}$ e A e o problema continua a ser invariante em relação a G.

A função de decisão pura,

$$\delta_j(\mathbf{x}) = \delta_j(x_1, x_2, \ldots, x_N) = x_{(j)}, \tag{5.21}$$

corresponde a tomar na amostra, $\mathbf{x}$, o valor que ocupa a ordem j-ésima quando os x_i se dispõem por ordem crescente; trata-se de uma função de decisão invariante[6] em relação a G, pois,

$$\delta_j[\psi(x_1), \psi(x_2), \ldots, \psi(x_N)] = \psi[x_{(j)}] = \psi[\delta_j(\mathbf{x})].$$

O cálculo da função risco,

$$R(F, \delta_j) = E_F\left\{\left[F(X_{(j)}) - \frac{1}{2}\right]^2\right\},$$

[6] De facto pode mostrar-se que $\delta_j(\mathbf{x})$, $j = 1, 2, \ldots, N$, são as únicas funções de decisão puras invariantes.

considerando a distribuição marginal de $X_{(j)}$, estatística de ordem j,

$$dG(x_{(j)}) = \frac{N!}{(j-1)!(N-j)!}[F(x_{(j)})]^{j-1}[1-F(x_{(j)})]^{N-j}dF(x_{(j)}),$$

conduz à expressão,

$$R(F,\delta_j) = \frac{N!}{(j-1)!(N-j)!}\int_{-\infty}^{+\infty}\left[F(x_{(j)})-\frac{1}{2}\right]^2[F(x_{(j)})]^{j-1}[1-F(x_{(j)})]^{N-j}dF(x_{(j)}).$$

Fazendo, $z = F(x_{(j)})$,

$$\begin{aligned} R(F,\delta_j) &= \frac{N!}{(j-1)!(N-j)!}\int_0^1\left(z-\frac{1}{2}\right)^2 z^{j-1}(1-z)^{N-j}dz \\ &= [(N+2)(N+1)]^{-1}\left[\left(j-\frac{N+1}{2}\right)^2\frac{N+1}{4}\right], \end{aligned}$$

resultado obtido recorrendo às propriedades da função Beta.

Assim, $\delta_j(\mathbf{x})$ minimiza $R(F,\delta_j)$ quando j é o maior inteiro contido em $(N+1)/2$, isto é, quando $x_{(j)}$ é a mediana[7] da amostra. Note-se a particularidade de a função risco não depender de F; note-se ainda que sendo as funções de decisão puras invariantes da forma (5.21) a que acaba de referir-se é a função de decisão pura invariante óptima. □

5.4 Função invariante máxima

O conceito de função invariante máxima que vai introduzir-se na presente secção destina-se, como a maior parte da matéria estudada no capítulo 5, a servir de base ao estudo a fazer nos dois capítulos seguintes. Como terá ocasião de verificar-se o conceito de invariante máxima facilita o tratamento de problemas invariantes quando a amostra não se reduz a uma observação.

Considere-se um espaço $\mathfrak{X}$ [não necessariamente um espaço da amostra] e um grupo G de transformações; a aplicação, $T\colon \mathfrak{X} \to \mathbf{R}^m$, $m \geqq 1$, diz-se função invariante máxima em relação a G quando verifica as duas condições seguintes:

[1] invariância:

$$T(gx) = T(x) \text{ para todo o } g \in G \text{ e todo o } x \in \mathfrak{X};$$

[2] maximalidade:

$$T(x_1) = T(x_2) \text{ implica } x_2 = gx_1 \text{ para algum } g \in G.$$

[7] Ou corresponde a uma possível definição de mediana.

A condição [1] estabelece que T é constante sobre cada órbita de $\mathcal{X}$; a condição [2] estabelece que T assume valores que diferem de órbita para órbita de $\mathcal{X}$.

Uma função que verifique [1] e não verifique [2] é invariante mas não invariante máxima, isto é, é constante sobre cada órbita mas assume o mesmo valor em mais do que uma órbita.

A distinção pode esclarecer-se recorrendo ao conceito de partição, aliás mais fundamental do que o conceito de função. Uma partição de $\mathcal{X}$, seja $\mathcal{U}$, é invariante se, para qualquer dos seus elementos, $U \in \mathcal{U}$, $x \in U$ implica $gx \in U$. Uma partição é invariante máxima quando resulta da intersecção de todas as partições invariantes, quer dizer, quando é a mais fina das partições invariantes. Uma função é invariante máxima quando gera em $\mathcal{X}$ uma partição invariante máxima; quando T é função invariante máxima, qualquer função biunívoca de T continua a ser invariante máxima.

Quando $\mathcal{X}$ é espaço de resultados e se observa a variável (escalar ou vector) aleatória, X, assumindo valores, $x \in \mathcal{X}$, $T(X)$ diz-se estatística invariante máxima se a função T satisfaz as condições [1] e [2].

Exemplo 5.14 — Suponha-se $\mathcal{X} = \mathbf{R}^N$ e G consistindo no grupo de translações,

$$g_c(x_1, x_2, \dots, x_N) = (x_1 + c, x_2 + c, \dots, x_N + c), \quad -\infty < c < +\infty.$$

A função,

$$T(\mathbf{x}) = (x_1 - x_N, x_2 - x_N, \dots, x_{N-1} - x_N), \quad T \in \mathbf{R}^{N-1},$$

é invariante máxima. Com efeito, verifica-se [1]:

$$T[g_c(\mathbf{x})] = T(\mathbf{x}),$$

isto é,

$$T(x_1 + c, x_2 + c, \dots, x_N + c) = T(x_1, x_2, \dots, x_N);$$

e verifica-se [2]:

$$T(\mathbf{x}) = T(\mathbf{x}') \text{ implica } x_i - x_N = x_i' - x_N', \quad i = 1, 2, \dots, N-1,$$

sendo, $g_c(\mathbf{x}') = \mathbf{x}$ quando se toma $c = x_N - x_N'$.

A função T não é a única invariante máxima; outro exemplo é,

$$T^*(\mathbf{x}) = (x_1 - x_2, x_2 - x_3, \dots, x_{N-1} - x_N).$$

□

Exemplo 5.15 — Com $\mathfrak{X} = \mathbf{R}^N$ e G consistindo no grupo de mudanças de escala,

$$g_c(x_1, x_2, \dots, x_N) = (cx_1, cx_2, \dots, cx_N), \quad c > 0,$$

tome-se,

$$z^2 = \sum x_i^2, \quad T(\mathbf{x}) = \begin{cases} 0 & \text{se } z = 0, \\ (x_1/z, x_2/z, \dots, x_N/z) & \text{se } z \neq 0. \end{cases}$$

A função T é invariante máxima:

[1] — $T(cx_1, cx_2, \dots, cx_N) = T(x_1, x_2, \dots, x_N)$;

[2] — se $T(\mathbf{x}) = T(\mathbf{x}')$, vem,

$$T(\mathbf{x}) = T(\mathbf{x}') = 0 \Rightarrow x_i = x_i' = 0,\ i = 1, 2, \dots, N;$$
$$T(\mathbf{x}) = T(\mathbf{x}') \neq 0 \Rightarrow x_i/z = x'/z',\ i = 1, 2, \dots, N,$$
$$\text{e } g_c(\mathbf{x}') = \mathbf{x} \text{ com } c = z/z'. \qquad \square$$

Exemplo 5.16 — Com $\mathfrak{X} = \mathbf{R}^N$ e G o grupo de transformações,

$$g_{\alpha,\beta}(x_1, x_2, \dots, x_N) = (\alpha + \beta x_1, \alpha + \beta x_2, \dots, \alpha + \beta x_N), \quad -\infty < \alpha < +\infty,\ 0 < \beta < +\infty,$$

defina-se,

$$\overline{x} = (1/N) \sum_{i=1}^{N} x_i \quad \text{e} \quad s^2 = (1/N) \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

É simples exercício mostrar que a função,

$$T(\mathbf{x}) = \begin{cases} 0 & \text{se } s = 0, \\ [(x_1 - \overline{x})/s, (x_2 - \overline{x})/s, \dots, (x_N - \overline{x})/s] & \text{se } s \neq 0, \end{cases}$$

é invariante máxima. □

Os dois teoremas seguintes estão na base do emprego das funções invariantes máximas:

Teorema 5.7 — Considere-se um espaço $\mathfrak{X}$ e um grupo de transformações, G, em relação ao qual $T(x)$ é função invariante máxima. A função $W(x)$ é invariante em relação a G se e somente se W é função de $T(x)$.

Dem. Se $W(x) = w[T(x)]$, sai,

$$W(gx) = w[T(gx)] = w[T(x)] = W(x),$$

e a condição é suficiente.

Se $W(x)$ é invariante, suponha-se $T(x_1) = T(x_2)$; então, $x_2 = gx_1$ para algum $g \in G$, e,

$$W(x_2) = W(gx_1) = W(x_1),$$

o que implica ser W função de $T(x)$: a condição é necessária. □□

Teorema 5.8 — Considere-se um problema de decisão invariante em relação a um grupo de transformações, G. Seja $\tau(\theta)$ invariante máxima em Θ em relação ao grupo induzido $\overline{G}$. Então, se $T(x)$ é função invariante em relação a G, a distribuição da estatística, $T(X)$, depende apenas de $\tau(\theta)$.

Dem. Para qualquer conjunto A (mensurável), do domínio de T, tem-se,

$$\begin{aligned} P_{\overline{g}\theta}[T(X) \in A] &= P_\theta[T(gX) \in A] \quad \text{(por (5.14))} \\ &= P_\theta[T(X) \in A] \quad \text{(invariância de } T(x)\text{)}; \end{aligned}$$

portanto, $\eta(\theta) = P_\theta[T(X) \in A]$ é função invariante em Θ. Pelo teorema anterior [com $\mathfrak{X}$ e G subsituídos por Θ e $\overline{G}$] segue-se que $\eta(\theta)$ é função de $\tau(\theta)$; como A é arbitrário, a demonstração fica concluída. □□

5.5 Medidas Haar invariantes[8]

A medida Haar invariante é definida em grupos topológicos localmente compactos. Um grupo topológico é um grupo abstracto que é também espaço topológico verificando-se ainda que a operação do grupo é contínua.

O estudo muito restrito que pode aqui fazer-se dirige-se apenas aos casos em que o grupo topológico, seja Ω, é subconjunto de um espaço euclideano, $\Omega \subset \mathbf{R}^k$ dotado da métrica habitual, com medida à Lebesgue positiva.

Ficando automaticamente garantido que Ω é localmente compacto [todo o ponto $\omega \in \Omega$ tem uma vizinhança aberta cujo fecho é conjunto compacto, isto é, no caso presente, conjunto limitado e fechado], convém enunciar o que se entende por operação contínua:

$$\lim_{n\to\infty} \omega_n \circ \omega'_n = \omega_0 \circ \omega'_0 \quad \text{e} \quad \lim_{n\to\infty} \omega_n^{-1} = \omega_0^{-1}$$

sempre que,

$$\lim_{n\to\infty} \omega_n = \omega_0 \quad \text{e} \quad \lim_{n\to\infty} \omega'_n = \omega'_0.$$

Com $\omega \in \Omega$ e $E \subset \Omega$, defina-se,

$$\begin{aligned} \omega \circ E &= \{\omega \circ \xi \,:\, \xi \in E\}, \\ E \circ \omega &= \{\xi \circ \omega \,:\, \xi \in E\}; \end{aligned}$$

os conjuntos, $\omega \circ E$ e $E \circ \omega$ são designados por translação esquerda e translação direita de E por ω.

[8] A leitura desta secção não é essencial para acompanhamento do texto.

Exemplo 5.17 — Seja Ω o grupo de transformações afins em $\mathbf{R}$:

$$\omega \equiv (\alpha, \beta) \in \Omega \quad \Rightarrow \quad \omega(z) = \alpha + \beta z,\ z \in \mathbf{R},$$

com,

$$\Omega = \{\omega \equiv (\alpha, \beta) \,:\, -\infty < \alpha < +\infty, 0 < \beta < +\infty\},$$

e $\Omega \subset \mathbf{R}^2$.

Como imediatamente se obtém,

$$\begin{aligned} \omega_1 \circ \omega &= (\alpha_1, \beta_1) \circ (\alpha, \beta) \\ &= (\alpha_1 + \beta_1\alpha, \beta_1\beta), \\ \omega \circ \omega_1 &= (\alpha, \beta) \circ (\alpha_1, \beta_1) \\ &= (\alpha + \beta\alpha_1, \beta\beta_1). \end{aligned}$$

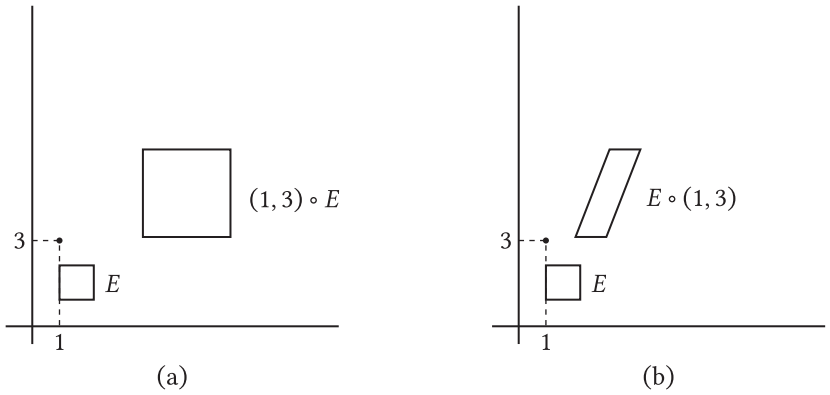


Fig. 5.5

Suponha-se

$$E = \{(\alpha, \beta) \,:\, 1 \leqq \alpha \leqq 2,\ 1 \leqq \beta \leqq 2\} \subset \Omega,$$

e $\omega = (1, 3)$; vem,

$$\begin{aligned} (1, 3) \circ E &= \{(1 + 3\alpha, 3\beta) \,:\, 1 \leqq \alpha \leqq 2,\ 1 \leqq \beta \leqq 2\} \\ &= \{(\alpha', \beta') \,:\, 4 \leqq \alpha' \leqq 7,\ 3 \leqq \beta' \leqq 6\}; \\ E \circ (1, 3) &= \{(\alpha + \beta, 3\beta) \,:\, 1 \leqq \alpha \leqq 2,\ 1 \leqq \beta \leqq 2\} \\ &= \{(\alpha', \beta') \,:\, 3 \leqq 3\alpha' - \beta' \leqq 6,\ 3 \leqq \beta' \leqq 6\}; \end{aligned}$$

os conjuntos $(1.3) \circ E$ e $E \circ (1, 3)$ são indicados na Fig. 5.5, (a) e (b), respectivamente.

□

Uma medida ν definida nos conjuntos (borelianos) do grupo topológico localmente compacto, Ω, diz-se Haar invariante esquerda quando satisfaz as seguintes condições:

{H1} — invariante em relação às translações esquerda, isto é, para todo o E (boreliano) $\subset \Omega$ e todo o $\omega \in \Omega$,

$$\nu(\omega \circ E) = \nu(E);$$

{H2} — para qualquer conjunto compacto (no caso presente, limitado e fechado), $C \subset \Omega$, $\nu(C) < \infty$;

{H3} — para qualquer conjunto aberto, $\mathcal{O} \neq \emptyset$, $\mathcal{O} \subset \Omega$, $\nu(\mathcal{O}) > 0$.

Uma medida Haar invariante direita define-se com a óbvia alteração de {H1}, isto é, pedindo invariância em relação às translacções direita.

Exemplo 5.18 — Quando $\Omega = \mathbf{R}$ e a operação do grupo é a adição, encontrando-se Ω dotado da topologia habitual, tem-se que a medida Haar invariante corresponde à medida à Lebesgue.

Com $E = \{\xi \,:\, a \leqq \xi \leqq b\} \subset \mathbf{R}$ e $\omega_1 \in \mathbf{R}$, tem-se

$$\begin{aligned} E \circ \omega_1 &= \{\omega_1 + \xi \,:\, a \leqq \xi \leqq b\} \\ &= \{\xi' \,:\, \omega_1 + a \leqq \xi' \leqq \omega_1 + b\}, \end{aligned}$$

sendo $\omega_1 \circ E = E \circ \omega_1$.

Designando, μ, a medida à Lebesgue em $\mathbf{R}$, sai,

$$\mu(E) = \mu(E \circ \omega_1) = \mu(\omega_1 \circ E) = b - a;$$

quer dizer, a medida à Lebesgue em $\mathbf{R}$ (ou $\mathbf{R}^k$) é medida Haar invariante (esquerda e direita)[9]. □

Exemplo 5.19 — Com $\Omega = (0, \infty)$, Ω é um grupo em relação à multiplicação.

Com, $0 < a < b$, $E = \{\xi \,:\, a \leqq \xi \leqq b\} \subset \Omega$, tome-se $\omega_1 \in \Omega$; vem,

$$\begin{aligned} E \circ \omega_1 &= \{\omega_1 \xi \,:\, a \leqq \xi \leqq b\} \\ &= \{\xi' \,:\, \omega_1 a \leqq \xi' \leqq \omega_1 b\} = \omega_1 \circ E. \end{aligned}$$

Se ν é medida Haar invariante (esquerda e direita),

$$\nu(E) = \nu(E \circ \omega_1) = \nu(\omega_1 \circ E), \quad E \subset \Omega,\ \omega_1 \in \Omega;$$

[9] É suficiente fazer a verificação para intervalos.

verifica-se que ν goza desta propriedade[10] quando a respectiva densidade (generalizada) em relação à medida à Lebesgue assume a forma, $h(\omega) = 1/\omega$ [quer dizer, a derivada de Radon-Nikodym, $d\nu/d\mu = 1/\omega$],

$$\begin{aligned}\nu(E) &= \int_E (1/\omega)\, d\mu \\ &= \int_E (1/\omega)\, d\omega;\end{aligned}$$

com efeito,

$$\nu(E) = \nu(E \circ \omega_1) = \nu(\omega_1 \circ E) = \log(b/a).$$

□

Repare-se que nos dois exemplos anteriores a medida Haar invariante atribui massa infinita a todo o espaço: $\nu(\Omega) = \infty$. Não é assim possível proceder a uma normalização que a converta em medida de probabilidade; a medida Haar invariante pode quando muito associar-se a distribuições impróprias.

Quando Ω é espaço compacto [propriedade que se verifica, por exemplo, quando $\Omega = [0, 1]$, mas não se verifica quando $\Omega = (-\infty, +\infty)$ ou $\Omega = (0, \infty)$], as medidas Haar invariantes esquerda e direita podem converter-se em medidas de probabilidade porquanto $\nu(\Omega)$ é finito.

Exemplo 5.20 — Com $\Omega = [0, 1]$, Ω é um grupo em relação à adição (módulo 1). A medida Haar invariante (esquerda ou direita) corresponde à distribuição uniforme em $[0, 1]$. Esta distribuição atribui a mesma probabilidade ao intervalo $[a, b] \subset [0, 1]$ e, com $\omega \in [0, 1]$, àquele dos conjuntos seguintes,

$$\begin{gathered}[a + \omega, b + \omega], \\ [a + \omega, 1] \cap [0, b + \omega - 1], \\ [a + \omega - 1, b + \omega - 1],\end{gathered}$$

que esteja contido no intervalo $[0, 1]$ [Watcher (1983)]. □

Para introduzir [no contexto restrito em que o grupo topológico localmente compacto, Ω, é subconjunto de $\mathbf{R}^k$] as densidades[10] Haar invariantes [isto é, as densidades em relação à medida à Lebesgue das medidas Haar invariantes], considerem-se, com ω_1 fixo, $\omega_1 \in \Omega$, as transformações,

$$\omega \to \omega_1 \circ \omega,\ \ \omega \to \omega \circ \omega_1,\ \ \omega \in \Omega,$$

e admita-se que existem as respectivas derivadas e, portanto, os Jacobianos [quase por toda a parte em Ω].

[10] Densidades generalizadas, i.e., de distribuições impróprias.

Designe-se $H^e_{\omega_1}(\omega)$ a matriz $[k \times k$ se $\Omega \subset \mathbf{R}^k]$, $[\partial(\omega_1 \circ \omega)/\partial\omega]$; o Jacobiano da transformação, $\omega \to \omega_1 \circ \omega$, escreve-se então,

$$J^e_{\omega_1}(\omega) = |\, H^e_{\omega_1}(\omega)\,|.$$

Analogamente,

$$H^d_{\omega_1}(\omega) = [\partial(\omega \circ \omega_1)/\partial\omega],\ J^d_{\omega_1}(\omega) = |\, H^d_{\omega_1}(\omega)\,|,$$

para a transformação, $\omega \to \omega \circ \omega_1$.

Considerando {H1}, a densidade Haar invariante esquerda, seja $h^e(\omega)$, é definida pela relação,

$$\int_E h^e(\omega)\, d\omega = \int_{\omega_1 \circ E} h^e(\xi)\, d\xi, \tag{5.22}$$

para todo o E (boreliano) $\subset \Omega$ e todo o $\omega_1 \in \Omega$. Analogamente, a densidade Haar invariante direita, seja $h^d(\omega)$, é definida por,

$$\int_E h^d(\omega)\, d\omega = \int_{E \circ \omega_1} h^d(\xi)\, d\xi. \tag{5.23}$$

Pode mostrar-se que as densidades Haar invariantes esquerda e direita existem e são únicas a menos de uma constante multiplicativa [Halmos (1950)].

Para calcular $h^e(\omega)$, tome-se o integral do segundo membro de (5.22) e faça-se a mudança de variável, $\xi = \omega_1 \circ \omega$; vem,

$$\int_{\omega_1 \circ E} h^e(\xi)\, d\xi = \int_E h^e(\omega_1 \circ \omega) J^e_{\omega_1}(\omega)\, d\omega.$$

Comparando com o primeiro membro de (5.22) sai,

$$h^e(\omega) = h^e(\omega_1 \circ \omega)\, J^e_{\omega_1}(\omega);$$

fazendo ω igual ao elemento identidade de Ω, seja $\mathcal{E}$,

$$h^e(\omega_1) = h^e(\mathcal{E})/J^e_{\omega_1}(\mathcal{E}).$$

Dado que $h^e(\mathcal{E})$ é uma constante, repondo ω no lugar de ω_1, tem-se finalmente,

$$h^e(\omega) = 1/J^e_{\omega}(\mathcal{E}),\ \omega \in \Omega, \tag{5.24}$$

para densidade Haar invariante esquerda.

Analogamente,

$$h^d(\omega) = 1/J^d_{\omega}(\mathcal{E}),\ \omega \in \Omega, \tag{5.25}$$

é a densidade Haar invariante direita.

Exemplo 5.17 — *Continuação.* A transformação, $\omega \to \omega_1 \circ \omega$, é, neste caso,

$$(\alpha, \beta) \to (\alpha_1 + \beta_1\alpha, \beta_1\beta),$$

donde,

$$H^e_{(\alpha_1,\beta_1)}[(\alpha, \beta)] = \begin{bmatrix} \beta_1 & 0 \\ 0 & \beta_1 \end{bmatrix},$$

com o Jacobiano,

$$J^e_{(\alpha_1,\beta_1)}[(\alpha, \beta)] = \beta_1^2.$$

Como $J^e_{(\alpha_1,\beta_1)}[(0, 1)] = \beta_1^2$, tem-se que a densidade Haar invariante esquerda é no caso presente,

$$h^e(\omega) \equiv h^e(\alpha, \beta) = 1/\beta^2, \quad (\alpha, \beta) \in \Omega.$$

Por outro lado,

$$H^d_{(\alpha_1,\beta_1)}[(\alpha, \beta)] = \begin{bmatrix} 1 & \alpha_1 \\ 0 & \beta_1 \end{bmatrix},$$

com o Jacobiano,

$$J^d_{(\alpha_1,\beta_1)}[(\alpha, \beta)] = \beta_1.$$

Como $J^d_{(\alpha_1,\beta_1)}[(0, 1)] = \beta_1$, tem-se que a densidade Haar invariante direita é,

$$h^d(\omega) \equiv h^d(\alpha, \beta) = 1/\beta, \quad (\alpha, \beta) \in \Omega.$$

Note-se que h^e e h^d são densidades impróprias e diferentes [Berger (1980)]. □

A teoria das densidades Haar invariantes é importante no contexto das distribuições a priori não informativas. Para explorar essa ideia convém estudar mais um exemplo.

Exemplo 5.21 — Seja X uma v. a. Normal com média μ e desvio padrão σ. Tem-se $\mathcal{X} = \mathbf{R}$ e,

$$\Theta = \{(\mu, \sigma) : -\infty < \mu < +\infty,\ 0 < \sigma < +\infty\}.$$

O grupo de transformações,

$$G = \{g_{\alpha,\beta} : -\infty < \alpha < +\infty,\ 0 < \beta < +\infty\},$$

com,

$$g_{\alpha,\beta}x = \alpha + \beta x,$$

induz em Θ o grupo de transformações,

$$\overline{G} = \{\overline{g}_{\alpha,\beta} : -\infty < \alpha < +\infty,\ 0 < \beta < +\infty\},$$

com,

$$\overline{g}_{\alpha,\beta}(\mu,\sigma) = (\alpha + \beta\mu, \beta\sigma).$$

O que interessa destacar neste exemplo é a correspondência biunívoca que existe entre $\overline{G}$ e Θ [quer dizer, $\overline{G}$ e Θ são isomorfos]. □

Exemplo 5.21 — *Continuação.* Considere-se uma amostra casual de dimensão N e a estatística suficiente mínima para (μ, σ), $T = (\overline{X}, S)$. Seja, G, o grupo de transformações em $\mathfrak{J} = \{(\overline{x}, s) : -\infty < \overline{x} < +\infty,\ 0 < s < +\infty\}$ (que está no lugar de $\mathfrak{X}$) com elementos,

$$g_{\alpha,\beta}(\overline{x}, s) = (\alpha + \beta\overline{x}, \beta s);$$

é fácil de ver que neste caso, $\mathfrak{J}$, G, Θ, $\overline{G}$ podem considerar-se o mesmo grupo, isto é, são isomorfos. □

1) Admita-se que $\overline{G}$ e Θ são isomorfos. Tratando-se do mesmo espaço, podem designar-se as densidades Haar invariantes sobre Θ por $h^e(\theta)$ e $h^d(\theta)$. Quer h^e, quer h^d, podem tomar-se como distribuições a priori (impróprias) não informativas. Para esclarecer este ponto convém fazer o confronto entre os Exs. 5.18 e 5.19 e as considerações [páginas 68 e 69] feitas para justificar (1.40) e (1.41); convém ainda comparar o Ex. 5.17 (parte final) com o Ex. 1.27. De facto, $h^e(\mu, \sigma) = 1/\sigma^2$ e $h^d(\mu, \sigma) = 1/\sigma$, correspondem às distribuições a priori não informativas consideradas por Jeffreys que acabou por optar pela densidade Haar invariante direita. Mas qual é, então, a escolha correcta?

2) Para avançar um pouco na resposta à questão tem de admitir-se que $\mathfrak{X}$, G, Θ e $\overline{G}$ são isomorfos. Consegue então provar-se [veja-se Berger (1980) para maior desenvolvimento] que a melhor função de decisão invariante e a função de decisão Bayes (generalizada) contra h^d coincidem. Assim, pelo menos nos problemas invariantes, a escolha da distribuição a priori não informativa deve recair na densidade Haar invariante direita. Outra consequência notável é a seguinte: numa vasta classe de problemas invariantes pode ser muito mais mais fácil determinar a função de decisão Bayes (generalizada) do que determinar a melhor função de decisão invariante, havendo portanto um caminho alternativo e mais acessível para chegar a esta [Berger (1980) indica referências para aprofundamento da matéria].

5.6 Funções de decisão invariantes minimax e admissíveis

A ideia de reduzir a classe D^* com o propósito de facilitar a escolha leva a pesquisar funções de decisão que sejam óptimas ou, pelo menos notáveis, na subclasse das invariantes. Recordam-se, a propósito, os Exs. 5.11, 5.12 e 5.13, embora as principais aplicações do conceito de invariância sejam feitas nos próximos dois capítulos.

Quando uma função de decisão goza de certas propriedades dentro da subclasse das invariantes, interessa naturalmente averiguar em que medida essas propriedades se mantêm na classe mais vasta composta pela totalidade das funções de decisão. Por exemplo, a melhor função de decisão invariante, se existir, é com certeza minimax dentro da classe das invariantes; mas, será minimax dentro da classe universal?

As condições em que as propriedades minimax e admissibilidade se conservam são especialmente importantes. Infelizmente somente podem aqui estudar-se com algum pormenor situações que envolvem invariância em relação a grupos finitos, $G = \{g_1, g_2, \dots, g_M\}$.

Teorema 5.9 — Se um problema de decisão é invariante em relação a um grupo finito de transformações, G, então, se existir uma função de decisão minimax, existe uma função de decisão invariante que é minimax. Se uma função de decisão é minimax na classe invariante também é minimax na classe universal.

Dem. Consiste em mostrar que para todo o $\phi \in \Phi$ existe uma função de decisão invariante, $\phi^* \in \Phi$, tal que,

$$\sup_\theta \hat{R}(\theta, \phi^*) \leqq \sup_\theta \hat{R}(\theta, \phi).$$

Seja, $G = \{g_1, g_2, \dots, g_M\}$, o grupo finito e com qualquer, $\phi \in \Phi$, defina-se,

$$\begin{aligned} \phi^*(a \mid x) &= \frac{1}{M} \sum_{m=1}^{M} \phi_{g_m}(a \mid x) \\ &= \frac{1}{M} \sum_{m=1}^{M} \tilde{g}_m^{-1} \phi(a \mid g_m x), \end{aligned} \tag{5.26}$$

onde ϕ_{g_m} corresponde a (5.19).

Começa por verificar-se que ϕ^* é invariante.

Sabe-se que, com qualquer $g_i \in G$, $\{g_1 \circ g_i, g_2 \circ g_i, \dots, g_M \circ g_i\}$, é uma permutação de $\{g_1, g_2, \dots, g_M\}$; logo, para todo o x,

$$\phi^*(a \mid x) = \frac{1}{M} \sum_{m=1}^{M} (\tilde{g}_m \circ \tilde{g}_i)^{-1} \phi(a \mid (g_m \circ g_i)x). \tag{5.27}$$

Em termos de probabilidades no espaço de acções o raciocínio é mais fácil de acompanhar. Seja, A_0 (mensurável) $\subset A$; tem-se que (5.22) é equivalente à relação,

$$P^*(Z \in A_0 \mid x) = \frac{1}{M} \sum_{m=1}^{M} P(Z \in \tilde{g}_m A_0 \mid g_m x),$$

e que (5.27) é equivalente a,

$$P^*(Z \in A_0 \,|\, x) = \frac{1}{M} \sum_{m=1}^{M} P[Z \in (\tilde{g}_m \circ \tilde{g}_i)A_0 \,|\, (g_m \circ g_i)x]$$
$$= P^*[Z \in \tilde{g}_i A_0 \,|\, g_i x].$$

Atendendo a (5.18′′) conclui-se que ϕ^* é invariante.

Por outro lado,

$$\sup_\theta \hat{R}(\theta, \phi^*) = \sup_\theta \left(\frac{1}{M}\right) \sum_{m=1}^{M} \hat{R}(\theta, \phi_{g_m}),$$

ou seja,

$$\sup_\theta \hat{R}(\theta, \phi^*) \leqq \sup_\theta \frac{1}{M} \sum_{m=1}^{M} \sup_\theta \hat{R}(\theta, \phi_{g_m})$$
$$= \frac{1}{M} \sum_{m=1}^{M} \sup_\theta \hat{R}(\overline{g}_m \theta, \phi) \qquad \text{[Teorema 5.5]}$$
$$= \frac{1}{M} \sum_{m=1}^{M} \sup_\theta \hat{R}(\theta, \phi)$$
$$= \sup_\theta \hat{R}(\theta, \phi),$$

como era preciso demonstrar. □□

A propriedade estabelecida pelo teorema anterior pode provar-se em condições mais gerais. Assim,

[a] Se G é grupo topológico compacto o Teorema 5.9 ainda é verdadeiro, devendo substituir-se na expressão (5.26), $(1/M)\sum$ por dv, onde v é a medida Haar invariante.

[b] Se G é grupo topológico localmente compacto, mas não compacto, a medida Haar invariante não pode associar-se a uma medida de probabilidade e a substituição referida em [a] não pode fazer-se. Introduzindo condições adicionais sobre a natureza do grupo G, Hunt e Stein demonstraram que as funções de decisão invariantes minimax continuam a ser minimax na classe universal. O Teorema de Hunt-Stein é aflorado no capítulo 7 (veja-se a secção 7.8). No Teorema 6.8, que diz respeito ao grupo de translações em **R**, mostra-se que a melhor função de decisão invariante é minimax desde que se introduzam condições adicionais de outro tipo [veja-se a secção 6.3 e Ferguson (1967)].

O exemplo seguinte devido a Stein mostra que o Teorema 5.9 não se aplica a grupos não compactos, isto é, que nem sempre uma função de decisão minimax é invariante.

Exemplo 5.22 — Suponha-se que $\mathbf{X}$ e $\mathbf{Y}$ são independentes e possuem distribuição Normal k-dimensional: $\mathbf{X} \sim N_k(\mathbf{0}, \Sigma)$, $\mathbf{Y} \sim N_k(\mathbf{0}, \Delta\Sigma)$, $k \geqq 2$; tem-se,

$$\Theta = \{(\Delta, \Sigma) \,:\, \Delta > 0 \text{ e } \Sigma \text{ matriz covariância não singular}\}.$$

O problema consiste na estimação de Δ — portanto $A = (0, \infty)$ — com função perca,

$$L[(\Delta, \Sigma), a] = L(\Delta, a) = \begin{cases} 0 & \text{se } |\,\Delta - a\,| \leqq \Delta/2, \\ 1 & \text{se } |\,\Delta - a\,| > \Delta/2. \end{cases}$$

Considere-se o grupo G de transformações,

$$g_B(\mathbf{x}, \mathbf{y}) = (B\mathbf{x}, B\mathbf{y}),$$

onde B é qualquer matriz $k \times k$ não singular, e $g_{B_1} \circ g_{B_2} = g_{B_1 B_2}$.

Mostra-se sem dificuldade que o problema é invariante em relação a G, sendo,

$$\overline{g}_B(\Delta, \Sigma) = (\Delta, B\Sigma B'), \tilde{g}_B(a) = a.$$

As funções de decisão invariantes devem satisfazer a condição,

$$\delta(B\mathbf{x}, B\mathbf{y}) = \delta[g_B(\mathbf{x}, \mathbf{y})] = \tilde{g}_B \delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x}, \mathbf{y}), \tag{5.28}$$

para todo o B, $\mathbf{x}$ e $\mathbf{y}$. Fazendo,

$$\mathbf{x} = \mathbf{e}_1 = (1, 0, \ldots, 0)' \text{ e } \mathbf{y} = \mathbf{e}_k = (0, 0, \ldots, 1)',$$

e designando por $\mathbf{b}^1$ e $\mathbf{b}^k$ a primeira e última coluna de B, respectivamente, obtém-se de (5.28),

$$\delta(\mathbf{b}^1, \mathbf{b}^k) = \delta(\mathbf{e}_1, \mathbf{e}_k).$$

A única restrição em relação a $\mathbf{b}^1$ e $\mathbf{b}^k$ é não serem múltiplos pois caso contrário B era singular. Conclui-se, portanto, que $\delta(\mathbf{x}, \mathbf{y}) = C$ (constante) com probabilidade um. A função risco de tal função de decisão assume a forma,

$$R[(\Delta, \Sigma), \delta] = \begin{cases} 0 & \text{se } |\,\Delta - C\,| \leqq \Delta/2, \\ 1 & \text{se } |\,\Delta - C\,| > \Delta/2, \end{cases}$$

e, assim,

$$\sup_{(\Delta, \Sigma)} R[(\Delta, \Sigma), \delta] = 1.$$

Qualquer função de decisão invariante tem risco minimax igual a 1 [$\delta(\mathbf{x}, \mathbf{y}) = C$ é função de decisão pura invariante mas pode verificar-se que a casualização não altera o supremo acima].

Por outro lado considere-se a função de decisão,

$$\delta_0(\mathbf{x}, \mathbf{y}) = |\, y_1/x_1 \,|;$$

tem-se,

$$\begin{aligned} R[(\Delta, \sharp), \delta_0] &= E_{(\Delta,\sharp)}\{L[(\Delta, \sharp), \delta_0(\mathbf{X}, \mathbf{Y})]\} \\ &= P_{(\Delta,\sharp)}[|\Delta - \delta_0(\mathbf{X}, \mathbf{Y})\,| > \Delta/2] \\ &= P_{(\Delta,\sharp)}\{|\,1 - [(|\,Y_1/\Delta\sigma_{11}\,|)/(|\,X_1/\sigma_{11}\,|)]\,| > 1/2\} \\ &= P_{(\Delta,\sharp)}[|\,1 - (|\,Z\,|/|\,Z^*\,|)\,| > 1/2], \end{aligned}$$

onde σ_{11} é o desvio padrão de X_1, $Z = Y_1/\Delta\sigma_{11}$ e $Z^* = X_1/\sigma_{11}$ são variáveis aleatórias independentes $N(0, 1)$. Esta última probabilidade é constante [independente de $(\Delta, \sharp)$] e menor do que um. Consequentemente, a função de decisão não invariante, δ_0, tem risco minimax inferior ao risco minimax de qualquer função de decisão invariante. □

A questão posta para a propriedade minimax pode também formular-se para a admissibilidade. Se a invariância é em relação a um grupo finito a resposta encontra-se no,

Teorema 5.10 — Se um problema de decisão é invariante em relação a um grupo finito de transformações, G, então, função de decisão admissível na classe invariante é admissível na classe universal.

Dem. Seja Φ_I a subclasse das funções de decisão aleatórias invariantes. Suponha-se que $\phi' \in \Phi_I$ é admissível em Φ_I mas não é admissível na classe universal, Φ. Existe, $\phi \in \Phi$, tal que $\hat{R}(\theta, \phi) \leqq \hat{R}(\theta, \phi')$ para todo o $\theta \in \Theta$ e desigualdade estrita para algum $\theta' \in \Theta$. Considere-se $\phi'' \in \Phi_I$ definida como em (5.26). Tem-se,

$$\begin{aligned} \hat{R}(\theta, \phi'') &= \frac{1}{M}\sum_{m=1}^{M} \hat{R}(\theta, \phi_{g_m}), \\ &= \frac{1}{M}\sum_{m=1}^{M} \hat{R}(\overline{g}_m\theta, \phi), \quad \text{[Teorema 5.5]} \\ &\leqq \frac{1}{M}\sum_{m=1}^{M} \hat{R}(\overline{g}_m\theta, \phi'); \end{aligned}$$

pelo Teorema 5.6, $\hat{R}(\overline{g}_m\theta, \phi') = \hat{R}(\theta, \phi')$, que implica, $\hat{R}(\theta, \phi'') \leqq \hat{R}(\theta, \phi')$. Logo, ϕ'', função de decisão invariante, é pelo menos tão boa quanto, ϕ', também invariante. Contudo,

$$\begin{aligned}\hat{R}(\theta', \phi'') &= \frac{1}{M}\sum_{m=1}^{M} \hat{R}(\overline{g}_m\theta', \phi) \\ &< \frac{1}{M}\sum_{m=1}^{M} \hat{R}(\overline{g}_m\theta', \phi'),\end{aligned}$$

visto que algum dos $\overline{g}_m$ é a transformação identidade em $\overline{G}$ e nesse caso, $\hat{R}(\theta', \phi) < \hat{R}(\theta', \phi')$. Finalmente, pelo Teorema 5.6,

$$\hat{R}(\overline{g}_m\theta', \phi') = \hat{R}(\theta', \phi'), \quad m = 1, 2, \ldots, M,$$

donde, $\hat{R}(\theta', \phi'') < \hat{R}(\theta', \phi')$, o que é absurdo porque, $\phi', \phi'' \in \Phi_I$ e ϕ' é, por hipótese, admissível em Φ_I. □□

Sobre a admissibilidade — na classe universal — das funções de decisão invariantes sem a restrição do teorema anterior podem avançar-se os seguintes comentários:

[a] Se a melhor função de decisão invariante é Bayes contra uma distribuição a priori própria (por exemplo quando a medida Haar invariante pode converter-se por normalização numa medida de probabilidade) a sua admissibilidade fica garantida.

[b] Se a melhor função de decisão invariante é Bayes generalizada — que é a situação usual — não se segue que seja admissível.

[c] As condições gerais de admissibilidade para funções de decisão invariantes são difíceis de obter.

Os exemplos que seguem ilustram aspectos diversos tratados na presente secção.

Exemplo 5.23 — Retome-se o Ex. 4.6. O problema de decisão considerado é invariante em relação ao grupo aditivo em $\mathbf{R}^k$, com elementos, $g_{\mathbf{c}}\mathbf{x} = \mathbf{x} + \mathbf{c}$, $\mathbf{c} = (c_1, c_2, \ldots, c_k)$, $-\infty < c_i < +\infty$, $i = 1, 2, \ldots, k$. É fácil de ver que a função de decisão Bayes (generalizada), $\delta_h(\mathbf{x}) = \mathbf{x}$, é a melhor função de decisão invariante [veja-se contudo a secção 6.6]; no entanto, é inadmissível quando se tem $k \geqq 3$. □

Exemplo 5.24 — No problema do «dado» com n faces — Blackwell e Girshick (1954) — uma das faces, não se sabe qual, tem probabilidade p e cada uma das outras tem probabilidade $(1 - p)/(n - 1) < p$. O problema consiste em decidir, com base em

N lançamentos do «dado», qual das faces é viciada. Suponha-se que as faces estão numeradas de 1 a n e que $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$, onde θ_j é o estado que corresponde à situação em que a face j é viciada. Seja, $A = \{a_1, a_2, \dots, a_n\}$, onde a_i indica a decisão que consiste em tomar a face i como viciada; a função perca natural é,

$$L(\theta_j, a_i) = \begin{cases} 1 & \text{se} \quad i \neq j, \\ 0 & \text{se} \quad i = j. \end{cases}$$

O resultado dos N lançamentos pode exprimir-se por $\mathbf{z} = (z_1, z_2, \dots, z_n)$ em que $\Sigma z_i = N$ e z_i designa o número de vezes que é obtida a face i. A variável aleatória, $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$, tem distribuição conhecida,

$$f(\mathbf{z} \,|\, \theta_j) = (N!/z_1!z_2! \dots z_n!) p^{z_j} [(1-p)/(n-1)]^{N-z_j}, \;\; z_i \geqq 0, \;\; \Sigma z_i = N.$$

Suponha-se que a distribuição a priori é,

$$\mathbf{h} = (h_1, h_2, \dots, h_n), \;\; h_j = 1/n, \;\; j = 1, 2, \dots, n,$$

em concordância com a simetria do problema. O risco a posteriori vem dado por,

$$\begin{aligned} r_{\mathbf{z}}(a_i) &= \sum_{j=1}^{n} L(\theta_j, a_i) h(\theta_j \,|\, \mathbf{z}) \\ &= \sum_{j=1}^{n} L(\theta_j, a_i) f(\mathbf{z} \,|\, \theta_j) h_j \Big/ \sum_{j=1}^{n} f(\mathbf{z} \,|\, \theta_j) h_j, \end{aligned}$$

donde,

$$r_{\mathbf{z}}(a_i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} f(\mathbf{z} \,|\, \theta_j) \Big/ \sum_{j=1}^{n} f(\mathbf{z} \,|\, \theta_j).$$

Consequentemente,

$$r_{\mathbf{z}}(a_i) \leqq r_{\mathbf{z}}(a_j) \quad \text{se} \quad [1 - f(\mathbf{z} \,|\, \theta_i)] \leqq [1 - f(\mathbf{z} \,|\, \theta_j)],$$

isto é, se,

$$f(\mathbf{z} \,|\, \theta_j) / f(\mathbf{z} \,|\, \theta_i) = [(n-1)p/(1-p)]^{z_j - z_i} \leqq 1.$$

Recorde-se ser $p > (1-p)/(n-1)$ e suponha-se que a frequência máxima é observada apenas para uma das faces. Nessa hipótese, a função de decisão aleatória Bayes contra $\mathbf{h}$ assume a forma,

$$\begin{aligned} \phi(a_i \,|\, \mathbf{z}) &= 1 \quad \text{se} \quad z_i = \max(z_1, z_2, \dots, z_n), \\ \phi(a_i \,|\, \mathbf{z}) &= 0 \quad \text{se} \quad z_i < \max(z_1, z_2, \dots, z_n), \end{aligned}$$

e corresponde a tomar a acção a_i — considerar a face i viciada — quando a face i é observada com frequência máxima. Na hipótese de aparecerem m faces com frequência máxima, $1 \leqq m \leqq n$, a função de decisão aleatória Bayes contra $\mathbf{h}$ assume a forma,

$$\begin{aligned} \phi(a_i \,|\, \mathbf{z}) &= 1/m \quad \text{se} \quad z_i = \max(z_1, z_2, \dots, z_n), \\ \phi(a_i \,|\, \mathbf{z}) &= 0 \qquad \text{se} \quad z_i < \max(z_1, z_2, \dots, z_n); \end{aligned} \tag{5.29}$$

quer dizer, se aparecerem m faces com frequência máxima, decide-se por sorteio qual delas deve considerar-se viciada. Nesse sorteio cada uma das m faces tem probabilidade $1/m$ de ser escolhida.

Seja $G = \{g_1, g_2, \dots, g_{n!}\}$ o grupo finito constituído pelas $n!$ permutações dos números, $(1, 2, \dots, n)$; para qualquer k, $1 \leqq k \leqq n!$,

$$g_k \mathbf{z} = g_k(z_1, z_2, \dots, z_n) = (z_{k_1}, z_{k_2}, \dots, z_{k_n}),$$

onde $(k_1, k_2, \dots, k_n)$ é a permutação associada com k; seja ainda,

$$\overline{g}_k \theta_j = \theta_{k_j}, \quad j = 1, 2, \dots, n; \quad \overline{g}_k a_i = a_{k_i}, \quad i = 1, 2, \dots, n.$$

Como é evidente, o problema de decisão é invariante em relação a G; também se verifica imediatamente que ϕ é invariante em relação a G.

Como Θ e D são finitos, existe, pelo Corolário 4.1, solução minimax. Logo, pelo Teorema 5.9, existe também uma solução invariante minimax. Mas, o conjunto $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ forma uma órbita, pois quaisquer que sejam, $\theta, \theta' \in \Theta$, existe sempre um $\overline{g}_k \in \overline{G}$ tal que $\theta' = \overline{g}_k \theta$: o grupo $\overline{G}$ é transitivo. Pelo Teorema 5.6, a função de decisão aleatória definida por (5.29), dita ϕ, possui risco constante sobre Θ, facto que aliado à qualidade de ser Bayes e ao Teorema 4.2 estabelece que ϕ é invariante minimax. Por outro lado, por ser Bayes própria ϕ é admissível. □

Exemplo 5.25 — Suponha-se que X tem densidade Cauchy,

$$f(x \,|\, \theta) = \{\pi[1 + (x - \theta)^2]\}^{-1}, \quad \Theta = (-\infty, +\infty),$$

e que pretende estimar-se θ com função perca quadrática. O problema é invariante em relação ao grupo aditivo com elementos $g_c x = x + c$, $c \in \mathbf{R}$ e as funções de decisão puras invariantes são da forma $\delta_c(x) = x + c$. Contudo,

$$R(\theta, \delta_c) = E_\theta\{[\theta - (X + c)]^2\} = \infty,$$

para todo o c, enquanto o estimador não invariante, $\delta_0(x) \equiv 0$, tem função risco,

$$R(\theta, \delta_0) = \theta^2 < R(\theta, \delta_c).$$

Neste caso, todos os estimadores invariantes são inadmissíveis [Berger (1980)]. □

Exemplo 5.26 — Tem-se $P(X = \theta + 1) = 1/2$ e $P(X = \theta - 1) = 1/2$, função perca,

$$L(\theta, a) = \begin{cases} |\theta - a| & \text{se } |\theta - a| \leqq 1, \\ 1 & \text{se } |\theta - a| > 1, \end{cases}$$

e pretende estimar-se θ: $A = \Theta = (-\infty, +\infty)$. As funções de decisão invariantes são mais uma vez da forma $\delta_c(x) = x + c$ quando se considera o grupo de transformações aditivo (θ é um parâmetro de localização). Como Θ é uma órbita, $R(\theta, \delta_c)$ é constante sobre Θ pelo que basta calcular $R(0, \delta_c)$,

$$R(0, \delta_c) = E_0\{L[0, \delta_c(X)]\} = \begin{cases} 1 - \frac{1}{2}|c| & \text{se } |c| \leqq 1, \\ \frac{1}{2}|c| & \text{se } 1 \leqq |c| \leqq 2, \\ 1 & \text{se } |c| \geqq 2. \end{cases}$$

A função risco é minimizada para $c = +1$ e $c = -1$; consequentemente, $\delta_1(x) = x+1$ e $\delta_{-1}(x) = x - 1$ são as melhores funções de decisão invariantes.

O estimador não invariante,

$$\delta^*(x) = \begin{cases} \delta_1(x) = x + 1 & \text{se} \quad x < 0, \\ \delta_{-1}(x) = x - 1 & \text{se} \quad x \geqq 0, \end{cases}$$

tem função risco,

$$R(\theta, \delta^*) = \begin{cases} 0 & \text{se } -1 \leqq \theta < 1, \\ \frac{1}{2} & \text{outros } \theta; \end{cases}$$

logo $\delta_1(x)$ e $\delta_{-1}(x)$ são inadmissíveis. Pode aliás demonstrar-se que, em regra, quando o melhor estimador invariante não é único, todos os estimadores invariantes são inadmissíveis [Berger (1980)]. □

5.7 Notas complementares

(i) A classe de funções de decisão invariantes pode ser demasiado restrita ou vazia (demasiada invariância) ou excessivamente ampla (pouca invariância). Uma ilustração do primeiro caso encontra-se no Ex. 5.22; uma possível saída consiste em não usar toda a invariância, isto é, trabalhar com o grupo mínimo de transformações para o qual o grupo induzido sobre Θ seja transitivo. Uma ilustração do segundo caso contém-se no Ex. 5.7 em que as considerações de invariância nos problemas que envolvem a Binomial não apresentam qualquer utilidade. O que se passa com a Poisson é ainda pior pois não há nenhum grupo de transformações

que deixe invariante as distribuições da família Poisson (repare-se que o Ex. 5.12 não envolve a estrutura dessas distribuições).

(ii) A necessidade de trabalhar com funções perca também invariantes tem como consequência que nos problemas tipo ensaio de hipóteses a invariância somente é aplicável quando a função perca é da forma «$0-K_i$» (veja-se o capítulo 7).

(iii) Na secção 5.5 foi referida a relação que existe entre a escolha da melhor função de decisão e a determinação da função de decisão Bayes generalizada (contra a medida Haar invariante direita). Pode até afirmar-se que os dois métodos são virtualmente equivalentes. Berger (1980), ao indicar preferência pelo segundo, aponta, entre outros, os seguintes motivos: (a) a pesquisa de procedimentos Bayes generalizados tem aplicações mais latas e (b) não se reveste das dificuldades causadas pelas situações em que há demasiada invariância. Em contrapartida o estudo da invariância tem o atractivo de apontar para distribuições a priori não informativas razoáveis (medida Haar invariante direita). Aliás, como nota Ferguson (1967), não é legítimo afirmar que um decisor actua «irracionalmente» quando não aplica funções de decisão invariantes nos casos em que a distribuição a priori não é invariante[11].

(iv) A análise da relação entre os conceitos de suficiência e de invariância não cabe nas presentes notas. A suficiência e a invariância são princípios virados para a redução dos dados, portanto, princípios com o mesmo objectivo genérico, que, em certo sentido, podem considerar-se complementares.

Levantam-se, no contexto, questões fundamentais como estas:

- É possível aplicar conjuntamente os dois princípios?
- Se a resposta é afirmativa, em que casos a aplicação da suficiência seguida da invariância é equivalente à aplicação da invariância seguida da suficiência?

A resposta à primeira questão é afirmativa e acontece que só raramente deixa de verificar-se a equivalência objecto da segunda questão. Sendo assim, a prática geralmente adoptada consiste em começar por reduzir os dados por meio de uma estatística suficiente mínima investigando depois a existência de grupos de transformações invariantes. A justificação para este procedimento baseia-se nas seguintes considerações:

- maior facilidade de cálculo;
- a suficiência é o principal elemento unificador da inferência estatística;
- a classe de funções de decisão baseadas numa estatística suficiente é essencialmente completa, propriedade que em geral se não verifica com a classe de funções de decisão invariantes.

[11] Alguns resultados importantes sobre invariância e medidas Haar invariantes são apresentados por Zacks (1971).

Notando que alguns problemas invariantes são tratados condicionando numa estatística invariante máxima (vejam-se os estimadores Pitman no capítulo seguinte), o Teorema 5.8, aplicado a uma tal estatística, $T(X)$, permite aprofundar o confronto entre suficiência e invariância. As estatísticas suficientes permitem simplificar o problema através da redução que operam no espaço da amostra, mas não interferem no espaço do parâmetro; as estatísticas, invariantes máximas, além de operarem também uma redução de $\mathcal{X}$, reduzem ainda o espaço do parâmetro na medida em que tornam possível ‹agregar› todos os pontos situados na mesma órbita de Θ.

(v) Na pesquisa da melhor função de decisão invariante há que ter em conta que pode acontecer que um dado problema de decisão seja invariante em relação a dois grupos de transformações. Em tais casos, a melhor função de decisão invariante pode não ser a mesma nos dois casos. Em Lehmann (1983) encontra-se um exemplo que embora interessante é demasiado extenso para ser aqui incluído.

CAPÍTULO 6

ESTIMAÇÃO

6.1 Introdução

O problema da estimação foi várias vezes aflorado em páginas anteriores. De forma selectiva vão apresentar-se mais alguns resultados.

Definida uma experiência consistindo na observação da variável ou vector aleatório, X, a estimação pode dizer-se que é um procedimento para em face do espaço de estados, Θ, «adivinhar» o verdadeiro estado com base na observação colhida, x. Nesse contexto tem-se, $A = \Theta$, e qualquer função de decisão, $\delta \in D$, faz corresponder a cada x o valor, $a = \delta(x)$, $a \in A$, proposto como estimativa ou valor aproximado do verdadeiro estado.

Interessa normalmente trabalhar com um conceito mais geral de parâmetro que inclua funções definidas em Θ, correntemente denotadas por $\tau(\theta)$.

Exemplo 6.1 — Se X_i, $i = 1, 2, \ldots, N$, X_i I.I.D. com distribuição de Poisson com média θ, tem-se, $\Theta = \{\theta : \theta > 0\}$. Desejando estimar-se,

$$P(X_i = 0) = e^{-\theta},\ i = 1, 2, \ldots, N,$$

ou,

$$P(\Sigma X_i = 0) = e^{-N\theta},$$

vem, no primeiro caso, $\tau(\theta) = e^{-\theta}$, no segundo caso, $\tau(\theta) = e^{-N\theta}$. □

Exemplo 6.2 — Com $\Theta = \{\theta : -\infty < \theta < +\infty\}$, pode ter-se,

$$\tau(\theta) = \theta^2 + 1 \quad \text{ou} \quad \tau(\theta) = \begin{cases} 1 & \text{se} \quad \theta \geqq 0, \\ 0 & \text{se} \quad \theta < 0. \end{cases}$$ □

6.2 Teorema de Rao-Blackwell

Um princípio que restringe a classe de funções de decisão — afinal o que se passa com o princípio da invariância — consiste em dar atenção apenas aos estimadores que sejam centrados ou não enviesados, isto é, que verifiquem

$$E_\theta\{\delta(X)\} = \tau(\theta) \quad \text{para todo o } \theta \in \Theta. \tag{6.1}$$

A selecção operada com o princípio de não enviesamento não está liberta de críticas[1] pois tem-se verificado que elimina estimadores triviais mas pode também eliminar estimadores dignos de consideração; no entanto, apesar das suas limitações, é frequentemente invocado na estimação[2].

Suponha-se que $\delta \in D$ é uma função de decisão ou estimador não enviesado de $\tau(\theta)$ e que se verifica, $E_\theta\{|\,\delta(X)\,|^2\} < \infty$, para todo o $\theta \in \Theta$; tem-se,

$$E_\theta\{[\delta(X) - \tau(\theta)]^2\} = V_\theta\{\delta(X)\}, \tag{6.2}$$

isto é, o erro quadrático médio é igual à variância de $\delta(X)$ calculada quando a distribuição de X é P_θ. A prática de minimizar o erro quadrático médio [= à função risco com perca quadrática] leva a investigar a existência de estimadores não enviesados com variância uniformemente mínima ou estimadores UMVU [convém manter as iniciais da expressão «*uniformly minimum variance unbiased*»]. Não se esqueça, porém, que um estimador UMVU pode ser muito inferior a um estimador ligeiramente enviesado mas de menor variância[3].

Daqui por diante faz-se $\tau(\theta) = \theta$ para facilitar a exposição, uma vez que os resultados obtidos nesse caso particular adaptam-se quando necessário ao caso geral. Salvo menção em contrário, θ é um parâmetro real, com $\Theta = \mathbf{R}^k$ ou = a subconjunto convexo de $\mathbf{R}^k$, $k \geqq 1$.

---

[1] Tal como o de erro quadrático médio é um conceito frequencista.

[2] A ideia de «imparcialidade» [veja-se pág. 145] de um estimador não enviesado traduz-se pelo equilíbrio entre os casos de subestimação e sobrestimação; assim, quando usado repetidamente um estimador não enviesado tende a reproduzir «em média» o verdadeiro valor de $\tau(\theta)$. Note-se que há casos em que não existem estimadores não enviesados. Lehmann (1983) cita o seguinte exemplo: com $X \sim B(N;\theta)$, suponha-se $\tau(\theta) = 1/\theta$; se $\delta(X)$ é estimador não enviesado, deve ter-se,

$$\sum_{x=0}^{N} \delta(x) \binom{N}{x} \theta^x (1-\theta)^{N-x} = \tau(\theta) \quad \text{para todo o } 0 < \theta < 1;$$

fazendo $\theta \to 0$, o primeiro membro tende para $\delta(0)$ e o segundo membro para ∞! Contudo (desde que N não seja muito pequeno), existem estimadores que assumem valores próximo de $1/\theta$ com elevada probabilidade, por exemplo, N/X (com algum ajustamento se for $X = 0$).

[3] Por outras palavras, um estimador não enviesado pode ser estritamente dominado por um estimador enviesado, isto é, pode ser inadmissível. Um caso notável é abordado na secção 6.6.

O método mais poderoso para encontrar estimadores UMVU é o conhecido Teorema de Rao-Blackwell. Antes de fazer a demonstração recorde-se que uma função de decisão diz-se baseada numa estatística suficiente para θ, seja $T(X)$, quando,

$$\delta(x) = \delta[T(x)] \quad \text{para todo o } x \in \mathfrak{X},$$

devendo a expressão «para todo o $x \in \mathfrak{X}$» interpretar-se em termos mais gerais como «excepto, quando muito, para um conjunto de valores de x com probabilidade zero». Mais rigorosamente, δ diz-se baseada em $T(X)$ se,

$$P_\theta\{\delta(X) = \psi[T(X)]\} = 1 \quad \text{para todo o } \theta \in \Theta. \tag{6.3}$$

Teorema 6.1 — Com T estatística suficiente para θ, seja δ' uma função de decisão não baseada em T; então se δ' é estimador não enviesado de θ, a função de decisão baseada em T definida pela relação,

$$\delta(x) = \delta[T(x)] = \delta(t) = E\{\delta'(X) \,|\, T = t\}, \tag{6.4}$$

é também estimador não enviesado de θ. Se $L(\theta, a)$ é, para θ fixo, função estritamente convexa de a,

$$R(\theta, \delta) < R(\theta, \delta') \quad \text{para todo o } \theta \in \Theta. \tag{6.5}$$

Em particular, com $L(\theta, a) = (\theta - a)^2$,

$$V_\theta\{\delta(X)\} < V_\theta\{\delta'(X)\} \quad \text{para todo o } \delta \in \Theta. \tag{6.6}$$

Dem. Como, por hipótese, T é suficiente para θ, a distribuição condicionada por $T = t$ de qualquer outra estatística e em particular de δ' não depende de θ [o que explica o não aparecimento de θ no índice inferior do valor esperado que figura na expressão (6.4)]. Se o valor esperado em (6.4) existe, o estimador δ, que define, conserva a propriedade de não enviesamento possuída por δ', pois,

$$E_\theta\{\delta(X)\} = E_\theta\{E[\delta'(X) \,|\, T]\} = E_\theta\{\delta'(X)\} = \theta.$$

Por outra parte, em consequência da convexidade estrita da função perca, tem-se de (2.24),

$$L[\theta, E\{\delta'(X) \,|\, t\}] < E\{L[\theta, \delta'(X)] \,|\, t\},$$

donde,

$$\begin{aligned} R(\theta, \delta) &= E_\theta\{L[\theta, \delta(X)]\} = E_\theta\{L[\theta, E\{\delta'(X) \,|\, T\}]\} \\ &< E_\theta\{E\{L[\theta, \delta'(X)] \,|\, T\}\} = E_\theta\{L[\theta, \delta'(X)]\} = R(\theta, \delta'), \end{aligned}$$

e (6.5) fica provado. A verificação de (6.6) é imediata. □□

A aplicação de (6.4) para obter δ a partir de δ' designa-se por Rao-Blackwellização de δ'. Como acaba de mostrar-se a operação mantém o não enviesamento e pode produzir um estimador com menor risco ou variância. De facto, a igualdade em (6.5) e (6.6) só se verificaria se,

$$P_\theta[\delta(X) = \delta'(X)] = 1 \quad \text{para todo o } \theta \in \Theta,$$

o que implicaria que tanto δ como δ' fossem baseados em T, contrariamente à hipótese de partida.

A Rao-Blackwellização é um importante passo na construção de estimadores UMVU. Antes de apresentar alguns exemplos importa fazer dois comentários.

(i) O Teorema de Rao-Blackwell pode demonstrar-se sem introduzir a hipótese de que δ' é estimador não enviesado, passando a ser (6.5) a única tese.

(ii) Salvo ligeiros pormenores referentes às hipóteses, é curioso assinalar que o Teorema de Rao-Blackwell pode obter-se directamente a partir dos Teoremas 4.4 e 4.6.

Exemplo 6.3 — Seja $\mathbf{X} = (X_1, X_2, \dots, X_N)$, X_i I.I.D. com função de distribuição (absolutamente) contínua,

$$\Theta = \{F : F \text{ é uma função de distribuição (absolutamente) contínua}\}.$$

Se Y_i, $i = 1, 2, \dots, N$, são as estatísticas de ordem [no exemplo 5.13 designaram-se por $X_{(i)}$], o estimador $\mathbf{T} = (Y_1, Y_2, \dots, Y_N)$ é suficiente para Θ. Com efeito, dado $\mathbf{T} = \mathbf{t}$, com $\mathbf{t} = (y_1, y_2, \dots, y_N)$, o vector $\mathbf{X} = (X_1, X_2, \dots, X_N)$ só pode assumir uma das $N!$ permutações de $(y_1, y_2, \dots, y_N)$ e cada uma dessas permutações tem probabilidade $1/N!$, independente da particular $F \in \Theta$ (de facto $\mathbf{T}$ é suficiente mínima pois, como é sabido, $\mathbf{T}$ é suficiente mínima para a distribuição de Cauchy que é um dos elementos da família Θ por ser absolutamente contínua).

Seja (restringindo Θ para garantir a existência de momentos) a estimação de,

$$\tau(F) = E_F\{X_i\} = \int x dF(x), \ i = 1, 2, \dots, N;$$

como é óbvio, $\delta'(\mathbf{X}) = X_1$, é estimador não enviesado. No entanto, como δ' ignora a informação dada pela observação de $X_2, \dots, X_N$, não parece difícil encontrar um estimador não enviesado melhor.

Aplicando o Teorema de Rao-Blackwell, procure determinar-se, δ, tal que,

$$\delta(x_1, x_2, \dots, x_N) = \delta(y_1, y_2, \dots, y_N) = \delta(\mathbf{t}),$$

com,

$$\delta(\mathbf{t}) = E\{X_1 \mid \mathbf{T} = \mathbf{t}\}.$$

Fixado, $\mathbf{T} = \mathbf{t}$, X_1 só pode assumir os valores, y_i, $i = 1, 2, \ldots, N$, e,

$$P(X_1 = y_1 \,|\, \mathbf{T} = \mathbf{t}) = (N-1)!/N! = 1/N,\ i = 1, 2, \ldots, N.$$

Agora,

$$\begin{aligned}\delta(\mathbf{t}) = E\{X_1 \,|\, \mathbf{T} = \mathbf{t}\} &= (y_1/N) + (y_2/N) + \cdots + (y_N/N) \\ &= \Sigma y_i/N = \Sigma x_i/N = \overline{x},\end{aligned}$$

donde se conclui que $\delta(\mathbf{X}) = \Sigma X_i/N = \overline{X}$ é estimador não enviesado de $\tau(F)$ com variância uniformemente inferior à de $\delta'(\mathbf{X})$. De facto,

$$V_F\{\delta(\mathbf{X})\} = (1/N)\ V_F\{\delta'(\mathbf{X})\}.$$

Se $\tau(F) = E_F\{X_i^2\}$, $i = 1, 2, \ldots, N$, o mesmo teorema esclarece, de modo semelhante, que $\Sigma X_i^2/N$ é estimador não enviesado com menor variância do que X_1^2 [Fraser (1957)]. □

Exemplo 6.4 — Com X_i I.I.D., $X_i \sim N(\theta, 1)$, $i = 1, 2, \ldots, N$, um estimador razoável para θ é a mediana da amostra,

$$\delta'(\mathbf{X}) = \begin{cases} Y_{k+1} & \text{se } N = 2k+1 \\ (Y_k + Y_{k+1})/2 & \text{se } N = 2k \end{cases}$$

onde Y_i, $i = 1, 2, \ldots, N$, são as estatísticas de ordem. Suponha-se N ímpar para não alongar o estudo. A distribuição de $Y_{k+1} - \theta$ é dada por,

$$dG(u) = [N!/(k!)^2][\Phi(u)]^k[1 - \Phi(u)]^k d\Phi(u),$$

onde $G(u) = P(Y_{k+1} - \theta \leqq u)$ e, como é usual,

$$\Phi(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} \exp\{-\xi^2/2\}\, d\xi.$$

Como $\int_{-\infty}^{+\infty} u dG(u) = 0$, vem, $E_\theta\{Y_{k+1} - \theta\} = 0$, donde,

$$E_\theta\{\delta'(\mathbf{X})\} = E_\theta\{Y_{k+1}\} = \theta,$$

o que mostra ser δ' estimador não enviesado de θ. Por outro lado é bem sabido que $T = \Sigma X_i/N$ é estatística suficiente para θ; procedendo à Rao-Blackwellização de δ', a condição $T = t$, isto é, $\Sigma x_i = Nt$, implica,

$$E\{\Sigma X_i \,|\, t\} = Nt \rightarrow E\{X_i \,|\, t\} = t,$$

e ainda,

$$E\{Y_{k+1} \mid t\} = E\{X_i \mid t\} = t,$$

como é possível demonstrar. Portanto, com $\delta(t) = E\{\delta'(\mathbf{X}) \mid t\} = t$, o estimador $\delta(\mathbf{X}) = T = \Sigma X_i/N$ é não enviesado e tem variância uniformemente inferior à de δ' (risco uniformemente inferior para função perca estritamente convexa). De facto,

$$V_\theta\{\delta(\mathbf{X})\} = 1/N; \quad V_\theta\{\delta'(\mathbf{X})\} \cong \pi/2N,$$

onde a segunda expressão vale como aproximação para valores grandes de N por ser resultado assintótico [Fraser (1957)]. □

Os teoremas seguintes são simples corolários do Teorema de Rao-Blackwell,

Teorema 6.2 — Se a função perca é convexa e a função de decisão $\delta \in D$ é definida como em (6.4), então,

$$R(\theta, \delta) \leqq R(\theta, \delta') \quad \text{para todo o } \theta \in \Theta,$$

qualquer que seja $\delta' \in D$. □□

Teorema 6.3 — Se a função perca é estritamente convexa a classe dos estimadores baseados numa estatística suficiente é completa[4]. □□

O Teorema de Rao-Blackwell pode alargar-se de forma a considerar o caso em que $\boldsymbol{\theta}$, $\boldsymbol{\tau}(\boldsymbol{\theta})$, $\boldsymbol{\delta}'$, $\boldsymbol{\delta}$ e $\mathbf{T}$ são vectores de $\mathbf{R}^k$, passando a designar-se por Teorema de Blackwell-Lehmann-Scheffé. Grande parte da demonstração do segundo duplica formalmente a do primeiro e não se trata aqui [veja-se Fraser (1957)]. A principal alteração diz respeito a (6.6) que não tem sentido em termos vectoriais.

Seja $\boldsymbol{\delta}'(X) = [\delta'_1(X), \delta'_2(X), \ldots, \delta'_k(X)]$ um estimador não enviesado de $\boldsymbol{\tau}(\boldsymbol{\theta}) = [\tau_1(\boldsymbol{\theta}), \tau_2(\boldsymbol{\theta}), \ldots, \tau_k(\boldsymbol{\theta})]$ e seja $[\sigma'_{ij}(\boldsymbol{\theta})]$ a respectiva matriz de covariâncias, com elementos,

$$\sigma'_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{[\delta'_i(X) - \tau_i(\boldsymbol{\theta})][\delta'_j(X) - \tau_j(\boldsymbol{\theta})]\},$$

onde,

$$E_{\boldsymbol{\theta}}\{\delta'_i(X)\} = \tau_i(\boldsymbol{\theta}).$$

Supondo que a matriz, $[\sigma'_{ij}(\boldsymbol{\theta})]$, é não singular, represente-se a sua inversa por $[\sigma'^{ij}(\boldsymbol{\theta})]$; sejam, $[\sigma_{ij}(\boldsymbol{\theta})]$ e $[\sigma^{ij}(\boldsymbol{\theta})]$ as matrizes correspondentes para o estimador $\boldsymbol{\delta}(X)$. O resultado que generaliza (6.6) é, então,

$$\sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij}(\boldsymbol{\theta})\xi_i\xi_j \leqq \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma'_{ij}(\boldsymbol{\theta})\xi_i\xi_j, \tag{6.7}$$

[4] Em correspondência com o Teorema 4.7. Releiam-se, na pág. 272, os comentários ao Teorema 6.1.

que pode escrever-se

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma^{ij}(\boldsymbol{\theta})\xi_i\xi_j \geqq \sum_{i=1}^{k}\sum_{j=1}^{k}\sigma'^{ij}(\boldsymbol{\theta})\xi_i\xi_j, \tag{6.8}$$

para todo $\boldsymbol{\theta} \in \Theta$, sendo ξ_i, $i = 1, 2, \ldots, k$, números reais arbitrários não todos nulos; a igualdade em (6.7) ou (6.8) verifica-se quando e só quando,

$$P_{\boldsymbol{\theta}}[\boldsymbol{\delta}(X) = \boldsymbol{\delta}'(X)] \quad \text{para todo o } \boldsymbol{\theta} \in \Theta.$$

Tem-se, claro,

$$\boldsymbol{\delta}(X) = E\{\boldsymbol{\delta}'(X) \,|\, \mathbf{T} = \mathbf{t}\} = [E\{\delta_1'(X) \,|\, \mathbf{t}\}, E\{\delta_2'(X) \,|\, \mathbf{t}\}, \ldots, E\{\delta_k'(X) \,|\, \mathbf{t}\}].$$

O elipsoide de concentração para um estimador vectorial, seja $\boldsymbol{\delta}(X)$, foi definido por Cramér e representa-se pela relação,

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma^{ij}(\boldsymbol{\theta})\xi_i\xi_j = k + 2, \tag{6.9}$$

onde se escreveu ξ_i em vez de $\delta_i(X) - \tau_i(\boldsymbol{\theta})$. A interpretação é a seguinte: considerando a distribuição Normal k-dimensional com vector de médias, $\boldsymbol{\tau}(\boldsymbol{\theta})$, e matriz covariâncias, $[\sigma_{ij}(\boldsymbol{\theta})]$, idênticos aos do estimador, o primeiro membro de (6.9) é o expoente da densidade correspondente aquela distribuição multiplicado por $-1/2$. Por outro lado, considerando o elipsoide tal que a distribuição Uniforme sobre o mesmo tenha ainda médias e covariâncias idênticas às da referida distribuição Normal k-dimensional obtém-se a justificação para o segundo membro. Quando duas variáveis aleatórias k-dimensionais possuem o mesmo vector de médias e o elipsoide de concentração da primeira está contido no elipsoide de concentração da segunda, diz-se que a primeira tem maior concentração do que a segunda, concentração entendida em relação ao centro de gravidade comum.

A relação (6.8), demonstrada no Teorema de Blackwell-Lehmann-Scheffé, mostra que dos estimadores, $\boldsymbol{\delta}(X)$ e $\boldsymbol{\delta}'(X)$, que possuem o mesmo vector de médias, é o primeiro o mais concentrado, porquanto a igualdade,

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma^{ij}(\boldsymbol{\theta})(\omega\xi_i)(\omega\xi_j) = \sum_{i=1}^{k}\sum_{j=1}^{k}\sigma'^{ij}(\boldsymbol{\theta})\xi_i\xi_j [= k + 2]$$

somente se verifica com $\omega \leqq 1$.

O Teorema de Rao-Blackwell — ou a sua extensão — não resolve inteiramente o problema dos estimadores UMVU. Deixa, no entanto, bem claro que se existir

estimador UMVU esse estimador tem de procurar-se na classe dos estimadores não enviesados que sejam baseados numa estatística suficiente. Se esta classe tem um único elemento ele é, obviamente, o estimador UMVU; é isso que de facto se verifica quando o estimador é baseado numa estatística suficiente completa.

A importância da completicidade na determinação do estimador UMVU é de fácil compreensão. Se a estatística T é suficiente completa para θ e se $\delta'(X)$ é um estimador não enviesado qualquer, seja $\delta(X)$ o estimador obtido por Rao-Blackwellização de δ', $\delta(T) = E\{\delta'(X)\,|\,T\}$. Se houver outro estimador não enviesado baseado em T, seja $\delta^*(T)$, tem-se,

$$E_\theta\{\delta(T)\} = E_\theta\{\delta^*(T)\} = \theta,$$

donde,

$$E_\theta\{\delta(T) - \delta^*(T)\} = 0 \quad \text{para todo o } \theta \in \Theta,$$

e pela completicidade de T (qualquer estimador não enviesado de zero[5] é igual a zero com probabilidade igual a um),

$$P_\theta[\delta(T) = \delta^*(T)] = 1 \quad \text{para todo o } \theta \in \Theta,$$

isto é, $\delta(t) = \delta^*(t)$, excepto quando muito sobre um conjunto com probabilidade zero, situação em que não se faz em geral distinção entre os estimadores.

Se não se conhece uma estatística suficiente completa e se opera com uma estatística suficiente, T, podem existir vários estimadores não enviesados baseados em T mas não há procedimento geral para comparar as respectivas variâncias e chegar ao estimador UMVU.

Note-se, de acordo com o Teorema de Rao-Blackwell, que se δ é o (único) estimador UMVU, então δ é também a única função de decisão para a qual se tem,

$$R(\theta, \delta) < R(\theta, \delta') \quad \text{para todo o } \theta \in \Theta,$$

qualquer que seja δ', desde que a função perca seja estritamente convexa.

Das considerações antecendentes fica estabelecido o Teorema de Lehmann-Scheffé,

Teorema 6.4 — Se existe uma estatística suficiente completa para θ, seja T, e pelo menos um estimador não enviesado de θ, então existe um único estimador UMVU. Este estimador é o único não enviesado que é função de T e o único estimador não enviesado com risco mínimo quando a função perca é estritamente convexa. □□

[5] Baseado em T. Sobre completicidade veja-se, além das obras indicadas, Silvey (1970).

Este teorema pode ser generalizado de modo a considerar parâmetros e estimadores k-dimensionais sendo a respectiva demonstração muito semelhante à que foi apresentada para o caso unidimensional,

Teorema 6.5 — Se existe uma estatística suficiente completa para $\boldsymbol{\theta}$, seja $\mathbf{T}$, e pelo menos um estimador não enviesado de $\boldsymbol{\theta}$, então existe um único estimador com elipsoide de concentração mínimo e risco mínimo (função perca estritamente convexa). Este estimador é o único não enviesado que é função de $\mathbf{T}$. □□

Exemplo 6.5 — Com $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. com distribuição Exponencial, pretendem determinar-se os estimadores UMVU para θ e para,

$$\tau(\theta) = P_\theta(X_i > m) = e^{-m\theta}.$$

Com $f(\mathbf{x} \mid \theta) = \theta^N \exp\{-\theta \Sigma x_i\}$, tem-se que $T = \Sigma X_i$ é estatística suficiente completa para θ. Por outra parte, $\overline{X}$, é estimador não enviesado de $1/\theta$; portanto, parece razoável conjecturar que $\delta(\mathbf{X}) = c/\Sigma X_i$, onde c é uma constante a determinar, é estimador não enviesado de θ. Com efeito,

$$E_\theta\{\delta(\mathbf{X})\} = cE_\theta\{1/T\} = c\int_0^\infty (1/t)[1/\Gamma(N)]\theta^N e^{-\theta t} t^{N-1} dt,$$

pois, como é sabido, T tem distribuição Gama. Assim,

$$E_\theta\{\delta(\mathbf{X})\} = \frac{c\theta\Gamma(N-1)}{\Gamma(N)} = \frac{c\theta}{N-1},$$

donde se conclui que $\delta(\mathbf{X}) = (N-1)/\Sigma X_i$ é estimador não enviesado de θ; como já é função da estatística suficiente completa T, δ é o estimador UMVU de θ.

Por outro lado, a função indicatriz, $I_{(m,\infty)}(X_1)$ é estimador não enviesado de $\tau(\theta) = e^{-m\theta}$, pois,

$$E_\theta\{I_{(m,\infty)}(X_1)\} = P_\theta(X_1 > m) = e^{-m\theta};$$

a Rao-Blackwellização de $I_{(m,\infty)}(X_1)$ conduz a,

$$\begin{aligned}
\delta(t) &= E\{I_{(m,\infty)}(X_1) \mid \Sigma X_i = t\} \\
&= P\{I_{(m,\infty)}(X_1) = 1 \mid \Sigma X_i = t\} \\
&= P\{X_1 > m \mid \Sigma X_i = t\} \\
&= \int_m^t f(x_1 \mid t)\, dx_1.
\end{aligned}$$

Ora,

$$\begin{aligned} f(x_1 \mid t) &= f(x_1 \mid \theta) f(t \mid x_1, \theta) / f(t \mid \theta) \\ &= \frac{\theta e^{-\theta x_1} [1/\Gamma(N-1)] \theta^{N-1} e^{-\theta(t-x_1)} (t-x_1)^{N-2}}{[1/\Gamma(N)] \theta^N e^{-\theta t} t^{N-1}} \\ &= \frac{(N-1)(t-x_1)^{N-2}}{t^{N-1}}, \quad (x_1 \leqq t \text{ e } N > 1). \end{aligned}$$

Assim,

$$\delta(t) = \frac{N-1}{t^{N-1}} \int_m^t (t-x_1)^{N-2}\, dx_1 = \left(\frac{t-m}{t}\right)^{N-1}, \quad t \geqq m,$$

donde, finalmente,

$$\delta(\mathbf{X}) = \left(\frac{\Sigma X_i - m}{\Sigma X_i}\right)^{N-1} I_{(m,\infty)}(\Sigma X_i),$$

é o estimador UMVU de $\tau(\theta)$ [Mood, Graybill e Boes (1974)]. □

Exemplo 6.6 — Suponha-se ser $\mathbf{X}$ uma amostra casual de dimensão N de uma população $N(\theta, 1)$, caso em que $T = \Sigma X_i$ é estatística suficiente completa para θ. Consequentemente, $\delta(\mathbf{X}) = \overline{X}$, por ser não enviesado e função de T, é o estimador UMVU de θ. Por ser, $E_\theta\{\overline{X}^2\} = (1/N) + \theta^2$,

$$\delta^*(\mathbf{X}) = \overline{X}^2 - (1/N),$$

é o estimador UMVU de $\tau(\theta) = \theta^2$. Curiosamente, δ^*, além de poder conduzir a estimativas negativas para θ^2, não é admissível quando a função perca é quadrática. Com efeito, o estimador,

$$\delta''(\mathbf{X}) = \max\left(0, \overline{X}^2 - \frac{1}{N}\right),$$

embora enviesado, verifica,

$$E_\theta\{[\theta - \delta''(\mathbf{X})]^2\} < E_\theta\{[\theta - \delta^*(\mathbf{X})]^2\},$$

para todo o θ, e é portanto estritamente melhor do que δ^* [Ferguson (1967)]. □

Exemplo 6.7 — Com X_i I.I.D., $X_i \sim N(0, \theta^2)$, $i = 1, 2, \ldots, N$, sabe-se que $T = \Sigma X_i^2$ é estatística suficiente completa para θ^2; $\delta(\mathbf{X}) = \Sigma X_i^2/N$, sendo não enviesado e função de T, é o estimador UMVU para θ^2. Contudo, δ não é admissível quando a função perca é quadrática.

Considere-se a classe de estimadores de θ^2,

$$\delta_c(\mathbf{X}) = c\Sigma X_i^2, \quad c > 0;$$

sabe-se que T/θ^2 tem distribuição χ^2 com N graus de liberdade, com,

$$E_{\theta^2}\{T\} = N\theta^2, \quad E_{\theta^2}\{T^2\} = N(N+2)\theta^4.$$

A função risco é dada por,

$$\begin{aligned} R(\theta^2, \delta_c) &= E_{\theta^2}\{[\theta^2 - \delta_c(\mathbf{X})]^2\} = E_{\theta^2}\{(\theta^2 - cT)^2\} \\ &= [c^2 N(N+2) - 2cN + 1]\theta^4. \end{aligned}$$

Esta função é minimizada quando $c = 1/(N+2)$; assim, o estimador,

$$\delta^*(\mathbf{X}) = \frac{\Sigma X_i^2}{N+2},$$

tem risco uniformemente inferior ao risco do estimador UMVU, δ, e é o melhor dentro da classe de estimadores, $\delta_c > 0$ [Ferguson (1967)]. □

Exemplo 6.8 — Suponha-se que o número de chamadas que afluem a um PBX num período de 10 minutos é uma variável aleatória, X, com distribuição de Poisson e média θ. Pretende estimar-se a probabilidade de no próximo período de 20 minutos não haver chamadas: $\tau(\theta) = e^{-2\theta}$.

Tem-se que X é estatística suficiente completa para θ; o estimador de $\tau(\theta)$, $\delta(X)$, é não enviesado se,

$$E_\theta\{\delta(X)\} = \sum_{x=0}^{\infty} \delta(x) e^{-\theta} \theta^x / x! = e^{-2\theta},$$

isto é, se,

$$\sum_{x=0}^{\infty} \delta(x)\theta^x / x! = e^{-\theta} = \sum_{x=0}^{\infty} (-1)^x \theta^x / x!,$$

para todo o $\theta > 0$. Para se verificar esta igualdade têm de ser iguais os coeficientes homólogos das sucessivas potências de θ. Estranhamente, tem-se como estimador UMVU de $\tau(\theta)$ — que é uma probabilidade,

$$\delta(x) = (-1)^x, \quad x = 0, 1, 2, \ldots;$$

se no período inicial de 10 minutos o número de chamadas registado é par, $\delta(x) = 1$; se é ímpar, $\delta(x) = -1$! [Ferguson (1967)]. □

No que diz respeito aos métodos de solução convém apontar o seguinte: (i) os Exs. 6.5 (estimação de θ), 6.6 e 6.7, ilustram como se obtém estimadores UMVU procurando conjecturar a forma dos estimadores que sendo função de uma estatística

suficiente completa são adicionalmente não enviesados; (ii) o Ex. 6.5 (estimação de $\tau(\theta) = e^{-m\theta}$) destaca o método em que se parte de um estimador não enviesado e se procede à respectiva Rao-Blackwellização condicionando numa estatística suficiente completa; (iii) no Ex. 6.8 volta a partir-se de uma estatística suficiente completa, seja $T(= X$ no exemplo), e tenta obter-se a função de decisão UMVU, seja $\delta(t)$, resolvendo o conjunto de equações,

$$E_\theta\{\delta(T)\} = \tau(\theta) \quad \text{para todo o } \theta \in \Theta,$$

isto é, no caso contínuo,

$$\int \delta(t) f(t \mid \theta)\, dt = \tau(\theta), \quad \theta \in \Theta,$$

no caso discreto,

$$\Sigma_t \delta(t) f(t \mid \theta) = \tau(\theta), \quad \theta \in \Theta.$$

[Outros exemplos podem estudar-se em Rohatgi (1976) e Lehmann (1983)].

Como os Exs.6.6, 6.7 e 6.8 mostram, os estimadores UMVU podem não ser bons estimadores. Se existe uma estatística suficiente completa e se existe um estimador não enviesado, o estimador UMVU é o melhor dentro de uma classe com um único elemento (a menos de diferenças em conjuntos com probabilidade zero) e as suas propriedades podem não ser interessantes.

O principal domínio de aplicação da estimação UMVU é na amostragem da família exponencial em consequência da conhecida proposição: se $\mathbf{X}$ é amostra casual de dimensão N de uma distribuição exponencial dotada de parametrização natural,

$$f(\mathbf{x} \mid \theta) = [C(\theta)]^N \exp\left\{\sum_{j=1}^{k} \theta_j \sum_{i=1}^{N} R_j(x_i)\right\} \Pi H(x_i),$$

e se o espaço dos parâmetros, Θ, contém um rectângulo k-dimensional [mais precisamente quando a família exponencial possui «*full rank*»], então,

$$\mathbf{T} = (T_1, T_2, \ldots, T_k), \; T_j = \sum_{i=1}^{N} R_j(x_i), \; j = 1, 2, \ldots, k,$$

é estatística suficiente completa.

Este teorema permite, por exemplo, encontrar facilmente estatísticas suficientes completas na amostragem das distribuições Binomial, Poisson e Normal[6]. Por outro lado, demonstra-se [veja-se Lehmann (1983)] que na amostragem da família

[6] Ou ainda, da Qui-quadrado, Gama e Beta, entre outras.

exponencial, verificadas certas condições de regularidade, os estimadores UMVU e os estimadores da máxima verosimilhança são assintoticamente equivalentes quando a dimensão da amostra tende para infinito. Assim, os estimadores UMVU partilham as propriedades assintóticas reconhecidas para os estimadores da máxima verosimilhança. Para pequenas amostras ambos os estimadores podem ser insatisfatórios, como aliás se ilustrou para os UMVU.

Os estimadores UMVU podem não ser admissíveis [Exs. 6.6 e 6.7]; no entanto, se a inadmissibilidade não for intensa pode aceitar-se que é compensada pelo não enviesamento e manter a preferência pelo UMVU considerado. Repare-se que há aqui um contraste em relação ao que se passa com os estimadores minimax. Se, $\tilde{\delta}$, minimax é inadmissível, é dominado estritamente por algum, $\tilde{\delta}_0$, que é também minimax necessariamente e, portanto, sempre preferível ao primeiro[7].

6.3 Melhores estimadores invariantes [I]

O critério de invariância é especialmente fecundo em problemas relacionados com a estimação de parâmetros de localização e de escala.

Considere-se uma variável aleatória real, X, com distribuição na família $\mathcal{F}$; seja, $F(x\,|\,\theta) = P_\theta(X \leqq x)$, a respectiva função de distribuição e considerem-se as três situações seguintes:

$$F(x\,|\,\theta) = F(x-\theta), \qquad \theta \in \Theta \subset \mathbf{R}; \tag{6.10}$$

$$F(x\,|\,\theta) = F(x/\theta), \qquad \theta \in \Theta \subset \mathbf{R}, \quad \theta > 0; \tag{6.11}$$

$$F(x\,|\,\mu,\sigma) = F[(x-\mu)/\sigma], \qquad \boldsymbol{\theta} = (\mu,\sigma) \in \Theta \subset \mathbf{R}^2, \quad \sigma > 0, \tag{6.12}$$

onde F é função de distribuição.

No caso (6.10), θ é parâmetro de localização; no caso (6.11), θ é parâmetro de escala; no caso (6.12), $\boldsymbol{\theta}$ é parâmetro de localização-escala[8]. Definições alternativas são, respectivamente, a distribuição de $X-\theta$ é independente de θ, a distribuição de X/θ é independente de θ e a distribuição de $(X-\mu)/\sigma$ é independente de $\boldsymbol{\theta}$.

Quando a variável aleatória, X, é contínua, as relações equivalentes a (6.10), (6.11) e (6.12), em termos da função de densidade, são, respectivamente,

$$f(x\,|\,\theta) = f(x-\theta), \tag{6.13}$$

$$f(x\,|\,\theta) = (1/\theta) f(x/\theta), \tag{6.14}$$

$$f(x\,|\,\mu,\sigma) = (1/\sigma) f[(x-\mu)/\sigma], \tag{6.15}$$

onde f é função de densidade.

[7] Na secção 6.4 fazem-se considerações adicionais sobre os estimadores UMVU.

[8] O caso dos parâmetros localização-escala não é aqui considerado [veja-se Lehmann (1983)].

Exemplo 6.9 — Se X tem distribuição Uniforme no intervalo $(\theta, \theta+1)$, θ é parâmetro de localização. Com efeito,

$$F(x \mid \theta) = (x-\theta)I_{(0,1)}(x-\theta) + I_{[1,\infty)}(x-\theta),$$
$$f(x \mid \theta) = I_{(0,1)}(x-\theta).$$

□

Exemplo 6.10 — Se X tem distribuição Exponencial,

$$F(x \mid \theta) = 1 - \exp\{-x/\theta\},\ \theta > 0, x > 0,$$
$$f(x \mid \theta) = (1/\theta)\exp\{-x/\theta\},$$

θ é parâmetro de escala. □

Exemplo 6.11 — Se $X \sim N(\mu, \sigma^2)$,

$$F(x \mid \mu, \sigma) = \Phi[(x-\mu)/\sigma],$$
$$f(x \mid \mu, \sigma) = (1/\sigma)\phi[(x-\mu)/\sigma],$$

portanto, $\boldsymbol{\theta} = (\mu, \sigma)$ é parâmetro de localização-escala. □

Considere-se um parâmetro de localização, θ, $\theta \in \Theta = \mathbf{R}$, e a estimação a partir de X, variável aleatória com função de densidade, $f(x \mid \theta) = f(x-\theta)$. O problema não é tão restrito como à primeira vista pode parecer porquanto X pode ser uma estatística suficiente [Não deixa, porém, de ser verdade que a doutrina da invariância se torna de aplicação mais difícil quando se lida directamente com variáveis N dimensionais (amostras) pese embora a simplificação introduzida pelo conceito de estatística invariante máxima estudado na secção 5.4].

Tem-se, $A = \Theta$; se a função perca é da forma, $L(\theta, a) = L(a-\theta)$, o problema é invariante em relação ao grupo de transformações,

$$G = \{g_c x = x + c\ :\ -\infty < c < +\infty\},$$

com $\overline{g}_c\theta = \theta + c$ e $\tilde{g}_c a = a + c$. $\overline{G}$ é grupo transitivo.

Uma função de decisão pura invariante deve satisfazer (5.17),

$$\delta(x+c) = \delta(x) + c, \tag{6.16}$$

para todo o x e todo o c. Escrevendo $x = 0$, vem $\delta(c) = \delta(0) + c$, para todo o c. Assim, qualquer função de decisão pura invariante tem a forma,

$$\delta_b(x) = x - b, \tag{6.17}$$

onde $b = -\delta_b(0)$ é um número real arbitrário. Pelo Teorema 5.3,

$$R(\theta, \delta) = R(\theta + c, \delta), \tag{6.18}$$

para todo o $\theta \in \Theta$ e todo o c; quer dizer, a função risco é independente de θ. Para a função de decisão δ_b definida por (6.17), vem,

$$R(\theta, \delta_b) = E_\theta\{L(X - b - \theta)\} = E_\theta\{L(X - b)\}, \tag{6.19}$$

pois como $R(\theta, \delta_b)$ não depende de θ pode tomar-se o valor esperado para $\theta = 0$. Deste modo, se $E_0\{L(X - b)\}$ existe e é finito, (6.17) é função de decisão pura invariante.

Na investigação da melhor função de decisão invariante pode confinar-se a análise às funções de decisão puras invariantes, pois,

Teorema 6.6 — Se todas as funções de decisão puras invariantes em relação a um grupo de transformações possuem risco constante[9], essas funções de decisão formam uma classe essencialmente completa dentro da classe de funções de decisão mistas invariantes.

Dem. Se $\delta^* \in D^*$ é função de decisão mista invariante, δ^* resulta da combinação de funções de decisão puras invariantes, $\delta \in D$, tais que, $R(\theta, \delta)$, não depende de θ. Sendo $R(\theta, \delta^*)$ média dos valores $R(\theta, \delta)$ para as $\delta \in D$ combinadas, há sempre um $\delta \in D$ para o qual, $R(\theta, \delta) \leqq R(\theta, \delta^*)$, para todo o $\theta \in \Theta$. □□

Teorema 6.7 — Na estimação de um parâmetro de localização, θ, com função perca, $L(\theta, a) = L(a-\theta)$, se $E_0\{L(X-b)\}$ existe e é finito para algum b e se existe, b_0, tal que,

$$E_0\{L(X - b_0)\} = \inf_b E_0\{L(X - b)\}, \tag{6.20}$$

então, $\delta(x) = x - b_0$ é a melhor função de decisão pura invariante. O respectivo risco, $R(\theta, \delta) =$ constante, é dado pelo 2.º membro de (6.20).

Dem. Consequência imediata das considerações que antecederam o Teorema 6.6. □□

Exemplo 6.9 — *Continuação.* Determine-se a melhor função de decisão pura invariante para estimar θ. Tem-se, $\delta(x) = x - b_0$, onde b_0 é o valor de b que minimiza, $E_0\{L(X - b)\}$; supondo, $L(\theta, a) = (a - \theta)^2$, vem,

$$\begin{aligned} E_0\{L(X - b)\} = E_0\{(X - b)^2\} &= \int_0^1 (x - b)^2 dx \\ &= (1 - 3b + 3b^2)/3, \end{aligned}$$

[9] Em situações em que $\overline{G}$ não é transitivo a casualização pode ser importante.

expressão que é mínima para $b = 1/2$. Assim, $\delta(x) = x - \frac{1}{2}$, é a função de decisão procurada. Note-se que $\delta(x) = x - E_0\{X\}$. □

Exemplo 6.12 — Sendo, $X_1, X_2, \ldots, X_N$, uma amostra casual de uma população com distribuição Exponencial com parâmetro de localização, θ, é fácil de verificar que $T = \min(X_i)$ é estatística suficiente para θ, com distribuição,

$$g(t\,|\,\theta) = N \exp\{-N(t-\theta)\},\ t > \theta;$$

pode, portanto, considerar-se T como variável a observar para estimar θ. A melhor decisão invariante é da forma, $\delta(t) = t - b_0$, onde b_0 é o valor de b que minimiza, $E_0\{L(T-b)\}$. Se a função perca é quadrática,

$$E_0\{L(T-b)\} = E_0\{(T-b)^2\} = \int_0^\infty (t-b)^2 N e^{-Nt} dt,$$

e, como já foi visto no exemplo anterior, tem-se

$$b_0 = E_0\{T\} = 1/N;$$

enfim, $\delta(t) = t - (1/N)$, ou, $\delta(\mathbf{x}) = \min(x_i) - (1/N)$. □

Exemplo 6.12 — *Continuação.* Considere-se outra função perca,

$$L(\theta, a) = \begin{cases} 0 & \text{se } |\,a-\theta\,| \leqq \Delta, \\ 1 & \text{se } |\,a-\theta\,| > \Delta. \end{cases} \qquad (\Delta > 0)$$

Tem-se,

$$\begin{aligned} E_0\{L(T-b)\} = P_0(|\,T-b\,| > \Delta) &= \int_0^{b-\Delta} N e^{-Nt} dt + \int_{b+\Delta}^\infty N e^{-Nt} dt \\ &= 1 - e^{-N(b-\Delta)} + e^{-N(b+\Delta)}, \end{aligned}$$

donde se conclui, por inspecção, que $b_0 = \Delta$ é o melhor valor para b, isto é, minimiza, $E_0\{L(T-b)\}$. Assim, $\delta(t) = t - \Delta$, ou $\delta(\mathbf{x}) = \min(x_i) - \Delta$, é a melhor função de decisão invariante para estimar θ quando a função perca tem a forma indicada. □

Este exemplo e o anterior mostram que quando se pode trabalhar com estatísticas suficientes unidimensionais se dá amplo campo de aplicação ao Teorema 6.7.

O facto da função de decisão invariante, $\delta(x) = x - b_0$, com b_0 determinado por (6.20), possuir risco constante — além de ser a melhor dentro da classe das invariantes — pode levar à conjectura[10], em presença do Teorema 4.22, de que δ é também minimax. Conhecem-se, no entanto, casos em que essa conjectura não se concretiza como no exemplo devido a Blackwell e Girshick (1954):

[10] Apesar do grupo aditivo não ser compacto.

Exemplo 6.13 — Suponha-se que $X - \theta$ tem função de probabilidade,

$$f(x) = 1/x(x+1),\ x = 1, 2, \ldots;$$

pretende estimar-se o parâmetro de localização, θ, com função perca dada por, $L(\theta, a) = \max(a-\theta, 0)$. As funções de decisão invariantes são da forma, $\delta(x) = x-b$, e possuem função risco,

$$\begin{aligned} R(\theta, \delta) = R(0, \delta) &= E_0\{\max(X-b, 0)\} \\ &= \sum_{x>b} \frac{x-b}{x(x+1)} = +\infty, \end{aligned}$$

para todo o b. Logo, as funções de decisão invariantes possuem todas função risco identicamente igual a $+\infty$.

Considere-se, por outro lado, a função de decisão não invariante, $\delta_0(x) = x - c\,|\,x\,|$, com $c > 1$. Tem-se,

$$R(\theta, \delta_0) = E_\theta\{L[\theta, \delta_0(X)]\} = \sum_{x > c\,|\,x+\theta\,|} (x - c\,|\,x+\theta\,|)/x(x+1).$$

Para, $\theta > -(c-1)/c$, a desigualdade, $x > c\,|\,x+\theta\,|$, é impossível donde sai que $R(\theta, \delta_0) = 0$; para $\theta \leq -(c-1)/c$, vem $r < x < s$, com $r = c\,|\,\theta\,|/(c+1)$ e $s = c\,|\theta\,|/(c-1)$. Assim,

$$\begin{aligned} R(\theta, \delta_0) &< \sum_{r<x<s} \frac{1}{x+1} < \int_r^{s+1} (1/x)\,dx \\ &= \log\left\{\frac{c+1}{c-1} + \frac{c+1}{c\,|\,\theta\,|}\right\} \leqq \log[2(c+1)/(c-1)]. \end{aligned}$$

Conclui-se assim que no caso presente nenhuma função de decisão invariante é minimax. □

Condições[11] em que a melhor função de decisão[12] é minimax encontram-se no seguinte,

---

[11] O que falha no exemplo anterior é verificar-se, para todo o $\delta \in D$,

$$\sup_\theta R(\theta, \delta) \geqq \inf_b E_0\{L(X-b)\} = E_0\{L(X-b_0)\};$$

o Teorema 6.8 estabelece que verificadas certas condições não muito restritivas sobre a função perca se obtém a desigualdade acima.

[12] Na estimação de um parâmetro de localização.

Teorema 6.8 — Nas condições do teorema anterior, se $F(x)$ designa a função de distribuição de X quando $\theta = 0$, se $L(\theta, a) = L(a - \theta)$ é função limitada inferiormente e se para todo $\xi > 0$ existe um número real M tal que,

$$\int_{-M}^{M} L(x - b)\, dF(x) \geqq E_0\{L(X - b_0)\} - \xi, \tag{6.21}$$

para todo o b, então $\delta(x) = x - b_0$ é minimax. A condição (6.21) é verificada quando $L(\theta, a) = L(a - \theta)$ é limitada ou quando, sendo função contínua de $(a - \theta)$, $L(a - \theta) \to +\infty$ se $(a - \theta) \to \pm\infty$.

A demonstração pode ver-se em Ferguson (1967). □□

Por outro lado, a verificar-se a admissibilidade da melhor decisão invariante, $\delta(x) = x - b_0$, que tem função risco constante, o Teorema 4.23 levaria a concluir que δ é minimax. Mas, como os Exs. 5.25 e 5.26 mostram, a admissibilidade pode falhar.

Considere-se seguidamente o caso em que θ, $\theta > 0$, é parâmetro de escala da distribuição de X e em que a função perca depende apenas de a/θ, $L(\theta, a) = L(a/\theta)$. O problema da estimação de θ é invariante em relação ao grupo multiplicativo,

$$G = \{g_c x = cx \,:\, c > 0\},$$

com, $\overline{g}_c\theta = c\theta$ e $\tilde{g}_c a = ca$. $\overline{G}$ é grupo transitivo.

Para facilitar a exposição somente se trata do caso, $P(X > 0) = 1$, probabilidade que a verificar-se para algum θ verifica-se para todo o θ em virtude de se tratar de um parâmetro de escala.

Fazendo, $X' = \log X$, $\theta' = \log\theta$, vem,

$$L'(\theta', a') = L(e^{\theta'}, e^{a'}) = L(e^{a' - \theta'}),$$

ou ainda, definindo, $L'(y) = L(e^y)$,

$$L'(\theta', a') = L'(a' - \theta'),$$

dá-se ao problema inicial estrutura semelhante à que se encontrou na estimação de um parâmetro de localização.

A melhor função de decisão invariante para estimar θ' é da forma $\delta(x') = x' - b'_0$, onde b'_0 minimiza,

$$E\{L'(X' - b') \,|\, \theta' = 0\};$$

em consequência, o melhor estimador invariante de $\log\theta$ é,

$$\delta^0(x) = \log x - \log b_0,$$

onde $b_0 = e^{b'_0}$ minimiza,

$$E\{L[\exp(X' - b')] \,|\, \theta = 1\} = E\{L(X/b) \,|\, \theta = 1\}.$$

Finalmente, a melhor função de decisão invariante para estimar o parâmetro de escala, θ, é, $\delta(x) = x/b_0$, onde b_0 é o valor de b que minimiza a expressão, $E\{L(X/b) \,|\, \theta = 1\}$. O Teorema 6.7 adapta-se facilmente ao caso presente.

Exemplo 6.14 — Com, X_i I.I.D., $X_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, N$, o problema consiste em estimar σ^2 quando a função perca é da forma,

$$L(\sigma^2, a) = [\log(a/\sigma^2)]^2. \tag{6.22}$$

Sabe-se que $T = \sum X_i^2$ é estatística suficiente para σ^2 e que T/σ^2 tem distribuição do χ^2 com N graus de liberdade. A melhor função de decisão invariante para estimar $\log \sigma^2$ é então, $\delta^0(t) = \log t - b'_0$ onde b'_0 é o valor de b' que minimiza,

$$E\{(\log T - b')^2 \,|\, \sigma^2 = 1\}.$$

Logo,

$$\begin{aligned} b'_0 &= E\{\log T \,|\, \sigma^2 = 1\} = [\Gamma(N/2)2^{N/2}]^{-1} \int_0^\infty e^{-t/2} t^{(N/2)-1} \log t \, dt \\ &= [\Gamma(N/2)]^{-1} \int_0^\infty e^{-z} z^{(N/2)-1} (\log z - \log 2) \, dz \\ &= \Psi(N/2) + \log 2, \end{aligned}$$

onde, $\Psi(u) = d \log \Gamma(u)/du$.

Finalmente, fazendo a necessária transformação,

$$\delta(t) = t/2 \exp\{\Psi(N/2)\},$$

ou seja,

$$\delta(\mathbf{x}) = \{\Sigma x_i^2\}/2 \exp\{\Psi(N/2)\},$$

é a melhor função de decisão invariante para estimar σ^2 quando a função perca tem a forma indicada [Ferguson (1967)]. □

Exemplo 6.15 — Seja a amostra casual, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. com função de densidade,

$$f(x \,|\, \theta) = [\Gamma(\alpha_0)\theta^{\alpha_0}]^{-1} e^{-x/\theta} x^{\alpha_0 - 1}, \; x > 0,$$

onde α_0 é conhecido e $\theta > 0$ é parâmetro de escala. Pretende determinar-se a melhor função de decisão invariante para estimar θ, quando a função perca é $L(\theta, a) = (a - \theta)^2/\theta^2$.

Sabe-se que $T = \Sigma X_i$ é estatística suficiente para θ e tem função de densidade,

$$g(t\,|\,\theta) = [\Gamma(N\alpha_0)\theta^{N\alpha_0}]^{-1} e^{-t/\theta} t^{N\alpha_0 - 1},\ t > 0,$$

que indica ser θ também parâmetro de escala para T. A melhor função de decisão invariante para estimar θ é então, $\delta(t) = t/b_0$, onde b_0 é o valor de b que minimiza,

$$E\left\{\left(\frac{T}{b} - 1\right)^2 \,|\, \theta = 1\right\},$$

isto é,

$$b_0 = \frac{E\{T^2\,|\,\theta = 1\}}{E\{T\,|\,\theta = 1\}} = N\alpha_0 + 1;$$

portanto, vem $\delta(t) = t/(N\alpha_0 + 1)$ ou $\delta(\mathbf{x}) = (\Sigma x_i)/(N\alpha_0 + 1)$. (Ferguson (1967)]. □

Exemplo 6.15 — *Continuação.* Considere-se a nova função perca,

$$L(\theta, a) = (a/\theta) - 1 - \log(a/\theta);$$

a melhor função de decisão invariante passa a ser, $\delta(t) = t/b_0$, onde b_0 é o valor de b que minimiza,

$$E\{[(T/b) - 1 - \log(T/b)]\,|\,\theta = 1\},$$

isto é, $b_0 = E\{T\,|\,\theta = 1\} = N\alpha_0$. Logo, $\delta(t) = t/N\alpha_0$ ou $\delta(\mathbf{x}) = (\Sigma x_i)/N\alpha_0$; note-se que,

$$E_\theta\{\delta(\mathbf{X})\} = E_\theta\{(\Sigma X_i)/N\alpha_0\} = \theta,$$

e como $\delta(\mathbf{X})$ é função da estatística suficiente completa, T, $\delta(\mathbf{X})$ é o estimador UMVU. □

O Teorema 6.8 estabelece condições em que o melhor estimador invariante é minimax. Quando se considera uma amostra casual de uma população Normal, $N(\theta, \sigma^2)$, com σ^2 conhecido, o Teorema 6.7 permite concluir que a função de decisão $\delta(\mathbf{x}) = \overline{x}$ [baseada na estatística suficiente ΣX_i] é a melhor invariante quando a função perca é quadrática [$b_0 = E_0\{\overline{X}\} = 0$] e o Teorema 6.8 permite concluir que também é minimax. O teorema seguinte cuja demonstração pode ver-se em Ferguson (1967) mostra que $\overline{X}$ continua a ser o melhor estimador invariante e a ser minimax para uma mais ampla classe de funções perca,

Teorema 6.9 — Se $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ é amostra casual de uma população Normal, $N(\theta, \sigma^2)$, com variância σ^2 conhecida, então $\overline{X}$ é o melhor estimador invariante e o estimador minimax de θ, sempre que a função perca, $L(\theta, a)$, seja função não decrescente de $|\,a - \theta\,|$ e $E_0\{L(\overline{X})\}$ exista e seja finito. □□

Retome-se a estimação da localização, θ, feita através da melhor função de decisão invariante, $\delta(x) = x - b_0$, dada pelo Teorema 6.7. Encontra-se aqui ilustrada a equivalência (que na secção 5.7 se afirmou ser virtual) entre tal função de decisão e a função de decisão Bayes (generalizada) contra a distribuição a priori não informativa.

Com $\Theta = (-\infty, +\infty)$, tome-se a distribuição a priori não informativa (correspondente à medida Haar invariante), $h(\theta) \propto 1$. Por (6.13),

$$f(x \mid \theta) = f(x - \theta),$$

e,

$$\begin{aligned} h(\theta \mid x) &= f(x-\theta) / \int_{-\infty}^{+\infty} f(x-\theta)\, d\theta \\ &= f(x-\theta); \end{aligned}$$

dada a função perca, $L(\theta, a) = L(a - \theta)$, o risco a posteriori assume a forma,

$$r_x(a) = \int_{-\infty}^{+\infty} L(a-\theta) f(x-\theta)\, d\theta.$$

Fazendo, $\theta = x - y$, $|\, d\theta/dy \,| = 1$,

$$\begin{aligned} r_x(a) &= \int_{-\infty}^{+\infty} L(y + a - x) f(y)\, dy \\ &= E_0\{L(X-b)\}, \qquad\qquad [b = x - a]. \end{aligned}$$

Dado x, a acção Bayes contra h, a_h, que minimiza $r_x(a)$ é, obviamente,

$$a_h = x - b_0,$$

onde b_0 é o valor de b que minimiza $E_0\{L(X-b)\}$. A equivalência é, pois, bem clara.

Na estimação da escala, θ, $\Theta = (0, +\infty)$, tome-se a distribuição a priori não informativa, $h(\theta) \propto 1/\theta$. Vem, $h(\theta \mid x) \propto (1/\theta^2) f(x/\theta)$, e, com $L(\theta, a) = L(a/\theta)$, fazendo $x/\theta = y$,

$$\begin{aligned} r_x(a) &= \int_0^{\infty} L(y/b) f(y)\, dy \\ &= E_{\theta=1}\{L(X/b)\} \qquad\qquad [b = x/a]. \end{aligned}$$

Dado x, a acção Bayes contra h, a_h, que minimizar $r_x(a)$ é, como se vê,

$$a_h = x/b_0,$$

onde b_0 é o valor de b que minimiza $E_{\theta=1}\{L(X/b)\}$.

6.4 Melhores estimadores invariantes [II]. Estimadores Pitman

Considere-se uma amostra de N observações, $X_1, X_2, \dots, X_N$, não necessariamente independentes, com distribuição conjunta em que θ é parâmetro de localização, quer dizer, $X_1 - \theta, X_2 - \theta, \dots, X_N - \theta$, têm distribuição independente de θ. Supondo, sem perda de generalidade, que $X_i, i = 1, 2, \dots, N$, são variáveis aleatórias contínuas, a respectiva função de densidade conjunta pode escrever-se,

$$f(x_1, x_2, \dots, x_N \,|\, \theta) = f(x_1 - \theta, x_2 - \theta, \dots, x_N - \theta). \tag{6.23}$$

Com $\mathfrak{X} = \mathbf{R}^N$ considere-se a estimação de θ com função perca, $L(\theta, a) = L(a - \theta)$. O problema é invariante em relação ao grupo de transformações,

$$g_c(x_1, x_2, \dots, x_N) = (x_1 + c, x_2 + c, \dots, x_N + c), \quad c \in \mathbf{R},$$

sendo $\overline{g}_c\theta = \theta + c$, $\tilde{g}_c a = a + c$. Evidentemente, $\overline{G}$ é grupo transitivo. Qualquer estimador invariante deve satisfazer a relação,

$$\delta(x_1 + c, x_2 + c, \dots, x_N + c) = \delta(x_1, x_2, \dots, x_N) + c,$$

para todo o $(x_1, x_2, \dots, x_N) \in \mathbf{R}^N$ e todo o $c \in \mathbf{R}$.

A estimação de θ pode transformar-se num problema unidimensional, embora condicional, de modo a ter acesso à doutrina da secção anterior. Um possível artifício — a que adiante se dá expressão mais geral — consiste em mudar de variáveis,

$$(X_1, X_2, \dots, X_N) \Rightarrow (Y_1, Y_2, \dots, Y_{N-1}, X_N),\ Y_j = X_j - X_N,\ j = 1, 2, \dots, N-1,$$

onde, $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{N-1})$, tem componentes cuja distribuição não depende de θ, porquanto,

$$Y_j = X_j - X_N = X_j - \theta - (X_N - \theta),\ j = 1, 2, \dots, N-1,$$

e em admitir que a experiência se realiza em duas fases: na primeira, observa-se $\mathbf{Y}$; na segunda, observa-se $X_N \,|\, \mathbf{Y}$. A decomposição em duas fases pode traduzir-se em termos da função de densidade conjunta de $(\mathbf{Y}, X_N)$ [note-se que o Jacobiano da transformação é igual a um],

$$f(y_1, y_2, \dots, y_{N-1}) f(x_N \,|\, \mathbf{y}, \theta),$$

onde θ é parâmetro de localização da distribuição de $X_N \,|\, \mathbf{Y}$.

A matéria da secção anterior — Teorema 6.7 — leva a conjecturar que o melhor estimador invariante no problema condicionado é,

$$\delta_0(X_1, X_2, \dots, X_N) = X_N - b_0(Y_1, Y_2, \dots, Y_{N-1}),$$

ou,

$$\delta_0(\mathbf{X}) = X_N - b_0(\mathbf{Y}), \tag{6.24}$$

onde, para cada $\mathbf{y}$, $b_0(\mathbf{y})$, se existir, é o número tal que,

$$E_0\{L(X_N - b_0) \,|\, \mathbf{y}\} = \inf_b E_0\{L(X_N - b) \,|\, \mathbf{y}\}. \tag{6.25}$$

De facto pode demonstrar-se que a conjectura é verdadeira. No entanto, a demonstração vai fazer-se numa perspectiva mais geral.

Nota-se, primeiramente, que na análise condicionada, em vez de tomar, $X_N \,|\, \mathbf{Y}$, pode tomar-se, $\delta(\mathbf{X}) \,|\, \mathbf{Y}$, onde δ é um qualquer estimador invariante; $\delta(\mathbf{X}) = X_N$ é caso particular do segundo porquanto é, manifestamente, estimador invariante. Assim, a função de decisão, δ_0, dada por (6.24), pertence à família introduzida no seguinte,

Lema 6.1 — Se δ é um estimador invariante qualquer, a condição necessária e suficiente para que δ_I seja invariante é que,

$$\delta_I(\mathbf{x}) = \delta(\mathbf{x}) + W(\mathbf{x}), \tag{a}$$

onde, $W(\mathbf{x})$ é qualquer função invariante,

$$W(\mathbf{x} + c\mathbf{1}) = W(\mathbf{x}),\ \mathbf{1} = (1, 1, \dots, 1), \tag{b}$$

para todo o $\mathbf{x} \in \mathbf{R}^N$ e $c \in \mathbf{R}$.

Dem. Verificando-se [a] e [b] tem-se,

$$\delta_I(\mathbf{x} + c\mathbf{1}) = \delta(\mathbf{x} + c\mathbf{1}) + W(\mathbf{x} + c\mathbf{1}) = \delta(\mathbf{x}) + c + W(\mathbf{x}) = \delta_I(\mathbf{x}) + c,$$

e δ_I é invariante. Por outra parte, se δ_I é invariante, seja,

$$W(\mathbf{x}) = \delta_I(\mathbf{x}) - \delta(\mathbf{x}).$$

Então,

$$\begin{aligned} W(\mathbf{x} + c\mathbf{1}) &= \delta_I(\mathbf{x} + x\mathbf{1}) - \delta(\mathbf{x} + x\mathbf{1}) \\ &= \delta_I(\mathbf{x}) + c - \delta(\mathbf{x}) - c = W(\mathbf{x}), \end{aligned}$$

e verificam-se [a] e [b]. □□

Combinando com o Teorema 5.7 [$W(\mathbf{x})$ sendo invariante é necessariamente função de invariante máxima] e notando que, no caso presente [veja-se Lehmann (1983)], qualquer invariante máxima é função das diferenças, $y_i = x_i - x_N$,

$i = 1, 2, \ldots, N-1$ — veja-se Ex. 5.14 — obtém-se a expressão geral dos estimadores invariantes da localização, θ [onde se tomou $W = -b$],

$$\delta_I(\mathbf{X}) = \delta(\mathbf{X}) - b(\mathbf{Y}),$$

donde sai, evidentemente, (6.24)[13].

A proposição que permite concluir que (6.24), com (6.25), é a melhor função de decisão invariante contém-se no,

Lema 6.2 — Se $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ tem distribuição (6.23), sejam $Y_i = X_i - X_N$, $i = 1, 2, \ldots, N-1$ e $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{N-1})$. Na estimação de θ com função perca, $L(\theta, a) = L(a - \theta)$, suponha-se que existe um estimador invariante, δ, com risco finito, e que para cada $\mathbf{y}$ existe um número, $b(\mathbf{y}) = b_0(\mathbf{y})$, que minimiza,

$$E_0\{L[\delta(\mathbf{X}) - b(\mathbf{y})] \,|\, \mathbf{y}\}.$$

Então, o melhor estimador invariante existe e tem por expressão,

$$\delta_I^0(\mathbf{X}) = \delta(\mathbf{X}) - b_0(\mathbf{y}).$$

Dem. Sendo δ_I um estimador invariante qualquer,

$$R(\theta, \delta_I) = E_0\{L[\delta(\mathbf{X}) - b(\mathbf{Y})]\} = E\{E_0\{L[\delta(\mathbf{X}) - b(\mathbf{Y})] \,|\, \mathbf{Y}\}\},$$

onde,

$$E\{E_0\{L[\delta(\mathbf{X}) - b(\mathbf{Y})] \,|\, \mathbf{Y}\}\} =$$
$$= \int \cdots \int E_0\{L[\delta(\mathbf{X}) - b(\mathbf{y})] \,|\, \mathbf{y}\} f(y_1, y_2, \ldots, y_{N-1})\, dy_1 dy_2 \ldots dy_{N-1}.$$

Mas,

$$\begin{aligned} E\{E_0\{L[\delta(\mathbf{X}) - b(\mathbf{Y})] \,|\, \mathbf{Y}\}\} &\geqq E\{E_0\{L[\delta(\mathbf{X}) - b_0(\mathbf{Y})] \,|\, \mathbf{Y}\}\} \\ &= E_0\{L[\delta(\mathbf{X}) - b_0(\mathbf{Y})]\} \\ &= R(\theta, \delta_I^0), \end{aligned}$$

e a demonstração fica completa pois a minimização enunciada faz sentido por δ ter risco finito, isto é, $E_0\{L[\delta(\mathbf{X})] \,|\, \mathbf{y}\} < \infty$. □□

[13] O artifício que conduziu a (6.24) consiste afinal em seleccionar uma função de decisão invariante — optou-se, para melhor intuição, por X_N — transformando o problema N-dimensional em unidimensional condicionado pela estatística invariante máxima,

$$\mathbf{Y} = (X_1 - X_N, \ldots, X_{N-1} - X_N),$$

que é, aliás, uma estatística ancilária pois a sua distribuição não depende de θ.

Repare-se que o Teorema 6.7 é caso particular da proposição acima obtida com $N = 1$. Note-se também que o Lema 6.2 modifica-se facilmente para dar cobertura a estimação invariante de parâmetros de escala [veja-se Lehmann (1983)].

A proposição seguinte, cuja demonstração pode estudar-se no autor referido, diz respeito ao problema de existência.

Lema 6.3 — Nas condições do Lema 6.2, se $L(\theta, a) = L(a - \theta)$ é função convexa e não monótona de $(a - \theta)$, então existe para θ um estimador IRM; se é função estritamente convexa, existe um estimador IRM e é único. □□

Uma análise moldada na que foi efectuada no final da secção anterior mostra que $\delta_0(\mathbf{X})$, tal como definida em (6.24) e (6.25), é função de decisão Bayes generalizada[14].

A função de densidade conjunta de $(Y_1, Y_2, \dots, Y_{N-1}, X_N)$, quando $\theta = 0$, pode escrever-se,

$$f(y_1, y_2, \dots, y_{N-1}, x_N) = f(y_1 + x_N, \dots, Y_{N-1} + x_N, x_N),$$

onde f é a função de densidade introduzida no 2.º membro de (6.23). A função de densidade marginal de $\mathbf{Y}$ vem dada por,

$$f(y_1, y_2, \dots, y_{N-1}) = \int f(y_1 + x_N, \dots, y_{N-1} + x_N, x_N)\, dx_N;$$

a função de densidade de X_N condicionada por $\mathbf{Y}$, quando $\theta = 0$, é,

$$f(x_N \,|\, \mathbf{y},\ \theta = 0) = \frac{f(y_1 + x_N, \dots, y_{N-1} + x_N, x_N)}{\int f(y_1 + x_N, \dots, y_{N-1} + x_N, x_N)\, dx_N}.$$

Por outro lado, com $\Theta = (-\infty, +\infty)$, $h(\theta) \propto 1$, a expressão do risco a posteriori é,

$$r_{\mathbf{x}}(a) = \frac{\int_{-\infty}^{+\infty} L(a - \theta) f(x_1 - \theta, \dots, x_{N-1} - \theta, x_N - \theta)\, d\theta}{\int_{-\infty}^{+\infty} f(x_1 - \theta, \dots, x_{N-1} - \theta, x_N - \theta)\, d\theta};$$

substituindo θ por $x_N - z$, e notando ser $x_i - x_N = y_i$, $i = 1, 2, \dots, N - 1$, vem,

$$\begin{aligned} r_{\mathbf{x}}(a) &= \int_{-\infty}^{+\infty} L(z + a - x_N) f(z \,|\, \mathbf{y}, \theta = 0)\, dz \\ &= E_0\{L(X_N - b) \,|\, \mathbf{y}\}, \qquad\qquad [b = x_N - a]. \end{aligned}$$

Confrontando com o caso unidimensional chega-se à conclusão de que no caso N-dimensional também se verifica a equivalência entre a melhor função de decisão invariante e a função de decisão Bayes generalizada[15].

[14] A prova podia fazer-se para os estimadores [a] do Lema 6.1.

[15] A demonstração quando θ é parâmetro de escala corre nas mesmas linhas.

Quando a função perca é quadrática, (6.24) é conhecido por estimador Pitman para a localização. Sendo, $L(\theta, a) = (a-\theta)^2$, tem-se, de (6.25),

$$E_0\{L(X_N - b) \,|\, \mathbf{Y}\} = E_0\{(X_N - b)^2 \,|\, \mathbf{Y}\},$$

expressão que é mínima para,

$$b_0 = E_0\{X_N \,|\, \mathbf{Y}\},$$

donde,

$$\delta_0(\mathbf{X}) = X_N - E_0\{X_N \,|\, \mathbf{Y}\}. \tag{6.26}$$

Recordando a forma de $f(x_N \,|\, \mathbf{y},\ \theta = 0)$, o estimador Pitman para a localização assume o aspecto,

$$\delta_0(\mathbf{X}) = X_N - \frac{\int x_N f(Y_1 + x_N, \dots, Y_{N-1} + x_N, x_N)\, dx_N}{\int f(Y_1 + x_N, \dots, Y_{N-1} + x_N, x_N)\, dx_N}; \tag{6.27}$$

finalmente, fazendo $x_N = X_N - \theta$, isto é, $\theta = X_N - x_N$, o estimador vem na forma habitual,

$$\delta_0(\mathbf{X}) = \frac{\int \theta f(X_1 - \theta, X_2 - \theta, \dots, X_N - \theta)\, d\theta}{\int f(X_1 - \theta, X_2 - \theta, \dots, X_N - \theta)\, d\theta}. \tag{6.28}$$

Se as variáveis, $X_1, X_2, \dots, X_N$, são I.I.D. com função de densidade f, tem-se, em vez de (6.28),

$$\delta_0(\mathbf{X}) = \frac{\int \theta \Pi f(X_i - \theta)\, d\theta}{\int \Pi f(X_i - \theta)\, d\theta}. \tag{6.29}$$

Os exemplos que vão ser apresentados foram recolhidos em Mood, Graybill e Boes (1974).

Exemplo 6.16 — Com X_i I.I.D., $X_i \sim N(\theta, 1)$, $i = 1, 2, \dots, N$, o estimador Pitman do parâmetro de localização, θ, é, por (6.29),

$$\begin{aligned}
\delta_0(\mathbf{X}) &= \int_{-\infty}^{+\infty} \theta(\sqrt{2\pi})^{-N} \exp\left\{-\frac{1}{2}\Sigma(X_i - \theta)^2\right\} d\theta \Big/ \int_{-\infty}^{+\infty} (\sqrt{2\pi})^{-N} \exp\left\{-\frac{1}{2}\Sigma(X_i - \theta)^2\right\} d\theta \\
&= \int_{-\infty}^{+\infty} \theta \exp\left\{-\frac{N}{2}\theta^2 + \theta\Sigma X_i\right\} d\theta \Big/ \int_{-\infty}^{+\infty} \exp\left\{-\frac{N}{2}\theta^2 + \theta\Sigma X_i\right\} d\theta \\
&= \int_{-\infty}^{+\infty} \theta \exp\left\{-\frac{N}{2}(\theta - \overline{X})^2\right\} d\theta \Big/ \int_{-\infty}^{+\infty} \exp\left\{-\frac{N}{2}(\theta - \overline{X})^2\right\} d\theta \\
&= \int_{-\infty}^{+\infty} \theta\sqrt{N/2\pi} \exp\left\{-\frac{N}{2}(\theta - \overline{X})^2\right\} d\theta \Big/ \int_{-\infty}^{+\infty} \sqrt{N/2\pi} \exp\left\{-\frac{N}{2}(\theta - \overline{X})^2\right\} d\theta,
\end{aligned}$$

donde, finalmente, $\delta_0(\mathbf{X}) = \overline{X}$, que é também o estimador UMVU. □

Exemplo 6.17 — Se $\mathbf{X} = (X_1, X_2, \dots, X_N)$, X_i I.I.D., com distribuição Uniforme no intervalo, $[\theta - (1/2), \theta + (1/2)]$, θ parâmetro de localização, o estimador Pitman tem por expressão,

$$\begin{aligned}\delta_0(\mathbf{X}) &= \int \theta \Pi I_{\left(\theta-\frac{1}{2},\,\theta+\frac{1}{2}\right)}(X_i)\,d\theta \Big/ \int \Pi I_{\left(\theta-\frac{1}{2},\,\theta+\frac{1}{2}\right)}(X_i)\,d\theta \\ &= \int \theta \Pi I_{\left(X_i-\frac{1}{2},\,X_i+\frac{1}{2}\right)}(\theta)\,d\theta \Big/ \int \Pi I_{\left(X_i-\frac{1}{2},\,X_i+\frac{1}{2}\right)}(\theta)\,d\theta,\end{aligned}$$

ou ainda, com $T_1 = \min(X_i)$, $T_2 = \max(X_i)$,

$$\delta_0(\mathbf{X}) = \int_{T_2-\frac{1}{2}}^{T_1+\frac{1}{2}} \theta\,d\theta \Big/ \int_{T_2-\frac{1}{2}}^{T_1+\frac{1}{2}} d\theta = (T_1 + T_2)/2,$$

□

Quando θ é parâmetro de escala, a conversão na estimação de um parâmetro de localização, tal como foi feito no final da secção anterior, conduz, com função perca,

$$L(\theta, a) = (a - \theta)^2/\theta^2,$$

ao estimador Pitman para a escala,

$$\delta_0^*(\mathbf{X}) = \frac{\int (1/\theta^{N+2}) f(X_1/\theta, X_2/\theta, \dots, X_N/\theta)\,d\theta}{\int (1/\theta^{N+3}) f(X_1/\theta, X_2/\theta, \dots, X_N/\theta)\,d\theta}, \tag{6.30}$$

ou, quando as variáveis X_i são I.I.D.,

$$\delta_0^*(X) = \frac{\int (1/\theta^{N+2}) \Pi f(X_i/\theta)\,d\theta}{\int (1/\theta^{N+3}) \Pi f(X_i/\theta)\,d\theta}. \tag{6.31}$$

Exemplo 6.18 — Com $\mathbf{X} = (X_1, X_2, \dots, X_N)$, X_i I.I.D. com função de densidade, $f(x\,|\,\theta) = (1/\theta) I_{(0,\theta)}(x)$, o estimador Pitman para o parâmetro de escala,

$$\begin{aligned}\delta_0^*(\mathbf{X}) &= \int_0^\infty (1/\theta^2) \Pi (1/\theta) I_{(0,\theta)}(X_i)\,d\theta \Big/ \int_0^\infty (1/\theta^3) \Pi (1/\theta) I_{(0,\theta)}(X_i)\,d\theta \\ &= \int_0^\infty \theta^{-N-2} \Pi I_{(X_i,\infty)}(\theta)\,d\theta \Big/ \int_0^\infty \theta^{-N-3} \Pi I_{(X_i,\infty)}(\theta)\,d\theta \\ &= \int_{\max(X_i)}^\infty \theta^{-N-2} d\theta \Big/ \int_{\max(X_i)}^\infty \theta^{-N-3} d\theta \\ &= [(N+2)/(N+1)] \max(X_i).\end{aligned}$$

Como $\max(X_i)$ é estatística suficiente completa para θ e

$$E_\theta\{\max(X_i)\} = [N/(N+1)]\theta,$$

tem-se que o estimador UMVU é $[(N+1)/N] \max(X_i)$. □

Exemplo 6.19 — Se $\mathbf{X} = (X_1, X_2, \dots, X_N)$, X_i I.I.D. com função de densidade, $f(x\,|\,\theta) = (1/\theta)\exp\{-x/\theta\}$, $x > 0$, o estimador Pitman de θ,

$$\begin{aligned}\delta_0^*(\mathbf{X}) &= \int_0^\infty (1/\theta^2)(1/\theta^N)\exp\{-\Sigma X_i/\theta\}\,d\theta \Big/ \int_0^\infty (1/\theta^3)(1/\theta^N)\exp\{-\Sigma X_i/\theta\}\,d\theta \\ &= \int_0^\infty (\omega/\Sigma X_i)^{N+2} e^{-\omega}(\Sigma X_i/\omega^2)\,d\omega \Big/ \int_0^\infty (\omega/\Sigma X_i)^{N+3} e^{-\omega}(\Sigma X_i/\omega^2)\,d\omega,\end{aligned}$$

onde se fez, $\omega = \Sigma X_i/\theta$. Assim,

$$\begin{aligned}\delta_0^*(\mathbf{X}) &= (\Sigma X_i)\int_0^\infty \omega^N e^{-\omega}d\omega \Big/ \int_0^\infty \omega^{N+1}e^{-\omega}d\omega \\ &= \Sigma X_i/(N+1).\end{aligned}$$

Como pode verificar-se, o estimador UMVU é $\Sigma X_i/N$. Tanto $\Sigma X_i/N$ como $\Sigma X_i/(N+1)$ são estimadores invariantes; no entanto, em face da doutrina exposta, se a função perca for, $L(\theta, a) = (a-\theta)^2/\theta^2$, o segundo tem função risco uniformemente inferior à do primeiro. □

Voltando a considerar parâmetros de localização, Girshick e Savage (1951) demonstraram [em condições mais gerais do que as indicadas] o seguinte,

Teorema 6.10 — Se $\mathbf{X} = (X_1, X_2, \dots, X_N)$ tem distribuição (6.23) e se o estimador Pitman definido por (6.26) tem variância finita, então o estimador Pitman é minimax.

Dem. Veja-se Lehmann (1983). □□

O teorema acima contém uma condição suficiente. Outra aproximação ao problema pode ser feita através da extensão do Teorema 6.8 ao caso condicionado [vejam-se Blackwell e Girshick (1954) e Ferguson (1967)].

A admissibilidade — propriedade mais forte do que a minimax — é naturalmente problema mais complexo. De facto, em termos gerais, uma função de decisão IRM mesmo que seja inadmissível tem uma razoável possibilidade de ser minimax [veja-se a secção 6.6]. O problema da admissibilidade do estimador Pitman foi estudado por Stein (1959) a quem se deve uma condição suficiente que pode ver-se em Zacks (1981). O seguinte resultado [Lehmann (1983)] é um caso especial,

Teorema 6.11 — Se X_i, $i = 1, 2, \dots, N$, são I.I.D. com densidade $f(x-\theta)$, e se existe um estimador invariante, δ_0, de θ, tal que $E_0\{|\,\delta_0(\mathbf{X})\,|^3\} < \infty$, então o estimador Pitman definido por (6.26) é admissível. □□

Como era de esperar o estimador Pitman é Bayes generalizado [veja-se análise pág. 290–295]. Esta conclusão é aliás imediata quando se atende à escrita habitual: de (6.28), com $L(\theta, a) = (\theta - a)^2$, conclui-se imediatamente ser, $\delta_0(\mathbf{X}) = E\{\theta \,|\, \mathbf{X}\}$; de (6.30), com $L(\theta, a) = (\theta - a)^2/\theta^2$, tem-se

$$\delta_0^*(\mathbf{X}) = \frac{E\{1/\theta \,|\, \mathbf{X}\}}{E\{1/\theta^2 | \, \mathbf{X}\}}.$$

Para entender bem a segunda expressão faça-se o confronto com (6.52) e recorde-se que no caso do parâmetro de escala a distribuição a priori não informativa é $h(\theta) \propto 1/\theta$, donde,

$$h(\theta \,|\, \mathbf{x}) \propto (1/\theta^{N+1}) f(x_1/\theta, x_2/\theta, \ldots, x_N/\theta),$$

ou,

$$h(\theta \,|\, \mathbf{x}) \propto (1/\theta^{N+1}) \Pi f(x_i \,|\, \theta),$$

quando a amostra é casual.

Nos Exs. 6.16, 6.18 e 6.19 fez-se ligeiro confronto entre estimadores UMVU e melhores estimadores invariantes. Em termos gerais não se pode ir além de algumas simples notas [veja-se Lehmann (1983)]:

(i) Os estimadores UMVU resultam bem no quadro das famílias exponenciais, mas não existem praticamente para as famílias grupo [famílias obtidas aplicando à variável experimental um grupo de transformações — notem-se, por exemplo, (6.13), (6.14) e (6.15)], domínio por excelência dos estimadores invariantes, havendo entre as duas famílias pouco em comum[16].

São exemplos de famílias grupo, a Normal [que pertence também à família exponencial], e, entre outras, a Exponencial Dupla,

$$f(x \,|\, \theta_1, \theta_2) = (1/2\theta_2) \exp\{-|\, x - \theta_1 \,|\, /\theta_2\}, \; -\infty < x < +\infty,$$

a Cauchy,

$$f(x \,|\, \theta_1, \theta_2) = (\theta_2/\pi)\{1/[\theta_2^2 + (x - \theta_1)^2]\}, \; -\infty < x < +\infty,$$

a Logística,

$$f(x \,|\, \theta_1, \theta_2) = (1/\theta_2)\{e^{-(x-\theta_1)/\theta_2}/[1 + e^{-(x-\theta_1)/\theta_2}]\}, \; -\infty < x < +\infty,$$

[16] Por exemplo, quando θ é parâmetro de localização, a variável aleatória, X, tem distribuição da família exponencial uniparamétrica e da família grupo (aditivo ou localização) quando e só quando, X tem distribuição Normal (com variância conhecida), ou $X = c \log Y$, $c \neq 0$, Y com distribuição Gama. Para outros aspectos veja-se Lehmann (1983).

a Uniforme,

$$f(x\,|\,\theta_1,\theta_2) = (1/\theta_2),\ \theta_1 - \frac{\theta_2}{2} < x < \theta_1 + \frac{\theta_2}{2},$$

sendo em qualquer dos casos, $-\infty < \theta_1 < +\infty$ e $\theta_2 > 0$.

(ii) Enquanto os estimadores UMVU dão bons resultados com função perca convexa, os melhores estimadores invariantes não se encontram dependentes de tal restrição. Em particular, se a função perca é limitada, não existem estimadores UMVU.

(iii) Os estimadores UMVU são frequentemente inadmissíveis; os melhores estimadores invariantes [veja-se o Teorema 6.11 que diz respeito aos estimadores Pitman] são admissíveis em condições não muito restritivas.

(iv) Com função perca quadrática demonstram-se as seguintes proposições:

[A] Quando $\delta(\mathbf{X})$ é um estimador invariante com enviesamento constante, v, então, $\delta(\mathbf{X}) - v$, é invariante, não enviesado e tem risco inferior ao de $\delta(\mathbf{X})$;
[B] O melhor estimador invariante, sendo único, é não enviesado;
[C] Se um estimador UMVU existe é o melhor estimador invariante [veja-se o Ex. 6.16; nos Exs. 6.18 e 6.19 não há coincidência pois a função perca não é quadrática].

6.5 Admissibilidade com função perca quadrática

Como certamente não passou despercebido, no tratamento dos problemas de estimação há forte tendência para introduzir funções perca quadráticas,

$$L(\theta, a) = \lambda(\theta)(\theta - a)^2,\ \lambda(\theta) > 0,$$

onde $\lambda(\theta)$ é um factor de ponderação que para maior simplicidade se toma muitas vezes idêntico à unidade.

Dado que para cada $\theta \in \Theta$ $[\Theta = \mathbf{R}]$ há sempre um $a \in A$ $[A = \mathbf{R}]$ para o qual o mínimo de $L(\theta, a)$ é igual a zero, a função perca é idêntica à função pesar. Em contrapartida, como se trata de função não limitada, pode conduzir a que a máxima perca esperada seja infinita para toda a classe de funções de decisão o que, naturalmente, facilita muito pouco o estudo das soluções minimax.

Há vários argumentos[17] para justificar ou desculpar o emprego de funções perca quadráticas. Primeiro, por darem uma boa aproximação de funções perca na vizinhança do respectivo mínimo desde que essas funções não sejam demasiado achatadas nessa região; aliás, nos problemas que envolvem a distribuição Normal, a função perca quadrática serve tão bem como qualquer função perca crescente

[17] Veja-se secção 4.2.

com $|\theta - a|$. Segundo, pode aduzir-se a relação da função perca quadrática com o erro quadrático médio ou com a variância se a função de decisão é não enviesada. Terceiro, e complementarmente, há que destacar a acessibilidade ao tratamento matemático.

O primeiro resultado que vai apresentar-se é consequência da bem conhecida desigualdade de Frechet-Cramér-Rao. Considere-se o vector aleatório, $\mathbf{X}$, com função de densidade na família, $\mathcal{F} = \{f(\mathbf{x}\,|\,\theta) : \theta \in \Theta\}$, satisfazendo as condições de regularidade [veja-se Cramér (1946)] requeridas para estabelecer essa desigualdade; seja, D, a classe das funções de decisão puras ou classe dos estimadores de θ; com $\delta \in D$, escreva-se,

$$E_\theta\{\delta(\mathbf{X})\} = \theta + v_\delta(\theta), \tag{6.32}$$

onde $v_\delta(\theta)$ é o enviesamento de δ, e retenha-se a desigualdade de Frechet-Cramér--Rao na forma de interesse para o caso presente,

$$\begin{aligned} R(\theta,\delta) &= E_\theta\{L[\theta,\delta(\mathbf{X})]\} = E_\theta\{[\delta(\mathbf{X}) - \theta]^2\} \\ &\geqq [v_\delta(\theta)]^2 + [(1 + v'_\delta(\theta))^2 / J(\theta)] = \Omega_\delta(\theta), \end{aligned} \tag{6.33}$$

onde, $v'_\delta(\theta) = dv_\delta(\theta)/d\theta$, $J(\theta)$ é a quantidade de informação de Fisher,

$$\begin{aligned} J(\theta) &= \int [\partial \log f(\mathbf{x}\,|\,\theta)/\partial\theta]^2 f(\mathbf{x}\,|\,\theta)\, d\mathbf{x} \\ &= -\int [\partial^2 \log f(\mathbf{x}\,|\,\theta)/\partial\theta^2] f(\mathbf{x}\,|\,\theta)\, d\mathbf{x}, \end{aligned}$$

e $\Omega_\delta(\theta)$ é a forma abreviada do 2.º membro da desigualdade.

O teorema seguinte, devido a Hodges e Lehmann (1951), estabelece uma condição suficiente de admissibilidade,

Teorema 6.12 — Se $\delta_0 \in D$ tem erro quadrático médio igual ao segundo membro da desigualdade FCR para todo o $\theta \in \Theta$ e se para qualquer outro $\delta_1 \in D$, a relação $\Omega_{\delta_1}(\theta) \leqq \Omega_{\delta_0}(\theta)$ para todo o $\theta \in \Theta$ implica $v_{\delta_1}(\theta) = v_{\delta_0}(\theta)$ para todo o $\theta \in \Theta$, então δ_0 é admissível quando a função perca é quadrática.

Dem. Se δ_0 não é admissível existe δ_1 tal que, $R(\theta,\delta_1) \leqq R(\theta,\delta_0)$ para $\theta \in \Theta$ e desigualdade estrita pelo menos para um valor de θ. Verificando-se, $R(\theta,\delta_0) = \Omega_{\delta_0}(\theta)$ para todo o $\theta \in \Theta$, tem-se,

$$\Omega_{\delta_1}(\theta) \leqq R(\theta,\delta_1) \leqq R(\theta,\delta_0) = \Omega_{\delta_0}(\theta),$$

ainda para todo o $\theta \in \Theta$. De acordo com a hipótese, $v_{\delta_1}(\theta) = v_{\delta_0}(\theta)$ para todo o $\theta \in \Theta$; portanto, $v'_{\delta_1}(\theta) = v'_{\delta_0}(\theta)$ e $\Omega_{\delta_1}(\theta) = \Omega_{\delta_0}(\theta)$ para todo o $\theta \in \Theta$ o que contradiz a dominação estrita de δ_0 por δ_1. Logo, δ_0 é admissível. □□

Note-se que o teorema permanece válido quando a função perca é da forma, $\lambda(\theta)(\theta - a)^2$.

Exemplo 6.20 — Se $\mathbf{X}$ é uma amostra casual de dimensão N de uma população $N(\theta, 1)$ — faz-se $\sigma^2 = 1$ para simplificar — o teorema anterior permite demonstrar que o estimador habitual, $\delta(\mathbf{X}) = \overline{X}$ é admissível. Tem-se que $\overline{X}$ tem erro quadrático médio, $1/N$, igual ao segundo membro da desigualdade FCR, com $v_{\overline{X}}(\theta) = 0$. Seja, δ, um estimador de θ tal que para todo o $\theta \in \Theta$,

$$R(\theta, \delta) < R(\theta, \overline{X}),$$

isto é,

$$[v_\delta(\theta)]^2 + [1 + v'_\delta(\theta)]^2/N \leqq 1/N, \tag{6.34}$$

para todo o θ. Como as parcelas do primeiro membro não podem ser negativas,

$$[1 + v'_\delta(\theta)]^2 \leqq 1 \quad \Rightarrow \quad 2v'_\delta(\theta) + [v'_\delta(\theta)]^2 \leqq 0 \quad \Rightarrow \quad v'_\delta(\theta) \leqq 0,$$

e ainda,

$$[v_\delta(\theta)]^2 \leqq 1/N \quad \Rightarrow \quad |\, v_\delta(\theta)\,| \text{ limitado.}$$

Se a função $v_\delta(\theta)$ é limitada e se $v'_\delta(\theta) \leqq 0$, pode mostrar-se que existe uma sucessão de valores de θ, seja $\{\theta_k\}$, $k = 1, 2, \ldots$, tal que $v'_\delta(\theta_k) \to 0$ quando $\theta_k \to \infty$. Com efeito, suponha-se que tal não se verifica e que $v'_\delta(\theta_k) \leqq -\varepsilon$ para todo o $k \geqq k_0$; $v_\delta(\theta)$ decrescia continuamente e não podia ser função limitada. De modo análogo se demonstra que existe uma sucessão de valores de θ, $\{\theta_j\}$, $j = 1, 2, \ldots$, tal que $v'_\delta(\theta_j) \to 0$ quando $\theta_j \to -\infty$. Consequentemente, por (6.34) verifica-se que também $v_\delta(\theta_k) \to 0$ e $v_\delta(\theta_j) \to 0$, quando $\theta_k \to \infty$ e $\theta_j \to -\infty$. No entanto, como $v_\delta(\theta)$ é monótona, deve ser $v_\delta(\theta) \equiv 0$, isto é, $v_\delta(\theta) \equiv 0 \equiv v_{\overline{X}}(\theta)$ e, pelo Teorema 6.12, $\overline{X}$ é estimador admissível de θ quando a função perca é quadrática [Hodges e Lehmann (1951); veja-se Lehmann (1983) para uma demonstração diferente]. □

Ao estudar em que termos, $T(X)$, estimador não enviesado de $\tau(\theta)$, $E_\theta\{T(X)\} = \tau(\theta)$, tem variância igual ao segundo membro da desigualdade de FCR, demonstra-se [ver, por exemplo, Zacks (1971)] que a condição necessária e suficiente é que X tenha distribuição da família exponencial,

$$f(x\,|\,\theta) = C(\theta) \exp\{\theta T(x)\} H(x).$$

Face ao Teorema 6.12 a questão que naturalmente se põe é a seguinte: se o estimador não enviesado de $\tau(\theta)$, seja $T(X)$, está nas condições acima, fica a respectiva admissibilidade garantida?

Para dar uma resposta considere-se primeiramente o importante,

Lema 6.4 — Seja X uma variável aleatória com média θ e variância σ^2. Com função perca quadrática, $\alpha X + \beta$ é estimador inadmissível de θ quando:

(i) $\alpha > 1$, ou
(ii) $\alpha < 0$, ou
(iii) $\alpha = 1$, $\beta \neq 0$.

Dem. Designe-se, $R[\theta, (\alpha, \beta)]$ o risco de $\alpha X + \beta$; tem-se,

$$\begin{aligned} R[\theta, (\alpha, \beta)] &= E\{(\alpha X + \beta - \theta)^2\} \\ &= \alpha^2\sigma^2 + [(\alpha - 1)\theta + \beta]^2. \end{aligned}$$

(i) Se $\alpha > 1$, então,

$$R[\theta, (\alpha, \beta)] \geqq \alpha^2\sigma^2 > \sigma^2 = R[\theta, (1,0)],$$

e $\alpha X + \beta$ é dominado por X.

(ii) Se $\alpha < 0$, então, $(\alpha - 1)^2 > 1$ e, consequentemente,

$$\begin{aligned} R[\theta, (\alpha, \beta)] &\geqq [(\alpha - 1)\theta + \beta]^2 = (\alpha - 1)^2 \left[\theta + \frac{\beta}{\alpha - 1}\right]^2 \\ &> \left[\theta + \frac{\beta}{\alpha - 1}\right]^2 = R[\theta, (0, -\beta/(\alpha - 1))], \end{aligned}$$

desigualdade que mostra ser $\alpha X + \beta$ dominado pelo estimador constante, $\delta \equiv \beta/(\alpha - 1)$.

(iii) Neste caso, $\alpha X + \beta = X + \beta$, estimador que é dominado por X, porquanto,

$$R[\theta, (1, \beta)] = R[\theta, (1,0)] + \beta^2.$$

□□

Suponha-se que X tem função de densidade,

$$f(x \mid \theta) = C(\theta) \exp\{\theta T(x)\} H(x), \tag{6.35}$$

sendo o espaço natural do parâmetro Θ formado pelo intervalo,

$$\Theta = (\theta', \theta''), \ -\infty \leqq \theta' \leqq \theta'' \leqq +\infty. \tag{6.35'}$$

Para estimar $E_\theta\{T\} = \tau(\theta)$, o lema anterior mostra que qualquer estimador admissível tem de ser da forma,

$$\delta(T) = [1/(1+\lambda)]T + [\gamma\lambda/(1+\lambda)], \tag{6.36}$$

com $0 \leqq \lambda < +\infty$ [caso correspondente a $0 < \alpha \leqq 1$, no Lema 6.4].

Pode agora apresentar-se a condição suficiente de admissibilidade devida a Karlin [vejam-se Zacks (1981) e Lehmann (1983)]:

Teorema 6.13 — Nas condições (6.35), (6.35′), uma condição suficiente de admissibilidade para o estimador (6.36) de $\tau(\theta) = E_\theta\{T\}$, com função perca quadrática, é para algum (e portanto para todos) $\theta' < \theta_0 < \theta''$, os dois integrais,

$$\int_{\theta_0}^{\theta} e^{-\gamma\lambda\theta}[C(\theta)]^{-\lambda}d\theta \quad \text{e} \quad \int_{\theta}^{\theta_0} e^{-\gamma\lambda\theta}[C(\theta)]^{-\lambda}d\theta,$$

tenderem para infinito quando θ tender para θ'' e θ' respectivamente.

Dem. Veja-se Lehmann (1983). □□

Como consequência imediata tem-se,

Teorema 6.14 — Se com (6.35) o espaço natural do parâmetro é toda a recta real, $\theta = \mathbf{R}$, então T é estimador admissível de $\tau(\theta) = E_\theta\{T\}$, com função perca quadrática.

Dem. Com $\lambda = 0$ e $\theta = 1$, os dois integrais do teorema anterior tendem para infinito quando $\theta \to \pm\infty$. □□

Estão nas condições deste teorema a Normal com variância conhecida (o Ex. 6.20 corresponde a este caso)[18], a Binomial e a Poisson; o mesmo não se passa com a Gama e a Binomial negativa.

6.6 Estimação da média da Normal *k*-dimensional[19]

A presente secção vai dedicar-se essencialmente ao estudo de um caso notável de inadmissibilidade com função perca quadrática, no contexto do qual se introduzem os estimadores de James-Stein já referidos de passagem (Ex. 4.6). Este estudo tem ainda a particularidade de mostrar novos aspectos do confronto entre clássicos e bayesianos.

[18] Com $T = \overline{X}$ e $\tau(\theta) = \theta$.

[19] A leitura da secção não é indispensável ao prosseguimento do estudo.

Considere-se a observação de uma variável aleatória k-dimensional, $\mathbf{X} = (X_1, X_2, \dots, X_k)$, com distribuição Normal, $N_k(\boldsymbol{\theta}, I_k)$, onde, $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, é o vector dos valores esperados, $\theta_i = E\{X_i\}$, $i = 1, 2, \dots, k$, e I_k a matriz identidade de ordem k. Quer dizer, as variáveis X_i são independentes, $X_i \sim N(\theta_i, 1)$, $i = 1, 2, \dots, k$.

Na prática cada X_i pode designar a média de N observações independentes, $X_{ij} \sim N(\theta_i, \sigma^2)$, $j = 1, 2, \dots, N$. Então, $X_i \sim N(\theta_i, \sigma^2/N)$ e supondo σ^2 conhecido, hipótese inicialmente admitida[20], uma mudança de escala permite substituir σ^2/N por 1 e obter a formulação canónica do parágrafo anterior. A simplificação foi já várias vezes adaptada ao proceder à estimação através de ‹uma› observação de uma estatística suficiente; presentemente, trata-se de trabalhar com ‹uma› observação de uma estatística k-dimensional suficiente para $\boldsymbol{\theta}$, em que se operou eventualmente uma mudança de escala.

O problema consiste em estimar $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \Theta = \mathbf{R}^k$, portanto, $\mathbf{a} \in A \equiv \mathbf{R}^k$, com função perca quadrática,

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{a}) &= \sum_{i=1}^{k} (\theta_i - a_i)^2 \\ &= (\boldsymbol{\theta} - \mathbf{a})'(\boldsymbol{\theta} - \mathbf{a}) \\ &= \| \boldsymbol{\theta} - \mathbf{a} \|^2; \end{aligned} \tag{6.37}$$

todo o estudo é pois condicionado por esta forma da função perca.

O estimador «clássico» (para todos os efeitos o vector das médias),

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X}) = (X_1, X_2, \dots, X_k) = \mathbf{X},$$

tem propriedades conhecidas: máxima verosimilhança, linear, não enviesado e melhor invariante em relação ao grupo aditivo em $\mathbf{R}^k$. A teoria em que se apoia vem de Gauss, embora, no que toca ao não enviesamento não haja justificação indiscutível.

Se $k = 1$, o Teorema 6.8 permite concluir que $\hat{\theta}$ é minimax; adicionalmente, o Teorema 6.12 [veja-se Ex. 6.20] permite demonstrar que $\hat{\theta}$ é admissível.

Com k inteiro positivo qualquer continua a demonstrar-se que $\hat{\boldsymbol{\theta}}$ é minimax: está-se em presença de uma função de decisão igualadora,

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{k} (\theta_i - X_i)^2 \right\} = E_{\boldsymbol{\theta}} \left\{ \| \boldsymbol{\theta} - \mathbf{X} \|^2 \right\} = k, \tag{6.38}$$

a que pode aplicar-se a doutrina do Teorema 4.26 [veja-se o Ex. 4.16 e alargue-se a k dimensões tomando, $h_\tau(\boldsymbol{\theta}) \equiv N_k(\mathbf{0}, \tau^2 I_k)$].

[20] Os casos, $\mathbf{X} \sim N_k(\boldsymbol{\theta}, \sigma^2 I_k)$ e $\mathbf{X} \sim N_k(\boldsymbol{\theta}, \Sigma)$ são adiante referidos.

Com a admissibilidade o panorama é diferente. Durante muitos anos pensou-se que $\hat{\boldsymbol{\theta}}$ era óptimo em todos os sentidos e, portanto, também admissível; no entanto, embora Stein (1956) tenha demonstrado que para $k = 2$, $\hat{\boldsymbol{\theta}}$ é ainda admissível, James e Stein (1960) mostraram que para $k \geqq 3$, $\hat{\boldsymbol{\theta}}$ deixa de ser admissível! Esta conclusão provocou perplexidade e controvérsia que não se pode considerar encerrada[21]. As reacções desfavoráveis são compreensíveis por não ser fácil aceitar a «dessacralização» da média aritmética[22] ainda por cima no contexto da distribuição Normal concebida por Gauss exactamente pelas propriedades que lhe confere.

James e Stein demonstraram que, para $k \geqq 3$, o estimador clássico, $\hat{\boldsymbol{\theta}}$, é estritamente dominado pelo estimador que hoje tem o seu nome,

$$\boldsymbol{\delta}^{JS}(\mathbf{X}) = \left(1 - \frac{k-2}{\|\mathbf{X}\|^2}\right)\mathbf{X}, \tag{6.39}$$

ou, componente a componente,

$$\delta_i^{JS}(\mathbf{X}) = \left(1 - \frac{k-2}{\|\mathbf{X}\|^2}\right)X_i. \tag{6.39′}$$

Para futura referência convém apresentar o estimador de James-Stein na expressão mais geral,

$$\boldsymbol{\delta}(\mathbf{X}) = \boldsymbol{\mu} + \left(1 - \frac{k-2}{\|\mathbf{X}-\boldsymbol{\mu}\|^2}\right)(\mathbf{X}-\boldsymbol{\mu}), \tag{6.40}$$

ou, componente a componente,

$$\delta_i(\mathbf{X}) = \mu_i + \left(1 - \frac{k-2}{\|\mathbf{X}-\boldsymbol{\mu}\|^2}\right)(X_i-\mu_i), \tag{6.40′}$$

A génese dos estimadores de James-Stein foi estudada por vários autores [citam-se, por exemplo, Lindley (1962) e Zellner e Vandaele (1975)]. Faz-se aqui apenas uma tentativa para descrever uma possível motivação do seu emprego (que deve ser complementada com o estudo da operação de «*shrinking*» adiante referida) antes da apresentação um tanto mais rigorosa das suas propriedades.

A motivação para a estrutura geral do estimador (6.40) [ou (6.39)] — não para a sua forma funcional específica — pode resultar das seguintes considerações: suponha-se que a priori se atribui uma elevada probabilidade às relações $\theta_i = \mu_i$,

[21] Repare-se na advertência feita há tempos por Efron (1975): «*An equivalent of the Surgeon-General's warning may be in order: these methods are not perfectely understood yet, and are still the subject of heated controversy among statisticians*».

[22] Insista-se, mais uma vez, que em $\hat{\boldsymbol{\theta}} = (X_1, X_2, \ldots, X_k)$, a i-ésima componente, X_i, pode estar no lugar da média das i-ésimas componentes de N vectores k-dimensionais independentes (amostra casual).

$i = 1, 2, \ldots, k$ [ou $\theta_i = 0, i = 1, 2, \ldots, k$]. Então parece coerente começar por ensaiar a hipótese,

$$H_0\colon \theta_1 = \mu_1, \ldots, \theta_k = \mu_k;$$

se H_0 for aceite, pode propor-se $\boldsymbol{\mu}$ como estimador de $\hat{\boldsymbol{\theta}}$; se H_0 não for aceite, pode propor-se $\mathbf{X}$ como estimador de $\hat{\boldsymbol{\theta}}$. A melhor região de aceitação para H_0 é da forma, $\Sigma(X_i - \mu_i)^2 = \| \mathbf{X} - \boldsymbol{\mu} \|^2 \leqq S$; assim, em síntese, o estimador assume o aspecto,

$$\boldsymbol{\delta}(\mathbf{X}) = \begin{cases} \boldsymbol{\mu} & \text{se } \| \mathbf{X} - \boldsymbol{\mu} \|^2 \leqq S, \\ \mathbf{X} & \text{se } \| \mathbf{X} - \boldsymbol{\mu} \|^2 > S. \end{cases}$$

Um procedimento mais «suave» consiste em avançar com um estimador definido pela expressão,

$$\boldsymbol{\delta}(\mathbf{X}) = W(\| \mathbf{X} - \boldsymbol{\mu} \|)\mathbf{X} + [1 - W(\| X - \boldsymbol{\mu} \|)]\boldsymbol{\mu},$$

onde a função de ponderação, W, em vez de tomar apenas os valores 0 e 1 como no caso anterior, é uma função contínua e monótona crescente tal que, $W(0) = 0$, e $W(\| \mathbf{X} - \boldsymbol{\mu} \|) = 1$ quando $\| \mathbf{X} - \boldsymbol{\mu} \| \to \infty$.

Voltando ao estimador de James-Stein na forma mais corrente[23] — (6.39) — mostra-se adiante [Teorema 6.16] que a respectiva função risco assume a forma,

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS}) = k - (k-2)^2 E\{1/\| \mathbf{X} \|^2\}. \tag{6.41}$$

Comparando com (6.38) tem-se, com $k \geqq 3$,

$$R(\boldsymbol{\delta}, \boldsymbol{\delta}^{JS}) < R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \text{ para todo o } \boldsymbol{\theta}, \tag{6.42}$$

relação que estabelece a inadmissibilidade de $\hat{\boldsymbol{\theta}}$.

Na Fig. 6.1 comparam-se as funções risco, $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS})$ e $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, construidas tomando $\boldsymbol{\theta}'\boldsymbol{\theta}$ como argumento.

Quando $\boldsymbol{\theta}'\boldsymbol{\theta} \to 0$ [quando $\boldsymbol{\theta} \to \mathbf{0} \equiv$ origem], a função risco de $\boldsymbol{\delta}^{JS}$ tende para 2; quando $\boldsymbol{\theta}'\boldsymbol{\theta} \to \infty$, a mesma função tende para k, isto é, tende a confundir-se com $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$. A vantagem de $\boldsymbol{\delta}^{JS}$ sobre $\hat{\boldsymbol{\theta}}$ tem expressão mais significativa quando $\boldsymbol{\theta}$ está próximo da origem.

O estimador, $\boldsymbol{\delta}^{JS}$, é não linear e enviesado, pois calcula-se,

$$E_{\boldsymbol{\theta}}\{\boldsymbol{\theta}^{JS}(\mathbf{X})\} = \boldsymbol{\theta} - (k-2)E\{1/\| \mathbf{X} \|^2\}\boldsymbol{\theta};$$

é ainda minimax: devido a (6.42) e ao facto de $\boldsymbol{\theta}$ ser minimax. Como se deduz da respectiva fórmula é também não invariante, facto que para alguns autores [por

[23] As considerações sobre (6.39) aplicam-se inteiramente a (6.40) desde que no espaço do parâmetro $\boldsymbol{\theta}$ se desloque a origem do ponto $\mathbf{0}$ para o ponto $\boldsymbol{\mu}$.

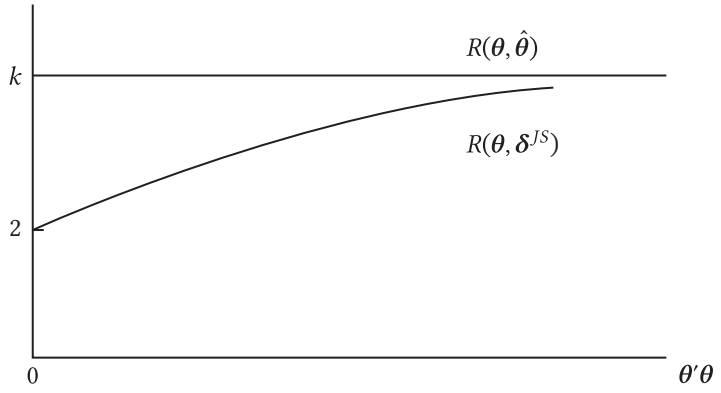


Fig. 6.1

exemplo, J. Tiago de Oliveira, comunicação verbal] é inconveniente inultrapassável. E, curiosamente, é inadmissível [veja-se Teorema 6.16].

O resultado (6.41) pode obter-se com um pouco mais de generalidade [veja-se Lehmann (1983)]:

Teorema 6.15 — Se X_i, $i = 1, 2, \ldots, k$, são I.I.D., $X_i \sim N(\theta_i, 1)$, o estimador de $\boldsymbol{\theta}$ dado por,

$$\boldsymbol{\delta}_c(\mathbf{X}) = \left(1 - c\frac{k-2}{\|\mathbf{X}\|^2}\right)\mathbf{X}, \tag{6.43}$$

tem função risco,

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}_c) = k - (k-2)^2 E\{[(2c - c^2)/\|\mathbf{X}\|^2]\}, \tag{6.44}$$

quando a função perca é quadrática.

Dem. Designe $\Psi(\boldsymbol{\theta})$ o primeiro membro de (6.44),

$$\Psi(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\|\boldsymbol{\theta} - \boldsymbol{\delta}_c\|^2\},$$

e $\Phi(\boldsymbol{\theta})$ o segundo membro. A demonstração de que $\Psi(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})$ para todo o $\boldsymbol{\theta}$ é feita em três passos:

[1] Tanto $\Psi(\boldsymbol{\theta})$ como $\Phi(\boldsymbol{\theta})$ dependem de $\boldsymbol{\theta}$ apenas através de $\boldsymbol{\theta}'\boldsymbol{\theta} = \Sigma\theta_i^2$.

[2] A média de $\Psi(\boldsymbol{\theta})$ e $\Phi(\boldsymbol{\theta})$, respectivamente, ponderada pela função,

$$h(\boldsymbol{\theta}) = (\sqrt{2\pi A})^{-k} \exp\{-\Sigma\theta_i^2/2A\}, \ A > 0, \tag{I}$$

é dada para todo o $B = 1/(A+1)$, por

$$\int \Psi(\boldsymbol{\theta})h(\boldsymbol{\theta})\, d\boldsymbol{\theta} = k - wB = \int \Phi(\boldsymbol{\theta})h(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{II}$$

com $w = (k-2)(2c - c^2)$. Essas médias podem interpretar-se como valores esperados de $\Psi(\boldsymbol{\theta})$ e $\Phi(\boldsymbol{\theta})$ quando $\boldsymbol{\theta}$ tem função de densidade [I].

[3] Considerando o passo [I] pode escrever-se,

$$\Psi(\boldsymbol{\theta}) = \Psi^*(\Sigma\theta_i^2), \quad \Phi(\boldsymbol{\theta}) = \Phi^*(\Sigma\theta_i^2).$$

Para a família de distribuições [I], $\Sigma\theta_i^2$ é uma estatística suficiente completa; portanto[24], devido a [II], $\Psi^*(\Sigma\theta_i^2) = \Phi^*(\Sigma\theta_i^2)$ quase por toda a parte e como estas funções são continuas [veja-se Teorema 7.16] tem-se $\Psi(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})$ para todo o $\boldsymbol{\theta}$ como pretende mostrar-se.

Voltando aos passos [1] e [2], afinal o que falta demonstrar, e começando pelo segundo, vem,

$$E\{\Phi(\boldsymbol{\theta})\} = k - (k-2)E\{E\{(w/\|\mathbf{X}\|^2) \,|\, \boldsymbol{\theta}\}\}, \qquad \text{[III]}$$

retomando a interpretação de $h(\boldsymbol{\theta})$ como densidade de $\boldsymbol{\theta}$. No segundo membro desta expressão, o valor esperado $E\{\cdot \,|\, \boldsymbol{\theta}\}$ é tomado em relação à distribuição dos X_i, $N(\theta_i, 1)$, dado $\boldsymbol{\theta}$, e o valor esperado exterior é tomado em relação a $h(\boldsymbol{\theta})$. Como, $w/\|\mathbf{X}\|^2$, não envolve $\boldsymbol{\theta}$, uma conhecida propriedade dos valores esperados condicionados permite escrever,

$$E\{E\{(w/\|\mathbf{X}\|^2) \,|\, \boldsymbol{\theta}\}\} = E\{w/\|\mathbf{X}\|^2\},$$

onde o valor esperado do segundo membro é agora tomado em relação à distribuição marginal dos X_i, que facilmente se verifica ser $N(0, 1/B)$. Como neste caso $B\|\mathbf{X}\|^2 \sim \chi_k^2$ e se sabe que $E\{1/\chi_\nu^2\} = 1/(\nu - 2)$, tem-se $E\{1/\|\mathbf{X}\|^2\} = B/(k-2)$, e a segunda parte de [II] fica demonstrada.

Para provar a primeira parte caminha-se em sentido inverso, considerando primeiro o valor esperado condicionado,

$$\Sigma E\{[\theta_i - (1 - \hat{B})\, X_i]^2 \,|\, \mathbf{X}\}, \qquad \text{[IV]}$$

onde,

$$\hat{B} = c(k-2)/\|\mathbf{X}\|^2. \qquad \text{[V]}$$

De [IV] obtém-se, para cada parcela,

$$V\{\theta_i \,|\, \mathbf{X}\} + [E\{\theta_i \,|\, \mathbf{X}\} - (1 - \hat{B})\, X_i]^2 = (1 - B) + (\hat{B} - B)^2\, X_i^2,$$

e, consequentemente, para a soma,

$$(1 - B)\, k + (\hat{B} - B)^2 \|\mathbf{X}\|^2.$$

[24] Recorde-se o conceito de estatística suficiente completa.

Desenvolvendo $(\hat{B} - B)^2$ com $\hat{B}$ dado por [V] e apelando para as relações,

$$E\{\| \mathbf{X} \|^2\} = k/B \quad \text{e} \quad E\{1/\| \mathbf{X} \|^2\} = B/(k-2),$$

obtém-se a primeira parte de [II].

Finalmente para provar o passo [1] tem-se quanto a $\Psi(\boldsymbol{\theta})$ o facto de ser o valor esperado de uma função de $\boldsymbol{\theta}$ e de $\mathbf{X}$ que depende apenas de ΣX_i^2, $\Sigma \theta_i X_i$ e $\Sigma \theta_i^2$, Aplicando uma transformação ortogonal, $(X_1, X_2, \dots, X_k) \to (Y_1, Y_2, \dots, Y_k)$, tal que $Y_1 = \Sigma \theta_i X_i / \sqrt{(\Sigma \theta_i^2)}$, os Y_i são ainda independentes com variância igual a um e com valores esperados, $E\{Y_1\} = \sqrt{(\Sigma \theta_i^2)}$ e $E\{Y_j\} = 0$, $j > 1$. Em termos dos Y_i, $\Psi(\boldsymbol{\theta})$ é o valor esperado de uma função de Y_i^2, $\Sigma_2^k Y_j^2$ e de $\Sigma \theta_i^2$ e o resultado segue-se. O mesmo se passa com $\Phi(\boldsymbol{\theta})$ que é o valor esperado de uma função de ΣX_i^2. □□

As proposições seguintes saiem imediatamente do teorema anterior,

Corolário 6.1 — O estimador $\boldsymbol{\delta}_c$ definido por (6.43) domina estritamente o estimador clássico, $\hat{\boldsymbol{\theta}}$ ($\boldsymbol{\delta}_c$ para $c = 0$) sempre que, $0 < c < 2$ e $k \geqq 3$.

Dem. Quando $0 < c < 2$, $2c - c^2 > 0$ e, consequentemente, $R(\boldsymbol{\theta}, \boldsymbol{\delta}_c) < k$ para todo o $\boldsymbol{\theta}$. □□

Corolário 6.2 — O estimador de James-Stein, $\boldsymbol{\delta}^{JS}$, que é igual a $\boldsymbol{\delta}_c$ quando $c = 1$ domina estritamente todos os estimadores $\boldsymbol{\delta}_c$ com $c \neq 1$.

Dem. O factor, $2c - c^2$ assume o valor máximo 1 se e somente se $c = 1$. □□

Considerando o último corolário pode pensar-se que o estimador de James--Stein é admissível. Que tal propriedade se não verifica pode depreender-se do seguinte,

Teorema 6.16 — Seja $\boldsymbol{\delta}_B$ um estimador da forma, $\boldsymbol{\delta}_B(\mathbf{X}) = (1 - \hat{B})\mathbf{X}$, onde $\hat{B}$ é uma função monótona decrescente dos X_i, $i = 1, 2, \dots, k$, e suponha-se que,

$$P_\theta(\hat{B} > 1) > 0.$$

Então, $R(\boldsymbol{\theta}, \boldsymbol{\delta}') < R(\boldsymbol{\theta}, \boldsymbol{\delta}_B)$, para todo o $\boldsymbol{\theta}$, onde,

$$\boldsymbol{\delta}'(\mathbf{X}) = (1 - \hat{B})_+ \mathbf{X},$$

$(w)_+ = \max\{0, w\}$.

Dem. Pode estudar-se em Lehmann (1983). □□

Como imediatamente se reconhece, o estimador $\boldsymbol{\delta}_1^{JS}$ definido adiante por (6.45) domina estritamente $\boldsymbol{\delta}^{JS}$. Infelizmente consegue provar-se que também $\boldsymbol{\delta}_1^{JS}$ é não admissível (porque não é Bayes nem Bayes generalizado) embora, como diz Lehmann (1983), não seja possível introduzir estimadores que apresentem nítida melhoria em relação a $\boldsymbol{\delta}_1^{JS}$.

Generalizando de certo modo o Corolário 6.1, Baranchik [veja-se Strawderman (1971)] demonstrou,

Teorema 6.17 — Um estimador de $\boldsymbol{\delta}$ da forma,

$$\boldsymbol{\delta}(\mathbf{X}) = \left[1 - c(\|\mathbf{X}\|)\frac{k-2}{\|\mathbf{X}\|^2}\right]\mathbf{X},$$

é minimax para $k \geqq 3$ quando: (i) $0 < c(\|\mathbf{X}\|) < 2$ e (ii) a função c é não decrescente. □□

É curioso comparar o caso unidimensional ($k = 1$) em que os estimadores minimax são quase sempre únicos, com o caso multidimensional ($k \geqq 3$) em que, como o Teorema 6.17 esclarece, há uma enorme riqueza de estimadores minimax.

O Teorema 6.16 aponta que os estimadores da família Baranchick são inadmissíveis quando, $c(\|\mathbf{X}\|)/\|\mathbf{X}\|^2 > 1/(k-2)$ com probabilidade positiva; no entanto, a mesma família contém estimadores admissíveis como o de Strawderman (1971)[25].

A admissibilidade é sempre definida com referência a uma função perca, no caso em apreço a definida em (6.35). Que se passa quando se toma outra função perca, por exemplo,

$$L^*(\boldsymbol{\theta}, \boldsymbol{\delta}) = \Sigma\,|\,\theta_i - \delta_i\,|\,?$$

Embora o tratamento matemático não seja tão claro [Efron (1975)] parece que as conclusões se mantêm nos termos seguintes: $\boldsymbol{\delta}^{JS}$ ou alguma das suas variantes

[25] Strawderman obteve um estimador admissível da família Baranchick (portanto, também minimax). Para o efeito, partindo de uma distribuição a priori própria,

$$h(\boldsymbol{\theta}) = \frac{(\|\boldsymbol{\theta}\|^2)^{(k-2)/2}(1-a)}{(2\pi)^{k/2}} \int_0^\infty \frac{\xi^{(k/2)-a} e^{-\xi/2}}{(\|\boldsymbol{\theta}\|^2 + \xi)^{2-a}}\, d\xi,$$

obteve um estimador Bayes, tipo James-Stein,

$$\boldsymbol{\delta}^S(\mathbf{X}) = \left\{1 - [(k+2-2a)/\|\mathbf{X}\|^2]\cdot \left[1 - 2(k+2-2a)^{-1}\left(\int_0^1 \lambda^{(k-2a)/2} \exp\{-(\lambda-1)\|\mathbf{X}\|^2/2\}\, d\lambda\right)^{-1}\right]\right\}\mathbf{X},$$

onde, $1/2 \leqq a < 1$ para $k = 5$ e, $0 \leqq a \leqq 1$ para $k \geqq 6$.
Lin (1974) mostrou que para $k \geqq 3$, sendo, $3 - (k/2) \leqq a \leqq 2$, $\boldsymbol{\delta}^S$ é ainda admissível embora deixe de ser Bayes (contra uma distribuição a priori própria) para ser Bayes generalizado (contra uma distribuição a priori imprópria).

domina estritamente $\hat{\boldsymbol{\theta}}$ enquanto a função perca for soma dos «erros» cometidos em relação a cada uma das componentes de $\boldsymbol{\theta}$[26].

Como os X_i, $i = 1, 2, \ldots, k$, são independentes, parece estranho que o estimador James-Stein [ver (6.38)] faça intervir na estimação de θ_i, valor esperado de X_i, as observações de X_j, $j \neq i$. A tese de que pode melhorar-se o estimador «clássico» em problemas independentes através da estimação combinada é conhecida por efeito Stein.

Não há dúvida tratar-se de um dos aspectos mais controversos do problema que já mereceu, aliás, o seguinte desabafo: «*Do you mean that if I want to estimate tea consumption in Taiwan I will do better to estimate simultaneously the speed of light and the weight of hogs in Montana*» [citado por Efron e Morris (1973)].

Uma possível contra argumentação [Efron e Morris (1973)] desenvolve-se nas seguintes linhas: porquê dizer que não há qualquer relação entre os θ_i nas situações em que existe uma relação de facto que é a de serem todos pequenos [$\boldsymbol{\theta}$ próximo da origem]? E é nessa situação que na verdade o estimador James-Stein se revela vantajoso!

Aliás a origem não tem nada de mítico pois considerando qualquer outro ponto de $\mathbf{R}^k$ seja $\boldsymbol{\mu}$, pode obter-se o estimador tipo James-Stein, já referido em (6.40) que é ainda uniformemente melhor do que $\hat{\boldsymbol{\theta}}$ e se revela particularmente vantajoso quando $\boldsymbol{\theta}$ está próximo de $\boldsymbol{\mu}$.

O choque entre as duas posições parece constituir forte motivação para dar entrada ao conceito de distribuição a priori. A discussão de Hill (1975) lança luz sobre a questão e põe em confronto a posição dos estatísticos «clássicos» e dos bayesianos.

Um bayesiano que considera os θ_i, $i = 1, 2, \ldots, k$, independentes a priori e trabalha com função perca quadrática, nunca será conduzido ao emprego de estimadores James-Stein. A perspectiva bayesiana coincide assim com o bom senso e recusa combinar dados respeitantes a problemas entre os quais não há qualquer relação.

Para um não bayesiano não faz sentido falar na independência a priori dos θ_i, $i = 1, 2, \ldots, k$ (os θ_i não são considerados aleatórios). Assim, como o estimador James-Stein é melhor do que o estimador «clássico» segundo um critério frequencista (valor esperado do erro quadrático calculado sobre todo o espaço das amostras), o não bayesiano parece ser conduzido ao extremo de combinar observações não relacionadas para reduzir o risco ou perca esperada. A forma que tem para evitar tal absurdo é limitar a classe de funções de decisão ou estimadores que aplica. Partindo nomeadamente da ideia de que os problemas são aparentemente

[26] Veja-se, contudo, Lehmann (1983).

não relacionados, pode restringir a atenção à classe de estimadores,

$$\boldsymbol{\delta}(\mathbf{X}) = [\delta_1(X_1), \delta_2(X_2), \ldots, \delta_k(X_k)],$$

em que θ_i é estimado exclusivamente a partir de X_i. E como Hill mostra, o estimador «clássico» é admissível dentro desta classe em relação à função perca quadrática.

A atitude de Box (1983) é, dentro da corrente bayesiana, concordante com a de Hill. Imaginando uma experiência em que está em causa a aplicação do estimador James-Stein, diz: «*Such an experiment could make sense when it is conducted to compare, for example, the levels of infestation of k different varieties of wheat, or the numbers of eggs laid by k different breeds of chickens or the yields of k sucessive batches of chemical; in general, that is, when a priori we expect similarities of one kind or another between the entities compared. But clearly, if similarities are in mind, they ought not to be denied by the form of the model. …The presence of a specific form of priori distribution allows the investigator to incorporate in the model precisely the* kind *of similarities he wishes to entertain*».

Box defende, portanto, que somente a explicitação de uma distribuição a priori, descrevendo adequadamente o problema em estudo, pode orientar na escolha do estimador a empregar, evitando o ridículo associado à ideia de que pode melhorar-se o erro quadrático médio englobando num só grupo variedades de trigo, raças de galinhas ou lotes de um produto químico[27].

Tanto (6.37), como (6.44), são exemplos de estimadores redutores — «*shrinking estimators*»; o efeito da redução ou «*shrinking*» é empurrar o estimador «clássico», $\mathbf{X}$, para a origem — (6.37) — ou para um dado ponto $\boldsymbol{\mu}$, [(6.44)].

Considere-se o «*shrinking*» para origem que é representativo do fenómeno do «*shrinking*» em geral. Semelhante redução é logicamente sensata quando há razões para pensar que $\boldsymbol{\theta}$ não está muito afastado da mesma origem. Mais uma vez se torna evidente a necessidade de introduzir uma distribuição a priori para $\boldsymbol{\theta}$. Se a vantagem dos estimadores tipo James-Stein só é na prática relevante numa região relativamente pequena do espaço do parâmetro $\boldsymbol{\theta}$ — fora desta região o ganho em termos de risco é modesto e pode não justificar o maior trabalho de cálculo — então é natural procurar dentro desse tipo os estimadores que melhor se comportam na região a priori mais credível para $\boldsymbol{\theta}$. Faz então igualmente sentido proceder ao «*shrinking*» para essa região ou para um ponto dessa região. Por outras palavras, qualquer ideia de que o estimador processa magicamente o «*shrinking*» para a região correcta é completamente errada.

Parece claro que a informação a priori é indispensável para orientar o processo de redução, seja para a origem ou para um ponto $\boldsymbol{\mu}$.

[27] A génese dos estimadores James e Stein [já se referiram Lindley (1962) e Zellner e Vandaele (1975)] contribui também para aclarar este problema.

Se, $\|\mathbf{X}\|^2 < k-2$, a expressão (6.37) mostra que $\boldsymbol{\delta}^{JS}$ leva a redução longe demais, isto é, opera a redução para além da origem. Adaptando a chamada «regra-mais» pode propor-se o estimador,

$$\boldsymbol{\delta}_1^{JS}(\mathbf{X}) = \left[1 - \frac{k-2}{\|\mathbf{X}\|^2}\right]_+ \mathbf{X}, \tag{6.45}$$

onde $[w]_+ = \max\{0, w\}$. Pelo Teorema 6.16, $\boldsymbol{\delta}_1^{JS}$ domina estritamente $\boldsymbol{\delta}^{JS}$; quer dizer, a «regra-mais» permite diminuição do risco para todo o $\boldsymbol{\theta}$ (finito).

Na hipótese mais geral, $\mathbf{X} \sim N_k(\boldsymbol{\theta}, \sigma^2 I_k)$, com σ^2 desconhecido, há que estimar σ^2, tarefa que, evidentemente, não pode pedir-se ao vector (das médias) $\hat{\boldsymbol{\theta}}$. Então, se for S uma estatística obtida através de uma experiência independente daquela que permite observar $\hat{\boldsymbol{\theta}}$ e tal que, $S \sim \sigma^2 \chi^2_N$, então,

$$\boldsymbol{\delta}_2^{JS}(\mathbf{X}) = \left[1 - \frac{k-2}{N+2}\frac{S}{\|\mathbf{X}\|^2}\right] \mathbf{X}, \tag{6.46}$$

domina estritamente $\hat{\boldsymbol{\theta}}$, sempre, é claro, para $k \geqq 3$.

Enfim, no caso ainda mais geral, $\mathbf{X} \sim N_k(\boldsymbol{\theta}, \mathfrak{Sigma})$, com matriz de covariâncias, $\mathfrak{Sigma}$, desconhecida, James e Stein mostraram, para uma função perca da forma,

$$L[(\boldsymbol{\theta}, \mathfrak{Sigma}), \mathbf{a}] = (\boldsymbol{\theta} - \mathbf{a})' \mathfrak{Sigma}^{-1} (\boldsymbol{\theta} - \mathbf{a}), \tag{6.47}$$

que o estimador,

$$\boldsymbol{\delta}(\mathbf{X}, S^*) = \left[1 - \frac{k-2}{(n-k+3)}(\mathbf{X}' S^{*-1} \mathbf{X})^{-1}\right] \mathbf{X}, \tag{6.48}$$

onde a matriz S^* — estimador de $\mathfrak{Sigma}$ — tem distribuição de Wishart, $W_k(n, \mathfrak{Sigma})$, é uniformemente melhor do que $\hat{\boldsymbol{\theta}}$ [a distribuição de Wishart pode estudar-se em Muirhead (1982) ou em qualquer obra sobre análise multivariada].

6.7 Estimadores Bayes

O problema dos estimadores Bayes, considerado já através de alguns exemplos, retoma-se para apresentação de algumas proposições de interesse mais geral.

Considerando o conceito de risco a posteriori, $r_{\mathbf{x}}(\delta)$ — veja-se a expressão (3.26) — recorde-se que dada a distribuição a priori, $h(\theta)$, e observada a amostra casual $\mathbf{X}$, $\mathbf{X} = \mathbf{x} \in \mathfrak{X}$, a estimativa Bayes, $\delta_h \in A \equiv \Theta$, é tal que,

$$r_{\mathbf{x}}(\delta_h) = \inf_\delta r_{\mathbf{x}}(\delta), \quad \mathbf{x} \text{ fixo.}$$

Concebendo que a análise se realiza ou repete para todo o $\mathbf{x} \in \mathcal{X}$ e que δ_h existe para todo o $\mathbf{x} \in \mathcal{X}$, a correspondência $\mathbf{x} \to \delta_h$ define a função de decisão Bayes contra h, $\delta_h(\mathbf{x})$, ou, como se queira, o estimador Bayes contra h, $\delta_h(\mathbf{X})$.

A possibilidade de obter estimadores Bayes depende das condições em que, para cada $\mathbf{x}$, existe $a \in A$ tal que $r_{\mathbf{x}}(a) < \infty$. Para investigar essas condições consideram-se dois tipos de função perca:

1) Quadrática: $L(\theta, a) = \lambda(\theta)(\theta - a)^2,\ \lambda(\theta) > 0$,

2) Erro absoluto: $L(\theta, a) = \lambda(\theta)|\,\theta - a\,|,\ \lambda(\theta) > 0$.

Em qualquer destes casos a função perca, para cada $\theta \in \Theta$, é função convexa de a. Além disso, ambas as funções verificam a condição adicional referida no Teorema 4.4 (veja-se também o Teorema 4.5), facto que permite limitar o estudo às funções de decisão puras.

Para aliviar a escrita x representa um escalar ou vector; seguindo a prática corrente, a notação vectorial, $\mathbf{x}$, só é empregue quando indispensável.

1) *Função Perca Quadrática.* Seja,

$$r_x(a) = E_h\{L(\theta, a) \,|\, x\} = \int_\Theta L(\theta, a) h(\theta \,|\, x)\, d\theta,$$

onde convém especificar o subíndice h para indicar que $E_h\{\cdot\}$ se toma em relação a $h(\theta)$ e $E_h\{\cdot \,|\, x\}$ em relação a $h(\theta \,|\, x)$. Entrando com a expressão de $L(\theta, a)$, vem,

$$\begin{aligned} r_x(a) &= E_h\{\lambda(\theta)(\theta - a)^2 \,|\, x\} \\ &= a^2 E_h\{\lambda(\theta) \,|\, x\} - 2a E_h\{\lambda(\theta)\theta \,|\, x\} + E_h\{\lambda(\theta)\theta^2 \,|\, x\}. \end{aligned} \tag{6.49}$$

Se os coeficientes desta função quadrática em a são finitos, também $r_x(a)$ é finito. A situação é perfeitamente esclarecida pela proposição de Girshick e Savage (1951),

Teorema 6.18 — Com função perca quadrática, seja h a distribuição a priori e $r_x(a)$ o respectivo risco a posteriori. Para cada $x \in \mathcal{X}$, verifica-se $r_x(a) < \infty$ para nenhum $a \in A$, para exactamente um $a \in A$ ou para todo o $a \in A$. O segundo caso implica $E_h\{\lambda(\theta) \,|\, x\} = \infty$ e o terceiro $E_h\{\lambda(\theta) \,|\, x\} < \infty$.

Dem. Se $r_x(a) < \infty$ para dois elementos de A, sejam a_1 e a_2, então $r_x(a) < \infty$ para todo o $a \in A$. Com efeito, suponha-se, $a_1 < a_2$; dado que, para todo o θ, $L(\theta, a)$ é função convexa de a, tem-se, com $a \in [a_1, a_2]$,

$$\lambda(\theta)(\theta - a)^2 \leqq \frac{a_2 - a}{a_2 - a_1}\lambda(\theta)(\theta - a_1)^2 + \frac{a - a_1}{a_2 - a_1}\lambda(\theta)(\theta - a_2)^2;$$

passando aos valores esperados fica provado ser $r_x(a) < \infty$ para $a \in [a_1, a_2]$.

Por outro lado, com $a' \in [a_1, a_2]$, $a' \neq 0$ e a arbitrário, $a \neq a'$, é fácil de verificar, para qualquer θ,

$$(a'-a)^2 - 2(a'-a)\,(\theta - a) \leqq (\theta - a')^2; \tag{6.50}$$

multiplicando por $\lambda(\theta)$ e tomando valores esperados, vem,

$$\begin{aligned}(a'-a)[E_h\{\lambda(\theta)\,|\,x\}(a'-a) + 2aE_h\{\lambda(\theta)\,|\,x\} - \\ - 2E_h\{\lambda(\theta)\theta\,|\,x\}] \leqq r_x(a') < \infty.\end{aligned}$$

Fazendo $a = 0$, obtém-se,

$$a'^2 E_h\{\lambda(\theta)\,|\,x\} - 2a' E_h\{\lambda(\theta)\theta\,|\,x\} < \infty,$$

forma quadrática em a', finita, logo com coeficientes finitos,

$$0 < E_h\{\lambda(\theta)\,|\,x\} < \infty,\ E_h\{\lambda(\theta)\theta\,|\,x\} < \infty. \tag{6.51}$$

Como $r_x(a')$ é finito, pois $a' \in [a_1, a_2]$, em face das desigualdades (6.51) e da expressão (6.49), sai $E_h\{\lambda(\theta)\theta^2\,|\,x\} < \infty$. Assim, em consequência de serem finitos os coeficientes de (6.49), tem-se $r_x(a) < \infty$ para todo o $a \in A$.

Se a_0 é o único elemento de A para o qual $r_x(a)$ é finito, tome-se um $a \in A$, $a \neq a_0$, $r_x(a) = \infty$, e escreva-se,

$$\begin{aligned}r_x(a) &= E_h\{\lambda(\theta)(\theta - a_0 + a_0 - a)^2\,|\,x\} \\ &= r_x(a_0) + 2(a_0 - a)E_h\{\lambda(\theta)(\theta - a_0)\,|\,x\} + (a_0 - a)^2 E_h\{\lambda(\theta)\,|\,x\};\end{aligned}$$

como $r_x(a_0) < \infty$, vem $E_h\{\lambda(\theta)(\theta - a_0)\,|\,x\} < \infty$, e, consequentemente, tem de ser $E_h\{\lambda(\theta)\,|\,x\} = \infty$. A demonstração fica assim completa. □□

Teorema 6.19 — Com função perca quadrática, se a distribuição a priori, h, é tal que existe um $a \in A$ com $r_x(a) < \infty$, então existe um estimador Bayes contra h, essencialmente único, definido por,

$$\delta_h(x) = \begin{cases} a_0 \text{ se } r_x(a) < \infty \text{ unicamente para } a = a_0, \\ E_h\{\lambda(\theta)\theta\,|\,x\}/E_h\{\lambda(\theta)\,|\,x\} \text{ se } r_x(a) < \infty \text{ para todo o } a \in A. \end{cases} \tag{6.52}$$

Dem. A primeira parte da definição é evidente; a segunda parte resulta da minimização da forma quadrática (6.49). □□

Teorema 6.20 — Com função perca quadrática, se a distribuição a priori, h, é tal que existe um $a \in A$ com $r_x(a) < \infty$, e se existe $E_h\{\lambda(\theta)\}$, então o estimador Bayes é enviesado ou o risco Bayes é nulo.

Dem. A existência de $E_h\{\lambda(\theta)\}$ implica a existência de $E_h\{\lambda(\theta)\,|\,x\}$ quase para todo o x, pois,

$$E_h\{\lambda(\theta)\} = \int_{\mathfrak{X}} E_h\{\lambda(\theta)\,|\,x\}\, f(x)\, dx.$$

Defina-se,

$$h_\lambda(\theta) = \frac{\lambda(\theta)h(\theta)}{E_h\{\lambda(\theta)\}}; \tag{6.53}$$

como imediatamente se verifica, o estimador em relação à função perca inicial que é Bayes contra h é idêntico ao estimador Bayes em relação à função perca $L^*(\theta, a) = (\theta - a)^2$ que é Bayes contra h_λ, Dado que h e h_λ admitem a existência de uma acção com risco a posteriori finito, o estimador Bayes contra h_λ existe e é definido por,

$$\delta_\lambda(x) = E_{h_\lambda}\{\theta\,|\,x\}.$$

Suponha-se que δ_λ é não enviesado, $E_\theta\{\delta_\lambda(X)\} = \theta$ para todo o $\theta \in \Theta$. Nessa hipótese,

$$\begin{aligned} E\{\theta\delta_\lambda(X)\} &= E\{\delta_\lambda(X) \cdot E\{\theta\,|\,X\}\} = E\{[\delta_\lambda(X)]^2\} \\ &= E\{\theta \cdot E\{\delta_\lambda(X)\,|\,\theta\}\} = E\{\theta^2\}; \end{aligned}$$

devendo notar-se que o valor esperado do primeiro membro, sem índice, refere-se à distribuição conjunta de X e de θ. Tem-se, então, para o risco Bayes,

$$\begin{aligned} R(h_\lambda, \delta_\lambda) &= E\{[\theta - \delta_\lambda(X)]^2\} \\ &= E\{\theta^2\} - 2E\{\theta\delta_\lambda(X)\} + E\{[\delta_\lambda(X)]^2\} = 0, \end{aligned}$$

como era preciso demonstrar. □□

Note-se que este teorema generaliza o raciocínio feito no Ex. 4.16 para mostrar que $\overline{X}$ sendo estimador não enviesado da média da Normal não era Bayes.

Exemplo 6.21 — Suponha-se $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. com distribuição de Bernoulli e considere-se a distribuição a priori,

$$h(\theta) = [B(\alpha, \beta)]^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ 0 \leq \theta \leqq 1.$$

A distribuição a posteriori é conhecida [(1.55)],

$$h(\theta\,|\,\mathbf{x}) = [B(\alpha + \Sigma x_i, \beta + N - \Sigma x_i)]^{-1}\theta^{\alpha+\Sigma x_i-1}\ (1-\theta)^{\beta+N-\Sigma x_i-1}.$$

Se $L(\theta, a) = (\theta - a)^2$, $r_x(a) < \infty$ para todo o $a \in [0, 1]$; de (6.52) tem-se que o estimador Bayes é,

$$\delta_h(\mathbf{x}) = E_h\{\theta\,|\,\mathbf{x}\} = \frac{\alpha + \Sigma x_i}{\alpha + \beta + N},$$

sendo a última expressão obtida de (1.56). Se $L(\theta, a) = \theta(\theta - a)^2$, o risco a posteriori continua a ser finito para todo o $a \in [0, 1]$, e o estimador Bayes assume a forma,

$$\begin{aligned}\delta_h(\mathbf{x}) &= E_h\{\theta^2 \,|\, \mathbf{x}\}/E_h\{\theta \,|\, \mathbf{x}\} \\ &= \frac{\alpha + \Sigma x_i}{\alpha + \beta + N} + \frac{\beta + N - \Sigma x_i}{(a + \beta + N)(\alpha + \beta + N - 1)},\end{aligned}$$

saindo a última expressão de (1.56) e (1.57). □

Exemplo 6.21 — *Continuação.* Suponha-se $N > 3$, $\alpha = 1$ e $\beta = 1$, isto é, $h(\theta)$ distribuição Uniforme em $[0, 1]$. Tem-se,

$$h(\theta \,|\, \mathbf{x}) = [B(\Sigma x_i + 1, N - \Sigma x_i + 1)]^{-1} \theta^{\Sigma x_i} (1 - \theta)^{N - \Sigma x_i}$$

Considere-se a função perca, $L(\theta, a) = (\theta - a)^2/\theta^2(1 - \theta)^2$. Vem,

$$r_x(a) = [B(\Sigma x_i + 1, N - \Sigma x_i + 1]^{-1} \int_0^1 (\theta - a)^2 \theta^{\Sigma x_i - 2} (1 - \theta)^{N - \Sigma x_i - 2} d\theta,$$

donde se conclui que para $\Sigma x_i = 0$ ou $\Sigma x_i = 1$ [só insucessos ou um sucesso nas N provas], $r_{\mathbf{x}}(a)$ só é finito para $a = 0$; para $1 < \Sigma X_i < N - 1$, $r_{\mathbf{x}}(a)$ é finito para todo o $a \in [0, 1]$; para $\Sigma x_i = N - 1$ ou $\Sigma x_i = N$, $r_{\mathbf{x}}(a)$ só é finito para $a = 1$. O estimador Bayes é, portanto, definido por,

$$\begin{aligned}\Sigma x_i = 0 \text{ ou } 1 \qquad &\to \delta_h(\mathbf{x}) = 0, \\ 1 < \Sigma x_i < N - 1 \qquad &\to \delta_h(\mathbf{x}) = E_h\{1/\theta(1 - \theta)^2\}/E_h\{1/\theta^2(1 - \theta)^2\} \\ &\qquad\quad = B(\Sigma x_i, N - \Sigma x_i - 1)/B(\Sigma x_i - 1, N - \Sigma x_i - 1) \\ &\qquad\quad = (\Sigma x_i - 1)/(N - 2); \\ \Sigma x_i = N - 1 \text{ ou } N \to &\delta_h(\mathbf{x}) = 1.\end{aligned}$$

□

2) *Função perca erro absoluto.* Em correspondência com o Teorema 6.14 tem-se,

Teorema 6.21 — Com função perca erro absoluto, seja h a distribuição a priori e $r_x(a)$ o respectivo risco a posteriori. Para cada $x \in \mathcal{X}$, tem-se $r_x(a) < \infty$ para nenhum $a \in A$, para exactamente um $a \in A$ ou para todo o $a \in A$. O segundo caso implica $E_h\{\lambda(\theta) \,|\, x\} = \infty$ e o terceiro $E_h\{\lambda(\theta) \,|\, x\} < \infty$.

Dem. Se $r_x(a_i) < \infty$, $i = 1, 2$, então $E_h\{\lambda(\theta) \,|\, x\} < \infty$; com efeito, supondo $a_1 < a_2$, da relação,

$$\lambda(\theta)(a_2 - a_1) \leqq \lambda(\theta)[|\, a_2 - \theta \,| + |\, \theta - a_1 \,|],$$

sai,

$$(a_2 - a_1)E_h\{\lambda(\theta)\,|\,x\} \leqq r_x(a_2) + r_x(a_1) < \infty.$$

Tome-se $a' \in A$, a' qualquer, e seja $a = a_1 - a'$; vem,

$$\lambda(\theta)|\,\theta - a\,| \leqq \lambda(\theta)|\,\theta - a_1\,| + \lambda(\theta)|\,a'\,|,$$

e, como, $E_h\{\lambda(\theta)\,|\,x\} < \infty$,

$$r_x(a) \leqq r_x(a_1) + |\,a'\,|E_h\{\lambda(\theta)\,|\,x\} < \infty.$$

Fica, portanto, demonstrado que se o risco a posteriori é finito para dois elementos de A então é finito para todo o $a \in A$, pois na última desigualdade a é arbitrário.

Se $a_0 \in A$ é o único elemento com risco a posteriori finito, verifica-se ser $E_h\{\lambda(\theta)\,|\,x\} = \infty$. De facto, dado qualquer $a \in A$, $a \neq a_0$, $r_x(a) = \infty$; escreva-se,

$$\begin{aligned} r_x(a) &= E_h\left\{\lambda(\theta)|\,\theta - a_0 + a_0 - a\,\|x\right\} \\ &\leqq |\,a_0 - a\,|E_h\{\lambda(\theta)\,|\,x\} + r_x(a_0), \end{aligned}$$

donde, $E_h\{\lambda(\theta)\,|\,x\} = \infty$, visto ser $r_x(a_0) < \infty$. A demonstração fica completa. □□

Tome-se, com $E_h\{\lambda(\theta)\,|\,x\} < \infty$, a distribuição a priori, h_λ, definida por (6.53). Obtém-se,

$$r_x(a) = E_h\left\{\lambda(\theta)\,|\,\theta - a\,\|x\right\} = E_{h_\lambda}\left\{|\,\theta - a\,\|x\right\};$$

o estimador Bayes contra h obtém-se através da minimização de $r_x(a)$. Mas, $r_x(a)$ é o desvio médio em relação ao ponto a da distribuição a posteriori, $h_\lambda(\theta\,|\,x)$, obtida da distribuição a priori, $h_\lambda(\theta)$,

$$h_\lambda(\theta\,|\,x) = f(x\,|\,\theta)\; h_\lambda(\theta)/f(x).$$

Como o desvio médio é mínimo quando tomado em relação à mediana (veja-se secção 3.5), tem-se,

Teorema 6.22 — Com função perca erro absoluto, se a distribuição a priori, h, é tal que existe um $a \in A$ com $r_x(a) < \infty$, então existe um estimador Bayes, essencialmente único, definido por,

$$\delta_h(x) = \begin{cases} a_0 \text{ se } r_x(a) < \infty \text{ unicamente para } a = a_0, \\ \text{mediana de } h_\lambda(\theta\,|\,x) \text{ se } r_x(a) < \infty \text{ para todo } a \in A. \end{cases}$$

□□

Exemplo 6.22 — Considere-se a estimação do parâmetro θ da Binomial quando $h(\theta) = I_{[0,1]}(\theta)$ e a função perca tem por expressão,

$$L(\theta, a) = \frac{|\,\theta - a\,|}{\theta(1-\theta)}.$$

Tem-se,

$$r_x(a) = \int_0^1 \left[\frac{|\,\theta - a\,|}{\theta(1-\theta)}\right] h(\theta \,|\, x)\, d\theta,$$

onde,

$$h(\theta \,|\, x) = [B(x+1, N-x+1)]^{-1}\theta^x(1-\theta)^{N-x}.$$

Suponha-se $x = 0$: $r_0(a)$ só é finito para $a = 0$; se $x = N$, $r_N(a)$ só é finito para $a = 1$; se $0 < x < N$, $r_x(a)$ é finito para todo o $a \in [0,1]$. O estimador Bayes vem definido por,

$$\begin{aligned}
x = 0 \quad &\rightarrow \quad \delta_h(x) = 0, \\
0 < x < N \quad &\rightarrow \quad \delta_h(x) = \text{ mediana da distribuição:} \\
&\qquad\qquad h_\lambda(\theta \,|\, x) = [B(x, N-x)]^{-1}\theta^{x-1}\,(1-\theta)^{N-x+1}, \\
x = N \quad &\rightarrow \quad \delta_h(x) = 1.
\end{aligned}$$

[Blackwell e Girshick (1954)]. □

CAPÍTULO 7

PROBLEMAS DE BIDECISÃO (ENSAIO DE HIPÓTESES)

7.1 Posição do problema

Considere-se a família de distribuições, $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$, da variável ou vector aleatório, X, observado no quadro da experiência planeada. Hipótese estatística, recorde-se, é proposição formulada sobre P_θ e que estabelece em $\mathcal{F}$ uma dicotomia:

$$\begin{aligned} \mathcal{F}_0 =&\{P_\theta : \theta \in \Theta_0\},\ \mathcal{F}_1 = \{P_\theta : \theta \in \Theta_1\}, \\ &\Theta_0 \cup \Theta_1 = \Theta, \Theta_0 \cap \Theta_1 = \emptyset, \end{aligned} \tag{7.1}$$

compreendendo $\mathcal{F}_0$ as distribuições para as quais a hipótese é verdadeira e $\mathcal{F}_1$ as distribuições para as quais a hipótese é falsa.

Matematicamente a hipótese é equivalente à proposição, $P_\theta \in \mathcal{F}_0$ ou $\theta \in \Theta_0$; designa-se por,

$$H_0\colon\ P_\theta \in \mathcal{F}_0 \ \text{ ou } \ \theta \in \Theta_0. \tag{7.2}$$

A classe de alternativas designa-se por,

$$H_1\colon\ P_\theta \in \mathcal{F}_1 \ \text{ ou } \ \theta \in \Theta_1. \tag{7.3}$$

O ensaio ou teste da hipótese H_0 consiste em decidir, após observar X, se a hipótese é aceite ou rejeitada. Seja, $A = \{a_0, a_1\}$, onde,

- a_0 corresponde à aceitação de H_0,
- a_1 corresponde à rejeição de H_0;

uma função de decisão pura, $\delta \in D$, faz corresponder a cada $x \in \mathfrak{X}$ uma das duas acções. Assim, cada $\delta \in D$ equivale a uma partição de $\mathfrak{X}$ com dois elementos, sejam W_δ^c e W_δ, tais que,

$$\begin{aligned} W_\delta^c &= \{x \in \mathfrak{X} \,:\, \delta(x) = a_0\}, \\ W_\delta &= \{x \in \mathfrak{X} \,:\, \delta(x) = a_1\}, \\ W_\delta \cup W_\delta^c &= \mathfrak{X},\ W_\delta \cap W_\delta^c = \emptyset. \end{aligned}$$

Os conjuntos W_δ e W_δ^c são mensuráveis pois pressupõe-se que todo o $\delta \in D$ é função mensurável. A W_δ também se chama região crítica ou região de rejeição: se $x \in W_\delta$ a hipótese H_0 é rejeitada; W_δ^c é a região de aceitação: se $x \in W_\delta^c$ a hipótese H_0 é aceite.

Considere-se a função perca, $L(\theta, a_i)$, $i = 0, 1$, e ponha-se para simplificar, $L(\theta, a_i) = L_i(\theta)$, $i = 0, 1$. A função risco ou perca esperada assume a forma,

$$\begin{aligned} R(\theta, \delta) &= L_0(\theta) P_\theta(X \in W_\delta^c) + L_1(\theta) P_\theta(X \in W_\delta) \\ &= L_0(\theta)[1 - P_\theta(W_\delta)] + L_1(\theta) P_\theta(W_\delta) \\ &= L_0(\theta) + [L_1(\theta) - L_0(\theta)] P_\theta(W_\delta), \end{aligned} \tag{7.4}$$

onde $P_\theta(W_\delta)$ é a probabilidade de rejeição, isto é, do acontecimento, $(X \in W_\delta)$, quando o estado ou verdadeiro valor do parâmetro é θ. Com W_δ fixo — para um dado $\delta \in D$ — a função de θ, $P_\theta(W_\delta)$, designa-se por função potência da região crítica W_δ ou do teste caracterizado por δ.

Uma função de decisão mista, $\delta^* \in D^*$, sendo a combinação de elementos, $\delta \in D$, segundo a distribuição de probabilidade associada com δ^*, consiste, no caso presente, na escolha, por processo casual ditado por essa distribuição, de um conjunto da classe $\{W_\delta \subset \mathfrak{X} \,:\, \delta \in D\}$. Devido à complexidade que envolve semelhante método de casualização, o tratamento através de funções de decisão aleatórias revela-se mais acessível e torna-se, por isso, mais relevante na teoria do ensaio de hipóteses.

Sendo $\phi \in \Phi$ função de decisão aleatória, tem-se,

$$\phi \equiv [\phi(a_0 \,|\, x),\ \phi(a_1 \,|\, x)],$$

onde $\phi(a_0 \,|\, x)$ é a probabilidade de aceitar H_0 quando se observa $X = x$ e $\phi(a_1 \,|\, x)$ é a probabilidade de rejeitar H_0 quando se observa $X = x$. Obviamente,

$$\phi(a_0 \,|\, x) \geqq 0,\ \phi(a_1 \,|\, x) \geqq 0,\ \phi(a_0 \,|\, x) + \phi(a_1 \,|\, x) = 1. \tag{7.5}$$

Em face da última relação adopta-se, correntemente, a seguinte convenção:

$$\phi(a_1 \,|\, x) = \phi(x),\ \phi(a_0 \,|\, x) = 1 - \phi(x),\ 0 \leqq \phi(x) \leqq 1, \tag{7.6}$$

sendo cada $\phi \in \Phi$, inteiramente caracterizado pela função $\phi(x)$ também designada função teste.

É de verificação imediata que a classe de funções de decisão aleatórias, Φ, contém a classe de funções de decisão puras, D: a função de decisão pura, δ, associada com a região crítica W_δ, é o elemento de Φ definido por,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad x \in W_\delta, \\ 0 & \text{se} \quad x \in W_\delta^c. \end{cases}$$

A função risco, $\hat{R}(\theta, \phi)$, passa a escrever-se sem chapéu «^» por não haver lugar para confusão. Tem-se, para qualquer $\phi \in \Phi$,

$$\begin{aligned} R(\theta, \phi) &= E_\theta\{L_0(\theta)[1 - \phi(X)] + L_1(\theta)\phi(X)\} \\ &= L_0(\theta) + [L_1(\theta) - L_0(\theta)]E_\theta\{\phi(X)\}, \end{aligned} \tag{7.7}$$

onde, $E_\theta\{\phi(X)\}$, probabilidade de rejeitar H_0 quando o estado é θ, considerada como função de θ, é agora a função potência da função de decisão ou função teste ϕ.

A formulação clássica do problema do ensaio de hipóteses obtém-se especializando em relação à função perca, isto é, tomando a função perca «0–1»,

$$L_0(\theta) = \begin{cases} 0 & \text{se} \quad \theta \in \Theta_0, \\ 1 & \text{se} \quad \theta \in \Theta_1, \end{cases} \qquad L_1(\theta) = \begin{cases} 1 & \text{se} \quad \theta \in \Theta_0, \\ 0 & \text{se} \quad \theta \in \Theta_1, \end{cases} \tag{7.8}$$

e tomando a função teste igual à função indicatriz da região crítica, $\phi(x) = I_W(x)$. Consequentemente,

$$R(\theta, \phi) = R(\theta, W) = \begin{cases} P_\theta(W) & \text{se} \quad \theta \in \Theta_0, \\ 1 - P_\theta(W) & \text{se} \quad \theta \in \Theta_1; \end{cases}$$

assim, para $\theta \in \Theta_0$, $R(\theta, W) = P_\theta(W)$, é a probabilidade de rejeitar a hipótese H_0 quando é verdadeira (erro de 1.ª espécie), para $\theta \in \Theta_1$, $R(\theta, W) = 1 - P_\theta(W)$, é a probabilidade de aceitar a hipótese H_0 quando é falsa (erro de 2.ª espécie). Neste contexto, a solução do problema consiste na determinação da região crítica, W, que permita controlar ou limitar, em algum sentido, as probabilidades de erro.

Em muitos problemas a função perca tem uma estrutura semelhante a (7.8) embora não tão simples,

$$L_0(\theta) = \begin{cases} = 0 & \text{se} \quad \theta \in \Theta_0, \\ > 0 & \text{se} \quad \theta \in \Theta_1, \end{cases} \qquad L_1(\theta) = \begin{cases} > 0 & \text{se} \quad \theta \in \Theta_0, \\ = 0 & \text{se} \quad \theta \in \Theta_1; \end{cases} \tag{7.9}$$

em semelhantes casos a função risco tem a forma,

$$R(\theta, \phi) = \begin{cases} L_1(\theta)E_\theta\{\phi(X)\} & \text{para } \theta \in \Theta_0, \\ L_0(\theta)[1 - E_\theta\{\phi(X)\}] & \text{para } \theta \in \Theta_1. \end{cases} \tag{7.10}$$

Não surpreende, portanto, que a teoria do ensaio de hipóteses seja desenvolvida, em muitos aspectos, concentrando a atenção na função potência, $E_\theta\{\phi(X)\}$, sem fazer referência explícita à função perca.

A hipótese a ensaiar H_0 (ou a hipótese alternativa H_1) diz-se simples quando Θ_0 (ou Θ_1) contém apenas um elemento; diz-se composta se Θ_0 (ou Θ_1) contém pelo menos dois elementos.

Na abordagem inicial considera-se o ensaio de uma hipótese simples contra uma alternativa também simples,

$$H_0\colon \theta = \theta_0 \quad \text{contra} \quad H_1\colon \theta = \theta_1;$$

quer dizer, supõe-se $\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 = \{\theta_0\}$ e $\Theta_1 = \{\theta_1\}$. Como Θ é finito pode recorrer-se ao conceito de conjunto risco casualizado [veja-se (3.9)] S^*_Φ, ou mais simplesmente, S^*,

$$S^* = \{[R(\theta_0, \phi), R(\theta_1, \phi)] \,:\, \phi \in \Phi\}, \tag{7.11}$$

com $R(\theta_i, \phi)$, $i = 0, 1$, definido por (7.7). Por outro lado, é natural admitir as seguintes relações,

$$L_0(\theta_0) < L_1(\theta_0);\ L_0(\theta_1) > L_1(\theta_1), \tag{7.12}$$

pois, de contrário, as decisões correctas seriam mais penalizadas do que as incorrectas.

Adaptando (7.7) ao caso presente, sai,

$$\begin{aligned} R(\theta_0, \phi) &= L_0(\theta_0) + [L_1(\theta_0) - L_0(\theta_0)]E_{\theta_0}\{\phi(X)\}, \\ R(\theta_1, \phi) &= L_1(\theta_1) + [L_0(\theta_1) - L_1(\theta_1)]E_{\theta_1}\{1 - \phi(X)\}. \end{aligned} \tag{7.13}$$

Definindo,

$$\alpha(\phi) = E_{\theta_0}\{\phi(X)\},\ \beta(\phi) = E_{\theta_1}\{1 - \phi(X)\}, \tag{7.14}$$

a mudança de origem e de escala das coordenadas permite substituir o conjunto risco casualizado, S^*, pelo conjunto equivalente,

$$S^*_0 = \{[\alpha(\phi), \beta(\phi)] \,:\, \phi \in \Phi\}. \tag{7.15}$$

Teorema 7.1 — O conjunto S^*_0 é convexo, contém os pontos $(1, 0)$ e $(0, 1)$ e é simétrico em relação ao ponto $(1/2, 1/2)$.

Dem. (a) S^*_0 é convexo: se (α', β') e (α'', β'') pertencem a S^*_0 e se $\lambda \in [0,1]$, também,

$$\lambda(\alpha', \beta') + (1 - \lambda)(\alpha'', \beta'') \in S^*_0,$$

porquanto, ϕ', $\phi'' \in \Phi$, arrasta, $\lambda\phi' + (1 - \lambda)\phi'' \in \Phi$; *(b)* S^*_0 contém (1,0) e (0,1): tomem-se, respectivamente, $\phi(x) \equiv 1$ e $\phi(x) \equiv 0$; *(c)* S^*_0 é simétrico em relação a $(1/2, 1/2)$: $\alpha(1 - \phi) = 1 - \alpha(\phi)$ e $\beta(1 - \phi) = 1 - \beta(\phi)$ e se $\phi \in \Phi$ também $1 - \phi \in \Phi$. □□

A $\alpha = \alpha(\phi) = E_{\theta_0}\{\phi(X)\}$, probabilidade do teste ϕ rejeitar H_0 quando é verdadeira, chama-se tamanho de ϕ; $\beta = \beta(\phi) = E_{\theta_1}\{1 - \phi(X)\}$, é a probabilidade de o teste ϕ aceitar H_0 quando é falsa.

Um teste ϕ — função de decisão aleatória — diz-se o melhor de tamanho α para ensaiar a hipótese simples H_0 contra a alternativa simples H_1, se $E_{\theta_0}\{\phi(X)\} = \alpha$ e se para qualquer outro teste ϕ' de tamanho não superior a α, $E_{\theta_0}\{\phi'(X)\} \leqq \alpha$, se tem,

$$E_{\theta_1}\{1 - \phi(X)\} \leqq E_{\theta_1}\{1 - \phi'(X)\}. \tag{7.16}$$

É interessante notar que se ϕ é admissível, então ϕ é o melhor teste do seu tamanho. Com efeito, se ϕ é admissível tem-se, qualquer que seja $\phi' \in \Phi$, que,

$$R(\theta_0, \phi') \leqq R(\theta_0, \phi) \quad \text{arrasta} \quad R(\theta_1, \phi') \geqq R(\theta_1, \phi),$$

isto é,

$$E_{\theta_0}\{\phi'(X)\} \leqq E_{\theta_0}\{\phi(X)\} \quad \text{arrasta} \quad E_{\theta_1}\{1 - \phi'(X)\} \geqq E_{\theta_1}\{1 - \phi(X)\},$$

e a conclusão segue-se.

A recíproca não é necessariamente verdadeira. Se $\phi \in \Phi$ é o melhor teste de tamanho α, ϕ pode não ser admissível. Com efeito, se

$$E_{\theta_1}\{1 - \phi(X)\} = 0,$$

pode existir $\phi' \in \Phi$, com $E_{\theta_0}\{\phi'(X)\} < \alpha$ e $E_{\theta_1}\{1-\phi'(X)\} = 0$; logo, ϕ não é admissível (veja-se Ex. 7.3).

7.2 Lema de Neyman-Pearson

Um método geral para determinar o melhor teste de tamanho α para ensaiar uma hipótese simples contra uma alternativa simples contém-se no importante Lema de Neyman-Pearson. Esta proposição, adiante demonstrada em termos decisionais, tem aplicações que transcendem largamente tal situação. Entretanto, a título introdutório, considere-se a variável aleatória discreta, X, com função de probabilidade,

$$f(x \mid \theta) = P_\theta(X = x), x \in \mathfrak{X},$$

$\mathfrak{X}$ conjunto finito ou infinito numerável. Sejam, H_0: $\theta = \theta_0$ e H_1: $\theta = \theta_1$ e retome-se o teste não casualizado associado com uma região crítica, W. A melhor região crítica de tamanho, α, W_o, deve satisfazer,

$$\begin{aligned} P_{\theta_0}(W_o) &= \sum_{x \in W_o} f(x \mid \theta_0) = \alpha, \\ P_{\theta_1}(W_o) &= \sum_{x \in W_o} f(x \mid \theta_1) \quad \text{máxima.} \end{aligned} \tag{7.17}$$

Se W é qualquer outra região crítica tal que $P_{\theta_0}(W) \leqq \alpha$, então deve ter-se, $P_{\theta_1}(W_o) \geqq P_{\theta_1}(W)$. Para seleccionar os pontos $x \in \mathfrak{X}$ que devem formar o conjunto $W_o \subset \mathfrak{X}$ pode pensar-se que cada x é avaliado segundo dois critérios: a probabilidade $f(x\,|\,\theta_0)$ e a probabilidade $f(x\,|\,\theta_1)$. No seu conjunto, os x incluídos em W_o não podem ter pontuação superior a α quando avaliados pelo primeiro critério e devem ter pontuação o maior possível quando avaliados pelo segundo critério. Parece então claro que a escolha deve ser orientada pela razão,

$$\lambda(x\,|\,\theta_0,\theta_1) = \frac{f(x\,|\,\theta_1)}{f(x\,|\,\theta_0)},$$

incluindo em W_o os pontos x com razão mais elevada, tantos quanto o permita a primeira relação (7.17). Formalmente, W_o é o conjunto de todos os pontos x tais que $\lambda(x\,|\,\theta_0,\theta_1) > c$, com c determinado pela condição,

$$P_{\theta_0}(W_o) = \Sigma f(x\,|\,\theta_0) = \alpha,$$

onde o somatório percorre o conjunto $\{x : \lambda(x\,|\,\theta_0,\theta_1) > c\}$.

Na sequência do processo pode surgir uma dificuldade: ao incluir um dado ponto, o valor α não é atingido, mas ao passar ao ponto seguinte, respeitando a ordem decrescente das razões, o valor α é excedido. Esta dificuldade pode resolver-se de dois modos: considerando como primeira condição em (7.17), $P_{\theta_0}(W_o) \leqq \alpha$, o que equivale a procurar a melhor região crítica de tamanho não superior a α; procedendo a uma casualização para obter a melhor região crítica exactamente de tamanho α. De acordo com a segunda óptica, o ponto cuja inclusão faz ultrapassar α, seja x_0, é «incluído com probabilidade γ», com γ saído da equação,

$$P_{\theta_0}(W_o^-) + \gamma f(x_0\,|\,\theta_0) = \alpha,$$

onde $P_{\theta_0}(W_o^-) < \alpha$ e $P_{\theta_0}(W_o^- \cup \{x_0\}) > \alpha$. Quando se diz que x_0 é incluído com probabilidade γ pretende significar-se que se for observado $X = x_0$ a hipótese é rejeitada com probabilidade γ (e aceite com probabilidade $1-\gamma$). Pode suceder — veja-se Ex.7.1 — que se apresentem vários pontos com a mesma prioridade de x_0, isto é, tais que,

$$\lambda(x_0\,|\,\theta_0,\theta_1) = \lambda(x_o'\,|\,\theta_0,\theta_1) = \lambda(x_0''\,|\,\theta_0,\theta_1) = \ldots;$$

nesse caso, sendo $W_o^+ = \{x_0, x_0', x_0'', \ldots\}$, determina-se γ de modo a verificar-se,

$$P_{\theta_0}(W_o^-) + \gamma P_{\theta_0}(W_o^+) = \alpha.$$

Exemplo 7.1 — Considere-se a amostra casual, $\mathbf{X} = (X_1, X_2)$, da Poisson com média θ:

$$f(\mathbf{x}\,|\,\theta) = e^{-2\theta}\theta^{x_1+x_2}/x_1!x_2!,\ \theta > 0,$$

e $\mathcal{X} = \{\mathbf{x} = (x_1, x_2) : x_i = 0, 1, 2, \ldots,\ i = 1, 2\}$. Seja, H_0: $\theta = 2$, H_1: $\theta = 4$ e considere-se a razão,

$$\lambda(\mathbf{x}\,|\,2, 4) = f(\mathbf{x}\,|\,4)/f(\mathbf{x}\,|\,2) = e^{-4}2^{x_1+x_2}.$$

Vem,

$$\{\mathbf{x} : \lambda(\mathbf{x}\,|\,2, 4) > c\} = \{\mathbf{x} : x_1 + x_2 > c^*\},$$

com $c^* = (\log c + 4)/\log 2$. Arbitrado um valor a α, seja $\alpha = 0{,}05$, os cálculos pertinentes são,

$$P_{\theta=2}(X_1 + X_2 > 8) = 0{,}021,\ P_{\theta=2}(X_1 + X_2 = 8) = 0{,}030;$$
$$P_{\theta=4}(X_1 + X_2 > 8) = 0{,}407,\ P_{\theta=4}(X_1 + X_2 = 8) = 0{,}140.$$

A melhor região crítica de tamanho $\alpha = 0{,}05$ é constituída, nos termos acima indicados, por,

$$W_o^- = \{\mathbf{x} : x_1 + x_2 > 8\},\ W_o^+ = \{\mathbf{x} : x_1 + x_2 = 8\},$$

verificando-se,

$$P_{\theta=2}(W_o^-) + \gamma P_{\theta=2}(W_o^+) = 0{,}05,$$

com $\gamma = 0{,}967$. Assim, se $\mathbf{X} \in W_o^-$ a hipótese H_0 é rejeitada com probabilidade um; se $\mathbf{X} \in W_o^+$ a hipótese H_0 é rejeitada com probabilidade 0,967 e aceite com probabilidade 0,033. Por outro lado, a probabilidade,

$$P_{\theta=4}(W_o^-) + (0{,}967)P_{\theta=4}(W_o^+) = 0{,}54,$$

é máxima no sentido já indicado, isto é, nenhuma outra região crítica de tamanho 0,05 tem maior probabilidade de rejeitar H_0 quando é falsa.

Note-se que a região crítica definida por W_o^- e W_o^+ é equivalente à função teste ou função de decisão,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se } \mathbf{x} \in W_o^- \quad (\text{ou se } x_1 + x_2 > 8), \\ 0{,}967 & \text{se } \mathbf{x} \in W_o^+ \quad (\text{ou se } x_1 + x_2 = 8), \\ 0 & \text{se } \mathbf{x} \in (W_o^- \cup W_o^+)^c \quad (\text{ou se } x_1 + x_2 < 8), \end{cases}$$

para a qual,

$$E_{\theta=2}\{\phi(\mathbf{X})\} = 0{,}05,\ E_{\theta=4}\{1 - \phi(\mathbf{X})\} = 0{,}46.$$

Nenhum outro teste de tamanho 0,05 tem probabilidade de um erro de 2.ª espécie inferior a 0,46. □

Demonstra-se seguidamente o conhecido Lema de Neyman-Pearson justamente considerado uma das pedras fundamentais da teoria da decisão estatística. O problema é ainda o de ensair uma hipótese simples, H_0: $\theta = \theta_0$, contra uma alternativa simples, H_1: $\theta = \theta_1$. Limita-se o tratamento ao caso contínuo uma vez que o caso discreto se obtém por decalque [Ferguson (1967)].

Teorema 7.2 — Suponha-se $\Theta = \{\theta_0, \theta_1\}$ e seja $f(x\,|\,\theta)$ a função de densidade da variável aleatória X:

(I) Qualquer teste, $\phi(x)$, da forma,

$$\phi(x) = \begin{cases} 1 & \text{se } f(x\,|\,\theta_1) > c f(x\,|\,\theta_0), \\ \gamma(x) & \text{se } f(x\,|\,\theta_1) = c f(x\,|\,\theta_0), \\ 0 & \text{se } f(x\,|\,\theta_1) < c f(x\,|\,\theta_0), \end{cases} \tag{7.18}$$

para algum $c \geqq 0$, com $0 \leqq \gamma(x) \leqq 1$, é o melhor do seu tamanho para ensaiar H_0 contra H_1. Correspondendo a $c = \infty$, o teste,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,\theta_0) = 0, \\ 0 & \text{se} \quad f(x\,|\,\theta_0) > 0, \end{cases} \tag{7.19}$$

é o melhor do tamanho 0 para ensaiar H_0 contra H_1.

(II) Para todo o α, $0 \leqq \alpha \leqq 1$, existe um teste da forma (7.18) ou (7.19), com $\gamma(x) = \gamma$, constante, para o qual $E_{\theta_0}\{\phi(X)\} = \alpha$.

(III) Se $\phi(x)$ é o melhor teste de tamanho α para ensaiar H_0 contra H_1, então tem a forma (7.18) ou (7.19), excepto, quando muito, sobre um conjunto de $\mathfrak{X}$ com probabilidade zero quando $\theta = \theta_0$ ou quando $\theta = \theta_1$.

Dem. (I) Seja $\phi(x)$ da forma (7.18) e considere-se outro teste qualquer, $\phi'(x)$, $0 \leqq \phi' \leqq 1$, com $E_{\theta_0}\{\phi'(X)\} \leqq E_{\theta_0}\{\phi(X)\}$. Tem-se,

$$\int_{\mathfrak{X}} [\phi(x) - \phi'(x)][f(x\,|\,\theta_1) - c f(x\,|\,\theta_0)]\, dx \geqq 0, \tag{7.20}$$

visto a função integranda ser não negativa. Em consequência,

$$E_{\theta_1}\{\phi(X)\} - E_{\theta_1}\{\phi'(X)\} \geqq c[E_{\theta_0}\{\phi(X)\} - E_{\theta_0}\{\phi'(X)\}] \geqq 0, \tag{7.21}$$

donde $E_{\theta_1}\{\phi(X)\} \geqq E_{\theta_1}\{\phi'(X)\}$ e $\phi(x)$ é melhor do que $\phi'(x)$.

No caso, $c = \infty$, $\phi(x)$ da forma (7.19), tem tamanho zero e qualquer outro teste de tamanho zero, $\phi'(x)$, tem necessariamente de ser $\phi'(x) = 0$ quando

$x \in \{x : f(x\,|\,\theta_0) > 0\}$, excepto quando muito num subconjunto com probabilidade zero. Portanto,

$$E_{\theta_1}\{\phi(X) - \phi'(X)\} = \int_{\mathfrak{X}_0} [1 - \phi'(x)] f(x\,|\,\theta_1)\, dx \geqq 0,$$

com $\mathfrak{X}_0 = \{x : f(x\,|\,\theta_0) = 0\}$, e a demonstração de (I) fica completa.

(II) Como (7.19) indica a forma do melhor teste de tamanho zero, pode limitar-se o estudo a $0 < \alpha \leqq 1$. O tamanho do teste (7.18), quando $\gamma(x) = \gamma$, é

$$\begin{aligned} E_{\theta_0}\{\phi(X)\} &= P_{\theta_0}[f(X\,|\,\theta_1) > c f(X\,|\,\theta_0)] + \gamma P_{\theta_0}[f(X\,|\,\theta_1) = c f(X\,|\,\theta_0)] \\ &= 1 - P_{\theta_0}(Y \leqq c) + \gamma P_{\theta_0}(Y = c), \end{aligned} \tag{7.22}$$

onde $Y = f(X\,|\,\theta_1)/f(X\,|\,\theta_0)$. Com α fixo, $0 < \alpha \leqq 1$, há que determinar c e γ de modo a ser $E_{\theta_0}\{\phi(X)\} = \alpha$, isto é,

$$P_{\theta_0}(Y \leqq c) - \gamma P_{\theta_0}(Y = c) = 1 - \alpha. \tag{7.23}$$

Se existe c_0 tal que $P_{\theta_0}(Y \leqq c_0) = 1 - \alpha$, toma-se $\gamma = 0$ e $c = c_0$ e o problema está resolvido. Caso contrário, existe c_0 tal que (veja-se Fig. 7.1),

$$P_{\theta_0}(Y < c_0) \leqq 1 - \alpha < P_{\theta_0}(Y \leqq c_0); \tag{7.24}$$

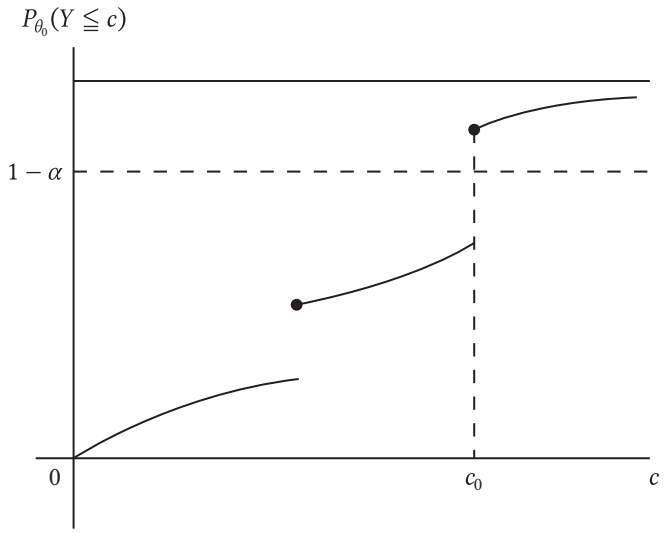


Fig. 7.1

nesse caso toma-se $c = c_0$ e,

$$\gamma = \frac{P_{\theta_0}(Y \leqq c_0) - (1-\alpha)}{P_{\theta_0}(Y = c_0)}, \tag{7.25}$$

que satisfaz (7.23), sendo ainda, $0 \leqq \gamma \leqq 1$.

(III) Se $\alpha = 0$, o argumento desenvolvido em (I) mostra que $\phi(x) = 0$ no conjunto, $\{x : f(x\,|\,\theta_0) > 0\}$, quase por toda a parte. Se $\phi'(x)$ tem probabilidade mínima de conduzir a um erro de 2.ª espécie, então deve ser $1 - \phi'(x) = 0$ no conjunto, $\{x : f(x\,|\,\theta_1) > 0\}$–$\{x : f(x\,|\,\theta_0) > 0\}$, quase por toda a parte. Assim, $\phi'(x)$ difere de $\phi(x)$ dado por (7.19) quando muito sobre um conjunto com probabilidade zero quando $\theta = \theta_0$ ou quando $\theta = \theta_1$. Se $\alpha > 0$, seja $\phi(x)$ um teste melhor do tamanho α da forma (7.18); então, por ser,

$$E_{\theta_0}\{\phi(X)\} = E_{\theta_0}\{\phi'(X)\},\ E_{\theta_1}\{\phi(X)\} = E_{\theta_1}\{\phi'(X)\},$$

o integral (7.20) deve ser igual a zero. Mas como a função integranda é não negativa resulta que $\phi(x) = \phi'(x)$ em quase todos os pontos do conjunto $\{x : f(x\,|\,\theta_1) \neq c f(x\,|\,\theta_0)\}$. Portanto, $\phi'(x)$ tem a forma (7.18), quase por toda a parte, com o mesmo valor de c que caracteriza $\phi(x)$ e a demonstração fica completa. □□

Exemplo 7.2 — Seja $(X_1, X_2, \ldots, X_N)$ amostra casual de uma população de Bernoulli. Pretende ensaiar-se H_0: $\theta = \theta_0$ contra H_1: $\theta = \theta_1$ ($\theta_1 > \theta_0$); sabe-se que $T = \Sigma X_i$ é estatística suficiente para θ com função de probabilidade Binomial,

$$f(t\,|\,\theta) = \binom{N}{t}\theta^t(1-\theta)^{N-t},\ t = 0, 1, \ldots, N,\ (t = \Sigma x_i).$$

Pelo Teorema 7.2, com $0 < \alpha < 1$, o melhor teste de tamanho α é da forma,

$$\phi(t) = \begin{cases} 1 & \text{se} \quad f(t\,|\,\theta_1) > c f(t\,|\,\theta_0), \\ \gamma & \text{se} \quad f(t\,|\,\theta_1) = c f(t\,|\,\theta_0), \\ 0 & \text{se} \quad f(t\,|\,\theta_1) < c f(t\,|\,\theta_0). \end{cases}$$

Mas, $f(t\,|\,\theta_1) > c f(t\,|\,\theta_0)$, implica,

$$\theta_1^t(1-\theta_1)^{N-t} > c\theta_0^t(1-\theta_0)^{N-t} \quad \Rightarrow \quad t > c'.$$

Assim,

$$\phi(t) = \begin{cases} 1 & \text{se} \quad t > c', \\ \gamma & \text{se} \quad t = c', \\ 0 & \text{se} \quad t < c', \end{cases}$$

com c' determinado pelas condições,

$$\sum_{t=c'+1}^{N} \binom{N}{t} \theta_0^t (1-\theta_0)^{N-t} \leqq \alpha, \quad \sum_{t=c'}^{N} \binom{N}{t} \theta_0^t (1-\theta_0)^{N-t} > \alpha,$$

e γ pela condição,

$$\sum_{t=c'+1}^{N} \binom{N}{t} \theta_0^t (1-\theta_0)^{N-t} + \gamma \binom{N}{c'} \theta_0^{c'} (1-\theta_0)^{N-c'} = \alpha.^1$$

□

Exemplo 7.3 — Suponha-se que a variável aleatória X tem função de densidade,

$$f(x \mid \theta) = I_{[\theta+1,\,\theta+2]}(x),$$

e que pretende ensaiar-se H_0: $\theta = 0$ contra H_1: $\theta = 0{,}5$ (veja-se Fig. 7.2). Para determinar o melhor teste de tamanho α, $0 \leqq \alpha \leqq 1$, considerem-se os seguintes casos:

(a) $\alpha = 0$. Pelo Teorema 7.2, o melhor teste de tamanho zero é da forma (7.19), ou seja, no caso presente,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad 2 < x \leqq 2{,}5, \\ 0 & \text{se} \quad 1 \leqq x \leqq 2, \end{cases} \qquad [A]$$

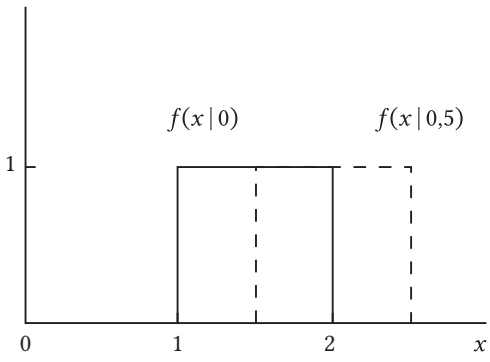


Fig. 7.2

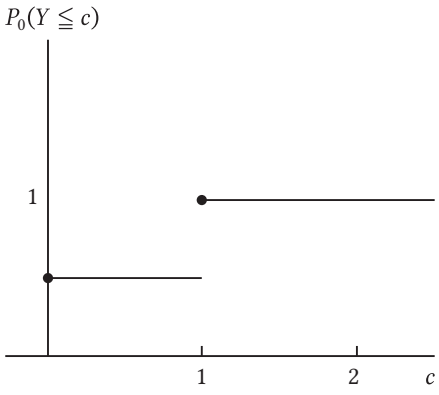


Fig. 7.3

[1] Sem casualização não seria possível, em geral, obter, neste caso, um teste exactamente de dimensão α.

Para tratar os restantes casos considere-se, $P_0(Y \leqq c)$, função de distribuição de $Y = f(X\,|\,0{,}5)/f(X\,|\,0)$, quando $\theta = 0$. Como facilmente se calcula (veja-se Fig. 7.3),

$$P_0(Y \leqq c) = \begin{cases} 0 & \text{se} \quad -\infty < c < 0, \\ 0{,}5 & \text{se} \quad 0 \leqq c < 1, \\ 1 & \text{se} \quad 1 \leqq c < +\infty. \end{cases}$$

(b) $0 < \alpha < 0{,}5$. Como, $0{,}5 < 1 - \alpha < 1$, o valor c_0 que satisfaz,

$$P_0(Y < c_0) \leqq 1 - \alpha < P_0(Y \leqq c_0),$$

é $c_0 = 1$, pois, $P_0(Y < 1) = 0{,}5 < 1 - \alpha < 1 = P_0(Y \leqq 1)$; para γ vem, por (7.25),

$$\gamma = \frac{P_0(Y \leqq c_0) - (1-\alpha)}{P_0(Y = c_0)} = \frac{P_0(Y \leqq 1) - (1-\alpha)}{P_0(Y = 1)} = 2\alpha.$$

O melhor teste de tamanho α é, então,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,0{,}5) > f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 2 < x \leqq 2{,}5, \\ 2\alpha & \text{se} \quad f(x\,|\,0{,}5) = f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 1{,}5 \leqq x \leqq 2 \\ 0 & \text{se} \quad f(x\,|\,0{,}5) < f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 1 \leqq x < 1{,}5. \end{cases} \qquad \text{[B]}$$

(c) $\alpha = 0{,}5$. Neste caso, $P_0(Y \leqq c) = 1 - 0{,}5 = 0{,}5$, para $0 \leqq c \leqq 1$ e, por (7.23), $\gamma = 0$. Então, o melhor teste de tamanho α, é

$$\phi(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,0{,}5) > c f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 1{,}5 \leqq x \leqq 2{,}5, \\ 0 & \text{se} \quad f(x\,|\,0{,}5) \leqq c f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 1 \leqq x < 1{,}5. \end{cases} \qquad \text{[C]}$$

(d) $0{,}5 < \alpha < 1$. Neste caso, $0 < 1 - \alpha < 0{,}5$, e o valor de c_0 que satisfaz,

$$P_0(Y < c_0) \leqq 1 - \alpha < P_0(Y \leqq c_0),$$

é $c_0 = 0$, pois, $P_0(Y < 0) = 0 < 1 - \alpha < 0{,}5 = P_0(Y \leqq 0)$; para γ vem, por (7.25),

$$\gamma = \frac{P_0(Y \leqq 0) - (1-\alpha)}{P_0(Y = 0)} = \frac{0{,}5 - (1-\alpha)}{0{,}5} = 2\alpha - 1,$$

donde, o melhor teste de tamanho α,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,0{,}5) > 0 \quad \text{ou} \quad \text{se} \quad 1{,}5 \leqq x \leqq 2{,}5, \\ 0 & \text{se} \quad f(x\,|\,0{,}5) = 0 \quad \text{ou} \quad \text{se} \quad 1 \leqq x < 1{,}5. \end{cases} \qquad \text{[D]}$$

(e) $\alpha = 1$. Neste caso, $\phi(x) = 1$ para $x \in \{x : f(x\,|\,0) > 0\}$; logo, o melhor teste de tamanho 1 é $\phi(x) \equiv 1$, $1 \leqq x \leqq 2{,}5$. □

O conjunto risco, S_0^* — veja-se (7.15) — além das propriedades a que se refere o Teorema 7.1, goza ainda da seguinte:

Teorema 7.3 — O conjunto risco, S_0^*, é fechado inferiormente.

Dem. Designe, na notação habitual, $\Lambda(S_0^*)$ a fronteira inferior de S_0^*; se $\mathbf{z} \in \Lambda(S_0^*)$, tem-se,

$$\{\mathbf{z}\} = Q_{\mathbf{z}} \cap \overline{S}_0^*, \tag{7.26}$$

onde $Q_{\mathbf{z}}$ é o quadrante inferior em $\mathbf{z}$ e $\overline{S}_0^*$ o fecho de S_0^*. Seja $\mathbf{z} = (\alpha, \beta)$; pelo Lema de Neyman-Pearson existe um teste, ϕ, melhor de tamanho α. A probabilidade de um erro de 2.ª espécie não pode ser inferior a β, pois se o fosse contradizia (7.26).

Se,

$$E_{\theta_1}\{1 - \phi(X)\} > \beta + \xi, \quad \xi > 0,$$

e $\alpha \neq 0$, então, devido à convexidade de S_0^* e ao facto de $(0,1), (1,0) \in S_0^*$, existem pontos, $(\alpha, \nu) \in S_0^*$, tais que, $\beta \leqq \nu < \beta + \xi$ qualquer que seja $\xi > 0$. Portanto, se $\alpha \neq 0$, $E_{\theta_1}\{1 - \phi(X)\} = \beta$. Se $\alpha = 0$, este argumento não é aplicável, mas, se $\phi'(x)$ é o melhor teste de tamanho $\beta + \xi$ para ensaiar H_1 contra H_0, a probabilidade de um erro de 2.ª espécie deve ser igual a zero em consequência do exposto e $1 - \phi'(x)$ é melhor do que $\phi(x)$ para ensaiar H_0 contra H_1, o que é uma contradição. A demonstração fica completa. □□

Este teorema — que se alarga imediatamente ao conjunto risco S^* — em conjugação com o Teorema 4.13 mostra que a classe formada pelos melhores testes de tamanho α [testes da forma (7.18) ou (7.19)] é uma classe completa mínima [a classe formada pelos testes (7.18) com $\gamma(x) = \gamma$ (constante) e (7.19) é essencialmente completa]. Para seleccionar um elemento daquela classe toma-se α, $0 \leqq \alpha \leqq 1$, e determina-se pelo Lema de Neyman-Pearson o melhor teste de tamanho α; o tamanho α pode interpretar-se como um índice por meio do qual se selecciona uma função de decisão admissível, excepto, é claro, quando se cai na situação em que $\beta = 0$, em que o melhor teste de tamanho α pode não ser admissível.

Exemplo 7.3 — *Continuação.* Para determinar o conjunto,

$$S_0^* = \{[\alpha(\phi), \beta(\phi)] : \phi \in \Phi\},$$

onde,

$$\alpha(\phi) = E_{\theta=0}\{\phi(X)\} = \int_1^2 \phi(x)\, dx,$$

$$\beta(\phi) = E_{\theta=0,5}\{1 - \phi(X)\} = 1 - \int_{1,5}^{2,5} \phi(x)\, dx,$$

note-se que S_0^* é convexo, simétrico em relação ao ponto $(1/2, 1/2)$ e contem os pontos $(0,1)$ e $(1,0)$. Por outro lado, pelo Teorema 7.3, a fronteira inferior pertence a S_0^* e é formada pelos pontos (α, β) correspondentes aos melhores testes de tamanho α. Se $\alpha = 0$, o melhor teste é dado por [A], para o qual se tem $\beta = 0.5$. O ponto $(0,1)$ corresponde ao teste $\phi(x) \equiv 0$. Com $0 < \alpha < 0{,}5$, o melhor teste é dado por [B], donde,

$$\beta = 1 - \int_{1,5}^{2} 2\alpha \, dx - \int_{2}^{2,5} dx = 0{,}5 - \alpha,$$

e os pontos (α, β) formam o segmento que une os pontos $(0, 0{,}5)$ e $(0{,}5, 0)$. Com $\alpha = 0{,}5$, tem-se, por [C], $\beta = 0$. Finalmente, com $0{,}5 < \alpha \leqq 1$, tem-se $\beta = 0$. Repare-se que neste último caso os melhores testes de tamanho α não são admissíveis. O conjunto S_0^* representa-se na Fig. 7.4.

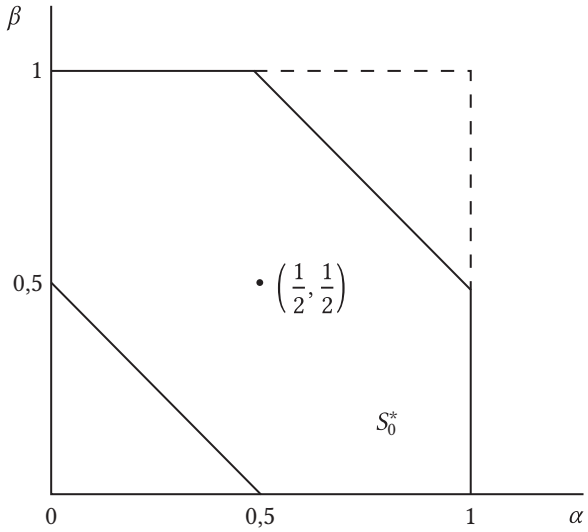


Fig. 7.4

□

Exemplo 7.4 — Suponha-se que a variável aleatória, X, tem função de densidade,

$$f(x \mid \theta) = (1/\theta)e^{-x/\theta}, \; x > 0, \; \theta > 0.$$

Pretende ensaiar-se H_0: $\theta = 1$ contra H_1: $\theta = 2$. A função de distribuição de $Y = f(X \mid 2)/f(X \mid 1)$, quando $\theta = 1$,

$$\begin{aligned} P_1(Y \leqq c) &= P_1[f(X \mid 2) \leqq c f(X \mid 1)] \\ &= P_1(X \leqq 2 \log 2c), \end{aligned}$$

é dada pela expressão,

$$P_1(Y \leqq c) = \begin{cases} 0 & \text{para} \quad 0 \leqq c < 1/2, \\ 1 - (1/4c^2) & \text{para} \quad c \geqq 1/2. \end{cases}$$

Fixe-se $\alpha = 0{,}05$; com $1 - \alpha = 0{,}95$ existe c_0 tal que $P_1(Y \leqq c_0) = 0{,}95$, pois de facto, $1 - (1/4c_0^2) = 0{,}95$, implica $c_0 = \sqrt{5} \cong 2{,}24$. Portanto, por (7.25), $\gamma = 0$, e o melhor teste de tamanho 0,05 é,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad f(x \mid 2) > 2{,}24 \; f(x \mid 1), \\ 0 & \text{se} \quad f(x \mid 2) \leqq 2{,}24 \; f(x \mid 1). \end{cases}$$

É fácil verificar que $f(x \mid 2) > 2{,}24 \; f(x \mid 1)$ quando $x > 3$; assim, o melhor teste de tamanho 0,05 pode escrever-se na forma,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad x > 3, \\ 0 & \text{se} \quad x \leqq 3. \end{cases}$$

Do ponto de vista prático importa notar que no caso presente não há necessidade de determinar a função de distribuição, $P_1(Y \leqq c)$, para calcular c_0. Com efeito, $f(x \mid 2) > c_0 \; f(x \mid 1)$ implica $x > c_0'$, e c_0' pode obter-se a partir do tamanho do teste, α.

$$\alpha = E_1\{\phi(X)\} = \int_{c_0'}^{\infty} f(x \mid 1)\, dx = \int_{c_0'}^{\infty} e^{-x}\, dx,$$

donde $e^{-c_0'} = \alpha$; com $\alpha = 0{,}05$ sai $c_0' = 3$.

Para obter o conjunto S_0^* considerem-se os melhores testes de tamanho α, $0 \leqq \alpha \leqq 1$, que são da forma,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad x > c', \\ 0 & \text{se} \quad x \leqq c', \end{cases}$$

com $c' \geqq 0$, $e^{-c'} = \alpha$. Assim,

$$\begin{aligned} \alpha &= E_1\{\phi(X)\} = e^{-c'}, \\ \beta &= E_2\{1 - \phi(X)\} = \int_0^{c'} (1/2) e^{-x/2}\, dx = 1 - e^{-c'/2} \\ &= 1 - \sqrt{\alpha}. \end{aligned}$$

A equação da fronteira inferior é $\beta = 1 - \sqrt{\alpha}$ e por simetria obtém-se imediatamente S_0^* — veja Fig. 7.5.

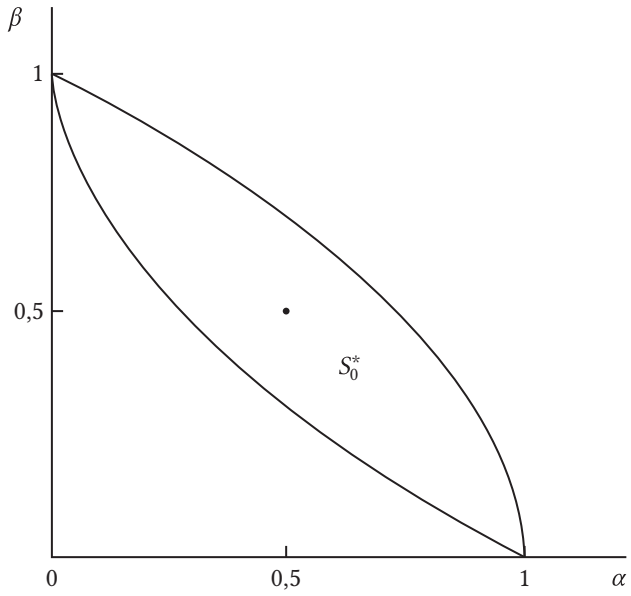


Fig. 7.5

□

Exemplo 7.5 — Com $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D., $X_i \sim N(\theta, 1)$, para ensaiar H_0: $\theta = \theta_0$ contra H_1: $\theta = \theta_1, \theta_1 > \theta_0$, através do Lema de Neyman-Pearson, verifica-se que o melhor teste de tamanho α é da forma,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \overline{x} > c', \\ 0 & \text{se} \quad \overline{x} \leqq c', \end{cases} \qquad [A]$$

pois,

$$f(\mathbf{x}\,|\,\theta_1)/f(\mathbf{x}\,|\,\theta_0) > c \text{ implica } \overline{x} > c';^{2}$$

o valor de c' determina-se facilmente a partir de α e da função de densidade de $\overline{X} \sim N(\theta, 1/N)$. Assim,

$$\alpha = E_{\theta_0}\{\phi(\mathbf{X})\} = P_{\theta_0}(\overline{X} > c') = 1 - \Phi[\sqrt{N}(c' - \theta_0)];$$

a probabilidade de um erro de 2.ª espécie é,

$$\beta = E_{\theta_1}\{1 - \phi(\mathbf{X})\} = P_{\theta_1}(\overline{X} \leqq c') = \Phi[\sqrt{N}(c' - \theta_1)].$$

Da expressão de $E_{\theta_0}\{\phi(\mathbf{X})\}$ obtém-se,

$$c' = \theta_0 + (1/\sqrt{N})\ \Phi^{-1}(1 - \alpha),$$

[2] A casualização não é necessária por $\overline{X}$ ser variável aleatória contínua.

onde $\Phi^{-1}(u)$ é a função inversa de $\Phi(u)$, função de distribuição da $N(0,1)$.[3] Conclui-se daqui que c' não depende da hipótese alternativa, H_1; isto é, qualquer que seja $\theta_1 > \theta_0$, o teste é sempre da forma [A] variando c' apenas com θ_0, com a dimensão da amostra N e com o tamanho do ensaio α. Este aspecto é importante e é adiante retomado. Por outro lado, substituindo c' na expressão de β, vem,

$$\beta = \Phi[\Phi^{-1}(1-\alpha) - \sqrt{N}(\theta_1 - \theta_0)],$$

equação da fronteira inferior de S_0^*. □

Em resumo, o procedimento usual no ensaio de H_0 contra H_1 consiste em atribuir um valor a α — correntemente, 0,01 ou 0,05 — fixando a probabilidade com que o decisor está disposto a cometer um erro de 1.ª espécie, e em procurar o melhor teste de tamanho α, isto é, o teste que minimiza a probabilidade de cometer um erro de 2.ª espécie. Na prática, com $\alpha = 0{,}01$ ou 0,05, raramente sucede sair $\beta = E_{\theta_1}\{1 - \phi(X)\} = 0$; se $\beta > 0$, o melhor teste de tamanho α é admissível.

Veja-se, seguidamente, o que acontece com as soluções minimax e Bayes.

Com $\Theta = \{\theta_0, \theta_1\}$, faça-se, em (7.13), $L_0(\theta_0) = L_1(\theta_1) = 0$, $L_0(\theta_1) = K_0$, $L_1(\theta_0) = K_1$, isto é, considere-se a função perca «0–K_i»; vem,

$$R(\theta_0, \phi) = K_1 E_{\theta_0}\{\phi(X)\},$$
$$R(\theta_1, \phi) = K_0 E_{\theta_1}\{1 - \phi(X)\}.$$

O teste minimax, $\tilde{\phi}$, deve verificar,

$$\max[R(\theta_0, \tilde{\phi}), R(\theta_1, \tilde{\phi})] \leqq \max[R(\theta_0, \phi), R(\theta_1, \phi)],$$

para todo o $\phi \in \Phi$.

Teorema 7.4 — Para ensaiar H_0: $\theta = \theta_0$ contra H_1: $\theta = \theta_1$, o teste minimax é definido por,

$$\tilde{\phi}(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,\theta_1) > C f(x\,|\,\theta_0), \\ \omega & \text{se} \quad f(x\,|\,\theta_1) = C f(x\,|\,\theta_0), \\ 0 & \text{se} \quad f(x\,|\,\theta_1) < C f(x\,|\,\theta_0), \end{cases} \tag{7.27}$$

onde C é determinado pela condição, $R(\theta_0, \tilde{\phi}) = R(\theta_1, \tilde{\phi})$.

Dem. Seja $\phi \in \Phi$ um teste arbitrário. Se $R(\theta_0, \tilde{\phi}) < R(\theta_0, \phi)$, tem-se,

$$\max[R(\theta_0, \tilde{\phi}), R(\theta_1, \tilde{\phi})] = R(\theta_0, \tilde{\phi}) < \max[R(\theta_0, \phi), R(\theta_1, \phi)].$$

[3] Não confundir Φ tal como empregado no presente exemplo com qualquer classe de testes.

Se, $R(\theta_0, \tilde{\phi}) \geqq R(\theta_0, \phi)$, vem,

$$E_{\theta_0}\{\tilde{\phi}(X)\} > E_{\theta_0}\{\phi(X)\};$$

mas, pelo Lema de Neyman-Pearson, $\tilde{\phi}$, sendo da forma (7.18), é o melhor teste do seu tamanho para ensaiar H_0 contra H_1. Logo, necessariamente,

$$E_{\theta_1}\{1 - \phi(X)\} \geqq E_{\theta_1}\{1 - \tilde{\phi}(X)\},$$

isto é, $R(\theta_1, \phi) \geqq R(\theta_1, \tilde{\phi})$. Portanto,

$$\max[R(\theta_0, \tilde{\phi}), R(\theta_1, \tilde{\phi})] = R(\theta_1, \tilde{\phi}) \leqq R(\theta_1, \phi) \leqq \max[R(\theta_0, \phi), R(\theta_1, \phi)],$$

e $\tilde{\phi}$ é minimax. □□

Exemplo 7.5 — *Continuação.* O teste minimax é da forma,

$$\tilde{\phi}(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \overline{x} > C', \\ 0 & \text{se} \quad \overline{x} \leqq C', \end{cases}$$

onde C' é determinado pela condição, $R(\theta_0, \tilde{\phi}) = R(\theta_1, \tilde{\phi})$, isto é,

$$K_1 E_{\theta_0}\{\tilde{\phi}(\mathbf{X})\} = K_0 E_{\theta_1}\{1 - \tilde{\phi}(\mathbf{X})\},$$

ou seja,[4]

$$K_1\{1 - \Phi[\sqrt{N}(C' - \theta_0)]\} = K_0 \Phi[\sqrt{N}(C' - \theta_1)];$$

dados, N, θ_0 , θ_1, K_0 , e K_1, o valor de C' obtém-se facilmente com uma tabela da Normal. A casualização não é necessária, donde $\omega = 0$. □

Exemplo 7.3 — *Continuação.* Verificando o teste minimax a condição,

$$K_1 E_{\theta_0}\{\tilde{\phi}(X)\} = K_0 E_{\theta_1}\{1 - \tilde{\phi}(X)\},$$

ou seja,

$$K_1 \alpha = K_0 \beta,$$

tem-se, considerando a equação da fronteira inferior de S_0^*,

$$\alpha + \beta = 0{,}5, \ \alpha > 0, \ \beta > 0,$$

que $\tilde{\phi}$ é o melhor teste de tamanho $\alpha = 0{,}5K_0/(K_0 + K_1)$, ou, por [B],

$$\tilde{\phi}(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,0{,}5) > f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 2 < x \leqq 2{,}5, \\ K_0/(K_0 + K_1), & \text{se} \quad f(x\,|\,0{,}5) = f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 1{,}5 \leqq x \leqq 2, \\ 0 & \text{se} \quad f(x\,|\,0{,}5) < f(x\,|\,0) \quad \text{ou} \quad \text{se} \quad 1 \leqq x < 1{,}5. \end{cases}$$

□

[4] Aqui Φ designa mais uma vez a distribuição da $N(0, 1)$.

Teorema 7.5 — Para ensaiar H_0: $\theta = \theta_0$ contra H_1: $\theta = \theta_1$, o teste Bayes contra a distribuição a priori, $\mathbf{h} = (\zeta, 1-\zeta)$, é definido por,

$$\phi_h(x) = \begin{cases} 1 & \text{se} \quad f(x\,|\,\theta_1) > K f(x\,|\,\theta_0), \\ 0 & \text{se} \quad f(x\,|\,\theta_1) \leqq K f(x\,|\,\theta_0), \end{cases} \tag{7.28}$$

onde,

$$K = \frac{\zeta K_1}{(1-\zeta)K_0}. \tag{7.29}$$

Dem. Mantendo a função perca na forma «0–K_i», anteriormente considerada, e recordando que na pesquisa das soluções Bayes não há necessidade de casualizar, tem-se que o teste Bayes contra **h** deve minimizar,

$$\begin{aligned} R(\mathbf{h}, \phi) &= \zeta R(\theta_0, \phi) + (1-\zeta) R(\theta_1, \phi) \\ &= \zeta K_1 E_{\theta_0}\{\phi(X)\} + (1-\zeta) K_0 E_{\theta_1}\{1 - \phi(X)\} \\ &= (1-\zeta)K_0 + \int_{\mathfrak{X}} \phi(x)[\zeta K_1 f(x\,|\,\theta_0) - (1-\zeta) K_0 f(x\,|\,\theta_1)]\, dx; \end{aligned}$$

da análise da última expressão saiem imediatamente (7.28) e (7.29). □□

Exemplo 7.5 — *Continuação.* Para determinar o teste Bayes, que é necessariamente da forma [A], basta notar que a desigualdade,

$$\frac{f(\mathbf{x}\,|\,\theta_1)}{f(\mathbf{x}\,|\,\theta_0)} > \frac{\zeta K_1}{(1-\zeta)K_0},$$

implica,

$$\overline{x} > \frac{\log[\zeta K_1/(1-\zeta)K_0]}{N(\theta_1 - \theta_0)} + \frac{\theta_0 + \theta_1}{2}.$$

Note-se que o risco a posteriori é, no caso presente, dado por,

$$\begin{aligned} r_{\mathbf{x}}(a_o) &= K_0 h(\theta_1\,|\,\mathbf{x}) = K_0(1-\zeta) f(\mathbf{x}\,|\,\theta_1)/[\zeta f(\mathbf{x}\,|\,\theta_0) + (1-\zeta) f(\mathbf{x}\,|\,\theta_1)], \\ r_{\mathbf{x}}(a_1) &= K_1 h(\theta_0\,|\,\mathbf{x}) = K_1 \zeta f(\mathbf{x}\,|\,\theta_0)/[\zeta f(\mathbf{x}\,|\,\theta_0) + (1-\zeta) f(\mathbf{x}\,|\,\theta_1)], \end{aligned}$$

e que, portanto, (7.28) e (7.29) equivalem a,

$$\phi_h(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad r_{\mathbf{x}}(a_0) > r_{\mathbf{x}}(a_1), \\ 0 & \text{se} \quad r_{\mathbf{x}}(a_0) \leqq r_{\mathbf{x}}(a_1). \end{cases}$$

Evidentemente, a casualização é irrelevante pois estaria a fazer-se após a observação de pontos **x** para os quais as duas acções têm o mesmo risco a posteriori. □

Exemplo 7.6 — Suponha-se que a qualidade de um produto é representada por uma variável aleatória $X \sim N(\theta, 1)$. Quando o processo de produção se encontra sob controlo o valor de θ deve ser θ_0. Em cada hora é observada uma unidade do produto escolhida ao acaso; ao fim de N horas a situação é descrita por $X_1, X_2, \ldots, X_N$, X_i independentes.

Suspeita-se que ao fim de k horas, $1 < k < N$, ocorreu uma avaria traduzida pelo deslocamento da média θ de θ_0 para $\theta_1 > \theta_0$. A perca ocasionada por tal deslocamento é avaliada em $(\theta_1 - \theta_0)$ por hora e se o mesmo de facto ocorreu a laboração deve ser interrompida para rectificação. Por outro lado, se o deslocamento não ocorreu e a laboração é suspensa, produz-se uma perca avaliada em L.

Considerando a distribuição a priori, $\mathbf{h} = (\zeta, 1-\zeta)$ — não ocorrência ou ocorrência de deslocamento — pretende-se encontrar, considerando as N observações, o teste Bayes contra $\mathbf{h}$ para ensaiar,

$$H_0\colon\ X_1, X_2, \ldots, X_N \text{ I.I.D. segundo } N(\theta_0, 1),$$

contra,

$$\begin{aligned} H_1\colon\ & X_1, \ldots, X_k \text{ I.I.D. segundo } N(\theta_0, 1) \text{ e} \\ & X_{k+1}, X_{k+2}, \ldots, X_N \text{ I.I.D. segundo } N(\theta_1, 1), k \text{ dado.} \end{aligned}$$

Tem-se,

$$\begin{aligned} \frac{f(\mathbf{x}\,|\,H_1)}{f(\mathbf{x}\,|\,H_0)} &= \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{k}(x_i-\theta_0)^2 + \sum_{i=k+1}^{N}(x_i-\theta_1)^2\right]\right\} \Big/ \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta_0)^2\right\} \\ &= \exp\left\{\theta_0\sum_{i=1}^{k}x_i + \theta_1\sum_{i=k+1}^{N}x_i - \frac{k}{2}\theta_0^2 - \frac{(N-k)}{2}\theta_1^2\right\} \Big/ \exp\left\{\theta_0\sum_{i=1}^{N}x_i - \frac{N}{2}\theta_0^2\right\} \\ &= \exp\left\{(\theta_1-\theta_0)(N-k)\overline{x}^*_{N-k} - \frac{(N-k)}{2}(\theta_1^2-\theta_0^2)\right\}, \end{aligned}$$

onde,

$$\overline{x}^*_{N-k} = \sum_{i=k+1}^{N}\frac{x_i}{(N-k)},$$

é a média das últimas $N-k$ observações.

Assim, sai $\phi_h(\mathbf{x}) = 1$ [rejeita-se H_0], quando,

$$\overline{x}^*_{N-k} > \frac{\log[\zeta L/(1-\zeta)(N-k)(\theta_1-\theta_0)]}{(N-k)(\theta_1-\theta_0)} + \frac{\theta_0+\theta_1}{2};$$

comparando com o resultado do Ex. 7.5 conclui-se que tudo se passa como no ensaio de $H_0\colon\ \theta = \theta_0$ contra $H_1\colon\ \theta = \theta_1$ com base nas últimas $N-k$ observações, isto é, as primeiras k observações são irrelevantes como aliás seria de esperar [Zacks (1981)]. □

7.3 Testes uniformemente mais potentes

O ensaio de hipóteses simples contra alternativas simples possui sobretudo interesse teórico, pois, em geral, o espaço do parâmetro é intervalo de $\mathbf{R}$ ou $\mathbf{R}^k$ e não conjunto com dois elementos, $\Theta = \{\theta_0, \theta_1\}$. A enorme importância do Lema de Neyman-Pearson deve-se precisamente ao facto de as suas aplicações transcenderem largamente esse caso restrito.

Para estudar o ensaio de hipótese composta, seja H_0: $\theta \in \Theta_0$, contra alternativa composta, seja H_1: $\theta \in \Theta_1$, há que adaptar alguns conceitos estudados nas secções anteriores. A maior complexidade da análise obriga, por vezes, a deixar a função perca na penumbra.

Função de decisão aleatória ou teste, $\phi \in \Phi$, para ensaiar H_0 contra H_1, tem tamanho α, com $0 \leqq \alpha \leqq 1$, se,

$$\sup_{\theta \in \Theta_0} E_\theta\{\phi(X)\} = \alpha; \tag{7.30}$$

um teste, $\phi \in \Phi$, diz-se uniformemente mais potente (UMP) de tamanho α se, além de verificar (7.30),

$$E_\theta\{\phi(X)\} \geqq E_\theta\{\phi'(X)\} \text{ para } \theta \in \Theta_1, \tag{7.31}$$

qualquer que seja o teste de tamanho α, $\phi' \in \Phi$.

Os testes UMP só existem em casos especiais em virtude de a condição (7.31) ser demasiado exigente. Por exemplo, supondo, $\Theta = (-\infty, +\infty)$, $\Theta_0 = (-\infty, \theta_0]$, $\Theta_1 = (\theta_0, +\infty)$, a Fig. 7.6 mostra a posição que a função potência do teste UMP, ϕ, deve ter em em relação a qualquer outro teste, ϕ', tendo como ϕ tamanho α.

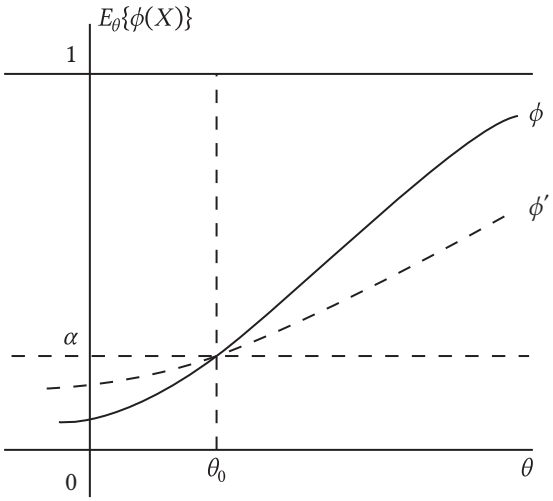


Fig. 7.6

Se H_0 e H_1 são hipóteses compostas e se $\Theta = \mathbf{R}$, a situação de mais fácil tratamento é aquela em que,

$$H_0\colon \theta \leqq \theta_0 \text{ contra } H_1\colon \theta > \theta_0; \tag{7.32}$$

diz-se neste caso que o ensaio é unilateral [o tratamento de (7.32) aplica-se, com modificação óbvia, a $H_0\colon \theta \geqq \theta_0$ contra $H_1\colon \theta < \theta_0$].

Infelizmente, nem mesmo para os ensaios unilaterais existem, em condições gerais, testes UMP. Este facto leva a explorar duas vias de investigação:

1) Procurar testes UMP depois de impor restrições sobre a família $\mathcal{F} = \{f(x\,|\,\theta) : \theta \in \Theta\}$.
2) Procurar testes que sejam UMP dentro de uma subclasse de Φ.

A primeira via conduz a famílias de distribuições com razão de verosimilhança monótona (RVM).

A família, $\{f(x\,|\,\theta) : \theta \in \Theta\}$, em que Θ é um intervalo de $\mathbf{R}$, tem RVM, quando,

$$f(x\,|\,\theta_2)/f(x\,|\,\theta_1), \quad \text{com } \theta_2 > \theta_1, \tag{7.33}$$

é função não decrescente de x. O que interessa de facto é a monotonicidade de $f(x\,|\,\theta_2)/f(x\,|\,\theta_1)$, pois se esta função é não crescente pode reduzir-se à primeira forma tomando $Y = -X$ ou operando a reparametrização $\omega = -\theta$.

São definições equivalentes,

$$f(x_2\,|\,\theta_2)/f(x_2\,|\,\theta_1) \geqq f(x_1\,|\,\theta_2)/f(x_1\,|\,\theta_1), \tag{7.34}$$

com $\theta_2 > \theta_1$ e $x_2 > x_1$, ou ainda,

$$f(x_2\,|\,\theta)/f(x_1\,|\,\theta), \quad \text{com } x_2 > x_1, \tag{7.35}$$

função não decrescente de θ.

Como $f(x\,|\,\theta)$ pode ser igual a zero para alguns valores de x e de θ, é mais conveniente definir RVM pela condição,

$$x_2 > x_1,\ \theta_2 > \theta_1 \ \Rightarrow\ f(x_2\,|\,\theta_2)f(x_1\,|\,\theta_1) - f(x_2\,|\,\theta_1)f(x_1\,|\,\theta_2) \geqq 0, \tag{7.36}$$

ou ainda,

$$x_2 > x_1,\ \theta_2 > \theta_1 \ \Rightarrow\ \begin{vmatrix} f(x_1\,|\,\theta_1) & f(x_1\,|\,\theta_2) \\ f(x_2\,|\,\theta_1) & f(x_2\,|\,\theta_2) \end{vmatrix} \geqq 0. \tag{7.37}$$

Exemplo 7.7 — A família Binomial tem RVM:

$$\frac{\binom{N}{x}\theta_2^x(1-\theta_2)^{N-x}}{\binom{N}{x}\theta_1^x(1-\theta_1)^{N-x}} = \left[\frac{(1-\theta_2)}{(1-\theta_1)}\right]^N \left[\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right]^x,$$

é função não decrescente de x quando,

$$\left[\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right] > 1 \text{ ou } \theta_2 > \theta_1.$$

□

Exemplo 7.8 — A família Normal com média θ e variância conhecida, seja $\sigma^2 = 1$, tem RVM:

$$(2\pi)^{-1/2}\exp\{-(x-\theta_2)^2/2\}/(2\pi)^{1/2}\exp\{-(x-\theta_1)^2/2\} =$$
$$= \exp\{-(\theta_2^2-\theta_1^2)/2\}\exp\{x(\theta_2-\theta_1)\},$$

e, com $\theta_2 > \theta_1$, $\exp\{x(\theta_2-\theta_1)\}$ é função não decrescente de x. □

Exemplo 7.9 — A família Cauchy não tem RVM, pois,

$$\frac{f(x\,|\,\theta_2)}{f(x\,|\,\theta_1)} = \frac{1+(x-\theta_1)^2}{1+(x-\theta_2)^2},$$

converge para 1 quando $x \to \pm\infty$. □

O conceito de RVM apresentado refere-se a uma variável aleatória, X, escalar. Se $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ e $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, e se,

$$\frac{f(\mathbf{x}\,|\,\theta_2)}{f(\mathbf{x}\,|\,\theta_1)},\ \theta_2 > \theta_1, \tag{7.38}$$

é função não decrescente da estatística, $T(\mathbf{x})$, diz-se que $\{f(\mathbf{x}\,|\,\theta) : \theta \in \Theta\}$ tem RVM em $T(\mathbf{x})$.

Exemplo 7.10 — Se X_i I.I.D., $X_i \sim N(\theta, 1)$, $i = 1, 2, ..., N$, é de verificação imediata que $f(\mathbf{x}\,|\,\theta)$ tem RVM em $T(\mathbf{x}) = \Sigma x_i$ ou em $T^*(\mathbf{x}) = \overline{x}$ (estatísticas suficientes). □

Exemplo 7.11 — Se X_i I.I.D., $i = 1, 2, \ldots, N$, com distribuição na família Exponencial, tem-se,

$$f(\mathbf{x}\,|\,\theta) = [C(\theta)]^N \exp\{Q(\theta)T(\mathbf{x})\}\Pi H(x_i);$$

logo, se $Q(\theta)$ é função não decrescente de θ, a família tem RVM em $T(\mathbf{x})$. □

A monotonicidade da razão de verosimilhanças traduz uma preservação de ordem entre os valores do parâmetro e os valores de X ou de $T(\mathbf{X})$ no caso vectorial.

Suponha-se que X representa o número de peças defeituosas numa amostra casual de N peças e que θ, $0 \leqq \theta \leqq 1$, designa a probabilidade de obter uma peça defeituosa. Evidentemente, a valores pequenos de θ correspondem, mais provavelmente, valores pequenos de X e a valores grandes de θ correspondem, mais provavelmente, valores grandes de X. Nos problemas de inferência interessa o caminho inverso, isto é, poder inferir que com valores pequenos (valores grandes) de X estão associados valores pequenos (valores grandes) de θ. Esse caminho está aberto se X tem distribuição com RVM, precisamente o que sucede quando X tem distribuição Binomial.

Com a distribuição de Cauchy, que não tem RVM, ainda se pode afirmar que a valores pequenos (valores grandes) de θ correspondem, com maior probabilidade, valores pequenos (valores grandes) de X. No entanto, devido ao comportamento das abas da Cauchy, não é legítimo inferir que valores pequenos (valores grandes) de X estão associados com valores pequenos (valores grandes) de θ.

As consequências da monotonicidade da razão de verosimilhanças podem apreciar-se, também, no contexto do Lema de Neyman-Pearson. Para ensaiar H_0: $\theta = \theta_0$ contra H_1: $\theta = \theta_1$, $\theta_1 > \theta_0$, a determinação do melhor teste está esquematizada na Fig. 7.7 [quando $f(x\,|\,\theta)$ tem RVM] e na Fig. 7.8 [quando $f(x\,|\,\theta)$ não tem RVM].

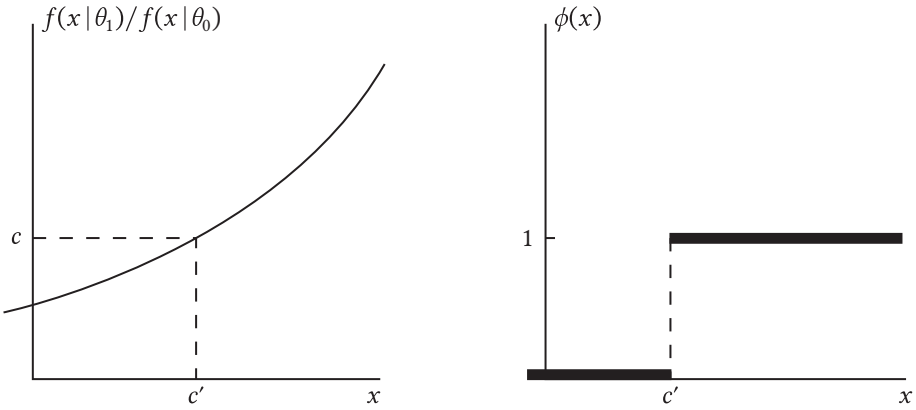


Fig. 7.7

Repare-se que no primeiro caso [quando $f(x\,|\,\theta)$ tem RVM] a função teste, $\phi(x)$ é não decrescente (monótona). A proposição seguinte, devida a Karlin, que adiante se aplica, é outra das consequências da RVM particularmente relevante no estudo da função potência dos testes unilateriais.

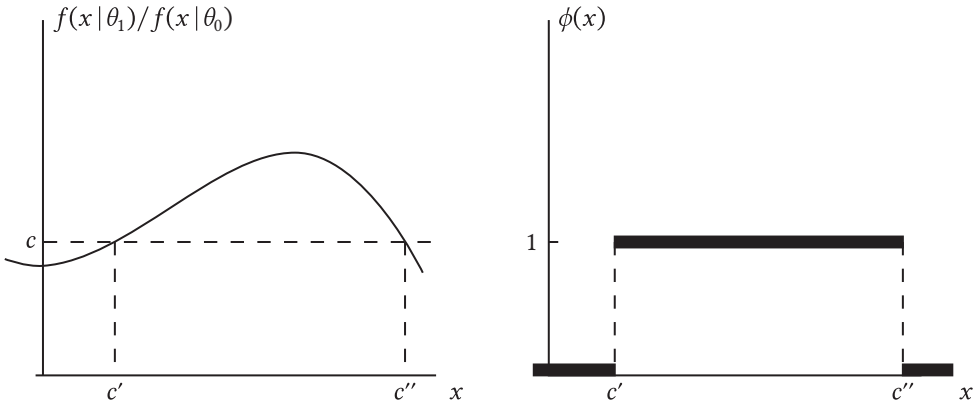


Fig. 7.8

Teorema 7.6 — Suponha-se que $\mathcal{F} = \{f(x\,|\,\theta) : -\infty < \theta < +\infty\}$ tem RVM em x. Se $\psi(x)$ é função não decrescente de x, então $E_\theta\{\psi(X)\}$ é função não decrescente de θ.

Dem. Pode ver-se em Zacks (1981) ou em Karlin (1957). □□

O teorema seguinte, devido a Karlin e Rubin (1956), estabelece em que condições existem testes UMP para ensaios unilateriais:

Teorema 7.7 — Se o vector aleatório, $\mathbf{X}$, tem distribuição com RVM em $T(\mathbf{x})$, então, para ensaiar H_0: $\theta \leqq \theta_0$ contra H_1: $\theta > \theta_0$, os testes da forma,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad T(\mathbf{x}) > t_0, \\ \gamma & \text{se} \quad T(\mathbf{x}) = t_0, \\ 0 & \text{se} \quad T(\mathbf{x}) < t_0, \end{cases} \tag{7.39}$$

possuem as seguintes propriedades:

(i) São UMP de tamanho $E_{\theta_0}\{\phi(\mathbf{X})\} = \alpha$, desde que seja $E_{\theta_0}\{\phi(\mathbf{X})\} > 0$.
(ii) Para todo o α, $0 \leqq \alpha \leqq 1$, e todo o θ_0, $-\infty < \theta_0 < +\infty$, existem números t_0 e γ, $-\infty \leqq t_0 \leqq +\infty$, $0 \leqq \gamma \leqq 1$, tais que o teste definido por (7.39) é UMP de tamanho α para ensaiar H_0 contra H_1.
(iii) A função potência do teste ϕ definido por (7.39), $E_\theta\{\phi(\mathbf{X})\}$, é não decrescente em θ.

Dem. Sejam, $\theta_0, \theta_1 \in \Theta$, $\theta_1 > \theta_0$. Para ensaiar, H_0^*: $\theta = \theta_0$ contra H_1^*: $\theta = \theta_1$, o Lema de Neyman-Pearson ensina que os testes da forma (7.18) são melhores do seu tamanho desde que seja $0 \leqq c < \infty$ e que o teste com $c = \infty$, (7.19), é o melhor

de tamanho zero. Como, por hipótese, $f(\mathbf{x}\,|\,\theta)$ tem RVM em $T(\mathbf{x})$, tem-se que os testes da forma (7.39) são também da forma (7.18) desde que seja $E_{\theta_0}\{\phi(\mathbf{X})\} > 0$, isto é, $c < \infty$; de facto, nesse caso,

$$T(\mathbf{x}) \lesseqgtr t \Leftrightarrow f(\mathbf{x}\,|\,\theta_1)/f(\mathbf{x}\,|\,\theta_0) \lesseqgtr c.$$

Os testes da forma (7.39) são, portanto, os melhores do seu tamanho para ensaiar H_0^* contra H_1^*, desde que seja $E_{\theta_0}\{\phi(\mathbf{X})\} > 0$.

Pelo teorema de Karlin, a função potência de ϕ é não decrescente

$$E_\theta\{\phi(\mathbf{X})\} \leqq E_{\theta_0}\{\phi(\mathbf{X})\} = \alpha,\ \theta \leqq \theta_0. \tag{7.40}$$

Como, porém, ϕ não depende de θ_1 (depende apenas de α e de θ_0: veja-se o caso do Ex. 7.5 para não mencionar outros), segue-se que ϕ é UMP de tamanho α para ensaiar $\theta = \theta_0$ contra $\theta > \theta_0$. Assim, ϕ é UMP dentro da classe de testes ϕ'' que verificam,

$$E_{\theta_0}\{\phi''(\mathbf{X})\} \leqq E_{\theta_0}\{\phi(\mathbf{X})\} = \alpha. \tag{7.41}$$

Ora, a classe de testes que satisfazem (7.40) está contida na classe de testes que satisfazem (7.41). Em consequência, ϕ, que é UMP na classe mais ampla que cumpre (7.41) é também UMP na classe mais restrita que cumpre (7.40). Enfim, desde que seja $\alpha > 0$, ϕ é UMP de tamanho α para ensaiar H_0: $\theta \leqq \theta_0$ contra H_1: $\theta > \theta_0$.

Provados (i) e (iii) falta provar (ii). Notando que,

$$E_{\theta_0}\{\phi(\mathbf{X})\} = P_{\theta_0}[T(\mathbf{X}) > t_0] + \gamma P_{\theta_0}[T(\mathbf{X}) = t_0],$$

a determinação de t_0 e de γ processa-se tal como a determinação de c_0 e de γ no Lema de Neyman-Pearson [reveja-se o que se passa a partir de (7.22) e substitua-se a função de distribuição de $Y = f(X\,|\,\theta_1)/f(X\,|\,\theta_0)$ pela distribuição de $T(\mathbf{X})$]. □□

Exemplo 7.12 — Os casos tratados nos Ex. 7.2, 7.3, 7.4 e 7.5 podem alargar-se imediatamente ao ensaio da hipótese unilateral: H_0: $\theta \leqq \theta_0$ contra a alternativa H_1: $\theta > \theta_0$, sendo fácil concluir que não há, em geral, qualquer alteração na estrutura dos testes que sendo anteriormente melhores de tamanho α passam a ser UMP de tamanho α, desde que seja $0 < \alpha \leqq 1$. Como a demonstração do teorema anterior procura esclarecer as dificuldades podem surgir quando $\alpha = 0$.

Retome-se o Ex. 7.3. Para ensaiar, H_0: $\theta \leqq 0$ contra H_1: $\theta > 0$, é evidente que os testes da forma,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad x > x_0, \\ 0 & \text{se} \quad x \leqq x_0, \end{cases} \tag{A}$$

têm tamanho $\alpha = 0$ desde que se tome $x_0 \geqq 2$. No entanto, só o teste com $x_0 = 2$ é UMP de tamanho zero. Para melhor esclarecimento comparam-se, na Fig. 7.9, as funções potência dos testes correspondentes a $x_0 = 2$ e $x_0 = 3$.

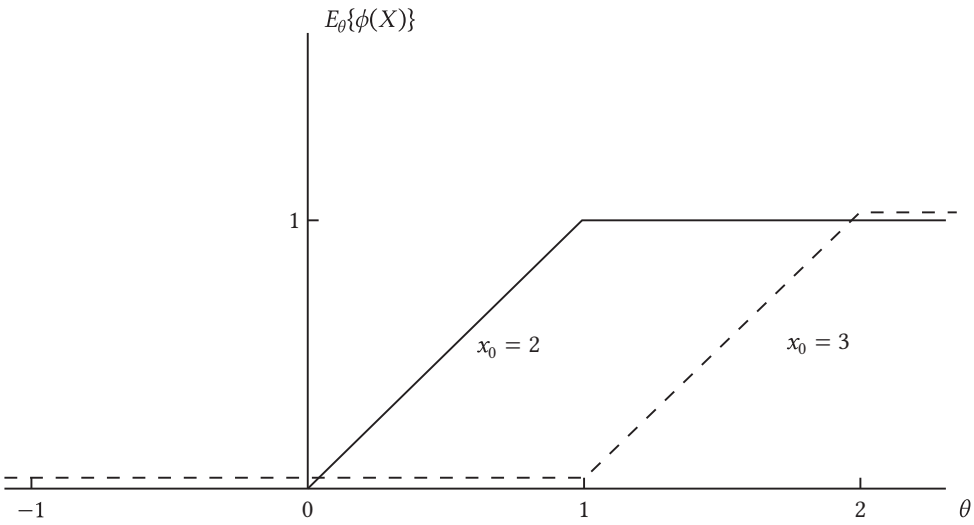


Fig. 7.9

Em resumo, para ensaiar H_0: $\theta \leqq 0$ contra H_1: $\theta > 0$, qualquer teste da forma [A] é UMP do seu tamanho se esse tamanho for maior do que zero. Os testes da forma [A] com tamanho zero não são em regra UMP; existe, no entanto, um teste da forma [A] que é UMP de tamanho zero. □

Exemplo 7.13 — Considere-se, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.l.D., $X_i \sim U(0, \theta)$, $\theta > 0$.

Tem-se,

$$f(\mathbf{x}\,|\,\theta) = (1/\theta^N) I_{(0,\theta)}[\max(x_i)];$$

com $\theta_2 > \theta_1$, a razão de verosimilhança vem dada por,

$$\frac{f(\mathbf{x}\,|\,\theta_2)}{f(\mathbf{x}\,|\,\theta_1)} = (\theta_1^N/\theta_2^N)\{I_{(0,\theta_2)}[\max(x_i)]/I_{(0,\theta_1)}[\max(x_i)]\}.$$

Com,

$$\begin{aligned}\Lambda(\mathbf{x}) &= I_{(0,\theta_2)}[\max(x_i)]/I_{(0,\theta_1)}[\max(x_i)] \\ &= \begin{cases} 1 & \text{se} \quad 0 < \max(x_i) < \theta_1, \\ \infty & \text{se} \quad \theta_1 \leqq \max(x_i) < \theta_2, \end{cases}\end{aligned}$$

e a convenção, $\Lambda(\mathbf{x}) = \infty$ se $\max(x_i) \geqq \theta_2$, conclui-se que $f(\mathbf{x}\,|\,\theta)$ tem RVM em $T(\mathbf{x}) = \max(x_i)$.

Para ensaiar H_0: $\theta \leqq \theta_0$ contra H_1: $\theta > \theta_0$, qualquer teste da forma,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \max(x_i) > t_0, \\ 0 & \text{se} \quad \max(x_i) \leqq t_0, \end{cases}$$

é UMP do seu tamanho se este for maior do que zero. Como $T = \max(X_i)$ tem função de densidade,

$$g(t\,|\,\theta) = \frac{Nt^{N-1}}{\theta^N},\ 0 < t < \theta,$$

tem-se que, para $0 < t_0 < \theta_0$, o tamanho do teste é,

$$E_{\theta_0}\{\phi(\mathbf{X})\} = \int_{t_0}^{\theta_0} \frac{Nt^{N-1}}{\theta_0^N}\,dt = 1 - (t_0/\theta_0)^N;$$

para $t_0 \geqq \theta_0$ o teste tem tamanho zero mas só com $t_0 = \theta_0$ é que é UMP. □

Na presente secção ainda não se fez qualquer emprego da função perca; para caminhar nesse sentido note-se o corolário do teorema anterior,

Teorema 7.8 — Se o vector aleatório $\mathbf{X}$ tem distribuição com RVM em $T(\mathbf{x})$, então, para qualquer teste $\phi' \in \Phi$ e para cada $\theta_0 \in \Theta$, existe um teste da forma (7.39) tal que,

$$E_\theta\{\phi(\mathbf{X})\} \geqq E_\theta\{\phi'(\mathbf{X})\} \quad \text{para} \quad \theta \geqq \theta_0, \tag{7.42}$$

$$E_\theta\{\phi(\mathbf{X})\} \leqq E_\theta\{\phi'(\mathbf{X})\} \quad \text{para} \quad \theta \leqq \theta_0, \tag{7.43}$$

Dem. Seja $\alpha = E_{\theta_0}\{\phi'(\mathbf{X})\}$, $0 \leqq \alpha \leqq 1$, e considere-se o teste da forma (7.39) UMP de tamanho α para ensaiar H_0: $\theta \leqq \theta_0$ contra H_1: $\theta > \theta_0$; tem-se imediatamente (7.42). Por simetria, o Teorema 7.7 implica que $1 - \phi(\mathbf{x})$ é UMP do seu tamanho para ensaiar H_0^*: $\theta \geqq \theta_0$ contra H_1^*: $\theta < \theta_0$, desde que seja $\alpha \neq 1$; portanto, tem-se (7.43). O caso $\alpha = 1$ obtém-se por simetria do caso $\alpha = 0$. □□

Se a_0 (aceitação de H_0) é a acção correcta quando $\theta < \theta_0$ e a_1 (rejeição de H_0) é a acção correcta quando $\theta > \theta_0$, parece razoável introduzir as seguintes relações para a função perca,

$$L_1(\theta) - L_0(\theta) \leqq 0 \quad \text{para } \theta > \theta_0, \tag{7.44}$$

$$L_1(\theta) - L_0(\theta) \geqq 0 \quad \text{para } \theta < \theta_0, \tag{7.45}$$

sem qualquer restrição quanto ao que se passa para $\theta = \theta_0$.

Teorema 7.9 — Se a função perca satisfaz (7.44) e (7.45) e se o vector aleatório, $\mathbf{X}$, tem distribuição com RVM em $T(\mathbf{x})$, a classe de testes (7.39) é essencialmente completa. Se o conjunto $\{x : f(x\,|\,\theta) > 0\}$ é independente de θ e se existem números $\theta_1, \theta_2 \in \Theta$, $\theta_1 \leqq \theta_0 \leqq \theta_2$, tais que

$$L_1(\theta_1) - L_0(\theta_1) > 0 \quad \text{e} \quad L_1(\theta_2) - L_0(\theta_2) < 0, \tag{7.46}$$

então os testes (7.39) são admissíveis.

Dem. Seja $\phi' \in \Phi$ um teste qualquer e $\phi \in \Phi$ um teste da forma (7.39) satisfazendo (7.42) para o valor θ_0 a que se referem (7.44)–(7.45). A diferença entre as funções risco, tal como definidas em (7.7), é,

$$R(\theta, \phi') - R(\theta, \phi) = [L_1(\theta) - L_0(\theta)][E_\theta\{\phi'(\mathbf{X})\} - E_\theta\{\phi(\mathbf{X})\}]; \tag{7.47}$$

se $\theta > \theta_0$, por (7.42) e (7.44), ambos os factores do segundo membro são não positivos; se $\theta < \theta_0$ ambos os factores são não negativos em virtude de (7.43) e (7.45); se $\theta = \theta_0$, $E_{\theta_0}\{\phi(\mathbf{X})\} = E_{\theta_0}\{\phi'(\mathbf{X})\}$. Portanto, é sempre, $R(\theta, \phi') \geqq R(\theta, \phi)$, e a classe de testes (7.39) é essencialmente completa.

Seja agora ϕ um teste da forma (7.39) e suponha-se que ϕ' é um teste para o qual $R(\theta, \phi') \leqq R(\theta, \phi)$ para todo o $\theta \in \Theta$; para provar que ϕ é admissível importa mostrar ser $R(\theta, \phi') \geqq R(\theta, \phi)$ para todo o $\theta \in \Theta$. De (7.42) e (7.43) sai, com θ_1 e θ_2 nas condições enunciadas,

$$E_{\theta_1}\{\phi'(\mathbf{X})\} - E_{\theta_1}\{\phi(\mathbf{X})\} \leqq 0; \; E_{\theta_2}\{\phi'(\mathbf{X})\} - E_{\theta_2}\{\phi(\mathbf{X})\} \geqq 0.$$

Se $E_{\theta_1}\{\phi(\mathbf{X})\} > 0$, ϕ é UMP do seu tamanho para ensaiar H_0': $\theta \leqq \theta_1$ contra a alternativa H_1': $\theta > \theta_1$ em virtude do Teorema 7.7; se $E_{\theta_1}\{\phi(\mathbf{X})\} = 0$, ϕ é ainda UMP do seu tamanho porque todos os testes de tamanho zero têm função potência constantemente nula quando $\{x : f(x\,|\,\theta) > 0\}$ não depende de θ. Como ϕ é UMP em todos os casos, tem-se $E_\theta\{\phi(\mathbf{X})\} \geqq E_\theta\{\phi'(\mathbf{X})\}$ para todo o $\theta > \theta_1$. Por simetria, vem $E_\theta\{\phi(\mathbf{X})\} \leqq E_\theta\{\phi'(\mathbf{X})\}$ para todo o $\theta < \theta_2$. Isso implica $R(\theta, \phi') \geqq R(\theta, \phi)$ como era preciso demonstrar. □□

Exemplo 7.14 — Com X_i I.I.D., $X_i \sim N(\theta, \sigma_0^2)$, $i = 1, 2, \ldots, N$, σ_0^2 conhecido, $f(\mathbf{x}\,|\,\theta)$ tem RVM monótona em

$$T(\mathbf{x}) = \frac{\Sigma x_i}{N} = \overline{x};$$

$\overline{X}$ é estatística suficiente, $\overline{X} \sim N(\theta, \sigma_0^2/N)$.

Para ensaiar H_0: $\theta \leqq \theta_0$ contra H_1: $\theta > \theta_0$ os testes da forma,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \overline{x} > t_0, \\ 0 & \text{se} \quad \overline{x} \leqq t_0, \end{cases} \tag{7.48}$$

são UMP de tamanho $\alpha > 0$ com t_0 determinado pela condição,

$$E_{\theta_0}\{\phi(\mathbf{X})\} = P_{\theta_0}(\overline{X} > t_0) = \alpha.$$

Suponha-se que a função perca é a representada na Fig. 7.10 e que se tem $\theta_0 = 31$, $\sigma_0^2 = 8$ e $N = 64$. Considerem-se os três testes, ϕ_1, ϕ_2, ϕ_3, correspondentes aos

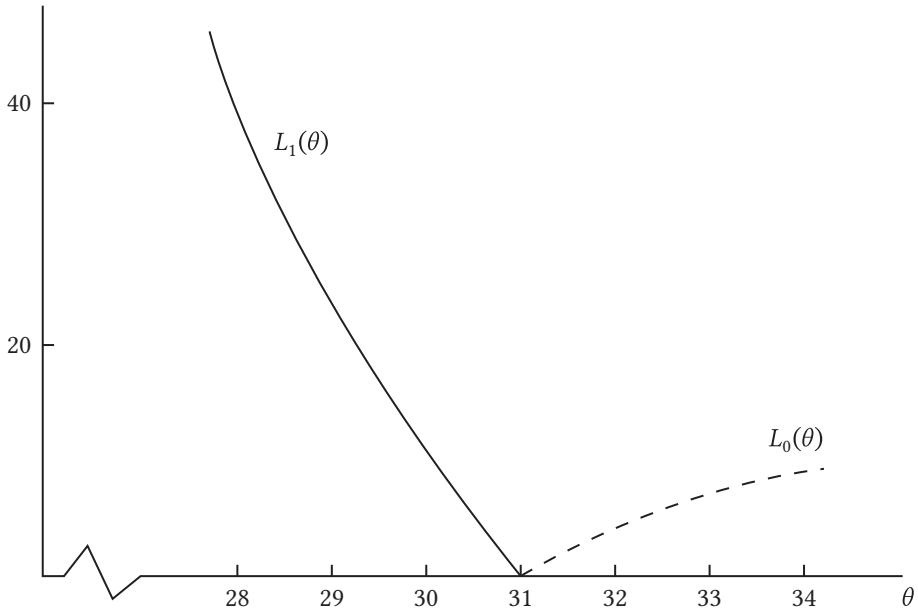


Fig. 7.10

valores $t_0 = 30{,}5, 31{,}25, 31{,}5$. É interessante comparar as funções risco, $R(\theta, \phi_i)$, $i = 1, 2, 3$. Para ϕ_1, tem-se,

$$R(\theta, \phi_1) = \begin{cases} L_1(\theta) E_\theta\{\phi_1(\mathbf{X})\} & \text{se} \quad \theta \leqq 31, \\ L_0(\theta) E_\theta\{1 - \phi_1(\mathbf{X})\} & \text{se} \quad \theta > 31, \end{cases}$$

donde,

$$R(\theta, \phi_1) = \begin{cases} L_1(\theta)\{1 - \Phi[\sqrt{64}(30{,}5 - \theta)/8]\} & \text{para } \theta \leqq 31, \\ L_0(\theta)\Phi[\sqrt{64}(30{,}5 - \theta)/8] & \text{para } \theta > 31. \end{cases}$$

Procedendo de modo semelhante para ϕ_2, e ϕ_3 obtém-se as funções risco representadas na Fig. 7.11.

Deixa-se como exercício a determinação do tamanho dos testes considerados: ϕ_1, ϕ_2, ϕ_3.

Repare-se que os testes da forma (7.48), além de formarem uma classe essencialmente completa, são admissíveis. De facto,

$$\{\mathbf{x} : f(\mathbf{x} \mid \theta) > 0\} = \mathbf{R}^N,$$

não depende de θ e a função perca verifica (7.46).

Os aspectos elementares do presente exemplo encontram-se expostos em Chernoff e Moses (1959).

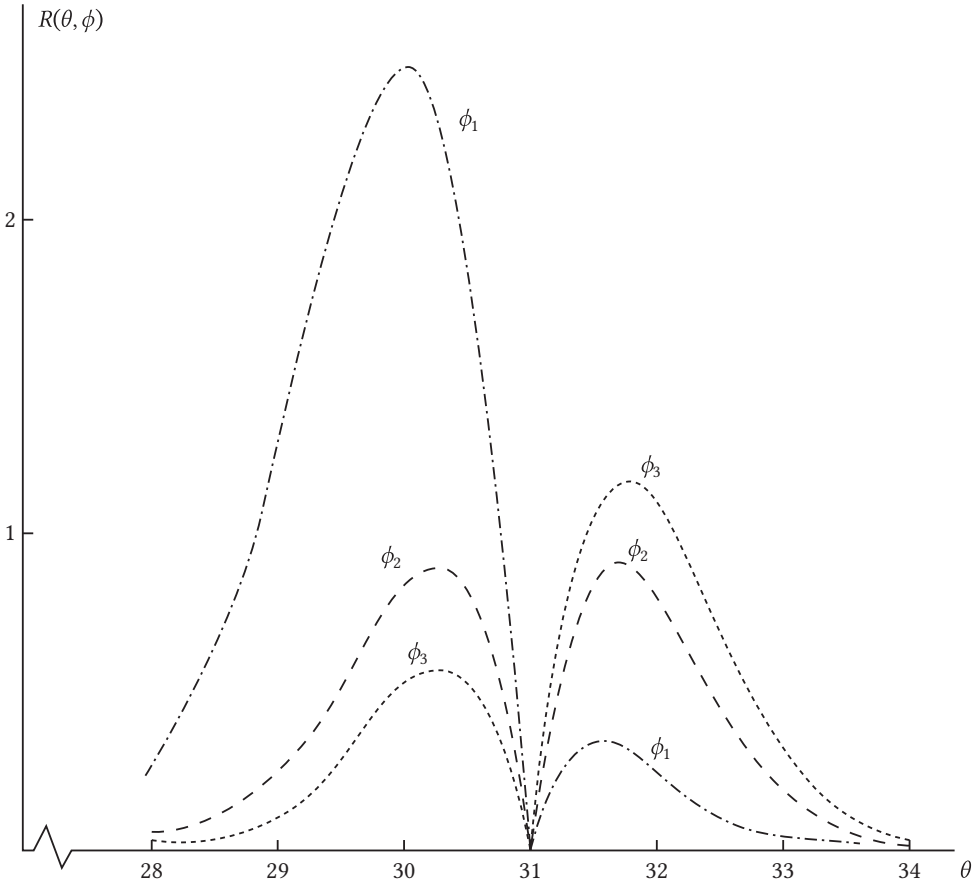


Fig. 7.11

□

7.4 Ensaios bilaterais

Os ensaios considerados na secção anterior são unilaterais. A situação complica-se quando se estudam ensaios bilaterais, como,

$$H_0\colon \theta = \theta_0 \ \text{ contra } \ H_1\colon \theta \neq \theta_0, \tag{7.49}$$

$$H_0\colon \theta_1 \leqq \theta \leqq \theta_2 \ \text{ contra } \ H_1\colon \theta < \theta_1 \ \text{ ou } \ \theta > \theta_2 \ (\theta_2 > \theta_1), \tag{7.50}$$

$$H_0\colon \theta \leqq \theta_1 \ \text{ ou } \ \theta \geqq \theta_2 \ (\theta_2 > \theta_1) \ \text{ contra } \ H_1\colon \theta_1 < \theta < \theta_2. \tag{7.51}$$

Os resultados que vão apresentar-se somente são válidos para distribuições do tipo exponencial,

$$f(x \mid \theta) = C(\theta) \exp\{Q(\theta) R(x)\} H(x), \tag{7.52}$$

ou,

$$f(x\,|\,\theta) = C(\theta)\exp\{\theta x\}H(x), \tag{7.53}$$

com $\theta \in \Theta$, Θ intervalo de $\mathbf{R}$. Estas distribuições são subclasse das distribuições com RVM o que implica uma maior restrição dentro da via de investigação 1) referida no início da secção anterior.

Mesmo com a restrição à família exponencial só existem testes UMP para o ensaio (7.51). Com efeito, Lehmann (1959) demonstra,

Teorema 7.10 — Se $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. com função de densidade (7.52) [com $Q(\theta)$ estritamente crescente], com $T(\mathbf{X})$ estatística suficiente, $T(\mathbf{X}) = \Sigma R(X_i)$, para ensaiar,

$$H_0\colon \theta \leqq \theta_1 \;\text{ ou }\; \theta \geqq \theta_2\ (\theta_2 > \theta_1) \;\text{ contra }\; H_1\colon \theta_1 < \theta < \theta_2,$$

existe um teste UMP dado por,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad t_1 < T(\mathbf{x}) < t_2, (t_2 > t_1), \\ \gamma_i & \text{se} \quad T(\mathbf{x}) = t_i,\ i = 1, 2, \\ 0 & \text{se} \quad T(\mathbf{x}) < t_1 \text{ ou } T(\mathbf{x}) > t_2, \end{cases} \tag{7.54}$$

onde t_i, γ_i, $i = 1, 2$, são determinados pelas condições,

$$E_{\theta_1}\{\phi(\mathbf{X})\} = E_{\theta_2}\{\phi(\mathbf{X})\} = \alpha. \tag{7.55}$$

□□

Exemplo 7.15 — Suponha-se que X tem distribuição Binomial com $N = 15$ e que para ensaiar $H_0\colon \theta \leqq 0{,}3$ ou $\theta \geqq 0{,}8$ contra $H_1\colon 0{,}3 < \theta < 0{,}8$ pretende determinar-se um teste UMP com $\alpha = 0{,}25$.

Como X é suficiente para θ o teste é da forma,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad x_1 < x < x_2, \\ \gamma_i & \text{se} \quad x = x_i,\ i = 1, 2, \\ 0 & \text{se} \quad x < x_1 \text{ ou } x > x_2, \end{cases}$$

com,

$$\begin{aligned} E_{\theta_1}\{\phi(X)\} = E_{\theta=0{,}3}\{\phi(X)\} &= \gamma_1 \tbinom{15}{x_1} (0{,}3)^{x_1}(0{,}7)^{15-x_1} + \\ &+ \sum_{x=x_1+1}^{x_2-1} \tbinom{15}{x} (0{,}3)^{x}(0{,}7)^{15-x} + \gamma_2 \tbinom{15}{x_2} (0{,}3)^{x_2}(0{,}7)^{15-x_2} \\ &= 0{,}25, \end{aligned}$$

$$E_{\theta_2}\{\phi(X)\} = E_{\theta=0,8}\{\phi(X)\} = \gamma_1 \binom{15}{x_1} (0{,}8)^{x_1}(0{,}2)^{15-x_1} + \\ + \sum_{x=x_1+1}^{x_2-1} \binom{15}{x} (0{,}8)^{x}(0{,}2)^{15-x} + \gamma_2 \binom{15}{x_2} (0{,}8)^{x_2}(0{,}2)^{15-x_2} \\ = 0{,}25,$$

donde se obtém, depois de algum cálculo,

$$x_1 = 6,\ x_2 = 11,\ \gamma_1 = 0{,}81,\ \gamma_2 = 0{,}46.$$

□

Exemplo 7.16 — Embora não se consigam estabelecer proposições gerais sobre a existência de testes UMP para os ensaios bilateriais (7.49) e (7.50), há casos particulares em que tais testes existem.

Com X_i I.I.D., $X_i \sim U(0,\theta)$, $i = 1, 2, \ldots, N$, o teste,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \max(x_i) < \theta_0(\alpha)^{1/N} \text{ ou } \max(x_i) > \theta_0, \\ 0 & \text{outros } \mathbf{x}, \end{cases}$$

é UMP de tamanho α para ensaiar H_0: $\theta = \theta_0$ contra H_1: $\theta \neq \theta_0$.

Com X_i I.I.D. com função de densidade $\exp\{-(x_i - \theta)\}$, $x_i > \theta$, existem testes UMP para ensaiar H_0: $\theta = \theta_0$ contra H_1: $\theta \neq \theta_0$, pois as variáveis aleatórias, $Y_i = \exp\{-X_i\}$, são I.I.D. $U(0, e^{-\theta})$. □

Os três teoremas seguintes contém resultados referentes aos ensaios (7.49) e (7.50) mas as respectivas demonstrações [veja-se Ferguson (1967)] são omitidas, quer por serem demasiado pesadas, quer por carecerem de proposições auxiliares — nomeadamente a que generaliza o Lema de Neyman-Pearson — cuja inclusão iria alongar demasiado o presente estudo.

Teorema 7.11 — Se a variável aleatória, X, tem distribuição (7.53), então dado qualquer $\phi' \in \Phi$, existe, para o ensaio (7.50), um teste $\phi \in \Phi$, definido por,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad x < x_1 \text{ ou } x > x_2, \\ \gamma_i & \text{se} \quad x = x_i,\ i = 1, 2, \\ 0 & \text{se} \quad x_1 < x < x_2, \end{cases} \tag{7.56}$$

que verifica,

$$E_{\theta_1}\{\phi(X)\} = E_{\theta_1}\{\phi'(X)\} \quad \text{e} \quad E_{\theta_2}\{\phi(X)\} = E_{\theta_2}\{\phi'(X)\}, \tag{7.57}$$

e,

$$E_{\theta}\{\phi(X)\} - E_{\theta}\{\phi'(X)\} = \begin{cases} \leqq 0 & \text{para} \quad \theta_1 < \theta < \theta_2, \\ \geqq 0 & \text{para} \quad \theta < \theta_1 \text{ ou } \theta > \theta_2. \end{cases} \tag{7.58}$$

□□

A proposição enunciada estabelece que dado qualquer teste ϕ' para ensaiar H_0 contra H_1, é sempre possível encontrar outro teste ϕ, da forma (7.56), que verificando (7.57) e (7.58) nunca é inferior a ϕ'. Repare-se, porém, que ϕ não é UMP, pois, como a Fig. 7.12 esclarece, pode haver um teste ϕ'', não satisfazendo (7.57), mas ainda de tamanho comum a ϕ e ϕ', com potência superior a ϕ quando $\theta > \theta_2$, por exemplo.

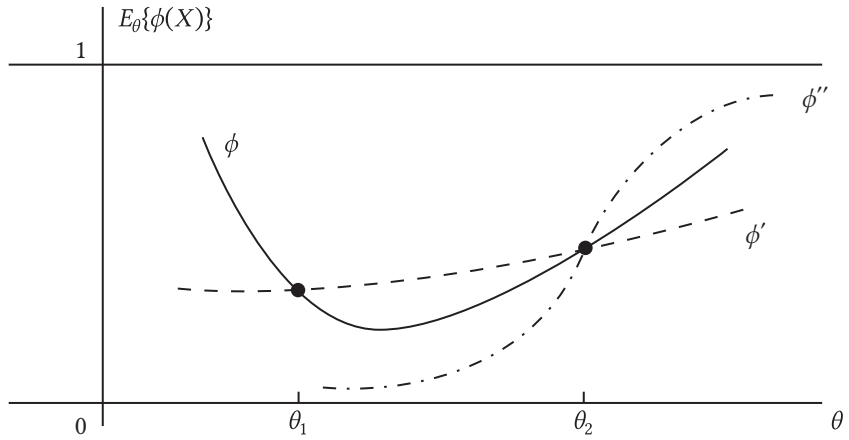


Fig. 7.12

Teorema 7.12 — Se a variável aleatória, X, tem distribuição (7.53), então, dado qualquer $\phi' \in \Phi$, existe, para o ensaio (7.49), um teste $\phi \in \Phi$, definido tal como em (7.56), tal que,

$$E_{\theta_0}\{\phi(X)\} = E_{\theta_0}\{\phi'(X)\}, \tag{7.59}$$

$$\frac{d}{d\theta}E_\theta\{\phi(X)\}\,|_{\,\theta=\theta_0} = \frac{d}{d\theta}E_\theta\{\phi'(X)\}\,|_{\,\theta=\theta_0}, \tag{7.60}$$

$$E_\theta\{\phi(X)\} \geqq E_\theta\{\phi'(X)\} \text{ para todo o } \theta \in \Theta. \tag{7.61}$$

□□

Repare-se que este teorema também não estabelece que o teste ϕ da forma (7.56) é UMP, mas, sim, que dado qualquer ϕ' existe um ϕ da forma (7.56), do mesmo tamanho e com função potência com a mesma inclinação para $\theta = \theta_0$, nunca inferior ao primeiro.

Quando se consideram dois testes ϕ e ϕ' do mesmo tamanho, é evidente que a condição necessária para ϕ não ter potência inferior a ϕ' para todo o $\theta \in \Theta$ é que a inclinação no ponto θ_0 (θ_0 ponto interior de Θ) da função $E_\theta\{\phi(X)\}$ seja igual à da função $E_\theta\{\phi'(X)\}$. O Teorema 7.12 mostra que essa condição é também suficiente.

Os Teoremas 7.11 e 7.12 na medida em que permitem sempre determinar um teste (7.56) não menos potente do que qualquer teste proposto, preparam naturalmente o caminho para instituir os testes da forma (7.56) em classe essencialmente completa. Para o passo final há que introduzir restrições sobre a função perca.

Admita-se que a função perca satisfaz as condições,

$$\begin{aligned} &L_1(\theta) - L_0(\theta) \geqq 0 \quad \text{se} \quad \theta_1 < \theta < \theta_2, \\ &L_1(\theta) - L_0(\theta) \leqq 0 \quad \text{se} \quad \theta < \theta_1 \ \text{ ou } \ \theta > \theta_2, \end{aligned} \tag{7.62}$$

onde $\theta_1 \leqq \theta_2$ e Θ é um intervalo de $\mathbf{R}$. O caso $\theta_1 < \theta_2$ refere-se ao ensaio definido por (7.50) e o caso $\theta_1 = \theta_2$ ao ensaio (7.49). Tem-se,

Teorema 7.13 — Se a variável aleatória, X, tem distribuição (7.53) e se a função perca satisfaz as condições (7.62) com $\theta_1 \leqq \theta_2$, a classe de testes definidos por (7.56) é essencialmente completa. Se além disso existirem números reais, $\theta_1' < \theta_2' < \theta_3'$, tais que,

$$L_1(\theta_1') - L_0(\theta_1') < 0, \ L_1(\theta_2') - L_0(\theta_2') > 0, \ L_1(\theta_3') - L_0(\theta_3') < 0,$$

todo o teste da forma (7.56) é admissível. □□

7.5 Procedimentos monótonos

Os Teoremas 7.9 e 7.13 são susceptíveis de generalização. Considerem-se duas acções, $A = \{a_0, a_1\}$, sem referir necessariamente qualquer ensaio de hipóteses [na presente secção as acções a_0 e a_1 aparecem trocadas em relação a alguns casos anteriores o que não levanta qualquer problema por as designações serem convencionais] e introduzam-se os seguintes pressupostos sobre a função perca:

[A] A diferença, $L_0(\theta) - L_1(\theta)$, é contínua por troços e muda de sinal exactamente n vezes, nos pontos isolados, $\theta_1^0, \theta_2^0, \ldots, \theta_n^0$, conjunto que se designa por Ω^0. Supõe-se que Ω^0 contém os zeros de $L_0(\theta) - L_1(\theta)$ e que $L_0(\theta) - L_1(\theta) > 0$ para $\theta < \theta_1^0$.

[B] Dois pontos sucessivos de Ω^0 podem ser iguais mas não mais do que dois; sendo $\theta_i^0 = \theta_{i+1}^0$, então,

$$[L_0(\theta) - L_1(\theta)][L_0(\theta') - L_1(\theta')] > 0,$$

para $\theta < \theta_i^0 < \theta'$ (θ e θ' na vizinhança de θ_i^0), e,

$$[L_0(\theta) - L_1(\theta)][L_0(\theta_i^0) - L_1(\theta_i^0)] < 0,$$

para $\theta < \theta_i^0$ (θ na vizinhança de θ_i^0). A igualdade de dois pontos sucessivos corresponde à situação em que uma acção é preferida na vizinhança de θ_i^0 excepto para $\theta = \theta_i^0$ onde a outra acção é a mais favorável.

Os pressupostos [A]–[B] têm as condições (7.44)–(7.45) e (7.62) como casos particulares. A situação correspondente a [A] representa-se na Fig. 7.13 e a correspondente a [B] na Fig. 7.14.

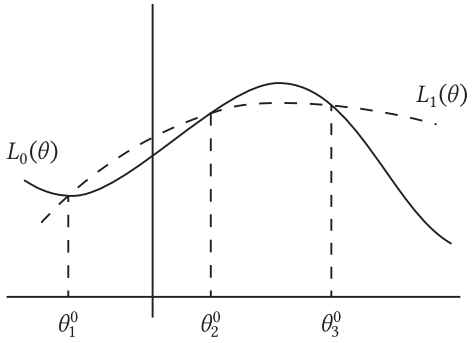


Fig. 7.13

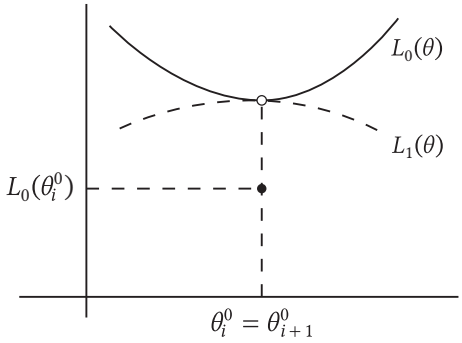


Fig. 7.14

Quando a função perca verifica os pressupostos [A]–[B] tem especial interesse a classe de procedimentos monótonos, isto é, a classe de funções de decisão aleatórias ou testes da forma,

$$\phi(x) = \begin{cases} 1 & \text{para} \quad x_{2i} < x < x_{2i+1},\ i = 0, 1, \dots, [n/2], \\ \gamma_j & \text{para} \quad x = x_j,\ j = 1, 2, \dots, n, \\ 0 & \text{para} \quad \text{outros } x, \end{cases} \tag{7.63}$$

onde, $[n/2]$, simboliza o maior inteiro contido em $n/2$, $x_0 = -\infty$ e $x_{n+1} = +\infty$. A classe de procedimentos monótonos com n pontos de casualização designa-se por $\mathfrak{M}_n$; evidentemente, os testes da forma (7.39) pertencem a $\mathfrak{M}_1$ e os testes da forma (7.56) pertencem a $\mathfrak{M}_2$.

Uma função de decisão da classe $\mathfrak{M}_n$, leva à acção a_1 sempre que o valor observado para X cai em qualquer dos intervalos,

$$(x_0, x_1), (x_2, x_3), \dots, (x_{2i}, x_{2i+1}), \dots, (x_n, x_{n+1}),$$

e à acção a_0 sempre que o valor de X pertence a qualquer dos intervalos,

$$(x_1, x_2), \dots, (x_{2i+1}, x_{2i+2}), \dots, (x_{n-1}, x_n).$$

Isso, é claro, se n for par; com n ímpar as modificações são imediatas. Se o valor observado coincidir com algum dos x_j, $j = 1, 2, \dots, n$, procede-se a uma casualização através da qual o decisor toma a_1 com probabilidade γ_j e a_0 com probabilidade $1-\gamma_j$.

Uma função de densidade ou de probabilidade, $f(x \mid \theta)$, dependente de um parâmetro θ cujo domínio, Θ, é intervalo, aberto ou fechado, de $\mathbf{R}$, pertence à classe $\mathscr{P}_n$

ou diz-se do tipo Pólya n, se,

$$\begin{vmatrix} f(x_1\,|\,\theta_1) & f(x_2\,|\,\theta_1) & \dots & f(x_m\,|\,\theta_1) \\ f(x_1\,|\,\theta_2) & f(x_2\,|\,\theta_2) & \dots & f(x_m\,|\,\theta_2) \\ \cdot & \cdot & \dots & \cdot \\ f(x_1\,|\,\theta_m) & f(x_2\,|\,\theta_m) & \dots & f(x_m\,|\,\theta_m) \end{vmatrix} \geqq 0 \tag{7.64}$$

para, $x_1 < x_2 < \cdots < x_m$, $\theta_1 < \theta_2 < \cdots < \theta_m$, e para todo o $m \leqq n$. Note-se que para $n = 1$ a condição é satisfeita por toda e qualquer função de densidade ou de probabilidade. Recordando (7.37) conclui-se imediatamente que as distribuições com RVM pertencem a $\mathscr{P}_2$.

Se para todo o $m \leqq n$ a desigualdade (7.64) é estrita diz-se que $f(x\,|\,\theta)$ pertence propriamente à classe $\mathscr{P}_n$.

Tem-se, como é claro,

$$\mathscr{P}_2 \supset \mathscr{P}_3 \supset \dots \supset \mathscr{P}_n \supset \dots;$$

a classe $\mathscr{P}_\infty$ resulta da intersecção de todas estas classes e contém as funções que verificam (7.64) para todo o m. É simples exercício mostrar que as distribuições da família exponencial pertencem propriamente a $\mathscr{P}_\infty$.

Os dois teoremas fundamentais sobre procedimentos monótonos são devidos a Karlin (1956, 1957a):

Teorema 7.14 — Se a função perca satisfaz os pressupostos [A]–[B] e se a variável aleatória, X, tem distribuição propriamente $\mathscr{P}_{n+1}$, dada a distribuição a priori, $h(\theta)$, ou todas as funções de decisão são Bayes ou a função de decisão Bayes contra h pertence a $\mathfrak{M}_n$ com valores γ_j, $j = 1, 2, \dots, n$, arbitrários. □□

Teorema 7.15 — Se a função perca satisfaz os pressupostos [A]–[B] e se a variável aleatória, X, tem distribuição propriamente $\mathscr{P}_{n+1}$, dada qualquer função de decisão $\phi' \notin \mathfrak{M}_n$, existe uma única função de decisão $\phi \in \mathfrak{M}_n$, tal que $R(\theta, \phi') > R(\theta, \phi)$, excepto para $\theta \in \Omega^0$ onde se verifica a igualdade. Além disso, $\mathfrak{M}_n$ constitui uma classe completa mínima. □□

A demonstração destes teoremas não cabe no presente estudo. Em relação ao segundo pode dar-se, no entanto, uma ideia geral do raciocínio de modo a destacar uma importante propriedade das distribuições do tipo Pólya.

Com $\phi' \notin \mathfrak{M}_n$ e $\phi \in \mathfrak{M}_n$, considere-se,

$$\begin{aligned} R(\theta, \phi') - R(\theta, \phi) &= [L_0(\theta) - L_1(\theta)] \int [\phi'(x) - \phi(x)] f(x\,|\,\theta)\, dx \\ &= [L_0(\theta) - L_1(\theta)] \int \Delta(x) f(x\,|\,\theta)\, dx; \end{aligned}$$

excepto para $\theta \in \Omega^0$ tem-se $R(\theta, \phi') > R(\theta, \phi)$ quando os dois factores do segundo membro tiverem o mesmo sinal. Dada a natureza de ϕ [por ϕ pertencer a $\mathfrak{M}_n$] a função de x, $\Delta(x)$, muda de sinal n vezes; se $f(x\,|\,\theta)$ é propriamente Pólya $n+1$, também o integral muda de sinal n vezes e pela mesma ordem de $\Delta(x)$.

Uma importante consequência do Teorema 7.15 é mostrar que se a variável aleatória, X, tem distribuição propriamente Pólya 3 não existe teste UMP para os ensaios bilaterais (7.49) e (7.50). Recordando o Ex. 7.16 há que notar que a distribuição Uniforme não é propriamente Pólya 3 daí a existência de um teste UMP para H_0: $\theta = \theta_0$ contra H_1: $\theta \neq \theta_0$. Note-se, por outro lado, que para a distribuição do exemplo não existe teste UMP para o ensaio (7.50) [veja-se Karlin (1957a)].

7.6 Testes uniformemente mais potentes não enviesados

No ensaio de hipóteses bilaterais, (7.49) e (7.50), a via de investigação que consiste em impor restrições à família, $\{f(x\,|\,\theta) : \theta \in \Theta\}$, não permite estabelecer a existência, em termos gerais, de testes UMP. Fica, no entanto, aberta, a pesquisa de testes UMP em classes mais restritas de funções de decisão, sem excluir restrições sobre a família $\{f(x\,|\,\theta) : \theta \in \Theta\}$.

Considere-se o ensaio de H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$; um teste $\phi \in \Phi$, de tamanho α,

$$\sup_{\theta=\Theta_0} E_\theta\{\phi(X)\} = \alpha,$$

diz-se não enviesado quando,

$$E_\theta\{\phi(X)\} \geqq \alpha \quad \text{para} \quad \theta \in \Theta_1. \tag{7.65}$$

A interpretação de (7.65) é a seguinte: se $\phi \in \Phi$ é um teste não enviesado, a probabilidade de rejeitar H_0 quando falsa nunca é inferior à probabilidade de rejeitar H_0 quando verdadeira. Os testes ϕ e ϕ' cujas funções potência se representam na Fig. 7.15 são não enviesados; o teste ϕ'' é enviesado.

Considere-se, H_0: $\theta_1 \leqq \theta \leqq \theta_2$ contra H_1: $\theta < \theta_1$ ou $\theta > \theta_2$, com $\theta_2 > \theta_1$. Se a variável aleatória X tem distribuição do tipo exponencial com parametrização natural, com θ parâmetro k-dimensional.

$$f(x\,|\,\boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^{k} \theta_j R_j(x)\right\} H(x),$$

uma amostra casual, $(X_1, X_2, \ldots, X_N)$, tem função de densidade,

$$f(\mathbf{x}\,|\,\boldsymbol{\theta}) = [C(\boldsymbol{\theta})]^N \exp\left\{\sum_{j=1}^{k} \theta_j \sum_{i=1}^{N} R_j(x_i)\right\} \Pi H(x_i). \tag{7.66}$$

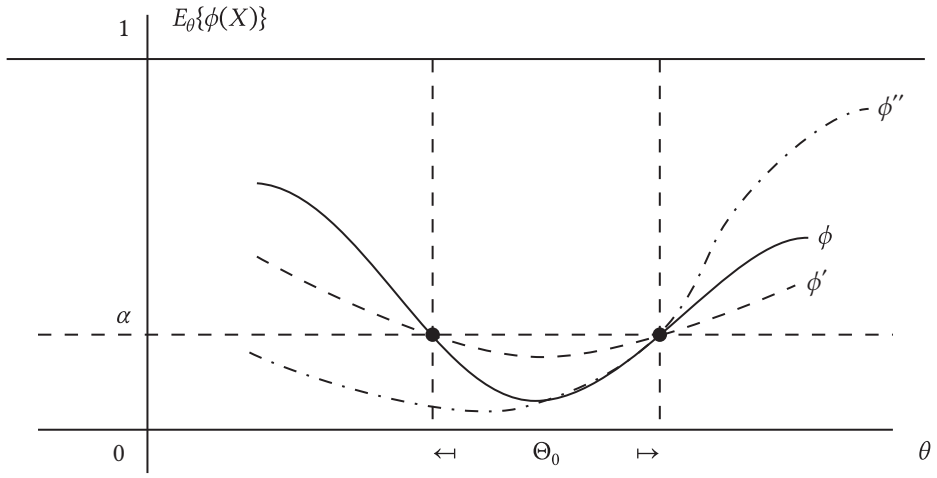


Fig. 7.15

Lehman (1959) demonstra o seguinte,

Teorema 7.16 — Se $\phi(\mathbf{x})$ é uma função para a qual, com $f(\mathbf{x}\,|\,\boldsymbol{\theta})$ dado por (7.66), o integral,

$$\int \ldots \int \phi(\mathbf{x}) f(\mathbf{x}\,|\,\boldsymbol{\theta})\, d\mathbf{x}, \tag{7.67}$$

existe para todo o $\boldsymbol{\theta} \in \Theta$, espaço natural do parâmetro, então o integral é função contínua de $\boldsymbol{\theta}$ em todos os pontos do interior de Θ e as derivadas de todas as ordens em relação a $\boldsymbol{\theta}$ podem permutar-se com a operação de integração. □□

A consequência deste teorema que aqui particularmente interessa é a seguinte: se X tem distribuição da família exponencial (7.53), a função potência de qualquer teste $\phi \in \Phi$ [o facto de ser $0 \leqq \phi(x) \leqq 1$ garante a existência do integral (7.67)] é contínua em θ e tem derivadas de todas as ordens permutáveis com a operação de integração. Portanto, qualquer teste ϕ não enviesado de tamanho α deve verificar,

$$E_{\theta_1}\{\phi(X)\} = E_{\theta_2}\{\phi(X)\} = \alpha, \tag{7.68}$$

pois, por definição, não pode ser $E_{\theta_i}\{\phi(X)\} > \alpha$, $i = 1, 2$, e também não pode ter-se $E_{\theta_i}\{\phi(X)\} < \alpha$, para $i = 1, 2$, porque isso implicaria, pela continuidade,

$$E_\theta\{\phi(X)\} < \alpha \text{ para } \theta_1 - \xi < \theta \leqq \theta_1 \text{ ou para } \theta_2 \leqq \theta < \theta_2 + \xi,$$

e ϕ não satisfazia (7.65).

Um teste não enviesado que é uniformemente mais potente dentro da classe dos testes não enviesados diz-se UMPU (uniformemente mais potente «*unbiased*»). A proposição seguinte é simples corolário do Teorema 7.11,

Teorema 7.17 — Se a variável aleatória, X, tem distribuição (7.53), o teste com estrutura (7.56), em que x_1, x_2, γ_1 e γ_2 são determinados pelas condições (7.68), é UMPU relativamente a H_0: $\theta_1 \leqq \theta \leqq \theta_2$. □□

Note-se que semelhante teste é, pelo Teorema 7.11, melhor do que qualquer teste não enviesado de tamanho α; como há pelo menos um teste não enviesado de tamanho α, $\phi(x) \equiv \alpha$, a existência fica garantida.

Analogamente, para H_0: $\theta = \theta_0$ contra H_1: $\theta \neq \theta_0$, qualquer teste não enviesado, ϕ, de tamanho α, deve verificar, $E_{\theta_0}\{\phi(X)\} = \alpha$, e,

$$\frac{d}{d\theta} E_\theta\{\phi(X)\}|_{\theta=\theta_0} = 0, \tag{7.69}$$

em virtude da continuidade. É consequência imediata do Teorema 7.12, o

Teorema 7.18 — Se a variável aleatória, X, tem distribuição (7.53), o teste com estrutura (7.56), em que x_1, x_2, γ_1 e γ_2 são determinados pelas condições,

$$E_{\theta_0}\{\phi(X)\} = \alpha \quad \text{e} \quad E_{\theta_0}\{X\phi(X)\} = \alpha E_{\theta_0}\{X\}, \tag{7.70}$$

é UMPU relativamente a H_0: $\theta = \theta_0$. □□

Semelhante teste existe porque, mais uma vez, $\phi(x) \equiv \alpha$ é um teste não enviesado de tamanho α. A segunda condição (7.70) é consequência de (7.69) e de (7.53), pois,

$$\begin{aligned}\frac{d}{d\theta} E_\theta\{\phi(X)\} &= \frac{d}{d\theta} \int \phi(x) C(\theta) \exp\{\theta x\} H(x)\, dx \\ &= [C'(\theta)/C(\theta)] E_\theta\{\phi(X)\} + E_\theta\{X\phi(X)\},\end{aligned}$$

e $C'(\theta)/C(\theta) = -E_\theta\{X\}$ [derive-se $\int C(\theta)\exp\{\theta x\}H(x)\,dx = 1$ em relação a θ].

No ensaio de H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$, se a função perca satisfaz as condições,

$$L_0(\theta) < L_1(\theta) \text{ se } \theta \in \Theta_0,\ L_1(\theta) < L_0(\theta) \text{ se } \theta \in \Theta_1, \tag{7.71}$$

pode verificar-se que o teste não enviesado, ϕ, sendo admissível na classe dos testes não enviesados, também é admissível na classe de todos os testes. Suponha-se o contrário, isto é, que existe ϕ' qualquer tal que se verifica, $R(\theta, \phi) \geqq R(\theta, \phi')$ para todo o θ e desigualdade estrita pelo menos para um valor de θ. Em consequência de (7.7), expressão da função risco, tem-se,

$$[E_\theta\{\phi(X)\} - E_\theta\{\phi'(X)\}][L_1(\theta) - L_0(\theta)] \geqq 0,$$

ou seja, em face de (7.71)

$$E_\theta\{\phi(X)\} \geqq E_\theta\{\phi'(X)\} \quad \text{para} \quad \theta \in \Theta_0,$$
$$E_\theta\{\phi(X)\} \leqq E_\theta\{\phi'(X)\} \quad \text{para} \quad \theta \in \Theta_1.$$

Como ϕ é não enviesado, $E_\theta\{\phi(X)\} \leqq \alpha$ para $\theta \in \Theta_0$ e para algum α; logo, também $E_\theta\{\phi'(X)\} \leqq \alpha$ para $\theta \in \Theta_0$ e ϕ' é não enviesado o que contradiz a hipótese de ϕ ser admissível na classe dos testes não enviesados. Assim, se a função perca obedece às condições (7.71), os ensaios UMPU são admissíveis.

Exemplo 7.17 — Com $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. $N(0, \sigma^2)$, no ensaio de H_0: $\sigma_1^2 \leqq \sigma^2 \leqq \sigma_2^2$ emprega-se a estatística suficiente para σ^2, $T = \Sigma X_i^2$, que tem função de densidade, $g(t \mid \sigma^2) = (1/\sigma^2)\chi_N^2(t \mid \sigma^2)$, onde,

$$\chi_N^2(t) = [2^{N/2}\Gamma(N/2)]^{-1} \exp\{-t/2\} t^{(N/2)-1},$$

é a função de densidade da distribuição-χ^2 com N graus de liberdade.

Considerando que a distribuição-χ^2 pertence à família exponencial, o Teorema 7.17 indica que o teste UMPU é da forma (7.56),

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \Sigma x_i^2 < t_1 \text{ ou } \Sigma x_i^2 > t_2, \\ 0 & \text{se} \quad t_1 \leqq \Sigma x_i^2 \leqq t_2, \end{cases}$$

sendo a casualização irrelevante por T ser variável aleatória contínua. Se α é a dimensão pretendida, os valores de t_1 e t_2 são obtidos a partir do sistema correspondente a (7.68),

$$E_{\sigma_1^2}\{\phi(\mathbf{X})\} = 1 - \int_{t_1/\sigma_1^2}^{t_2/\sigma_1^2} \chi_N^2(t)\, dt = \alpha,$$

$$E_{\sigma_2^2}\{\phi(\mathbf{X})\} = 1 - \int_{t_1/\sigma_2^2}^{t_2/\sigma_2^2} \chi_N^2(t)\, dt = \alpha,$$

que se resolve facilmente com auxílio de tabelas da distribuição-χ^2.

Para ensaiar H_0: $\sigma^2 = \sigma_0^2$, o teste UMPU tem ainda a mesma estrutura mas t_1 e t_2 são determinados a partir do sistema correspondente a (7.70). Assim, por uma parte,

$$E_{\sigma_0^2}\{\phi(\mathbf{X})\} = 1 - \int_{t_1/\sigma_0^2}^{t_2/\sigma_0^2} \chi_N^2(t)\, dt = \alpha \tag{7.72}$$

por outra parte,

$$E_{\sigma_0^2}\{T\} = \int_0^\infty t(1/\sigma_0^2)\chi_N^2(t/\sigma_0^2)\, dt = N\sigma_0^2,$$

e,

$$E_{\sigma_0^2}\{T\phi(\mathbf{X})\} = N\sigma_0^2 - \left[\int_{t_1/\sigma_0^2}^{t_2/\sigma_0^2} t\,\chi_N^2(t)\,dt\right]\sigma_0^2;$$

portanto, a segunda condição (7.70) conduz a,

$$\int_{t_1/\sigma_0^2}^{t_2/\sigma_0^2} t\,\chi_N^2(t)\,dt = N(1-\alpha),$$

ou seja, notando que $t\chi_N^2(t) = N\chi_{N+2}^2(t)$,

$$\int_{t_1/\sigma_0^2}^{t_2/\sigma_0^2} \chi_{N+2}^2(t)\,dt = 1-\alpha. \tag{7.73}$$

Em resumo, t_1 e t_2 são determinados pelas relações (7.72) e (7.73) recorrendo a tabelas da distribuição-χ^2 [Lehmann (1959)]. □

7.7 Testes similares[5]

Na óptica de reduzir também a classe de testes Φ para optimizar em classe mais restrita — ponto de vista conducente aos testes UMPU — apresentam-se dignos de referência os testes similares. O estudo destes testes encontra em geral motivação no objectivo de chegar aos testes não enviesados quando o espaço do parâmetro, Θ, é k-dimensional, $k \geqq 2$.

A análise feita na secção 7.3 mostra que mesmo quando θ é escalar os testes UMP só existem em situações muito particulares. Se θ é vector, a condição (7.31) é ainda de mais difícil verificação e os testes UMPU constituem já uma boa solução.

Um teste $\phi \in \Phi$ é α-similar num subconjunto Ω do espaço do parâmetro Θ ($\Omega \subset \Theta$), se,

$$E_\theta\{\phi(X)\} = \alpha \text{ para todo o } \theta \in \Omega; \tag{7.74}$$

um teste ϕ é similar em Ω se é α-similar para algum α, $0 \leqq \alpha \leqq 1$.

Considere-se o ensaio de H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$ e $\Theta_0 \cap \Theta_1 = \emptyset$; $\phi \in \Phi$ diz-se UMP α-similar se é UMP na subclasse de testes que são α-similares em $\Omega = \overline{\Theta}_0 \cap \overline{\Theta}_1$, onde $\overline{\Theta}_0$ e $\overline{\Theta}_1$ são os fechos de Θ_0 e Θ_1 respectivamente.

Exemplo 7.18 — Designe $\mathbf{X}$ uma amostra casual de dimensão N de uma população $N(\mu, \sigma^2)$ e sejam $\Theta_0 = \{\boldsymbol{\theta} = (\mu, \sigma^2) : \sigma^2 \leqq \sigma_0^2\}$ e $\Theta_1 = \{\boldsymbol{\theta} = (\mu, \sigma^2) : \sigma^2 > \sigma_0^2\}$, conjuntos representados na Fig. 7.16. O conjunto $\Omega = \overline{\Theta}_0 \cap \overline{\Theta}_1$ corresponde aos pontos da recta $\sigma^2 = \sigma_0^2$.

[5] A presente secção é de natureza complementar.

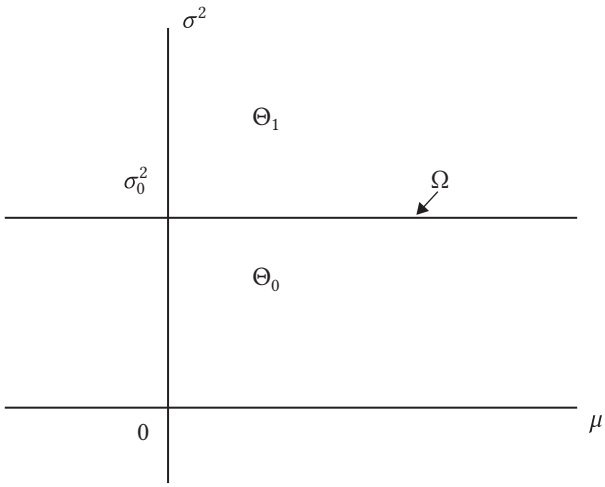


Fig. 7.16

Por outro lado, sabe-se que,

$$\frac{NS^2}{\sigma^2} = \frac{\Sigma(X_i - \overline{X})^2}{\sigma^2},$$

tem distribuição-χ^2 com $N-1$ graus de liberdade, independente de μ. O teste,

$$\phi(\mathbf{x}) = \phi(s^2) = \begin{cases} 1 & \text{se} \quad s^2 > s_0^2, \\ 0 & \text{se} \quad s^2 \leqq s_0^2, \end{cases}$$

onde $s^2 = \Sigma(x_i - \overline{x})^2/N$ e s_0^2 é determinado pela condição,

$$\int_{Ns_0^2/\sigma_0^2}^{\infty} \chi_{N-1}^2(t)\, dt = \alpha,$$

é α-similar em Ω. De facto,

$$E_{(\mu,\sigma^2)}\{\phi(S^2)\} = E_{\sigma_0^2}\{\phi(S^2)\} = \alpha,$$

donde,

$$E_{(\mu,\sigma^2)}\{\phi(S^2)\} = \alpha \text{ para todo o } (\mu,\sigma^2) \in \Omega.$$

□

Se ϕ é teste não enviesado de tamanho α, a respectiva função potência verifica as desigualdades,

$$\begin{aligned} &E_\theta\{\phi(X)\} \leqq \alpha \text{ para } \theta \in \Theta_0, \\ &E_\theta\{\phi(X)\} \geqq \alpha \text{ para } \theta \in \Theta_1; \end{aligned} \tag{7.75}$$

se $E_\theta\{\phi(X)\}$ é função contínua de θ as relações (7.75) implicam,

$$E_\theta\{\phi(X)\} = \alpha \text{ para } \theta \in \Omega = \overline{\Theta}_0 \cap \overline{\Theta}_1, \tag{7.76}$$

quer dizer, ϕ é α-similar em Ω. De facto, se $\theta^* \in \Omega$, θ^* é ponto limite de Θ_0 e de Θ_1; logo, existem duas sucessões, $\{\theta_{0m}\}$ e $\{\theta_{1n}\}$, a primeira de pontos de Θ_0 e a segunda de pontos de Θ_1, ambas tendendo para θ^*, tais que, pela continuidade,

$$E_{\theta_{0m}}\{\phi(X)\} \Rightarrow E_{\theta^*}\{\phi(X)\} \leqq \alpha, \; E_{\theta_{1n}}\{\phi(X)\} \Rightarrow E_{\theta^*}\{\phi(X)\} \geqq \alpha.$$

No domínio das relações entre testes não enviesados e testes similares tem-se,

Teorema 7.19 — Se todo o teste de H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$ tem função potência contínua e se ϕ é UMP α-similar e tem tamanho α, então ϕ é UMPU.

Dem. Por um lado, se θ é UMP α-similar de tamanho α então ϕ é não enviesado, pois:

(1) $E_\theta\{\phi(X)\} \leqq \alpha$ para $\theta \in \Theta_0$ por ser ϕ de tamanho α e
(2) $E_\theta\{\phi(X)\} \geqq \alpha$ para $\theta \in \Theta_1$ por ser ϕ UMP α-similar, portanto, uniformemente mais potente do que o teste $\phi'(x) \equiv \alpha$ que, como é óbvio, é α-similar.

Por outro lado, se ϕ'' é qualquer teste não enviesado de tamanho α, tem-se, como acima foi dito, que ϕ'' é α-similar em $\Omega = \overline{\Theta}_0 \cap \overline{\Theta}_1$. Logo, ϕ não pode ser inferior a ϕ'' onde se conclui que ϕ é UMPU. □□

A relevância deste teorema resulta do seguinte facto: a determinação de testes UMP similares é analiticamente mais simples do que a determinação de testes UMPU; basta notar que os testes não enviesados são definidos por desigualdades [veja-se (7.75)].

A construção de testes α-similares pode fazer-se sem grande dificuldade quando existe uma estatística suficiente T — unidimensional ou multidimensional — para a família, $\{f(x\,|\,\theta) : \theta \in \Omega\}$ onde $\Omega = \overline{\Theta}_0 \cap \overline{\Theta}_1$.

Com efeito, desde que ϕ verifique,

$$E\{\phi(X)\,|\,T = t\} = \alpha \text{ para quase todo o } t \text{ em relação a } \Omega, \tag{7.77}$$

isto é, excepto quando muito para valores t num conjunto com probabilidade 0 para todo o $\theta \in \Omega$, tem-se que ϕ é α-similar em Ω. De facto,

$$E_\theta\{\phi(X)\} = E_\theta\{E\{\phi(X)\,|\,T\}\} = \alpha \text{ para } \theta \in \Omega. \tag{7.78}$$

Um teste ϕ que é α-similar em Ω e satisfaz (7.77) diz-se que tem estrutura Neyman. Caracteriza-se, então, pela propriedade seguinte: a probabilidade de rejeição de H_0

condicionada por $T = t$ é igual a α. Como T é suficiente para $\theta \in \Omega$, a distribuição de X em cada subconjunto, $\{x : T(x) = t\}$, é independente de θ para $\theta \in \Omega$; assim, a condição (7.77) permite essencialmente reduzir o problema de ensaiar H_0 ao problema de ensaiar uma hipótese simples para cada valor t assumido por T.

Exemplo 7.18 — *Continuação.* Considere-se,

$$H_0\colon \begin{cases} \sigma^2 = \sigma_0^2, \\ \mu \text{ qualquer}, \end{cases} \qquad H_1\colon \begin{cases} \sigma^2 > \sigma_0^2, \\ \mu \text{ qualquer}, \end{cases}$$

onde H_0 é obviamente hipótese composta. Sendo $T = \overline{X}$ estatística suficiente para μ, a distribuição de $\mathbf{X}$ sobre cada hiperplano,

$$(x_1 + x_2 + \cdots + x_N)/N = t,$$

é independente de μ. Em consequência, qualquer teste $\phi(\mathbf{X})$ condicionado por $\overline{X} = t$, tem a mesma distribuição para $(\mu, \sigma^2) \in \Omega$, onde,

$$\Omega = \{(\mu, \sigma^2) : \mu \text{ qualquer}, \sigma^2 = \sigma_0^2\}.$$

O teste ϕ introduzido tem estrutura Neyman, pois, para todo o t,

$$E_{\sigma_0^2}\{\phi(S^2) \,|\, \overline{X} = t\} = P[\Sigma(X_i - t)^2/\sigma_0^2 > Ns_0^2/\sigma_0^2] = \alpha.$$

Portanto, sobre cada hiperplano pode proceder-se ao ensaio da hipótese simples, $H_0'\colon \sigma^2 = \sigma_0^2$ contra a alternativa $H_1'\colon \sigma^2 > \sigma_0^2$. □

Por vezes é fácil mostrar que um teste é UMP na subclasse dos testes similares que possuem estrutura Neyman. Interessa, então, investigar em que condições um tal teste é também UMP na classe mais ampla dos testes similares.

Teorema 7.20 — Se T é estatística suficiente completa limitada para a família, $\{f(x \,|\, \theta) : \theta \in \Omega\}$, então todo o teste similar em Ω tem estrutura Neyman.

Dem. Se $E_\theta\{\phi(X)\} = \alpha$ para todo o $\theta \in \Omega$, tem-se,

$$E_\theta\{E\{\phi(X) \,|\, T\}\} = \alpha \text{ para } \theta \in \Omega,$$

ou seja,

$$E_\theta\{E\{\phi(X) \,|\, T\} - \alpha\} = 0 \text{ para } \theta \in \Omega;$$

ora, $E\{\phi(X) \,|\, T\} - \alpha$, é função só de T, visto que o valor esperado não depende de $\theta \in \Omega$ por T ser suficiente para $\theta \in \Omega$, e além disso é função limitada, pois $0 \leqq \phi(x) \leqq 1$. Logo, se T for estatística suficiente completa limitada para $\theta \in \Omega$, tem-se, $E\{\phi(X) \,|\, T\} - \alpha = 0$, excepto para valores de t formando um conjunto com probabilidade zero quando $\theta \in \Omega$. A demonstração fica assim completa. □□

Quando a hipótese a ensaiar é composta e o espaço do parâmetro é pelo menos bidimensional, pode recapitular-se o caminho a seguir para a eventual determinação de testes UMPU com base na doutrina exposta.

Suponha-se, para concretização,

$$\boldsymbol{\theta} = (\eta, \nu_1, \nu_2, \dots, \nu_k) = (\eta, \boldsymbol{\nu}),$$

pressupondo que $\Theta \subset \mathbf{R}^{k+1}$ e que não está contido em qualquer subespaço linear com menos de $k+1$ dimensões.

Tome-se, inicialmente,

$$H_0\text{: } \boldsymbol{\theta} \in \Theta_0 \text{ onde } \Theta_0\text{: } \begin{cases} \eta = \eta_0, \\ \nu_1, \nu_2, \dots, \nu_k, \text{ quaisquer,} \end{cases}$$

e, portanto,

$$H_1\text{: } \boldsymbol{\theta} \in \Theta_1 \text{ onde } \Theta_1\text{: } \begin{cases} \eta \neq \eta_0, \\ \nu_1, \nu_2, \dots, \nu_k, \text{ quaisquer,} \end{cases}$$

tem-se, nesse caso,

$$\Omega = \overline{\Theta}_0 \cap \overline{\Theta} = \Theta_0.$$

O referido caminho contém os seguintes passos:

1.º] Procura determinar-se uma estatística, $\mathbf{T} = (T_1, T_2, \dots, T_k)$, suficiente completa limitada para $\boldsymbol{\nu} = (\nu_1, \nu_2, \dots, \nu_k)$, ou, o que é o mesmo, para $\boldsymbol{\theta} \in \Omega$.

2.º] A partir de $\mathbf{T}$ procuram construir-se testes com estrutura Neyman. Assim, para cada valor $\mathbf{t} = (t_1, t_2, \dots, t_k)$ de $\mathbf{T}$ há que determinar testes condicionados, ϕ, verificando,

$$E_{\eta_0}\{\phi(\mathbf{X}) \,|\, \mathbf{T} = \mathbf{t}\} = \alpha,$$

nas condições já indicadas. A consideração de um teste condicionado para todos os possíveis valores de $\mathbf{t}$ conduz a uma função definida em todo o domínio de $\mathbf{X}$ que se for mensurável pode tomar-se como teste não condicionado.

3.º] Pelo Teorema 7.20 os testes similares têm estrutura Neyman.

4.º] Procura-se um teste UMP similar.

5.º] Pelo Teorema 7.19 o teste UMP similar é também UMPU se a função potência de todo o teste $\phi \in \Phi$ for função contínua de θ.

A família exponencial com parametrização natural é particularmente indicada para aplicação deste processo. Seja $\mathbf{X}$ uma amostra casual de dimensão $N > k+1$, com

$$f(\mathbf{x} \,|\, \boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp\{\eta U(\mathbf{x}) + \sum_{j=1}^{k} \nu_j T_j(\mathbf{x})\} H(\mathbf{x}), \tag{7.79}$$

com ligeira modificação da simbologia e notando que $\boldsymbol{\theta} = (\eta, \boldsymbol{\nu})$.

A acessibilidade da família exponencial com parametrização natural resulta de: (1.º) se Θ contém um rectângulo $(k+1)$-dimensional, conforme sucede nos casos mais correntes, a estatística $(U, \mathbf{T})$ é suficiente completa para $\boldsymbol{\theta}$ [Lehmann (1959)] e, como imediatamente se reconhece, $\mathbf{T}$ é estatística suficiente completa para $\boldsymbol{\nu}$; (2.º) sendo $0 \leqq \phi(\mathbf{x}) \leqq 1$, $\phi \in \Phi$, pelo Teorema 7.16, $E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\}$ é função contínua de $\boldsymbol{\theta}$.

Considere-se agora o ensaio de, H_0: $\eta \leqq \eta_0$ contra H_1: $\eta > \eta_0$, com $\nu_1, \nu_2, \ldots, \nu_k$ quaisquer e $(\eta, \boldsymbol{\nu}) \in \Theta$ e suponha-se que existem pontos em Θ com $\eta < \eta_0$ e $\eta > \eta_0$.

Como $(U, \mathbf{T})$ é estatística suficiente para $\boldsymbol{\theta} \in \Theta$ a análise pode conduzir-se em termos da observação de $(U, \mathbf{T})$ que tem função de densidade,

$$g(u, \mathbf{t} \,|\, \eta, \boldsymbol{\nu}) = C_0(\eta, \boldsymbol{\nu}) \exp\left\{\eta u + \sum_{j=1}^{k} \nu_j t_j\right\} H_0(u, \mathbf{t}), \quad (\eta, \boldsymbol{\nu}) \in \Theta. \tag{7.80}$$

Quando $\mathbf{T} = \mathbf{t}$ é dado, a distribuição de U condicionada por $\mathbf{T} = \mathbf{t}$ é ainda exponencial,

$$g(u \,|\, \mathbf{t}, \eta) = \left[\int \exp\{\eta u\} H_0(u, \mathbf{t})\, du\right]^{-1} \exp\{\eta u\} H_0(u, \mathbf{t}),$$

como se conclui notando,

$$g(\mathbf{t} \,|\, \eta, \boldsymbol{\nu}) = C_0(\eta, \boldsymbol{\nu}) \left[\int \exp\{\eta u\} H_0(u, \mathbf{t})\, du\right] \exp\left\{\sum_{j=1}^{k} \nu_j t_j\right\},$$

e,

$$g(u \,|\, \mathbf{t}, \eta) = g(u, \mathbf{t} \,|\, \eta, \boldsymbol{\nu}) / g(\mathbf{t} \,|\, \eta, \boldsymbol{\nu}).$$

No problema condicionado tem-se apenas uma variável aleatória, U, e um parâmetro unidimensional η e trata-se de ensaiar H_0: $\eta \leqq \eta_0$ contra a alternativa H_1: $\eta > \eta_0$. Possuindo $g(u \,|\, \mathbf{t}, \eta)$ RVM, o Teorema 7.7 é directamente aplicável; existe, portanto, um teste UMP de tamanho α, da forma,

$$\phi(u, \mathbf{t}) = \begin{cases} 1 & \text{se} \quad u > u_0(\mathbf{t}), \\ \gamma_0(\mathbf{t}) & \text{se} \quad u = u_0(\mathbf{t}), \\ 0 & \text{se} \quad u < u_0(\mathbf{t}), \end{cases} \tag{7.81}$$

onde as funções, $u_0(\mathbf{t})$ e $\gamma_0(\mathbf{t})$, são determinadas pela condição,

$$E_{\eta_0}\{\phi(U, \mathbf{T}) \,|\, \mathbf{T} = \mathbf{t}\} = \alpha \text{ para todo o } \mathbf{t}. \tag{7.82}$$

O teste ϕ definido por (7.81) é condicionado por $\mathbf{T} = \mathbf{t}$. Reintrepretando ϕ como teste definido no domínio de $(U, \mathbf{T})$ para a hipótese inicial [referente aos parâmetros da distribuição de $(U, \mathbf{T})$ e não ao parâmetro da distribuição de U condicionada por $\mathbf{T} = \mathbf{t}$], tem-se,

Teorema 7.21 — O teste definido por (7.81) e (7.82) é UMPU de tamanho α para H_0: $\eta \leqq \eta_0$ contra H_1: $\eta > \eta_0$.

Dem. Sendo, $\Omega = \{(\eta, \boldsymbol{\nu}) : (\eta, \boldsymbol{\nu}) \in \Theta \text{ e } \eta = \eta_0\}$, $\mathbf{T}$ que é suficiente para $\boldsymbol{\nu}$ é suficiente para $\boldsymbol{\theta} \in \Omega$. Além disso, Θ é conjunto convexo [Ferguson (1967), Lema 3.5.2], e, por hipótese, contém um rectângulo $(k+1)$-dimensional e pontos $(\eta, \boldsymbol{\nu})$ com $\eta < \eta_0$ e $\eta > \eta_0$. Portanto, Ω é convexo, k-dimensional e contém um rectângulo k-dimensional. Logo, $\mathbf{T}$ é estatística suficiente completa para $\boldsymbol{\theta} \in \Omega$. Por outro lado, como ϕ verifica (7.82) para todo o $\mathbf{t}$, tem-se que ϕ tem estrutura Neyman sendo consequentemente similar, mais precisamente α-similar em Ω.

A função potência, $E_\theta\{(U, \mathbf{T})\}$, é, pelos motivos já referidos, função contínua de $\boldsymbol{\theta}$. Em vista do Teorema 7.19, para provar que ϕ é UMPU basta provar que ϕ é UMP α-similar de tamanho α, isto é, mostrar que ϕ é UMP entre todos os testes que satisfazem (7.82). Ora,

$$E_{\boldsymbol{\theta}}\{\phi(U, \mathbf{T})\} = E_{\boldsymbol{\theta}}\{E_{\eta_0}\{\phi(U, \mathbf{T}) \,|\, \mathbf{T}\}\}; \tag{7.83}$$

desde que ϕ maximiza, $E_{\eta_0}\{\phi(U, \mathbf{T}) \,|\, \mathbf{T} = \mathbf{t}\}$, para todo o $\mathbf{t}$, também maximiza $E_{\boldsymbol{\theta}}\{\phi(U, \mathbf{T})\}$.

Para completar a demonstração é necessário mostrar que $\phi(u, \mathbf{t})$ é mensurável em u e $\mathbf{t}$, existindo, portanto, o valor esperado (7.83) [veja-se Lehmann (1959)]. □□

Recorrendo a raciocínio do mesmo tipo pode demonstrar-se, sempre no âmbito da família exponencial com parametrização natural, que, para,

$$H_0\colon \eta_1 \leqq \eta \leqq \eta_2 \quad \text{contra} \quad H_1\colon \eta < \eta_1 \text{ ou } \eta > \eta_2,$$

o teste da forma,

$$\phi(u, \mathbf{t}) = \begin{cases} 1 & \text{se} \quad u < u_1(\mathbf{t}) \quad \text{ou} \quad u > u_2(\mathbf{t}), \\ \gamma_j(\mathbf{t}) & \text{se} \quad u = u_j(\mathbf{t}),\ j = 1, 2 \\ 0 & \text{se} \quad u_1(\mathbf{t}) < u < u_2(\mathbf{t}), \end{cases} \tag{7.84}$$

com $u_j(\mathbf{t})$ e $\gamma_j(\mathbf{t})$, $j = 1, 2$, determinados pelas condições,

$$E_{\eta_1}\{\phi(U, \mathbf{T}) \,|\, \mathbf{T} = \mathbf{t}\} = E_{\eta_2}\{\phi(U, \mathbf{T}) \,|\, \mathbf{T} = \mathbf{t}\} = \alpha, \tag{7.85}$$

é UMPU.

Para,

$$H_0\colon \eta = \eta_0 \quad \text{contra} \quad H_1\colon \eta \neq \eta_0,$$

o teste da forma (7.84), com $u_j(\mathbf{t})$ e $\gamma_j(\mathbf{t})$, $j = 1, 2$, determinados pelas condições,

$$E_{\eta_0}\{\phi(U, \mathbf{T}) \,|\, \mathbf{T} = \mathbf{t}\} = \alpha, \tag{7.86}$$

$$E_{\eta_0}\{U\phi(U, \mathbf{T}) \,|\, \mathbf{T} = \mathbf{t}\} = \alpha E_{\eta_0}\{U \,|\, \mathbf{T} = \mathbf{t}\}, \tag{7.87}$$

é UMPU.

Exemplo 7.19 — Considere-se uma amostra casual de uma população Normal,

$$f(\mathbf{x} \,|\, \mu, \sigma^2) = (\sigma^2 2\pi)^{-N/2} \exp\{-\Sigma(x_i - \mu)^2/2\sigma^2\},$$

e o ensaio,

$$H_0\colon \mu \leqq 0 \quad \text{contra} \quad H_1\colon \mu > 0.$$

A $f(\mathbf{x} \,|\, \mu, \sigma^2)$ pode dar-se a forma,

$$f(\mathbf{x} \,|\, \mu, \sigma^2) = (\sigma^2 2\pi)^{-N/2} \exp\{-N\mu^2/2\sigma^2\} \exp\{(\mu/\sigma^2)\Sigma x_i - (1/2\sigma^2)\Sigma x_i^2\};$$

como é sabido, $(U, T) = (\Sigma X_i, \Sigma X_i^2)$, é estatística suficiente completa limitada para $\boldsymbol{\theta} = (\mu, \sigma^2)$. Com a reparametrização,

$$\eta = \mu/\sigma^2, \ \nu = -1/2\sigma^2,$$

a distribuição de (U, T) pode escrever-se,

$$g(u, t \,|\, \eta, \nu) = C_0(\eta, \nu) \exp\{\eta u + \nu t\},$$

e em termos dos novos parâmetros o ensaio equivalente é $H_0\colon \eta \leqq 0$ contra $H_1\colon \eta > 0$.

No problema condicionado por $T = t$, isto é, por $\Sigma X_i^2 = t$, o teste,

$$\phi(u, t) = \begin{cases} 1 & \text{se} \quad u > u_0(t), \\ 0 & \text{se} \quad u \leqq u_0(t), \end{cases} \tag{7.88}$$

com $u_0(t)$ determinado pela condição,

$$E_{\eta=0}\{\phi(U, T) \,|\, T = t\} = \alpha, \tag{7.89}$$

é UMP para H_0 contra H_1. Pelo Teorema 7.21, $\phi(u, t)$, definido agora para todo o t, é UMPU em relação à hipótese inicial que envolve os dois parâmetros.

Para obter a função $u_0(t)$ através da condição (7.89), tem-se,

$$P_{\eta=0}[U > u_0(t) \,|\, T = t] = \alpha;$$

quando $\eta = 0$, também $\mu = 0$, e a distribuição de $\mathbf{X}$ condicionada por $\Sigma X_i^2 = t$ é uniforme nesta hiper-esfera. Assim, $\sqrt{t}$ é um parâmetro de escala para essa distribuição condicionada o que leva a concluir que $u_0(t)/\sqrt{t}$ deve ser constante, seja $u_0(t) = k\sqrt{t}$, com k conveniente. Então, tem-se,

$$\phi(u,t) = \begin{cases} 1 & \text{se} \quad u/\sqrt{t} > k, \\ 0 & \text{se} \quad u/\sqrt{t} \leqq k, \end{cases}$$

com k fixado de modo a que o teste não condicionado tenha tamanho α. Depois de alguns cálculos, vem, finalmente, repondo as variáveis iniciais, $(x_1, x_2, \ldots, x_N)$, com $\overline{x} = \Sigma x_i/N$, $s^2 = \Sigma(x_i - \overline{x})^2/N$,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad \sqrt{N-1}\,\overline{x}/s > k', \\ 0 & \text{se} \quad \sqrt{N-1}\,\overline{x}/s \leqq k', \end{cases}$$

onde k' se calcula notando que $\sqrt{N-1}\,\overline{X}/S$ tem distribuição t-Student com $N-1$ graus de liberdade.

Nas Figs. 7.17 e 7.18 faz-se a apresentação do exemplo para $N = 2$, não sendo difícil imaginar a configuração quando N é qualquer [Ferguson (1967)].

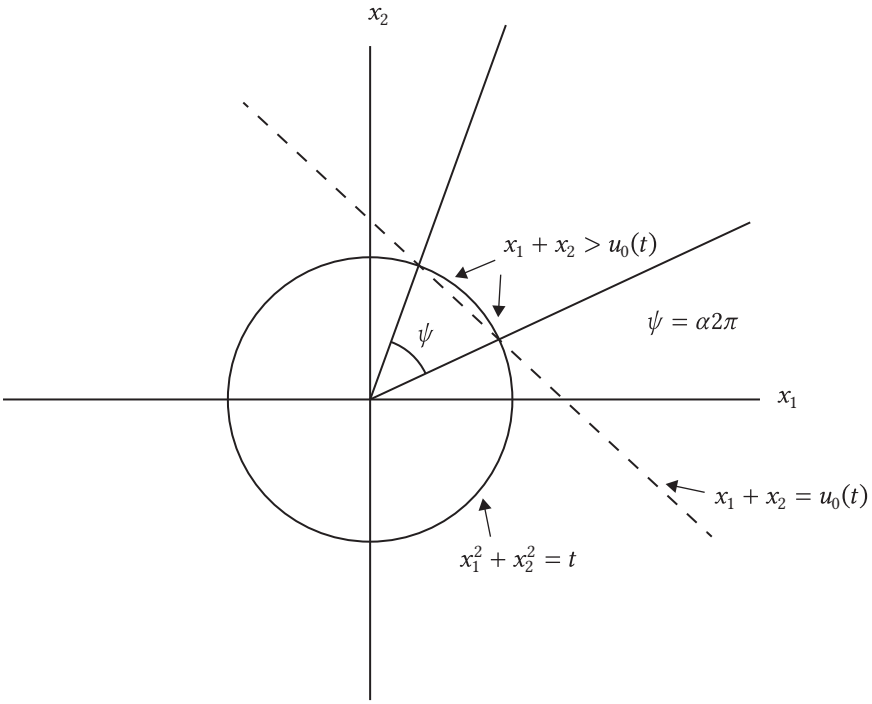


Fig. 7.17

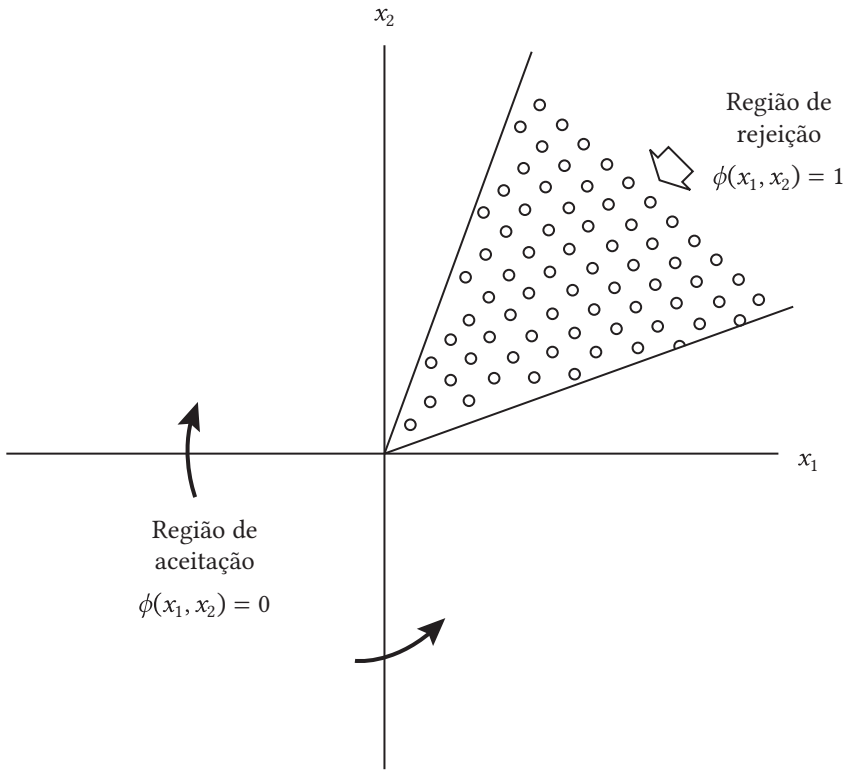


Fig. 7.18

□

Exemplo 7.19 — *Continuação.* Considere-se o ensaio,

$$H_0\colon\ \mu = 0 \quad \text{contra} \quad H_1\colon\ \mu \neq 0;$$

conclui-se que o teste UMPU é dado por,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{se} \quad |\sqrt{N-1}\,\overline{x}/s| > k'', \\ 0 & \text{se} \quad |\sqrt{N-1}\,\overline{x}/s| \leqq k'', \end{cases} \tag{7.90}$$

onde k'' se obtém imediatamente em função de α com auxílio de tabelas da distribuição t-Student com $N-1$ graus de liberdade.

Para se chegar à forma (7.90) há que notar, primeiro, que a condição (7.87) pode agora ser satisfeita quando $u_1(t) = -u_2(t)$, e, segundo, que a condição (7.86) pode ser satisfeita com $u_2(t) = c\sqrt{t}$ para alguma constante c. A Fig. 7.19 compara com a Fig. 7.18.

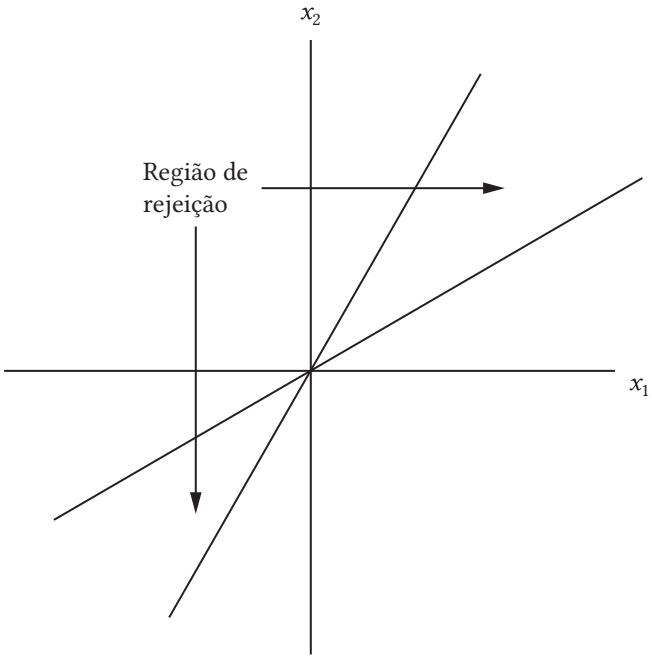


Fig. 7.19

□

As potencialidades de aplicação da doutrina dos testes similares à família exponencial com parametrização natural não ficam esgotadas pelo Teorema 7.21 referente ao ensaio H_0: $\eta \leqq \eta_0$ contra H_1: $\eta > \eta_0$ e pelos teoremas análogos que podem demonstrar-se para H_0: $\eta = \eta_0$ contra H_1: $\eta \neq \eta_0$ [a que aliás se refere o Ex. 7.19 (cont.)] ou para H_0: $\eta_1 \leqq \eta \leqq \eta_2$ contra H_1: $\eta < \eta_1$ ou $\eta > \eta_2$.

Retomando (7.80), distribuição da estatística suficiente, $(U, \mathbf{T})$, quando $\mathbf{X}$ tem distribuição (7.79), raciocínio muito semelhante ao utilizado na demonstração do Teorema 7.21 pode aplicar-se na construção de testes para o ensaio de hipóteses sobre o parâmetro,

$$\eta^* = m_0\eta + m_1\nu_1 + \cdots + m_k\nu_k,\ m_0 \neq 0,$$

obtido por combinação linear dos parâmetros $(\eta, \boldsymbol{\nu})$ de (7.80). De facto, fazendo,

$$U^* = U/m_0,\ T_j^* = T_j - (m_j/m_0)U,$$

tem-se,

$$g(u^*, \mathbf{t}^* \mid \eta^*, \boldsymbol{\nu}) = C_0^*(\eta^*, \boldsymbol{\nu}) \exp\left\{\eta^* u^* + \sum_{j=1}^{k} \nu_j t_j^*\right\} H_0^*(u^*, \mathbf{t}^*),$$

distribuição formalmente análoga a (7.80).

Exemplo 7.20 — Suponha-se que $\mathbf{X} = (X_1, X_2, \dots, X_N)$ e $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{N'})$ são amostras casuais independentes de populações com distribuição de Poisson e médias λ e μ, respectivamente. A função de probabilidade conjunta de $\mathbf{X}$ e $\mathbf{Y}$ é,

$$f(\mathbf{x}, \mathbf{y} \,|, \lambda, \mu) = \exp\{-(N\lambda + N'\mu)\}\lambda^{\Sigma x_i}\mu^{\Sigma y_i}/\Pi x_i!\Pi y_j!,$$

ou, depois de transformação evidente,

$$f(\mathbf{x}, \mathbf{y} \,|\, \lambda, \mu) = \exp\{-(N\lambda + N'\mu)\}\exp\{(\log\lambda)\Sigma x_i + (\log\mu)\Sigma y_j\}/\Pi x_i!\Pi y_i!,$$

devendo notar-se que $S_1 = \Sigma X_i$ e $S_2 = \Sigma Y_j$ são estatísticas suficientes para λ e μ, respectivamente.

A comparação de λ e μ, através de ensaios do tipo,

$$H_0\colon \lambda \leqq \mu \quad \text{contra} \quad H_1\colon \lambda > \mu,$$

ou do tipo,

$$H_0\colon \lambda = \mu \quad \text{contra} \quad H_1\colon \lambda \neq \mu,$$

tem grande interesse prático.

Dado que S_1 e S_2 são variáveis aleatórias com distribuição de Poisson com médias $N\lambda$ e $N'\mu$, respectivamente, a sua função de probabilidade conjunta é,

$$\begin{aligned} g(s_1, s_2 \,|\, \lambda, \mu) &= \exp\{-(N\lambda + N'\mu)\}\exp\{(\log N\lambda)s_1 + (\log N'\mu)s_2\}/s_1!s_2! \\ &= \exp\{-(N\lambda + N'\mu)\}\exp\left\{\left(\log\frac{N'\mu}{N\lambda}\right)s_2 + (\log N\lambda)(s_1 + s_2)\right\}/s_1!s_2!. \end{aligned}$$

Fazendo,

$$\eta = \log(N'\mu/N\lambda) \quad \text{e} \quad \nu = \log N\lambda,$$
$$U = S_2,\ T = S_1 + S_2,$$

tem-se, no que se refere ao segundo factor exponencial do segundo membro, $\exp\{\eta u + \nu t\}$; a hipótese de que $\lambda \leqq \mu$ é equivalente a,

$$H_0\colon \eta \geqq \log(N'/N) = \eta_0 \quad \text{contra} \quad H_1\colon \eta < \eta_0.$$

A distribuição de U $(= S_2)$ condicionada por T $(= S_1 + S_2)$ é uma Binomial,

$$\begin{aligned} g(u \,|\, t, \eta_0) &= P_{\eta_0}(U = u \,|\, T = t) \\ &= P_{\eta_0}(U = u, T = t)/P_{\eta_0}(T = t) \\ &= P_{\eta_0}(S_2 = u, S_1 = t - u)/P_{\eta_0}(S_1 + S_2 = t) \\ &= \binom{t}{u}[1/(1 + e^{\eta_0})]^u[e^{\eta_0}/(1 + e^{\eta_0})]^{t-u} \\ &= \binom{t}{u}p_0^u(1 - p_0)^{t-u}, \quad p_0 = 1/(1 + e^{\eta_0}). \end{aligned}$$

Ensaiar H_0: $\eta \geqq \eta_0$ é equivalente a ensaiar H_0: $p \leqq p_0$ a partir da observação de uma variável aleatória U com distribuição que, para cada t fixo, é Binomial. Tem-se, portanto, o teste UMPU,

$$\phi(u,t) = \begin{cases} 1 & \text{se} \quad u > u_0(t), \\ \gamma(t) & \text{se} \quad u = u_0(t), \\ 0 & \text{se} \quad u < u_0(t), \end{cases}$$

onde $u_0(t)$ e $\gamma(t)$ são determinados pela relação,

$$P_{p_0}(U > u_0(t) \,|\, T = t) + \gamma(t) P_{p_0}(U = u_0(t) \,|\, T = t) = \alpha,$$

isto é,

$$\sum_{u=u_0(t)+1}^{t} \binom{t}{u} p_0^u (1-p_0)^{t-u} + \gamma(t) \binom{t}{u_0(t)} p_0^{u_0(t)} (1-p_0)^{t-u_0(t)} = \alpha;$$

assim, para cada t, $u_0(t)$ e $\gamma(t)$ são calculados com tabelas da Binomial [Lehmann (1959)]. □

Exemplo 7.21 — Para ensaiar a preferência dos consumidores relativamente a dois tipos, A e B, de certo produto, fez-se um inquérito junto de N indivíduos escolhidos ao acaso de uma dada população. Designando por Π_A, Π_B e Π_O, respectivamente, as probabilidades de preferência pelo tipo A, pelo tipo B e de indiferença e por, X, Y e Z o número de indivíduos que optam por essas posições, estas três variáveis aleatórias têm distribuição Multinomial.

$$(N!/x!y!z!)\Pi_A^x \Pi_B^y \Pi_O^z, \quad x + y + z = N,$$

e a hipótese a ensaiar é, H_0: $\Pi_A = \Pi_B$. A função de probabilidade acima pode escrever-se,

$$f(x,z \,|\, \Pi_A, \Pi_O) = \frac{N!}{x!z!\,(N-x-z)!} \left(\frac{\Pi_A}{1-\Pi_A-\Pi_O}\right)^x \left(\frac{\Pi_O}{1-\Pi_A-\Pi_O}\right)^z (1-\Pi_A-\Pi_O)^N,$$

verificando-se que é do tipo exponencial com parametrização natural quando se toma $U = X$, $T = Z$,

$$\eta = \log[\Pi_A/(1-\Pi_A-\Pi_O)] \quad \text{e} \quad \nu = \log[\Pi_O/(1-\Pi_A-\Pi_O)].$$

A hipótese, H_0: $\Pi_A = \Pi_B$ é equivalente a H_0: $\Pi_A = 1-\Pi_A-\Pi_O$, isto é, corresponde a H_0: $\eta = 0$. Existe, portanto, um teste UMPU obtido através do método exposto,

em que se começa por considerar o melhor teste não enviesado condicionado por $Z = z$. Como a distribuição de X condicionada por $Z = z$ é uma Binominal,

$$f(x \mid z, \Pi) = \binom{N-x}{x} \Pi^x (1-\Pi)^{N-z-x},$$
$$x = 0, 1, \ldots, N - z, \quad \Pi = \Pi_A/(\Pi_A + \Pi_B),$$

tem-se o teste UMPU para H_0: $\Pi = 1/2$,

$$\phi(x, z) = \begin{cases} 1 & \text{se} \quad |x - (N-z)/2| > x_0(z), \\ \gamma(z) & \text{se} \quad |x - (N-z)/2| = x_0(z), \\ 0 & \text{se} \quad |x - (N-z)/2| < x_0(z), \end{cases}$$

onde as funções $x_0(z)$ e $\gamma(z)$ são determinadas a partir do valor de α recorrendo a tabelas da Binomial [Lehmann (1959)]. □

7.8 Testes invariantes

O critério da invariância, já ilustrado nos problemas de estimação, é também relevante no ensaio de hipóteses ou problema de bidecisão.

No ensaio de H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$, suponha-se que a função perca é da forma,

$$\begin{aligned} L_0(\theta) &= \begin{cases} L_{00} & \text{se} \quad \theta \in \Theta_0, \\ L_{01} & \text{se} \quad \theta \in \Theta_1, \end{cases} \qquad L_{00} < L_{01}, \\ L_1(\theta) &= \begin{cases} L_{10} & \text{se} \quad \theta \in \Theta_0, \\ L_{11} & \text{se} \quad \theta \in \Theta_1, \end{cases} \qquad L_{10} > L_{11}. \end{aligned} \tag{7.91}$$

Não admitindo esta estrutura simples a análise torna-se bem mais complexa.

Considere-se o caso em que o problema de decisão é invariante em relação ao grupo de transformações, G, e sejam, como habitualmente, $\overline{G}$ e $\tilde{G}$ os grupos de transformações induzidos por G em Θ e $A = \{a_0, a_1\}$.

Um primeiro aspecto é a invariância da família de distribuições, $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$: se X tem distribuição P_θ, gX tem distribuição $P_{\overline{g}\theta}$. Daqui decorre, como caso particular de (5.15),

$$E_\theta\{\phi(gX)\} = E_{\overline{g}\theta}\{\phi(X)\}, \tag{7.92}$$

para todo o teste $\phi \in \Phi$.

Um segundo aspecto é a invariância da função perca,

$$L(\theta, a) = L(\overline{g}\theta, \tilde{g}a), \tag{7.93}$$

que pode exprimir-se em termos mais simples se os valores $L_{i,j}$, $i, j = 0, 1$, de (7.91) forem diferentes, pressuposto que se aceita sem esforço. Assim, (7.93) verifica-se se e somente se $\tilde{g}a = a$ para todo o $\tilde{g} \in \tilde{G}$ e

$$\overline{g}\Theta_0 = \Theta_0,\ \overline{g}\Theta_1 = \Theta_1 \ \text{ para todo o } \ \overline{g} \in \overline{G}. \tag{7.94}$$

Deste modo, $\tilde{G}$ compreende apenas a transformação identidade e a invariância da função perca é traduzida por (7.94).

Em resumo, no ensaio de H_0 contra H_1, um grupo de transformações, G, deixa o problema invariante quando as duas subfamílias,

$$\{P_\theta \,:\, \theta \in \Theta_0\} \quad \text{e} \quad \{P_\theta \,:\, \theta \in \Theta_1\},$$

são invariantes em relação a $\overline{G}$.

No presente contexto, função de decisão ou teste, ϕ, é invariante quando,

$$\phi(gx) = \phi(x) \text{ para todo o } g \in G \text{ e todo o } x \in \mathcal{X}, \tag{7.95}$$

proposição que é simples consequência de (5.18) e de ser $\tilde{g} \equiv \tilde{e}$ [transformação identidade]. A expressão (7.95) indica que ϕ é uma função invariante; em consequência do Teorema 5.7, $\phi(x)$ é invariante se e somente se,

$$\phi(x) = \psi[T(x)],$$

onde $T(X)$ é uma estatística invariante máxima. Quer dizer, a classe dos testes invariantes é equivalente à classe dos testes que são função de uma estatística invariante máxima, classe esta conceitualmente mais simples.

Por outro lado, qualquer teste invariante é construído a partir de $T(X)$ que, pelo Teorema 5.8, tem distribuição que depende apenas de $\nu(\theta)$. Nas aplicações, os invariantes máximos, $T(x)$ e $\nu = \nu(\theta)$, em relação a G e a $\overline{G}$, assumem valores em $\mathbf{R}$, pelo menos em muitos casos, e a família de funções de densidade, $f(t \,|\, \nu)$, tem razão de verosimilhanças monótona. Assim, para ensaiar uma hipótese, H_0: $\nu \leqq \nu_0$ contra H_1: $\nu > \nu_0$ existe um teste UMP entre aqueles que dependem somente de T e, consequentemente, existe um teste UMP invariante.

Exemplo 7.22 — Considerem-se N variáveis aleatórias, $X_1, X_2, \ldots, X_N$, não necessariamente independentes. Pretende ensaiar-se,

$$H_0\colon\ f(\mathbf{x} \,|\, \theta) = f_0(x_1 - \theta, x_2 - \theta, \ldots, x_N - \theta),$$

contra,

$$H_1\colon\ f(\mathbf{x} \,|\, \theta) = f_1(x_1 - \theta, x_2 - \theta, \ldots, x_N - \theta),$$

onde f_0 e f_1 são funções de densidade dadas (f_0 pode ser Normal e f_1 Cauchy) e θ é parâmetro de localização, $-\infty < \theta < +\infty$.

O problema é invariante em relação ao grupo de transformações,

$$g_c(x_1, x_2, \dots, x_N) = (x_1 + c, x_2 + c, \dots, x_N + c), -\infty < c < +\infty,$$

pois, como é evidente, se $(X_1, X_2, \dots, X_N)$ tem função de densidade f_i, $i = 0, 1$, também $(X_1 + c, X_2 + c, \dots, X_N + c)$ tem função de densidade f_i. Como foi indicado no Ex. 5.14,

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_{N-1}), \quad Y_i = X_i - X_N,$$

é estatística invariante máxima; logo, a classe dos testes invariantes é a classe dos testes que são função de $\mathbf{Y}$.

Por outro lado, a distribuição de $\mathbf{Y}$ é independente de θ, pois a transformação induzida, $\overline{g}_c\theta = \theta + c$, é transitiva, isto é, o espaço Θ constitui uma órbita. Mais precisamente, a densidade de $\mathbf{Y}$ segundo H_i, $i = 0, 1$, é

$$\int_{-\infty}^{+\infty} f_i(y_1 + z, y_2 + z, \dots, y_{N-1} + z, z)\, dz;$$

em termos de $\mathbf{Y}$, H_0 e H_1 são hipóteses simples e o Lema de Neyman-Pearson pode aplicar-se. O melhor teste a que esse Lema conduz é independente de θ, e, em consequência, é UMP na classe dos testes invariantes ou UMP invariante. Semelhante teste, que se descreve em forma simplificada, é o seguinte,

$$\phi(y_1, \dots, y_{N-1}) = 1 \quad \text{quando} \quad \frac{\int_{-\infty}^{+\infty} f_1(y_1 + z, \dots, y_{N-1} + z, z)\, dz}{\int_{-\infty}^{+\infty} f_0(y_1 + z, \dots y_{N-1} + z, z)\, dz} > k,$$

ou,

$$\phi(x_1, \dots, x_N) = 1 \quad \text{quando} \quad \frac{\int_{-\infty}^{+\infty} f_1(x_1 + u, \dots, x_N + u)\, du}{\int_{-\infty}^{+\infty} f_0(x_1 + u, \dots, x_N + u)\, du} > k,$$

Lehmann (1959). □

Exemplo 7.23 — Para estudar os efeitos de certo tratamento tomam-se N pares de cobaias. De cada par escolhe-se uma cobaia ao acaso e aplica-se-lhe o tratamento, ficando a outra cobaia a servir de controlo. Seja X_i igual a 1 ou 0 consoante no i-ésimo par o tratamento se revelou benéfico ou não e tome-se $p_i = P(X_i = 1)$. A hipótese de que o tratamento não produz efeito,

$$H_0\colon\ p_i = \frac{1}{2},\ i = 1, 2, \dots, N,$$

deve ser ensaiada contra a alternativa,

$$H_1\colon\ p_i > \frac{1}{2},\ i = 1, 2, \dots, N.$$

O problema é invariante em relação ao grupo de permutações das N variáveis $X_1, X_2, \dots, X_N$, pois a ordem por que se observam os pares é completamente irrelevante para o estudo dos efeitos do tratamento. Um invariante máximo em relação a esse grupo é $Z = \Sigma X_i$, variável que exprime o número de casos em que o tratamento se revelou vantajoso. A função de probabilidade de Z é,

$$P(Z = z) = q_1 q_2 \dots q_N [\Sigma (p_{i_1}/q_{i_1})(p_{i_2}/q_{i_2}) \dots (p_{i_z}/q_{i_z})],$$
$$z = 0, 1, 2, \dots, N,$$

onde $q_i = 1 - p_i$ e o somatório se refere a todas as $\binom{N}{z}$ escolhas dos índices, $i_1 < i_2 < \dots < i_z$.

Para qualquer alternativa particular — hipótese simples,

$$H_1^*\colon\ P(X_i = 1) = p_i^* > \frac{1}{2},\ i = 1, 2, \dots, N,$$

tem-se,

$$\frac{P(Z = z \mid H_1^*)}{P(Z = z \mid H_0)} = \frac{q_1^* \dots q_N^* [\Sigma (p_{i_1}^*/q_{i_1}^*) \dots (p_{i_z}^*/q_{i_z}^*)]}{(1/2)^N \binom{N}{z}} = \psi \Delta(z),$$

onde ψ não depende de z. Pelo Lema de Neyman-Pearson o melhor teste para ensaiar H_0 contra H_1^* é,

$$\phi(z) = \begin{cases} 1 & \text{se} \quad \Delta(z) > k, \\ \gamma & \text{se} \quad \Delta(z) = k, \\ 0 & \text{se} \quad \Delta(z) < k. \end{cases}$$

No entanto, pode mostrar-se que $\Delta(z + 1) > \Delta(z)$, $z = 0, 1, \dots, N - 1$; logo,

$$\phi(z) = \begin{cases} 1 & \text{se} \quad z > k', \\ \gamma & \text{se} \quad z = k', \\ 0 & \text{se} \quad z < k'. \end{cases}$$

com k' e γ determinados em função da dimensão, α, requerida,

$$E\{\phi(Z) \mid H_0\} = \sum_{z = k' + 1}^{N} \binom{N}{z} 2^{-N} + \gamma \binom{N}{k'} 2^{-N} = \alpha.$$

O teste ϕ sendo independente da alternativa particular H_1^* é, consequentemente, UMP invariante para H_0 contra H_1 (Lehmann (1959)]. □

Exemplo 7.24 — Seja $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, X_i I.I.D. $N(\mu, \sigma^2)$; o ensaio da hipótese, H_0: $\sigma^2 \geqq \sigma_0^2$, é invariante em relação ao grupo de transformações,

$$g_c(x_1, x_2, \ldots, x_N) = (x_1 + c, x_2 + c, \ldots, x_N + c), \quad -\infty < c < +\infty.$$

De facto,

$$\overline{g}_c(\mu, \sigma^2) = (\mu + c, \sigma^2),$$

e é de verificação imediata que $\overline{g}_c\Theta_0 = \Theta_0$ e $\overline{g}_c\Theta_1 = \Theta_1$ para todo o c. Como $(T_1, T_2) = [\overline{X}, \Sigma(X_i - \overline{X}^2)]$ é estatística suficiente para (μ, σ^2), o problema pode conduzir-se, sem perca de informação, a partir de (T_1, T_2). Relativamente a (T_1, T_2) o grupo de transformações exprime-se por,

$$g_c T_1 = T_1 + c; \; g_c T_2 = T_2,$$

expressões que mostram ser T_2 uma estatística invariante máxima; note-se que a distribuição de T_2 não varia quando o ponto (μ, σ^2) percorre qualquer órbita de Θ, isto é, a distribuição de T_2 é independente de μ. A classe dos testes invariantes é portanto, a classe dos testes que são função de T_2. Dentro desta classe é fácil de ver que o teste,

$$\phi(t_2) = \begin{cases} 1 & \text{se} \quad t_2 < k, \\ 0 & \text{se} \quad t_2 \geqq k, \end{cases}$$

ou,

$$\phi(x_1, x_2, \ldots, X_N) = \begin{cases} 1 & \text{se} \quad \Sigma(x_i - \overline{x})^2 < k, \\ 0 & \text{se} \quad \Sigma(x_i - \overline{x})^2 \geqq k, \end{cases}$$

é UMP; para tanto basta recordar o Teorema 7.7. Logo, ϕ é UMP invariante.

Repare-se que k tal que, $P_{\sigma_0^2}(T_2 < k) = \alpha$, se determina sem dificuldade porque T_2/σ_0^2 tem distribuição-χ^2 com $N - 1$ graus de liberdade. □

Exemplo 7.25 — Na inspecção de lotes por amostragem a aceitação ou rejeição de cada lote depende do número de elementos defeituosos encontrados numa amostra casual retirada do mesmo lote. Nos esquemas mais simples — inspecção por atributos — cada elemento é directamente classificado como defeituoso ou não defeituoso em resultado de uma avaliação puramente qualitativa; nos casos mais delicados — inspecção por variáveis — há uma prévia avaliação quantitativa de uma característica relevante, representada por uma variável Y, sendo um elemento considerado satisfatório quando, por exemplo, Y excede uma constante, y_0, especificada de antemão.

A probabilidade de obter um elemento defeituoso é, $p = P(Y \leqq y_0)$, e o problema consiste em ensaiar, H_0: $p \geqq p_0$. Quando a distribuição de Y é conhecida

pode aplicar-se a inspecção por variáveis que, para a mesma dimensão da amostra e para testes do mesmo tamanho, conduz a resultados mais potentes do que a inspecção por atributos.

Numa amostra de N elementos suponha-se que as respectivas medidas, Y_1, $Y_2, \ldots, Y_N$, são independentes e identicamente distribuídas com $Y_i \sim N(\nu, \sigma^2)$. Assim,

$$p = (1/\sigma\sqrt{2\pi}) \int_{-\infty}^{y_0} \exp\{-(y-\nu)^2/2\sigma^2\}\, dy = \Phi[(y_0 - \nu)/\sigma],$$

e a hipótese a ensaiar pode escrever-se,

$$H_0\colon (y_0 - \nu)/\sigma \geqq \Phi^{-1}(p_0).$$

Em termos das variáveis, $X_i = Y_i - y_0$, $i = 1, 2, \ldots, N$, $X_i \sim N(\mu, \sigma^2)$, onde se fez, $\mu = \nu - y_0$, tem-se,

$$H_0\colon \mu/\sigma \leqq \theta_0 \quad \text{com} \quad \theta_0 = -\Phi^{-1}(p_0).$$

Para ensaiar H_0 pode considerar-se a redução de dados operada pela estatística suficiente para (μ, σ),

$$(T_1, T_2) = \left[\overline{X}, \sqrt{\{\Sigma(X_i - \overline{X})^2\}}\right],$$

onde T_1 e T_2 são independentes e possuem distribuições bem conhecidas.

O grupo de transformações, $g_c(x_1, x_2, \ldots, x_N) = (cx_1, cx_2, \ldots, cx_N)$, $c > 0$, conduz a,

$$g_c T_1 = cT_1, \; g_c T_2 = cT_2;$$
$$\overline{g}_c(\mu, \sigma) = (c\mu, c\sigma);$$

verifica-se, então, que T_1/T_2 é uma estatística invariante máxima cuja distribuição depende apenas de $\theta = \mu/\sigma$. A classe de testes invariantes é a classe de testes que são função de T_1/T_2, ou, por conveniência, função de

$$T = \frac{\sqrt{N}\, T_1}{T_2/\sqrt{N-1}} = \frac{\overline{X}}{S/\sqrt{N-1}}, \quad S^2 = \Sigma(X_i - \overline{X})^2/N,$$

sendo a hipótese a ensaiar,

$$H_0\colon \theta \leqq \theta_0.$$

A variável aleatória, T, tem distribuição-t não central[6] com parâmetro de não centralidade, $\delta = \sqrt{N}\theta$. Pode mostrar-se que tal distribuição tem RVM; logo, o teste,

$$\phi(t) = \begin{cases} 1 & \text{se} \quad t > k, \\ 0 & \text{se} \quad t \leqq k, \end{cases}$$

[6] Veja Apêndice A.2.

é UMP invariante. Em termos das variáveis iniciais, Y_i, tem-se, que,

$$\phi(y_1, y_2, \dots, y_N) = 1 \text{ quando } \frac{\sqrt{N}(\overline{y} - y_0)}{[\Sigma(y_i - \overline{y})^2/(N-1)]^{1/2}} > k,$$

é UMP invariante para H_0: $(y_0 - \nu)/\sigma \geqq \Phi^{-1}(p_0)$ [Lehmann (1959)]. □

Verifica-se muitas vezes que os testes UMP não enviesados e os testes UMP invariantes são coincidentes; assim sucede no Ex. 7.24. Casos há, todavia, em que existem os primeiros e não os segundos e casos há em que se verifica o inverso. Por exemplo, no Ex. 7.25, em que existem testes UMP invariantes, a existência de testes UMP não enviesados parece ser questão em aberto.

A coincidência acima referida não é acidental. Para enunciar o teorema [Lehmann (1959)] que esclarece este ponto é necessário introduzir o conceito de teste quase invariante. Um teste $\phi \in \Phi$ é quase invariante em relação a um grupo de transformações, G, se,

$$\phi(gx) = \phi(x) \text{ para todo o } x \in \mathcal{X} - N_g \text{ e todo o } g \in G,$$

onde $P_\theta(X \in N_g) = 0$ para todo o $\theta \in \Theta$ e todo o $g \in G$. Note-se que um teste invariante é quase invariante.

Posto isto,

Teorema 7.22 — Se, num problema de ensaio de hipóteses, existe um teste UMP não enviesado único, a menos de conjuntos de medida nula, e existe um teste UMP quase invariante em relação a um grupo de transformações, G, então o segundo é também único, a menos de conjuntos de medida nula, e coincide com o primeiro quase por toda a parte. □□

Os testes UMP não enviesados apresentam vantagens em relação aos testes UMP invariantes no domínio da admissibilidade. Não é difícil mostrar que os testes UMPU são admissíveis: se ϕ' domina estritamente ϕ que é UMPU cai-se numa contradição porque ϕ' é também não enviesado. Em contrapartida a admissibilidade dos testes UMP invariantes tem de ser estabelecida caso a caso [veja-se Lehmann (1959) para mais pormenores].

Finalmente uma referência ao Teorema de Hunt-Stein.

Seja G um grupo de transformações sobre $\mathcal{X}$ que deixa o problema de ensaiar H_0: $\theta \in \Theta_0$ contra H_1: $\theta \in \Theta_1$ invariante. Uma sucessão de medidas de probabilidade em G, seja $\{\nu_n\}$, diz-se assintoticamente invariante à direita se,

$$\lim_{n \to \infty} [\nu_n(B \circ g) - \nu_n(B)] = 0,$$

para qualquer B (mensurável) $\subset G$ e todo o $g \in G$.

Teorema 7.23 — Se o problema de ensaiar H_0 contra H_1 é invariante em relação ao grupo G e se existe em G uma sucessão de medidas de probabilidade assintoticamente invariante à direita, então existe um teste invariante que é minimax [veja-se Lehmann (1959) e a excelente discussão de Giri (1983)]. □□

Repare-se, por comparação com o Teorema 5.9 e com as considerações subsequentes, que para garantir a existência de uma função de decisão minimax — no caso particular em que o problema é de bidecisão ou ensaio de hipóteses — não se pede que o grupo de transformações, G, seja finito ou que seja compacto; pede-se apenas que exista uma sucessão de medidas de probabilidade assintoticamente invariante à direita.

7.9 Invariância no ensaio de hipóteses lineares[7]

Um importante domínio de aplicação do princípio da invariância é o das hipóteses lineares no qual se verifica, com frequência, a não existência de testes UMP não enviesados.

A hipótese linear geral refere-se a N variáveis aleatórias independentes, X_1, $X_2, \ldots, X_N$, com distribuição Normal, médias, $\mu_1, \mu_2, \ldots, \mu_N$, e variância comum, σ^2.

Supõe-se, anteriormente ao ensaio estatístico, que o vector das médias,

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)' = (E\{X_1\}, E\{X_2\}, \ldots, E\{X_N\})' = E\{\mathbf{X}\}, \tag{7.96}$$

pertence a um sub-espaço linear k-dimensional de $\mathbf{R}^N$, $k < N$, seja L_Ω. Este pressuposto, $\boldsymbol{\mu} \in L_\Omega$, pode escrever-se

$$\begin{aligned} \mu_1 &= b_{11}\nu_1 + b_{12}\nu_2 + \cdots + b_{1k}\nu_k, \\ \mu_2 &= b_{21}\nu_1 + b_{22}\nu_2 + \cdots + b_{2k}\nu_k, \\ &\cdots \\ \mu_N &= b_{N1}\nu_1 + b_{N2}\nu_2 + \cdots + b_{Nk}\nu_k, \end{aligned} \tag{7.97}$$

onde $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_k)'$ é um vector arbitrário e os b_{ij} formam uma matriz $\mathbf{B} = [b_{ij}]$, $N \times k$, de característica k. Em termos matriciais (7.97) traduz-se por,

$$\boldsymbol{\mu} = \mathbf{B}\boldsymbol{\nu}. \tag{7.98}$$

A hipótese linear a cujo ensaio se pretende proceder consiste em admitir que $\boldsymbol{\mu}$ é ponto de um sub-espaço $(k-r)$-dimensional, L_ω, do sub-espaço L_Ω, com $1 \leqq r \leqq k$,

$$H_0\colon \mathbf{C}\boldsymbol{\nu} = \mathbf{0}, \tag{7.99}$$

[7] A presente secção é de natureza complementar.

onde $\mathbf{C} = [c_{ij}]$ é uma matriz $r \times k$ de característica r e $\mathbf{0}$ é o vector com r componentes iguais a zero.

Exemplo 7.26 — Considerem-se duas amostras independentes, $(X_{11}, X_{12}, \dots, X_{1N_1})$ e $(X_{21}, X_{22}, \dots, X_{2N_2})$, X_{1i} I.I.D., $X_{1i} \sim N(\mu_1, \sigma^2)$, X_{2i} I.I.D., $X_{2i} \sim N(\mu_2, \sigma^2)$.

O clássico problema de ensaiar a igualdade das médias,

$$H_0\colon \mu_1 = \mu_2,$$

é caso particular de hipótese linear. Com efeito, tomando,

$$\begin{aligned}\mathbf{X} &= (X_{11}, \dots, X_{1N_1}, X_{21}, \dots, X_{2N_2})',\\ \boldsymbol{\mu} &= (\mu_1, \dots, \mu_1, \mu_2, \dots, \mu_2)',\end{aligned}$$

tem-se,

$$\boldsymbol{\mu} = \mu_1(1, \dots, 1, 0, \dots, 0)' + \mu_2(0, \dots, 0, 1, \dots, 1)';$$

assim, $\boldsymbol{\mu}$ é ponto do sub-espaço gerado pelos vectores,

$$(1, \dots, 1, 0, \dots, 0)' \quad \text{e} \quad (0, \dots, 0, 1, \dots, 1)',$$

e, portanto, $k = 2$, $\boldsymbol{\nu} = (\mu_1, \mu_2)$. Em correspondência com (7.98) vem,

$$\boldsymbol{\mu} = \begin{bmatrix} 1 & 0 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}.$$

Por outro lado, H_0 é equivalente a,

$$H_0\colon (1, -1) \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = 0;$$

logo, $\mathbf{C} = (1, -1)$ e $r = 1$. Quer dizer, segundo H_0, $\boldsymbol{\mu}$, é ponto do sub-espaço unidimensional gerado pelo vector $(1, 1, \dots, 1)'$ com $N_1 + N_2$ componentes. □

Exemplo 7.27 — Outro caso particular de hipótese linear ocorre na análise da variância (classificação simples ou «*oneway*»). Pretende estudar-se o efeito de k tratamentos através de uma experiência na qual cada tratamento i, $i = 1, 2, \dots, k$, é aplicado a uma sub-amostra ou grupo de N_i elementos. Designe, X_{ij}, a observação correspondente ao j-ésimo elemento do i-ésimo grupo; tem-se o seguinte quadro,

Grupos					
1	2	...	i	...	k
X_{11}	X_{21}	...	X_{i1}	...	X_{k1}
X_{12}	X_{22}	...	X_{i2}	...	X_{k2}
.	.	...	.	...	.
X_{1N_1}	X_{2N_2}	...	X_{iN_i}	...	X_{kN_k}

A dimensão total da amostra é $N = N_1 + N_2 + \cdots + N_k$. Segundo o modelo adoptado na análise da variância, X_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, N_i$, são variáveis aleatórias I.I.D., $X_{ij} \sim N(\mu_i, \sigma^2)$. Pode então escrever-se,

$$X_{ij} = \mu_i + \xi_{ij},\ i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, N_i, \tag{7.100}$$

onde ξ_{ij} são I.I.D., $\xi_{ij} \sim N(0, \sigma^2)$. A expressão (7.100) indica que cada X_{ij} é soma de duas componentes, a primeira representando o efeito do tratamento e a segunda o «erro de observação».

Fazendo $\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{1N_1}, \ldots, X_{kN_k})'$, vem,

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_k, \ldots, \mu_k)',$$

e $\boldsymbol{\mu}$ é ponto do sub-espaço k-dimensional de $\mathbf{R}^N$ gerado pelos k vectores,

$$\begin{aligned}
&(1, \ldots, 1, 0, \ldots, 0, \ldots, 0, \ldots, 0)',\\
&(0, \ldots, 0, 1, \ldots, 1, \ldots, 0, \ldots, 0)',\\
&\ldots\\
&(0, \ldots, 0, 0, \ldots, 0, \ldots, 1, \ldots, 1)'.
\end{aligned}$$

A hipótese a ensaiar é a de que não há diferença entre os efeitos dos k tratamentos,

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_k,$$

isto é, que $\boldsymbol{\mu}$ é ponto do sub-espaço unidimensional gerado pelo vector com N componentes iguais a 1. Em termos da matriz $\mathbf{C}$ tem-se, com $\boldsymbol{\nu} = (\mu_1, \mu_2, \ldots, \mu_k)$,

$$H_0\colon\ \mathbf{C}\boldsymbol{\nu} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ . & . & . & \ldots & . & . \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \ldots \\ \mu_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}$$

sendo $\mathbf{C}$ matriz $(k-1) \times k$ de característica $k - 1$. □

Por meio de uma mudança de variável pode reduzir-se a hipótese linear à chamada forma canónica. Com $\mathbf{X} = (X_1, X_2, \dots, X_N)'$, seja,

$$\mathbf{Y} = \mathbf{D}\mathbf{X}, \tag{7.101}$$

onde $\mathbf{D}$ é uma matriz ortogonal[8], $N \times N$, construída do seguinte modo:

(a) as primeiras r linhas são vectores de L_Ω perpendiculares a L_ω;
(b) as linhas $r+1$ a k são vectores gerando L_ω;
(c) as últimas $N-k$ linhas são vectores perpendiculares a L_Ω (i.e., perpendiculares às colunas de $\mathbf{B}$).

Note-se que os vectores das primeiras k linhas de $\mathbf{D}$ geram o sub-espaço L_Ω e, como é evidente, as N linhas de $\mathbf{D}$ formam uma base de $\mathbf{R}^N$.

A transformação ortogonal, $\mathbf{Y} = \mathbf{D}\,\mathbf{X}$, conduz a variáveis Y_i I.I.D. com $Y_i \sim N(\lambda_i, \sigma^2)$, onde,

$$\boldsymbol{\lambda} = \mathbf{D}\boldsymbol{\mu} = \mathbf{D}\,\mathbf{B}\boldsymbol{\nu},$$

e, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)'$.

Fazendo a partição de $\mathbf{D}$ em blocos de acordo com a forma de construção referida nas alíneas (a), (b) e (c), sai,

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \end{bmatrix}, \qquad \mathbf{DB} = \begin{bmatrix} \mathbf{D}_1\mathbf{B} \\ \mathbf{D}_2\mathbf{B} \\ \mathbf{D}_3\mathbf{B} \end{bmatrix};$$

[8] Tem-se,

$$\begin{aligned} Y_1 &= d_{11}X_1 + d_{12}X_2 + \cdots + d_{1N}X_N, \\ Y_2 &= d_{21}X_1 + d_{22}X_2 + \cdots + d_{2N}X_N, \\ &\cdots \\ Y_N &= d_{N1}X_1 + d_{N2}X_2 + \cdots + d_{NN}X_N, \end{aligned}$$

verificando a matriz $N \times N$, $\mathbf{D} = [d_{ij}]$, as condições, $\mathbf{D}'\mathbf{D} = \mathbf{D}\mathbf{D}' = I_N$ (matriz identidade de ordem N), ou seja,

$$\sum_{i=1}^{N} d_{ij}d_{ik} = \begin{cases} 1 & \text{se} \quad j = k, \\ 0 & \text{se} \quad j \neq k, \end{cases}$$

$$\sum_{i=1}^{N} d_{ji}d_{ki} = \begin{cases} 1 & \text{se} \quad j = k, \\ 0 & \text{se} \quad j \neq k. \end{cases}$$

Como é sabido, uma transformação ortogonal corresponde a uma rotação em torno da origem, que, portanto, deixa invariantes as distâncias à origem:

$$\Sigma X_i^2 = \Sigma Y_i^2.$$

agora: em consequência de (c) tem-se $\mathbf{D}_3\mathbf{B} = \mathbf{0}$, quer dizer,

$$\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_N = 0; \tag{7.102}$$

se a hipótese H_0 é verdadeira, vem, de (a) e de (7.99),

$$H_0\colon\ \lambda_1 = \lambda_2 = \cdots = \lambda_r = 0.^{9} \tag{7.103}$$

Em resumo, a mudança de variável (7.101) permite reformular o problema da hipótese linear do seguinte modo: a partir da observação das variáveis aleatórias, $Y_1, Y_2, \ldots, Y_N$, independentes, com distribuição Normal, todas com a mesma variância, σ^2, possuindo as últimas $N - k$ média igual a zero — veja-se (7.102) — ensaiar a hipótese de as primeiras r, $1 \leqq r \leqq k$, possuírem também média igual a zero — veja (7.103). Depois de reformulado o problema da hipótese linear apresenta-se na chamada forma canónica, e é invariante em relação ao grupo de transformações gerado pelos seguintes grupos de transformações:[10]

G_1: $g(\mathbf{y}) = \mathbf{y}^o$ onde $y_i^o = y_i + c_i$ para $i = r+1, \ldots, k$, $y_i^o = y_i$ para outros valores de i;
G_2: $g(\overset{*}{\mathbf{y}}) = \mathbf{H}\overset{*}{\mathbf{y}}$ onde $\mathbf{H}$ é uma matriz ortogonal de ordem r e $\overset{*}{\mathbf{y}} = (y_1, \ldots, y_r)$;
G_3: $g(\mathbf{y}) = m\mathbf{y}$ onde $m > 0$.

Começa por reduzir-se o problema por meio de estatísticas suficientes: como, $\lambda_{k+1} = \cdots = \lambda_N = 0$, tem-se que,

$$Y_1, Y_2, \ldots, Y_k \quad \text{e} \quad \sum_{i=k+1}^{N} Y_i^2,$$

são suficientes para $\lambda_1, \lambda_2, \ldots, \lambda_k$ e σ^2.

[9] Quer dizer: notando ser,

$$\begin{aligned}
\lambda_1 &= d_{11}\mu_1 + d_{12}\mu_2 + \cdots + d_{1N}\mu_N,\\
\lambda_2 &= d_{21}\mu_1 + d_{22}\mu_2 + \cdots + d_{2N}\mu_N,\\
&\cdots\\
\lambda_N &= d_{N2}\mu_1 + d_{N_2\mu_2} + \cdots + d_{NN}\mu_N,
\end{aligned}$$

e atendendo às condições (a), (b) e (c) que presidiram à construção da matriz ortogonal $\mathbf{D}$, tem-se,

(1) $\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_N = 0$ se e somente se $\boldsymbol{\mu} \in L_\Omega$;
(2) $\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_N = \lambda_1 = \lambda_2 = \cdots = \lambda_r = 0$ se e somente se $\boldsymbol{\mu} \in L_\omega$.

[10] O grupo G gerado pelos subgrupos, G_1, G_2 e G_3, entende-se como o menor grupo contendo G_1, G_2 e G_3. Demonstra-se [Lehmann (1959)] que se o processo de determinação de um invariante máximo em passos sucessivos é viável — como aqui sucede — então conduz a um invariante máximo em relação a G, no caso presente, $\sum_{i=1}^{r} Y_i^2 / \sum_{i=k+1}^{N} Y_i^2$.

Limitando a análise aos testes invariantes e recorrendo primeiro a G_1 verifica-se que,

$$Y_1, Y_2, \ldots, Y_r \quad \text{e} \quad \sum_{i=k+1}^{N} Y_i^2,$$

constitui uma estatística $r+1$-dimensional invariante máxima. A aplicação de G_2 conduz à estatística invariante máxima,

$$\sum_{i=1}^{r} Y_i^2 \quad \text{e} \quad \sum_{i=k+1}^{N} Y_i^2;$$

finalmente, aplicação de G_3 conduz à estatística invariante máxima,

$$\sum_{i=1}^{r} Y_i^2 / \sum_{i=k+1}^{N} Y_i^2$$

sendo mais conveniente tomar,

$$F = \frac{\sum_{i=1}^{r} Y_i^2/r}{\sum_{i=k+1}^{N} Y_i^2/(N-k)}. \tag{7.104}$$

A variável, F, é, portanto, função da estatística invariante máxima em relação ao grupo de transformações gerado pela aplicação sucessiva de G_1, G_2 e G_3 [veja-se Lehmann (1959)].

A variável aleatória, F, tem distribuição-F não central[11] com r e $N-k$ graus de liberdade e parâmetro de não centralidade,

$$\gamma^2 = \sum_{i=1}^{r} \lambda_i^2/\sigma^2.$$

A hipótese H_0: $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 0$ transforma-se na hipótese simples, H_0: $\gamma^2 = 0$ sendo a alternativa H_1: $\gamma^2 > 0$. Como pode demonstrar-se que a distribuição-F não central possui RVM em γ^2, a análise com base no Lema de Neyman-Pearson, várias vezes aqui repetida, leva à conclusão que um teste $\phi(\mathbf{y})$ tal que $\phi(\mathbf{y}) = 1$ quando $F > k'$ e $\phi(\mathbf{y}) = 0$ quando $F \leqq k'$, é UMP invariante para o ensaio de H_0. Por outro lado, quando H_0 é verdadeira, $\gamma^2 = 0$, a variável (7.104) tem distribuição-F com r e $N-k$ graus de liberdade, sendo fácil determinar k'. Para obter a função potência há que recorrer à distribuição-F não central.

[11] Veja-se Apêndice A.3.

Repare-se que se $r = 1$ a condição $F > k'$ pode escrever-se,

$$|Y_1| \Big/ \Big| \sum_{i=k+1}^{N} Y_i^2/(N-k) \Big|^{1/2} > k'',$$

e corresponde a um teste bilateral operado com a distribuição-t. Este teste é também UMP não enviesado. Quando $r > 1$, não existe qualquer teste UMP não enviesado para a hipótese linear, embora existam, como se viu, testes UMP invariantes.

Nas aplicações práticas não há necessidade de proceder à mudança de variável que levou à forma canónica. A expressão de F em termos das variáveis iniciais, $X_1, X_2, \ldots, X_N$, consegue fazer-se de modo expedito com o auxílio do método dos mínimos quadrados. Com efeito, mostra-se que (7.104) é equivalente a,

$$F = \frac{\left[\sum_{i=1}^{N}(X_i - \hat{\hat{\mu}}_i)^2 - \sum_{i=1}^{N}(X_i - \hat{\mu}_i)^2\right]/r}{\sum_{i=1}^{N}(X_i - \hat{\mu}_i)^2/(N-k)} \tag{7.105}$$

onde: $\hat{\mu}_i$ se obtém minimizando a soma de quadrados,

$$\sum_{i=1}^{N}(X_i - \mu_i)^2,$$

com a condição de $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$ ser ponto de L_Ω; $\hat{\hat{\mu}}_i$ se obtém minimizando a mesma soma de quadrados com a condição de $\boldsymbol{\mu}$ ser ponto de L_ω.

Exemplo 7.28 — Retome-se o Ex. 7.27. Escreva-se, para melhor ligação com a análise antecedente,

$$X_{ij} = \mu_{ij} + \xi_{ij}, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, N_i.$$

As condições,

$$\mu_{ij} = \mu_{i\ell}(= \mu_i), \quad i = 1, 2, \ldots, k, \quad j, \ell = 1, 2, \ldots, N_i,$$

são anteriores ao ensaio e exprimem que,

$$\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \ldots, \mu_{1N_1}, \ldots, \mu_{kN_k})',$$

é ponto de L_Ω. Portanto, tem-se, $\hat{\mu}_{ij} = \hat{\mu}_i$, saindo os $\hat{\mu}_i$ da minimização de

$$\sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{ij} - \mu_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_{i\bullet})^2 + \sum_{i=1}^{k} N_i(\overline{X}_{i\bullet} - \mu_i)^2,$$

com $\overline{X}_{i\bullet} = \sum_{j=1}^{N_i} X_{ij}/N_i$, notação corrente na análise da variância. A minimização é obtida quando se toma $\hat{\mu}_i = \overline{X}_{i\bullet}$. Por outro lado, as condições,

$$\mu_{ij} = \mu, \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, N_i,$$

decorrem de H_0 e exprimem que μ é ponto de L_ω. Portanto, tem-se, $\hat{\hat{\mu}}_{ij} = \hat{\hat{\mu}}$, saindo $\hat{\hat{\mu}}$, da minimização de,

$$\sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{ij} - \mu)^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_{\bullet\bullet})^2 + N(\overline{X}_{\bullet\bullet} - \mu)^2,$$

com $\overline{X}_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{N_i} X_{ij}/N$, $N = N_1 + N_2 + \cdots + N_k$. A minimização consegue-se com $\hat{\hat{\mu}} = \overline{X}_{\bullet\bullet}$.

Utilizando a expressão,

$$F = \frac{\sum_{i=1}^{N}(\hat{\mu}_i - \hat{\hat{\mu}}_i)^2/r}{\sum_{i=1}^{N}(X_i - \hat{\mu}_i)^2/(N-k)}, \tag{7.106}$$

que se verifica ser equivalente a (7.105), tem-se,[12]

$$F = \frac{\sum_{i=1}^{k} N_i(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_{i\bullet})^2/(N-k)}, \quad (k-1=r) \tag{7.107}$$

variável que quando H_0 é verdadeira tem distribuição-F com $k-1$ e $N-k$ graus de liberdade. A expressão (7.107) é a que se encontra na análise da variância no caso da classificação simples. O que se não pode explicar na abordagem elementar é que se trata de um teste UMP invariante. A hipótese H_0 de que as médias são todas iguais é, portanto, rejeitada quando sai $F > k'$, com k' determinado em função do tamanho do teste. Para estabelecer a potência do teste há que recorrer à distribuição-F não central com $k-1$ e $N-k$ graus de liberdade e parâmetro de não centralidade,

$$\gamma^2 = \left\{\sum_{i=1}^{k} N_i\left[\mu_i - \left(\sum_{i=1}^{k} N_i\mu_i/N\right)\right]^2\right\}/\sigma^2.$$

□

[12] Depois de notar que relativamente a (7.105) e (7.106) se tem $\mathbf{X} = (X_1, X_2, \dots, X_N)$ e que relativamente a (7.107) se tem $\mathbf{X} = (X_{11}, X_{12}, \dots, X_{1N_1}, \dots, X_{kN_k})$, $N_1 + N_2 + \cdots + N_k = N$.

CAPÍTULO 8

PROBLEMAS DE MULTIDECISÃO

8.1 Introdução

Quando o espaço de acções é finito, $A = \{a_1, a_2, \dots, a_m\}, m \geqq 3$, fala-se em problemas de multidecisão. Vários exemplos foram anteriormente apresentados; estudam-se agora algumas proposições gerais.

Uma função de decisão pura, $\delta \in D$, adequada ao presente caso pode interpretar-se com uma partição mensurável de $\mathfrak{X}$,

$$\{W_1^\delta, W_2^\delta, \dots, W_m^\delta\}, \quad \cup\, W_i^\delta = \mathfrak{X},\ W_i^\delta \cap\ W_j^\delta = \emptyset,\ i \neq j, \tag{8.1}$$

tal que,

$$W_i^\delta = \{x\ :\ x \in \mathfrak{X},\ \delta(x) = a_i\}, \quad i = 1, 2, \dots, m. \tag{8.2}$$

As funções mistas, $\delta^* \in D^*$, revelam-se, como era de esperar em face do que foi dito na secção 7.1, de trato difícil, porquanto correspondem a uma casualização sobre a família das m-partições mensuráveis de $\mathfrak{X}$. Pelo contrário, a exemplo do que sucede nos problemas de bidecisão, as funções de decisão aleatórias, $\phi \in \Phi$, têm estrutura simples,

$$\{\phi(a_1 \mid x),\ \phi(a_2 \mid x), \dots,\ \phi(a_m \mid x)\},\ 0 \leqq \phi(a_i \mid x) \leqq 1,\ \sum_{i=1}^{m} \phi(a_i \mid x) = 1, \tag{8.3}$$

onde $\phi(a_i \mid x)$ indica a probabilidade de o decisor escolher a_i quando observa $X = x$ e usa a função de decisão aleatória, ϕ.

Para simplificar toma-se $\phi(a_i \mid x) = \phi_i(x)$, sendo,

$$0 \leqq \phi_i(x) \leqq 1,\ \Sigma\phi_i(x) = 1. \tag{8.4}$$

Quando se emprega ϕ a função risco assume a forma,

$$\begin{aligned} R(\theta, \phi) &= \Sigma L_i(\theta) \int \phi_i(x) f(x \mid \theta)\, dx \\ &= \Sigma L_i(\theta) E_\theta\{\phi_i(X)\}, \end{aligned} \tag{8.5}$$

onde, $L_i(\theta) = L(\theta, a_i)$, $i = 1, 2, \ldots, m$, e se escreve $R(\theta, \phi)$ em vez de $\hat{R}(\theta, \phi)$.

8.2 Procedimentos monótonos

No domínio da multidecisão têm particular relevo os problemas monótonos.

Os problemas monótonos têm a seguinte caracterização: Θ é um intervalo de $\mathbf{R}$, e, para certa ordenação das acções, seja $a_1, a_2, \ldots, a_m$, existem $m - 1$ valores de θ, $\theta_1 \leqq \theta_2 \leqq \ldots \leqq \theta_{m-1}$, tais que,

$$\begin{array}{ll} L_i(\theta) - L_{i+1}(\theta) \leqq 0 & \text{para} \quad \theta < \theta_i, \\ L_i(\theta) - L_{i+1}(\theta) \geqq 0 & \text{para} \quad \theta > \theta_i. \end{array} \qquad i = 1, 2, \ldots, m-1, \tag{8.6}$$

Esta caracterização implica que, para $\theta_{i-1} < \theta < \theta_i$ [podendo ser $\theta_0 = -\infty$ ou $\theta_m = +\infty$], a acção mais favorável é a_i, porquanto,

$$L_i(\theta) = \inf_{1 \leqq j \leqq m} L_j(\theta) \quad \text{para} \quad \theta_{i-1} < \theta < \theta_i.$$

Na Fig. 8.1 exemplificam-se as funções perca típicas dos problemas monótonos.

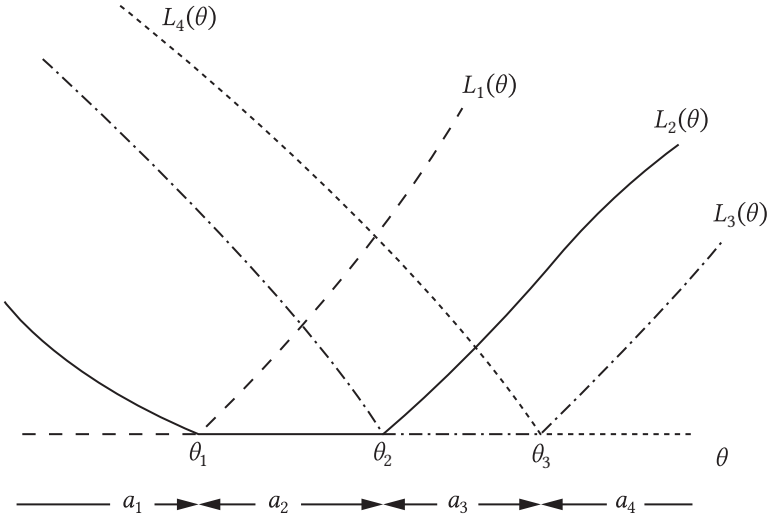


Fig. 8.1

As relações (8.6) constituem uma generalização de (7.44)–(7.45); pode, portanto, dizer-se que os problemas monótonos de multidecisão generalizam os ensaios unilaterais tratados na secção 7.3. Para estes ensaios deduziu-se que as funções de decisão aleatórias da forma (7.39) tinham propriedades importantes; no caso presente, a (7.39) correspondem funções de decisão monótonas ou procedimentos monótonos, definidos por números x_i, γ_i, γ_i', $i = 1, 2, \ldots, m-1$, com,

$$-\infty = x_0 \leqq x_1 \leqq x_2 \leqq \ldots \leqq x_m = +\infty, \quad 0 \leqq \gamma_i \leqq 1,\ 0 \leqq \gamma_i' \leqq 1,$$

e,

$$\phi_i(x) = \begin{cases} 0 & \text{se} \quad x < x_{i-1}, \\ \gamma_{i-1}' & \text{se} \quad x = x_{i-1}, \\ 1 & \text{se} \quad x_{i-1} < x < x_i, \\ \gamma_i & \text{se} \quad x = x_i, \\ 0 & \text{se} \quad x > x_i, \end{cases} \tag{8.7}$$

para $i = 1, 2, \ldots, m$.

No exemplo apresentado na Fig. 8.1, em que $A = \{a_1, a_2, a_3, a_4\}$, um procedimento monótono tem a seguinte estrutura:

- para $x < x_1$ opta-se por a_1;
- para $x = x_1$ casualiza-se e decide-se por a_1 com probabilidade γ_1 e por a_2 com probabilidade $\gamma_1' = 1 - \gamma_1$;
- para $x_1 < x < x_2$ opta-se por a_2;
- para $x = x_2$ casualiza-se e decide-se por a_2 com probabilidade γ_2 e por a_3 com probabilidade $\gamma_2' = 1 - \gamma_2$;
- para $x_2 < x < x_3$ opta-se por a_3;
- para $x = x_3$ casualiza-se e decide-se por a_3 com probabilidade γ_3 e por a_4 com probabilidade $\gamma_3' = 1 - \gamma_3$;
- para $x > x_3$ opta-se por a_4.

O principal resultado da presente secção é o seguinte,

Teorema 8.1 — Se a variável aleatória, X, tem distribuição com RVM, a classe dos procedimentos monótonos é essencialmente completa relativamente aos problemas monótonos de multidecisão.

Dem. Seja $\psi = \{\psi_1(x), \psi_2(x), \ldots, \psi_m(x)\}$ um elemento de Φ e tome-se,

$$\xi_j(x) = \sum_{i=j+1}^{m} \psi_i(x),\ j = 0, 1, \ldots, m-1,\ \xi_m(x) \equiv 0. \tag{8.8}$$

Dado que $0 \leqq \xi_j(x) \leqq 1$, $\xi_j(x)$ pode considerar-se como teste para ensaio, por exemplo, da hipótese, H_0: $\theta \leqq \theta_j$. Mas, pelo Teorema 7.7, existe para este ensaio um teste, seja $\zeta_j(x)$, da forma,

$$\zeta_j(x) = \begin{cases} 1 & \text{se } x > x_j, \\ \gamma_j'' & \text{se } x = x_j, \\ 0 & \text{se } x < x_j, \end{cases} \qquad j = 1, 2, \ldots, m-1, \tag{8.9}$$

pelo menos tão bom como $\xi_j(x)$ e tal que,

$$E_\theta\{\zeta_j(X)\} - E_\theta\{\xi_j(X)\} \begin{cases} \leqq 0 & \text{para} \quad \theta < \theta_j, \\ = 0 & \text{para} \quad \theta = \theta_j, \\ \geqq 0 & \text{para} \quad \theta > \theta_j. \end{cases} \tag{8.10}$$

Adicionalmente, seja $\zeta_0(x) \equiv 1$ e $\zeta_m(x) \equiv 0$. Dado que,

$$E_{\theta_{j-1}}\{\zeta_{j-1}(X)\} = E_{\theta_{j-1}}\{\xi_{j-1}(X)\} \geqq E_{\theta_{j-1}}\{\xi_j(X)\} \geqq E_{\theta_{j-1}}\{\zeta_j(X)\},$$

pode tomar-se, modificando x_j e γ_j'' se necessário, $\zeta_{j-1}(x) \geqq \zeta_j(x)$ para todo o x, isto é, pode tomar-se, $x_1 \leqq x_2 \leqq \ldots \leqq x_m$. Assim,

$$\phi_j(x) = \zeta_{j-1}(x) - \zeta_j(x), \ j = 1, 2, \ldots, m,$$

é componente de um procedimento monótono.

Importa agora mostrar que ϕ é pelo menos tão bom como ψ. De (8.5) tem-se,

$$\begin{aligned} R(\theta, \psi) - R(\theta, \phi) &= \sum_{i=1}^{m} L_i(\theta)[E_\theta\{\psi_i(X)\} - E_\theta\{\phi_i(X)\}] \\ &= \sum_{i=1}^{m} L_i(\theta)\{[E_\theta\{\zeta_i(X)\} - E_\theta\{\xi_i(X)\}] - [E_\theta\{\zeta_{i-1}(X)\} - E_\theta\{\xi_{i-1}(X)\}]\}, \end{aligned}$$

donde,

$$R(\theta, \psi) - R(\theta, \phi) = \sum_{i=1}^{m-1} [L_i(\theta) - L_{i+1}(\theta)][E_\theta\{\zeta_i(X)\} - E_\theta\{\xi_i(X)\}]. \tag{8.11}$$

Os termos do somatório do segundo membro são não negativos, pois, para $\theta < \theta_i$, $L_i(\theta) - L_{i+1}(\theta) \leqq 0$ e $E_\theta\{\zeta_i(X)\} - E_\theta\{\xi_i(X)\} \leqq 0$; para $\theta > \theta_i$ verificam-se as desigualdades com $\geqq$; para $\theta = \theta_i$, tem-se, por (8.9), que todos os termos são nulos. A demonstração fica completa. □□

Exemplo 8.1 — Considere-se a variável aleatória, X, com distribuição Binomial, $B(N;\theta)$. Pretendem ensaiar-se as três alternativas,

$$H_1 : 0 \leqq \theta \leqq 0{,}4;\ H_2 : 0{,}4 < \theta < 0{,}6;\ H_3 : 0{,}6 \leqq \theta \leqq 1,$$

num caso em que $N = 10$. A função perca,

Estados	Acções		
	a_1	a_2	a_3
$0 \leqq \theta \leqq 0{,}4$	0	10	100
$0{,}4 < \theta < 0{,}6$	10	0	10
$0{,}6 \leqq \theta \leqq 1$	100	10	0

corresponde a um problema monótono. Além disso, a distribuição Binomial tem RVM e o Teorema 8.1 pode aplicar-se. Suponha-se que é proposta a seguinte função de decisão,

	$\psi_1(x)$	$\psi_2(x)$	$\psi_3(x)$
$x = 0, 1, 2$	0,6	0,3	0,1
$x = 3, 4, 5, 6, 7$	0,2	0,6	0,2
$x = 8, 9, 10$	0,1	0,3	0,6

O aspecto fulcral do Teorema 8.1 é indicar a forma como pode construir-se uma função de decisão, ϕ, pelo menos tão boa como ψ. Para o efeito, começam por formar-se as funções, $\xi_j(x)$, definidas por (8.8):

$$\xi_0(x) \equiv 1;\ \xi_1(x) = \psi_2(x) + \psi_3(x);\ \xi_2(x) = \psi_3(x);\ \xi_3(x) \equiv 0.$$

A função $\xi_1(x)$ pode considerar-se como teste para o ensaio da hipótese $H : \theta \leqq 0{,}4$; pelo Teorema 7.7 existe, para a referida hipótese, um teste,

$$\zeta_1(x) = \begin{cases} 1 & \text{para} \quad x > x_1, \\ \gamma_1'' & \text{para} \quad x = x_1, \\ 0 & \text{para} \quad x < x_1, \end{cases}$$

verificando (8.10). Os valores x_1 e γ_1'' determinam-se a partir da condição

$$E_{\theta=0{,}4}\{\xi_1(X)\} = E_{\theta=0{,}4}\{\zeta_1(X)\}.$$

Como,

$$\begin{aligned}E_{\theta=0,4}\{\xi_1(X)\} &= E_{\theta=0,4}\{\psi_2(X)+\psi_3(X)\} = E_{\theta=0,4}\{\psi_2(X)\}+E_{\theta=0,4}\{\psi_3(X)\}\\ &= 0{,}3\sum_{x=0}^{2}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}+0{,}6\sum_{x=3}^{7}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}\\ &\quad+0{,}3\sum_{x=8}^{10}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}+0{,}1\sum_{x=0}^{2}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}\\ &\quad+0{,}2\sum_{x=3}^{7}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}+0{,}6\sum_{x=8}^{10}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}\\ &= 0{,}7343,\end{aligned}$$

conclui-se que x_1 e γ_1'' devem satisfazer,

$$\begin{aligned}E_{\theta=0,4}\{\zeta_1(X)\} &= \sum_{x=x_1+1}^{10}\binom{10}{x}(0{,}4)^x(0{,}6)^{10-x}+\gamma_1''\binom{10}{x_1}(0{,}4)^{x_1}(0{,}6)^{10-x_1}\\ &= 0{,}7343,\end{aligned}$$

o que dá $x_1 = 3$ e $\gamma_1'' = 0{,}542$.

Prosseguindo note-se que a função $\xi_2(x)$ pode ser tomada como teste da hipótese $H : \theta \leqq 0{,}6$; pelo Teorema 7.7 existe, para a mesma hipótese, um teste,

$$\zeta_2(x) = \begin{cases} 1 & \text{para} \quad x > x_2,\\ \gamma_2'' & \text{para} \quad x = x_2,\\ 0 & \text{para} \quad x < x_2,\end{cases}$$

verificando (8.10), com valores x_2 e γ_2'' determinados pela condição,

$$E_{\theta=0,6}\{\xi_2(X)\} = E_{\theta=0,6}\{\zeta_2(X)\}.$$

Facilmente se calcula,

$$E_{\theta=0,6}\{\xi_2(X)\} = E_{\theta=0,6}\{\psi_3(X)\} = 0{,}2657,$$

donde,

$$\begin{aligned}E_{\theta=0,6}\{\zeta_2(X)\} &= \sum_{x=x_2+1}^{10}\binom{10}{x}(0{,}6)^x(0{,}4)^{10-x}+\gamma_2''\binom{10}{x_2}(0{,}6)^{x_2}(0{,}4)^{10-x_2}\\ &= 0{,}2657;\end{aligned}$$

logo, $x_2 = 7$ e $\gamma_2'' = 0{,}458$.

Finalmente, com $\zeta_0(x) \equiv 1$ e $\zeta_3(x) \equiv 0$, o procedimento monótono, ϕ, pelo menos tão bom como ψ, tem por estrutura,

$$\phi_1(x) = \zeta_0(x) - \zeta_1(x) = \begin{cases} 1 & \text{para } x = 0, 1, 2, \\ 0{,}458 & \text{para } x = 3 \\ 0 & \text{para } x = 4, 5, 6, 7, 8, 9, 10, \end{cases}$$

$$\phi_2(x) = \zeta_1(x) - \zeta_2(x) = \begin{cases} 0 & \text{para } x = 0, 1, 2, \\ 0{,}542 & \text{para } x = 3, \\ 1 & \text{para } x = 4, 5, 6, \\ 0{,}542 & \text{para } x = 7 \\ 0 & \text{para } x = 8, 9, 10, \end{cases}$$

$$\phi_3(x) = \zeta_2(x) - \zeta_3(x) = \begin{cases} 0 & \text{para } x = 0, 1, 2, 3, 4, 5, 6, \\ 0{,}458 & \text{para } x = 7, \\ 1 & \text{para } x = 8, 9, 10. \end{cases}$$

Note-se que a estrutura do procedimento monótono não depende da função perca, o que sempre acontece quando esta satisfaz as condições (8.6). [Blackwell e Girshick (1954)]. □

No caso de duas acções demonstra-se em condições não muito restritas (vejam-se Teoremas 7.9 e 7.15) que os procedimentos monótonos são admissíveis. Quando o número de acções é superior a dois o problema da admissibilidade é mais complexo e não cabe no âmbito destas notas [veja-se Karlin (1957b)].

8.3 Soluções Bayes no caso geral

Em multidecisão, os problemas monótonos — função perca satisfazendo (8.6) — constituem caso particular. No entanto, no caso geral, não há grande dificuldade em caracterizar as soluções Bayes.

Considere-se, $A = \{a_1, a_2, \ldots, a_m\}$, e seja h, $h \in \mathcal{H}$, uma distribuição a priori; dada a função de decisão aleatória, $\phi = \{\phi_1(x), \phi_2(x), \ldots, \phi_m(x)\}$, a respectiva função risco,

$$\begin{aligned} R(h, \phi) &= \int_\Theta R(\theta, \phi) h(\theta)\, d\theta \\ &= \int_\Theta \left\{ \sum_{i=1}^{m} [L_i(\theta) \int_{\mathcal{X}} \phi_i(x) f(x \mid \theta)\, dx] \right\} h(\theta)\, d\theta, \end{aligned}$$

pode escrever-se, admitindo legítima a permuta dos integrais e tomando $f(x \mid \theta) h(\theta) = h(\theta \mid x) f(x)$,

$$R(h,\phi) = \int_{\mathfrak{X}} \left\{ \sum_{i=1}^{m} \left[\int_{\Theta} L_i(\theta) h(\theta \mid x)\, d\theta \right] \phi_i(x) \right\} f(x)\, dx$$
$$= \int_{\mathfrak{X}} \left\{ \sum_{i=1}^{m} r_x(i)\phi_i(x) \right\} f(x)\, dx,$$

onde $r_x(i)$ é o risco a posteriori. Para minimizar, $R(h,\phi)$, basta tomar,

$$\phi_i(x) = \begin{cases} 1 & \text{para } x \text{ tal que } r_x(i) = \min_j r_x(j), \\ 0 & \text{para outros } x. \end{cases} \tag{8.12}$$

Fica assim demonstrado,

Teorema 8.2 — Num problema de multidecisão com m acções, dada a distribuição a priori, h, o procedimento (8.12) é Bayes contra h. □□

Esta proposição foi antecipadamente ilustrada através dos Exs. 3.12 e 5.24. Para finalizar apresenta-se mais o seguinte:

Exemplo 8.2 — A variável aleatória, X, tem uma das distribuições, $f_j(x)$, $j = 1, 2, \ldots, m$, e pretende-se a partir da observação de X classificar a respectiva distribuição. A acção i consiste em atribuir a X a distribuição $f_i(x)$; tem-se, $A = \{1, 2, \ldots, m\}$ e $\Theta = \{1, 2, \ldots, m\}$. A função perca é da forma,

$$L_i(j) = \begin{cases} L & \text{se} \quad i \neq j, \\ 0 & \text{se} \quad i = j, \end{cases} \tag{8.13}$$

isto é, o custo de uma classificação errada é constante.

A função risco a posteriori assume a expressão,

$$r_x(i) = L \left[\sum_{\substack{j=1 \\ j \neq i}}^{m} f_j(x) h_j \right] \Big/ \left[\sum_{j=1}^{m} f_j(x) h_j \right],$$

quando a distribuição a priori é $\mathbf{h} = (h_1, h_2, \ldots, h_m)$, pois,

$$h(j \mid x) = f_j(x)h_j / f(x) = f_j(x)h_j \Big/ \left[\sum_{j=1}^{m} f_j(x) h_j \right],$$

e tem de atender-se a (8.13). Logo, pelo Teorema 8.2,

$$\phi_i(x) = \begin{cases} 1 & \text{se } r_x(i) \leqq r_x(j),\ j = 1, 2, \ldots, m, \\ 0 & \text{outros } x, \end{cases}$$

isto é,

$$\phi_i(x) = \begin{cases} 1 & \text{se } f_i(x)/f_j(x) \leqq h_j/h_i,\ j = 1, 2, \ldots, m, \\ 0 & \text{outros } x. \end{cases}$$

□

Uma boa referência para outros exemplos é Ferguson (1967).

APÊNDICE

A.1 Distribuição do qui-quadrado não central

Se $X_1, X_2, \ldots, X_k$ são variáveis aleatórias I.I.D., $X_i \sim N(0,1)$, $i = 1, 2, \ldots, k$, sabe-se que,

$$Z = \sum_{i=1}^{k} X_i^2, \tag{A1}$$

tem distribuição do qui-quadrado (central) com k graus de liberdade; simbolicamente, $Z \sim \chi_k^2$, sendo a função de densidade de Z,

$$f_k(z) = [2^{k/2}\Gamma(k/2)]^{-1} z^{(k/2)-1} \exp\{-z/2\},\ z > 0. \tag{A2}$$

Se $X_i \sim N(\theta_i, 1)$, $i = 1, 2, \ldots, k$, diz-se que a variável aleatória Z, definida por (A1) tem distribuição do qui-quadrado não central com parâmetro de não centralidade,

$$\lambda = \sum_{i=1}^{k} \theta_i^2; \tag{A3}$$

simbolicamente, $Z \sim \chi_{k,\lambda}^2$. A função de densidade é dada pela expressão,

$$f_k(z \mid \lambda) = \sum_{j=0}^{\infty} \frac{(\lambda/2)^j e^{-\lambda/2}}{j!} f_{k+2j}(z). \tag{A4}$$

A distribuição do qui-quadrado não central foi obtida por Fisher em 1928 ao estudar um caso limite da distribuição do coeficiente de correlação múltipla. Patnaik, em 1949, pôs em relevo a sua importância na determinação aproximada da potência dos ensaios do qui-quadrado. Hoje em dia as aplicações são numerosas, quer na estatística, quer na física matemática, quer ainda na teoria da comunicação.

Como se verifica por (A4), a distribuição do qui-quadrado não central pode entender-se como a média ponderada de qui-quadrados centrais, χ_{k+2j}^2, $j = 0, 1,$

2, ..., em que para cada j os coeficientes de ponderação são probabilidades de uma Poisson com média $\lambda/2$.

Repare-se que Z pode escrever-se,

$$\begin{aligned} Z &= \sum_{i=1}^{k} [(X_i - \theta_i) + \theta_i]^2 \\ &= \sum_{i=1}^{k} (U_i + \theta_i)^2, \end{aligned}$$

onde $U_i \sim N(0,1)$, $i = 1, 2, \ldots, k$; considerando a última expressão é interessante constatar que a distribuição de Z depende dos k parâmetros, $\theta_1, \theta_2, \ldots, \theta_k$, unicamente através da soma dos respectivos quadrados [$= \lambda$].

A.2 Distribuição-*t* não central

Se $U \sim N(0,1)$ e se $V \sim \chi^2_k$ são variáveis aleatórias independentes sabe-se que,

$$T = \frac{U}{\sqrt{V/k}}, \tag{A5}$$

tem distribuição-*t* com k graus de liberdade; simbolicamente, $T \sim t_k$. A função de densidade de T é bem conhecida,

$$f_k(t) = \left[\sqrt{k}\, B\left(\frac{1}{2}, \frac{k}{2}\right)\right]^{-1} \left[1 + \frac{t^2}{k}\right]^{-(k+1)/2}, \; t \in \mathbf{R}. \tag{A6}$$

Quando no lugar de (A5) se toma, com $\delta \in \mathbf{R}$,

$$T = \frac{U + \delta}{\sqrt{V/k}}, \tag{A7}$$

diz-se que T tem distribuição-*t* não central com parâmetro de não centralidade δ (ou δ^2 ou $\delta^2/2$); simbolicamente $T \sim t_{k,\delta}$. A função de densidade tem a forma,

$$\begin{aligned} f_k(t \mid \delta) &= \frac{e^{\delta^2/2}}{2^{(k-1)/2}\sqrt{k\pi}\,\Gamma(k/2)} \int_0^\infty y^k \exp\left\{-\frac{1}{2}[(1 + t^2 k^{-1})y^2 - 2(t\sqrt{k})y]\right\} dy \\ &= \frac{k!}{2^{(k-1)/2}\sqrt{k\pi}\,\Gamma(k/2)} \exp\{-[k\delta^2/(k + t^2)]\} \cdot \\ &\quad \cdot [k/(k + t^2)]^{(k+1)/2}\, Hh_k[-\delta t/(\sqrt{k + t^2})], \end{aligned} \tag{A8}$$

onde a função Hh tem por expressão,

$$Hh_k(x) = (1/k!)\int_0^{\infty} \xi^k \exp\left\{-\frac{1}{2}(\xi + x)^2\right\} d\xi.$$

A distribuição-*t* não central foi obtida por Fisher em 1931 e tem largas aplicações, nomeadamente na determinação da função potência do ensaio de hipóteses sobre a média da Normal quando a variância não é conhecida.

A.3 Distribuição-*F* não central

Se $U \sim \chi^2_m$ e se $V \sim \chi^2_n$ são variáveis aleatórias independentes sabe-se que,

$$F = \frac{U/m}{V/n}, \tag{A9}$$

tem distribuição-*F* com m e n graus de liberdade, com função de densidade,

$$f_{m,n}(z) = \frac{1}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \cdot \frac{\left(\frac{m}{n}z\right)^{(m/2)-1}}{\left(1 + \frac{m}{n}z\right)^{(m+n)/2}} \cdot \frac{m}{n}, \; z > 0;$$

simbolicamente, $F \sim F_{m,n}$.

Se U e V, variáveis aleatórias independentes, possuem distribuição do qui-quadrado não central, respectivamente, $U \sim \chi^2_{m,\lambda_1}$ e $V \sim \chi^2_{n,\lambda_2}$, o rácio (A9) diz-se que tem distribuição-*F* não central com m e n graus de liberdade e parâmetros de não centralidade, λ_1 e λ_2. Em muitas aplicações tem-se $\lambda_2 = 0$, quer dizer, V tem distribuição qui-quadrado central; fala-se então em distribuição-*F* não central simples [a designação simples é em geral omitida] com m e n graus de liberdade e parâmetro de não centralidade λ_1, que tem por função de densidade,

$$\begin{aligned} f_{m,n}(z \mid \lambda_1) = {} & \frac{e^{-\lambda_1/2} m^{m/2} n^{n/2}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \cdot \frac{z^{(m/2)-1}}{(n + mz)^{(m+n)/2}} \cdot \\ & \cdot \sum_{j=0}^{\infty} \left\{ \left(\frac{\lambda_1 mz/2}{n + mz} \right)^j \frac{(m+n)(m+n+2)\ldots[m+n+2(j-1)]}{j!\, m(m+2)\ldots[m+2(j-1)]} \right\}. \end{aligned} \tag{A10}$$

Fisher, em 1928, obteve a distribuição Beta não central que está relacionada com a distribuição-*F* não central; a obtenção desta deve-se a Tang, em 1938, embora a designação tenha sido usada pela primeira vez em 1949 por Patnaik.

O estudo desenvolvido das distribuições apresentadas — e de muitas outras — pode fazer-se com proveito na excelente monografia de Norman L. Johnson e Samuel Kotz publicada em quatro volumes:

(1969) *Distributions in Statistics: Discrete Distributions*, Houghton Mifflin Co., Boston;
(1970) *Distributions in Statistics: Continuous Univariate Distributions-1*, Idem.
(1970) *Distributions in Statistics: Continuous Univariate Distributions-2*, Idem.
(1972) *Distributions in Statistics: Continuous Multivariate Distributions*, J. Wiley & Sons, Inc., Nova Iorque.

O volume que trata das distribuições não centrais é o terceiro.

REFERÊNCIAS

Anscombe, F. J. e Aumann, R. J. (1963): «A definition of subjective probability», *Ann. Math. Stat.*, 34, 199–205.

Barnard, G. A. e Godambe, V. P. (1982): «Memorial Article, Allan Birnbaum 1923–1976», *Ann. Stat.*, 10, 1033–1039.

Barnett, Vic (1982): *Comparative Statistical Inference*, 2nd Ed., Wiley, Nova Iorque.

Berger, James O. (1980): *Statistical Decision Theory*, Springer-Verlag, Nova Iorque.

Berger, James O. (1984): «The Robust Bayesian Viewpoint», in *Robustness of Bayesian Analysis*, Ed. J. B. Kadane, North-Holland, Amsterdão.

Berger, James O. (1985): *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, Nova Iorque.

Berger, James O. e Sellke, Thomas (1987): «Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence», *J. Amer. Stat. Assoc.*, 82, 112–139.

Berger, James O. e Wolpert, Robert (1984): *The Likelihood Principle*, Institute of Mathematical Statistics, Hayward, Califórnia.

Birnbaum, Allan (1962): «On the foundations of statistical inference (with discussion)», *J. Amer. Stat. Assoc.*, 57, 269–326.

Birnbaum, Allan (1972): «More on Concepts of Statistical Evidence», *J. Amer. Stat. Assoc.*, 67, 858–861.

Blackwell, D. e Girshick, M. A. (1954): *Theory of Games and Statistical Decisions*, Wiley, Nova Iorque.

Box, George E. P. (1976): «Science and Statistics», *J. Amer. Stat. Assoc.*, 71, 791–799.

Box, George E. P. (1979): «Robustness in the strategy of scientific model building», in *Robustness in Statistics*, Ed. R. L. Launer e G. N. Wilkinson, Academic Press, Nova Iorque.

Box, George E. P. (1983): «An apology for ecumenism in statistics», in *Scientific Inference, Data Analysis and Robustness*, Ed. G. E. P. Box, Tom Leonard e Chien-Fu Wu, Academic Press, Nova Iorque.

Box, George E. P. (1984): «The importance of practice in the development of statistics», *Technometrics*, 26, 1–8.

Box, George E. P. e Tiao, G. C. (1973): *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.

Casella, G. e Berger, R. L. (1987): «Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem», *J. Amer. Stat. Assoc.*, 82, 106–111.

Castro, Gustavo de (1952): «A Estatística Matemática: uma Alfaia Científica», *Rev. Med. Veterinária*, XLVII, 52–64.

Castro, Gustavo de (1980): *Inferência e Decisão em Medicina*, Fasc. IV, Secção Editorial da AEFML, Lisboa.

Chernoff, H. e Moses, L. E. (1959): *Elementary Decision Theory*, Wiley, Nova Iorque.

Cox, D. R. (1958): «Some problems connected with statistical inference», *Ann. Math. Stat.*, 29, 357–372.

Cox, D. R. e Hinkley, D. V. (1974): *Theoretical Statistics*, Chapman & Hall, Londres.

DeGroot, M. H. (1970): *Optimal Statistical Decisions*, McGraw-Hill, Nova Iorque.

Dempster, A. P. (1971): «Model searching and estimation in the logic of inference», in *Foundations of Statistical Inference*, Ed. V. P. Godambe e D. A. Sprott, Holt, Rinehart and Winston of Canada, Ltd, Toronto.

Dickey, J. M. (1982): «Conjugate Families of Distributions», in *Encyclopedia of Statistical Sciences*, Vol. 2, Ed. S. Kotz e N. L. Johnson, Wiley, Nova Iorque.

Dunnett, C. W. (1972): «Drug Screening: The Never-Ending Search for New and Better Drugs», in *Statistics: A Guide To The Unknown*, Ed. Judith Tanur e outros, Holden-Day, S. Francisco.

Edwards, W., Lindman, H. e Savage, L. J. (1963): «Bayesian Statistical Inference for Psychological Research», *Psychological Review*, 70, 193–242 [Reproduzido em *Robustness of Bayesian Analysis*, Ed. J. B. Kadane, North-Holland, Amsterdão (1984)].

Efron, B. (1975): «Biased versus unbiased estimation», *Advances in Mathematics*, 16, 259–277.

Efron, B. (1978): «Controversies in the foundations of statistics», *Am. Math. Monthly*, 85, 231–246.

Efron, B. e Morris, C. (1973): «Combining possibly related estimation problems (with discussion)», *J. R. S. S. B*, 35, 379–421.

Ferguson, T. S. (1967): *Mathematical Statistics. A Decision Theoretic Approach*, Academic Press, Nova Iorque.

Fine, Terrence L. (1973): *Theories of Probability*, Academic Press, Nova Iorque.

Fisher, R. A. (1956): *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edimburgo.

Fraser, D. A. S. (1957): *Nonparametric Methods in Statistics*, Wiley, Nova Iorque.

Gass, S. I. (1969): *Linear Programming*, 3rd Ed., McGraw-Hill, Nova Iorque.

Giri, N. C. (1983): «Hunt-Stein Theorem», in *Encyclopedia of Statistical Sciences*, Vol. 3, Ed. S. Kotz e N. L. Johnson, Wiley, Nova Iorque.

Girshick, M. A. e Savage, L. J. (1951): «Bayes and minimax estimates for quadratic loss functions», *2nd Berkeley Symposium on Math. Stat. and Prob.*, Univ. Calif. Press., Berkeley.

Gomes, Maria Ivette (1981): «Verosimilhança e Inferência Estatística», in *Actas do II Colóquio de Estatística e Investigação Operacional*, Soc. Port. de Estatística e Inv. Operacional, Fundão/Covilhã.

Halmos, P. R. (1950): *Measure Theory*, D. Van Nostrand, Nova Iorque.

Hays, W. L. e Winkler, R. L. (1970): *Statistics: Probability, Inference, and Decision*, Vol. 1, Holt, Rinehart and Winston, Nova Iorque.

Hodges, W. L. e Lehmann, E. L. (1950): «Some problems in minimax point estimation», *Ann. Math. Stat.*, 21, 182–197.

Hodges, W. L. e Lehmann, E. L. (1951): «Some applications of the Cramér-Rao inequality», *2nd Berkeley Symposium on Math. Stat. and Prob.*, Univ. Calif. Press, Berkeley.

Huber, P. J. (1981): *Robust Statistics*, Wiley, Nova Iorque.

James, W. e Stein, C. (1960): «Estimation with quadratic loss», *4th Berkeley Symposium on Math. Stat and Prob.*, Vol. 1, Univ. Calif. Press, Berkeley.

Jeffreys, H. (1961): *Theory of Probability*, 3rd Ed., Oxford Univ. Press, Londres.

Joshi, V. M. (1984): «Likelihood Principle», in *Encyclopedia of Statistical Sciences*, Vol. 4, Ed. S. Kotz e N. L. Johnson, Wiley, Nova Iorque.

Kalbfleish, J. G. (1985): *Probability and Statistical Inference. Vol. 2: Statistical Inference*, 2nd Ed., Springer-Verlag, Nova Iorque.

Karlin, S. (1956): «Decision theory for Pólya type distributions. Case of two actions, I», *3rd Berkeley Symposium on Math. Stat. and Prob.*, Vol. 1, Univ. Calif. Press, Berkeley.

Karlin, S. (1957a): «Pólya type distributions, II», *Ann. Math. Stat.*, 28, 281–308.

Karlin, S. (1957b): «Pólya type distributions, III: Admissibility for multi-action problems», *Ann. Math. Stat.*, 28, 839–860.

Karlin, S. e Rubin, H. (1956): «The theory of decision procedures for distributions with monotone likelihood ratio», *Ann. Math. Stat.*, 27, 272–299.

Kempthorne, O. e Folks, L. (1971): *Probability, Statistics, and Data Analysis*, The Iowa State Univ. Press, Ames, Iowa.

Kendall, M. G. e Stuart, A. (1967): *The Advanced Theory of Statistics*, Vol. 2, C. Griffin, Londres.

Kiefer, J. (1982): «Conditional Inference», in *Encyclopedia of Statistical Sciences*, Vol. 1, Ed. S. Kotz e N. L. Johnson, Wiley, Nova Iorque.

Kyburg Jr., H. E. e Smokler, H. E. (Ed.) (1964): *Studies in Subjective Probability*, Wiley, Nova Iorque.

Lehmann, E. L. (1959): *Testing Statistical Hypothesis*, Wiley, Nova Iorque.

Lehmann, E. L. (1983): *Theory of Point Estimation*, Wiley, Nova Iorque.

Lin, P. E. (1974): «Admissible minimax estimators of the multivariate normal mean with squared error loss», *Comun. in Statistics*, 3, 95–100.

Lindley, D. V. (1962): «Discussion on Professor Stein's Paper», *Journal of the Royal Statistical Society*, Ser. B, 24, 285–287.

Lindley, D. V. (1965): *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Vols. 1 e 2, Cambridge Univ. Press, Cambridge.

Lindley, D. V. (1971a): «Bayesian Statistics, a Review», *SIAM*, Filadélfia.

Lindley, D. V. (1971b): *Making Decisions*, Wiley-Interscience, Londres.

Lindley, D. V., and L. D. Phillips (1976): «Inference for a Bernoulli Process (a Bayesian View).» *American Statistician*, 30, 112–19.

Lindley, D. V. (1982): «Bayesian Inference», in *Enclyclopedia of Statistical Sciences*, Vol. 1, Ed. S. Kotz e N. L. Johnson, Wiley, Nova Iorque.

Lindgren, B. W. (1971): *Elements of Decision Theory*, Mcmillan, Nova Iorque.

Lindgren, B. W. (1976): *Statistical Theory*, 3rd Ed., Mcmillan, Nova Iorque.

Luce, R. D. e Raiffa, H. (1957): *Games and Decisions*, Wiley, Nova Iorque.

Mallows, C. L. e Tukey, J. W. (1982): «An overview of techniques of data analysis, emphasizing its exploratory aspects», in *Some Recent Advances in Statistics*, Ed. J. Tiago de Oliveira e B. Epstein, Academia das Ciências de Lisboa, Lisboa.

Mood, A. M., Graybill, F. A. e Boes, D. (1974): *Introduction to the Theory of Statistics*, 3rd Ed., McGraw-Hill, Nova Iorque.

Muirhead, R. J. (1982): *Aspects of Multivariate Statistical Theory*, Wiley, Nova Iorque.

Murteira, B. J. F. (1980): *Probabilidades e Estatística*, Vol. II, McGraw-Hill, Lisboa.

Neumann, J. e Morgenstern, O. (1944): *Theory of Games and Economic Behavior*, 3rd Ed. 1953, Princeton Univ. Press, Princeton.

Neyman, J. (1950): *First Course in Probability and Statistics*, H. Holt, Nova Iorque.

Prat, J. W. (1962): Discussão de Birnbaum (1962), *J. Amer. Stat. Assoc.*, 57, 269–326.

Prat, J. W., Raiffa, H. e Schlaifer, R. (1964): «The foundations of decision under uncertainty: an elementary exposition», *J. Amer. Stat. Assoc.*, 59, 353–375.

Raiffa, H. (1968): *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison-Wesley, Reading.

Raiffa, H. e Schlaifer, R. (1961): *Applied Statistical Decision Theory*, Harvard University, Boston.

Ramsey, F. P. (1926): «Truth and Probability», reproduzido em Kyburg e Smokler (1964).

Roberts, A. W. e Varberg, D. E. (1973): *Convex Functions*, Academic Press, Nova Iorque.

Rohatgi, V. K. (1976): *An Introduction to Probability Theory and Mathematical Statistics*, Wiley, Nova Iorque.

Savage, J. L. (1954): *The Foundations of Statistics*, Wiley, Nova Iorque.

Savage, J. L. (1961): «The foundations of statistics reconsidered», *4th Berkeley Symposium on Math. Stat. and Prob.*, Vol. 1, Univ. Calif. Press, Berkeley.

Savage, J. L. e outros (1962): *The Foundations of Statistical Inference*, Methuen, Londres.

Silvey, S. D. (1970): *Statistical Inference*, Penguin Education Library of University Mathematics, Londres.

Stein, C. (1956): «Inadmissibility of the usual estimator for the mean of a multivariate normal distribution», *3rd Berkeley Symposium on Math. Stat. and Prob.*, Vol. 1, Univ. Calif. Press, Berkeley.

Stein, C. (1959): «The admissibility of Pitman's estimator of a single location parameter», *Ann. Math. Stat.*, *30*, 970–979.

Strawderman, W. (1971): «Proper Bayes minimax estimation of multivariate normal mean», *Ann. Math. Stat.*, 42, 385–388.

Tiago de Oliveira, J. (1981): «Resumo das intervenções e comentários», Sessão de Inferência Estatística, in *Actas do II Colóquio de Estatística e Investigação Operacional*, Soc. Port. de Estatística e Inv. Operacional, Fundão/Covilhã.

Turkman, Maria A. A. (1981): «Inferência Bayesiana», in *Actas do II Colóquio de Estatística e Investigação Operacional*, Soc. Port. de Estatística e Inv. Operacional, Fundão/Covilhã.

Watcher, K. W. (1983): «Haar Distributions», in *Encyclopedia of Statistical Sciences*, Vol. 3, Ed. S. Kotz e N. L. Johnson, Wiley, Nova Iorque.

Wald, A. (1950): *Statistical Decision Functions*, Wiley, Nova Iorque.

Wald, A. e Wolfowitz, J. (1951): «Two methods of randomization in statistics and the theory of games», *Ann. Math. Stat.*, 53, 581–586.

Zacks, S. (1971): *The Theory of Statistical Inference*, Wiley, Nova Iorque.

Zacks, S. (1981): *Parametric Statistical Inference*, Pergamon Press, Oxford.

Zellner, A. (1971): *An Introduction to Bayesian Inference in Econometrics*, Wiley, Nova Iorque.

Zellner, A. e Vandaele, W. (1975): «Bayes-Stein estimators for k-means, regression and simultaneous equation models», in *Studies in Bayesian Econometrics and Statistics*, North-Holland, Amsterdão.

«*Obra mais arrojada de Bento Murteira, pela profunda reflexão crítica sobre as metodologias estatísticas que constitui, foi no nosso meio um marco impulsionador de uma consciência crítica da teoria e prática estatística que, ainda hoje, deve ser leitura assídua e reflexiva de quem faz da Estatística a sua ocupação profissional.*»

Carlos Daniel Paulino
«A Evolução da Estatística Bayesiana em Portugal»,
Memorial da Sociedade Portuguesa de Estatística, 2005, Eds. F. Rosado, pp. 215–218.

«*Professor Bento Murteira was a towering figure in Portugal (and highly recognized in the wider world) in the sciences of statistics, forecasting, econometrics and decision making. This book, which he wrote in 1988, is a wonderful testament to his brilliance.*

I wish I had been able to interact with Professor Murteira; the depth of his understanding was phenomenal.»

James Berger
Arts and Sciences Distinguished Professor Emeritus of Statistics, Duke University

APOIOS

SPE Sociedade Portuguesa de Estatística

fct Fundação para a Ciência e a Tecnologia

Instituto Nacional de Estatística
Statistics Portugal

CEMAPRE
Centro de Matemática Aplicada à Previsão e Decisão Económica

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa