A MAXIMIZAÇÃO DA ENTROPIA E SUA UTILIZAÇÃO NA PROCURA DE DISTRIBUIÇÕES *A PRIORI* EM INFERÊNCIA BAYESIANA

Dissertação apresentada como requisito parcial para a obtenção do grau de mestre em Métodos Matemáticos Aplicados à Economia e Gestão de Empresas

O candidato agradece muito reconhecido a boa vontade, interesse e saber que o Ex.^{mo} Prof. Bento José Ferreira Murteira continuamente lhe facultou na orientação deste trabalho.

1. ENTROPIA: UM SÓ OU VÁRIOS CONCEITOS DISTINTOS?

1.1. Da Termodinâmica à Mecânica Estatística: Clausius e Maxwell. O conceito de entropia foi introduzido nas décadas de 50 e 60 do século XIX pelo físico alemão R. Clausius (1822-1888) no decurso de estudos sobre irreversibilidade em Termodinâmica[1].

A definição de entropia em Termodinâmica é dada na forma diferencial por,

$$dS = \frac{dQ}{T}, \tag{1.1}$$

onde dQ representa a quantidade de calor transferido (reversivelmente) a uma determinada temperatura T medida na escala absoluta (graus Kelvin). Trata-se, portanto, de uma grandeza física mensurável em laboratório.

A entropia ganhou relevância através da polémica gerada à volta das implicações teleológicas da segunda lei da Termodinâmica: o chamado "heat death" do Universo, que tanto cativou a imaginação científica do fim do século XIX

1 O termo entropia (do grego ἐντροπία, equivalente a transformação) foi utilizado pela primeira vez, também por Clausius, em 1865. Nessa altura Clausius deu o seguinte célebre enunciado das duas leis da Termodinâmica: "A energia do mundo é constante [1ª lei], e a sua entropia tende para um máximo [2ª lei] ". Curiosamente, Clausius comentaria então que a segunda lei seria de entendimento bem mais difícil do que a primeira.

e princípio do século XX[2].

Se bem que a lei do aumento irreversível da entropia possa ser aceite como um postulado[3], o ideal então vigente de conseguir uma explicação integral do mundo físico em termos mecânicos ditou a procura de uma analogia mecânica para o conceito de entropia[4]. A Teoria Cinética dos Gases, então em desenvolvimento, iria proporcionar um bom ponto de partida para tal estudo. Desde logo se reconheceu que a abordagem dos problemas a partir da hipótese molecular imporia a adopção de uma nova metodologia matemática. De facto, os números proibitivos em jogo nos agregados de átomos e

2 Já neste século P. Bridgman referia um aspecto em que as leis da Termodinâmica se distinguiam das restantes leis da Física: "... there is something more palpably verbal about them - they smell more of their human origins" (Bridgman, 1941, p. 3).

3 As leis da Termodinâmica são puramente fenomenológicas e livres de qualquer modelo sobre a constituição microscópica da matéria. A sua justificação última reside na sua constante verificação em processos naturais.

4 "This reduction seemed an urgent necessity for those physicists in the nineteenth century who believed that all properties of matter and energy were ultimately explicable by mechanical models, using Newton's laws of motion". (S. Brush, s.u. "Irreversibility", *The Encyclopaedia of Physics*, p. 617)

moléculas[5] tornavam à partida inutilizáveis as técnicas dinâmicas de Newton e Laplace, que tanto sucesso tinham tido em Mecânica Celeste. A nova metodologia adoptada, e que resultaria de uma combinação dos princípios mecânicos com a Teoria da Probabilidade e Estatística, acabaria por tomar o nome de Mecânica Estatística. Marcos fundamentais no desenvolvimento de tal ciência foram os trabalhos de Clausius em 1857 e 1858 (derivação da lei do gás perfeito a partir dos valores médios dos agregados de átomos em estudo e introdução do conceito de caminho livre médio) e de J.C. Maxwell (1831-1879) em 1859 (obtenção da lei de probabilidade que rege a distribuição das velocidades das partículas de um gás em equilíbrio termodinâmico). No artigo de 1859, que marca de certa forma o verdadeiro início da utilização das técnicas probabilísticas em Física, Maxwell obtem a célebre lei de densidade para as velocidades,

$$f(v) = \frac{4a^{3/2}}{\pi^{1/2}} v^2 \exp(-a v^2), \tag{1.2}$$

que leva o seu nome.

A lei de Maxwell, para a validade da qual existem inúmeras

5 Agregados de átomos e moléculas numa ordem de grandeza de 10^{23} em Física microscópica *versus* alguns poucos corpos celestes no caso do estudo do Sistema Solar.

provas indirectas, revelar-se-á como sendo a distribuição de velocidades de entropia máxima correspondente ao estado de equilíbrio térmico de um gás.

É também a Maxwell, em 1867, que se deve a introdução de um célebre *Gedankenexperiment* tendo como protagonista um pequeno ser, que passou à História da Ciência com o nome de "Demónio de Maxwell" e que tinha o fim expresso de tentar "furar" a segunda lei da Termodinâmica[6]. É significativo que tal "Demónio", essencialmente um *brain-teaser* sobre o conceito de entropia, só tenha sido exorcizado por L. Szilard, em 1929, com recurso a uma linha de raciocínio que prefigurava certos aspectos da Teoria da Informação[7].

---

6 O demónio de Maxwell tinha o fim explícito, confessado pelo seu progenitor, de: "To show that the 2nd Law of Thermodynamics has only a statistical certainty" (carta de Maxwell a P. G. Tait, não datada, citada em Daub, 1970, *in Darwin to Einstein, Historical Studies on Science and Belief*, p. 228). Tratava-se essencialmente de um pequeno ser que, devido às suas reduzidas dimensões, teria a faculdade de separar as moléculas em movimento rápido num gás das moléculas lentas, assim possibilitando a acumulação das primeiras num espaço de dimensão mais reduzida, em contradição flagrante com a lei do aumento da entropia.

7 Szilard refere mais tarde: [o seu artigo de 1929] "Was a radical departure in thinking because I said that the essential thing here is that the demon utilizes information - to be precise

1.2. A Mecânica Estatística: Boltzmann e Gibbs. A partir dos meados da década de 1860, o físico austríaco L. Boltzmann (1844-1906) revelou-se um dos espíritos mais tenazes na procura de uma explicação mecânica para a entropia. Começando por raciocinar em termos determinísticos puros[8] (1866), Boltzmann seria progressivamente levado (1871-1872) a abandonar tal ideal[9].

---

7 (cont. pág. ant.) information which is not really in his possession because he guesses it. I said there is a relationship between information and entropy, and I computed what this relationship was". (*Leo Szilard: His Version of the Facts. Selected Recollections and Correspondence*, S. Weart e G. Szilard, eds., 1978, Cambridge, Mass., MIT Press).

O ponto essencial para que Szilard chamou a atenção era o de que o Demónio teria de utilizar meios físicos no processo de distinção de quais as moléculas rápidas a separar. Esses meios físicos criariam uma quantidade de entropia pelo menos igual àquela que seria eliminada no processo de separação das moléculas. Assim estava salva a segunda lei da Termodinâmica!

8 Nos próprios termos de Boltzmann surgidos num artigo de 1866: "to give a completely general proof of the second law of the theory of heat, as well as to discover the theorem in mechanics that corresponds to it" (citado em A. Pais, 1982, p. 61)

9 "The problems of the mechanical theory of heat... are problems in the theory of probability" (Boltzmann, 1872, citado em A. Pais, 1982, p. 61).

Um primeiro resultado fundamental foi obtido por Boltzmann (1872) ao estabelecer a analogia entre o carácter de crescimento monótono da entropia com o evoluir do tempo (de acordo com a segunda lei) e uma característica da descrição mecânica de um agregado de moléculas por si descoberta e possuindo tal propriedade de monotonia. Boltzmann mostrou que a quantidade,

(1.3) $$H = \int \ldots \int f \ln f \, dq_1 \ldots dq_s \, dp_1 \ldots dp_s,$$

[onde, $f(q_1,\ldots,q_s,p_1,\ldots,p_s)$, nos dá o número de sistemas cujas coordenadas generalizadas ocupam no espaço de fase os intervalos, (q_i, q_i+dq_i), (p_i, p_i+dp_i), $i=1,2,\ldots,s]$, nunca aumenta em resultado das colisões atómicas. No caso da lei das velocidades ser a de Maxwell, o valor de H mantém-se constante. Tal quantidade H, que satisfaz,

(1.4) $$\frac{dH}{dt} \leq 0,$$

pode então identificar-se com o negativo da entropia de um gás. Este notável resultado, conhecido como *Teorema H*, não foi provado rigorosamente por Boltzmann. Muitas das hipóteses estatísticas sobre as colisões moleculares careciam de justificação, como o própio Boltzmann viria a reconhecer. O carácter estatístico da segunda lei, já entrevisto por Maxwell e revelado na falta de certeza quanto ao aumento da entropia devido à possibilidade de flutuações, só seria reconhecido

por Boltzmann alguns anos mais tarde[10].

Em 1877, Boltzmann conseguiu um novo resultado notável de tipo quantitativo, ao supor o espaço de fase de uma molécula subdividido num número de células de magnitude,

(1.5) $$\tau = dq_1 \ldots dq_s \, dp_1 \ldots dp_s .$$

Sendo n_i o número de moléculas que num dado instante ocupam a célula i, pode substituir-se o integral (1.3) pela soma,

(1.6) $$H = \sum_i (n_i \ln n_i) \, \tau$$

Tendo em conta que só a variação de H pode ser determinada experimentalmente e fazendo, $\sum_i n_i = N$, (1.6) pode escrever-se na forma,

(1.7) $$H = \tau \Big[\sum_i (n_i \ln n_i - n_i) - (N \ln N - N) \Big] .$$

Usando em (1.7) a aproximação $\ln n! = n \ln n - n$, válida para n grande, obtem-se,

(1.8) $$H = -\tau \ln \frac{N!}{n_1! n_2! \ldots n_r!} .$$

---

10 Para tal muito contribuiu o célebre paradoxo proposto por Loschmit em 1876: como se explica que, sendo reversíveis as leis da Mecânica que regem as colisões moleculares (i. e., invariantes para uma simetria no parâmetro tempo), os fenómenos macroscópicos não sejam?

Em (1.8) o coeficiente multinomial $N!/n_1!n_2!...n_r!$ dá o número de possibilidades de N moléculas se distribuirem entre as células do espaço de fase, de forma a que a célula i contenha n_i moléculas. A este coeficiente chamou Boltzmann o número de configurações, W, correspondente aos números de ocupação, $n_1, n_2, \ldots, n_r$.

Assim, sendo H proporcional ao negativo da entropia S, tem-se,

(1.9) $$S = k \ln W.$$

A quantidade W em (1.9)[11] tomou o nome de probabilidade termodinâmica, se bem que represente um número inteiro de grande dimensão.

Boltzmann identificou então o estado de equilíbrio termodinâmico com o estado de máxima probabilidade termodinâmica.

No problema matemático de maximizar W, com as condições restritivas que fisicamente se devem exigir,

(1.10) $$\sum_i n_i = N,$$

11 A equação S = k ln W é devida a Planck. Aí figura a constante de Boltzmann, k, primeiramente introduzida por Planck em 1900 na sua célebre lei da radiação que marcou o início da revolução quântica.

e

$$\sum_i n_i E_i = U, \tag{1.11}$$

tem-se pela primeira vez bem explícito um problema de maximização de entropia sujeita a restrições.

A restrição (1.10) exige a conservação do número de moléculas e a restrição (1.11), onde E_i representa a energia característica da i-ésima célula do espaço de fase, exige a conservação da energia.

A solução, obtida pela técnica dos multiplicadores de Lagrange, é,

$$n_i^* = \frac{N}{Z(\beta)} \exp \left\{ -\beta E_i \right\}, \tag{1.12}$$

onde,

$$Z(\beta) = \sum_i \exp \left\{ -\beta E_i \right\}, \tag{1.13}$$

e β é tal que a restrição respeitante à conservação de energia, (1.11), seja satisfeita.

A distribuição de máxima entropia obtida em (1.12) é assim aquela que, obedecendo às restrições (1.10) e (1.11), se pode verificar através de um número máximo de configurações. Um sistema físico tenderá a evoluir de um estado inicial de menor probabilidade para um estado final de probabilidade máxima correspondente ao estado de equilíbrio

térmico[12]. Assim, a noção de entropia tinha sido reinterpretada em termos probabilísticos. Por outro lado, sendo possível obter a distribuição de Maxwell a partir de (1.12), confirmava-se a propriedade que esta tinha de maximizar a entropia. É notável como os cálculos de Boltzmann conduziram pela via combinatória à mesma distribuição que Maxwell obtivera por uma linha de raciocínio completamente diferente[13].

Se bem que os argumentos de Maxwell sugerissem já o carácter estatístico do aumento da entropia, este físico não chegou a apresentar uma relação quantitativa entre a segunda lei e a Estatística. O grande mérito de Boltzmann reside em ter descrito quantitativamente o grau de desordem de um sistema em termos da probabilidade de este se verificar e de relacionar esta probabilidade com a entropia do sistema.

---

12 "The initial state will in most cases be an highly improbable one; from it, the system will hasten to ever more probable states until it finally reaches the most probable of all, i. e. thermal equilibrium" (Boltzmann, citado por Porter, 1986, p. 212)

13 A elegante derivação feita por Maxwell em 1859, e que ocupa alguns poucos parágrafos, usava essencialmente a isotropia da distribuição, que se deve verificar devido ao elevado número de colisões verificado. Uma equação funcional dava então o resultado desejado (1.2).

Entre 1873 e 1878 o engenheiro norte-americano J. W. Gibbs (1839-1903) criou, e virtualmente completou no que respeita ao estudo do equilíbrio, a ciência da Termodinâmica Química. Gibbs associou ao estado termodinâmico de um sistema a probabilidade de este ocupar no espaço de fase qualquer um dos pontos consistentes com as restrições macroscópicas impostas. Introduziu no contínuo do espaço de fase uma função de densidade de probabilidade na forma,

$$w(q_1, \ldots, q_s; p_1, \ldots, p_s) = \exp\left\{\eta(q_1, \ldots, q_s; p_1, \ldots, p_s)\right\}, \tag{1.14}$$

chamando a função η "índice de probabilidade de fase". Obteve então a distribuição canónica e grande canónica, correspondentes ao estado de equilíbrio termodinâmico, a partir de restrições sobre, respectivamente, a energia média e a energia média e o número de partículas. Utilizou como critério a minimização do valor médio (ponderado através de w) do índice de probabilidade de fase η[14].

Tal como no caso do problema de Boltzmann, na minimização de, $\int \ldots \int w \ln w \, dq_1 \ldots dp_s$, sujeito a restrições

14 "The distribution in phase space which without violating this condition gives the least value of the average index of probability of phase $\bar{\eta}$..." (Gibbs, *Elementary Principles of Statistical Mechanics*, in Jaynes, 1968, p. 92)

de energia e do número de partículas, tem-se já um problema correspondente ao da maximização da entropia sujeita a restrições.

Gibbs não deu, no entanto, qualquer razão para a escolha de tal função a minimizar e, em consequência, se bem que o processo desse resultados em concordância com valores experimentais observados[15], ao nível da própria justificação lógica o método permaneceu durante algumas décadas com fundamentação pouco clara, aguardando o aparecimento da Teoria da Informação[16].

1.3. A Teoria da Informação: Shannon e Jaynes. A contribuição fundamental de Shannon proporcionou a síntese intelectual que marca o aparecimento da Teoria da Informação[17]. Esta teoria que é matemática e de carácter

15 Resistindo mesmo, com adaptações mínimas, ao teste da teoria quântica.

16 "It was not until the work of Shannon in our own time that the full significance and generality of Gibbs method could be appreciated. Once we had Shannon's theorem establishing the uniqueness of entropy as, an "information measure", it was clear that Gibbs procedure was an example of a general method for inductive inference, whose applicability is in no way restricted to equilibrium thermodynamics or to physics" (Jaynes, 1968, p.92).

17 O matemático N. Wiener (1894-1964) foi outra figura relevante no aparecimento da Teoria da Informação. Trabalhando na mesma instituição científica que Shannon (o MIT), Wiener parece ter

formal, pode ser considerada como um ramo da Teoria da Probabilidade e da Estatística. Manteve, no entanto, indeléveis os traços provenientes do campo original de trabalho em que foi criada: a Teoria das Comunicações Eléctricas.

No seu célebre artigo (1948), Shannon descreve as características esquemáticas a que deve obedecer qualquer sistema de comunicação de informação. Reconhece também a natureza estatística do processo de comunicação. Seguidamente propõe a medida logarítmica de informação[18],

$$H = - \sum_{i=1}^{n} p_i \ln p_i , \tag{1.15}$$

onde n é o número de símbolos distintos do "alfabeto" usado na transmissão de uma mensagem e p_i é a probabilidade de um dado símbolo transmitido ser do i-ésimo tipo. H dá, portanto, uma medida de informação por símbolo transmitido. A medida

---

17 (cont. pág. ant.) também intuído a relação entre entropia e informação. A sua linha de investigação orientou-se, porém, no sentido de uma nova ciência por ele chamada Cibernética. De alguma forma a ideia andava no ar. Mas nem por isso a contribuição de Shannon deixou de causar a maior admiração devido ao seu carácter quantitativo rigoroso e completo.

18 Em 1928 R. Hartley propusera já como medida de informação o logaritmo do número de símbolos equiprováveis utilizado na transmissão. Tratava-se já da medida de Shannon, mas restrita ao caso particular de equiprobabilidade.

logarítmica é demonstravelmente a única que obedece a um conjunto de condições tidas como de exigência natural. Tal derivação axiomática dá à expressão (1.15) uma relevância especial como medida de informação.

Quando Shannon decidiu dar à medida obtida o nome de entropia[19], imediatamente se levantou o problema de saber se se estaria perante o mesmo conceito da entropia termodinâmica ou antes se haveria unicamente uma coincidência no facto de tanto a Teoria da Informação como a Mecânica Estatística necessitarem de funções com a mesma estrutura. O problema

---

19 Alguns anos mais tarde (1961), numa entrevista concedida a Myron Tribus, Shannon confessou ter ficado perplexo com o nome a dar à medida logarítmica de informação. Aconselhando-se com o matemático J. von Neumann, recebeu deste a sugestão do termo entropia com a seguinte justificação: "first, the function is already in use in thermodynamics under that name; second and more importantly, most people don't know what entropy really is and if you use the word 'entropy' in an argument you will win every time!" (Tribus, 1979, pp. 2-3). No entanto, tal sugestão bem humorada está longe de suscitar consenso: "Thus confusion entered in and von Neumann had done science a disservice!... It would, therefore, have been better if the H - measure had been given an entirely neutral name. Tisza (1966) suggest that it should have been called "the dispersal" of the particular distribution. This would have been admirable and would have resulted in the avoidance of much confusion". (Denbigh e Denbigh, 1985, pp. 104-105)

continua a gerar controvérsia[20] e as posições tomadas reflectem, regra geral, as várias linhas de interpretação probabilística em Mecânica Estatística e Termodinâmica. A linha de interpretação subjectiva nestas ciências está já naturalmente inclinada a aceitar a identificação de conceitos, uma vez que a entropia termodinâmica era já interpretada como ausência de conhecimento sobre o estado físico de um sistema[21]. Por outro lado, os defensores de uma interpretação

---

20 Confira-se a opinião de N. Georgescu-Roegen, na linha da de Denbigh e Denbigh, referida na nota anterior: "A real imbroglio involving the entropy concept grew... But the mere coincidence of formulae was not a basis for justifying the terminological transfer. We would not call "kinetic energy" the second moment of a distribution just because both formulae are a sums of squares: $\Sigma\ \eta_i X_i^2$. With that transfer, the concept started travelling from one domain to another with hardly any justification (s.u. "Entropy", *The New Palgrave Dictionary of Economics*, vol. II, p. 156). No entanto, o próprio Shannon aconselhou sempre a maior circunspecção na aplicação dos seus resultados fora do campo da Teoria das Comunicações. Na tentativa de refrear excessos foi ao ponto de publicar em 1956 um artigo caracteristicamente intitulado "The Bandwagon".

21 "Now, the crux of the problem which the subjectivist probability school is facing is the justification, that is the understanding of how it happens that Shannon's "missing information" is also, in physical problems, the entropy - a perfectly objective and measurable quantity in systems. This question which is essentially the same one

probabilística objectiva recusam-se a identificar a entropia termodinâmica, uma grandeza física calorimetricamente mensurável em laboratório e independente do estado de conhecimento do observador, com a noção de entropia da Teoria da Informação, onde não existe qualquer possibilidade de localização física ou de estabelecimento de relações biunívocas com quantidades de calor.

A utilização da medida de Shannon em conjugação com a técnica, erigida por E. T. Jaynes em princípio, de maximizar H sujeito às restrições particulares operativas em cada situação[22], possibilitaria a Jaynes (1957) e a outros autores (e. g. Tribus, Katz, Hobson) reobter o formalismo da Mecânica Estatística por uma via original[23]. Tal via evitava o terreno da

---

21 (cont. pág. ant.) as explaning how the subjective probability comes out as objective frequency, is by no means a trivial one, as unsophisticated common sense would have it" (O. Costa de Beauregard, 1971, pp. 11-12)

22 O próprio Shannon tinha já sugerido que na construção de um código se considerasse a fonte de informação com entropia máxima, sujeita às condições estatísticas que se desejem reter. Seria essa entropia máxima que determinaria a capacidade necessária e suficiente do canal (cf. Jaynes, 1979, p. 40). No entanto, Shannon não chegou a formalizar e desenvolver quantitativamente a questão.

23 A aceitação de tal metodologia está longe de ser pacífica: "But it was E. T. Jaynes who after the spread of Shannon's theorem set out to

Teoria Ergódica[24], eriçado das maiores dificuldades matemáticas e de passagem obrigatória para os seguidores de uma interpretação frequencista ou objectiva das probabilidades em Mecânica Estatística. O preço a pagar pelo uso de tal "atalho" será o de ter de abdicar de uma interpretação frequencista da probabilidade e de voltar ao ponto de vista original de Bernoulli e Laplace (cf. Cap. 2.2), mas com o princípio da razão insuficiente (ibid.) generalizado e substituído pelo princípio da maximização da entropia. Tal preço, envolvendo questões de princípio na própria interpretação do conceito fundamental da probabilidade, é por alguns reputado de muito alto[25].

---

23 (cont. pág. ant.) erect thermodynamics on that basis alone. In spite of bizarreness, or perphaps because of it, the idea is still running in some circles." (N. Georgescu-Roegen, s.u. "Entropy", *The New Palgrave Dictionary of Economics*, vol. II, p. 156).

24 Cf. Jaynes, 1985, pp. 24-25.

25 Jaynes recorda (1979, pp. 41-42) como não conseguiu sensibilizar G. Uhlenbeck para os seus pontos de vista durante conversas tidas entre ambos no princípio dos anos 50. Contrapunha Uhlenbeck: "Entropy cannot be a measure of "amount of ignorance", because different people have different amounts of ignorance; entropy is a definite physical quantity that can be measured in the laboratory with thermometers and calorimeters". Cerca de 30 anos depois Jaynes lamenta não ter sabido responder na altura: "Certainly, different

A noção de informação quantificada está intimamente ligada ao conceito de probabilidade e incerteza e, como tal, não é de estranhar que os desenvolvimentos da Teoria da Informação acabassem por ter um impacto directo no próprio campo da Teoria da Probabilidade e da Inferência Estatística.

Jaynes, munido da sua técnica de maximização da entropia sujeita a restrições, erigida agora em princípio de inferência, iria assim abordar o problema básico da Inferência Bayesiana: a obtenção das probabilidades *a priori*. Tal será o assunto a tratar mais detalhadamente no terceiro capítulo.

O conceito de entropia, a caminho de século e meio de existência, tem-se assim revelado um verdadeiro Proteu, dando origem às mais acesas controvérsias quando dos seus múltiplos aparecimentos em diversos campos científicos. Está-se verdadeiramente perante um único conceito, por isso mesmo de transcendente significado, ou antes a levar longe demais certos paralelismos formais em campos que, pela sua própria

---

25 (cont. pág. ant.) people have different amounts of ignorance. The entropy of a thermodynamic system is the degree of ignorance of a person whose sole knowlegde about its microstate consists of the values of the macroscopic quantities X_i which define its thermodynamic state. This is a completely 'objective' quantity, in the sense that it is a function only of the X_i, and does not depend on anybody's personality".

diversidade, impedem à partida a unidade do conceito? No entanto, com a perspectiva proporcionada por 40 anos passados sobre a publicação do trabalho seminal de Shannon, poucos duvidarão do seu carácter fundamental e do seu notável contributo para a compreensão de assunto tão difícil e por isso mesmo tão fascinante da entropia e probabilidade.

2. O USO DA INFORMAÇÃO *A PRIORI* EM INFERÊNCIA

2.1. Introdução. Viu-se no capítulo anterior como uma linha de pensamento originada no campo da Termodinâmica Teórica tinha, por via da Mecânica Estatística, levado ao desenvolvimento da técnica da maximização da entropia. A perspectiva proporcionada pela Teoria da Informação facilitaria o retomar de um fio de ideias, que acabou por ter impacte na própria Inferência Estatística.

A passagem de uma proposição particular a uma proposição geral introduz no processo de inferência, de que é a própria definição, a noção de incerteza e, por essa via, deve fazer apelo ao conceito de probabilidade. É assim que só nos séculos XVI e XVII, com o início do estudo matemático das Probabilidades, foram lançadas as bases necessárias a uma abordagem formal aos processos de indução e inferência. Esta via de estudo focalizou-se na análise do que poderíamos chamar "induções locais", isto é, de inferências de tipo específico aplicáveis em situações bem delimitadas e, por isso mesmo, acessíveis a uma formalização matemática. Tal metodologia, que em sentido lato pode ser identificada com o campo da Estatística Teórica e Aplicada, tem contribuído fortemente para a elucidação de certos aspectos do problema da inferência, se bem que pela sua própria especificidade não possa pretender dar resposta às questões de índole filosófica mais geral.

O problema da delimitação das situações susceptíveis de serem formalizadas matematicamente tem subjacente a questão fulcral das várias interpretações a dar ao conceito de probabilidade. É assim que as várias escolas de Inferência que é possível entrever no panorama da Estatística do nosso século podem ser *grosso modo* associadas com correspondentes escolas de interpretação da probabilidade.

Tal assunto evoca preferencialmente a controvérsia, muitas vezes com aspectos de querela, entre frequencistas e Bayesianos. Tendo sempre presente que ambos os campos são largas coligações de tendências unidas por afinidades interpretativas mais do que coortes organizadas, não se está longe da verdade se se disser que os frequencistas invocam as probabilidades físicas, enquanto que os Bayesianos defendem o credo mais lato das probabilidades epistemológicas[1] (aqui incluindo a acepção subjectiva e a lógica).

No entanto, o controverso problema da interpretação a dar ao conceito de probabilidade tem preocupado ao longo do tempo algumas das maiores figuras da Teoria da Probabilidade.

---

1 Traduziu-se o termo "epistemic probability" utilizado por Good. Veja-se o artigo deste autor, "Subjective Probability", no *The New Palgrave Dictionary of Economics*, vol. IV, pp. 537-543.

2.2. O princípio da razão insuficiente: **Bernoulli, Bayes e Laplace.** O pioneiro Jacob I Bernoulli (1654-1705), nas três primeiras das quatro partes do seu livro *Ars Conjectandi* (1713), trata de problemas essencialmente relacionados com as probabilidades de jogos de azar, o que estava bem no espírito dos esforços desenvolvidos pelos seus contemporâneos probabilistas. Porém, na quarta parte desce a águas mais profundas e, depois de reconhecer o carácter limitativo da probabilidade associada aos jogos de azar[2], reconhece o interesse de tratar outros casos de índole mais geral, não sem pressentir os delicados problemas aí implicados[3].

Com o fim explícito de tentar quantificar as probabilidades de tais acontecimentos, Bernoulli introduz então a sua célebre

---

2 "But here, finally, we seem to have met our problem, since this may be done only in a very few cases and almost nowhere other than in games of chance, the inventors of which, in order to provide equal chances to the players, took pains to set up so that the numbers of cases would be known and ... so that all these cases could happen with equal ease." (Bernoulli, *Ars Conjectandi*, *in* Jaynes, 1979, p. 17)

3 "But what mortal will ever determine, for example, the number of diseases ... these and other such things depend upon causes completely hidden from us ..." (*Idem*, *ibid.*)

lei dos grandes números, pedra base na definição frequencista da probabilidade.

É também no *Ars Conjectandi* que surge explicitado pela primeira vez o princípio da razão insuficiente, assim chamado pelo próprio Bernoulli, e que em essência diz o seguinte: se a atribuição de uma probabilidade a uma proposição for considerada como descrevendo um certo estado de conhecimento sobre essa mesma proposição, e, se a evidência disponível não nos levar a favorecer a proposição P_1 em relação à proposição P_2, sendo ambas exaustivas, a atribuição correcta de probabilidades deverá ser de 1/2 para cada uma das proposições.

Se a aplicação de tal princípio é pacífica em certos casos de simetria aparente, como nos casos dos jogos de azar, tal não acontece em situações mais fluidas e imponderáveis, conforme Bernoulli assinalou (cf. nota 3). A vida atribulada de tal princípio até aos nossos dias não nos parece que possa ser imputada, como afirma Jaynes (1979, p. 16) a uma escolha infeliz de Bernoulli para o seu nome. Keynes (1921) não conseguiu acalmar a controvérsia ao mudar o nome para princípio da indiferença. Não há dúvida de que a designação de *desideratum* de consistência proposto por Jaynes (1979, p. 16) soa ainda algo melhor. No entanto, o problema parece ser de

tipo essencial[4] e as dificuldades associadas a um conceito não podem ser exorcizadas com recurso a qualquer cosmética nominalista.

As dificuldades do princípio estão associadas às da própria definição clássica da probabilidade para a qual ele é aliás considerado como um suplemento necessário. A justificação para o princípio está ligada ao problema da simetria na partição das alternativas elementares consideradas.

---

4 Consideremos a crítica feita por H. Reichenbach, filósofo empiricista e defensor de uma interpretação frequencista da probabilidade, crítica dirigida segundo o clássico eixo de contenda filosófica entre empiricismo e racionalismo. Afirma Reichenbach: "The principle [a que chama "of indifference" ou "of no reason to the contrary"] [...] is considered by the rationalist as a postulate of logic. It appears to him self-evident, like logical principles. The difficulty with this interpretation of probability is that it abandons the analytic character of logic and introduces a synthetic a priori. [...] The principle of indifference leads rationalism into all the familiar difficulties known from the history of philosophy [...] Does nature conform to human ignorance? Questions of this type cannot be given a positive answer - or the philosopher has to believe in a harmony between reason and nature, that is, a synthetic a priori. [...] The rationalist interpretation of probability must be regarded as a remnant of speculative philosophy and has no place in scientific philosophy" (Reichenbach, 1951, *in Space, Time and the New Mathematics*, pp. 60-61)

Meio século depois do aparecimento do *Ars Conjectandi* surgiu a contribuição de Th. Bayes (1702-1761) intitulada "An Essay towards solving a problem in the doctrine of chances" (1763). O artigo pode ter sugerido aos contemporâneos a marca de um diletante competente debruçando-se sobre um assunto difícil. Daí talvez a falta de interesse que suscitou, só vindo a sair da obscuridade alguns anos mais tarde, pela atenção que Laplace lhe dedicou. Hoje, com mais de duzentos anos, o Essay, que seguramente continua a dispor de uma invejável situação em termos de citações na literatura estatística e filosófica, continua a gerar grande controvérsia[5].

No entanto, o problema a tratar está claramente enunciado nas próprias palavras de Bayes, aparecidas no início do artigo e que bem poderiam figurar como um moderno *abstract*:

---

5 O próprio grau de competência e sofisticação matemática e lógica de Bayes é um assunto relativamente controverso. Veja-se, por exemplo, a opinião de George Barnard, em conversa com Morris de Groot transcrita no *Statistical Science* (Vol. 3, Nº 2, 1988, p. 203): "And he [Bayes] wasn't that bad a mathematician; in fact he was a good one, contrary to what Steve Stigler argues [Sthephen M. Stigler. *The History of Statistics*, Harvard University Press, Cambridge, Massachussetts]. You know Steve says Bayes was just a minor figure and it was really Laplace who did it all. In fact, I think Bayes had a subtle sense of logic.".

"*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named" (*in International Encyclopedia of the Social Sciences*, s.u. "Bayes, Thomas", I, p. 26). Para a sua solução, que conduziu à distribuição beta, Bayes apresentou duas ideias originais. A primeira envolve uma expressão para a probabilidade condicional, que ainda hoje, mas numa forma mais generalizada, tem o nome de teorema de Bayes. A segunda, bem mais controversa, encontra-se na linha do princípio da razão insuficiente proposto por Bernoulli, e consiste em postular[6] uma lei uniforme para a probabilidade do acontecimento, anteriormente ao número de sucessos e insucessos verificados ou, na terminologia Bayesiana, utiliza uma distribuição *a priori* uniforme.

6 É no famoso *Scholium* que Bayes adianta uma justificação para o seu postulado: "that the ... rule is the proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, seems to appear from the following consideration; viz. that concerning such an event I have no reason to think that, in a certain number of trials it should rather happen any one possible number of times than another." (*in International Encyclopedia of the Social Sciences*, s.u. "Bayes, Thomas", I, pp. 26-27).

Em 1774, o matemático francês Laplace (1749--1827), publicava o *Mémoire sur la Probabilité des causes par les événements* [7].

Sem mencionar Bayes[8], Laplace propõe no seu *Mémoire* a expressão,

$$p_j \Big/ \sum_{j=1}^{n} p_j \,, \tag{2.1}$$

como dando a probabilidade de que a causa j (de entre n causas exaustivas) tenha sido operativa na realização de um dado acontecimento e onde p_j é a probabilidade desse acontecimento, sendo j a causa operativa. Nesta expressão está implícito que a probabilidade de qualquer das causas seja igual.

Em 1778, Laplace justifica a uniformidade na probabilidade das causas, numa linha de pensamento paralela à

7 Sthephen Stigler afirma: "The memoir is an explosion of ideas that left an indelible imprint on statistics. In this one article we can recognize the roots of modern decision theory, Bayesian Inference with nuisance parameters, and the asymptotic approximation of posterior distributions" (1986, p. 359).

8 Ver Stigler (1986) para a questão do conhecimento de Laplace sobre a obra de Bayes, em 1774.

do *Scholium* de Bayes[9].

Na sua célebre *Théorie analytique des probabilités* de 1812, Laplace reapresenta o seu resultado de 1774, o qual desde então tinha utilizado com o maior sucesso em problemas de Mecânica Celeste. Na segunda edição (1814), generaliza a expressão para o caso de as causas não serem equiprováveis, apresentando a expressão,

$$w_j \; p_j \; / \; \sum_{j=1}^{n} w_j \; p_j \; , \qquad (2.2)$$

onde w_j representa a probabilidade *a priori* da causa j. No entanto, foi utilizando a expressão mais particular, aquela que pressupunha a uniformidade das probabilidades *a priori* , que Laplace obteve alguns dos seus mais brilhantes resultados. Porque prescindiu Laplace, nas suas aplicações, da maior generalidade da sua segunda expressão de 1814? O facto é que a utilização desta expressão pressupõe o cálculo das probabilidades *a priori* em situações não triviais de uniformidade, cálculo que na generalidade, sendo sumamente

9 "Lorsqu'on n'a aucune donnée a priori sur la possibilité d'un événement, il faut supposer toutes les possibilités depuis zéro jusqu'à l'unité également probables ..." (Laplace, p. 27 da o. c. p. 26).

difícil, ainda hoje permanece como um dos maiores obstáculos a uma mais corrente utilização das técnicas Bayesianas.

A via indicada por Laplace ficou abandonada durante mais de um século, até ao aparecimento da *Theory of Probability* de Harold Jeffreys em 1939, que pode ser considerado como um marco fundamental no ressurgir de um linha Bayesiana em Inferência Estatística. A que é devido um hiato de mais de cem anos no reatar de uma linha de investigação que já tinha dado os seus frutos com Laplace? O problema da dificuldade técnica da eliciação das probabilidades *a priori* não é de forma alguma justificação única. Na verdade, as ideias de Laplace vinham sendo alvo do estigma de *ad hoc* e acusadas de não obedecerem a um *desideratum* com boa fundamentação lógica. Desde a década de 1840 que se gera nos círculos científicos e filosóficos ingleses uma oposição consistente às ideias de Laplace sobre as probabilidades *a priori* por parte de figuras que iriam ser os proponentes de uma estrita interpretação frequencista do conceito de probabilidade. Ellis (em 1842), Boole (em 1854), Venn (em 1866), Chrystal (em 1891) recusam liminarmente qualquer postulado do tipo do princípio da razão insuficiente como sendo uma fantasia de natureza metafísica, não susceptível de ser justificada racionalmente. Alguma aceitação pontual no continente europeu por parte de figuras

conceituadas, como von Kries (1886), por exemplo, não é suficiente para inverter o sentido de relativo descrédito científico que passa a pesar sobre a concepção de Bayes - - Laplace sobre a probabilidade *a priori* [10].

No segundo quartel deste século, o fiel da balança desequilibra-se totalmente, por via do ascendente ganho em Estatística pela *praxis* Fisheriana e pela escola de Neyman - - Pearson - Wald, chegando as ideias de Bayes - Laplace a apresentar o estigma de anátema. Só no início do terceiro quartel se assistirá ao retomar do interesse pela perspectiva Bayesiana, dando origem a um movimento de reapreciação de ideias que continua nos nossos dias[11].

---

10 Veja-se a controvérsia gerada à volta da regra da sucessão de Laplace como o exemplo mais em destaque na contestação feita ao uso não crítico da teoria da probabilidade inversa.

11 "The modern Bayesian movement has increased the interest in the philosophy of science among statisticians during the last several decades, although most statisticians are still naturally more concerned with techniques such as the theory and application of linear models. Also several philosophers of science have become interested in Bayesianism" (Good, "Scientific Method and Statistics", *Encyclopaedia of Statistical Sciences*, Vol. VIII, p.295).

2.3. As várias escolas de inferência perante o uso da informação *a priori*. Focando, em primeira aproximação, três escolas no panorama actual da Inferência - a escola de Fisher, a de Neyman - Pearson - Wald (NPW) e a Bayesiana -, passa a referir-se a posição que tomam a respeito do problema da incorporação da informação *a priori* no processo de inferência.

2.3.1. A escola de Fisher. A escola Fisheriana é de entre as três a que tem um carácter menos formalizado, sendo geralmente bem aceite entre os praticantes das ciências experimentais. A atitude de base de Fisher quanto ao problema da inferência[12] reflectiu sempre o ponto de vista do investigador confrontado com problemas concretos nos

12 Joan Fisher Box refere na biografia científica que escreveu sobre o pai: "Fisher's scientific interests continually confronted him with the basic problem of Statistics: how to make inferences from the particular to the general. From the first paper in 1912 to his last 50 years later, he was exercized by the need to express with precision what might justly be concluded from the variable data about uncertain events. Much of his early work had been devoted to what he came to regard as the lowest level of scientific inference - to tests of significance which make a dichotomy between hypotheses that are discredited by the data and those that are not. At the highest level

domínios das várias ciências[13]. No entanto, as implicações de tipo mais filosófico suscitadas pelas técnicas por si propostas também lhe mereceram a atenção[14].

A atitude de Fisher em relação à aplicabilidade geral da metodologia Bayesiana em Estatística foi sempre de oposição[15]. Clara foi também a sua rejeição de qualquer postulado na linha do princípio da razão insuficiente, só aceitando a incorporação das probabilidades *a priori* no

---

12 (cont. pág. ant.) under certain conditions the whole probability distribution of the parameter of interest could be derived from the data using Bayes' theorem. However this was only possible if the probability distribution of the parameter was known a priori. (J. Fisher Box, 1978, pp. 447-448).

13 Não será disso indicativo o título escolhido por Fisher para o livro que permanece como a sua obra mais conhecida: *Statistical Methods for Research Workers* (1925)?

14 Contudo, não há unanimidade no julgar das suas aptidões neste último domínio. Veja-se o juízo de Good: "He [Fisher] was a great praticioner, but a mediocre philosopher" (Good, "Subjective Probability", *The New Palgrave Dictionary of Economics*, vol. IV, p. 542).

15 "Fisher had always been derisory of the estimates and inferences resulting from the Bayes' inverse-probability approach" (Bartlett, *Int. Encycl. of the Soc. Sc.*, s.u. "Fisher, R. A.", vol. V, p. 488).

mecanismo de inferência em casos bem específicos e só quando houvesse sólidas bases científicas para a sua obtenção (e. g. em certos estudos genéticos)[16].

Fisher reconhecia, no entanto, o carácter insuficiente e não satisfatório da inferência baseada em testes de significância, a que o estatístico se vê obrigado a recorrer nos casos mais comuns em que o conhecimento *a priori* é considerado inutilizável.

Numa terceira situação[17] poderia utilizar-se o método fiducial. Tal situação exige, contudo, a falta de conhecimento *a priori* sobre o parâmetro e a existência de estatísticas suficientes e quantidades "pivotais" nas quais se verifique uma

16 "Fisher accepted Bayes' theorem and honored its originator as the first to use the teory of probability as an instrument of inductive reasoning. But from the beginning he rejected Bayes' postulate. He argued that not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge. Then he accepted and used the Bayesian method only in those cases in which the prior probabilities were known, as they sometimes are, for example, in genetical work". (J. Fisher Box, 1978, p. 449).

17 cf. final do cap. 28 em Fisher (1956) e pp. 447 - 448 de J. Fisher Box (1978).

relação especial de reciprocidade entre estatística e parâmetros. Em tal caso, o método proporciona como resultado de inferência, e à semelhança do formalismo Bayesiano, uma distribuição de probabilidade para o parâmetro.

A situação actual do sistema fiducial é problemática, sendo este sistema pouco utilizado e havendo dificuldades no próprio conceito de probabilidade fiducial[18].

Assim, a experiência acumulada por Fisher ao longo de

---

18 Não deixa, no entanto, de suscitar uma certa curiosidade admirativa: "... its formal bypassing of Bayes' theorem was a masterly stroke which received attention outside statistical circles" (o. c. na nota 15, ibid.). Mas a crítica mais saborosa, vinda de um Bayesiano personalista confesso e que nela não deixa de introduzir uma homenagem, talvez seja a de J. Savage: "Fisher's school, with its emphasis on fiducial probability - a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs - may be regarded as an exception to the rule that frequentists leave great latitude for subjective choice in statistical analysis" (Savage, 1961, p. 578). No entanto, é interessante que o argumento fiducial tenha sido recentemente retomado por investigadores como Fraser, Wilkinson e Barnard. Em conversa com Morris DeGroot publicada no *Statistical Science* Barnard refere: "I realize yet again that it's a pity I didn't read carefully what Fisher wrote me back in the late 40's because I think the pivotal idea in a sense allows you to be a Bayesian in so far as you need to be. My opinion now is that the

uma vida de investigação criativa parece tê-lo levado a uma posição de "sagesse" reconciliada com o reconhecimento da inexistência de uma solução unitária para os problemas da inferência[19].

2.3.2. A escola de Neyman-Pearson-Wald. A segunda escola de inferência considerada, a de NPW[20], teve origem na colaboração verificada a partir dos finais dos anos 20 entre Egon Pearson (1895-1980) e Jerzy Neyman (1894--1981). Neyman e Pearson alargaram o conceito Fisheriano de teste de significância, ao considerarem explicitamente, para além da hipótese nula, a sua hipótese alternativa. Trabalhando no quadro mais inclusivo da teoria do ensaio de hipóteses,

---

18 (cont. pág. ant.) proprer process for statistical inference is conditioning on known values of variables whose distribution is known" (1988, Vol. 3, Nº 2, p. 207).

19 "Fisher did not offer a universal solution to the problems of inference. Instead, looking back over a lifetime of pondering these problems, he considered various inferencial techniques, each appropriate to certain kinds of problems but none appropriate to all" (J. Fisher Box, 1978, p. 448).

20 É em certa medida uma versão formalizada da *praxis* estatística Fisheriana. Veja-se a opinião expressa por Lindley no artigo "Statistical Inference", *The New Palgrave Dictionary of Economics*, vol. IV, p. 491.

procederam então a um estudo de optimalidade, o que tinha características inéditas em Estatística, com base num critério de contenção das probabilidades de cometer erros de tipo bem especificado[21].

Não foi de ânimo leve que Neyman e Pearson se decidiram pela não incorporação da informação *a priori* no mecanismo inferencial. Assim, só depois de reconhecerem as dificuldades práticas na quantificação das probabilidades *a priori*, é que estes autores optaram por uma aderência estrita a medidas de probabilidade que pudessem ser relacionadas com frequências relativas[22].

---

21 Cerca de vinte e cinco anos depois da publicação dos artigos fundamentais de colaboração entre Neyman e Pearson, este último recorda: "What I think we found [...] was a dissatisfaction with the logical basis, or lack of it which seemed to underly the choice and development of statistical tests [...] We tried therefore to find a set of principles with a mathematical basis which it seemed to us would lead to a rational choice of statistical procedures when faced with certain types of problem in the analysis of data [...] No doubt, because the scope of application of statistical methods in those days was narrower, the emphasis which we gave to certain types of situation may now seem out of balance" (*The Foundations of Statistical Inference*, pp. 54-55).

22 Pearson refere ainda sobre este ponto: "I think I am right in saying that it was Neyman, brought up in the continental mathematical

Wald (1902-1950), a outra grande figura desta escola, viria a reunir as teorias da estimação e do ensaio de hipóteses estatísticas no quadro único e mais geral da decisão estatística em face da incerteza. Um artigo fundamental (1939) permaneceu relativamente desconhecido durante alguns anos. Os seus trabalhos culminaram no livro *Statistical Decision Functions* (1950). É assim que os últimos anos da década de 40 viram o magnífico esforço desenvolvido por Wald erigir uma construção que, se bem que supostamente obediente a uma traça frequencista, se veio a revelar ter pontos de contactos com conceitos de tipo Bayesiano. Isto, aliás, viria a ser reconhecido pelo próprio autor, dentro de uma linha de seriedade intelectual que não poderia deixar de ser característica de um arquitecto de tal classe[23].

---

22 (cont. pág. ant.) school, who held longest to the idea of retaining in our theory measures of prior probability" (ibid., p. 55).

23 "It is one of these ironies that make history of Science so interesting, that the missing Bayes-optimality proofs, which Laplace and Jeffreys had failed to supply, were at last found inadvertently, while trying to prove the opposite, by an ardent disciple of the von Mises' collective approach. It is also a tribute to Wald's intellectual honesty that he was able to recognize this, and in his final work (Wald, 1950) he called these optimal rules "Bayes Strategies" (Jaynes, 1979, p. 26).

2.3.3. A escola Bayesiana. Eis-nos assim chegados à década de 50, que veria o ressurgimento da ideia Bayesiana em Estatística, após, durante o 2º quartel do século XX, o fiel da balança se ter desiquilibrado totalmente a favor das concepções frequencistas.

O terreno para o ressurgimento Bayesiano do 3º quartel do século XX, e que continua nos nossos dias, vinha a ser preparado, nos aspectos mais gerais da interpretação e fundamentação axiomática da probabilidade, por Keynes (1883--1946), Ramsey (1903-1930) e de Finetti (1906-1985) e, nos aspectos metodológicos e de utilização operacional, por Jeffreys (1891-). A metodologia seguida por Jeffreys, que, à semelhança de Laplace, tinha pela frente alguns problemas concretos, agora não de Mecânica Celeste mas de Geofísica, foi alvo de muitas críticas vindas dos círculos frequencistas. Estas críticas passaram por cima dos resultados alcançados por Jeffreys, que, tal como no caso de Laplace, tinham valor científico indiscutível. Centraram-se na falta de *desiderata* e de rigor lógico, no espírito *ad hoc* dos métodos, aos quais faltaria uma justificação lógica conveniente. Mais uma vez o dedo era apontado ao "pecado original" Bayesiano: a obtenção das probabilidades *a priori*.

No entanto, Jeffreys tinha tentado enfrentar corajosamente

esse difícil problema, retomando a questão clássica de como exprimir formalmente a ignorância completa com respeito a um parâmetro contínuo. Estudando os casos de um parâmetro de localização μ susceptível de assumir qualquer valor real e de um parâmetro de escala σ com domínio positivo, Jeffreys propôs a consideração dos elementos de probabilidade proporcionais a, $d\mu$, e a, $(1/\sigma)\, d\sigma$, respectivamente. O facto de tais distribuições *a priori* serem impróprias fez que tais resultados fossem de imediato muito atacados. Na 2ª edição, de 1948, da sua *Theory of Probability*, Jeffreys propõe uma teoria de invariância bastante mais geral[24] para a determinação de distribuições *a priori* não informativas.

A figura mais em destaque no ressurgimento Bayesiano da década de 50 é, sem dúvida, L. J. Savage (1917-1971), autor do livro *The Foundations of Statistics*, cuja publicação em

24 "[Jeffreys] showed armazing prevision by coming within a hair's breadth of discovering both the principles of Maximum Entropy and Transformation Groups. He wrote down the actual entropy expression (note the date!), but then used it only to generate a quadratic form by expanssion about its peak. Jeffrey's Invariance Theory is still of great importance today, and the question of its relation to other methods that have been proposed is still under study" (Jaynes, 1979, p. 25).

1954[25] é um marco no reabrir desta via da Inferência.

Convém agora delinear os contornos com que a escola Bayesiana se apresenta na disputa com as outras escolas da primazia no terreno da Estatística. O sistema Bayesiano é, em comparação com as escolas de Fisher e de NPW, aquele que possui uma estrutura mais formalizada. Com efeito, toda a mecânica inferencial Bayesiana pode ser expressa dentro do quadro único do cálculo das probabilidades, beneficiando directamente do carácter altamente formalizado deste último. No entanto, e aqui a diferença é verdadeiramente crucial, a perspectiva Bayesiana aceita e, mais do que isso, necessita de que uma distribuição de probabilidade, $\pi(\theta)$, para o parâmetro Θ, sobre o qual se quer fazer inferência, seja reconhecida como

25 "Many of the papers he wrote between 1954 and his death [1971] are concerned with his increasing understanding of the power of Bayesian Methods, and to read these in sequence is to appreciate a great mind making new discoveries. His concern is usually with matters of principle[...] Proper judgement of the importance of Savage's work can only come when the Bayesian paradigm is established as the Statistical method or when it is shown to be defective. Even if the latter happens, Savage can be credited with creating a method that led to important new ideas of lasting value. He was a true originator" (Lindley, *Encyclopaedia of Statistical Sciences*, s.u. "Savage, Leonard J.", vol VIII, p. 266).

podendo exprimir um conhecimento *a priori*, i. e., anterior ao conhecimento dos dados X , sobre esse mesmo parâmetro, Θ. É só na base desta premissa que o sistema de inferência Bayesiana pode ser posto em movimento, beneficiando então da transmissão integral da potência do cálculo das probabilidades.

O resultado da inferência é uma distribuição de probabilidade para o parâmetro, $\pi(\theta|x)$, a chamada distribuição *a posteriori* e que exprime a alteração introduzida pelo conhecimento dos dados x na distribuição *a priori* $\pi(\theta)$. A passagem da distribuição *a priori* à distribuição *a posteriori* faz-se mediante a utilização do teorema de Bayes,

$$\pi(\theta|x) \propto f(x|\theta).\ \pi(\theta)\ , \tag{2.3}$$

ou, tendo a preocupação de apresentar a distribuição *a posteriori* normalizada,

$$\pi(\theta|x) = \frac{f(x|\theta).\pi(\theta)}{\int f(x|\theta)\ \pi(\theta)\ d\theta}\ , \tag{2.4}$$

onde f (x|θ) é a chamada verosimilhança.

Importa reconhecer o carácter essencialmente recursivo de tal sistema de inferência, em que uma distribuição *a posteriori* entretanto obtida pode ser utilizada como uma distribuição *a priori* anterior à consideração de novos dados, os quais, uma vez incorporados na fórmula de Bayes por via da verosimilhança, darão origem a uma nova distribuição *a*

posteriori, podendo tal mecânica inferencial repetir-se sucessivamente.

Em maior aproximação podem distinguir-se, dentro da escola Bayesiana, duas formas distintas de interpretar a distribuição *a priori* e que correspondem directamente às duas formas de considerar uma probabilidade epistemológica - a subjectiva e a lógica.

Na primeira, que caracteriza o que pode chamar-se de Bayesianismo subjectivo ou personalista, as probabilidades *a priori* são entendidas como exprimindo um grau de credibilidade que pode variar de indivíduo para indivíduo. Tais probabilidades poderão ser obtidas por recurso a uma maiêutica que consista num sistema de apostas, com o comportamento individual obedecendo a um certo tipo de coerência[26].

---

26 As probabilidades individuais devem ser tais que nunca possibilitem o aparecimento de uma aposta de prejuizo certo: o chamado "Dutch book". Savage, a figura mais em destaque nesta escola, resume em poucas palavras o seu credo: "For many of us, it is most stimulating to think of the odds that Mr. So-and-So would be just willing to offer in favor of an event measuring his personal probability through the formula probability = odds/(1+odds) [...] The concept of personal probability [...] seems to those of us who have worked with it an excellent model for the concept of opinion [...] I will

A segunda corrente, o Bayesianismo lógico, defende que as probabilidades *a priori* representem uma relação quase lógica entre a evidência e uma hipótese. A probabilidade medirá o grau de implicação entre proposições, o qual deverá ter carácter racional, impessoal e, por isso mesmo, objectivo: qualquer indivíduo na posse da mesma informação deve obter uma mesma distribuição *a priori* que assim adquire um carácter de necessidade lógica.

O caso do desconhecimento total *a priori* sobre o parâmetro tem sido assunto de especial interesse para a escola Bayesiana lógica. Já atrás foram referidos os trabalhos de Jeffreys sobre este difícil problema, na linha de uma tradição com raízes em Bernoulli, Bayes e Laplace. Hartigan (1964) e Jaynes (1968) voltaram ao problema, sendo o último dos autores citados o proponente principal da utilização do princípio da maximização da entropia em inferência.

---

26 (cont. pág. ant.) confess however that I and some other Bayesians hold this to be the only valid concept of probability and therefore the only one needed in statistics, physics or other applications of the idea. In particular, we radical Bayesians claim that all that is attractive about the frequency theory of probability is subsumed in the theory of personal probability" (Savage, 1961, pp. 581-582).

A conjunção das técnicas de invariância com as de maximização de entropia constitui um método prometedor na procura de probabilidades *a priori* objectivas. Tal será o assunto a tratar mais detalhadamente no próximo capítulo.

Apesar de serem evidentes as diferenças, nas concepções filosóficas de base sobre a probabilidade, entre a linha subjectiva e a lógica, parece não estar em causa o facto de se considerarem ambas de raíz Bayesiana[27].

Os seguidores da linha lógica, se bem que desejando utilizar o formalismo Bayesiano em Inferência[28], recusam a

---

27 "Harold Jeffreys holds what I call a necessary view of the theory of probability, but such a view is Bayesian in the broad sense that is makes use of Bayes' theorem and therefore demands a thorough exploration of Bayes' theorem for its application to statistics." (Savage, 1961, p. 582).

28 Note-se que, se por um lado a estrutura lógica da escola inferencial frequencista revela a ausência de princípios orientadores gerais para a geração de técnicas estatísticas (se bem que existam propostas por esta escola técnicas específicas para o tratamento de problemas de tipo bem determinado que são de utilização eficaz e de interpretação intuitiva), por outro lado a escola Bayesiana segue uma via totalmente distinta.. A linha inferencial Bayesiana procura técnicas estatísticas a partir de critérios de razoabilidade estabelecidos *ab initio* utilizando o método dedutivo, o que é, em certa medida, mais consentâneo com a metodologia geral seguida em matemática.

atitude radical de Savage de aceitar *ab initio*, e sem qualquer sentido de reserva mental, uma distribuição subjectiva e pessoal de probabilidades *a priori* como sustentáculo de uma investigação científica[29]. Neste sentido, os Bayesianos lógicos compartilham da inquietação sentida pelos frequencistas de verem em eminente risco de desmoronamento, como resultado da infiltração do subjectivismo no terreno científico, de veneráveis edifícios construidos em obediência a uma estrita fé objectiva,[30] que reconhecem como única.

Assim , a conjugação entre o desejo de utilizar o mecanismo inferencial Bayesiano, não abandonando por outro

---

29 "Nevertheless, the author must agree with the conclusions of orthodox statisticians, that the notion of personalistic probability belongs to the field of psychology and has no place in applied statistics [...] An infortunate impression has been created that rejection of personalistic probability automatically means the rejection of Bayesian methods in general." (Jaynes, 1968, p. 88).

30 "To have a unifield mathematical model of the mind's working in all these varied situations is certainly intellectualy attractive. But is it always meaningful? I think that there is always this question at the back of my mind: can it really lead to my own clear thinking to put at the very foundation of the mathematical structure used in acquiring knowledge, functions about whose forms I have often such imprecise ideas?" (E. Pearson, "Some thoughts on statistical inference", *in* Vic Barnett, 1973, p. 199).

lado o ideal científico de objectividade, deixa aberto à escola Bayesiana lógica o árduo caminho da procura de probabilidades *a priori* objectivas. Tal busca, directamente ligada à obtenção de distribuições não informativas, foi já considerada como uma verdadeira procura do Graal em Estatística[31].

31 "The formalization of ignorance thus remains the central object of a continuing quest by the knights of the Bayesian round table: inspiring them to imaginative feats of daring, while remaining, perhaps, forever unattainable" (Dawid, "Invariant Prior Distributions", *Encyclopaedia of Statistical Sciences*, vol. IV, p. 235).

3. A MAXIMIZAÇÃO DA ENTROPIA NA PROCURA DE DISTRIBUIÇÕES *A PRIORI*

3.1. Introdução. A controvérsia sobre a inferência Bayesiana e sobre a sua aceitabilidade como metodologia científica centra-se na posição tomada por esta escola de exprimir a informação *a priori* sobre um parâmetro na forma explícita de uma distribuição de probabilidade.

A posição subjectiva e personalista de Savage, se bem que pareça cortar cerce o nó górdio do problema da obtenção das probabilidades *a priori*, não deixou de suscitar as maiores reservas entre muitos dos praticantes das ciências exactas, como já foi referido. Por outro lado, sendo patentes os atractivos da metodologia Bayesiana, permaneceu arreigado dentro da corrente estatística a que se chamou de Bayesiana lógica, necessária ou objectiva, o desejo de basear a inferência em distribuições *a priori* objectivas[1]. A procura de tais distribuições leva invariavelmente à velha e difícil questão da formalização probabilística da ignorância, problema com raízes históricas já longas e que se tentou expor no capítulo anterior.

Ir-se-á agora reconhecer no método da maximização da entropia (MAXENT) uma técnica específica para a obtenção de

1 "The value of an objective theory would be so great that it seems... well worth trying to see wheter some objective and natural choice of the prior distribution is possible". (Lindley, 1961, p. 465)

distribuições *a priori* objectivas que representam um estado de conhecimento livre de qualquer "contaminação" subjectiva. Para isso há que restringir à partida o tipo de informação que será considerada utilizável no mecanismo de inferência. Ver-se-á então que o MAXENT é, considerando em segundo plano dificuldades técnicas de tipo calculatório que por vezes lhe podem estar associadas, um método interessante e dispondo de uma boa justificação lógica para a procura de probabilidades *a priori*.

O MAXENT permitirá obter, dentro da classe de distribuições *a priori* que incorporam as peças de informação objectiva disponíveis, aquela que é menos informativa no sentido em que maximiza uma determinada funcional, a entropia, a qual mede precisamente o carácter não informativo de uma distribuição.

Nos casos, surgidos com frequência nas Ciências Físicas, em que a informação *a priori* toma a forma do conhecimento de momentos, o método é então de utilidade comprovada com soluções já conhecidas do cálculo das variações.

Quando o domínio de variação do parâmetro em estudo é contínuo, surge então uma dificuldade técnica na definição da entropia de uma distribuição, invariante para reparametrizações. O recurso a distribuições *a priori* de tipo

não informativo geradas por técnicas de invariância, terreno de investigação fundamental para o Bayesianismo lógico, pode então revelar-se de utilidade para o MAXENT.

3.2. Informação *a priori* testável. Considere-se em primeiro lugar o *desideratum* básico que deve orientar a procura de distribuições *a priori*, enunciado por Jaynes nos seguintes termos: "in two problems where we have the same prior information, we should assign the same prior probabilities" (1968, p. 88).

Parece fútil pretender discernir a igualdade ou desigualdade entre peças de informação, se não for de alguma maneira possível medi-las sem qualquer ambiguidade numa escala quantitativa. Ficam, portanto, afastadas à partida quaisquer informações de cariz subjectivo. Jaynes propõe que se considere unicamente informação dita testável ("testable", cf. Jaynes, 1968, p. 89), i. e., aquela que, uma vez obtida a medida de probabilidade, $\pi(\theta)d\theta$, permita testar inequivocamente se a medida obtida respeita ou não tal informação. São claramente peças testáveis de informação:

(3.1.a) I_1: $a<\Theta<b$;

(3.1.b) I_2: $E[\psi(\Theta)] = c$;

(3.1.c) I_3: $P(\Theta\leq d) = \alpha$;

(3.1.d) I_4: a distribuição é simétrica em relação à abcissa $\theta = c$.

Por outro lado, as informações:

(3.1.e) I_5: há razões para crer que $E[\psi(\Theta)] = c$;

(3.1.f) I_6: com alta probabilidade Θ assume valores entre a e b,

se bem que relevantes, não sendo testáveis, não podem ser utilizadas como restrições num problema de MAXENT.

Dentro da classe de distribuições *a priori* que satisfaçam as peças de informação testável, i.e., restrições com carácter numérico à forma funcional de, $\pi(\theta)$, deve então escolher-se aquela que maximiza a entropia. Como tal, a distribuição *a priori* MAXENT será aquela que representa uma incerteza máxima sobre o parâmetro, tomando, por outro lado, em consideração o conhecimento *a priori* testável na situação em estudo.

3.3. Entropia de uma distribuição discreta. Analise-se com mais pormenor a definição de entropia de uma distribuição, noção introduzida por Shannon (1948) (cf. cap. 1, p. 13). O estudo será restrito por agora ao caso de uma distribuição de probabilidade para o parâmetro Θ, definida sobre um conjunto discreto $\tilde{\Theta}$ (o espaço do parâmetro).

A entropia de $\pi(\theta)$, designada por $En(\pi)$, é definida como,

(3.2) $$En(\pi) = -\sum_{\Theta} \pi(\theta_i) \ln \pi(\theta_i) = -E[\ln \pi(\Theta)]$$

Se $\pi(\theta_k) = 0$, então a quantidade $\pi(\theta_k) \ln \pi(\theta_k)$ é definida como sendo 0, de acordo com

(3.3) $$\lim_{\pi(\theta_k)\to 0} \pi(\theta_k) \ln \pi(\theta_k) = 0.$$

Caso a distribuição de probabilidade, $\pi(\theta)$, tenha um número finito de pontos de massa, a entropia será função dessas massas e independente das abcissas em que se situem,

(3.4) $$En(\pi) = f\left(\pi(\theta_1), \pi(\theta_2), \ldots, \pi(\theta_n)\right).$$

Tem-se assim que $En(\pi)$, dependendo funcionalmente de $\pi(\theta_i)$, será invariante para qualquer transformação biunívoca dos valores de Θ, i. e., para qualquer reparametrização dada. Como tal, $En(\pi)$ pode ser considerada como uma quantidade descritiva sobre a distribuição $\pi(\theta)$, indicando numericamente o grau de repartição da unidade de massa, independente da localização num eixo numérico dos pontos de massa da distribuição[2].

2 Tisza (1966) sugere para $En(\pi)$ o nome de "dispersal" de uma distribuição. Denbigh e Denbigh referem: "This would have been admirable and would have resulted in the avoidance of much confusion. Unfortunately the name entropy for Shannon's measure has become so widespread that it seems hopeless to try to put the clock back" (1985, p.105). (cf. também notas das pp.19 e 20, cap.1)

Especialmente importante no contexto deste estudo é o facto de $En(\pi)$ poder ser tomada como medida de quantidade de incerteza inerente a uma distribuição, ligando-se tal interpretação directamente à noção de informação. Para tal pode ser útil a consideração de uma experiência aleatória em que uma variável com tal distribuição é observada. Anteriormente à realização da experiência, o grau de incerteza quanto ao seu resultado é igual a $En(\pi)$. Sendo a incerteza igual a zero logo após o conhecimento do resultado, e interpretando informação como tendo o efeito de diminuir a incerteza, $En(\pi)$ poderá então ser considerada como uma medida natural de informação[3].

Têm sido apresentadas em Estatística várias outras medidas de informação directamente relacionadas com a entropia. Entre as mais conhecidas estão os números de informação de Kullback - Leibler, $I(1,2)$ e $J(1,2)$, definidos como,

$$I(1,2) = \sum_i \pi_1(\theta_i) \ln \frac{\pi_1(\theta_i)}{\pi_2(\theta_i)}, \tag{3.5}$$

3 "To speak about information or about uncertainty means essentially the same thing: in the first case we consider an experiment which has been performed, in the second case an experiment not yet performed" (Rényi, 1970, pp. 553-554)

(3.6) $$J(1,2) = I(1,2) + I(2,1).$$

I(1,2) mede a diferença, divergência ou distância estatística entre duas distribuições, sendo igual a zero se, e só se, $\pi_1(\theta_i) \equiv \pi_2(\theta_i)$. A potência de testes estatísticos está directamente relacionada com (3.5), devendo notar-se que a expressão considerada tem a estrutura do valor esperado do logaritmo de uma razão de verosimilhanças. J(1,2) é uma medida simetrizada da distância estatística entre as duas distribuições.

Rényi (1961) introduziu a noção de entropias de ordem α definidas por

(3.7) $$En_\alpha(\pi) = (1-\alpha)^{-1} \ln \left(\sum_i \pi(\theta_i)^\alpha \right) .$$

Quando em (3.7) $\alpha \to 1$, $En_\alpha(\pi)$ reduz-se à noção habitual de entropia. Kemp (1975) debruçou-se sobre a utilidade de $En_2(\pi)$ como medida descritiva de uma distribuição.

É uma característica notável da entropia de Shannon, $En(\pi)$, o facto de poder ser deduzida com carácter único na base de sistemas simples de axiomas. Tais sistemas de axiomas devem exprimir certas condições razoavelmente esperadas de uma medida de informação.

Tome-se como exemplo o sistema de axiomas apresentado pelo próprio Shannon (1948). Seja agora $En(\pi)$ uma funcional

de incerteza sobre uma distribuição $\pi(\theta)$ e que satisfaça as seguintes três condições:

(3.8) Ax. 1: $En(\pi) = f(\pi_1, \pi_2, \ldots, \pi_n)$ é uma função contínua dos π_i ;

Ax 2: Em caso de equiprobabilidade, i.e., quando $\pi_i = \frac{1}{n}$, tem-se que $En(\pi)$ é uma função monótona crescente com n ;

Ax 3: Sendo $0 \leq \lambda \leq 1$,

$$En[\pi_1, \pi_2, \ldots, \pi_{n-1}, \lambda\pi_n, (\lambda-1)\pi_n] =$$
$$= En(\pi_1, \pi_2, \ldots, \pi_{n-1}, \pi_n) + \pi_n En(\lambda, 1-\lambda).$$

Shannon mostrou que $En(\pi)$ deve então tomar necessariamente a forma

$$(3.2) \qquad En(\pi) = -\sum_i \pi_i \ln \pi_i$$

Khinchin (1957), Feinstein (1958), Hobson (1971) e outros apresentaram sistemas de axiomas equivalentes ao de Shannon. Por outro lado, outras funcionais de informação podem ser obtidas consoante se relaxe o axioma 3 em (3.8) de várias formas. Não parecem, contudo, possuir uma relevância semelhante à da entropia de Shannon.

Supondo ainda a unidade de massa distribuída num número finito de pontos $\{\theta_1, \theta_2, \ldots, \theta_n\}$, pode ver-se com facilidade que $En(\pi)$ assume valores entre um mínimo igual a zero, no

caso de se ter $\pi(\theta_k) = 1$ e $\pi(\theta_i) = 0$ com $i \neq k$, i.e., no caso em que a unidade de massa está concentrada num só ponto, no caso o ponto θ_k e um máximo igual a ln n, no caso de se ter $\pi(\theta_i) = 1/n$, i.e., no caso de equiprobabilidade. Esta última distribuição é, portanto, de máxima entropia, quando se considera como única restrição a condição de normalização

$$\sum_{i=1}^{n} \pi(\theta_i) = 1 \quad . \tag{3.9}$$

Tendo em conta que a distribuição uniforme é aquela que o princípio da razão insuficiente indica para o caso de um número finito de possibilidades, tal poderia parecer indicar uma "dedução" de tal princípio. Note-se, no entanto, que em qualquer sistema de axiomas que leve univocamente à medida logarítmica (3.2), está já implícito que dê origem por maximização à medida equiprovável.

Pode ser elucidativa a consideração das probabilidades $\{\pi_i\}$ com i=1, 2, ..., n finito, como coordenadas de um ponto P num espaço n-dimensional. As condições $\pi_i \geq 0$ e $\sum_i \pi_i = 1$ obrigam o ponto P a estar num sector triangular D, pertencente ao hiperplano (n-1) dimensional $\sum_i \pi_i = 1$. Sobre este *simplex*, $En(\pi)$ é uma função contínua das coordenadas $\{\pi_i\}$,

que assume o valor máximo igual a ln n no ponto $\pi_i = 1/n$ (i=1, ..., n), i.e., no centro geométrico do *simplex*, e o valor mínimo igual a zero sobre qualquer um dos n vértices do *simplex*, pontos onde $\pi_k=1$ e $\pi_i=0$ se $i \neq k$.

Qualquer informação de tipo testável restringe P a uma sub-região D' do *simplex*. Se D' for fechado, sendo $En(\pi)$ limitada superiormente em D (uma vez que $En(\pi) \leq \ln n$), deve portanto ter máximo, não necessariamente único. Caso D' seja aberto, qualquer ponto maximizante que pertença ao fecho de D' pode ser aceite como solução do problema.

3.4. Alguns exemplos. Vai inicialmente considerar-se como exemplo de aplicação do MAXENT o seguinte problema: deseja-se determinar as probabilidades discretas, $\pi(\theta_i)$, que, satisfazendo restrições do tipo,

$$(3.10) \quad P(\Theta \in Q_j) = \sum_{\{\theta_i \in Q_j\}} \pi(\theta_i) = q_j \qquad (j=1, 2, \ldots, m),$$

maximizam $En(\pi)$. Pela utilização de métodos Lagrangianos, é-se levado à solução,

$$(3.11) \quad \pi(\theta_i) = \prod_{\{j:\theta_i \in Q_j\}} \lambda_j ,$$

onde λ_j (j=1, 2, ..., m) são constantes satisfazendo,

$$(3.12) \quad \sum_{\{i:\theta_i \in Q_j\}} \prod_{\{k:\theta_i \in Q_k\}} \lambda_k = q_j .$$

Assim, o método de solução pede que se associe uma incógnita λ_j com cada agregado (acontecimento) Q_j. Seguidamente, exprime-se cada probabilidade $\pi(\theta_i)$ como o produto dos λ_j correspondentes aos agregados que contenham o ponto θ_i. Substituindo estas expressões nas equações que definem os λ_j obtém-se um sistema de m equações a m incógnitas λ_j. Na posse da solução do sistema, exprimem-se os $\pi(\theta_i)$ como produtos dos λ_j entretanto obtidos.

Em muitos casos de interesse conseguem-se soluções explícitas para as equações (3.12). Quando tal se revela impossível, podem utilizar-se métodos numéricos. Caso exista uma solução MAXENT única, o método de Newton-Raphson convergirá garantidamente para essa solução.

Considere-se agora o caso em que a informação *a priori* de tipo testável sobre Θ assume a forma,

$$E[g_k(\Theta)] = \sum_{i=1}^{n} g_k(\theta_i)\ \pi(\theta_i) = \mu_k \,, \tag{3.13}$$

com, k = 1, 2, ..., m, sendo os valores, μ_k, conhecidos para um dado problema. A procura da distribuição $\pi(\theta)$ que maximiza $En(\pi)$, sujeita a (3.13) e também, é claro, à condição de normalização, $\sum_{i=1}^{n} \pi(\theta_i) = 1$, é problema conhecido do cálculo das variações, resolúvel pela técnica dos

multiplicadores de Lagrange. Tem a seguinte solução:

$$(3.14) \qquad \bar{\pi}(\theta_i) = \frac{\exp\left\{\sum_{k=1}^{m} -\lambda_k\, g_k(\theta_i)\right\}}{\sum_{i=1}^{n} \exp\left\{\sum_{k=1}^{m} -\lambda_k\, g_k(\theta_i)\right\}} .$$

O denominador em (3.14),

$$(3.15) \qquad Z(\lambda_1, \lambda_2, \ldots, \lambda_m) = \sum_{i=1}^{n} \exp\left\{\sum_{k=1}^{m} -\lambda_k\, g_k(\theta_i)\right\},$$

tem o nome de função de partição. Os multiplicadores de Lagrange $\{\lambda_k\}$, são escolhidos de forma a satisfazer as restrições (3.13). Para que tal aconteça, deve verificar-se,

$$(3.16) \qquad \mu_k = -\frac{\partial}{\partial \lambda_k} \ln Z = -\frac{1}{Z}\frac{\partial Z}{\partial \lambda_k} =$$

$$= \frac{\sum_{i=1}^{n} \exp\left\{\sum_{k=1}^{m} -\lambda_k\, g_k(\theta_i)\right\} . g_k(\theta_i)}{Z} ,$$

com k = 1, 2,..., m. Em (3.16) tem-se um sistema de m equações nas m incógnitas λ_k. O valor máximo obtido para a entropia é função dos μ_k,

$$(3.17) \qquad S(\mu_1, \mu_2, \ldots, \mu_m) = \ln Z + \sum_{k=1}^{m} \lambda_k\, \mu_k .$$

Note-se que com o conhecimento da função, $S(\mu_1, \mu_2, \ldots, \mu_m)$, se pode obter,

$$(3.18) \qquad \lambda_k = \frac{\partial S}{\partial \mu_k} .$$

3.5. Uma interpretação combinatória. A solução (3.14) pode ter uma interpretação combinatória na linha dos resultados obtidos por Boltzmann (1877) (cf. Cap. 1, pp.7--10). Assim, suponha-se que uma dada variável é observada N vezes, sendo obtido em cada uma das repetições um de n símbolos θ_i. No final, o resultado θ_i registou uma frequência absoluta N_i. Cada vector de frequências relativas $F = (f_1, f_2, \ldots, f_n)$, onde $f_i = N_i/N$ tem multiplicidade W(F), pode obter-se a partir de um número de possibilidades,

$$W(F) = \frac{N!}{(Nf_1)!\ (Nf_2)!\ \ldots\ (Nf_n)!} \quad . \tag{3.19}$$

Quando $N \to \infty$, tem-se, utilizando aproximação assimptótica de Stirling,

$$\frac{1}{N} \ln W(F) \longrightarrow - \sum_{i=1}^{n} f_i \ln f_i = H(f) \ . \tag{3.20}$$

Em (3.20) encontra-se novamente a entropia de Shannon, calculada agora com base em frequências relativas f_i. Sabendo suplementarmente que as frequências ponderadas de m quantidades $g_k(\theta_i)$ assumiram os valores μ_k, ou seja,

$$\sum_{i=1}^{n} g_k(\theta_i)\, f_i = \mu_k \ , \tag{3.21}$$

a solução de máxima entropia será aquela com multiplicidade (3.19) máxima, sujeita às restrições (3.21). O problema é formalmente equivalente àquele que foi apresentado

anteriormente em (3.13), e terá uma solução correspondente a (3.14), onde figuram agora os f_i em substituição dos $\pi(\theta_i)$.

A distribuição de probabilidade de máxima entropia será assim aquela que pode ser realizada a partir de um número máximo de formas consistentes com a informação testável tomada em consideração nas restrições.

Raciocinando em termos de frequências observadas, seja, $f' = (f'_1, f'_2, \ldots, f'_n)$, vector distinto de frequências relativas que satisfaçam as m condições (3.21). Tem-se, é claro, que $H(f')$ definida de acordo com (3.20) deve ser menor que $H(f)$. Considerada a razão entre as multiplicidades das distribuições $\{f_i\}$ e $\{f'_i\}$ de acordo com (3.19), tem-se o resultado assimptótico,

$$(3.22) \qquad \frac{W}{W'} \sim A \exp N \left[H(f) - H(f')\right] . \left[1 + \frac{B}{N} + O(N^{-2})\right],$$

que evidencia o crescimento exponencial da razão (3.22) em função de N.

Como ilustração do formalismo apresentado, vai considerar-se a solução de um exemplo simples.

Pretende obter-se a distribuição de máxima entropia da variável aleatória - número de pontos obtidos no lançamento de um dado -, quando se tem o conhecimento de que a média de tal distribuição é igual a 4,5 e não 3,5, como seria de esperar no caso de o dado ser equilibrado. Tem-se assim de (3.13),

$$\sum_{i=1}^{6} i\,\pi_i = 4.5 , \tag{3.23}$$

e de (3.14),

$$\bar{\pi}_i = \frac{e^{-\lambda i}}{\sum_{i=1}^{6} e^{-\lambda i}} \quad (i = 1, \ldots, 6) . \tag{3.24}$$

A função de partição (3.15) é

$$Z(\lambda) = \sum_{i=1}^{6} e^{-\lambda i} = \frac{e^{-\lambda}(1 - e^{-6\lambda})}{1 - e^{-\lambda}} . \tag{3.25}$$

Para que a restrição (3.23) se verifique, é necessário que, de acordo com (3.16), seja,

$$-\frac{\partial}{\partial\lambda} \ln Z = \frac{1 - 7e^{-6\lambda} + 6e^{-7\lambda}}{(1 - e^{-\lambda})(1 - e^{-6\lambda})} = 4,5 . \tag{3.26}$$

Tal conduz à equação algébrica de grau 7 em $e^{-\lambda}$,

$$3e^{-7\lambda} - 5e^{-6\lambda} + 9e^{-\lambda} - 7 = 0, \tag{3.27}$$

sendo a raíz numérica que é solução do problema, $e^{-\lambda} = 1,44925$, donde vem imediatamente $\lambda = -0,37105$, $Z = 26,66365$, $\ln Z = 3,28330$. Assim, a distribuição MAXENT sujeita à condição (3.23) é dada pelo conjunto de seis probabilidades

$$\{\bar{\pi}_1, \ldots, \bar{\pi}_6\} = \{0,05435;\ 0,07877;\ 0,11416;\ 0,16545;\ 0,23977;\ 0,34749\}, \tag{3.28}$$

com entropia $En(\pi) = S(4,5) = 1,61358$, obviamente menor do

que a entropia da distribuição uniforme que seria obtida se não fosse tomada em conta a informação (3.23). Neste último caso a entropia seria evidentemente S = ln 6 = 1,79176. Note-se ainda que, caso a restrição sobre a média fosse,

$$\sum_{i=1}^{6} i\ \pi_i = 3{,}5\ , \tag{3.29}$$

a distribuição MAXENT seria necessariamente a distribuição uniforme, i.e., $\bar{\pi}_i = 1/6$ para i = 1,...,6, uma vez que (3.29) é compatível com esta última distribuição, a qual é, como vimos, a distribuição de probabilidade de máxima entropia na ausência de informação complementar. Geometricamente, a informação (3.29) restringe P a uma sub-região D', a qual contém o ponto $\pi_i = 1/n$ para i = 1,...,n, com n = 6. Por outro lado, tal não acontece com a informação (3.23), que restringe P a uma outra região D". Esta já não contém o ponto central do *simplex* $\pi_i = 1/6$ para i = 1, 2, ..., 6, o qual, deixa então de ser solução possível do problema.

Faça-se agora uma perturbação na distribuição (3.21), por forma a que a nova distribuição continue a respeitar (3.23), por exemplo, passando uma massa δ da abcissa i=5 para i=4, compensada em termos de média da distribuição com a passagem de uma massa $\delta/3$ da abcissa i=3 para i=6. Supondo $\delta = 0{,}03$, obtém-se a nova distribuição,

(3.30) $\{\pi_1', \ldots, \pi_6'\} = \{\bar{\pi}_1, \bar{\pi}_2, \bar{\pi}_3 - \frac{\delta}{3}, \bar{\pi}_4 + \delta, \bar{\pi}_5 - \delta, , \bar{\pi}_6 + \frac{\delta}{3}\} = \{0,05435; 0,07877; 0,10416; 0,19545; 0,20977; 0,35749\}$.

À nova distribuição $\{\pi_i'\}$ apresentada em (3.30) corresponde, obviamente, uma entropia menor que a entropia da distribuição MAXENT $\{\bar{\pi}_i'\}$ em (3.28):

(3.31) $En(\pi') = 1,60845 < En(\bar{\pi}) = 1,61358.$

Passando a raciocinar em termos de frequências observadas, calcule-se, tal como em (3.22) a razão assimptótica entre as multiplicidades definidas de acordo com (3.19). As probabilidades $\{\bar{\pi}_i\}$ e $\{\pi_i'\}$ devem, é claro, ser agora interpretadas como frequências relativas observadas na repetição por N vezes da experiência aleatória. Tem-se de (3.22) a razão,

(3.32) $\frac{W}{W'} \sim A \exp\{N\,[En(\bar{\pi}) - En(\pi')]\} \cdot \{1 + \frac{B}{N} + O(N^{-2})\}$,

onde,

(3.33) $A = \prod_{i=1}^{6} \left(\frac{\pi_i'}{\bar{\pi}_i}\right)^{1/2} = 0,98496$,

(3.34) $B = \frac{1}{12} \sum_{i=1}^{6} \left(\frac{1}{\pi_i'} - \frac{1}{\bar{\pi}_i}\right) = 0,03575$,

(3.35) $En(\bar{\pi}) - En(\pi') = 0,00513$,

são independentes de N e representam correcções dos termos de

ordem superior na aproximação de Stirling. Na análise de (3.32) vê-se facilmente que a parte essencial do comportamento assimptótico de $\frac{W}{W'}$ é dada por,

(3.36) $$\exp\left\{N \cdot 0{,}00513\right\}.$$

Se, por exemplo, se fizer N = 10.000, $\frac{W}{W'}$ será um número da ordem de grandeza de 10^{21}. Tal pode servir como justificação heurística para a adopção de uma distribuição MAXENT. Com efeito, tomadas duas distribuições de frequências relativamente próximas e que verificam a mesma informação testável (3.23), vê-se como a distribuição MAXENT (3.28) dispõe de uma multiplicidade relativa à da outra distribuição, multiplicidade que cresce exponencialmente de acordo com (3.32). É essa a razão da ubiquidade de tais distribuições em Mecânica Estatística, onde, como se sabe, N, é de elevada ordem de grandeza, e, como tal, faz com que os factores combinatórios se concentrem de tal forma no ponto de máxima entropia, que fica praticamente excluída a verificação de qualquer outra distribuição, para um dado conjunto de restrições físicas verificadas.

Quando a distribuição de probabilidade tem um conjunto discreto mas infinito de pontos de massa, deixa de existir um limite superior finito para $En(\pi)$ [(cf. (3.16)]. Em alguns casos a função de partição $Z'(\lambda_1, \ldots, \lambda_m)$ pode divergir para

qualquer λ_i real, ou podem as restrições (3.16) não ter qualquer solução. Tal será indicativo da impossibilidade de obter a distribuição MAXENT devido à insuficiência de informação nas restrições consideradas.

3.6. Entropia de uma distribuição contínua. Surge um problema de tipo diferente quando se considera uma distribuição contínua. Note-se, de acordo com (3.2), que a entropia de uma distribuição discreta é essencialmente um número associado a uma partição, formando então os acontecimentos elementares $\{\Theta=\theta_i\}$ uma partição natural do espaço do parâmetro Θ. Por outro lado, quando Θ é contínuo, os pontos θ, não sendo contáveis, não formam uma partição.

Uma das vias para a definição de entropia será a consideração de uma distribuição discretizada, obtida pelo arredondamento de θ, por exemplo na forma,

(3.37) $$\theta_\delta = n\delta \qquad \text{se} \quad (n-1)\delta \leq \theta \leq n\delta.$$

Tem-se então,

(3.38) $$P(\Theta_\delta = n\delta) = P[(n-1)\delta < \Theta \leq n\delta] = \int_{(n-1)\delta}^{n\delta} \pi(\theta)\, d\theta = \delta\pi(\bar{\theta}),$$

onde $\pi(\bar{\theta})$ é um número entre o máximo e o mínimo assumidos por $\pi(\theta)$ sobre o intervalo $](n-1)\delta;\ n\delta[$. Aplicando (3.2) à distribuição discretizada, com a função de probabilidade dada

em (3.38), obtém-se:

$$(3.39) \qquad En(\pi_\delta) = - \sum_i \delta \cdot \pi(\bar{\theta}_i) \ln [\delta \cdot \pi(\bar{\theta}_i)].$$

Tendo em conta,

$$(3.40) \qquad \sum_i \delta \cdot \pi(\bar{\theta}_i) = \int_{-\infty}^{+\infty} \pi(\theta)\, d\theta = 1,$$

tem-se de (3.39),

$$(3.41) \qquad En(\pi_\delta) = - \ln \delta - \sum_i \delta \cdot \pi(\bar{\theta}_i) \ln \pi(\bar{\theta}_i).$$

Quando $\delta \to 0$, a distribuição discretizada definida em (3.37), tende para uma distribuição contínua com densidade $\pi(\theta)$. Ao mesmo tempo, de (3.41) vem,

$$(3.42) \qquad \lim_{\delta \to 0} En(\pi_\delta) = - \lim_{\delta \to 0} \ln \delta - \int_{-\infty}^{+\infty} \pi(\theta) \ln \pi(\theta)\, d\theta.$$

Vê-se assim, da primeira parcela do segundo membro de (3.42), que $En(\pi_\delta)$ tende para ∞ quando $\delta \to 0$.

Tomando a segunda parcela do segundo membro de (3.42),

$$(3.43) \quad En^*(\pi) = - \int_{-\infty}^{+\infty} \pi(\theta) \ln \pi(\theta)\, d\theta = \lim_{\delta \to 0} [En(\pi_\delta) + \ln \delta],$$

alguns autores definem então o que chamam de entropia reduzida (Ventsel, 1973), ou simplesmente entropia

[(Papouilis, 1984); (Harris, 1982)]. Note-se, contudo, que (3.43) não é invariante para reparametrizações. Assim, por exemplo, sendo

(3.44) $\pi(\theta) = e^{-\theta}$ com $\theta > 0$,

fazendo a reparametrização

(3.45) $\Psi = c\theta$ com $c > 0$,

resultam de (3.42),

(3.46) $En^*[\pi(\theta)] = - \int_0^{+\infty} e^{-\theta} \ln e^{-\theta} \, d\theta = 1$,

(3.47) $En^*[\pi'(\Psi)] = - \int_0^{+\infty} \frac{1}{c} e^{-\frac{\Psi}{c}} \ln \left(\frac{1}{c} e^{-\frac{\Psi}{c}} \right) d\Psi =$

$= 1 + \ln c$.

O valor obtido em (3.47) depende, evidentemente, da reparametrização considerada em (3.45).

Jaynes (1968, p. 95; 1979, pp. 86-87) propõe que não se tome $En^*(\pi)$ como medida de informação de uma distribuição contínua. Voltando a (3.2) e passando ao limite, ele obteve,

(3.48) $En_c(\pi) = - \int_{-\infty}^{+\infty} \pi(\theta) \ln \frac{\pi(\theta)}{\pi_0(\theta)} \, d\theta$,

onde $\pi_0(\theta)$ é uma função de medida invariante, proporcional à densidade limite de pontos discretos (cf. Jaynes, 1968, p. 95). Uma vez que $\pi(\theta)$ e $\pi_0(\theta)$ se transformam da mesma maneira para uma mudança de variáveis, $En_c(\pi)$ será invariante (cf. Kullback, 1959, pp.18-22). Note-se que a expressão (3.48), é

o simétrico da informação de Kullback - Leibler ou ganho de informação de Rényi, para duas distribuições contínuas com densidades $\pi(\theta)$ e $\pi_0(\theta)$ e que é definido, por analogia com (3.5), na forma,

$$(3.49) \qquad I(1,2) = \int_{-\infty}^{+\infty} \pi_1(\theta) \ln \frac{\pi_1(\theta)}{\pi_2(\theta)} \, d\theta \, .$$

No entanto, permanece a seguinte dificuldade prática: não sendo o espaço do parâmetro o resultado de um processo de passagem ao limite óbvio, o que determina a correspondente função de medida $\pi_0(\theta)$? Como justificação heurística para a escolha que se irá efectuar, pode considerar-se o problema de maximizar (3.48) sem qualquer restrição adicional. A solução será tomar $\pi(\theta) = A.\pi_0(\theta)$, onde A é uma constante de normalização. Assim, $\pi_0(\theta)$ deverá representar a ignorância completa sobre o parâmetro Θ, mostrando, por outro lado, que o problema da maximização da entropia, quando o espaço do parâmetro é contínuo, reconduz à velha questão de Bernouilli, Bayes e Laplace da representação numérica do estado de ignorância completa. Note-se que, de acordo com (3.49), (3.48) representa o ganho de informação de Rényi entre $\pi(\theta)$ e a distribuição *a priori* não informativa $\pi_0(\theta)$, reforçando a boa justificação heurística para a definição da entropia $En_c(\pi)$ como medida de informação. No entanto, como já se referiu, o

problema da procura de uma distribuição *a priori* não informativa num dado contexto é dos mais controversos e avessos a solução de consenso. Mesmo quando não se questiona *ab initio* o bom fundamento de tal problema e se entra directamente na questão da procura numérica de soluções concretas, as dificuldades subsistem. Ironicamente, um dos aspectos mais preocupantes não é tanto a falta de soluções em dados contextos, como antes a pluralidade de soluções distintas para o mesmo problema quando se utilizam técnicas diversas. Por outro lado, é igualmente notável que em alguns casos as mesmas técnicas coincidam na distribuição *a priori* proposta.

Quando o espaço do parâmetro é conjunto finito com n elementos, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, a escolha óbvia para a distribuição *a priori* não informativa é dada pela medida discreta de equiprobabilidade com, $\pi(\theta_i) = 1/n$, para qualquer, $i = 1, \ldots, n$. Tal é, aliás, a solução clássica sugerida pelo princípio da razão insuficiente.

No caso do espaço do parâmetro ser contínuo e de amplitude finita, $\Theta = \{\theta: a < \theta < b\}$, a correspondente utilização de uma medida de equiprobabilidade leva à consideração da densidade $\pi_0(\theta) = 1/(b-a)$, para $a < \theta < b$. Quando se considera como espaço do parâmetro toda a recta

real, $\Theta = R^1$, uma generalização óbvia da medida uniforme levará a considerar $\pi_0(\theta) = c$. Note-se, porém, que tal medida é imprópria, no sentido em que atribui massa infinita a todo o espaço Θ. Com efeito, verificando-se,

$$(3.50) \qquad \int_{-\infty}^{+\infty} \pi_0(\theta)\, d\theta = c \int_{-\infty}^{+\infty} d\theta = +\infty ,$$

e havendo divergência no integral, é habitual fazer a constante c igual à unidade.

No entanto, a impropriedade da medida uniforme sobre R^1 não impede que pela aplicação do teorema de Bayes se obtenha uma distribuição *a posteriori*, $\pi(\theta|x)$, própria. Tal justificaria formalmente o uso de medidas uniformes em uma ou mais dimensões como distribuições *a priori* de ignorância no mecanismo de inferência Bayesiana. Foi referido o seu uso por Laplace e a severa crítica que a prática indiscriminada de tal metodologia suscitou.

Fisher (1956) e outros autores assinalaram outro problema no facto de que o uso da medida uniforme em diferentes parametrizações de um mesmo problema conduz imediatamente a inconsistências. Com efeito, considerando um outro parâmetro $\eta = \exp \theta$, a nova densidade $\pi^*(\eta)$ será

$$(3.51) \qquad \pi^*(\eta) = \eta^{-1}\, \pi \ln(\eta).$$

Assim, $\pi(\theta) \propto 1$ implica que tenha de ser $\pi^*(\eta) \propto \frac{1}{\eta}$. Tal

mostra a impossibilidade de manter consistentemente a uniformidade de uma distribuição *a priori* para diferentes reparametrizações, quando o espaço do parâmetro é contínuo.

3.7. Distribuições *a priori* invariantes. É com base na constatação anterior que se desenvolveram esforços na procura de distribuições *a priori* não informativas que gozem da propriedade de invariância para certas transformações. Como já foi referido, Jeffreys (1948, 1961), Hartigan (1964) e Jaynes (1968) deram importantes passos nesta via.

Para análise mais formal da questão da exigência de invariância em distribuições *a priori*, considere-se um modelo, $\mathcal{M} = (X, \Theta, \mathcal{P})$, em que X é observável, Θ é um parâmetro (escalar ou vector) e $\mathcal{P}$ é uma família parametrizada, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, onde P_θ representa a distribuição do observável X. À família parametrizada $\mathcal{P}$ chamar-se-á modelo de distribuição associado com $\mathcal{M}$. Pretende-se fazer corresponder a cada modelo $\mathcal{M}$ uma distribuição *a priori* representando ignorância (i. e., não informativa) sobre o parâmetro e que se designa por $\Pi_{\mathcal{M}}$. Admite-se a possibilidade de que a ignorância não possa ser representável por uma distribuição de probabilidade própria, exigindo-se unicamente que a medida considerada seja σ-finita, podendo atribuir uma

massa infinita a todo o espaço do parâmetro.

Podem agora exigir-se várias condições que proporcionem a atribuição de distribuições *a priori* $\Pi_{\mathcal{M}}$ a $\mathcal{M}$ de forma invariante. Consideram-se inicialmente as seguintes três regras:

1) Invariância no parâmetro (PI): Seja $\mathcal{M} = (X, \Theta, \mathcal{P})$ e seja $\Phi = \phi(\Theta)$ uma transformação invertível. O modelo $\mathcal{M}_1 = (X, \Phi, \mathcal{P})$ difere de $\mathcal{M}$ unicamente na sua parametrização. Deve então exigir-se que seja,

$$\Pi_{\mathcal{M}_1}(\Phi \in A) = \Pi_{\mathcal{M}_1}(\Theta \in \phi^{-1}(A)) , \tag{3.52}$$

ou, supondo a existência de densidades, que se verifique,

$$\pi_{\mathcal{M}_1}(\varnothing) = \pi_{\mathcal{M}}(\theta) . |J(\theta)|^{-1} , \tag{3.53}$$

onde $J(\theta)$ é o Jacobiano $\det\left(\frac{\partial \phi(\theta)}{\partial \theta}\right)$.

2) Invariância nos dados (DI): Sendo $Y = y(X)$ uma transformação de X, seja $\mathcal{M}_2 = (Y, \Theta, Q)$, modelo induzido para o observável Y, ainda parametrizado por Θ. É claro que a família parametrizada de $\mathcal{M}_2$, Q, é já possivelmente distinta da família $\mathcal{P}$ de $\mathcal{M}$. Será agora naturalmente exigível que seja,

$$\Pi_{\mathcal{M}_2}(\Theta \in A) = \Pi_{\mathcal{M}}(\Theta \in A). \tag{3.54}$$

Antes de considerar um terceiro tipo de invariância, note-

-se que a exigência de (PI) e (DI) releva do contexto habitual do cálculo das probabilidades, sendo feita igualmente em relação a distribuições informativas, e. g., no caso de distribuições *a priori* subjectivas. Já característica do contexto da ignorância é a seguinte exigência:

3) Invariância de contexto (CI): Sendo $\mathcal{M} = (X, \Theta, \mathcal{P})$ e $\mathcal{M}' (X', \Theta', \mathcal{P})$ dois modelos distintos partilhando da mesma família parametrizada $\mathcal{P}$, exige-se a seguinte igualdade:

(3.55) $$\Pi_{\mathcal{M}} (\Theta \in A) = \Pi_{\mathcal{M}'} (\Theta' \in A),$$

ou seja, que se tome em conta unicamente o modelo de distribuição, abstraindo-se de quaisquer outros aspectos de estrutura ou significado. Note-se que a exigência de (CI) possibilita que se escreva $\Pi_{\mathcal{P}}$ em vez de $\Pi_{\mathcal{M}}$.

Jeffreys (1961) propôs como distribuição *a priori* de ignorância aquela que tem a seguinte densidade:

(3.56) $$\pi_{\mathcal{P}} (\theta) = | I(\theta) |^{1/2},$$

onde, $I(\theta) = E\left[\frac{\partial \ln f(X|\theta)}{\partial \theta_i} \frac{\partial \ln f(X|\theta)}{\partial \theta_j}\right] = -E\left[\frac{\partial^2 \ln f(X|\theta)}{\partial \theta_i \partial \theta_j}\right]$

é a matriz de informação de Fisher. A distribuição *a priori* de Jeffreys (3.56) goza em simultâneo das três características de invariância (PI), (DI) e (CI), além de outras que podem ser

suplementarmente exigidas (cf. Hartigan, 1964, pp. 837-838). Entre estas últimas tem relevo especial o facto de a distribuição de Jeffreys não ser afectada por restrições feitas no espaço do parâmetro. Esse relevo resulta de que as distribuições *a priori* não informativas têm precisamente nos problemas com espaço do parâmetro restrito um dos domínios preferenciais de aplicação.

Zellner (1971), dentro do espírito da teoria da Informação, propôs como medida de informação em $f(x|\theta)$ a expressão,

$$I_x(\theta) = \int f(x|\theta) \ln f(x|\theta)\, dx , \tag{3.57}$$

cujo valor médio em θ calculou, obtendo a informação *a priori* média, na forma,

$$\bar{I}_x = \int I_x(\theta)\, \pi(\theta)\, d\theta . \tag{3.58}$$

Após ter definido a medida de ganho de informação G na forma,

$$G = \bar{I}_x - \int \pi(\theta) \ln \pi(\theta)\, d\theta , \tag{3.59}$$

considerou como sendo uma distribuição *a priori* de informação mínima aquela que maximiza G para uma dada função $f(x|\theta)$.

A distribuição *a priori* obtida por maximização de G pode coincidir ou não com a distribuição de Jeffreys (3.56). Assim, na estimação de $\theta = (\mu, \sigma)$ em universos normais, a regra de Jeffreys (3.56) obtém $\pi(\theta) \propto 1/\sigma^2$, enquanto que o método de

Zellner dá por maximização de G, $\pi(\theta) \propto 1/\sigma$. No entanto, o próprio Jeffreys desaconselhou neste problema o uso de $1/\sigma^2$ em favor de precisamente, $1/\sigma$, que obteve, assumindo independência. Jeffreys, aliás, defendeu sempre um uso crítico da sua regra, reconhecendo que esta por vezes propunha distribuições *a priori* inaceitáveis, em particular na situação de se ter estimação simultânea de parâmetros de localização e de escala, e. g. no caso da estimação simultânea do vector das médias e de σ^2 de uma normal multidimensional com hipótese de homoscedasticidade.

Uma outra via para a justificação da distribuição *a priori* de Jeffreys (cf. Box e Tiao, 1973, pp. 53-54) pode ser a de exigir que numa dada reparametrização $\Phi = \phi(\Theta)$, a matriz de informação de Fisher, $I(\phi)$, se revele independente de ϕ, o que, com base no resultado assimptótico que permite aproximar a função de verosimilhança por uma normal multidimensional, pode ser tomado como indicando o carácter não informativo de tal parâmetro Φ. Em seguida, resulta imediatamente das propriedades de transformação da matriz de informação de Fisher para uma dada reparametrização e da exigência de (PI) que a densidade *a priori* não informativa no parâmetro θ deve ter precisamente a forma $\pi(\theta) \propto |I(\theta)|^{1/2}$.

Considere-se agora uma transformação invertível dos dados $Y = g_0X$, com a propriedade de que, sempre que X tem distribuição em $\mathcal{P}$, o mesmo acontece com Y e reciprocamente. Tal transformação induz uma reparametrização de Θ, $\Phi = \bar{g}_0\Theta$, de tal forma que X tem distribuição P_θ, se e só se g_0X tem distribuição $P_{\bar{g}_0\theta}$. Note-se que o modelo $\mathcal{M}' = (Y, \Phi, \mathcal{P})$ tem exatamente a mesma distribuição que o modelo $\mathcal{M}$. Em tal caso pode designar-se g como uma transformação equivariante de X, tal como $\bar{g}$ é uma transformação equivariante de Θ, sendo então $\mathcal{M}$ ou $\mathcal{P}$ equivariantes para g e $\bar{g}$.

Os conjuntos de todas as transformações equivariantes de X e de Θ têm estrutura de grupo. Serão designados respectivamente por $\mathcal{G}$ e $\bar{\mathcal{G}}$. Sendo $\Phi = \bar{g}_0\Theta$, com $\bar{g} \in \bar{\mathcal{G}}$, deriva de (DI) e (PI) que se obtenha,

(3.60) $$\Pi_{\mathcal{M}'}(\Phi \in \bar{g}_0A) = \Pi_{\mathcal{M}}(\Theta \in A).$$

Por outro lado, exigindo acessoriamente (CI), conclui-se,

(3.61) $$\Pi_{\mathcal{M}'}(\Phi \in \bar{g}_0A) = \Pi_{\mathcal{M}}(\Theta \in \bar{g}_0A).$$

Assim, tem-se da exigência conjunta de (PI), (DI) e (CI) que $\Pi_{\mathcal{M}}$ (= $\Pi_{\mathcal{P}}$)deve ser invariante em $\bar{\mathcal{G}}$, verificando-se para qualquer $\bar{g} \in \bar{\mathcal{G}}$,

(3.62) $$\Pi_{\mathcal{P}}(\bar{g}_0A) = \Pi_{\mathcal{P}}(A).$$

Se o grupo $\bar{\mathcal{G}}$ for transitivo, i. e., se, para quaisquer θ_1 e

θ_2 pertencentes a Θ, existir $\bar{g} \in \bar{G}$ tal que $\theta_2 = \bar{g}o\theta_1$, a condição de invariância sob $\bar{G}$ determinará Π_p univocamente a uma constante multiplicativa, e, como tal, coincidindo com a distribuição *a priori* de Jeffreys, que, como se viu, satisfaz simultaneamente (PI), (DI) e (CI). No entanto, muitas vezes não haverá transitividade no grupo $\bar{G}$, verificando-se então a existência de uma pluralidade de distribuições invariantes. Brillinger (1963) estudou as condições necessárias e suficientes de invariância para grupos de transformação.

Suponha-se, por exemplo, que a densidade de X tem a forma $f(x-\theta)$. Ao depender unicamente de $(x-\theta)$, a densidade diz-se de localização, sendo Θ o parâmetro de localização. Considere-se um novo observável $Y = goX = X+c$. Definindo um novo parâmetro $\Phi = \bar{g}o\Theta = \Theta+c$, obtém-se para Y uma densidade de localização na forma $f(y-\phi)$.

A aplicação de (3.60) - (3.62) permite obter,

(3.63) $$\pi(\theta) = \pi(\theta-c),$$

para qualquer c, ou em particular, fazendo $c = \theta$,

(3.64) $$\pi(c) = \pi(0).$$

Tendo em conta a arbitrariedade de c, tem-se imediatamente que $\pi(\theta)$ deve ser constante. Princípios semelhantes de equivariância permitem obter para a densidade de escala

(unidimensional) da forma $(1/\sigma)\, f(x/\sigma)$, com o parâmetro de escala, σ, positivo, a densidade *a priori* não informativa $\pi(\sigma) = 1/\sigma$, também proposta por Jeffreys.

Devido à sua impropriedade, as densidades obtidas não são únicas. Nesse sentido, em vez de se exigir, por exemplo, (3.62), deve antes pedir-se,

(3.65) $$\pi(\theta) = h(c) \,.\, \pi(\theta - c).$$

Fazendo $\theta = c$ em (3.65) calcula-se $h(c) = \pi(\theta) \,/\, \pi(0)$. Substituindo h(c) nessa mesma expressão obtém-se,

(3.66) $$\pi(\theta - c) = \frac{\pi(0) \,.\, \pi(\theta)}{\pi(c)} \,.$$

Muitas outras densidades impróprias, para além da uniforme, satisfazem (3.66). Veja-se, por exemplo, a função $\pi(\theta) = \exp(\theta' z)$, onde z é um vector fixo. Uma densidade *a priori* que satisfaça (3.65) ou, equivalentemente, (3.66), diz-se ser uma invariante relativa de localização.

3.8. O método da marginalização. Um outro método que parece conter algumas potencialidades na procura da distribuições *a priori* não informativas poderá ser o método da marginalização. Se bem que Jeffreys (1939; § 3.8) tenha já antecipado alguns aspectos desta teoria, só nos anos 70 se parece ter reconhecido as virtualidades de tal método para a construção, por uma via independente, das distribuições *a*

priori de ignorância.

A origem deste método deriva de uma solução proposta por Jaynes para o chamado paradoxo da marginalização, surgido num artigo de 1973 da autoria conjunta de Dawid, Stone e Zidek. Este paradoxo parecia pôr a claro as inconsistências resultantes do uso de distribuições *a priori* impróprias em inferência Bayesiana. Uma linha de explicações para o paradoxo poderia residir na falta de aplicabilidade, na presença de medidas de probabilidade impróprias, do teorema de Fubini, o qual permite a permuta da ordem de integração com invariância do resultado final.

Dawid, Stone e Zidek propunham então a exploração da estrutura de grupo da distribuição amostrada e a escolha para distribuição *a priori* da medida invariante de Haar.

Jaynes encara o problema de maneira diferente ao identificar a origem do paradoxo, que não considera verdadeiro, na deficiente análise Bayesiana de um dos ângulos do problema. Por outro lado, reconhece que a distribuição *a priori* que "evita o paradoxo" tem a possibilidade de ser interpretada como tendo a característica de ser completamente não informativa para o problema[4]. Vai agora apresentar-se mais

4 "The marginalization problem is then turn to advantage by showing

de perto a análise feita por Jaynes da questão.

Suponha-se o observável X decomposto na forma $X = (Y, Z)$ e o parâmetro Θ na forma $\Theta = (\gamma, \eta)$. Deseja fazer-se inferência sobre γ, sabendo-se, além disso, que a distribuição amostral de Z depende unicamente de γ, i.e.,

$$f(z|\gamma, \eta) = \int f(y, z|\gamma, \eta)\, dy = f(z|\gamma) \,. \tag{3.67}$$

Se escrevermos a distribuição *a priori* na forma $\pi(\theta) = \pi(\gamma) \,.\, \pi(\eta)$, a aplicação do teorema de Bayes dá a seguinte distribuição *a posteriori* marginal,

$$\pi(\gamma|x) \propto \pi(\gamma) \int f(y, z|\gamma, \eta)\, \pi(\eta)\, d\eta \,, \tag{3.68}$$

a qual em geral deverá depender da informação *a priori* disponível sobre η.

Suponha-se agora que um outro investigador, ignorante das componentes y e η, aplica o teorema de Bayes, obtendo,

$$\pi(\gamma|z) \propto \pi(\gamma)\, f(z|\gamma). \tag{3.69}$$

Geisser e Cornfield (1963) chamam à função $\pi(\gamma|z)$, "pseudoposterior distribution".

É claro que os dois investigadores chegarão, regra geral,

---

4 (cont. pág. ant.) that it leads to a new means for defining what is meant by "uninformative" and for constructing noninformative priors as the solution of an integral equation. This method draws only upon the universally accepted principles of probability theory, making no appeal to such additional desiderata as entropy, group invariance, or Fisher information" (Jaynes, 1980, p. 43).

a conclusões diferentes, uma vez que o primeiro investigador tomou em consideração informação extra sobre (y,η). No entanto, se os investigadores chegarem às mesmas conclusões, é óbvio que o primeiro deles também não incorporou, afinal, qualquer informação sobre η. Assim, a distribuição *a priori* $\pi(\eta)$ que tornar as conclusões dos dois investigadores concordantes deve ser completamente não informativa. A condição matemática para a igualdade de (3.68) e (3.69) é a equação integral de Fredholm,

$$\int f(y,z|\gamma,\eta)\,\pi(\eta)\,d\eta = \lambda(y,z)\,f(z|\gamma)\,, \tag{3.70}$$

onde $\lambda(y,z)$ é uma função a determinar. A análise matemática de (3.70) não é simples[5], mas conhecem-se alguns resultados isolados. Assim, se y e η forem números positivos, sendo η um parâmetro de escala para y, a função de Jeffreys, $\pi(\eta) = \frac{1}{\eta}$, é uma solução (de 3.70), devendo então ser $\lambda(y,z) = \frac{1}{y}$. Podem considerar-se modelos específicos para os quais $\pi(\eta) = \frac{1}{\eta}$ seja única. É assim interessante que o princípio da marginalização dê resultados consistentes com outros métodos, ao mesmo tempo que utiliza unicamente os

5 Desconhecem-se ainda condições necessárias e suficientes gerais sobre $f(y, z|\gamma, \eta)$ para a existência ou unicidade de soluções.

princípios básicos da probabilidade sem fazer qualquer apelo à procura de grupos de transformação.

Por outro lado, este método estaria livre de uma das críticas mais pertinentes feita aos métodos de invariância: a de que o processo de derivação da distribuição *a priori* não informativa depende da estrutura experimental. Com efeito, desejando-se representar uma distribuição de ignorância total, não deixa de ser problemático que esta dependa do modelo de probabilidade adoptado[6].

Box e Tiao advogam o uso de distribuições que representem não uma ignorância total, estado cuja possibilidade de existência negam, mas antes uma quantidade de informação que seja pequena em relação àquela que é esperada por via dos dados experimentais[7]. Neste sentido, o tipo de distribuições propostas por estes autores (uniformes para um parâmetro tal

---

6 "Why should one's knowledge, or ignorance, of a quantity depend on the experiment being used to determine it. Incidently, if this view is accepted it shows a danger in developping "ready-made" Bayesian analysis in which θ is just a parameter". (Lindley, 1972, p. 71).

7 "It is important to bear in mind that one can never be in a state of *complete* ignorance; further, the statement "knowing little *a priori* can only have meaning relative to the information provided by an experiment" (Box e Tiao, 1973, p. 25).

que a verosimilhança seja "data translated", i.e., completamente determinada à excepção da sua localização, que será dada pelos dados a observar) podem, com boa justificação, depender do valor esperado da verosimilhança[8].

3.9. Alguns exemplos. Retome-se agora o problema da procura da distribuição MAXENT no caso em que o parâmetro Θ é contínuo. Suponha-se ainda que a informação *a priori* assume a forma do conhecimento de m momentos,

$$(3.71) \quad E\left[g_k(\Theta)\right] = \int_{\tilde{\Theta}} g_k(\theta)\, \pi(\theta)\, d\theta = \mu_k \,, \quad k=1, 2, \ldots, m \,.$$

A densidade que maximiza $En_c(\pi) = -\int_{\tilde{\Theta}} \pi(\theta) \ln \frac{\pi(\theta)}{\pi_o(\theta)}\, d\theta$, onde, de acordo com a sugestão de Jaynes (cf. p.), $\pi_o(\theta)$ é a distribuição *a priori* não informativa invariante natural para o problema, será dada por,

$$(3.72) \quad \bar{\pi}(\theta) = \frac{\pi_o(\theta)\, \exp\left\{\sum_{k=1}^{m} -\lambda_k\, g_k(\theta)\right\}}{\int_{\tilde{\Theta}} \pi_o(\theta)\, \exp\left\{\sum_{k=1}^{m} -\lambda_k\, g_k(\theta)\right\} d\theta} \,,$$

8 "... we seek to represent not total ignorance but an amount of prior information which is small relative to what the particular projected expected experiment can be expected to provide. The form of the prior *must* then depend on the expected likelihood" (Box e Tiao, 1973, p. 44).

onde os λ_k são constantes a determinar a partir das restrições (3.71). A densidade apresentada em (3.72), e que tem uma correspondência formal evidente com a solução obtida para o caso discreto em (3.14), deriva de técnicas do cálculo das variações. Como exemplo, veja-se primeiro o caso em que $\tilde{\Theta} = \mathrm{R}^+$, Θ é um parâmetro de localização e em que o único momento conhecido da distribuição é o primeiro, i. e., $\mathrm{E}(\Theta) = \mu$. A natureza do parâmetro pode levar a escolher a densidade *a priori* não informativa $\pi_0(\theta) = 1$. Nesse caso, tem-se então de (3.72),

$$(3.73) \qquad \bar{\pi}(\theta) = \frac{\exp(-\lambda_1\theta)}{\int_0^{+\infty} \exp(-\lambda_1\theta)\, d\theta} = \lambda_1 \exp(-\lambda_1\theta),$$

distribuição exponencial, que satisfará a restrição $\mathrm{E}(\Theta) = \mu$ fazendo $\lambda_1 = \frac{1}{\mu}$. Note-se, contudo, que caso fosse $\tilde{\Theta} = \mathrm{R}$, o problema não teria solução.

Como segundo exemplo, veja-se o caso em que Θ é um parâmetro de localização, com espaço do parâmetro $\tilde{\Theta} = \mathrm{R}$, e supondo-se ainda conhecidas a média e a variância, i. e.,

$$(3.74) \qquad g_1(\theta) = \theta\,; \qquad \mu_1 = \mu$$
$$g_2(\theta) = (\theta - \mu)^2\,; \qquad \mu_2 = \sigma^2\,.$$

A solução (3.72) é para este caso, onde se considerou novamente $\pi_0(\theta) = 1$,

$$(3.75) \qquad \bar{\pi}(\theta) = \frac{\exp\left\{-\lambda_1\theta - \lambda_2 (\theta-\mu)^2\right\}}{\int_{-\infty}^{+\infty} \exp\left\{-\lambda_1\theta - \lambda_2 (\theta-\mu)^2\right\} d\theta} .$$

Havendo a possibilidade de transformar algebricamente o numerador de (3.75) ao completar o quadrado, vê-se com facilidade que a densidade MAXENT obtida neste caso é a normal, com as restrições (3.71; 3.74) respeitadas desde que se faça $\lambda_1 = 0$ e $\lambda_2 = \frac{1}{2\sigma^2}$.

O problema tem um tratamento fácil no caso do conhecimento de momentos, o que de certa forma explica a sua aplicação com sucesso nas ciências físicas, onde, com frequência, o conhecimento das restrições se apresenta nessa forma. Noutros casos, porém, existem reais dificuldades na aplicação da técnica. Assim, sendo Θ não limitado e tomando as restrições consideradas forma de conhecimento sobre quantis, a densidade MAXENT não existe. Esta última situação é bastante séria na prática, uma vez que em muitos contextos fora das ciências físicas, o conhecimento subjectivo sobre a distribuição de Θ se exprime mais naturalmente na forma de quantis e não na forma de momentos[9]. Tal poderá explicar o

9 Uma outra dificuladade reside no facto de a especificação subjectiva de momentos poder levar a problemas de falta de robustez para as distribuições MAXENT (cf. Berger, 1985, Cap. 4).

facto de que a "explosão" referida por Berger (1985, p. 94) do uso do MAXENT se verifique preferencialmente em campos directamente relacionados com as ciências físicas (aqui incluindo disciplinas tais como a Geofísica, a análise de imagens, a análise espectral, etc.), sendo a sua aplicação ainda incipiente em campos mais ligados às ciências sociais (aqui incluindo a Economia).

3.10. Conclusão. Depois de, no primeiro capítulo, se ter seguido a evolução do conceito de entropia desde o seu aparecimento em Termodinâmica até à sua utilização na Teoria da Informação, ao mesmo tempo que se chamou a atenção para o carácter controverso de querer reconhecer um mesmo conceito na base de paralelismos formais em campos de conhecimento distintos, tratou-se, no segundo capítulo, do problema da utilização do conhecimento *a priori* em inferência. Esta última questão é de facto central, sendo uma verdadeira pedra de toque para a consideração de várias escolas em Inferência Estatística. Finalmente, no terceiro capítulo, viu-se como o MAXENT pode proporcionar em alguns casos uma resposta satisfatória a tal questão. Para além do seu sucesso em certos campos de aplicação, o MAXENT possui um invejável atractivo devido à sua simplicidade lógica.

Sucintamente o método pode ser considerado como uma

tentativa de resposta a um velho *desideratum* da inferência: como exprimir na forma menos enviesada possível o estado de conhecimento *a priori* sobre uma dada situação? Viu-se então como tal problema tinha ligação directa com a procura de distribuições *a priori* não informativas. Esta última questão é sumamente difícil e provavelmente não dispõe de resposta absolutamente satisfatória. A procura de uma regra universal para as probabilidades *a priori* [10] teve de ceder o lugar a uma busca mais realista de regras próprias para aplicação em situações mais delimitadas[11].

Por outro lado, havendo uma grande actividade e interesse nas aplicações a problemas concretos em diversos domínios, é natural que daí advenham novas ideias de cariz teórico. A própria utilização do conceito de entropia, tão central e controverso em Ciência, contribui para dar a este método um

10 "I had hopes for a time of getting a universal rule for prior initial probabilities" (Jeffreys, 1980, p.451)

11 "It may be emphasized here that it is not logically necessary to produce a single invariance rule which will be satisfactorily applicable to all distributions. In fact there seems to be little hope for such a rule of universal application, just as it is unlikely to discover a single scientific law to explain satisfactorily all physical phenomena." (Huzurbazar, 1980, p.448)

lugar privilegiado no debate epistemológico sobre Inferência Estatística.

BARD, Y. - "Maximum Entropy Principle, Classical Aproach", *in Encyclopaedia of Statistical Sciences*, John Wiley and Sons, 1985, vol. V, pp. 336-338.

BARNETT, V. - *Comparative Statistical Inference* (Cap. VI), John Wiley and Sons, 1973.

BARTLETT, M. S. - "Irreversibility and Statistical Theory", *in Essays on Probability and Statistics*, Methuen and Co., 1962, pp. 104-111.

BARTLETT, M. S. - "Fisher, R. A.", *in International Encyclopedia of the Social Sciences*, The Macmillan Co. and the Free Press, 1968, vol. V, pp. 485-491.

BEAUREGARD, O. C. de - "Information and Irreversibility Problems", *in Time in Science and Philosophy*, Elsevier Publishing Company, 1971, pp. 11-25.

BERGER, J. O. - *Statistical Decision Theory and Bayesian Analysis*, 2ª ed., Springer-Verlag, 1985.

BERGER, T. - "Information Theory and Coding Theory", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1983, vol. IV, pp. 124-141.

BOX, G. E.
e
TIAO, G. C. - *Bayesian Inference in Statistical Analysis* (Cap. I), Addison-Wesley Publ. Co., 1973.

BRIDGMAN, P. W. - *The Nature of Thermodynamics*, Harvard University Press, 1941.

BRUSH, S. G. - "Irreversibility", *in The Encyclopaedia of Physics*, 3ª ed., Van Nostrand Reinhold Company, 1985, pp. 614-617.

CAMPBELL, J. - *Grammatical Man*, Penguin Books, 1984.

DAUB, E. E. - "Maxwell's Demon", *Studies in History and Philosophy of Science*, 1, 1970, pp. 213-227, *in Darwin to Einstein, Historical Studies on Science and Belief*, Longman, 1980, pp. 222--235.

DAWID, A. P. - "Inference, Statistical: I", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1983, vol. IV, pp. 89-105.

DAWID, A. P. - "Invariant Prior Distributions", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1983, vol. IV, pp. 228-236.

DE GROOT, M. H. - "A Conversation with George A. Barnard", *Statistical Science*, 1988, vol. 3, nº2, pp. 196-212.

DENBIGH, K. G.
DENBIGH, J. S. - *Entropy in Relation to Incomplete Knowledge*, Cambridge University Press, 1985.

FEINSTEIN, A. - *Foundations of Information Theory*, Mc Graw-Hill, 1958.

FISHER, R. A. - *Statistical Methods for Research Workers*, Oliver and Boyd, 1925.

FISHER, R. A. - *Statistical Methods and Scientific Inference*, Oliver and Boyd, 1956.

FISHER BOX, J. - *R. A. Fisher: The Life of a Scientist*, John Wiley and Sons, 1978.

FRÉCHET, M. - "Laplace, Pierre Simon de", *in Int. Encycl. of the Soc. Sc.*, The Macmillan Co. and the Free Press, 1968, vol. IX, pp. 23-27.

FREEMAN, H. - "Wald, Abraham", *in Int. Encycl. of the Soc. Sc.*, The Macmillan Co. and the Free Press, 1968, vol. XVI, pp. 435-438.

GEISSER, S.
CORNFIELD, J. - "Posterior Distributions for Multivariate Normal Parameters", *Journal of the Royal Statistical Society*, 1963, B 25, pp. 368--376.

GEORGESCU-ROEGEN, N. - "Entropy", *in The New Palgrave Dictionary of Economics*, The Macmillan Press Ltd., 1987, vol. II, pp. 153-156.

GOOD, I, J. - "Subjective Probability", *in The New Palgrave Dict. of Econ.*, The Macmillan Press Ltd., 1987, vol. IV, pp. 537-543.

GOOD, I, J. - "Scientific Method and Statistics", *in Encycl. Stat. Sc.*, John Wiley and Sons, 1988, vol. VIII, pp. 291-304.

HARRIS, B. - "Entropy", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1982, vol. II, pp. 512-516.

HARTIGAN, J. - "Invariant Prior Distributions", *Ann. Math. Statist.*, 1964, vol. 35, pp. 836-845.

HOBSON, A. - *Concepts in Statistical Mechanics*, Gordon & Breach, 1971.

HUZURBAZAR, V. S. - "Bayesian Inference and Invariant Prior Probabilities", *in Bayesian Analysis in Econometrics and Statistics*, North-Holland Publ. Co., 1980, pp. 445-449.

JAYNES, E. T. - "Information Theory and Statistical Mechanics", *Physical Review*, 1957,

vol. 106, nº 4, pp. 620-630; vol. 108, nº 2, pp. 171-190.

JAYNES, E. T. - "Prior Probabilities", *IEE Trans. Syst. Sci. Cybern.* SSC-4, 1968, pp.227-241, *in Concepts and Applications of Modern Decision Models*, Michigan State University Business Studies Series, 1976, pp. 87-101.

JAYNES, E. T. - "Where do we stand on Maximum Entropy?", *in The Maximum Entropy Formalism*, The M. I. T. Press, 1979, pp. 15-117.

JAYNES, E. T. - "Marginalization and Prior Probabilities", *in Bayesian Analysis in Econometrics and Statistics*, North-Holland Publ. Co., 1980.

JAYNES, E. T. - "Where do we go from here?", *in Maximum - - Entropy and Bayesian Methods in Inverse Problems*, D. Reidel Publ. Co., 1985, pp. 21-58.

JEFFREYS, H. - *Theory of Probability*, 3ª ed., (Cap. III), Oxford at the Clarendon Press, 1961.

JEFFREYS, H. - "Some General Points in Probability Theory", *in Bayesian Analysis in Econometrics and Statistics*, North-Holland Publ. Co., 1980, pp. 451-453.

KEMP, A. W. - *Bull. Inst. Int. Statist.*, 1973, nº 45, pp. 45--51

KINCHIN, A. I. - *Mathematical Foundations of Information Theory*, Dover, 1957.

KULLBACK, S. - *Information Theory and Statistics*, John Wiley and Sons, 1959.

KULLBACK, S. - "Minimum Discrimination Information (MDI) Estimation", *in Encycl. of Stat. Sc.*, 1985, vol. V, pp. 527-529.

KYBURG, H. - "Logic of Statistical Reasoning", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1985, vol. V, pp. 117-122.

LEHMANN, E. L. - "Statistics: An Overview", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1988, vol. VIII, pp. 638-702.

LINDLEY, D. V. - "The use of Prior Probability Distributions in Statistical Inference and Decisions", *in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1961, vol. I, pp. 453-468.

LINDLEY, D. V. - *Bayesian Statistics, a Review*, Society for Industrial and Applied Mathematics, 1972.

LINDLEY, D. V. - "Bayes, Thomas", *in The New Palgrave Dict. of Econ.*, The Macmillan Press Ltd, 1987, vol. I, pp. 207-208.

LINDLEY, D. V. - "Fisher, Ronald Aylmer", *in The New Palgrave Dict. of Econ.*, The Macmillan Press Ltd, 1987, vol. II, pp. 376-377.

LINDLEY, D. V. - "Statistical Inference", *in The New Palgrave Dict. of Econ.*, The Macmillan Press Ltd, 1987, vol. IV, pp. 490-493.

LINDLEY, D. V. - "Savage, Leonard J.", *in Encycl. Stat. Sc.*, John Wiley and Sons, 1988, vol. VIII, pp. 265-267.

MURTEIRA, B. J. F. - *Estatística: Inferência e Decisão*, Instituto Superior de Economia, Lisboa, 1986.

PAIS, A. - *Subtle is the Lord...* (Cap. IV), Oxford University Press, 1982.

PAPAIOANNOU, T. - "Measures of Information", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1985, vol. V, pp. 391-397.

PAPOULIS, A. - *Probability, Random Variables and Stochastic Processes*, Mc Graw-Hill, 1984.

PEARSON, E. S. - "Prepared Contribution", *in The Foundations of Statistical Inference*, Methuen and Co., 1962, pp. 53-58.

PORTELA, A. G. e MURTEIRA, B. J. F. - "Entropia e Distribuições Conjugadas", *Estudos de Economia*, vol. II, nº2, 1982.

PORTER, T. M. - *The Rise of Statistical Thinking* (Cap. VII), Princeton University Press, 1986.

REICHENBACH, H. - "Logic and Predictive Knowledge", *The Rise of Scientific Philosophy*, 1951, *in Space, Time and the New Mathematics*, Bantam Books, 1964, pp. 46-71.

RÉNYI, A. - "On Measures of Entropy and Information", *in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1961, vol. I, pp. 547-561.

RÉNYI, A. - *Probability Theory* (Cap. IX), North - Holland Publish. - Co., 1970.

RUSHBROOKE, G. S. - "Statistical Mechanics", *in Encyclopaedia Britannica*,

1966, vol. XXI, pp. 340B--342.

SAVAGE, L. J. - "The Foundations of Statistics Reconsidered", *in Proceedings of the the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1961, vol. I, pp. 575-586.

SAVAGE, L. J. - "Bayesian Statistics", *in Recent Developments in Information and Decision Processes*, The Macmillan Co., 1962, pp. 161-194.

SCOTT, E. L. - "Neyman, Jerzy", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1985, vol. VI, pp. 215-223.

SEAL, H. L. - "Bayes, Thomas", *in Int. Encycl. of the Soc. Sc.*, The Macmillan Co. and the Free Press, 1968, vol. I, pp. 26-28.

STIGLER, S. M. - "Laplace's 1774 Memoir on Inverse Probability", *Statistical Science*, 1986, vol. 1, nº 3, pp. 359-378.

TRIBUS, M. - *Décisions Rationelles dans L'Incertain*, Masson e C^{ie} Editeurs, 1972.

TRIBUS, M. - "Thirty Years of Information Theory", *in The Maximum Entropy Formalism*, The M. I. T. Press, 1979, pp. 1-14.

TRIBUS, M. - "Entropy", *in The Encyclopaedia of Physics*, 3ª ed., Van Nostrand Reinhold Company, 1985, pp. 404-406.

VENTSEL, H. - *Théorie des Probabilités*, Editions MIR, 1973.

WALD, A. - *Statistical Decision Functions*, John Wiley and Sons, 1950.

WEISS, L. - "Wald, Abraham", *in Encycl. of Stat. Sc.*, John Wiley and Sons, 1988, vol. IX, pp. 514-517.

ZELLNER, A. - *An Introdution to Bayesian Inference in Econometrics*, John Wiley and Sons, 1971.

ÍNDICE

1. ENTROPIA: UM SÓ OU VÁRIOS CONCEITOS DISTINTOS? - 1

1.1. Da Termodinâmica à Mecânica Estatística: Clausius e Maxwell. - 1

1.2. A Mecânica Estatística: Boltzmann e Gibbs. - 5

1.3. A Teoria da Informação: Shannon e Jaynes. - 12

2. O USO DA INFORMAÇÃO *A PRIORI* EM INFERÊNCIA - 20

2.1. Introdução. - 20

2.2. O princípio da razão insuficiente: Bernoulli, Bayes e Laplace. - 22

2.3. As várias escolas de inferência perante o uso da informação *a priori*. - 31

2.3.1. A escola de Fisher. - 31

2.3.2. A escola de Neyman-Pearson-Wald. - 35

2.3.3. A escola Bayesiana. - 38

3. A MAXIMIZAÇÃO DA ENTROPIA NA PROCURA DE DISTRIBUIÇÕES *A PRIORI* - 47

3.1. Introdução. - 47

3.2. Informação *a priori* testável. - 49

3.3. Entropia de uma distribuição discreta. - 50

3.4. Alguns exemplos. - 56

3.5. Uma interpretação combinatória. - 59

3.6. Entropia de uma distribuição contínua. - 65

3.7. Distribuições *a priori* invariantes. - 71

3.8. O método da marginalização. - 78

3.9. Alguns exemplos. - 83

3.10. Conclusão. - 86

BIBLIOGRAFIA - 89