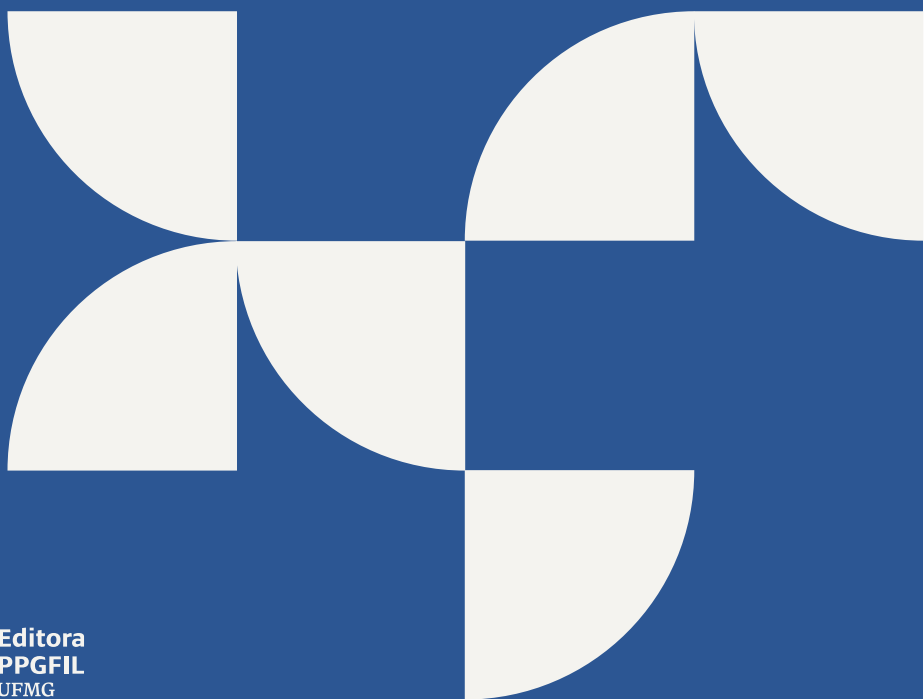


Probabilidade

Gerhard Schurz

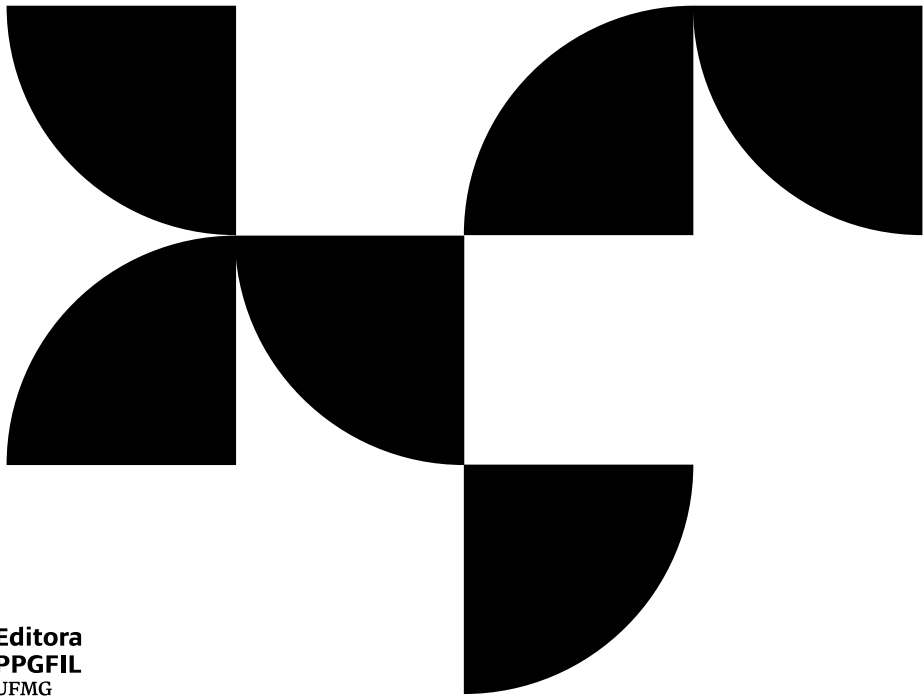
Tradução de
Marcos Sousa



Probabilidade

Gerhard Schurz

Tradução de
Marcos Sousa



UNIVERSIDADE FEDERAL DE MINAS GERAIS

Reitora: Sandra Regina Goulart Almeida

Vice-Reitor: Alessandro Fernandes Moreira

FACULDADE DE FILOSOFIA E CIÊNCIAS HUMANAS

Diretora: Thais Porlan de Oliveira

Vice-Diretor: Rogério Duarte do Pateo

DEPARTAMENTO DE FILOSOFIA

Chefe: Leonardo de Mello Ribeiro

Subchefe: Tadeu Mazzola Verza

PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

Coordenador: Ernesto Perini Frizzera da Mota Santos

Subcoordenador: Rogério Antônio Lopes

EDITORA PPGFIL-UFMG

Tadeu M. Verza

André J. Abath

Abílio A. Rodrigues Filho

Henrique B. Barreto

CONSELHO EDITORIAL

Daniel Pucciarelli

Miriam Campolina Diniz Peixoto

Rogério Antônio Lopes

Verlaine Freitas

CONSELHO CIENTÍFICO

Alessandro Pinzani (UFSC)

Anderson Bogéa da Silva (UFPR)

André Leclerc (UNB)

André Nascimento Pontes (UFAM)

Antônio Jorge Gomes Bento (Universidade da Beira Interior)
Beatriz Cecilia Bossi López (UCM)
Catarina Belo (Universidade Americana no Cairo)
Gisele Amaral dos Santos (UFRN)
Jacques Poulain (Université Paris VIII)
José Leonardo Annunziato Ruivo (UEMA)
Marco Antônio Caron Ruffino (UNICAMP)
Marcus Sacrini Ayres Ferraz (USP)
Maria Aparecida Paiva Montenegro (UFCE)
Maurício Pagotto Marsola (UNIFESP)
Sara Juliana Pozzer da Silveira (UFMT)
Vinicius Berlendis de Figueiredo[†] (UFPR)

<https://www.edppgfil.fafich.ufmg.br>

<https://www.editorappgfilufmg.com>

Avenida Presidente Antônio Carlos, 6627

FAFICH, sala 4047, 4º Piso

Pampulha, Belo Horizonte - MG.

CEP: 31270-901

Dados internacionais de catalogação na publicação (CIP)

Schurz, Gerhard.

Probabilidade [recurso eletrônico] / Gehard Schurz ; Marcos Sousa (trad.). – Belo Horizonte : PPGFIL, 2025.

S394w.Ps

1 recurso online (343 p.): pdf

Tradução de: Wahrscheinlichkeit.

ISBN: 978-65-01-77089-5

DOI: 10.5281/zenodo.17478340

1. Probabilidades. I. Sousa, Marcos Roberto de. II. Título.

CDD: 121.63

CDU: 519.2

Elaborada por Vilma Carvalho de Souza - CRB-6/1390

Copyright © 2015 Walter de Gruyter GmbH (edição alemã)

Copyright © 2025 Marcos Sousa (para a presente edição)

Capa e projeto gráfico: Ami Comunicação & Design

Revisão: Alice Carvalho Diniz Leite

Diagramação: Henrique B. Barreto

Finalização: Tadeu M. Verza



PPGFIL-UFMG, 2025

Esta obra foi selecionada pelo Conselho Editorial do Selo PPGFIL/UFMG após avaliação por pareceristas *ad hoc*.
O acesso e a leitura deste livro estão condicionados ao aceite dos termos de uso do Selo do PPGFIL/UFMG, disponíveis em: <https://www.edppgfil.fafich.ufmg.br>

Sumário

Prefácio pessoal	7
Notas técnicas	11
1 Introdução	13
2 Probabilidade objetiva e epistêmica	17
3 Fundamentos matemáticos da probabilidade	27
3.1 Leis da probabilidade	27
3.2 Distribuição binomial e a lei dos grandes números	38
3.3 Estruturas formais da teoria das probabilidades	44
3.4 Aditividade sigma: prós e contras	54
4 Justificativa das modalidades de inferência dentro da teoria das probabilidades	59
4.1 Modalidades de inferência	59
4.2 Raciocínio dedutivo	61
4.3 Condicionais incertas	65
4.4 Raciocínio indutivo	68
4.5 Raciocínio abdutivo	70

5 Problemas do conceito objetivo-estatístico de probabilidade 73

5.1 Problemas de justificação	73
5.2 Problemas de definição	80
5.3 Conteúdo empírico	98
5.4 Aleatoriedade objetiva, determinismo e indeterminismo	100
5.5 Propensões singulares	108

6 Problemas do conceito subjetivo-epistêmico de probabilidade 113

6.1 Problemas de definição	113
6.2 Problemas de justificação: quocientes de apostas justas coerentes	114

7 Relações entre probabilidades objetivas e epistêmicas: uma abordagem dualista 125

7.1 Princípio de coordenação (ou “princípio principal”)	125
7.2 O conteúdo empírico-indutivo das hipóteses estatísticas	133
7.3 Probabilidades iniciais independentes da experiência	135
7.4 Das probabilidades basais às crenças atuais: condicionalização à evidência geral	138
7.5 Probabilidades de suporte e o problema da evidência antiga	141
7.6 Permutabilidade e teorema da representação de De Finetti	146
7.7 Regularidade e aprendizagem indutiva	149
7.8 A justificativa das classes de referência mais estreitas	153
7.9 Tipos de classes de referência mais estreitas e calibração	156

8 Teste de hipóteses estatísticas **165**

8.1 Teste de verificação da verdade – o método dos intervalos de aceitação	166
8.2 Encontrando hipóteses estatísticas e intervalos de confiança	172
8.3 Teste de verificação da relevância – o método das diferenças significativas	174
8.4 Representatividade estatística e tipos de hipóteses estatísticas	181
8.5 Estatística de teste e estatística de inferência	188
8.6 Distribuições de probabilidade e métodos estatísticos para variáveis contínuas	192
8.7 Fontes de erro nas estatísticas: representatividade, interpretação causal e caso individual	207

9 Estatística bayesiana e bayesianismo **221**

9.1 A intuição da verossimilhança	221
9.2 Justificação bayesiana da intuição de verossimilhança	229
9.3 Bayesianismo objetivo e o princípio da indiferença: inferência indutiva I	233
9.4 Probabilidades de hipóteses sem o princípio da indiferença?	243
9.5 Bayesianismo subjetivo e convergência de graus subjetivos de crença: inferência indutiva II	251
9.6 Evidências consistentes e independentes	259
9.7 Justificativa probabilística do raciocínio indutivo? O paradoxo de Goodman	262
9.8 Teorias gerais bayesianas de confirmação	269
9.9 Pseudo-confirmação através de redução do espaço de possibilidades (truncamento) versus confirmação genuína	272

9.10 Ajuste de curvas	283
9.11 Probabilidade e aceitação	288
10 Apêndice lógico-matemático	295
10.1 Fundamentos lógicos	295
10.2 Construção lógica de funções estatísticas de probabilidade sobre experimentos aleatórios combinados	307
10.3 Provas	312
Bibliografia	329
Lista de figuras, definições e teoremas	343
Índice de pessoas	346

Prefácio pessoal

Nos mais de 30 anos em que me dediquei à teoria da ciência e à filosofia científica, a teoria da probabilidade (entre outros métodos formais) tem sido minha companheira constante. Isso se explica, por um lado, pelo fato de eu sempre ter sido um filósofo da ciência próximo das ciências especializadas (devido à minha dupla formação em ciências naturais e em filosofia) e, por outro lado, pela natureza das ciências e de seus próprios temas. Primeiro, a maioria das leis pesquisadas nas ciências (possivelmente com exceção da física clássica) não é estritamente determinística, mas sim de natureza estatística. Em segundo lugar, o conhecimento científico, exceto nas áreas científicas formais, é fundamentalmente incerto. Pela primeira razão, precisa-se da teoria estatística da probabilidade; e, pela segunda razão, precisa-se da teoria epistêmica da probabilidade. Uma filosofia teórica que não quer permanecer na torre de marfim filosófica e que antes procura resultados significativos e aplicáveis às ciências reais não pode permanecer no quadro da lógica dedutiva, mas deve incluir a teoria das probabilidades. Essa afirmação não significa que a lógica dedutiva não seja, entretanto, importante ou que possa ser dispensada. Não, a lógica dedutiva também está no cerne da teoria da probabilidade (como um componente necessário, mas não suficiente), e quem ignora a primeira também não consegue compreender a última.

Isso explica por que a teoria das probabilidades era minha companheira constante: porque eu precisava dela para aplicações. Devo agradecer a Dieter Birnbacher pelo fato de essa situação ter se

tornado o ímpeto para escrever as minhas experiências e os meus conhecimentos acumulados nessa área na forma de um pequeno livro. O que há de especial neste livrinho é – pelo menos penso assim – que não olho e nem desenvolvo a teoria das probabilidades de um ângulo específico, como sempre, mas antes tento reunir as diferentes perspectivas, terminologias e campos científicos.

A meu ver, o campo da teoria da probabilidade tem permanecido, há décadas, num estado infeliz de segregação. Enquanto as ciências empíricas falam quase exclusivamente sobre probabilidade estatística, os bayesianos, que são influentes na filosofia da ciência, entendem fundamentalmente a probabilidade no sentido subjetivo de graus racionais de crença, ao passo que os teóricos da probabilidade matemática ignoram sistematicamente esse conflito de interpretação. Como filósofo da ciência orientado para a aplicação, deparo-me frequentemente com a situação desagradável de ter de utilizar o conceito estatístico de probabilidade para obter resultados aplicáveis, enquanto os colegas me contrapõem dizendo: “foi demonstrado” que o conceito “frequentista” de “probabilidade” não funciona e é melhor usar o conceito bayesiano do grau de crença ou o conceito metafísico da propensão objetiva para casos individuais. No entanto, as ciências aplicadas utilizam facilmente os métodos da estatística – que, como se verá, funcionam muito bem se o conceito de limite de frequência for devidamente compreendido –, enquanto o conceito subjetivo de probabilidade, bem como o conceito metafísico de propensão para casos individuais, desempenha um papel comparativamente subordinado no jogo das ciências. Essa situação complicada foi outro impulso para eu escrever este livro.

Em última análise, cheguei à conclusão de que ambos os conceitos de probabilidade são necessários, razão pela qual

desenvolvo uma posição dualista neste livro, que se preocupa sobretudo com a elaboração dos princípios de ligação entre os dois conceitos de probabilidade. Isso começa com o fato de eu desenvolver a função de probabilidade sobre a álgebra das propriedades ou das proposições de uma linguagem interpretada, porque só assim a diferença lógica entre a probabilidade estatística e a epistêmica (a primeira tem fórmulas abertas e a segunda tem sentenças como argumentos) é explicitada, o que é pré-requisito para a elaboração dos princípios-ponte que as conectam.

Com base em uma passagem bem conhecida de Kant, a posição dualística pode ser formulada da seguinte forma: a teoria subjetiva sem a teoria estatística da probabilidade é cega (pois é irracional), e a teoria estatística sem a teoria subjetiva da probabilidade é vazia (pois não tem referência empírica). Contudo, a posição dualista não significa que tudo o que foi reivindicado em ambas as posições possa agora ser adotado – isso levaria rapidamente a contradições. Pelo contrário, certas partes de cada uma dessas duas posições devem ser abandonadas como “difícilmente sustentáveis”. Por exemplo, vejo que a posição subjetivista radical no bayesianismo é tão insustentável quanto a tese de que os sujeitos racionais devem ter graus de crença racionais a priori na forma de quocientes de apostas justas em todas as questões factuais (mesmo que não tenham tido experiência nisso). Por outro lado, acredito que a visão tradicional sobre o conceito estatístico de limite de frequência ser um conceito com conteúdo empírico pelo menos “aproximado” também precisa ser corrigido, uma vez que o conteúdo empírico desse conceito não é de natureza dedutiva, mas indutiva, e o conceito epistêmico de probabilidade é necessário para formulá-lo. Muitos outros *insights* desse tipo surgem da abordagem dualista, sobre a qual não quero revelar mais nada

Probabilidade

agora, mas em vez disso desejo ao leitor muita diversão na leitura deste pequeno livro.

Gerhard Schurz

Düsseldorf, maio de 2014.

Notas técnicas

O *apêndice lógico-matemático* contém adições essenciais, algumas de natureza introdutória e outras de natureza mais avançada. O Apêndice 10.1 apresenta pressupostos lógicos elementares para referência do iniciante. O Apêndice 10.2 explica os detalhes da construção lógica de funções de probabilidade usando álgebras e linguagens lógicas para o leitor avançado. Finalmente, o Apêndice 10.3 apresenta as demonstrações de todos os teoremas apresentados neste livro.

As *figuras* são numeradas de acordo com o seguinte padrão: "Fig. Capítulo Nº - Nº...". Exemplo: "Fig. 2-2" é a 2ª figura do Capítulo 2. Definições, frases-chave e ênfases são numeradas de forma semelhante. Exemplo: "(Def. 4-2)" é a 2ª definição do Capítulo 4. "(3-4)" é a 4ª ênfase do Capítulo 3. Aspas simples indicam citações estilísticas e aspas duplas indicam citações literais.

1 Introdução

O conceito intuitivo de probabilidade (*wahrscheinlichkeit*) envolve, ao mesmo tempo, algo objetivo (verdade) e algo subjetivo (aparente). Os primeiros fundadores da teoria da probabilidade não haviam notado suficientemente essa ambiguidade. Somente no século XX é que a natureza diferente dos dois conceitos de probabilidade foi revelada.

A teoria da probabilidade surgiu nos séculos XVI e XVII, predominantemente no contexto dos jogos de azar.¹ Documentos importantes desse período incluem um manuscrito de Galileu do início do século XVII, uma troca de cartas entre Pascal e Fermat de 1654 e um artigo de Huygens publicado em 1657, todos os três sobre a questão de como calcular as probabilidades de resultados em jogos de azar com múltiplos dados. Em 1713, apareceu o famoso artigo de Bernoulli, no qual ele provou a distribuição binomial e a lei fraca dos grandes números; e, em 1763, o Teorema de Bayes foi publicado (Bayes; Price, 1763). Foi apenas meio século depois, em 1814, que Laplace publicou os seus ensaios sobre probabilidade, ocasião frequentemente associada ao início da história da teoria da probabilidade. Da perspectiva de hoje, a abordagem probabilística de Laplace era epistêmica, uma vez que Laplace atribuiu a mesma probabilidade a todos os resultados epistêmicos (isto é, cognitivamente) igualmente possíveis de um experimento ou de um processo físico. No entanto, Laplace não distinguiu esse “princípio de distribuição igual”, subjetivo, da probabilidade objetiva igual dos resultados do lançamento de um

¹ Veja Gillies (2000, capítulo 1) e David (1998).

dado regular (Laplace, 1814, p. 6). Von Mises (1928, p. 69) foi o primeiro a deixar clara a diferença quando questionou como lidar com uma irregularidade; por exemplo, deveria ser evitado um cubo magnetizado unilateralmente, cujos resultados também seriam epistemicamente desconhecidos, mas estatisticamente falando não seriam igualmente prováveis. Dentro da teoria de Laplace, não foi possível diferenciar os dois casos, porque nenhuma distinção foi feita entre a probabilidade subjetiva e a objetiva.²

A atual teoria da probabilidade é caracterizada por uma divisão persistente de campos. Enquanto as ciências empíricas abordam quase exclusivamente a probabilidade estatística objetiva, os bayesianos, que são influentes nas teorias da ciência e do conhecimento, entendem a probabilidade no sentido subjetivo de graus racionais de crença, ao passo que os teóricos da probabilidade matemática ignoram sistematicamente esse conflito de interpretação. Além desses três grupos principais, existem alguns subgrupos filosoficamente importantes. No campo objetivo, por exemplo, os teóricos da propensão, que veem as probabilidades como tendências físicas objetivas; no campo subjetivo, os representantes da teoria lógica da probabilidade ou do bayesianismo objetivo, segundo o qual o “objetivo” não deve ser entendido no sentido de “externo ao sujeito [ou independente do sujeito]”, mas sim de “intersubjetivo”. *“Last but not least”*, deve ser mencionado o grupo das teorias dualísticas das probabilidades, no qual incluo também a minha abordagem.

Embora pareça natural contrastar as concepções objetivas de probabilidade com as concepções “subjetivas”, sigo Gillies (2000, p. 2) e me refiro às concepções de probabilidade como um grau racional de

² Laplace menciona o problema do cubo assimétrico em um ponto, mas o ignora em sua teoria (Gillies, 2000, p. 18).

crença, como a família das teorias epistêmicas da probabilidade. Além dos subjetivistas, essa família também abarca os teóricos da probabilidade “lógica” e os bayesianos “objetivos”, que rejeitam a designação de suas probabilidades como “subjetivas”, porque o termo contém o significado de “subjetivamente variável”, o que é verdade para algumas (por exemplo, para as personalistas), mas não para todas as teorias epistêmicas de probabilidade.

Os principais fundadores da teoria estatística da probabilidade incluem: von Mises (1964), Reichenbach (1935, 1949) e Fisher (1956) (Hays/Winkler, 1970, e Bortz, 1985, são recomendados como literatura introdutória em estatística). Os principais fundadores da teoria subjetiva incluem: Ramsey (1926) e De Finetti (1970) (a literatura introdutória pode ser encontrada, por exemplo, em Earman, 1992, e Howson/Urbach, 1996). Keynes (1921) e Carnap (1950, 1971, 1980) fundam a teoria lógica da probabilidade, que acrescenta outros axiomas ou princípios aos axiomas básicos da teoria da probabilidade, bem como que pretende fixar os graus de crença “a priori” para todos os sujeitos racionais. No entanto, existe um consenso generalizado de que esses axiomas adicionais vão muito além do âmbito do lógico e analiticamente válido, razão pela qual coloco o 'lógico' dessas teorias de probabilidade entre aspas (para o conceito de verdade/validade lógica ou analítica consulte o apêndice 10.1.3). Introduções representativas da teoria matemática da probabilidade são desenvolvidas por Bauer (1996) ou por Billingsley (1995). Visões gerais de diferentes abordagens de probabilidade são fornecidas por T. Fine (1973), Stegmüller (1973a, 1973b), Kutschera (1972, Capítulo 2), Howson/Urbach (1996) e especialmente Gillies (2000).

2 Probabilidade objetiva e epistêmica

A probabilidade objetiva expressa uma propriedade da realidade que é independente do sujeito. A probabilidade subjetiva, por outro lado, expressa o grau (real ou hipotético) de crença de um sujeito racional. Se esta envolve graus intersubjetivos de crença, falamos mais genericamente de probabilidade “epistêmica” (mais sobre isso abaixo). A Def. 2-1 apresenta a definição básica de probabilidades estatística e subjetiva, utilizando notação simbólica. “ Fx ” significa “ x é um F ” e “ Fa ” significa “ a é um F ”; “ F ” é um predicado que denota uma característica ou um evento do tipo F ; “ x ” é uma variável individual e “ a ” é uma constante individual, respectivamente, que descreve uma variável ou que designa um indivíduo específico. Uma breve introdução às notações lógicas e aos tipos de termos pode ser encontrada no Apêndice 10.1.

(Def. 2-1) A probabilidade *estatística (objetiva)* de uma característica ou tipo de evento repetível, por exemplo, Fx , é a frequência relativa de sua ocorrência ou o limite de sua frequência relativa no longo prazo. Formalmente, escrevemos um “minúsculo” $p(-)$: $p(Fx)$ para representar a frequência ou o limite de frequência que qualquer indivíduo x em uma determinada área possui a propriedade F . – Exemplo: A frequência de dias ensolarados em Düsseldorf.

A probabilidade *epistêmica* (*subjetiva*) de um determinado evento ou estado de coisas, por exemplo Fa , é o grau racional de crença em que um determinado sujeito ou todos os sujeitos de um determinado tipo de racionalidade acreditam que um evento ocorrerá. Formalmente, escrevemos um “maiúsculo” $P(-)$: $P(Fa)$ para representar, portanto, o grau de crença de que o indivíduo específico a possui a propriedade F . – Exemplo: O grau de nossa crença de que amanhã será um dia ensolarado em Düsseldorf.

A frequência relativa $h(Fx)$ de um tipo de evento Fx , em um domínio individual finito D , é o número de todos os F em D dividido pelo número de todos os indivíduos em D . Para domínios individuais finitos, identificamos a probabilidade estatística com a frequência relativa, ou seja, $p(Fx) = h(Fx)$. Por outro lado, se D for infinito, a frequência relativa é indefinida. Nesse caso, referimo-nos a um arranjo aleatório dos indivíduos em D na forma de uma sequência aleatória (infinita) (d_1, d_2, \dots) . A frequência relativa $h_n(Fx)$ de F 's entre os primeiros n membros da sequência é definida como $h_n(F) =_{\text{def}} a_n(F)/n$, com “ $a_n(F)$ ” como o número de F 's nos primeiros n termos da sequência. Determina-se agora a probabilidade estatística $p(Fx)$ como o limite das frequências relativas $h_n(Fx)$ de F 's em seções iniciais de n membros da sequência aleatória, para n em direção ao infinito: $p(Fx) = \lim_{n \rightarrow \infty} h_n(Fx)$. O conceito de limite é definido da seguinte forma:

(Def. 2-2) O limite da frequência relativa do tipo de evento Fx em uma dada sequência aleatória (d_1, d_2, \dots) é r ou, em resumo, $\lim_{n \rightarrow \infty} h_n(Fx) = r$, “se e somente se” [de agora em diante “sse”] para cada $\epsilon > 0$, não importa quão pequeno seja, houver um número de dígitos n , de

modo que para todo $m \geq n$ a frequência relativa $h_m(Fx)$ se desvie do limite r em menos que ϵ .

Para ilustrar isso, a convergência das frequências relativas de um evento com o limite de frequência $p(Fx) = 0,6$ em duas seqüências aleatórias é mostrada na Figura 2-1.

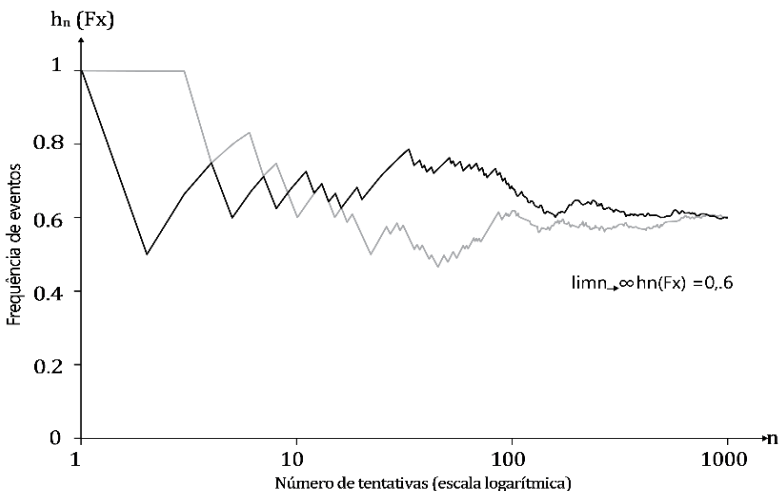


Fig. 2-1: Convergência das frequências relativas de um evento com limiar de frequência $p(Fx) = 0,6$ em duas seqüências aleatórias (programadas em Visual Basic).

Os limites de frequência são idealizações teóricas que vão além das frequências factualmente observáveis. As seqüências aleatórias são efetivadas através da realização repetida dos chamados experimentos aleatórios (mais detalhes no Capítulo 5). Os resultados dessas realizações formam os componentes das seqüências aleatórias.

Quando se fala em limites de frequência trata-se, portanto, da afirmação de que o experimento aleatório subjacente tem uma certa disposição para produzir o resultado em questão, F_x , com uma frequência que converge no longo prazo. Essa disposição também é chamada de “propensão genérica” (Seção 5.2). Falar sobre limites de frequência vai muito além, ontologicamente, das frequências reais, e é por isso que prefiro falar sobre probabilidade “estatística” – em vez de probabilidade “frequentista”, como é muitas vezes o caso no mundo de língua inglesa. O sorteio aleatório de um indivíduo de um intervalo de indivíduos D também representa um experimento aleatório. Isso resulta na seguinte conexão com frequências finitas: a frequência da característica F , em um intervalo finito D , corresponde exatamente à probabilidade estatística de escolher aleatoriamente um F no sorteio de um indivíduo de D , desde que cada indivíduo em D tenha a mesma chance estatística de ser sorteado – esse é o requisito básico para sorteios aleatórios. Formalmente falando, o seguinte deve valer para todo $d \in D$: $p(x=d) = 1/|D|$ (com “ $x=d$ ” para “o resultado do sorteio x foi d ” e “ $|D|$ ” para a cardinalidade de D).

Sempre que as probabilidades estatísticas referem-se aos resultados de experiências que podem ser repetidas qualquer número de vezes, não faz muito sentido identificá-las com as frequências finitas, uma vez que os resultados de apenas um número finito de execuções experimentais são sempre determinados pela aleatoriedade que distorce as probabilidades subjacentes. Por essa razão, a referência às tendências de frequência no limite parece ser a única solução aceitável. Somente se as probabilidades estatísticas referirem-se explicitamente a distribuições numa população finita, que não foram geradas por qualquer experiência aleatória conhecida

ou processo causal, é que a sua identificação com frequências finitas fará sentido.

Para que o conceito estatístico de probabilidade assim definido possa expressar uma propriedade objetiva da realidade, o conceito de sequência aleatória deve ser caracterizado de forma adequada e suficiente. Esse é o principal problema da teoria estatística, do qual trataremos no Capítulo 5. O principal problema da teoria da probabilidade epistêmica, por outro lado, é que diferentes sujeitos, ainda que tenham as mesmas experiências, podem atribuir diferentes graus de crença às mesmas proposições. O problema da teoria epistêmica é, portanto, racionalizar e objetivar graus subjetivos de crença, dos quais trataremos nos Capítulos 6 e 7.

Sobre a interpretação de probabilidades um e zero: no caso epistêmico, a afirmação $P(A) = 1$ significa simplesmente que um determinado sujeito tem certeza sobre a afirmação ou proposição A , ou seja, de forma alguma dúvida da verdade de A – qualquer que seja o valor de verdade factual de A . No caso estatístico, entretanto, o significado de $p(Fx) = 1$ requer uma explicação mais detalhada. Somente no caso de um domínio finito de indivíduos $p(Fx) = 1$, sse o teorema universal estrito ou sem exceções $\forall xFx$ (todos os indivíduos são F), ou $p(Fx) = 0$ equivalente a $\forall x\neg Fx$. No caso de um domínio infinito de indivíduos, entretanto, $p(Fx) = 1$ é mais fraco que $\forall xFx$ (e $p(Fx) = 0$ é mais fraco que $\forall x\neg Fx$). Afinal, dada uma sequência aleatória infinita (d_1, d_2, \dots) e um tipo de evento Fx , $p(Fx) = 0$ não implica que não haja indivíduo d_i nessa sequência que tenha a característica F , mas apenas que as frequências $h_n(Fx)$ convergem para zero. Por exemplo, seja a sequência aleatória a ordem dos números naturais \mathbb{N} , ou seja, $1, 2, \dots$, e deixe Fx denotar o predicado “ x é uma potência inteira de 2”. Existem infinitas potências inteiras de 2 entre

os números naturais, isto é, todos os números da forma 2^k para $k \in \mathbb{N}$. No entanto, $\lim_{k \rightarrow \infty} p(Fx) = \lim_{k \rightarrow \infty} (k/2^k) = 0$, porque entre os primeiros 2^k números naturais existem k potências de dois, e a razão de k para 2^k se aproxima de zero quando k se torna infinitamente grande.

A probabilidade estatística de que um número natural seja (ou não) uma potência de dois é, portanto, zero (ou um). A hipótese estatística $p(Fx) = 1$ permite qualquer número de exceções e até um número infinito de exceções, desde que sua frequência convirja apenas para zero; é, dessa maneira, significativamente mais fraca que a afirmação geral $\forall xFx$.

As probabilidades estatísticas sempre se referem a um tipo de evento ou tipo de fato repetível, expresso por um predicado ou uma fórmula aberta, por exemplo: Fx .³ A probabilidade subjetiva, por outro lado, refere-se a um evento ou fato específico, expresso por uma frase ou fórmula fechada, por exemplo: Fa . Isso porque apenas sentenças com significado definido, mas não fórmulas abertas com significado indefinido, podem ser objeto de crença e ter graus de crença. Um exemplo: se for dito que a probabilidade de chover em Düsseldorf amanhã é de $3/4$, então *prima facie* isso não pode ser uma afirmação de frequência, mas apenas uma afirmação de probabilidade epistêmica. Afinal, o amanhã só vem uma vez – ou chove amanhã ou não chove amanhã. *Prima facie*, com uma probabilidade de caso individual $P(Fa)$, apenas uma afirmação de probabilidade epistêmica pode ser feita – uma afirmação sobre, por exemplo, meu nível de crença em Fa . A probabilidade estatística correspondente $p(Fx)$, por outro lado, só pode ser atribuída a um evento repetível do tipo Fx , por

³ Uma fórmula de lógica de predicado é chamada aberta se ela contém variáveis individuais que são livres, ou seja, não são limitadas por um quantificador; caso contrário, a fórmula é considerada fechada. Veja Apêndice 10.1.

exemplo, que chova em qualquer dia x em Düsseldorf. Minha probabilidade subjetiva $P(Fa_i)$ pode variar arbitrariamente para diferentes indivíduos a_i ; o limite de frequência $p(Fx)$, por outro lado, é determinado pela classe de todos os Fs e pelo domínio D ou pelo experimento aleatório subjacente, não dependendo de nenhuma instanciamento individual Fa_i . Sintaticamente, isso significa que a função de probabilidade estatística $p(A)$ liga todas as variáveis livres na fórmula A (semelhante ao modo de um quantificador).

A expressão “ $p(Fx)$ ” não contém nenhuma variável livre; sse $p(\{x:Fx\})$ (com “ $\{x:Fx\}$ ” para “o conjunto de todos os x que são Fs ”) denotar o limite de frequência de Fx em sequências D -aleatórias (cf. Bacchus, 1990, Capítulo 3, que escreve “ $p_x(Fx)$ ” para isso). Quantificar as variáveis em probabilidades estatísticas, ou seja, $\forall x(p(Fx) = 0,5)$ [p minúscula], seria, portanto, uma confusão sintática, enquanto a quantificação universal faz sentido para probabilidades subjetivas: $\forall x(P(Fx) = 0,5)$ [P maiúscula] afirma que, para cada indivíduo d em D , o grau de crença na proposição “ d é um F ” é $0,5$.

Apesar dessas diferenças fundamentais, existem conexões entre os dois conceitos de probabilidade. O princípio mais conhecido para transferir probabilidades estatísticas para probabilidades subjetivas de casos individuais é o seguinte, que remonta a Reichenbach⁴ (1949, § 72):

⁴ Nota do tradutor: Reichenbach disse ao cunhar a frase “problema da classe de referência” que ela deveria ser “a classe mais estreita para a qual estatísticas confiáveis podem ser compiladas”, o que é correto, exceto que não se estreita uma classe de referência relevante dividindo-a de acordo com atributos irrelevantes. Hans Reichenbach, *The Theory of Probability* (1949). Em alemão, o termo usado é “*engste*” e em inglês “*narrowest*”, ambos para se referirem a classes mais estreitas, mais restritas.

(Def. 2-3) *Princípio da classe de referência mais estreita:* A probabilidade subjetiva $P(Fa)$ de um evento individual é determinada como a probabilidade estatística condicional (estimada) $p(Fx|R_x)$ do tipo de evento correspondente Fx na mais estreita classe de referência (nomológica) ou classe de referência R , da qual o sujeito subjacente sabe, ou assume com certeza, que a reside nela (ou seja, R_a se aplica).⁵

O princípio da classe de referência mais estreita é usado consistentemente na vida cotidiana e nas ciências. Por exemplo, para determinar a probabilidade subjetiva de uma determinada pessoa seguir uma determinada carreira (Fa), baseamo-nos nas características dessa pessoa que conhecemos como a classe de referência mais estreita (R_a) e na probabilidade estatística de uma pessoa x com as características R_x seguir nessa carreira ($p(Fx|R_x)$).

Na previsão meteorológica acima “a probabilidade de chover amanhã é de $3/4$ ”, a classe de referência mais estreita é a evolução meteorológica anterior considerada pelo meteorologista. Segundo o princípio de Reichenbach, essa previsão meteorológica tem a seguinte interpretação: a probabilidade estatística de chover num dia precedido por uma evolução meteorológica tipologicamente igual à de hoje é de $3/4$. É isso que os meteorologistas querem dizer quando fazem previsões meteorológicas probabilísticas.

O princípio de Reichenbach está intimamente relacionado ao seguinte princípio de inferência:

(2-1) *Inferência de especialização indutiva* (cf. Carnap, 1950, p. 207):

Premissa 1: $r\%$ de todos os Fs são Gs

⁵ A rigor, teríamos que distinguir a classe “ R ” do caractere “ R ” e definir $I(R) = R$. Veja o Apêndice 10.1.3 para semântica lógica.

Premissa 2: Este é um F

===== [com r % de probabilidade de crença]

Conclusão: Este é um G

A inferência de especialização indutiva também é chamada de “inferência direta” (Levi, 1977). O travessão duplo “==” indica que essa conclusão é incerta, ou seja, só leva a uma conclusão verdadeira com uma dada probabilidade (anotada à direita do travessão duplo) a partir de premissas verdadeiras. Tal como acontece com todas as conclusões incertas, aplica-se o princípio da evidência total: a premissa singular deve conter todas as evidências relevantes para a conclusão (cf. Schurz, 2006, p. 56, Ms. 2.6-4). Considerando essa condição adicional, o princípio da classe de referência mais estreita de Reichenbach é uma aplicação da inferência de especialização indutiva, que por sua vez tem a sua justificação mais profunda na Seção 7.1, onde se encontra a discussão sobre o princípio da coordenação estatística (mais informações sobre o raciocínio indutivo na Seção 4.1.4).

Com a ajuda do princípio da classe de referência mais estreita de Reichenbach, apenas a probabilidade subjetiva de sentenças singulares (ou seja, sentenças que contêm constantes individuais) pode ser determinada por probabilidades estatísticas, algo que não é possível em relação à probabilidade subjetiva de hipóteses gerais como: B. “Todos os corvos são pretos” (formalmente $\forall x(Fx \rightarrow Gx)$) ou “50% de todos os lançamentos de moeda dão coroa” (formalmente $p(Fx) = 0,5$). A probabilidade subjetiva das hipóteses gerais depende da sua probabilidade inicial subjetiva, que não pode ser atribuída a probabilidades estatísticas (ver Capítulo 7.3).

Explicar o princípio da classe de referência mais estreita traz consigo uma série de problemas e refinamentos, que serão explicados no Capítulo 7. O requisito na Def. 2-3 (colchetes), de que a classe de referência deve ser “nomológica”, também será explicado com mais detalhes lá.

3 Fundamentos matemáticos da probabilidade

3.1 Leis da probabilidade

Os conceitos estatísticos e epistêmicos de probabilidade obedecem às mesmas leis matemáticas básicas que foram axiomatizadas, pela primeira vez, por Kolmogorov (1933). Para formular essas leis, usamos os símbolos lógicos usuais e da teoria dos conjuntos, em particular: \neg (negação), \wedge (conjunção), \vee (disjunção), \rightarrow (implicação material), \leftrightarrow (equivalência), \forall (quantificador universal), \exists (quantificador existencial), \in (relação de elemento [ou de pertencimento a um conjunto]), \cup (união de conjunto), \cap (interseção de conjunto), $-$ (diferença de conjunto), \emptyset (conjunto vazio), \subseteq (subconjunto próprio ou impróprio), \subset (subconjunto próprio). $f:A \rightarrow B$ (f é uma função do conjunto A para o conjunto B). “ $=_{\text{def}}$ ” significa “é idêntico por definição”. Para mais informações sobre simbolismo e fundamentos da lógica e da teoria dos conjuntos, consulte o Apêndice 10.1. Frequentemente, para fins práticos, repetiremos em palavras o conteúdo de declarações formalizadas.

Kolmogorov apresentou os axiomas de probabilidade na representação algébrica matemática usual para conjuntos. Sejam A, B, \dots representações de subconjuntos (que também são chamados de

“eventos”) de um chamado espaço de possibilidades Ω . Aqui, Ω é o conjunto de todos os resultados possíveis de um experimento aleatório. Um exemplo é lançar um dado ($\Omega = \{1,2,3,4,5,6\}$) ou sortear um indivíduo de um determinado domínio de indivíduos D , que também é chamado de população ($\Omega = D$). Subconjuntos de elementos Ω são vistos como disjunções de possíveis resultados do experimento; então a quantidade $\{1,3,5\}$, ao lançar os dados, corresponde à afirmação “foi lançado 1, 3 ou 5, isto é, um número ímpar foi lançado”, etc. Na representação algébrica do conjunto, a negação $\neg A$ deve ser o complemento de A $A^c =_{\text{def}} \Omega - A$, a disjunção $A \vee B$ pode ser lida como a união $A \cup B$, e a conjunção $A \wedge B$ como a interseção $A \cap B$; em particular $A \vee \neg A = \Omega$ e $A \wedge \neg A = \emptyset$.

Assume-se que a álgebra dos subconjuntos de Ω é fechada sob essas operações (detalhes na Seção 3.3). A estrutura algébrica definida pode ser interpretada tanto estatisticamente quanto subjetivamente: no caso estatístico, $p(A)$ [p minúscula] representa o limite de frequência do evento A em uma sequência aleatória de realizações experimentais e, no caso subjetivo, $P(A)$ [P maiúscula] representa a probabilidade de crença do evento A em uma única realização do experimento.

Para a filosofia da ciência, a representação linguística da teoria da probabilidade é preferível, porque torna explícita a diferença entre eventos individuais e tipos de eventos. Aqui A, B, \dots representam fórmulas abertas se a probabilidade for entendida num sentido estatístico, e sentenças se for entendida num sentido subjetivo.

As variáveis A, B, \dots podem ser lidas abaixo ou como subconjuntos de Ω (matemática), ou como fórmulas abertas (estatísticas), ou como sentenças (epistêmicas); as leis da probabilidade são sempre as mesmas. O fato de A e B serem disjuntos

significa, em termos algébricos definidos, que $A \cap B$ é vazio; na leitura estatística de que a extensão de $A \wedge B$ é factualmente (ou seja, considerado nesse modelo como ‘factual’) vazia; e na leitura subjetiva de que $A \wedge B$ é inatingível em todos os modelos de linguagem considerados epistemicamente possíveis. O conjunto de todos os modelos possíveis (ou ‘mundos’), ou “Mod.” abreviadamente, que fundamenta a função de probabilidade subjetiva não precisa necessariamente coincidir com o conjunto de todos os modelos logicamente possíveis, mas pode ser um subconjunto deste. Abaixo escrevemos $\Box A$ (para “A é necessário”) se A for satisfeito por todos os modelos em Mod., e similarmente “ $\Diamond A$ ” para “A é possível” se A for satisfeito por pelo menos um modelo em Mod. “ \Box ” e “ \Diamond ” são os dois operadores básicos de frases modais; é dado que $\Diamond A$ sse $\neg \Box \neg A$, em palavras: A é possível sse a negação de A não é possível. No caso estatístico, “ $\Box A$ ” afirma que A é exaustivo e “ $\Diamond A$ ” afirma que a extensão de A não é vazia.

A quantificação universal “para todo A” refere-se: na leitura algébrica de conjuntos, a todos os subconjuntos de Ω contidos na álgebra de conjuntos (Seção 3.3); na leitura estatística, a todas as fórmulas abertas; na leitura subjetivo-epistêmica, a todas as sentenças da linguagem subjacente.

(Def. 3-1) Axiomas básicos de probabilidade

Para todos os A, B, ... , onde houver “p” também pode haver “P”:

(A1) $p(A) \geq 0$ (não-negatividade)

Em palavras: As probabilidades são sempre maiores ou iguais a zero.

(A2) $p(A \vee \neg A) = 1$ (padronização para 1)

Em palavras: A probabilidade do espaço de possibilidade total é 1.

(A₃) Se A, B são disjuntos: $p(A \vee B) = p(A) + p(B)$ (aditividade-finita)
 Em palavras: Para eventos disjuntos (tipos), as probabilidades se somam.

Uma função que satisfaz os axiomas (A₁₋₃) é chamada de função de probabilidade (Kolmogoroviana).

Os axiomas básicos de probabilidade resultam em vários Teoremas, dos quais os mais importantes são mencionados no Teorema 3.1. Uma fórmula A em n variáveis livres é chamada exaustiva, no caso estatístico, se A, no modelo dado, é satisfeito por todas as n-tuplas possíveis de indivíduos em D. No caso subjetivo, uma sentença A é chamada exaustiva se torna verdadeiros todos os modelos epistemicamente possíveis de linguagem A. Finalmente, na leitura algébrica do conjunto, A é exaustivo se A coincide com todo o espaço de possibilidades Ω . Uma sequência de n pares disjuntos A_i ($1 \leq i \leq n$) é chamada de partição ou decomposição de Ω se a disjunção $A_1 \vee \dots \vee A_n$ é exaustiva. Todos os Teoremas listados abaixo se aplicam – salvo indicação em contrário – a todos os A, B, ..., e “P” pode ser usado no lugar de “p”: nos pouparemos dessas adições a seguir.

(Teorema 3-1) *Teoremas de probabilidade incondicional* (Apêndice de Prova 10.3.1)

(T₁) $p(\neg A) = 1 - p(A)$ (probabilidade complementar)
Em palavras: a probabilidade de negação de um evento é 1 menos a do evento.

(T₂) $p(A) \leq 1$ (limite superior)
Em palavras: A probabilidade de cada evento é menor ou igual a 1.

(T₃) $p(A \wedge \neg A) = 0$ (contradição)
Em palavras: Uma contradição tem probabilidade zero.

(T4) Para cada partição A_1, \dots, A_n : $\sum_{1 \leq i \leq n} p(A_i) = 1$ e $p(B) = \sum_{1 \leq i \leq n} p(B \wedge A_i)$

Em palavras: A soma das probabilidades dos eventos de uma partição (A_i : $1 \leq i \leq n$) de Ω é igual a 1, e os eventos ($A_i \wedge B$: $1 \leq i \leq n$) formam uma partição de B cujas probabilidades se adicionam a $p(B)$.

(T5) $p(A_1 \vee A_2) = p(A_1) + p(A_2) - p(A_1 \wedge A_2)$ (lei geral da adição)

(T6) Se $A_1 \rightarrow A_2 =_{\text{def}} \neg A_1 \vee A_2$ é exaustivo, então $p(A_1) \leq p(A_2)$ (monotonicidade)

Em palavras: Se A_1 implica necessariamente A_2 , então a probabilidade de A_1 é menor ou igual à de A_2 .

(T7) Se $A_1 \leftrightarrow A_2$ for exaustivo, então $p(A_1) = p(A_2)$ (equivalência)

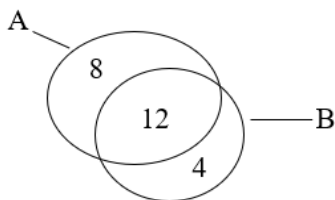
A probabilidade de A sob a suposição de que B está presente é chamada de probabilidade condicional de A dado B. Isso é escrito $p(A|B)$ ou $P(A|B)$ e, geralmente, define-se essa expressão da seguinte forma:

(Def. 3-2) *Probabilidade condicional:* $p(A|B) =_{\text{def}} \frac{p(A \wedge B)}{p(B)}$, desde que

$p(B) > 0$.

(Analogamente para “P” em vez de “p”.)

Em $p(A|B)$, B é chamado de evento condicional ou antecedente, e A é chamado de evento condicionado ou conseqüente. No caso estatístico finito, $p(A|B)$ coincide com a frequência relativa de A no conjunto B, que por sua vez é um subconjunto do domínio [Domínio] D – veja a Figura 3-1. No caso estatístico infinito, $p(A|B)$ coincide com o limite de frequência de A's em uma sequência aleatória de indivíduos-B.



$$\begin{aligned}
 p(Bx|Ax) &= 12/20 = 3/5 \\
 p(Ax|Bx) &= 12/16 = 3/4 \\
 |D| &= 24, \quad p(Ax) = 20/24 = 5/6 \\
 p(Bx) &= 16/24 = 2/3
 \end{aligned}$$

Figura 3-1: Probabilidades estatísticas condicionais

No caso epistêmico subjetivo, $P(A|B)$ representa o grau hipotético de crença em A, sob a suposição de que B é certo, ou seja, $P(B) = 1$. Se B for realmente acreditado com certeza, então $P(B) = 1$, então $P(A) = P(A|B)$. Para evitar mal-entendidos, deve-se notar que $P(A) = 1$, isto é, a certeza subjetiva em relação a A não implica nem lógica nem analiticamente que A é verdadeira: a certeza subjetiva é falível e a função de crença P é, portanto, independente da função de valor de verdade I (ver Apêndice 10.1.3 para semântica lógica). Claro, é possível fazer suposições adicionais a esse respeito: por exemplo, às vezes se assume que o assunto em questão é infalível pelo menos no que diz respeito à evidência empírica (experiência) E, de modo que aqui $p(E) = 1$ também implica $I(E) = 1$. Contudo, tais suposições vão além da teoria das probabilidades.

A definição usual de $p(A|B)$ tem a desvantagem de que $p(A|B)$ não é definido para um evento B com probabilidade 0. Não apenas eventos impossíveis têm probabilidade zero. Eventos contingentes também podem ter probabilidade zero, como o exemplo já mencionado sobre a probabilidade de extrair aleatoriamente um número de um número infinito de números naturais, que é a potência de 2 de um número natural. Portanto, foram desenvolvidos métodos

para axiomatizar diretamente a probabilidade condicional – em vez de a definir pela probabilidade incondicional – com o objetivo de poder estendê-la a eventos antecedentes contingentes com probabilidade zero. Carnap (1971, p. 38) propôs o seguinte sistema de axiomas:

(Def. 3-3) Axiomatização direta da probabilidade condicional:
 Suposição: Os eventos antecedentes (que ocorrem nos axiomas) são não vazios ou possíveis:

(B1) $p(A|B) \geq 0$ (não negatividade)

(B2) $p(A \vee B|B) = 1$ (conclusão)

(B3) Para A e B disjuntos: $p(A \vee B|C) = p(A|C) + p(B|C)$ (aditividade finita)

(B4) $p(A \wedge B|C) = p(B|C) \cdot p(A|B \wedge C)$ (princípio geral de multiplicação)

Se o evento antecedente for impossível, então a probabilidade condicional também é indefinida em sua axiomatização direta. Afinal, se fosse definida, resultaria na seguinte contradição: $p(A|B \wedge \neg B) = p(\neg A|B \wedge \neg B) = 1$, porque qualquer coisa segue logicamente de uma contradição $B \wedge \neg B$, isto é, A e $\neg A$. Ademais, isso resultaria em $p(A \vee \neg A | \perp) = 2$, em contradição com o Teorema 3-1 (T2).

Outra proposta para a axiomatização direta de probabilidades condicionais, apresentada na Seção 4.2, vem de Popper (1935).

Probabilidades condicionais diretamente axiomatizadas são usadas para definir probabilidades incondicionais, condicionando as primeiras a tautologias. Portanto, define-se $p(A) =_{\text{def}} p(A|\tau)$ (com “ τ ” para uma tautologia, por exemplo, $B \vee \neg B$).

Carnap provou o seguinte Teorema, que garante que a probabilidade condicional diretamente axiomatizada concorda com aquela definida convencionalmente (de acordo com a Def. 3-2):

(Teorema 3-2) Seja p uma função de probabilidade incondicional axiomatizada por A_1 -3 [na Def. 3-1], e seja p^* uma função de probabilidade condicional axiomatizada por B_1 -3 [na Def. 3-3]. Então, para todo A e B não vazio:
 $p(B \wedge A) = p^*(B|A) \cdot p(A)$, ou seja, p e p^* concordam sse $p(A) = p^*(A|\tau)$.

A prova do Teorema 3-2 pode ser encontrada em Carnap (1971, p. 41, T1-5). De $p(B \wedge A) = p^*(B|A) \cdot p(A)$ obtemos $p(B|A) =_{\text{def}} p(B \wedge A)/p(A) =_{\text{def}} p^*(B|A)$, desde que $p(A) > 0$. Os dois termos da probabilidade condicional concordam, desde que a probabilidade antecedente seja maior que zero. Por uma questão de simplicidade, referimo-nos à definição padrão (Def. 3-2) com probabilidades condicionais, salvo indicação em contrário.

Para muitos propósitos, é necessário o conceito de (in)dependência probabilística:

(Def. 3-4) Dois eventos A , B são chamados probabilisticamente independentes um do outro, abreviados como $A \perp B$, sse $p(A \wedge B) = p(A) \cdot p(B)$.

A definição de independência probabilística difere ligeiramente, dependendo se se adota a definição padrão $p(-|-)$ ou a axiomatização direta $p^*(-|-)$ de probabilidade condicional. Nesse último caso,

escrevemos $A \perp B$, abreviado para $p^*(A \wedge B) = p^*(A) \cdot p^*(B)$. Como se pode facilmente ver, aplica-se o seguinte (apêndice de prova 10.3.2):

(3-1) $A \perp B$ sse $p(A|B) = p(A)$ ou $p(B) = 0$ (em palavras: [A é independente de B] sse a suposição de B não altera a probabilidade de A, ou se a probabilidade de B é zero); [e também] sse $p(B|A) = p(B)$ ou $p(A) = 0$.

(3-2) $A \perp B$ sse $p^*(A|B) = p^*(A)$ ou $\square \neg B$ (em palavras: [A é independente de B] sse a suposição de B não altera a probabilidade de A, ou se B é impossível); [e também] sse $p(B|A) = p(B)$ ou $\square \neg A$.

De acordo com essas equações, dois eventos prováveis ou possíveis diferentes de zero A, B são probabilisticamente dependentes sse $p(A|B) \neq p(A)$ se aplica, ou seja, se a presença de um evento altera a probabilidade do outro. Em particular, A, B são considerados positivamente dependentes se $p(A|B) > p(A)$ (ou $p(A \wedge B) > p(A) \cdot p(B)$) e negativamente dependentes se $p(A|B) < p(A)$ (ou $p(A \wedge B) < p(A) \cdot p(B)$).

O importante é a não monotonicidade das probabilidades condicionais: um valor alto de $p(A|B)$ não implica de forma alguma um valor alto de $p(A|B \wedge C)$; em vez disso, $p(A|B \wedge C) = 0$ pode ser aplicado ao mesmo tempo. A Figura 3-2 mostra esse exemplo.

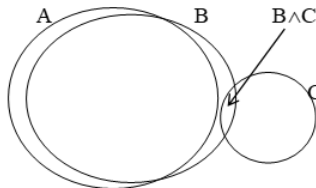


Figura 3-2: Não monotonicidade das probabilidades condicionais: $p(A|B)$ é alta, mas $p(A|B \wedge C)$ é zero.

Existem vários Teoremas derivados para probabilidades condicionais, os mais importantes são mencionados aqui:

(Teorema 3-3) *Teoremas de probabilidade condicional* (Apêndice de Prova 10.3.3):

Suposição: Para as fórmulas X na posição antecedente, $p(X) > 0$ é assumido (mais especificamente com axiomatização direta de acordo com o Teorema 3-2 $\diamond X$)

(TB1): Para a função de probabilidade $p_B(A) \stackrel{\text{def}}{=} p(A|B)$ condicionada em B , todas as leis da probabilidade incondicional aplicam-se.

(TB2:) Se $A \rightarrow B$ é exaustivo, então $p(B|A) = 1$. O inverso não é válido.

(TB3) $p(A \wedge B) = p(A|B) \cdot p(B)$

(TB4) Para cada partição B_1, \dots, B_n , aplica-se o seguinte: $p(A) = \sum_{1 \leq i \leq n} p(A|B_i) \cdot p(B_i)$ (princípio geral de multiplicação). Especificamente, $p(A) = p(A|B) \cdot p(B) + p(A|\neg B) \cdot (1-p(B))$

(TB5) $p(A|B) = p(B|A) \cdot p(A)/p(B)$ (Teorema de Bayes, 1ª versão)

(TB6) Para cada partição A_1, \dots, A_n , aplica-se o seguinte: $p(A_i|B) = p(B|A_i) \cdot p(A_i) / \sum_{1 \leq i \leq n} p(B|A_i) \cdot p(A_i)$ (Teorema de Bayes, 2ª versão)

(TB7) Simetria da dependência probabilística (desde que $1 > p(B)$, $p(A) > 0$):

$p(A|B) > p(A)$ sse $p(B|A) > p(B)$ sse $p(A|B) > p(A|\neg B)$ (análogo para \geq)

(TB1) é fundamental para o princípio Bayesiano de condicionalização (ver Capítulo 7.4). (TB2) nos mostra o já mencionado: que uma implicação estrita (sem exceção) implica uma probabilidade condicional de 1, mas não vice-versa. (TB3) é elementar. (TB4) inclui o importante caso especial $p(A) = p(A|B) \cdot p(B) + p(A|\neg B)$

· $p(B)$, mostrando que $p(A)$ é uma média ponderada de $p(A|B)$ e $p(A|\neg B)$, com os pesos $p(B)$ e $p(\neg B)$ (que somam 1). Portanto, em termos de tamanho, $p(A)$ deve estar estritamente entre os dois valores $p(A|B)$ e $p(A|\neg B)$.

A importância dos dois Teoremas bayesianos (TB5) e (TB6) é notada nas situações em que se está interessado em $P(A_i|B)$, mas apenas a probabilidade condicional inversa $P(B|A_i)$ é praticamente acessível. Esse é o caso, por exemplo, de quando A_i são hipóteses rivais e B é resultado de uma amostra empírica ou de um conjunto de dados – uma aplicação importante da teoria bayesiana das probabilidades. Outro caso são os problemas de diagnóstico, nos quais B desempenha o papel de um indicador para uma condição A a ser medida. Por exemplo, B poderia representar um resultado positivo para teste de câncer e A para câncer de fato. A única coisa que pode ser facilmente medida, de modo experimental, é a probabilidade de resultado de um indicador, dado que a doença A está presente ou não. Nesse contexto, $p(B|A)$ também é chamado de sensibilidade e $p(\neg B|\neg A)$ de especificidade do indicador B para A . Para fins prognósticos, interessa a probabilidade inversa de câncer, dado um resultado de um indicador, ou seja, para as probabilidades $p(A|B)$ e $p(\neg A|\neg B)$; esses valores também são chamados de confiabilidade ou eficiência do indicador como instrumento de previsão (cf. Sachs, 1992, p. 84-8). Com o Teorema de Bayes, a confiabilidade e a eficiência de um indicador podem ser calculadas a partir de sua sensibilidade, especificidade e probabilidade inicial $p(A)$ na população.

Assim, o importante é a simetria das dependências probabilísticas, expressa em (TB7): se A aumenta a probabilidade de B , então B também aumenta a probabilidade de A . Em contraste, as relações causais são fundamentalmente assimétricas – mostrando que

a conclusão de dependências probabilísticas para causais não pode ter validade geral (ver Capítulo 8.7).

3.2 Distribuição binomial e a lei dos grandes números

Um experimento aleatório é um processo repetível que leva a um dos vários resultados possíveis a cada vez, mas nem sempre ao mesmo resultado.

Exemplos são o lançamento de um dado e o sorteio aleatório de um indivíduo [ou de um objeto individual] de um domínio de indivíduos [ou de objetos individuais] ('urna'), mas também a ocorrência diária ou não de certos eventos, tais como a chuva. Os resultados possíveis de uma experiência aleatória não têm de ser completamente "aleatórios" no sentido de serem igualmente prováveis. Basta haver vários resultados possíveis para que se possa dizer que o acaso também está envolvido.

Repetições independentes dos mesmos experimentos aleatórios (ou 'idênticos') são particularmente importantes para a teoria estatística da probabilidade. Isso se refere à execução sucessiva do mesmo experimento aleatório, no qual as execuções individuais são fisicamente e, portanto, probabilisticamente independentes umas das outras. Exemplos seriam os resultados de n lançamentos de moeda (e_1, \dots, e_n) , onde e_i significa "cara" ou "coroa", ou seja, $e_i \in \{\text{cara, coroa}\}$. Linguisticamente, esses resultados podem ser representados como n conjunções $E_1x_1 \wedge \dots \wedge E_nx_n$; " E_ix_i " significa "cara(x_i)" ou "coroa(x_i)", e a variável individual x_i refere-se ao resultado da i -ésima execução do experimento aleatório. Note-se que, no entanto, os índices de variáveis

individuais não têm função de designação, mas servem apenas para diferenciar: Fx_1 e Fx_2 designam, portanto, o mesmo tipo de evento F com a extensão $\{x:Fx\}$ e a probabilidade $p(Fx_i) = p(Fx)$. De modo geral, uma fórmula aberta com n variáveis individuais distintas descreve a combinação de resultados de um experimento aleatório realizado n vezes. Concorde-se que a i -ésima variável individual, disposta da esquerda para a direita na fórmula, refere-se ao resultado da i -ésima experiência (para mais informações, ver apêndice 10.2). Isso resulta na lei da permutação, segundo a qual a probabilidade estatística de um tipo de evento é invariante quando suas variáveis individuais são trocadas ou permutadas, quer dizer, por exemplo, $p(A(x_1, x_2, x_3)) = p(A(x_2, x_1, x_3)) = p(A(x_3, x_1, x_2))$ (etc.). Por outro lado, $p(A(x_1, x_2)) = p(A(x_2, x_1))$.

Do ponto de vista filosófico, independência significa que a experiência aleatória não muda as suas disposições ao longo de execuções repetidas.⁶ A probabilidade estatística de lançar um número com uma moeda normal, que não se desgasta, não depende do que foi lançado em lançamentos anteriores. De forma mais geral, $p(E_2 x_2 | E_1 x_1) = p(E_2 x)$ ou, *em palavras*, a probabilidade de alcançar o resultado E_2 de um experimento aleatório não muda porque E_1 foi alcançado em outro experimento. Mesmo que cara fosse lançada dez vezes, a probabilidade estatística de sair coroa após tal série ainda é $\frac{1}{2}$. Essa é a base da chamada *impossibilidade de sistemas de jogo* em jogos aleatórios. Outro exemplo é o sorteio aleatório de indivíduos de uma “urna”. É fundamental que os indivíduos sejam recolocados depois de

⁶ Se a distribuição de probabilidade estatística mudar após a realização do experimento aleatório, a condição de independência será violada. Os resultados desses experimentos que dependem do próprio passado são chamados de cadeias de Markov.

sorteados; caso contrário, a distribuição de probabilidade muda e as repetições não são independentes. A lei de independência estatística decorre do pressuposto de independência física:

(3-3) *Lei de independência estatística para combinações de eventos:*
 $Fx_1 \perp Gx_2$, ou seja, $p(Fx_1 \wedge Gx_2) = p(Fx) \cdot p(Gx)$; isso também é chamado de *lei do produto*.⁷

Em palavras: A probabilidade estatística de alcançar F e depois G, em duas implementações do mesmo experimento aleatório, é igual ao produto das probabilidades de alcançar F e G, respectivamente, em uma única implementação.

Segue-se: $p(Gx_2 | Fx_1) = p(Gx)$ e $p(Fx_1 | Gx_2) = p(Fx)$.

Por exemplo, a probabilidade de lançar um número par e depois um seis em dois lançamentos de dados é $p(\text{Número Par}(x_1) \wedge \text{Seis}(x_2)) = p(\text{Número Par}(x)) \cdot p(\text{Seis}(x)) = (1/2) \cdot (1/6) = 1/12$. A probabilidade de lançar um seis e um número par em dois lançamentos, em qualquer ordem, é exatamente o dobro disso, ou seja, $1/6$, porque existem duas maneiras disjuntas de obter esse resultado: primeiro um número par e depois um seis, ou vice-versa. Em vez de um após o outro, os dois experimentos aleatórios também podem ser realizados simultaneamente, por exemplo, usando dois dados semelhantes lançados ao mesmo tempo: então $p(Fx_1 \wedge Gx_2)$ denota a probabilidade estatística de obter o resultado F com o dado 1 e o resultado G com o dado 2.

⁷ Algebricamente, em vez de " $p(Fx_1 \wedge Gx_2) = p(Fx) \cdot p(Gx)$ ", " $p(F_1, G_2) = p(F_1) \cdot p(G_2)$ ". Assim, o " \wedge " é substituído por uma vírgula e as variáveis individuais x_i pelos índices i , que se referem às álgebras correspondentes às execuções experimentais individuais.

A lei da independência (ou lei do produto) geralmente não se aplica às probabilidades subjetivas de eventos combinados. Pelo contrário, uma vez que a medida epistêmica da probabilidade é indutiva, o nosso grau de crença de que o próximo indivíduo é um F deve aumentar com a frequência de indivíduos F previamente observados. Portanto, $P(Fa|Fb) > P(Fa)$ e, portanto, $P(Fa \wedge Fb) > P(Fa) \cdot P(Fb)$ devem ser aplicados, o que contradiz a lei do produto. Essa diferença pode ser explicada da seguinte forma: na teoria subjetiva da probabilidade, presume-se que não se conhece a probabilidade estatística. Por exemplo, não é certo se uma determinada moeda seria uma moeda simétrica ($p = 1/2$) ou uma moeda assimétrica com tendência, tal como uma moeda magnetizada com $p(\text{número}) = 1/3$. Nesse último caso, faz sentido indutivamente concluir, a partir da ocorrência frequente de cara, que é mais provável que a moeda dê cara do que coroa. Na teoria da probabilidade estatística, por outro lado, não falamos sobre o nosso grau de crença em relação a uma probabilidade estatística desconhecida, mas sim sobre a probabilidade em si, que se supõe ser dada ou “conhecida”.

Devido à suposição de independência física, a lei do produto aplica-se à segunda situação. Ou seja, se for verdade que a moeda dá cara com frequência relativa r , então isso acontece independentemente dos lançamentos anteriores da moeda; então, por exemplo, pode-se concluir que, se a moeda for lançada duas vezes, ela dará cara com frequência relativa r^2 , etc. Essa consideração mostra que existem diferenças profundas entre probabilidades objetivas e subjetivas.

A conhecida lei binomial (ou lei de Bernoulli), para extrair amostras aleatórias de n elementos ou realizar um experimento aleatório n vezes, é derivada da lei estatística do produto. Seja $p =_{\text{def}}$

$p(Fx)$ e seja $h_n(Fx)$ a denotar a frequência relativa de um evento Fx em uma amostra aleatória de n elementos, então:

$$(3-4) \text{ F\u00f3rmula binomial: } p(h_n(Fx) = \frac{k}{n}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Aqui $\binom{n}{k}$ (“n sobre k”) \u00e9 definido como $\frac{n!}{k! \cdot (n-k)!}$, e $k!$ (“k fatorial”) como $1 \cdot 2 \cdot \dots \cdot (k-1) \cdot k$. Como sabemos, $\binom{n}{k}$ \u00e9 o n\u00famero de possibilidades para selecionar k entre n indiv\u00edduos. Isso explica rapidamente a f\u00f3rmula binomial: cada sele\u00e7\u00e3o espec\u00edfica de k entre n indiv\u00edduos com a propriedade F , e os indiv\u00edduos restantes $\neg F$, tem a probabilidade $p^k \cdot (1-p)^{n-k}$, de acordo com a lei do produto. Como existem exatamente $\binom{n}{k}$ possibilidades, h\u00e1 como resultado a f\u00f3rmula binomial (3-4).

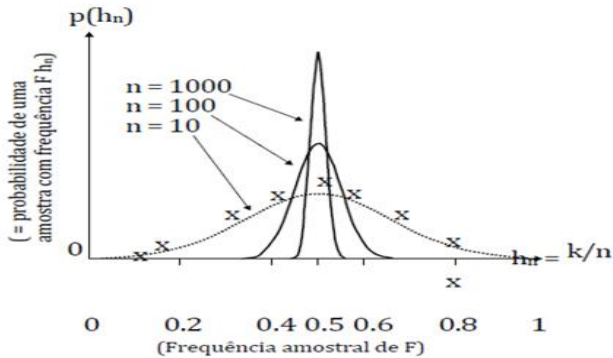


Fig. 3-3: Tr\u00eas distribui\u00e7\u00f5es binomiais $p(h_n=k/n)$ para $p = 1/2$ (aproximadas por b\u00f4i\u00e7\u00f5es normais).

A distribui\u00e7\u00e3o binomial \u00e9 mostrada na Fig. 3-3 ao longo do intervalo unit\u00e1rio $[0,1]$ (ela assume apenas valores diferentes de zero para n\u00fameros $r = k/n$). Obviamente, a distribui\u00e7\u00e3o torna-se cada vez mais acentuada \u00e0 medida que o tamanho das amostras n aumenta. O desvio esperado da frequ\u00eancia da amostra em rela\u00e7\u00e3o \u00e0 probabilidade

na população tende, portanto, a se tornar cada vez menor. Para $n \rightarrow \infty$, a distribuição binomial tende a uma distribuição gaussiana contínua infinitamente íngreme, cujos valores $p(h)$ tendem a zero para $h \neq p$ e a 1 para $h=p$ (Hays;Winkler, 1970, p. 222, 609). Isso resulta nas duas leis dos grandes números (Bauer, 1978, Capítulos 34, 38; Howson;Urbach, 1996, p. 47):

(Teorema 3-4) *Leis dos grandes números:*

(3-4.1) *Lei fraca* dos números grandes: Para cada número positivo ϵ , não importa quão pequeno seja, a probabilidade de $h_n(F)$ se desviar de $p(F)$ em menos que ϵ tende para 1, para n em direção ao infinito. Formalmente: $\forall \epsilon > 0: \lim_{n \rightarrow \infty} p(|h_n(F) - p(F)| < \epsilon) = 1$

(3-4.2) *Lei forte* dos números grandes: A probabilidade de que o limite de frequência de F , em uma sequência aleatória infinita, coincida com a probabilidade de F é 1. Formalmente: $p(\lim_{n \rightarrow \infty} h_n(Fx) = p(Fx)) = 1$

A lei forte fala diretamente sobre a probabilidade de (classes de) sequências aleatórias infinitas. A lei fraca, por outro lado, trata apenas da probabilidade de sequências aleatórias finitas arbitrariamente longas e de seu limite. A lei forte implica a lei fraca, mas não vice-versa. A lei fraca decorre do fato de que o desvio padrão de uma distribuição binomial é $p \cdot (1-p) / \sqrt{n}$ (ver Capítulo 8.6) e, portanto, tende a zero quando $n \rightarrow \infty$. Provar a lei forte requer a suposição mais forte da aditividade σ , que será discutida na próxima Seção. Para a prova das leis dos grandes números, ver Bauer (1978, § 19, p. 36-38) ou Stegmüller (1973b, p. 191 e seg).

Embora estejamos intuitivamente inclinados a ver as leis dos grandes números como uma “confirmação” da teoria estatística das

probabilidades, isso, como veremos na Seção 5.2, não é necessariamente o caso. Em primeiro lugar, essas leis são Teoremas formais, o que pode ser visto no fato de a convergência de frequências contra a probabilidade só se aplicar com uma probabilidade de 1, e isso significa coisas diferentes dependendo de como a “probabilidade” é interpretada. Se as probabilidades forem interpretadas subjetivamente, a lei forte afirma que se acredita com certeza subjetiva ($P = 1$) que o limite de frequência do evento do tipo F, em uma sequência infinita de eventos igualmente prováveis e mutuamente independentes (\neg) Fa_i (ou seja, $P(Fa_i) = P(Fa_j)$ e $P(Fa_i|Fa_j) = P(Fa_i)$ para todo $i \neq j \in \mathbb{N}$), corresponde à probabilidade de crença $P(Fa_i)$.

3.3 Estruturas formais da teoria das probabilidades

Existem duas maneiras formais fundamentalmente diferentes de construir funções de probabilidade: a construção *matemática*, que atribui probabilidades aos elementos de uma álgebra, ignorando a diferença entre probabilidade estatística e epistêmica, e a construção linguística, que atribui funções de probabilidade a fórmulas abertas no caso estatístico e a sentenças no caso epistêmico. A construção linguística divide-se em um método de construção *semântico* e um puramente *sintático*. Esses métodos de construção serão agora explicados.

(1.) *Matematicamente*, os modelos de probabilidade são definidos como triplos da forma (Ω, AL, p) (ver Bauer, 1878, Capítulo 1; Billingsley, 1995, Capítulo 2). Aqui, Ω é o *espaço de possibilidades* ou o

espaço de resultados, formalmente representado por um conjunto não vazio de resultados possíveis, e AL é a chamada *álgebra* sobre Ω , ou seja, um conjunto de subconjuntos Ω que é fechado em relação à formação de complemento, união e interseção. Isto é, com $A, B \in AL$, $A^c =_{\text{def}} \Omega - A$, B^c , assim como $A \cup B$ e $A \cap B$ também estão contidos em AL . Conforme mencionado, os elementos da AL são lidos como *disjunções* de possibilidades. Por exemplo, “ $\{1,3\}$ ” representa um resultado de dado em que foi lançado 1 ou 3 e tem uma probabilidade de $1/3$ (etc.). Se o espaço de possibilidades for finito ou contável, o conjunto de potências $\text{Pot}(\Omega)$ é geralmente escolhido como álgebra, ou seja, o conjunto de todos os subconjuntos de Ω (a álgebra provavelmente maior sobre Ω). Por outro lado, se Ω for incontável – por exemplo, o conjunto \mathbb{R} de todos os números reais em um intervalo numérico –, então $\text{Pot}(\Omega)$ é inadequado, porque nem todos os subconjuntos de \mathbb{R} podem receber medidas de probabilidade significativas. Nesse caso, escolhe-se a álgebra de Borel-Lebesgue (ver Capítulo 8.6). A seguir, também chamaremos os elementos de Ω de resultados ou eventos “completos” e os de AL de “disjuntivos”. A função de probabilidade $p:AL \rightarrow [0,1]$ atribui a cada elemento da álgebra AL um valor de probabilidade com valor real no intervalo fechado de números reais entre 0 e 1 ($[0,1] =_{\text{def}} \{r \in \mathbb{R}: 0 \leq r \leq 1\}$).

Dependendo se os elementos de AL são entendidos como eventos ou estados de coisas específicos ou como tipos repetíveis de eventos ou estados de coisas, essa medida de probabilidade é epistêmica (P) ou estatística (p).

(2.) Na estrutura *semântica da linguagem*, assume-se uma linguagem interpretada (“linguagem”) \mathcal{L} com interpretações (D,I) ; onde D é o Domínio e I é a função de interpretação que atribui a expressões da linguagem \mathcal{L} a sua extensão (ou seja, constantes

individuais, indivíduos em D , subconjuntos de predicados de n lugares de D^n ; mais detalhes no apêndice 10.1).

(2.1) A estrutura *linguística semântico-estatística* (cf. Adams, 1974; Bacchus 1990, Capítulo 3; Schurz; Leitgeb, 2008, § 6) refere-se a uma interpretação *específica* ou a um modelo específico $M = (D, I)$ da língua, que pretende refletir o mundo real. Por uma questão de simplicidade, consideramos primeiro fórmulas em *apenas uma* variável individual x . A probabilidade estatística de fórmulas com diversas variáveis individuais leva à consideração de espaços de produtos, que serão apresentados a seguir. O Domínio D funciona aqui como um espaço de possibilidades, $\Omega = D$, cujos elementos são vistos como resultados de um sorteio aleatório ou de um experimento aleatório. Como álgebra AL , escolhe-se um conjunto de álgebra sobre D que contém todas as extensões das fórmulas abertas $A(x)$ na variável individual x (o conjunto dessas extensões forma uma álgebra, porque o conjunto de todas as fórmulas x é fechado sob operações lógicas proposicionais; veja o Teorema 3-5). A medida de probabilidade $p:AL \rightarrow [0,1]$ é escolhida de modo que, para todo $A \in AL$ $p(A)$, coincida com a frequência de A s em D e, no caso de um D s infinito com o limite de frequência de um A , resulte em uma dada sequência aleatória de D indivíduos. Uma sequência aleatória é uma sequência infinita de “desenhos aleatórios” de indivíduos D com reposição, ou seja, todos os indivíduos são sorteados com o mesmo limite de frequência ($p(\{d_i\}) = p(\{d_j\})$ para todos $d_i, d_j \in D$). Somente nessa última ocorrência, os sorteios com substituição fornecem uma medida estatística que corresponde à frequência finita no caso de um D finito.⁸ Essa medida

⁸ Se as probabilidades $p(\{d_i\})$ forem diferentes para diferentes $d_i \in D$, então este não é um sorteio aleatório, mas um experimento de sorteio com viés. Neste caso, $p(Fx)$ não coincide com a frequência de F para intervalos finitos de indivíduos. Outra

é transferida para fórmulas agora abertas, comparando a probabilidade de uma fórmula $A(x)$ com a probabilidade de sua extensão e sendo identificada em D , que é referido abaixo como $\|A(x)\|^D$ (definição exata no Apêndice 10.2). Então define-se $p(A(x)) =_{\text{def}} p(\|A(x)\|^D)$, e, assim, a probabilidade estatística de todas as fórmulas abertas, na linguagem \mathcal{L} , é definida em apenas uma variável x . Observe que isso também define a probabilidade estatística de sortear um único indivíduo em D , denominado a_i , como a probabilidade da fórmula “ $x=a_i$ ”, onde o seguinte se aplica para probabilidades estatísticas: $p(x=a_i) = 1/|D|$ (ver nota de rodapé número 7). Para faixas infinitas de indivíduos, o limite de frequência $p(x=a_i)$ é zero, o que leva à violação da aditividade σ , descrita na Seção 3.4.

(2.2) Na estrutura semântico-epistêmica da linguagem (por exemplo, Carnap, 1971; 1980; Kutschera, 1972, p. 124 e seg.; Bacchus, 1990, Capítulo 2), escolhe-se como espaço de possibilidade Ω o conjunto Mod. de todas as interpretações ou modelos da linguagem que se considera epistemicamente possível, e como Álgebra AL sobre Mod. o conjunto dos subconjuntos Mod. que são conjuntos modelo de sentenças da linguagem subjacente \mathcal{L} . A seguir, $\|A\|$ denota o conjunto de \mathcal{L} -modelos que verificam o Teorema A, ou formalmente expresso $A =_{\text{def}} \{(D,I) \in \text{Mod.} : (D,I) \models A\}$.

O conjunto de modelos é chamado $\|A\|$, assim como a proposição denotada pela sentença “A”. A álgebra AL é, portanto, definida como o conjunto de proposições $\{\|A\| : A \in \text{Enviado}(\mathcal{L})\}$.

maneira de definir probabilidades estatísticas é desenhar sem substituição. Aqui a probabilidade $p(\{d_i\})$ é por definição para cada indivíduo $d_i \in D$ igual a $1/|D|$, com $|D|$ como a cardinalidade de D (Schurz/Leitgeb 2008, §3). Aqui também, no caso $|D| = \infty$ requer que os sorteios sejam aleatórios, caso contrário, este é o caso da seção. 5.2. O problema de reordenação de sequências aleatórias discutido na seção 5.2 é introduzido.

Devido ao fechamento do conjunto de todos os conjuntos de \mathcal{L} sob operações proposicionais, o conjunto de conjuntos de modelos definidos dessa forma é uma álgebra. Em AL, assume-se uma função de probabilidade $P: \text{Mod} \rightarrow [0,1]$, que se transfere para as sentenças de \mathcal{L} identificando a probabilidade da sentença A com a probabilidade de sua classe de modelo ou proposição: $P(A) =_{\text{def}} P(\|A\|)$.

A diferença semanticamente fundamental entre as probabilidades estatística e epistêmica é a seguinte: enquanto a probabilidade estatística é uma propriedade do mundo real e, portanto, se refere a um modelo específico (D, I) postulado como 'real' (ou vigente), a probabilidade epistêmica é algo sobre nossos graus de crença e, portanto, refere-se a todo o espaço de modelos epistemicamente possíveis Mod . O que ambas as versões da estrutura semântica linguística têm em comum é: a função de probabilidade é, primeiro, definida sobre a álgebra gerada linguisticamente (via D ou via Mod .) e, depois, é transferida para fórmulas ou frases. Essa transferência baseia-se na conhecida conexão entre operações lógicas e teóricas de conjuntos, o que garante que as classes de equivalência lógica de fórmulas/Teoremas, na álgebra de suas extensões/conjuntos de modelos, tenham uma imagem isomórfica:⁹

(Teorema 3-5) *Operações lógicas e algébricas de conjuntos:*

(a) A negação de uma fórmula ou frase corresponde ao complemento da extensão ou conjunto de modelos correspondente:

$$\|\neg A(x)\|^D = D - \|A(x)\|^D \text{ bzw. } \|\neg A\| = \text{Mod} - \|A\|.$$

⁹ Cada conjunto de fórmulas/sentenças L-equivalentes corresponde exatamente a um conjunto de extensão/modelo.

(b) A disjunção de duas fórmulas ou sentenças corresponde à união das extensões ou conjuntos de modelos correspondentes:

$$||A(x) \vee B(x)||^D = ||A(x)||^D \cup ||B(x)||^D \text{ e } ||A \vee B|| = ||A|| \cup ||B||.$$

(c) Da mesma forma, a conjunção (\wedge) corresponde à interseção (\cap).

Espaços de possibilidade ou domínios são (nem sempre, mas geralmente) infinitamente grandes. Uma álgebra de conjuntos é chamada σ -álgebra (pronuncia-se álgebra “sigma”) se seus conjuntos são fechados sob união *infinita* contável (\cup) ou sob interseção (\cap), isto é, se AL contém uma família infinita de conjuntos $M_i \in AL$, então a união infinita $\cup_{i \in \mathbb{N}} M_i$ e a interseção infinita $\cap_{i \in \mathbb{N}} M_i$ também são elementos de AL . Nas teorias matemáticas da medida e da integração, geralmente assume-se σ -álgebras. As álgebras geradas pela linguagem das extensões/conjuntos de modelos de fórmulas/sentenças, por outro lado, não são σ -álgebras, desde que um número infinito de fórmulas mútua e logicamente não equivalentes possa ser formado na linguagem dada. Esse é sempre o caso para linguagens com um número infinito de constantes individuais e/ou predicados unários (para linguagens com relações n -árias um número infinito de variáveis individuais é suficiente). Isso ocorre porque as expressões linguísticas são consideradas entidades algoritmicamente decidíveis, que só podem ter um comprimento finito. As fórmulas de uma linguagem \mathcal{L} , que correspondem a uniões ou interseções, podem, portanto, consistir apenas em disjunções ou conjunções de fórmulas finitamente longas, e, uma vez que há um número infinito de fórmulas não equivalentes aos pares em linguagens com uma oferta infinita de caracteres, as fórmulas \mathcal{L} podem ser usadas. A álgebra gerada de proposições não pode ser uma σ -álgebra. Expressões com o quantificador universal podem reproduzir algumas, mas não todas, conjunções infinitamente

longas (e, analogamente, sentenças com o quantificador existencial só podem reproduzir algumas disjunções infinitas). Por exemplo, a conjunção infinitamente longa $Fa_1 \wedge Fa_2 \wedge \dots$ corresponde à cláusula universal $\forall x Fx$, mas a conjunção infinita $F_1a \wedge F_2a \wedge \dots$ (com F_i como um conjunto indexado de predicados unários) não pode ser representada por nenhuma cláusula universal.

Pela mesma razão, para linguagens com um conjunto infinito de caracteres, não existe espaço de possibilidade Ω das possibilidades logicamente mais fortes ou “mundos possíveis” que possa ser expresso em \mathcal{L} , cujas unidades estariam contidas na linguagem-álgebra gerada. Para representá-los como fórmulas, seriam necessárias conjunções infinitamente longas. Tais “mundos possíveis” só podem ser representados em linguagens formais comuns pelos conjuntos infinitos de fórmulas, nomeadamente, pelos chamados conjuntos de fórmulas maximamente consistentes [conjunto maximal consistente].¹⁰ É claro que existe uma maneira formalmente simples de tornar a linguagem tão expressiva que a linguagem-álgebra gerada passa a ter fechamento σ : introduzindo nessas linguagens infinitas conjunções/disjunções como fórmulas abstratas. No entanto, nem as regras de forma e nem as regras de derivação de tais linguagens são decidíveis (uma vez que se referem a conjuntos infinitos de premissas), o que denota que o significado do cálculo como algoritmos de prova foi perdido.¹¹

¹⁰ Se a linguagem tiver nomes padrão, então os elementos de Ω podem ser representados por “diagramas”: estes são conjuntos de frases básicas com a máxima consistência.

¹¹ As regras de forma adicionais de tais linguagens são: se $X \subseteq \text{Form}(\mathcal{L})$, então $\bigwedge X, \bigvee X \in \text{Form}(\mathcal{L})$, com X como um conjunto arbitrário de fórmulas e “ \wedge ” e “ \vee ” para o potencialmente conjunção ou disjunção infinita dos termos de um conjunto de

(3.) Na estrutura *sintática-epistêmica da linguagem* (p.ex., Carnap, 1950), a função de probabilidade P é definida diretamente sobre as sentenças de uma linguagem \mathcal{L} e é caracterizada axiomáticamente, p.ex., usando os axiomas da Def. 3-1 ou da Def. 3-3.

Uma estrutura sintática é possível para probabilidades estatísticas, mas não conheço ninguém que tenha feito isso. Devido ao Teorema (T7) do Teorema 3-1, sentenças logicamente equivalentes têm a mesma probabilidade, razão pela qual a função de probabilidade P construída sintaticamente pode ser convertida em uma semântica, atribuindo P de forma idêntica aos conjuntos modelo de sentenças logicamente equivalentes. Conforme explicado, para línguas com uma oferta infinita de caracteres, os elementos logicamente mais fortes (completos) do espaço de possibilidades não podem mais ser representados linguisticamente. Por essa razão, Carnap (1950) definiu sua função de probabilidade sintático-epistêmica apenas para linguagens monádicas finitas (ou seja, linguagens com um número finito de constantes individuais e predicados de lugar único). Carnap (1971) preferiu a construção semântica por sua maior expressividade.

Na estrutura sintática, também normalmente se assume que a função de probabilidade p ou P é aplicada a expressões de linguagem objeto, mas é expressa na metalinguagem matemática. Também é possível introduzir a função de probabilidade p ou P diretamente em uma linguagem de objeto lógica de predicados estendida. Esse árduo caminho é descrito em Bacchus (1990, Capítulos 2.3, 3.2) e em Halpern (2003, Capítulos 7.3, 7.7). Além dos axiomas básicos de probabilidade, os axiomas para números reais também devem ser expressos em

fórmulas. Analogamente, as regras de derivação devem ser estendidas; por exemplo, a regra da conjunção infinita é: se para todo $A \in X \subseteq \text{Form}(\mathcal{L})$: $\vdash A$, então $\vdash \bigwedge X$.

linguagem objetiva. A teoria lógica de predicados assim obtida está correta, mas comprovadamente incompleta (Bacchus, 1990, p. 62, Teorema 15).

Seguindo as explicações acima, especificamos o conceito de inferência probabilística da seguinte forma:

(Def. 3-5) *Inferência probabilística:*

(a) Uma sentença de probabilidade é uma sentença na linguagem matemática, que é formada a partir de termos, usando variáveis e funções matemáticas e símbolos de relação, da seguinte forma: (i) constantes para conjuntos de números ou números e (ii) termos da forma $P(X)$, onde X representa um conjunto de uma álgebra AL , ou uma frase, ou uma fórmula de uma linguagem objeto \mathcal{L} . As sentenças de probabilidade são, portanto, por exemplo, sentenças da forma $P(X) = r$, $P(X)/(P(Y) > 2 \cdot P(Z))$, ou $P(Fa|Fb) > P(Fa)$ (análogo para p).

(b) Uma sentença de probabilidade S segue probabilisticamente de um conjunto de sentenças de probabilidade Δ , se S segue logicamente de Δ e dos axiomas básicos (A_{1-3} da Def. 3-1) com a ajuda das leis de cálculo para números reais. Escrevemos isso abreviado como $\Delta \Vdash_{A_{1-3}} S$. Semanticamente, isso significa o seguinte: $\Delta \Vdash_{A_{1-3}} S$ é válido sse em todos os modelos de probabilidade (Ω, AL, P) , nos quais Δ é verdadeiro, S também é verdadeiro.¹²

¹² Os modelos permitidos devem ser restritos a modelos matemáticos padrão que atribuem as interpretações corretas a constantes matemáticas, símbolos de função e relação, por exemplo, termos de números reais são elementos do modelo padrão de números reais, “+” a operação de adição aritmética, etc.

Exemplos: De $P(A) = 0,5$, segue probabilisticamente que $P(A \vee B) \geq 0,5$; ou de $P(A) = 0,8$ e $P(B) = 0,9$, segue probabilisticamente $P(A \wedge B) \geq 1 - (1 - 0,8) - (1 - 0,9) = 0,7$ (para o último resultado ver Teorema 4-1).

Finalmente, chegamos à representação de *combinações* de experimentos aleatórios (independentes). Com um experimento aleatório n vezes, queremos dizer a execução do mesmo experimento aleatório n vezes, por exemplo, por execução um após o outro. Quanto às probabilidades epistêmicas, essas não requerem tratamento especial, uma vez que o espaço de possibilidades coincide com o conjunto de todos os modelos, e isso determina a probabilidade de todas as sentenças com qualquer número de constantes individuais, incluindo a probabilidade das sentenças que se referem aos resultados individuais dos experimentos aleatórios. As probabilidades estatísticas combinadas, por outro lado, requerem tratamento separado, porque exigem leis adicionais, como as da Seção 3.2 e a lei de independência estatística mencionada (3-3) se aplica. Citamos aqui apenas os princípios básicos mais importantes e passamos os detalhes lógicos para o Apêndice 10.2.

Por meio de um experimento aleatório simples, os limites de frequência das fórmulas em uma variável, $A(x)$, são determinados por referência a uma sequência infinita de sorteios aleatórios de $\Omega = D$. Em uma fórmula com duas variáveis individuais $A(x_1, x_2)$, x_1 refere-se ao resultado do primeiro e x_2 ao do segundo experimento aleatório (independente) em algum arranjo fixo. Um exemplo é lançar uma moeda duas vezes. O espaço de possibilidade do experimento aleatório duplo é, portanto, $D^2 = D \times D$: sorteamos aleatoriamente pares (d_1, d_2) de $D \times D$ e verificamos se a atribuição da variável $[x_1: d_1, x_2: d_2]$ (qual das variáveis individuais x_1 atribui ao indivíduo d_1 e x_2 atribui ao indivíduo d_2) satisfaz a fórmula $A(x_1, x_2)$. A probabilidade estatística

$p(A(x_1, x_2))$ é o limite de frequência de pares que satisfazem $A(x_1, x_2)$ em uma sequência infinita de sorteios duplos ou em execuções duplas do experimento aleatório determinado $((d_{1,1}, d_{1,2}), (d_{2,1}, d_{2,2}), \dots)$. Observe que as probabilidades de resultados únicos também são determinadas por um experimento duplo: por causa da lei da independência, a probabilidade estatística de Fx é igual à probabilidade de sortear um par (x_1, x_2) , onde x_1 é um F e x_2 é arbitrário. Isso é chamado de “projeção” do espaço de resultados bidimensional (x_1, x_2) na dimensão x_1 e é chamado também de lei $p(Fx_1 \wedge x_2 = x_2) = p(Fx)$ – ou, na notação de conjunto, $p(\{(x_1, x_2): Fx_1\}) = P(\{x:Fx\})$ –, uma lei de projeção. A implementação dessa construção para experimentos aleatórios combinados, de diversas maneiras, pode ser encontrada no Apêndice 10.2.

3.4 Aditividade sigma: prós e contras

Uma suposição para medidas de probabilidade sobre σ -álgebras, que vai além dos axiomas básicos de Kolmogorov, é a σ -aditividade:

(Def. 3-6) *σ -aditividade*: Uma função de probabilidade $p: AL \rightarrow [0,1]$ é chamada σ -aditiva sse para cada sequência infinita $(A_i: i \in \mathbb{N})$ de elementos disjuntos aos pares A_i de AL : $p(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} p(A_i)$, ou em *palavras*: a probabilidade de sua união infinita é a soma infinita de suas probabilidades.

A soma infinita $\sum_{i \in \mathbb{N}} p(A_i)$ é explicada como o limite da sequência de somas finitas, $\lim_{n \rightarrow \infty} \sum_{1 \leq i \leq n} p(A_i)$. Uma função de probabilidade σ -aditiva também é chamada de *medida* de probabilidade. A σ -aditividade é uma suposição fundamental da teoria matemática da

medida sobre espaços de possibilidades contínuos (com valor real) e da teoria das integrais de Lebesgue (ver Capítulo 8.6). No entanto, ela não é um requisito geralmente adequado, porque tal suposição impõe um *viés* em qualquer distribuição de probabilidade sobre um espaço de possibilidades *infinito contável*. Deixe esse espaço de possibilidades ser, por exemplo, o conjunto de números naturais, $\Omega = \mathbb{N}$, e seja $p(\{i\})$ a probabilidade de extrair um certo número i de uma urna infinita, por assim dizer. Definitivamente será sorteado algum número, ou seja, $p(\mathbb{N}) = p(\cup_{i \in \mathbb{N}} \{i\}) = 1$. Contudo, a distribuição uniforme de p sobre \mathbb{N} significa que, para cada número natural, sua probabilidade de sorteio $p(\{i\})$ deve ser zero (porque se $p(\{i\}) > 0$, então $p(\cup_{i \in \mathbb{N}} \{i\}) = \sum_{i \in \mathbb{N}} p(\{i\})$ seria infinito). A soma das probabilidades de sorteio de todos os números individuais é, portanto, zero, porque uma soma de zeros de qualquer comprimento resulta em zero: $\sum_{i \in \mathbb{N}} p(\{i\}) = 0 + 0 + \dots = 0$. Isso significa que $\sum_{i \in \mathbb{N}} p(\{i\}) = 0 < p(\cup_{i \in \mathbb{N}} \{i\}) = 1$, ou seja, que uma função de probabilidade uniformemente distribuída sobre \mathbb{N} não pode ser σ -aditiva. Por essa razão, nem Kolmogorov (ver 1933, § 1-2) e nem De Finetti (ver 1970, Capítulo III.11.6) consideraram a σ -aditividade como um axioma básico. Spielmann (1977) objetou ao exemplo de que desenhos aleatórios reais devem estar relacionados a uma seção inicial finita de \mathbb{N} .

Entretanto, também se pode ver o experimento de desenho infinito como um limite de experimentos de desenho finito cada vez mais abrangentes sobre $\{1, \dots, n\}$; também nesse caso, as probabilidades limites resultantes para $n \rightarrow \infty$ violam a σ -aditividade.

A soma infinita $\sum_{i \in \mathbb{N}} p(\{i\})$ só pode assumir o valor 1, ou qualquer valor maior que zero e menor que infinito, se a sequência de probabilidades $(p(\{i\}); i \in \mathbb{N})$ começar com valores positivos e, então, se aproximar de zero com rapidez suficiente.¹³ Portanto, para cada distribuição de probabilidade σ -aditiva sobre \mathbb{N} , quase toda a massa de probabilidade é concentrada em uma seção inicial finita de \mathbb{N} e desaparece para $n \rightarrow \infty$ contra zero (cf. Howson; Urbach, 1996, p. 34). Isso é mostrado na Figura 3-4:

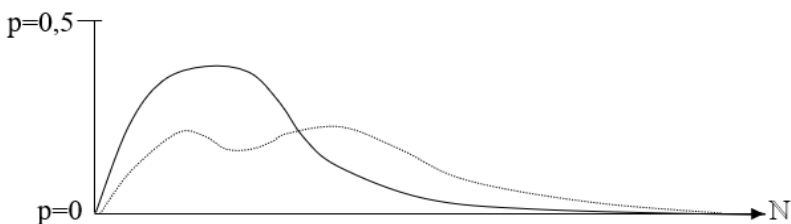


Figura 3-4: Medidas de probabilidade σ -aditivas sobre \mathbb{N} .

Kelly (1996, p. 321 e seg.) mostrou que a σ -aditividade das probabilidades *subjetivas* envolve uma suposição *indutiva* fraca. Para cada hipótese universal $\forall x A(x)$ sobre uma região infinita arranjada arbitrariamente $D = \{d_1, d_2, \dots\}$, devido à aditividade σ , a probabilidade de que o n -ésimo indivíduo seja a primeira instância falsificadora de $\forall x A(x)$ vai aproximando-se rapidamente de zero à medida que n aumenta. Portanto, se houver casos de falsificação de uma hipótese universal, então eles deverão ocorrer muito em breve. A consequência é a seguinte

¹³ Por exemplo, a soma infinita da sequência $1/2 + 1/4 + 1/8 + \dots$ tem o valor 1, mas a soma infinita $1/2 + 1/3 + 1/4 + \dots$ tem o valor ∞ .

(3-5) *Indutividade de limites* para todas as hipóteses estritas:

Se $P(\forall xA(x)) > 0$: $\lim_{n \rightarrow \infty} P(\forall xA(x) | A(a_1) \wedge \dots \wedge A(a_n)) = 1$.

ou seja, se o número de confirmações positivas de uma hipótese geral aumentar suficientemente, o nosso grau de crença na hipótese $\forall xA(x)$ (se for inicialmente considerada possível) aumenta até ao ponto de certeza. Um cético humano da indução nunca poderia concordar com isso: ele objetaria que, depois de cada número finito de observações confirmatórias $A(a_1), \dots, A(a_n)$, não importa quão grande seja, ainda resta um número infinito de indivíduos não observados que podem falsificar a hipótese, porque para ele a probabilidade de a hipótese ser falsa não diminui nem um pouco. Como mostra Kelly, a indutividade de limites se desfaz assim que P não é mais σ -aditivo.

A seguir, portanto, não consideramos a aditividade σ como um axioma geral da teoria das probabilidades. As medidas de probabilidade não σ -aditivas foram estudadas, entre outros, em Bhaskara Rao & Rao (1983) e em Schurz & Leitgeb (2008): elas satisfazem uma série de leis interessantes, mas mais fracas do que as medidas σ -aditivas. Por exemplo, o seguinte também se aplica a funções de probabilidade não aditivas:

(3-6) *Para p não- σ -aditivo*: $p(\cup_{i \in \mathbb{N}} \{i\}) \geq \sum_{i \in \mathbb{N}} p(\{i\})$.

Como a σ -aditividade refere-se a uniões ou disjunções infinitas, ela não pode ser totalmente reproduzida na estrutura linguística usual. Uma maneira de expressar completamente o princípio da σ -aditividade na linguagem são as linguagens abstratas infinitas explicadas acima com operações infinitas de conjunção e disjunção. Para sentenças quantificadas, entretanto, a σ -aditividade de uma

medida de probabilidade implica o seguinte princípio da continuidade, desde que seja assumida uma linguagem com intervalo individual infinito contável e com nomes padrão (ver Schurz; Leitgeb, 2008, fato 6). O último significa que cada indivíduo d_i de D em \mathcal{L} tem exatamente um nome padrão a_i ($I(a_i)=d_i$).

(3-7) *Princípio da continuidade* (consequência da σ -aditividade para linguagens com nomes padrão):

(a) *Versão epistêmica*: $P(\forall x A(x) = \lim_{n \rightarrow \infty} P(A(a_1) \wedge \dots \wedge A(a_n)))$. (Análogo para \exists)

Em palavras: O grau de crença de que todos os indivíduos possuem a propriedade A é idêntico ao limite do grau de crença de que os primeiros n indivíduos do domínio individual possuem a propriedade A , para n em direção ao infinito.

(b) *Versão estatística*: $p(\forall y R(x,y) = \lim_{n \rightarrow \infty} p(R(x,a_1) \wedge \dots \wedge R(x,a_n)))$. (Análogo para \exists).

4 Justificativa das modalidades de inferência dentro da teoria das probabilidades

4.1 Modalidades de inferência

No âmbito da teoria da probabilidade epistêmica, diferentes modalidades de inferência podem ser reconstruídas e, às vezes, justificadas. É feita uma distinção básica entre inferências *dedutivas* (ou lógicas) e *não dedutivas*. As inferências dedutivas são *certas*: transferem a verdade das suas premissas para a conclusão com certeza ou em todos os mundos possíveis. As inferências indutivas (ou não dedutivas), por outro lado, são *incertas*: apenas transferem a verdade das suas premissas para a conclusão em mundos suficientemente uniformes ou uniformes; estes são mundos nos quais o futuro do passado ou do não observado é suficientemente semelhante ao que já foi observado. Aqui estão dois exemplos:

(4-1)

Raciocínio dedutivo

Todos os peixes respiram por guelras.

Raciocínio indutivo

Todos os peixes observados até agora ($n^{\circ} 1, 2, \dots, n$) respiram por guelras.

Este animal é um peixe.

Portanto, este animal respira por guelras.

Certeza: transmissão da verdade em todos os mundos possíveis.

Então (provavelmente), todos os peixes respiram por guelras.

Incerteza: apenas transmissão da verdade em mundos possíveis 'uniformes' o suficiente.

A linha simples final indica segurança, a linha espessa (dupla) indica incerteza. As inferências indutivas transferem as conexões observadas para os casos novos e não observados. Diz-se também que as inferências indutivas expandem o conteúdo.

Nem todas as conclusões não dedutivas são da mesma natureza indutiva do sentido acima. Outro tipo de raciocínio não dedutivo é a abdução, ou a inferência para a melhor explicação. Grosso modo, um efeito observado é baseado em uma causa suspeita, por exemplo, de uma trilha sinuosa na areia até uma víbora de areia que passou por ali. A forma final da abdução remonta a C. S. Peirce. Para o Peirce posterior, era essencial que novos conceitos e modelos teóricos pudessem ser introduzidos nas ciências através do raciocínio abduativo (Peirce, 1903, § 170). Por exemplo, Newton concluiu abduativamente a existência de uma força gravitacional proveniente do movimento dos planetas ao redor do sol. Como enfatizou Peirce, o status de validade de uma hipótese inferida abduativamente é muito incerto e provisório: a hipótese abduida deve ser testada empiricamente por dedução e por indução para assumir o caráter de uma hipótese provável (1903, § 171). Além disso, há sempre diversas hipóteses explicativas possíveis para o fato que necessita de explicação, e o procedimento de raciocínio abduativo seleciona a melhor delas. Nesse sentido, Harman

(1965) reconstruiu a abdução peirciana como a inferência para a melhor explicação. Como mostrado em Schurz (2008a), a abdução entendida dessa forma abrange toda uma família de tipos de inferência que têm em comum o seguinte esquema.

(4-2) Esquema final da abdução (cf. Niiniluoto, 1999):

Premissa 1: Um fato (singular ou geral) que precisa de explicação E.

'Premissa' 2: Um conhecimento prévio W que implica uma certa hipótese H: H é uma explicação plausível e, entre as explicações candidatas atualmente conhecidas, a melhor explicação para E.

Conjectura abdutiva: H é verdadeiro

Quando se reconstroem probabilisticamente modalidades de inferência – sejam dedutivas, indutivas ou abdutivas –, não se pergunta principalmente sobre a preservação (completa ou parcial) da verdade, mas sim sobre o nível da probabilidade epistêmica condicional da conclusão, dadas as premissas, bem como sobre a (dependendo disso) transferência completa ou parcial de uma alta probabilidade das premissas para a conclusão. Essa consideração deve agora ser realizada para as diferentes modalidades de inferência.

4.2 Raciocínio dedutivo

Conforme explicado no apêndice da Seção 10.1, definimos “ $\|$ —” para a relação da sequência lógica, ou seja, $A_1, A_2 \|$ — B significa “B segue logicamente de A_1, A_2 ”. Devido às provas de completude para lógica proposicional e de predicados de primeira ordem, o termo de inferência semântica pode ser substituído, de forma equivalente, pelo

termo de derivação sintática “|—”. As conexões fundamentais entre a lógica dedutiva e a teoria das probabilidades são refletidas no seguinte teorema:

(Teorema 4-1) Teoria da probabilidade e conclusão lógica:

Seja \mathcal{P} o conjunto de todas as funções de probabilidade epistêmicas possíveis sobre a álgebra $AL(\mathcal{L})$ das proposições de uma linguagem \mathcal{L} . Deixando-se $U(A) =_{\text{def}} 1 - P(A)$ representar a chamada *incerteza* P da sentença A , então, para todas as sentenças A, A_1, \dots, A_n, B :

(4-1.1) (i) $A_1, \dots, A_n \parallel - B$ sse

(ii) $\forall P \in \mathcal{P}: P(B|A_1 \wedge \dots \wedge A_n) = 1$ sse

(iii) $\forall P \in \mathcal{P}: P(B) \geq P(A_1 \wedge \dots \wedge A_n)$ sse

(iv) $\forall P \in \mathcal{P}$: se $P(A_1 \wedge \dots \wedge A_n) = 1$, então $P(B) = 1$.

Em palavras: (i) Uma conclusão segue de um conjunto de premissas, sse (ii) a probabilidade condicional de conclusão, dada a conjunção de premissas em todos os modelos de probabilidade, é 1, sse (iii) a probabilidade da conclusão em todos os modelos de probabilidade é maior ou igual à probabilidade da conjunção de premissas, sse (iv) a probabilidade da conclusão é 1 em todos os modelos de probabilidade em que a probabilidade da conjunção de premissas é 1.

(4-1.2) $\forall P \in \mathcal{P}: U(A_1 \wedge \dots \wedge A_n) \leq U(A_1) + \dots + U(A_n)$.

Em palavras: Em todos os modelos de probabilidade, a incerteza de uma conjunção de sentenças é menor ou igual à soma das incertezas das sentenças individuais.

(4-1.3) (segue de 4.1+2) $A_1, \dots, A_n \parallel - B$ sse $\forall P \in \mathcal{P}: U(B) \leq U(A_1) + \dots + U(A_n)$.

Em palavras: Uma conclusão segue de um conjunto de premissas, se a soma das incertezas das premissas em todos os modelos de probabilidade é menor ou igual à incerteza da conclusão.

As conexões entre inferência lógica e probabilidade referem-se sempre ao que se aplica a todos os modelos de probabilidade. O Teorema 4-1.1 mostra que, para compreender a conexão entre raciocínio lógico e probabilidade, é preciso, antes de mais nada, conhecer não apenas as probabilidades das premissas, mas também a probabilidade de sua conjunção. Então tudo é simples, porque a probabilidade condicional da conclusão, dada a conjunção de todas as premissas, é necessariamente 1 e, assim, a probabilidade da conclusão é sempre maior ou igual à probabilidade da conjunção das premissas. As altas probabilidades são, portanto, transferidas inteiramente da conjunção de premissas para a conclusão. O significado de (iv) é que (iv) não é apenas suficiente, mas é também necessário para (i): uma probabilidade 1 só é transferida da conjunção de premissas para a conclusão em todos os modelos de probabilidade possíveis se a conclusão também for lógica e válida. Uma prova do Teorema (4-1.1) pode ser encontrada no apêndice matemático 10.3.4.

O Teorema (4-1.2) também é chamado de regra da soma das incertezas e fornece um limite superior significativo para a incerteza da conjunção de todas as premissas, que se aplica a todos os modelos de probabilidade, ou seja, à soma das incertezas das premissas. A prova do Teorema (4-1.2) remonta a Suppes (1966, 54). O Teorema (4-1.3) é uma consequência direta de (4-1.1) e de (4-1.2).

A estrutura sintática linguística, explicada na Seção 3.3, ainda inclui o conceito de conclusão lógica nos conceitos de disjuntividade e de exaustividade. Popper (1935; 1976, Apêndice II*; Apêndice IV*)

mostrou que uma axiomatização sintática da probabilidade pode ser realizada mesmo sem o pré-requisito de um conceito de inferência sobre as sentenças da linguagem-objeto, introduzindo axiomas de probabilidade para sentenças da linguagem-objeto de tal forma que elas contenham as leis da lógica proposicional em uma forma implícita.

As probabilidades assim axiomatizadas também são chamadas de *funções de Popper*.¹⁴

(Def. 4-1) *Axiomatização de funções de Popper* $P: \text{Sent}(\mathcal{L}) \times \text{Sent}(\mathcal{L}) \rightarrow \mathbb{R}$ para uma linguagem lógica proposicional \mathcal{L} : Para todo $A, B, C \in \text{Sent}(\mathcal{L})$:

(PA₁) $\exists X, Y \in \text{Sent}(\mathcal{L}): P(X|Y) \neq 1$.

(PA₂) $P(A|A) = 1$.

(PA₃) $P(A|B \wedge C) = P(A|C \wedge B)$.

(PA₄) $P(A \wedge B|C) = P(B \wedge A|C)$.

(PA₅) $P(A|B) + P(\neg A|B) = 1$ ou $P(C|B) = 1$ (para qualquer C).

(PA₆) $P(A \wedge B|C) = P(A|B \wedge C) \cdot P(B|C)$.

Definição: $P(A) =_{\text{def}} P(A|B \vee \neg B)$

(Teorema 4-2) *Funções de Popper*:

(4-2.1) A classe de funções condicionais de Popper (Def. 4-1) corresponde à classe de funções de probabilidade condicionais da Def 3-3, desde que “ $\square B$ ” seja definido por “ $P(C|\neg B) = 1$ para qualquer C ”.

¹⁴ Para a axiomatização (Def. 4-1) e a prova do Teorema 4-2, veja Popper (*ibid.*) e Hawthorne (1996, Teorema 1+1).

(4-2.2) A classe das funções de Popper incondicionais coincide com a classe das funções de probabilidade de Kolmogorov.

(4-2.3) $A \parallel \text{---} B$ sse para todas as funções de Popper P for aplicado o seguinte: $P(B|A) = 1$ sse o seguinte se aplicar a todas as funções de Popper: $P(B) \geq P(A)$.

O Teorema 4-2.2 é uma consequência do Teorema 4-2.1 e da Def. 3-3 da axiomatização direta de Carnap. O Teorema 4-2.3 resulta do Teorema 4-2.1 e do Teorema 4-1.1. O Teorema 4-2.3 permite definir o conceito de conclusão lógica usando condições puramente probabilísticas.

É possível argumentar que, dessa forma, a lógica pode ser rastreada até a teoria das probabilidades. No entanto, isso só é verdade para a linguagem-objeto, porque todas as provas metalinguísticas sobre funções de Popper (por exemplo, a prova do Teorema 4-2), por sua vez, pressupõem a validade das leis da dedução lógica na metalinguagem.

4.3 Condicionais incertas

Uma extensão interessante da justificação probabilística do raciocínio dedutivo remonta a Adams (1975): ele mostrou como se pode raciocinar dedutivamente com condicionais incertas – aqui expressas pela seta dupla $A \Rightarrow B$. Embora uma alta probabilidade de uma condicional 'material' comum $A \rightarrow B$ (logicamente equivalente a $\neg A \vee B$) signifique que $P(A \rightarrow B) = P(\neg A \vee B)$ é alta, uma alta probabilidade de uma condicional é incerta (por definição) de que a probabilidade condicional $p(B|A)$ é alta. Ambas não são iguais. É verdade que é necessário

$$(4-3) P(A \rightarrow B) \geq P(B|A),^{15}$$

mas inversamente, apesar de um $P(A \rightarrow B)$ alto, a probabilidade condicional $P(B|A)$ pode ser muito pequena. Por exemplo, $P(\text{chanceler} \rightarrow \text{palhaço de circo})$ é muito alta, porque a maioria das pessoas não são chanceleres, mas $P(\text{palhaço de circo} | \text{chanceler})$ é muito baixa.

Aplicam-se leis lógicas mais fracas a altas probabilidades condicionais do que a condicionais estritas. Por exemplo, a lei da transitividade aplica-se a condicionais estritas: “Todos os Fs são Gs” e “Todos os Gs são Hs” implicam “Todos os Fs são Hs”. E, caso ambas as premissas sejam altamente prováveis, a conclusão também será altamente provável, porque $p(Fx \rightarrow Gx) \geq 1 - \epsilon_1$ e $p(Gx \rightarrow Hx) \geq 1 - \epsilon_2$ implicam, com base no Teorema 4-1.3, $p(Fx \rightarrow Hx) \geq 1 - \epsilon_1 - \epsilon_2$. Em contraste, “A maioria dos Fs são Gs” e “A maioria dos Gs são Hs” geralmente não ocasionam “A maioria dos Fs são Hs”. Um contra-exemplo: a maioria dos alemães não vivem em Munique, e a maioria das pessoas que não vivem em Munique não são alemães. Usando a transitividade obter-se-ia: “A maioria dos alemães são não-alemães”, o que é obviamente errado.

No sistema P da lógica de probabilidade condicional que remonta a Adams (1975), conclui-se haver um conjunto de condicionais incertas para uma condicional incerta resultante. As

¹⁵ *Prova:* $P(\neg A \vee B) = P(\neg A) + P(B) - P(\neg A \wedge B) \geq P(B|A) = P(A \wedge B) / P(A)$ sse $P(\neg A) + P(A \wedge B) \geq P(A \wedge B) / P(A)$ (então $P(B) - P(\neg A \wedge B) = P(A \wedge B)$) sse $P(\neg A) \cdot P(A) \geq P(A \wedge B) - P(A \wedge B) \cdot P(A)$ sse $(1 - P(A)) \cdot P(A) \geq (1 - P(A)) \cdot P(A \wedge B)$, o que é obviamente verdade porque $P(A) \geq P(A \wedge B)$.

regras de derivação são as seguintes, com “ $\mid\text{---}P$ ” para “pode ser derivado no cálculo P”:

(Teorema 4-3) <i>Regras da lógica de probabilidade condicional (Sistema P):</i>	
Transitividade cautelosa VT:	$A \Rightarrow B, A \wedge B \Rightarrow C \mid\text{---}P A \Rightarrow C$
Monotonicidade cautelosa	$A \Rightarrow B, A \Rightarrow C \mid\text{---}P A \wedge B \Rightarrow C$
VM:	
Disjunção cautelosa VD:	$A \Rightarrow C, B \Rightarrow C \mid\text{---}P A \vee B \Rightarrow C$
Supraclassicalidade SK:	Se $A \parallel\text{---} B$, então $\mid\text{---}P A \Rightarrow B$.
<i>Algumas regras derivadas:</i>	
Conjunção K:	$A \Rightarrow B, A \Rightarrow C \mid\text{---}P A \Rightarrow B \wedge C$
Equivalência lógica à esquerda	Se $\mid\text{---} A \leftrightarrow B$, então $A \Rightarrow C \mid\text{---}P B \Rightarrow C$
LLÄ:	
Atenuação direita RA:	Se $\mid\text{---} B \rightarrow C$, então $A \Rightarrow B \mid\text{---}P A \Rightarrow C$
Prova condicional cautelosa	$A \wedge C \Rightarrow B \mid\text{---}P A \Rightarrow (B \rightarrow C)$
VKP:	

Adams (1975) (aluno de Suppes) provou um teorema da soma das incertezas semelhantes para o raciocínio no sistema P, tal como Suppes (1966) provou para o raciocínio lógico comum: a incerteza da conclusão não pode ser maior que a soma das incertezas das premissas. A incerteza de uma condicional incerta é definida como $U(A \Rightarrow B) =_{\text{def}} U(B|A) =_{\text{def}} 1 - P(B|A)$:

(Teorema 4-4) $A_1 \Rightarrow K_1, \dots, A_n \Rightarrow K_n \mid\text{---}_P A \Rightarrow K$ sse para todas as funções de probabilidade P, sobre as proposições da linguagem
--

subjacente \mathcal{L} , sem um operador condicional \Rightarrow , aplicar-se o seguinte: $U(K|A) \leq U(K_1|A_1) + \dots + U(K_n|A_n)$.

A prova do Teorema 4-4 pode ser encontrada dispersa em Adams (1975), e a prova de uma generalização do Teorema 4-4 pode ser encontrada em Schurz (1998).

4.4 Raciocínio indutivo

As inferências indutivas podem assumir muitas formas. Aqui é fornecida uma visão geral das formas *probabilísticas* mais importantes de inferências indutivas:

(4-4) *Conclusão da generalização indutiva:*

(a) *Estatisticamente:* $r\%$ de todos os Fs observados até agora eram Gs, então provavelmente cerca de $r\%$ de todos os Fs são Gs.

Versão semiformal (com “[$r \cdot n$]” como um arredondamento inteiro de $r \cdot n$, e “[$r \pm \epsilon$]” como um intervalo simétrico de 2ϵ em torno do valor r , para um número arbitrariamente pequeno ϵ): O valor de $P(p(Fx) \in [r \pm \epsilon] \mid h_n(F) = [r \cdot n]/n)$ é mais ou menos alto (dependente de ϵ e de n) e tende a 1 para $n \rightarrow \infty$.

(b) *Estrito:* (Caso especial de (a)): Todos os Fs observados até agora foram Gs, então provavelmente todos os Fs são Gs.

Versão semiformal: O valor de $P(\forall x Fx \mid Fa_1 \wedge \dots \wedge Fa_n)$ é mais ou menos alto (dependente de n) e se aproxima de 1 para $n \rightarrow \infty$.

Versões formais dessas conclusões serão discutidas nas Seções 7.7, 9.3 e 9.5. Para que as inferências indutivas sejam válidas probabilisticamente, a função de probabilidade subjacente deve satisfazer certas condições indutivas adicionais (por exemplo,

permutabilidade ou indiferença), as quais serão mais esclarecidas na Seção 7.6 e, com mais detalhes, na Seção 9.3.

Uma justificação não circular de inferências indutivas, sem inserir certos princípios indutivos na função de probabilidade, não é possível (ver Capítulo 9.7).

(4-5) *Raciocínio indutivo de previsão:*

(a) *Estatisticamente:* r% de todos os Fs observados eram Gs, então com uma probabilidade próxima de r% o próximo F também será um G.

Versão semiformal: O valor de $P(F_{n+1} | h_n(F) = [r \cdot n]/n)$ é ϵ -próximo de r (ϵ dependente de n) e tende a r para $n \rightarrow \infty$.

(b) *Estrito* (caso especial de (a)): Todos os Fs observados até agora foram Gs, portanto há uma alta probabilidade de que o próximo F também seja um G.

Versão semiformal: $P(F_{n+1} | F_1 \wedge \dots \wedge F_n) =$ bastante alto (dependente de n), e tende a 1 para $n \rightarrow \infty$.

(4-6) *Conclusão da especialização indutiva* (ver também (2-1)):

(a) *Estatisticamente:* r% de todos os Fs são Gs, este é um F, então, este será um G com r% de probabilidade.

Formalmente: $P(Ga | p(Gx|Fx) = r \wedge Fa) = r$.

(b) *Estrita* – esta conclusão é dedutivamente válida: $\forall x(Fx \rightarrow Gx)$, Fa/Ga.

A conclusão da especialização estatística indutiva também se baseia numa suposição indutiva de uniformidade. Nessa conclusão, a tendência de frequência da população é transferida para um único indivíduo ou amostra. Isso só funciona se o indivíduo não for uma exceção à tendência geral ou se a amostra for representativa. A versão

estrita da conclusão de especialização, por outro lado, não é de natureza indutiva, mas sim dedutiva. Os tipos de raciocínio indutivo listados acima são – embora incertos – tipos de raciocínio formais. Aplicam-se independentemente do conteúdo, e a sua correção é, portanto, concluída pela substituição dos seus símbolos não lógicos por símbolos sintaticamente idênticos.¹⁶ O mesmo se aplica aos princípios probabilístico-indutivos adicionais subjacentes a eles, tais como a permutabilidade ou a indiferença.

4.5 Raciocínio abduativo

A justificação probabilística do raciocínio abduativo também requer suposições probabilísticas adicionais. No entanto, se essas conclusões vão além das conclusões de generalização indutiva, já não podem ser justificadas por princípios formais adicionais, mas requerem suposições específicas de conteúdo (isto é, relacionadas a hipóteses específicas) para a função de probabilidade P .

No caso mais simples, as inferências abduativas têm a forma probabilística $P(H_1|E) > P(H_2|E)$, com H_1 e H_2 como hipóteses rivais, ambas implicando logicamente ou tornando prováveis os dados empíricos E . Se H_1 é a hipótese considerada mais provável por E entre um conjunto de hipóteses explicativas concorrentes, então, de acordo com o raciocínio abduativo, H_1 é provisoriamente aceita. Conforme o teorema de Bayes (TB₅ do Teorema 3-3), $P(H_1|E) = P(E|H_1) \cdot P(H_1)/P(E)$ e assim:

¹⁶ Em contraste, a substituição de predicados primitivos por predicados complexos é geralmente possível no raciocínio dedutivo, mas não no indutivo. No raciocínio indutivo, os “predicados de Goodman” devem ser excluídos (ver Seção 9.7).

(4-7) *Teorema de Bayes e raciocínio abdutivo*:

$$P(H_1|E) > P(H_2|E) \text{ sse } P(E|H_1) \cdot P(H_1) > P(E|H_2) \cdot P(H_2).$$

Em palavras: A evidência torna uma primeira hipótese mais provável do que uma segunda, sse o produto da verossimilhança e da probabilidade inicial da primeira hipótese for maior que o da segunda.

De acordo com (4-7), a hipótese H_i a ser considerada mais provável pela evidência E depende de dois fatores:

(1) da chamada verossimilhança de H_i , $P(E|H_i)$, que mede a probabilidade inversa de E dado H_i e, portanto, a *força* da relação explicativa entre a hipótese explicativa H_i e a evidência E , bem como

(2) da *probabilidade inicial* $P(H_i)$ da premissa explicativa.

No quadro bayesiano, essa hipótese H_i é preferível a uma dada partição $\{H_1, \dots, H_n\}$ cuja probabilidade final é máxima e depende dos fatores (1) e (2). Ambos os fatores dependem da natureza do conteúdo de E e de H_i , assim como do conhecimento prévio. Uma justificação probabilística do raciocínio abdutivo como um tipo formal de raciocínio não é possível dessa forma. Mais detalhes sobre vários métodos de remoção parcial de probabilidades e sobre probabilidades iniciais da arbitrariedade subjetiva são encontrados nos Capítulos 7 e 9.

O fator (1) mede a força da relação explicativa, mas não captura nenhum critério de qualidade capaz de distinguir explicações de possíveis previsões, como por exemplo a questão da relação causal entre *explanans* e *explanandum*. Lipton (1991, p. 59) chamou de uma explicação o aspecto qualitativo enriquecido pelo último aspecto de “*loveliness*”, enquanto Lipton entendeu a “*likeliness*” de uma explicação como a probabilidade das premissas explicativas dado E , ou seja, $P(H_i|E)$ em (4-7).

5 Problemas do conceito objetivo-estatístico de probabilidade

As questões filosóficas sobre conceitos de probabilidade podem ser divididas em questões de definição e questões de justificação:

Questões de definição: O que é probabilidade? Como pode ser explicada?

Questões de justificação: Como podem ser justificados os axiomas da probabilidade? Por que o conceito de probabilidade assim explicado é cientificamente relevante?

Consequentemente, os problemas de conceitos de probabilidade (se houver) são divididos em problemas de definição e problemas de justificação. Os problemas do conceito estatístico de probabilidade são sobretudo problemas de definição, enquanto os seus problemas de justificação são mais fáceis de resolver. Quando se trata do conceito subjetivo de probabilidade, ocorre exatamente o oposto, como se verá no Capítulo 6. As questões de justificação do conceito estatístico de probabilidade serão discutidas nas próximas Seções deste Capítulo.

5.1 Problemas de justificação

A justificação dos axiomas básicos (A_{1-3}) não é problemática para o conceito estatístico de probabilidade: decorre da definição de probabilidade estatística como uma frequência ou como um limite de

frequência (cf. Gillies, 2000, p. 109). Isso é óbvio para frequências. Para limites de frequência, apenas os axiomas A1 e A2 são óbvios, enquanto a justificativa do axioma da aditividade A3 não é muito simples e resulta da seguinte forma: tem-se $\lim_{n \rightarrow \infty} h_n(Fx \vee Gx) = \lim_{n \rightarrow \infty} (h_n(Fx) + h_n(Gx))$, devido à aditividade das frequências finitas h_n nos primeiros n termos da sequência (h_n) . *Assumimos* que os tipos de eventos Fx e Gx também possuem um limite de frequência. Segue-se que $\lim_{n \rightarrow \infty} (h_n(Fx) + h_n(Gx)) = \lim_{n \rightarrow \infty} (h_n(Fx)) + \lim_{n \rightarrow \infty} (h_n(Gx))$ deve ser válido devido à seguinte consideração: suponha que $\lim_{n \rightarrow \infty} (h_n(Fx)) = r_1$ e $\lim_{n \rightarrow \infty} (h_n(Gx)) = r_2$.

Então, de acordo com a Def. 2-2, existe um n_1 para todo (não importa quão pequeno) $\epsilon > 0$, de modo que é válido para todo $m \geq n_1$, $|h_m(Fx) - r_1| < \epsilon$, e um n_2 tal que se aplica para todo $m \geq n_2$, $|h_m(Gx) - r_2| < \epsilon$. Definimos $n =_{\text{def}} \max(n_1, n_2)$ e $\epsilon' =_{\text{def}} 2 \cdot \epsilon$. Isso significa que para todo $\epsilon' > 0$ existe um n , de modo que seja válido para todo $m \geq n$ $|h_m(Fx) + h_m(Gx) - (r_1 + r_2)| < \epsilon'$, o que de acordo com a Def. 2-2 significa ser válido $\lim_{n \rightarrow \infty} (h_n(Fx) + h_n(Gx)) = r_1 + r_2$.

Isso significa que o problema da justificação dos axiomas básicos já foi resolvido. A questão que permanece é: até que ponto se pode assumir que todos os elementos da álgebra de conceitos complexos (fórmulas) realmente têm um limite de frequência. O fato de serem possíveis sequências de resultados sem limite de frequência, cujas frequências oscilam eternamente entre dois valores $a < b$, pode ser visto ao construí-las para o caso de um experimento de lançamento de moeda (com “1” para cara e “0” para coroa): a sequência existe inicialmente a partir de 1 (portanto, $h_1(1) = 1$), depois de zeros até a primeira posição x_1 , na qual $h_{x_1}(1)$ se torna menor ou igual a a , então de uns para a primeira posição x_2 a x_1 , na qual $h_{x_2}(1)$ se torna maior ou

igual a b novamente, etc. Por exemplo, $a = 1/3$ e $b = 2/3$, tal sequência se parece com:

(5-1) *Sequência zero-um sem limite de frequência* (“ $n:1$ ” representa uma subsequência consecutiva de uns n vezes 1, e analogamente para “ $n:0$ ”):

Sequência: $2^0:1, 2^1:0, 2^2:1, 2^3:0, 2^4:1, \dots, 2^{2^n}:1, 2^{2^{n+1}}:0, \dots$
 Frequências: $1, 1/3, 5/7, 1/3, 21/31, \dots, \geq 2/3 = 1/3, \dots$

Com um pouco de esforço matemático, é possível ver que, após cada bloco de 2^{2^n} unidades (posição par, ou seja, 2^{2^n}), a frequência de unidades é $(1+2 \cdot X) / (1+3 \cdot X)$ e, portanto, excede $2/3$. Aqui X é a soma de todos os 2^i para i ímpar $\leq 2 \cdot n$. Após cada bloco de zeros, a frequência dos uns cai para exatamente $1/3$.¹⁷

Sequências de eventos cujas frequências não têm limite apresentam propriedades muito especiais: suas frequências oscilam, para sempre, entre dois valores – seus “limites superiores” e “limites inferiores”¹⁸ – com períodos de oscilação cuja duração mínima aumenta exponencialmente com o número de períodos (para prova, consulte o Apêndice 3.10.5). Alguém poderá perguntar se consequências artificiais desse tipo deveriam ser levadas a sério: afinal, a lei forte dos grandes números (Teorema 3-4) afirma que as frequências com probabilidade 1 convergem para a probabilidade

17 Os cálculos são baseados nas transformações $\sum_{i \leq x, i \text{ par}} 2^i = 1 + 2 \cdot \sum_{i \leq x, i \text{ ímpar}} 2^i$, caso x for par, e $\sum_{i \leq x, i \text{ ímpar}} 2^i = 2 \cdot \sum_{i \leq x, i \text{ ímpar}} 2^i$, caso x for ímpar.

18 O limite superior (ou inferior) de uma sequência infinita de números é definido como o supremo (ou ínfimo) de todos os números reais que são excedidos (ou subcotados) infinitamente frequentemente pelos membros da sequência. Esses dois limites sempre existem; para sequências com um limite de frequência elas coincidem no mesmo limite.

Schurz e Leitgeb (*ibid.*) contrapõem que o problema pode ser facilmente resolvido.

Primeiro, como mencionado, os tipos de eventos cujas frequências observadas não convergem, apesar do aumento do tamanho da amostra, são muito improváveis. Em segundo lugar, mesmo que existam tais tipos de eventos, eles não violam os axiomas básicos; só então o conjunto de todos os eventos, aos quais as probabilidades estatísticas podem ser atribuídas de forma significativa, não forma mais uma álgebra. Schurz e Leitgeb (2008, § 3, Teorema 1) mostram que tal conjunto forma um chamado *sistema pré-Dynkin*: é sempre fechado sob formação de complemento e união finita disjunta. No caso de elementos ilimitados, pode-se construir a função de probabilidade p sobre um sistema pré-Dynkin em vez de uma álgebra. Contudo, também existem duas maneiras pelas quais se pode obter uma álgebra de um sistema pré-Dynkin D . (1.) Qualquer um escolhe a maior álgebra $A^-(D) \subset D$, que está contida em D (nesse caso, lançam-se, no exemplo (5-2), X e Y fora da álgebra). Ou (2.) forma-se o fechamento algébrico $A^+(D)$ de D , ou seja, a menor álgebra $A^+(D) \supset D$, que contém D , e atribui-se limites de frequência 'contrafactuais' aos eventos ilimitados de $A^+(D)$ em algum lugar entre seus membros inferiores e superiores (o que é possível sem contradição; cf. Schurz; Leitgeb, 2008, § 3, Teorema 2).

O problema da justificação também inclui a questão: até que ponto o conceito de probabilidade é científico e de significado prático. Isso é óbvio para o conceito estatístico objetivo, porque, de acordo com a conhecida fórmula da teoria da decisão (por exemplo, Raiffa, 1973), o valor esperado para a utilidade de uma ação depende das probabilidades das possíveis circunstâncias que são relevantes para os seus efeitos.

Se h_1, \dots, h_n denotam os cursos de ação possíveis, u_1, \dots, u_m as circunstâncias possíveis, e se $U(h_i, u_j)$ é a utilidade do curso de ação h_i nas circunstâncias u_j ,¹⁹ então, a utilidade esperada (UE) dos possíveis cursos de ação é dado da seguinte forma:

(5-3) *Conceitos básicos da teoria da decisão:*

Cursos de ação possíveis: h_1, \dots, h_n Circunstâncias possíveis: u_1, \dots, u_m

Para a utilidade esperada da ação h_i : $UE(h_i) = \sum_{1 \leq j \leq m} U(h_i, u_j) \cdot p(u_j)$

Em palavras: a utilidade esperada de uma ação é a soma de seus valores de utilidade em todas as circunstâncias possíveis, cada um multiplicado pelo valor de probabilidade da circunstância.

Se as probabilidades forem de natureza estatística, o valor esperado coincide com a utilidade média objetiva do curso de ação no longo prazo. Para maximizar essa utilidade média, ou seja, para selecionar aqueles com maior utilidade média do conjunto de possíveis cursos de ação, é necessário que as probabilidades estatísticas sejam conhecidas pelo menos aproximadamente – e é aí que reside a sua relevância. Um exemplo: suponha que eu esteja diante da decisão de comprar, ou não, um carro. Em suma, nas viagens urbanas, o valor utilitário do automóvel é apenas metade do valor do transporte público (custos elevados, sem poupança de tempo), enquanto, nas viagens rurais, o primeiro excede o último em 50%. Se a utilidade do transporte público for definida como 2 em ambos os casos, a matriz de utilidade e os valores esperados ficarão assim:

¹⁹ As circunstâncias são consideradas independentes das ações. Uma extensão que inclui circunstâncias dependentes de ação pode ser encontrada em Skyrms (1980, § II.C).

(5-4)	Com o carro	Com o transporte público
Benefícios das viagens urbanas (frequência p)	1	2
Benefícios das viagens rurais (frequência 1-p)	3	2
UE (carro) = $1 \cdot p + 3 \cdot (1-p)$ = $3 - 2 \cdot p$	UE (público) = $p \cdot 2 + (1-p) \cdot 2 = 2$	

Se eu estimar p em 1/3 e comprar um carro, mas a verdadeira frequência de viagens pela cidade for 2/3, tomarei uma decisão errada, pois, nesse caso, $UE(\text{Carro}) = 5/3$ e, portanto, abaixo do valor (independente de p) de $UE(\text{público}) = 2$.

As ações são de natureza epistêmica, assim, por exemplo, para possíveis previsões, o mesmo se aplica. Nesse caso, a previsão com maior chance de ser verdadeira no longo prazo depende das probabilidades estatísticas das possibilidades do evento a ser previsto. Suponha que devamos prever quais dos k resultados possíveis e_1, \dots, e_k de um experimento aleatório ocorrerão e que as possíveis regras de predição que devemos escolher sejam todas da forma “Prever e_1 com frequência h_1 , e_2 com frequência h_2, \dots (etc.)”, para frequências de predição variadas h_1, \dots, h_k . Então (assumindo a independência física do experimento aleatório das previsões), a regra de predição que sempre prevê (ou seja, com frequência 1) o resultado de probabilidade máxima leva ao sucesso máximo da verdade no longo prazo.²⁰ Essa regra de

²⁰ Prova: Devido à independência, o resultado verdadeiro de nossas previsões é dado como $h_1 \cdot p(e_1) \dots + h_n \cdot p(e_n)$. Este valor é máximo sse para o (ou para um) evento e_k com

predição também é chamada a *regra do máximo* (cf. Reichenbach, 1938, p. 310; Greeno, 1970). Para aplicar a *regra do máximo*, deve-se conhecer as probabilidades estatísticas dos eventos ou, pelo menos, as suas proporções. Se eu considerar o evento e_1 mais provável que o e_2 , mas o limite de frequência de e_1 for menor que o de e_2 , a *regra do máximo* me levará a um sucesso de previsão abaixo do ideal.

Em resumo, estimativas precisas de probabilidades estatísticas estão no centro de todos os métodos de previsão e ação bem-sucedidos. Também pode ser demonstrado que otimizamos o sucesso das nossas previsões e ações se condicionarmos as nossas probabilidades estatísticas à classe de referência mais estreita no sentido do princípio de Reichenbach (Def. 2-3) (ver Capítulo 7.8).

5.2 Problemas de definição

A principal dificuldade do conceito estatístico de probabilidade reside em encontrar uma definição adequada de probabilidade estatística. Em primeiro lugar, o conceito de limite de frequência é uma idealização teórica, porque, na realidade, nenhum experimento aleatório pode ser repetido infinitamente, pelo menos não em nosso mundo – afinal, todo dado desgasta-se com o tempo. Quando se fala sobre limites de frequência, faz-se uma declaração contrafactual da seguinte forma:

(5-5) $p(Fx) = r$ significa: se alguém repetisse o experimento aleatório subjacente (com resultado possível F) infinitamente, as frequências de F convergiriam para o valor limite r.

probabilidade máxima $p(e_k)$ (entre os $p(e_1), \dots, p(e_n)$) h_k tem o valor 1 e os h_i restantes, $i \neq k$, tem o valor 0.

A afirmação (5-5) é semelhante a uma lei sobre a disposição do experimento aleatório subjacente ou do processo físico descrito por ele. Afirmações contrafactuais desse tipo nunca podem ser verificadas ou falsificadas por meio de observações de frequências finitas, mas podem, na melhor das hipóteses, ser confirmadas indutivamente, o que será discutido em breve. *Prima facie*, no entanto, isso ainda não é um problema sério, uma vez que, de acordo com ideias teórico-científicas difundidas, mesmo declarações (leis) estritas como “O açúcar é solúvel em água” são semanticamente explicadas por condicionais contrafactuais; nesse exemplo, pela condicional “Sempre que alguém colocasse o açúcar na água, ele se dissolveria nela” (cf. Schurz, 2006, Capítulo 3.10.1). Essa afirmação condicional permanece verdadeira ainda que o açúcar nunca tenha sido adicionado à água e, nesse caso, não poderia ser verificada ou falsificada, mas só poderia ser inferida indutivamente a partir de observações anteriores realizadas com o açúcar.

Um segundo problema que não tem contrapartida no caso estrito é muito mais difícil, a saber: ao realizar repetidamente um experimento aleatório, um número potencialmente infinito de sequências de resultados potencialmente crescentes infinitamente $(e_1, e_2, \dots) \in \Omega_\infty$ pode ser produzido. Existem muitos dados aparentemente simétricos que diferentes pessoas podem lançar. Existe a sequência total de todos os lançamentos de todas as pessoas, então, cada pessoa produz sua própria sequência e cada dado tem sua própria sequência, mas também se pode olhar todos os lançamentos de dados nos meses de janeiro, etc. Por que todas essas sequências (idealizadas) deveriam ter o mesmo limite de frequência $p(Fx)$ e por que todas elas deveriam ter um limite de frequência? O problema é que os limites de frequência dependem da ordem dos eventos em uma

determinada sequência de eventos. Se forem permitidas permutações arbitrárias (rearranjos) ou seleções de posição de uma determinada sequência de eventos, o limite de frequência pode mudar drasticamente. Seja, por exemplo, $(1,0,0,1,1,0,1,0,\dots)$ uma sequência aleatória com $p(\{1\}) = 1/2$, então, é possível reorganizar a sequência de modo que se tenha cinco uns e, depois, o primeiro zero e, depois, os próximos cinco uns e, depois, o próximo zero, etc. Isso apenas permuta a sequência original e, ainda assim, o limite de frequência na sequência permutada é agora $5/6$. Como existem infinitos números de unidades na sequência original, os números necessários para esse propósito podem nunca acabar.²¹ Da mesma maneira, pode-se usar a permutação para formar uma sequência que não tem limite de frequência, porque a frequência de um oscila eternamente entre $1/3$ e $2/3$. Para isso, como no exemplo (5-1), primeiro pegamos um número suficiente de uns em sequência a fim de aumentar a frequência alcançada $h_n(1)$ para pelo menos $2/3$, então, depois, pegamos um número suficiente de zeros em sequência para reduzir a frequência $h_n(1)$ novamente para $1/3$, etc. É ainda mais fácil gerar essas sequências “estranhas” selecionando posições em vez de permutações: por exemplo, de uma dada sequência aleatória, simplesmente selecionamos apenas aquelas posições com o resultado “1” e obtemos uma sequência estrita de uns.²²

²¹ Ao colocar 2^k unidades após cada k -ésimo zero, é possível até mesmo aumentar o limite de frequência de unidades para 1 por meio de reorganização.

²² Como toda sequência infinita com um limite de frequência diferente de 0 e 1 ainda contém, após qualquer número (finito) de posições, infinitos zeros e infinitos uns, pode-se, a partir dela, gerar qualquer outra sequência com um limite de frequência diferente de 0 e 1 – tanto por permutações quanto por seleções de posições. Por meio de seleções de posições – mas não por permutações –, é possível também eliminar elementos da sequência e, assim, gerar sequências compostas apenas de uns ou apenas de zeros. Aplicado a sequências infinitas, isso significa que, por seleções de

É claro que as sequências geradas por tais transformações dependentes de resultados (permutações ou seleções de posição) de uma sequência aleatória não seriam mais chamadas de sequências aleatórias, porque a aplicação da transformação pressupõe que já se saiba qual resultado foi produzido em qual posição. No entanto, não seria possível que sequências tão estranhas de resultados também pudessem ser alcançadas com uma moeda normal, como uma coincidência quase astronomicamente improvável? Essa é uma questão controversa. A (o que chamo de) teoria estatística ingênua responde a essa questão com a lei (forte) dos grandes números e argumenta o seguinte: a afirmação “ $p(Fx) = r$ ” não diz, de acordo com essa lei, que em todos os números aleatórios de sequências o limite de frequência de Fx é r , mas apenas que é r com probabilidade 1. Contudo, os críticos argumentam acertadamente contra essa tentativa de definição, alegando que ela é circular: ao definir a expressão “a probabilidade de Fx é r ” a frase “com probabilidade 1” aparece novamente. Conforme tal proposta, as probabilidades não são reduzidas a limites de frequência, mas, em última análise, a probabilidades.²³

Dado que ocasionalmente se ouve a opinião de que as definições circulares não são tão “inúteis” como tradicionalmente se afirma, deve ficar claro aqui o porquê de a definição circular ser inútil: porque não determina o conceito de probabilidade, nem sequer limita o seu conteúdo e, portanto, as frequências não se conectam com ele. “Probabilidade”, segundo a definição circular, poderia igualmente

posições, podem-se construir ainda mais sequências do que por permutações; por essa razão, von Mises (1964) restringiu-se à consideração das seleções de posições.

²³ Veja Skyrms (1980, p. 29), Eagle (2004, p. 396), Hájek (1999, p. 223), Kutschera (1972, p. 104). Stegmüller (1973b, p. 37) vê isso como uma “objeção fatal”.

significar “o grau em que um determinado sujeito racional deseja algo”, com a desejabilidade racional obedecendo aos axiomas de Komolgorov. Então, a lei forte dos grandes números afirma: esse sujeito racional deseja ao máximo que as frequências de um evento convirjam para o grau de sua desejabilidade – uma afirmação que nem sequer toca na conexão real entre probabilidades e limites de frequência.

Por outro lado, caso se tente transformar a condição “ $p = 1$ ” novamente usando a lei dos grandes números, então, obter-se-á a definição para a afirmação “ $p(Fx) = r$ ” com probabilidade 1 em uma sequência aleatória de sequências aleatórias, sendo esse o limite de frequência daquelas sequências com limite de frequência $p(Fx) = r$ igual a 1”. De todo modo, a condição circular “com probabilidade 1” não é eliminada por essa proposta, mas apenas adiada por um nível de iteração. Chamo isso de problema de iteração e o considero um subcaso do problema de circularidade. Finalmente, é preciso lembrar também que a lei dos grandes números expressa uma verdade puramente matemática. Portanto, não pode expressar uma propriedade específica das probabilidades estatísticas, embora seja válida (como explicado no Capítulo 3.2), na teoria subjetiva das probabilidades, como uma afirmação sobre graus “racionais” de crença em relação aos limites de frequência de sequências aleatórias.

Uma saída para o problema da circularidade é desenvolvida por von Mises (1964). Von Mises limita-se à suposição de uma única sequência básica de realizações experimentais – pode-se, por exemplo, imaginar a sequência de todos os lançamentos de dados do mesmo tipo físico já feitos, organizados no tempo e hipoteticamente estendidos ao futuro infinito. Os termos $i \in \mathbb{N}$ correspondem, portanto, a situações concretas de lançamentos de dados, organizadas no

tempo. Von Mises chama essa sequência básica de coletivo estatístico. As consequências individuais reais são caracterizadas por von Mises por meio do conceito de seleção de posição independente de resultados. A independência do resultado de uma função de seleção de posições é explicada, de acordo com a continuação da teoria de von Mises por Wald e Church, usando o conceito de função computável (algoritmicamente).²⁴ Uma seleção de posições previsível e independente do resultado é chamada de seleção de posições permissível. A seguir, $(g:\mathbb{N}\rightarrow\Omega) = (e_1, e_2, \dots)$ denota a sequência básica assumida, $g\uparrow n \stackrel{\text{def}}{=} (e_1, \dots, e_n)$ denota a sequência inicial de n membros da sequência básica, e s_i (para $i \in \mathbb{N}$) variam entre as seleções de posição de g . Então se define:

(Def. 5-1) Uma *seleção de posição admissível* s de g é uma função computável que, aplicada a qualquer posição $n \in \mathbb{N}$ de g , diz se aquela posição deve ser selecionada (+) ou não (-). Os resultados anteriores da sequência, $g\uparrow(n-1)$, mas não o resultado e_n , atuam como uma entrada adicional para s , ou seja, como uma posição de argumento de s .²⁵ Seja $s(g)$ a sequência que surge da seleção das posições s de g .

Exemplo: Seja s_1 a seleção de dígitos que seleciona todos os dígitos ímpares e s_2 a seleção de dígitos que seleciona todos os dígitos após “1”. Então:

²⁴ Veja von Mises (1964, cap. 1); Howson/Urbach (1996, p. 324); Church (1940); Kutschera (1972, p. 140); Salmon (1984, p. 57); Gillies (2000, 105-9).

²⁵ Expresso formalmente, $s: \{(n, g\uparrow(n-1)): n \in \mathbb{N}\} \rightarrow \{+, -\}$.

$$s_1((0,1,1,0,1,0,0,1,1,1,0\dots)) = (0,1,1,0,1,0,\dots),$$

$$s_2((0,1,1,0,1,0,0,1,1,1,0\dots)) = (1,0,0,1,1,0,\dots).$$

A restrição a funções computáveis em (Def. 5-1) é essencial, porque é claro que existem funções – no sentido teórico-abstrato dos conjuntos de sequências infinitas de pares argumento-valor – que selecionam exatamente aquelas posições da sequência básica e que são únicas ou que possuem alguma outra propriedade. Contudo, no caso de uma sequência aleatória genuína, não se pode calcular esses dígitos sem conhecer o seu resultado.

Com base na (Def. 5-1), von Mises (1964, Capítulo 1) define o seu conceito central do coletivo estatístico da seguinte forma:

(Def. 5-2) (5-2.1) Uma sequência básica $g:\mathbb{N}\rightarrow\Omega$ é uma sequência *estatística* básica (um “coletivo estatístico”) sse satisfaz as duas condições a seguir:

- (a) (*Condição de convergência*). Todo resultado possível (disjuntivo) E , na álgebra AL sobre Ω (expresso por uma fórmula aberta), tem um limite de frequência em g que identifica a probabilidade $p(E)$, e
- (b) (*Condição de aleatoriedade*): esse limite de frequência é insensível às seleções de posições admissíveis, ou seja, tem o mesmo limite de frequência em todas as subsequências $s(g)$ geradas pelas seleções de posição admissíveis s .

(5-2.2) Todas as subsequências de g , obtidas por meio de seleções de posições admissíveis, são chamadas de *sequências aleatórias*.

No caso de espaços de possibilidades finitos, é suficiente exigir que todos os resultados completos ($e \in \Omega$) tenham limites de frequência; disso segue (como mostrado na Seção 5.1) a existência de

limites de frequência para todas as disjunções finitas de resultados completos e, portanto, para todos os elementos da álgebra de potências sobre Ω .

Ambas as condições (a) e (b) da Def. 5-2.1 não se aplicam a priori, mas devem ser entendidas como declarações disposicionais sobre a natureza real do experimento aleatório subjacente. Com a definição (5-2), von Mises mata dois coelhos com um só golpe. Primeiro, ele define o conceito de sequência aleatória de uma forma natural: sequências aleatórias são todas sequências selecionadas de g por meio de escolhas de posição admissíveis. Em segundo lugar, isso garante a condição de independência (estatística) das repetições do experimento aleatório (von Mises, 1964, p. 27). A prova desse fato é brevemente esboçada: sejam $F_x, G_x \in AL$ dois eventos (representados linguisticamente) em AL , então, $p(F_{x_1} \wedge G_{x_2})$ é, por definição, o limite de frequência de $F_{x_1} \wedge G_{x_2}$ em uma sequência básica $g^2: \mathbb{N} \rightarrow \Omega^2$ de execuções duplas, por exemplo, execuções sucessivas do experimento aleatório. No entanto, a sequência básica simples g já contém uma sequência básica do tipo g^2 – nomeadamente, a sequência de todos os pares $((e_k, e_{k+1}): k \in \mathbb{N})$ de termos sucessivos –, ou seja, uma sequência $((e_1, e_2), (e_2, e_3), \dots)$. O limite de frequência do par de eventos (F_{x_1}, G_{x_2}) , nessa sequência infinita de pares, é determinado da seguinte forma: na sequência básica g , selecionam-se todas as localizações cuja localização predecessora realizou o resultado F_x e considera-se a subsequência resultante selecionada pela localização $s(g)$ a frequência de G_x . Se “ $h_n(F_x \text{ em } f)$ ” representa a frequência de F_x nos primeiros n termos da sequência f , então:

(5-6) *Prova de independência de von Mises:*

Temos $h_n((F_{x_1}, G_{x_2}) \text{ em } g) = h_n(Fx \text{ em } g) \cdot h_{s(n)}(Gx \text{ em } s(g))$, com $s(n) =_{\text{def}} h_n(Fx) \cdot n$.²⁶

Disso, segue: $\lim_{n \rightarrow \infty} h_n((F_{x_1}, G_{x_2}) \text{ em } g) = \lim_{n \rightarrow \infty} h_n(Fx \text{ em } g) \cdot \lim_{n \rightarrow \infty} h_{s(n)}(Gx \text{ em } s(g))$. Como a seleção de posição s é permitida, tem-se: $\lim_{n \rightarrow \infty} h_{s(n)}(Gx \text{ em } s(g)) = \lim_{n \rightarrow \infty} h_n(Gx \text{ em } g)$. Isso resulta em independência: $p(F_{x_1} \wedge G_{x_2}) = p(Fx) \cdot p(Gx)$.

Considerações análogas aplicam-se a experimentos aleatórios combinados n vezes: para cada n , a sequência básica g contém a sequência infinita de todas as sequências de n termos $((e_k, \dots, e_{k+n-1}): k \in \mathbb{N})$.²⁷ Isso permite a derivação da fórmula binomial dentro da estrutura von Misesiana. Além disso, a sequência básica contém até mesmo a sequência de deslocamento infinito de todas as sequências infinitas ou as “seções à direita” de g , cada uma começando com os termos 1, 2, 3... etc., ou seja, $((e_{k+i}: i \in \mathbb{N}): k \in \mathbb{N})$ (com $(e_{k+i}: i \in \mathbb{N}) =_{\text{def}} (e_k, e_{k+1}, e_{k+2}, \dots)$). Assim, o limite de frequência de eventos em Ω_∞ , ou seja, o limite de frequência de subsequências infinitas da sequência básica com certas propriedades, também é fixo, o que significa que as leis dos grandes números também podem ser interpretadas na estrutura de von Mises.

Foi objetado a von Mises que sua condição de aleatoriedade (Def. 5-2(b)) é desnecessariamente forte: é suficiente exigir apenas que quase todas as subsequências resultantes da seleção de posição admissível converjam para o limite da sequência básica, onde o termo

²⁶ Para explicar: Se e_n é um termo de g com predecessor $F(e_{n-1})$ e número de dígitos n , então $s(n) =_{\text{def}} h_n(Fx) \cdot n$ é o número de dígitos que este termo tem na subsequência $s(g)$, e que é idêntico ao número de Fs até o $(n-1)$ ésimo termo de g .

²⁷ Para $n = 4$, então $((e_1, e_2, e_3, e_4), (e_2, e_3, e_4, e_5), (e_3, e_4, e_5, e_6), \dots)$.

“quase todas” deve ser entendido como “com probabilidade 1” (Kutschera, 1972, p. 101). Uma vez que essa “melhoria” levaria diretamente de volta ao problema da circularidade já explicado, abstermo-nos de o fazer. Ainda chegaremos à possibilidade alternativa de interpretar “quase tudo” como probabilidade epistêmica.

Não são adotadas aqui duas peculiaridades da teoria de von Mises. Primeiro, em seu trabalho posterior, von Mises exige que os limites de frequência dos elementos de Ω sejam σ -aditivos (1964, p. 12). Pelas razões explicadas em 3.4, vemos isso apenas como uma suposição possível, mas não como uma condição necessária. Em segundo lugar, devido a “escrúpulos verificacionistas”, von Mises restringe a álgebra $AL(\Omega^\infty)$ aos chamados conjuntos de Jordan (1964, p. 59-92) e, portanto, rejeita a lei forte dos grandes números (1964, p. 240, n. 7). Não adotamos essa restrição, mas escolhemos a álgebra de Borel usual (Seção 8.6) como uma álgebra sobre Ω^∞ , onde codificamos Ω^∞ por um intervalo de números reais (o que é sempre possível para Ω finito).²⁸ Assim, em von Mises, a lei forte dos grandes números também pode receber significado dentro dessa estrutura. Se essa lei for restrita a uma sequência infinita de sequências infinitas (f_1, f_2, \dots) , todas elas geradas por seleções de posições admissíveis, então, essa lei só tem conteúdo trivial, porque para cada sequência f_i aplica-se: na abordagem von Misesiana, por definição, seus limites de frequência para os elementos da álgebra concordam com os limites de frequência na sequência básica g . Então, o limite de frequência das sequências que concordam estatisticamente com g é trivialmente 1. Entretanto, a lei forte dos grandes números também pode ser provada para

²⁸ Se Ω tem 2^k elementos e cada termo da sequência é representado por um número de k dígitos na representação binária, então Ω^∞ é mapeado bijetivamente para o intervalo de números reais $[0,1]$ na representação binária (cf. Kelly 1996, p. 57).

seqüências de deslocamento de seqüências não computáveis, desde que seus termos iniciais formem uma escolha permissível de posições:

(Teorema 5-1) *Leis dos grandes números na estrutura de von Mises* (apêndice de prova 10.3.6):

Seja $\kappa =_{\text{def}} (k_i; i \in \mathbb{N})$ uma seleção não computável de posições da seqüência básica g (ordenada por tamanho: $k_i \leq k_j$ sse $i \leq j$), e seja $(f_n; n \in \mathbb{N})$ uma seqüência de seqüências cujos termos iniciais são $f_n(1) =_{\text{def}} a_n$ e são definidos por uma seleção de posição permitida $\alpha =_{\text{def}} (a_n; n \in \mathbb{N})$, bem como cujos termos adicionais surgem adicionando κ a a_n : $f_n = (a_n + k_i; i \in \mathbb{N})$. Então, para cada evento $E \in \text{AL}$ com $p(E) =_{\text{def}} p$:

(5-1.1) *Lei fraca*: Para cada ϵ , não importa quão pequeno seja, o limite de frequência dessas seqüências de m termos $((a_n + k_i; i \leq m); n \in \mathbb{N})$, cuja frequência E desvia-se de p , em menos que ϵ , para $m \rightarrow \infty$, tende em direção a p .

(5-1.2) *Lei forte*: Se p é σ -aditivo sobre AL , então, o limite de frequência daquelas seqüências em $(f_n; n \in \mathbb{N})$ cujo limite de frequência E que concorda com p é 1 .

A seqüência κ no Teorema 5-1 pode ser escolhida de modo que sua adição, por exemplo, para o primeiro termo inicial a_n , possua um valor limite de frequência que se desvie de p (por exemplo, apenas uns); no entanto, o limite de frequência de tais seqüências (entre as seqüências deslocadas $a_n + \kappa$) é zero de acordo com o Teorema 5-1.2. Dessa forma, as leis dos grandes números também possuem conteúdo não trivial na abordagem von Misesiana.

Chamamos a caracterização von Misesiana da aleatoriedade (ou da irregularidade) de uma seqüência de aleatoriedade interna, porque

as seleções de posição que são insensíveis aos limites de frequência dependem apenas de eventos anteriores dentro da sequência ou do seu espaço de resultados Ω . Reichenbach (1935, p. 148 e seg.; 1949, Capítulo 32) e Salmon (1984, p. 61) estendem essa caracterização a seleções de posição que também podem depender de eventos externos anteriores. Pode-se formalizar isso assumindo um espaço de possibilidades refinado Ω^* , cujos elementos consistem em combinações de Ω resultados com eventos externos (epistemicamente acessíveis) ao mesmo tempo. A sequência básica correspondente de eventos Ω^* é denotada por g^* . Se Ω^* contém todos os tipos de eventos externos que são prognosticamente relevantes para os resultados de Ω , e se os limites de frequência em g também são insensíveis às seleções de posições que dependem de eventos Ω^* anteriores a g^* , então, g não é apenas interno, mas também aleatório objetivo, porque o resultado do experimento aleatório Ω não pode ser previsto por nenhuma informação (epistemicamente acessível). Por exemplo, o processo de lançar um dado (com uma distância de lançamento suficientemente longa) é um processo aleatório objetivo. O processo de colocar deliberadamente um dado, por outro lado, também pode gerar uma sequência internamente aleatória de resultados, mas isso não é objetivamente aleatório, porque o resultado pode ser previsto pela decisão volitiva da pessoa. Se os resultados de um experimento aleatório Z são objetivamente aleatórios, então, o predicado “ x é um resultado de Z ” atua como classe de referência objetivamente mais estreita estatística e causalmente relevante para cada resultado individual.

Na Seção 5.4, explicaremos que o conceito de aleatoriedade objetiva não pressupõe necessariamente a suposição do indeterminismo metafísico, mas também pode ser explicado pelo que

é chamado de “instabilidade determinística” na física. Conforme explicado em (5-5), o conceito estatístico de probabilidade assim explicado é um conceito disposicional. Com Popper (1959), também podemos descrever a probabilidade estatística $p(Fx)$ como a propensão genérica do experimento aleatório subjacente para produzir o resultado Fx . Concordamos com Howson e Urbach (1996, p. 338) que essa visão disposicional já era inerente à teoria de von Mises. Os experimentos aleatórios são tipos de processos repetíveis, e as propensões genéricas de Popper (1959) referem-se às disposições de tais tipos de processos para produzir sequências aleatórias com certos limites de frequência (cf. Gilles, 2000, p. 114-8). As chamadas propensões singulares, introduzidas por Popper (1990) e discutidas na Seção 5.5, comportam-se de maneira completamente diferente das propensões genéricas. No entanto, a visão disposicional explícita tem uma vantagem em relação à abordagem de von Mises: as propensões genéricas manifestam-se apenas em limites de frequência de sequências infinitas; contudo, elas não são propriedades dessas sequências, mas sim propriedades do experimento aleatório subjacente. Caso seja uma sequência aleatória objetiva, então, faz sentido entender a propensão genérica do tipo de experimento como um tropo de propensão a cada implementação individual do experimento (no início da implementação, antes que o resultado seja conhecido), porque a adesão do experimento aleatório fornecido funciona aqui como classe de referência objetiva mais estreita estatística e causalmente relevante. Entende-se por tropo a realização individual de uma propriedade repetível²⁹ – nesse caso, a propriedade

²⁹ No chamado nominalismo de tropos, os universais de propriedade são explicitados como classes de tropos semelhantes a tipos (ver Loux, 1998, capítulos 1-2).

de entregar um resultado de um experimento aleatório com uma certa propensão.

Isso resolve a seguinte objeção de Popper (1959) a von Mises: se o k -ésimo lançamento de uma série infinita de lançamentos de moedas foi lançado desviando-se do resto com uma moeda irregular, com uma tendência de $2/3$ para coroa, isso não muda o limite de frequência de $1/2$ do número na sequência, mas a propensão correta do k -ésimo lançamento da moeda para o número não é $1/2$, mas $2/3$. Na nossa terminologia, tal sequência “mista” não é uma sequência objetiva, mas sim uma sequência aleatória interna.

A reconstrução da abordagem de von Mises, apresentada até aqui, resolve quase todas as objeções às probabilidades frequentistas conhecidas na literatura. Para ilustrar isso, explico agora a proeminente lista de Hájek (1999) de quinze objeções ao conceito estatístico de probabilidade. As objeções de Hájek 13-15 dizem respeito às da Seção 3.4 sobre o fato de que os limites de frequência podem ocasionalmente violar a condição de σ -aditividade. Pelas razões já explicadas, esse fato não representa uma objeção real à teoria estatística da probabilidade, mas antes uma visão positiva dessa teoria. As objeções 10-12 de Hájek dizem respeito a problemas marginais que não são importantes para o nosso contexto. Suas objeções 4-9, por sua vez, têm todas a ver com o fato discutido nesta seção de que, a partir de sequências dadas, sequências com limites de frequência diferentes ou mesmo ausentes podem ser construídas a partir de sequências dadas por meio de rearranjos e de seleções de posição. Todas essas objeções são abordadas satisfatoriamente pela teoria das sequências aleatórias de von Mises. Infelizmente, não há espaço aqui para citações detalhadas de Hájek (1999) – o leitor pode verificar minhas afirmações lendo Hájek.

Passo agora às objeções 1-3 de Hájek, que requerem uma consideração mais profunda além do que já foi dito. A objeção número 1 afirma que a visão da probabilidade estatística como uma disposição contrafactual (uma vez que sequências aleatórias infinitas são irrealizáveis) equivale a abandonar o empirismo tradicional. Vemos isso como uma visão válida e não como uma objeção, porque a epistemologia do empirismo tradicional teve de ser abandonada na filosofia da ciência por muitas razões ou ser enfraquecida para um “empirismo mínimo” (ver Schurz, 2006, Capítulo 2). Tal como explicado no início da Seção 5.2, leis estritas também são explicadas como disposições contrafactuais: um pedaço de açúcar é solúvel em água, mesmo que nunca tenha sido colocado na água durante a sua existência. Da mesma forma, os resultados dos experimentos de lançamento de moeda convergem para o limite de frequência de $1/2$, ainda que nenhuma moeda seja lançada um número infinito de vezes.

Contudo, é essencial para a natureza física de disposições, como a solubilidade em água ou como a tendência estatística, que a explicação das condicionais contrafactuais subjacentes refira-se a todos os mundos fisicamente possíveis nos quais certas condições contrafactuais (mas fisicamente legais) sejam realizadas. Essas condicionais, portanto, não dependem de “ordens de similaridade” questionáveis entre mundos fisicamente possíveis, como geralmente fazem as condicionais Lewisianas (Lewis, 1973). Assim como um pedaço de açúcar colocado na água dissolve-se em todos os mundos fisicamente possíveis à temperatura e pressão normais, a frequência de caras numa experiência de lançamento de moeda converge para $1/2$ em todos os mundos fisicamente possíveis – ou pelo menos em “quase todos” (cf. abaixo). Em contraste, a aceitação de condicionais contrafactuais como “se você não tivesse se casado comigo, isso e

aquilo teria acontecido” depende de ordens de similaridade entre condições fisicamente possíveis para as quais não existem critérios objetivos suficientes, razão pela qual a aceitabilidade de tais condicionais é um componente subjetivo e convencional (para críticas, cf. Schurz, 2013b, Capítulo 6.6.3).

É nesse contexto que a objeção número 2 de Hájek torna-se relevante. Ele afirma que sequências aleatórias infinitas não só não existem de fato, mas também são fisicamente impossíveis, uma vez que todos os mundos fisicamente possíveis são finitos. Não sei se Hájek está correto aqui, já que universos em expansão e em contração infinitas não são fisicamente excluídos, sobretudo de acordo com o estado atual do conhecimento. Entretanto, mesmo que todos os mundos fisicamente possíveis fossem finitos, essa objeção não seria realmente um problema: afinal, existem infinitos mundos fisicamente possíveis, e se pode construir infinitas sequências aleatórias por combinações aleatórias de sequências aleatórias finitas de infinitos mundos fisicamente possíveis.

A única objeção realmente problemática, a meu ver, é a objeção número 3 de Hájek, que tem sido apresentada, por muitos teóricos da probabilidade, contra a explicação da probabilidade como um limite de frequência e que afirma o seguinte: ao contrário da suposição de von Mises, não há apenas uma sequência infinita fisicamente possível de realizações experimentais, em vez disso, há infinitas sequências básicas que são fisicamente possíveis e extensões de uma sequência inicial finita para o futuro indefinido.³⁰ Parece arbitrário, como fez von Mises, destacar uma dessas consequências e a declarar como

³⁰ No sentido de von Mises, é possível unir todas as consequências do nosso mundo em uma sequência geral organizada no espaço-tempo, mas não é possível fazer isso para consequências em diferentes mundos ou espaços-tempos possíveis.

consequência básica. A suposição de von Mises baseada na existência de apenas uma sequência básica infinita ideal, da qual todas as outras são derivadas por escolha admissível de posições, parece ser uma suposição quase insustentável a esse respeito.

Há, a meu ver, apenas dois métodos viáveis para responder a essa única objeção fundamental à teoria de von Mises:

Método 1: Assume-se que, em rigorosamente *todas* as continuações fisicamente possíveis de sequências aleatórias reais, as frequências convergem para o mesmo limite. Do ponto de vista probabilístico, entretanto, existem sequências aleatórias com limiar de frequência desviante ou inexistente; eles só têm probabilidade zero. Essa suposição implicaria, portanto, que o conceito de fisicamente possível aqui utilizado exclui a existência de mundos ou de sequências aleatórias com probabilidade zero. Isso equivale a uma suposição indutiva fraca: mundos de probabilidade zero são considerados fisicamente impossíveis.

Método 2: Ou mundos de probabilidade zero ou sequências aleatórias são considerados como fisicamente possíveis e as probabilidades estatísticas são explicadas como limites de frequência em que as sequências aleatórias possuem “probabilidade 1”. Se essa probabilidade fosse estatisticamente entendida de novo, isso tornaria a explicação do conceito de probabilidade circular, como já explicado. Portanto, é necessário definir o conceito de probabilidade na frase “possuem probabilidade 1” de forma independente, não estatística, mas sim epistêmica. Essa proposta foi apresentada por Kolmogorov (1933) e foi desenvolvida por Cramér (1946) na forma de seu conceito de “segurança prática” (cf. Gillies, 2000, p. 161). De acordo com tal proposta, a frase “com probabilidade 1” significa que, como sujeitos racionais, estamos certos (isto é, acreditamos com probabilidade

subjetiva 1) de que as frequências em sequências aleatórias arbitrárias convergem contra o limite de frequência ancorado na disposição do experimento aleatório. Isso também equivale a uma suposição indutiva, mas dessa vez não está escondida no conceito físico de possibilidade, mas é explicitada epistemicamente.

O Método 1 está (não mais textualmente, mas ainda assim) no espírito de von Mises e parece ter o mérito de levar a uma abordagem *puramente* estatística de probabilidade. A frase “com probabilidade 1”, conforme a lei forte dos grandes números, é interpretada estatisticamente no Método 1; no entanto, a afirmação da lei permanece trivial, uma vez que sequências aleatórias zero-prováveis são excluídas da classe de possibilidades físicas. Em contraste, o Método 2 leva a um conceito dualista de probabilidade, porque a definição de probabilidades estatísticas, nesse sentido, refere-se ao conceito de certeza epistêmica (como um caso especial de probabilidade epistêmica).

Contudo, a vantagem da *pureza* estatística do Método 1 é apenas superficial. Assim que se volta para a questão do *conteúdo empírico das declarações estatísticas de probabilidade*, é necessário fazer suposições indutivas na forma de pressupostos de probabilidade epistêmica. Portanto, adotamos o Método 2 como o método mais honesto e defendemos um conceito estatístico *dualista* de probabilidade: isto é, um conceito estatístico de probabilidade capaz de integrar pressupostos epistêmicos de probabilidade até certo ponto, mas sem contemplar todo o arcabouço do bayesianismo (Cap. 9). Passemos agora ao tema do conteúdo empírico.

5.3 Conteúdo empírico

Uma declaração estatística de probabilidade é uma *hipótese real*. No entanto, qual é exatamente o seu conteúdo empírico e como ele é testado empiricamente? O *problema com o conteúdo empírico* é não haver nenhuma afirmação observacional que logicamente decorre de uma afirmação sobre o limite de frequência: o fato de que um evento tipo E em uma sequência aleatória tenha um certo valor limite de frequência r é, para cada membro da sequência, não importa quão tarde (em um ponto tão tardio no tempo), n e é logicamente compatível com qualquer valor de frequência $h_n(E) \neq r$ alcançado até aquele ponto, uma vez que os n membros iniciais não têm importância em comparação aos infinitos membros seguintes, de acordo com a posição n. Matematicamente expresso: sendo " $a_n(E)$ " o número de resultados de E após n termos sequenciais, $\lim_{n \rightarrow \infty} a_n(E)/n = r$ também segue $\lim_{n \rightarrow \infty} (a_n(E)+k)/n = r$ para cada constante $k \in \mathbb{R}$.³¹ Esse fato é frequentemente formulado pelos bayesianos como uma objeção à teoria estatística da probabilidade (cf. Howson e Urbach, 1996, p. 331). Mas isso apenas expressa o fato bem conhecido de que nenhuma declaração de observação decorre logicamente de hipóteses estatísticas sobre intervalos infinitos de indivíduos.

O problema do conteúdo empírico não é resolvido satisfatoriamente nem por von Mises e nem por von Reichenbach. Ambos os autores estão satisfeitos com o critério de convergência analiticamente válido, segundo o qual para cada ϵ haverá algum ponto no tempo ou alguma consequência, por menor que seja, a partir da

³¹ *Prova:* $\lim_{n \rightarrow \infty} (f(n)+k)/n = \lim_{n \rightarrow \infty} f(n)/n + \lim_{n \rightarrow \infty} k/n = \lim_{n \rightarrow \infty} f(n)/n$, então $\lim_{n \rightarrow \infty} k/n = 0$.

qual a frequência se desviará de seu limite menor que ϵ (von Mises, 1964, 59; 91; von Reichenbach, 1949, Cap. 11). No entanto, esse critério de convergência nos deixa no escuro sobre o quão próximos estaríamos do valor-limite para cada conseqüente (ver Lenz, 1974, p. 99 e seg.).

O fato permanece: hipóteses estatísticas não têm conseqüências lógicas observáveis pelas quais possam ser testadas. Seguindo Popper (1935, Cap. 35) e Hempel (1965, p. 211), o conteúdo empírico de uma hipótese é convencionalmente definido como o conjunto de Teoremas observacionais logicamente implicados por ela. De acordo com essa definição, hipóteses estatísticas não têm conteúdo empírico e não são verificáveis. Para atribuir conteúdo empírico a hipóteses estatísticas, o conteúdo empírico deve ser estendido a conseqüências empíricas epistemicamente *prováveis*. Isso é feito com a ajuda do método descrito na Seção 4.4: *uma inferência de especialização indutiva* em que o limite de frequência das frequências amostrais é transferido, como uma probabilidade racional de crença, para amostras individuais, reais ou potenciais. Nesse sentido, a probabilidade de obter caras, em um experimento de lançamentos de moeda, com uma moeda regular, em 10.000 arremessos, entre 4.900 e 5.100 caras, é de 95% (ver Seção 8-4). Essa afirmação pertence ao conteúdo indutivo-empírico da hipótese estatística $p(\text{cara}) = 1/2$ (ver Seção 7.2 para mais detalhes).

O conteúdo empírico das hipóteses estatísticas não é, portanto, dedutivo, mas indutivo por natureza e requer certos pressupostos da teoria epistêmica da probabilidade para explicitá-la. Esse conteúdo indutivo-empírico é a base de toda a análise descrita no Capítulo 8, que se baseia nessa ideia básica: uma hipótese estatística H só pode ser aceita se as frequências observadas não forem muito improváveis sob a suposição da verdade *de H*.

O importante para o problema deste capítulo (a definição estatística de probabilidade) é a seguinte consequência: mesmo que se permaneça “estatisticamente puro” e se aplique o Método 1, as probabilidades epistêmicas devem ser usadas, o mais tardar, ao explicitar o conteúdo empírico. E *independentemente de* o Método 1 ou *de* o Método 2 serem usados para definir probabilidades estatísticas, o conteúdo empírico-indutivo de uma hipótese estatística permanece exatamente o mesmo, porque possibilidades com probabilidade epistêmica zero não desempenham um papel e podem sim ser epistemicamente negligenciadas. A probabilidade de acreditar que se vai obter caras, entre 4.900 e 5.100 caras, em 10.000 lançamentos, com uma moeda regular, é de 95%, independentemente de os lançamentos de moeda de probabilidade zero serem considerados possíveis ou impossíveis. Essa é outra razão para a superioridade do conceito dualista de probabilidade.

5.4 Aleatoriedade objetiva, determinismo e indeterminismo

Na Seção 5.2, tem-se uma sequência aleatória objetiva como uma sequência de resultados de um experimento aleatório realizado repetidamente, cujo limiar de frequência não pode ser alterado por qualquer seleção de posição calculável que se refira apenas a eventos anteriores (internos ou externos à sequência). Com base nessas informações, e tal como explicado no ponto 5.1, os resultados completos podem, então, ser previstos em Ω sem uma taxa de sucesso a longo prazo superior ao máximo dos seus limites de frequência. Por exemplo, se houver uma sequência aleatória binária objetiva com

limiares de frequência $p(Fx) = 0,8$ e $p(\neg Fx) = 0,2$, então, a porcentagem de previsões verdadeiras pode ser, na melhor das hipóteses, 80% e é alcançada se o evento Fx for sempre previsto de acordo com a regra máxima.

De acordo com a visão filosófica convencional, consequências aleatórias objetivas só são possíveis se a suposição do *determinismo* inerente à física *clássica* estiver errada, ou seja, se as próprias leis fundamentais da natureza fizerem apenas afirmações probabilísticas, como na física quântica. Argumentaremos nesta seção que a visão convencional está incorreta: mesmo na física clássica, há aleatoriedade objetiva. Começemos pelos termos básicos.

O determinismo é comumente associado a Laplace (1814), mas é a visão de mundo dominante em toda a física pré-clássica e clássica até o advento da física quântica. De acordo com a visão determinista do mundo, não existe coincidência objetiva: todo o curso do mundo é completamente determinado e previsto pelas suas condições iniciais e pelas leis da natureza. Se um processo aleatório for considerado, a exemplo do resultado de um lançamento de moeda ou do decaimento de um átomo radioativo, então, de acordo com essa visão de mundo, a natureza estatística desses eventos é meramente uma consequência da nossa ignorância e não da indeterminação objetiva do curso futuro dos eventos. O indeterminismo, nessa visão, é meramente *epistêmico*, não ontológico.

A física quântica moderna, no entanto, ensina *um indeterminismo objetivo* – segundo ela, existem processos aleatórios objetivos, cujo resultado, mesmo com *o conhecimento completo* das condições iniciais, é objetivamente indeterminado e determinado apenas na forma de probabilidades estatísticas. O decaimento radioativo, por exemplo, é um processo aleatório objetivo: se e quando

um átomo radioativo de césio-137 decairá não pode ser previsto por nenhum conhecimento físico, não importa o quão completo seja, mas a probabilidade de um átomo de césio-137 decair em 30 anos é exatamente $1/2$; e, portanto, após 30 anos, 50%, com certeza prática, de uma amostra de césio-137, terá decaído. A indeterminação objetiva está presente em todos os processos *microfísicos*, para partículas da ordem de nanômetros, enquanto os *macroprocessos* físicos são descritos aproximada e corretamente pela física clássica e, em altas velocidades, pela física relativística, ambas baseadas em leis *determinísticas* da natureza.

Se essa visão estivesse correta, então, o decaimento radioativo teria de ser considerado um processo aleatório objetivo, enquanto o experimento (clássico) de lançamento de moedas teria de ser considerado apenas um processo aleatório epistêmico. Essa também é a visão de Coffa (1974) e de Salmon (1989, p. 75). Por que, então, processos aleatórios aparentemente objetivos desempenham um papel tão importante em nosso mundo cotidiano, ou seja, nas áreas da física clássica? Por que ninguém jamais foi capaz de prever os resultados de um lançamento regular de dados ou de um jogo de roleta com sucesso significativamente maior do que o acaso? Em outras palavras, como explicar os processos aleatórios macrofísicos?

A primeira parte da explicação é fornecida pelo fato de a física clássica também permitir estados de sistema indeterminados: sistemas fisicamente fechados cujo desenvolvimento no futuro não é determinado pelo seu estado inicial completo. Isso é impressionantemente demonstrado em Earman (1986, Cap. III); apresentam-se aqui somente os fatos básicos. Na verdade, apenas as leis naturais fundamentais da física clássica são determinísticas, mas isso nem sempre é válido para as suas soluções, ou seja, para as leis que

as seguem (estrita ou aproximadamente) e que descrevem os possíveis cursos de desenvolvimento temporal ou ‘trajetórias’ de um sistema. Uma equação diferencial expressa uma dependência funcional entre o estado $s(x)$ de um sistema x e as suas mudanças no tempo, dependendo dos parâmetros α_i , ou seja, $s = f(ds/ dt, \alpha_i)$.³²

Essa, ou *mesmo qualquer* equação funcional $Y = f(X)$, é “determinística” no sentido de que o valor da variável independente X determina *inequivocamente* o da variável dependente Y . Equações estatísticas (ou estocásticas) têm uma distribuição de probabilidade como variável dependente: $p(Y) = f(X)$. Nesse sentido, *todas as* leis básicas da física clássica são deterministas, enquanto as leis básicas da mecânica quântica *também* incluem leis estatísticas, por exemplo, aquelas que atribuem “densidade de probabilidade de encontrar” aos quadrados de amplitude. É possível, no entanto, que as soluções das equações diferenciais determinísticas ordinárias tenham *singularidades* ou *pontos instáveis* (bifurcações), nos quais o curso futuro do desenvolvimento não é determinado pelo presente completo. Se, por exemplo, uma esfera perfeitamente elástica é colocada exatamente na ponta de um hemisfério igualmente perfeito, será indeterminada a direção para a qual essa esfera se moverá, não apenas prática, mas também teoricamente, porque as funções da solução têm um ponto instável ali.

Entre singularidades e pontos instáveis, há uma diferença significativa. No caso de uma singularidade verdadeira, mesmo assumindo que o estado inicial descrito pelos números reais é exato e, portanto, infinitamente determinado, o desenvolvimento futuro é indeterminado, porque diferentes trajetórias surgem do mesmo ponto

³² Equações diferenciais com derivadas mais altas podem ser reduzidas a sistemas de equações diferenciais que contêm apenas derivadas de primeira ordem.

singular. Assim, mesmo assumindo uma descrição infinitamente precisa do sistema (os números reais têm um número infinito de casas decimais), uma singularidade é um estado do sistema indeterminístico. No caso de um ponto instável, por outro lado, supondo uma descrição infinitamente precisa do sistema, não há um verdadeiro indeterminismo, porque, nessa situação, todo estado futuro também é completamente determinado pelas leis da natureza. O determinismo de Laplace aplica-se, assim, a descrições infinitamente precisas de estados. Isso também é chamado de “caos determinístico”, segundo o qual os estados do sistema caótico são caracterizados por um alto grau de acumulação de pontos instáveis (Schuster, 1994).

No entanto, a variante aproximada de Popper do determinismo de Laplace não se aplica, segundo a qual um conhecimento aproximadamente preciso do estado inicial também permite uma previsão aproximadamente precisa de estados futuros (Popper 1982, p. 34; cf. Earman 1986, p. 8). Em vez disso, sistemas em estados instáveis são altamente sensíveis a variações mínimas nas condições iniciais: para cada variação, não importa quão pequena seja, as trajetórias resultantes desviam-se umas das outras, tanto quanto possível, após um tempo suficientemente longo. O exemplo ilustrativo é o a esfera ideal, mencionada acima, colocada no topo de um hemisfério ideal: o lado para o qual a bola rolará é determinado por flutuações incomensuravelmente pequenas e, portanto, é impossível de ser previsto.³³

³³ Matematicamente, trajetórias divergentes satisfazem a seguinte condição: $\forall \Delta > 0 \exists \epsilon > 0 \exists s, s' \in S \exists t \geq 0: |s(0) - s'(0)| \leq \epsilon \wedge |s(t) - s'(t)| > \Delta$. Em palavras: para qualquer Δ arbitrariamente grande e qualquer ϵ pequeno, existem dois estados de sistema s, s' (no

Os filósofos favoráveis ao determinismo gostam de argumentar que isso é apenas uma impossibilidade prática, e não de princípios. No entanto, isso é um erro na medida em que as flutuações relevantes dos estados iniciais são tão pequenas a ponto de estarem nas dimensões da microfísica, nas quais as relações de incerteza da mecânica quântica tornam-se relevantes e estabelecem limites fundamentais para a mensurabilidade arbitrariamente precisa. Em outras palavras, pontos instáveis implicam: o princípio do indeterminismo microfísico penetra no nível da macrofísica. Portanto, há apenas diferenças matematicamente, mas não fisicamente, significativas entre singularidades e meras instabilidades. Ambas as situações implicam a existência de processos objetivamente indeterminados, mesmo na escala da macrofísica. Estados instáveis ocorrem frequentemente em sistemas complexos e estão aqui relacionados à emergência espontânea, isto é, indeterminada, de estados macrofísicos de ordem (cf. Haken, 1983) – o que, aliás, também permite explicar o fenômeno da liberdade mental em processos cerebrais complexos.

Além das singularidades e das instabilidades discutidas, há outras diferenciações de instabilidades e de indeterminismos na física clássica (para detalhes, ver Earman, 1986, Cap. III; Weingartner; Schurz, 1996). Independentemente dessas diferenças, a consequência relevante para nossa preocupação reside no fato de que a física clássica também permite estados de sistema indeterminados, cujo desenvolvimento posterior é objetivamente indeterminado.

Essa, no entanto, é apenas metade da explicação da aleatoriedade macrofísica. Até agora, apenas explicamos por que o lançamento de dados macrofísico é um processo aleatório realmente

espaço de estados S) e um tempo final t , tal que o estado inicial de ambos os sistemas não difere em mais do que ϵ e o estado final difere um do outro em mais de Δ .

imprevisível: porque – se o lançamento foi feito regularmente, e não “forjado” – flutuações mínimas nas condições iniciais precisas do lançamento (nos movimentos de ar, etc.) decidem qual lado do dado estará no topo. Em outras palavras, justamente a dependência sensível das condições iniciais também é a responsável pelo “caos determinista” que está presente. Contudo, ainda não podemos explicar por que os cubos construídos macroscópica e simetricamente resultam nessa notável *distribuição uniforme* estável de probabilidades contra as quais as frequências tendem a ser limítrofes. Mesmo que os resultados do lançamento sejam objetivamente indeterminados, por que não é o caso de pessoas diferentes terem “vieses de arremesso” diferentes e, portanto, rolaem probabilidades diferentes? Afinal, elas jogam de formas diferentes, umas mais violentamente, outras mais suavemente, etc.

Apesar dessas diferenças, o motivo de todas as diferentes pessoas e técnicas de arremesso ainda produzirem os mesmos limiares de frequência requer um passo explicativo adicional, a saber: pode-se mostrar que – não todas, mas *quase todas* (99,99...% todas) – as distribuições de frequência sobre as condições macrofísicas iniciais do lançamento de dados levam a uma distribuição igual das frequências de resultado do lançamento de dados. Uma abordagem explicativa nesse sentido é apresentada por Strevens (2008, p. 370 e seg.) e, a seguir, apresento um novo desenvolvimento de suas discussões.

Consideramos a dependência dos resultados do lançamento de dados com as condições macrofísicas iniciais, que resumimos de forma simplificada em uma variável X . Ao rolar os dados, essa variável compreende várias dimensões (velocidades de translação e de rotação, movimentos do ar, etc.). Um exemplo mais simples é o da roleta, onde a variável X inclui apenas a velocidade inicial da roleta. A

variável dependente Y é discreta e inclui os valores 1, ..., 6 ao rolar os dados; na roleta, os valores são 0, ..., 36. A observação crucial é o que Stevens chama de “microperiodicidade” dessa dependência: mudanças mínimas na variável X causam mudanças máximas e, portanto, um ciclo periódico da variável Y. O gráfico da função de Y em função de X é, assim, extremamente íngreme, quase vertical. Isso é mostrado na Fig. 5-1(a). Uma distribuição estatística da variável X macrofísica, que leva a uma distribuição desigual dos resultados, ou seja, prefere certos valores de Y, teria de ser ainda mais íngreme; teria de ser algo como a forma mostrada na Fig. 5-1(b), isto é, também teria de ser “microperiódica”, nas palavras de Stevens. As distribuições macrofísicas, no entanto, nunca são microperiódicas, como Stevens supõe sem maiores justificativas (*ibidem*, p. 372). Contudo, por quê? Proponho a seguinte resposta: *quase todas as* distribuições macrofísicas de probabilidade são contínuas e variam em dimensões macrofísicas. Têm, assim, inclinações muito *mais* planas do que a dependência microfísica na Fig. 5-1(a). Portanto, observam-se apenas

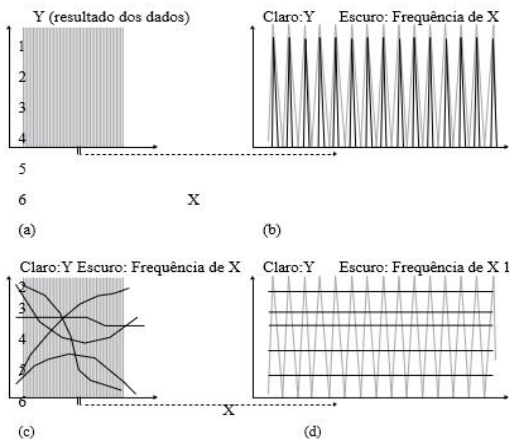


Fig. 5-1: Explicação da distribuição igual sobre os resultados do lançamento de dados, apesar das distribuições iniciais macro-físicas arbitrárias.

(a) Y microperiodicamente depende de X; distribuição extremamente íngreme

(b) X em representação extremamente esticada. Escuro: uma distribuição inicial que favoreceria um certo resultado Y – precisa ser ainda mais íngreme.

(c) O mesmo que (a). Escuro: distribuições macrofísicas sobre X. Todos eles levam a distribuições iguais sobre Y: ver (d).

(d) X em representação extremamente esticada. Escuro: A frequência dos valores X é distribuída uniformemente em intervalos X extremamente pequenos. Isso leva a uma distribuição igual sobre a frequência dos valores de Y.

frequências de X que são uniformemente distribuídas ao longo dos intervalos X microfisicamente pequenos, os quais compõem um período Y e, nesse sentido, levam a uma distribuição igual à dos valores de Y (ver Figs. 5-1c e 5-1d).

5.5 Propensões singulares

Na Seção 5.2, argumentamos que as probabilidades estatísticas são melhor entendidas como propensões genéricas, cuja expressão numérica é definida por limites de frequência que têm consequências indutivas para frequências finitas observáveis (ver também Rosenthal, 2004).

Alternativamente, alguns autores sugerem que probabilidades objetivas devem ser entendidas como propensões caso a caso, sem referência intrínseca a propensões genéricas e a experimentos aleatórios subjacentes. Popper (1990) faz tal sugestão em seus últimos anos, quando entende propensões singulares como as tendências causais graduais de eventos individuais para produzir certos efeitos

sem referência intrínseca à frequência (cf. a reconstrução em Miller, 1994, p. 182-6). Um exemplo é a tendência objetiva *dessa* aeronave de cair durante *esse* voo. De modo semelhante, as propensões objetivas caso a caso são caracterizadas por Lewis (1980). Para Lewis, a propensão caso a caso de um evento é condicionalmente relativizada ao estado geral do universo no momento em questão. Isso pode ser explicado da seguinte forma: na descrição de um experimento aleatório genérico, as condições físicas de contorno permitidas são especificadas com precisão suficiente. Em um lançamento de dados comum, por exemplo, é impossível que uma telha caindo enterre minha mão de arremesso no momento do meu lançamento.³⁴ Por outro lado, a propensão singular *desse* lançamento de dados depende muito do fato de essa telha estar a cair no momento em questão, tal como a propensão singular da queda *dessa* aeronave depende de um meteoro estar a entrar agora na região espaço-temporal em questão.

Além disso, as propensões caso a caso de Lewis são duplamente relativas ao tempo: a chance objetiva no tempo t^* de o evento ocorrer no tempo t depende do tempo t^* , no qual essa chance é determinada e que funciona como o “presente”. Enquanto a chance de eventos futuros ($t^* < t$) é identificada como a tendência causal do mundo no tempo t^* de produzir esse evento no tempo posterior t , de acordo com Lewis, a chance de eventos passados ($t^* \geq t$) é 1 ou 0, dependendo de o evento realmente ocorrer ou não.

³⁴ Em termos gerais, são permitidos apenas lançamentos com dados simétricos, sob influência exclusiva da gravidade, com local de aterrissagem nivelado e dentro de um intervalo permitido de velocidades e alturas iniciais, em temperatura normal e pressão atmosférica não muito alta. Esta é uma condição exclusiva e definitiva, *ceteris paribus*, no sentido de Schurz (2002).

Há uma variedade de objeções às propensões singulares. Eagle (2004) e Gillies (2000, p. 126-136) fornecem uma visão geral. Para ilustrar, um argumento de Humphreys (1985) afirma que as propensões singulares não devem ser identificadas como tendências causais, porque as relações de propensão entre eventos não são dirigidas, enquanto as relações causais são sempre direcionadas de causa para efeito. Por exemplo, hoje eu tenho uma propensão a estar vivo ontem, algo que não pode ser interpretada causalmente.

Uma objeção amplamente discutida por Eagle (*ibid.*) é o problema da *classe de referência*. De acordo com essa objeção, a propensão de um único evento só pode ser atribuída a limites de frequência se esse único evento for atribuído a uma *sequência aleatória* objetiva produzida por um experimento aleatório (repetível) que atua como uma classe de referência genérica. No entanto, há sempre *várias* dessas classes de referência. Por exemplo, a probabilidade anual de morte de Pedro pode ser determinada como a probabilidade de morte de pessoas com mais de 50 anos, a de homens não fumantes com mais de 50 anos, a de homens não fumantes com mais de 50 anos que trabalharam na mineração a vida toda, e assim por diante. A escolha de uma classe de referência apropriada é um problema importante para todas as aplicações da teoria estatística de probabilidade.

No entanto, como objeção a propensões singulares, essa objeção é descabida por duas razões. Primeiro, como é afirmado na Seção 5.2, só é possível atribuir *propensões genéricas* às reações individuais do experimento, como “tropos objetivos de propensão”, se for uma sequência aleatória objetiva, ou seja, se a afiliação ao experimento aleatório dado formar uma classe de referência *objetivamente mais estreita*. Nesse caso, no entanto, o problema de múltiplas classes de

referência não se coloca mais, porque há apenas uma classe de referência objetivamente mais estreita.

Em segundo lugar, a objeção não se aplica realmente às propensões *singulares* como as entendemos aqui. Porque, assim que a propensão de um evento singular Fa (“este é um F ”) é definida por um experimento aleatório Z , $p(Fa)$ é definido como $p(Fx|Zx)$, um conceito *genérico* de propensão já está disponível. E isso é verdade, mesmo que a propensão genérica do experimento aleatório seja atribuída aos eventos individuais como tropos de propensão. Isso ocorre porque os tropos de propensão ainda são os exemplos individuais de propensões genéricas. Essa é a única razão pela qual eles estão relacionados a sequências aleatórias e a suas tendências de frequência. Em outras palavras, os tropos de propensão não são propriedades intrínsecas dos sequenciadores individuais, mas sim disposições do experimento aleatório que se manifestam nas propriedades dos sequenciadores individuais.

Isso é bem diferente no caso das propensões singulares genuínas, que são ontologicamente entendidas como propriedades intrínsecas dos eventos individuais ou como membros individuais da sequência. É apenas nesse momento que a objeção às propensões singulares vem à tona: elas não têm relação com frequências ou limiares de frequência. Isso é mostrado pelo seguinte argumento: se as propensões singulares fossem sobrepostas sobre os eventos individuais ou sobre os membros da sequência individual e se o experimento fosse objetivamente aleatório, a propensão singular de um possível evento E em AL teria de ser a mesma para todas as execuções a_i do experimento aleatório. Por exemplo, no experimento de lançamento de moedas, a propensão de caras (a_i) deve ser de 0,5 para todos os a_i ($i \in \mathbb{N}$).

Nesse caso, entretanto, é impossível que as propensões singulares sejam sistematicamente relacionadas ao limiar de frequência $p(\text{Ex}) = 0,5$ da sequência aleatória $(a_i; i \in \mathbb{N})$. Isso ocorre porque é possível reorganizar a sequência fazendo uma seleção de posição inadmissível, de tal forma que o limiar de frequência se desvie arbitrariamente de 0,5. Contudo, as propensões singulares – ou seja, as propriedades intrínsecas do a_i – permanecem exatamente as mesmas. Isso mostra que não pode haver uma correlação entre propensões singulares e pontos de corte de frequência, nem mesmo uma relação indutiva. Portanto, não parece haver nada na abordagem da propensão singular capaz de impedir reivindicações de propensão singular totalmente irracionais, como a afirmação:

(5-7) Nesse sorteio, o mentalista Uri Geller conseguiu usar seu poder mental para fazer a moeda (com propensão 1) cair em cara; no entanto, ele só consegue isso em 50% de todos os casos.

Afirmações de propensão singular desse tipo são pura especulação, empiricamente não verificáveis e, portanto, sem valor cognitivo (cf. Gillies, 2000, p. 127).

6 Problemas do conceito subjetivo-epistêmico de probabilidade

6.1 Problemas de definição

No caso do conceito subjetivo de probabilidade, não é o problema da definição que causa grandes dificuldades: probabilidades subjetivas são simplesmente definidas como os graus epistêmicos de crença de pessoas, os quais satisfazem os axiomas básicos de probabilidade de Kolmogorov, de acordo com o Teorema 1 (ou a condição de coerência equivalente para quocientes de apostas justas, ver Seção 6.2). No entanto, é um achado persistente da psicologia cognitiva (ver Kahneman *et al.*, 1982) o fato de as crenças reais dos sujeitos de teste geralmente não atenderem aos axiomas da probabilidade. Os defensores do bayesianismo geralmente interpretam os axiomas básicos da probabilidade como condições de racionalidade. Entretanto, é controverso se essas condições de racionalidade são realmente adequadas para pessoas reais com suas limitações de complexidade cognitiva, ou seja, se são pelo menos aproximadamente satisfazíveis de acordo com o princípio “deveria-ser capaz” ou se se aplicam apenas a sujeitos abstratos ideais.

Se esse último fosse o caso, colocaria fortemente em questão a relevância prática e científica da teoria subjetiva da probabilidade.

Para a teoria subjetiva da probabilidade, portanto, surge um problema de justificação: por quais razões os graus racionais de crença devem satisfazer os axiomas básicos A_{1-3} , e qual deve ser o significado de tais graus de crença em relação ao objetivo prático e científico de encontrar verdades reais substantivas? É a esse problema da justificação que nos voltamos agora.

6.2 Problemas de justificação: quocientes de apostas justas coerentes

A justificativa subjetiva mais conhecida dos axiomas de probabilidade A_{1-3} é o método de explicar graus subjetivos de crença sobre o comportamento de apostas de pessoas racionais como quocientes de apostas justos, remontando a Frank Ramsey e Bruno De Finetti. A seguir, esse método é brevemente descrito.³⁵ Uma aposta W , em uma proposição A (de uma dada álgebra ou linguagem), é definida abstratamente como um triplo $W = (A, g, v)$. Onde g é o valor monetário dos ganhos que o apostador ganha e que o apostador contrário perde se A for verdadeiro, e onde v é o valor da perda que o apostador perde e que o apostador contrário ganha se A for falso. Tanto g quanto v são representados como números reais positivos. Chamam-se aposta $e =_{\text{def}} g+v$ e quociente de apostas $q =_{\text{def}} v/(g+v)$ da aposta (A, g, v) . O lucro e a perda podem ser representados como $g = (1-q) \cdot e$, $v = q \cdot e$, e a aposta como $(A, (1-q) \cdot e, q \cdot e)$. Para cada aposta

³⁵ Para detalhes, cf. Carnap (1971, Cap. 8); Skyrms (1999, cap. 6); Howson/Urbach (1996, Cap. 5); Gillies (2000, cap. 4).

$W = (A, g, v)$, $Wc = (-A, v, g)$ é a contra aposta correspondente, quer dizer, o oponente da aposta W aposta na contra aposta Wc (ou é o apostador de Wc).

Quando é racional para o apostador aceitar uma aposta $W = (A, g, v)$? De acordo com o princípio bayesiano, isso depende do valor esperado subjetivo $E(W)$ do lucro da aposta, que é determinado pela fórmula

$$(6-1) E((A,g,v)) = g \cdot P(A) - v \cdot P(\neg A)$$

Se a probabilidade estatística “ p ” é tomada em vez do grau subjetivo de crença “ P ”, então, $E(W)$ expressa o lucro médio objetivo de apostas do tipo W no longo prazo. É importante notar: ainda não inserimos a lei da probabilidade $P(\neg A) = 1 - P(A)$ na formulação de (6-1).

Do ponto de vista do apostador, a aposta W é chamada de *favorável* se $E(W) > 0$, de *desvantajosa* se $E(W) < 0$ e de *justa* se $E(W) = 0$.³⁶ Aposta e contra aposta, portanto, têm chances iguais de ganhar se os valores esperados de ambas as apostas forem zero. Para apostas justas, no entanto, verifica-se que o grau subjetivo de crença sob a suposição $P(\neg A) = 1 - P(A)$ corresponde ao quociente de apostas $q = v/g + v$ definido acima, porque

$$(6-2) E(W) = E((A, (1-q) \cdot e, q \cdot e)) = P(A) \cdot (1-q) \cdot e - (1-P(A)) \cdot q \cdot e = 0 \text{ sse } P(A) \cdot (1-q) = (1-P(A)) \cdot q \text{ sse } P(A) - q \cdot P(A) = q - q \cdot P(A) \text{ sse } P(A) = q.$$

³⁶ *Provas:* $E((A,g,v)) = g \cdot P(A) - v \cdot P(\neg A) = g \cdot P(\neg \neg A) - v \cdot P(\neg A) = (-v) \cdot P(\neg A) - (-g) \cdot P(\neg \neg A) = E((\neg A, v, g))$. Este resultado aplica-se independentemente de $P(\neg A) = 1 - P(A)$.

Isso dá origem à ideia de que se pode identificar os graus subjetivos de crença de uma pessoa com seus quocientes de apostas justas e, assim, operacionalizá-los empiricamente. No entanto, o que garante que os quocientes de apostas justos de uma pessoa atendam aos axiomas da teoria da probabilidade? Essa questão também surge para a lei de probabilidade $P(\neg A) = 1 - P(A)$ inserida no cálculo de (6-2). A normalização das probabilidades para os valores entre 0 e 1 e, portanto, para o axioma (A1) de Def. 3-1, já está garantida pela definição de quocientes de apostas. Contudo, o que garante os axiomas 2 e 3? Na verdade, os quocientes de apostas justas das pessoas não precisam atender a esses axiomas. Entretanto, como Ramsey e De Finetti mostraram independentemente, o cumprimento desses axiomas termina sendo equivalente a uma condição fundamental para o comportamento racional de apostas – a condição de coerência. Por um sistema de apostas WS, queremos dizer um conjunto finito de apostas $WS = \{W_i =_{\text{def}} (A_i, g_i, v_i) : 1 \leq i \leq n, n \in \mathbb{N}\}$, com A_i como proposições ou sentenças de uma dada álgebra AL ou linguagem \mathcal{L} . Para cada estado possível do mundo w , o *lucro total* (ou perda) do sistema de apostas WS é definido como: $g(WS, w) = \sum_{1 \leq i \leq n} g(W_i, w)$, onde $g(W_i, w) = g_i$ se A_i é verdadeiro em w , e, caso contrário, se $g_i(W_i, w) = -v_i$. Seja $q_X: AL \rightarrow [0, 1]$ a função de crença de uma pessoa X identificada com quocientes de apostas justas, que atribui às proposições P de uma álgebra AL os quocientes de apostas $q_X(P)$ entre 0 e 1, onde P estaria igualmente disposto a aceitar a aposta e a contra aposta em P. Assim, a coerência é definida da seguinte forma:

(Def. 6-1) Coerência: A função de crença $q_X: AL \rightarrow [0, 1]$ de um apostador X, explicada por quocientes de apostas justas, é chamada de coerente sse não há um sistema de apostas finitas que consiste

em (no que diz respeito a qX) apostas simples justas $W_s = \{W_i: 1 \leq i \leq n\}$ que leva a uma perda total $g(W_s, w) < 0$ em cada estado de mundo possível w para X .

Uma pessoa que faz apostas com uma função de crença inconsistente aceitaria um sistema de apostas justo que lhe garantiria uma perda certa – um sistema desses é chamado de “*Dutch book*”. Por exemplo, você aceitaria um *Dutch book* se aceitasse duas apostas comigo:

- . Apostar com cotas de $\frac{1}{2}$ que vai chover amanhã;
- . E, ao mesmo tempo, apostar com cotas de $\frac{3}{4}$ que não vai chover amanhã.

Com 'e' sendo o valor da aposta, seu retorno total seria $0,5e - 0,75e = -0,25e$ se chover amanhã e $-0,5e + 0,25e = -0,25e$ se não chover; portanto, você perderia um quarto do valor da aposta em qualquer caso."

Ramsey (1926) e De Finetti (1937) provam o seguinte Teorema:

(Teorema 6-1) Coerência: Uma função de crença subjetiva $q: A \rightarrow R$, explicada por quocientes de apostas justas, satisfaz os três axiomas de probabilidade (A1)-(A3) (ver Def. 3-1) se for coerente.

A definição de probabilidade da abordagem subjetiva é, portanto, a seguinte: probabilidades subjetivas são os quocientes de apostas justas de apostadores coerentes. Uma prova do Teorema 6-1 pode ser encontrada no Apêndice 10.3.7 (ver também a literatura mencionada pela nota 34). Carnap (1971) prova o Teorema 6-1 para probabilidades condicionais axiomatizadas de forma independente. Para justificar a σ -aditividade por meio da coerência, é necessário um

sistema fictício de apostas infinitas (ver Earman, 1992, p. 40 e Howson; Urbach, 1996, p. 81-84). Outra extensão do teorema 6-1 é a seguinte:

(Def. 6-2) (a) Uma função de crença $q: AL \rightarrow [0,1]$ é chamada de estritamente coerente sse não houver um sistema de apostas que consista em apostas simples justas que não produzam um lucro em nenhum mundo possível e uma perda em pelo menos um mundo possível.

(b) Uma função de probabilidade $P: AL \rightarrow [0,1]$ é chamada de regular sse AL for uma álgebra sobre um espaço de possibilidade contável e que atribui um grau de crença maior que 0 a todas as proposições possíveis (elementos não vazios de AL).

Teorema 6-2: Uma função de crença explicada por quocientes de apostas justas é estritamente coerente sse satisfizer os axiomas ($A1-3$) e for regular.

Para a prova do Teorema 6-2, conferir o Apêndice 10.3.8. Para espaços contínuos de possibilidade com um número incontável de elementos, há necessariamente sempre (incontáveis muitos) elementos com probabilidade zero, porque, caso contrário, a probabilidade de Ω excederia o valor de um infinitamente (ver Seção 7.5 e Seção 9.3); portanto, o conceito de regularidade deve ser limitado a espaços contáveis de possibilidade.

A justificação da teoria subjetiva da probabilidade dos quocientes de apostas justas está sujeita a muitas críticas (ver Earman, 1992, Seção 2.4; Howson, 1995, Cap. 5; Gillies, 2000, Cap. 4). Primeiro apresentamos três objeções, para as quais há respostas aceitáveis,

mesmo que elas façam uso de fortes idealizações, e, então, chegamos a três objeções básicas.

(1.) O valor esperado de uma aposta justa é zero. Por que, então, os maximizadores racionais de utilidade aceitariam uma aposta justa? O ponto é válido, mas não constitui uma objeção. Com certeza, pessoas racionais apenas aceitam apostas com valor esperado positivo, ou seja, cuja odds é menor que o grau de crença. Em vez disso, a odds justa de uma pessoa é determinada por meio de perguntas sobre uma situação hipotética, na qual a pessoa é forçada a aceitar uma aposta ou sua aposta contrária. Pergunta-se qual das duas apostas é preferida, aumentando sistemática e gradualmente a odds até que a preferência da pessoa mude da aposta para a aposta contrária. A odds na qual a pessoa está exatamente “indecisa” entre a aposta e a aposta contrária representa seu grau de crença racional.

(2.) A utilidade de uma aposta para o apostador, normalmente, não depende linearmente do lucro da aposta, mas aumenta menos do que linearmente para ganhos mais altos (em unidades monetárias). Por outro lado, a nocividade das altas perdas de apostas aumenta mais do que linearmente devido à natureza finita dos recursos do apostador. Existem diferentes métodos para considerar essa não-linearidade, os quais requerem novas idealizações. Por exemplo, propõe-se que as situações hipotéticas de apostas sobre as quais a pessoa é questionada sejam realizadas com apostas tão pequenas, em comparação com a riqueza da pessoa, que se possa supor uma linearidade aproximada (cf. Gillies, 2000, p. 56). No entanto, coloca-se a questão de saber se as preferências expressas por indivíduos em apostas muito baixas são subjetivamente confiáveis.

(3.) Apostas reais só podem ser feitas em proposições empiricamente verificáveis. Portanto, não é possível fazer apostas

reais nas próprias proposições que são mais importantes para a aplicação do bayesianismo – ou seja, as hipóteses científicas não verificáveis. Aqui, também, as idealizações devem ser usadas, por exemplo, com a ajuda de perguntas da seguinte forma: quantos euros você apostaria que a teoria da relatividade é verdadeira se houvesse um especialista perfeito e capaz de dar a você e ao oponente da aposta, depois de a aposta ser feita, uma resposta certamente verdadeira para essa pergunta? Obviamente, essas já são idealizações bastante estranhas que podem rapidamente sobrecarregar pessoas reais ou levar a inconsistências.

Passamos agora às objeções mais difíceis. A primeira delas já foi apresentada na Seção 6.1.

(4.) Surge a questão de saber se as pessoas razoáveis possuem graus quantitativos de crença sobre toda e qualquer proposição, e se é racional possuir tais graus de crença. Não conheço nenhuma pessoa que tenha graus numéricos de crença em questões como se a teoria da relatividade de Einstein seria verdadeira, se houve um ou mais Big Bangs ou se Deus existiria. Em vez de graus quantitativos, a maioria das pessoas criaria julgamentos qualitativos como “acho que isso está suficientemente comprovado”, “acho que isso é mais provável do que o seu oposto” ou “isso é muito improvável para eu considerar a possibilidade na prática”. As crenças ou os quocientes de apostas das pessoas parecem ser razoavelmente estáveis apenas quando são baseadas em experiências de frequências. Caso contrário, pode-se supor que os quocientes hipotéticos de apostas expressos por pessoas são declarados comparativa e arbitrariamente, bem como flutuam relativa e fortemente em situações de teste repetidas, mesmo para uma mesma pessoa. Seria uma tarefa importante investigar essa questão em experimentos psicológicos. (5.) A última objeção guia-me

diretamente à primeira objeção principal, isto é, à teoria subjetiva da probabilidade baseada em quocientes de apostas justas: quocientes de apostas justas coerentes estão longe de ser racionais no sentido de objetivamente orientados para a verdade. A taxa de sucesso real não é afetada pela justificativa de apostas justas. Deve-se recapitular que a definição de quociente de apostas justo refere-se apenas às preferências subjetivas de apostas dos indivíduos e é independente das frequências objetivas de verdade. Suponhamos, por exemplo, um subjetivista apostando entusiasticamente 1:1 que ele vai rolar um seis com um dado regular, em uma situação na qual o quociente de apostas seja justo. Assim, ele também estaria disposto a tomar a contra aposta 1:1 de que ele não vai rolar um seis, porque a probabilidade subjetiva de rolar um seis é $1/2$. Nosso subjetivista permanecerá coerente, mesmo que tenha perdido toda a sua fortuna, e não será capaz de ver qualquer falha em seu comportamento de apostas. É claro que ele ficará surpreso em relação ao fato de que as contra apostas, consideradas por ele tão justas, nunca lhe tenham sido tiradas. No entanto, ele não conseguirá explicar por que ele, de todas as pessoas, perdeu toda a fortuna enquanto outros a colheram, desde que ele não leve em conta as probabilidades objetivas de frequência do tipo de evento no qual apostou. Isso mostra que os axiomas A1-3 fornecem, na melhor das hipóteses, condições mínima para graus racionais de crença, que são muito fracas para excluir o comportamento irracional de apostas de um ponto de vista objetivo. Seguindo Howson (2000, p. 133), podemos comparar a condição de coerência para graus de crença com o princípio lógico de não contradição para afirmações: assim como a observância deste último não garante a racionalidade das afirmações, a coerência não assegura a racionalidade dos graus de crença.

(6.) Como a objeção anterior aponta, as crenças subjetivas podem ser de indivíduos diferentes, mesmo que todas sejam intersubjetivamente coerentes, e podem diferir umas das outras à vontade. Nesse contexto, surge uma segunda objeção principal, formulada por Ryder (1981). Afinal, a partir do momento em que várias pessoas possuem diferentes graus de crença na mesma proposição, um *Dutch book* pode ser construído contra o grupo de pessoas.

Existe, então, um sistema de apostas justas que, para todas as condições de mundos possíveis, conduz a uma perda total para o grupo e a um lucro total para quem apostar contra o grupo. Por exemplo, se a pessoa X apostar, com um quociente de apostas justo de $\frac{1}{2}$, que vai chover amanhã e se a pessoa Y apostar contra ela, com um quociente de apostas justo de $\frac{3}{4}$, e ainda se eu aceitar ambas as apostas como oponente, então, eu ganho um quarto da aposta em todos os mundos possíveis, porque:

(i) se chover amanhã, eu vou receber metade de e de X, bem como vou ter de pagar a Y um quarto de e ,

(ii) se não chover amanhã, eu vou receber três quartos de e de Y, bem como vou ter de pagar a X metade de e .

Assim, X e Y juntos perdem um quarto de e em qualquer caso, mesmo que ambos os quocientes de apostas sejam coerentes. Como argumenta Ryder (1981), uma regra de comportamento de aposta, se seguida por várias pessoas, podendo levar a uma perda necessária dessas pessoas, dificilmente pode ser chamada de “racional”. Gillies (2000, p. 170 e seg.) conclui, a partir do problema do *Dutch book* coletivo, que as pessoas orientadas para a cooperação devem ter interesse em estabelecer uma convergência de seus graus de crença. Entretanto, é claro que surge a questão de *como* tal acordo intersubjetivo pode ser estabelecido de forma *não arbitrária*. Afinal,

mesmo que um acordo artificial dos graus de crença seja alcançado por um ditador, isso não resolve o problema principal (problema (5.)): os graus intersubjetivamente congruentes de crença não seriam mais suscetíveis a um *Dutch book* coletivo e não teriam nenhuma conexão com frequências objetivas, de modo que todo o coletivo, apesar da coerência coletiva, transformar-se-ia em uma série de erros e de previsões coletivas, pois os graus intersubjetivos de crença não são combinados ou “calibrados” com as frequências objetivas.

A maior fraqueza da abordagem da coerência subjetiva é: ela não tem conexão intrínseca com probabilidades estatísticas. A partir do momento em que se tenta fazer tal referência, há, a meu ver, uma maneira mais ponderada de justificar os axiomas básicos (A_{1-3}) da teoria subjetiva da probabilidade: afirmando que os graus subjetivos racionais de crença pretendem refletir as probabilidades estatísticas objetivas. Eles só podem fazer isso se cumprirem os axiomas básicos (A_{1-3}). Chamo isso de justificativa de frequência intencional dos axiomas básicos para probabilidades subjetivas (Carnap, 1950, p. 167 ou Earman, 1992, p. 46 argumentam similarmente). No entanto, graus subjetivos de crença só podem concordar com frequências objetivas se relações funcionais de ponte entre probabilidades subjetivas e probabilidades estatísticas forem estabelecidas. Já vimos essa relação na Seção 2: o princípio da classe de referência mais estreita (Def. 2-3). O Capítulo seguinte apresenta um exame aprofundado das relações entre probabilidades objetivas e probabilidades epistêmicas.

7 Relações entre probabilidades objetivas e epistêmicas: uma abordagem dualista

7.1 Princípio de coordenação (ou “princípio principal”)

Alguns defensores da abordagem da probabilidade subjetiva têm procurado estabelecer ligações entre probabilidades subjetivas e probabilidades objetivas. Essas conexões são geralmente explicadas na forma de princípios *adicionais* ou “axiomas” que vão além dos axiomas básicos (A1-3) e, sucessivamente, que transformam abordagens subjetivas de probabilidade em abordagens “intersubjetivas”, “objetivas” ou “lógicas”. O princípio mais elementar desse tipo é o chamado de “princípio principal”, segundo o qual as probabilidades subjetivas, se as probabilidades objetivas são conhecidas, devem concordar com elas. Esse princípio existe em duas versões, as quais são muito diferentes em suas naturezas.³⁷

³⁷ Strevens (2004) introduziu o conceito de “princípio de coordenação probabilística”. O princípio singular, conhecido como princípio principal, também foi chamado de “princípio de Miller”, pois Miller (1966) tentou demonstrar sua inconsistência. No entanto, a objeção de Miller foi refutada por Jeffrey (1970). Como veremos, tanto o

O *princípio singular da coordenação* remonta a Lewis (1980) e combina probabilidades subjetivas com propensões de casos individuais. Como explicado na Seção 5.5, Lewis chama as propensões singulares de *chances*. Que “ $ch_t(Fa_i)$ ” denote a propensão singular, no tempo t , de que o evento possível Fa_i ocorre em um tempo (diferente) t_i ; onde “ a_i ” é uma sequência de constantes individuais contendo a constante de tempo t_i .

Se $t_i > t$, então, é a chance de um evento futuro e , se $t \geq t_i$, é a chance de um evento já passado ou presente, que, de acordo com Lewis, é 1 se o evento ocorreu e 0 se não. Se $E_{<t}$ é chamada de *proposição admissível*: essa é uma proposição que fala apenas sobre eventos *antes* de t , então, o “princípio principal” de Lewis afirma o seguinte (ver Lewis, 1980, p. 87 e seg.; Earman, 1992, p. 52):

(Def. 7-1) *Princípio singular da coordenação* (ou “princípio principal”):

$$(a) P(Fa_i | ch_t(Fx)=r \wedge E_{<t}) = r.$$

Em palavras: O grau racional de crença de que um determinado evento ocorre no tempo t_i , dada a chance objetiva desse evento no tempo t e dadas as proposições arbitrárias sobre eventos antes de t , é idêntico ao valor dessa chance objetiva.

(b) *Reforço*: Em vez de “ $ch_t(Fx)=r$ ”, pode-se estabelecer uma hipótese arbitrária sobre chances objetivas de eventos no tempo t , que “ $ch_t(Fx)=r$ ” implica probalisticamente (cf. Lewis, 1980, p. 97).

princípio de coordenação singular quanto o estatístico precisam ser restringidos a “evidências admissíveis” para evitar inconsistências. Williamson (2013, p. 299) propõe um princípio de “calibração” como uma generalização do princípio principal, mas que deixa em aberto como evidências frequenciais devem restringir as possíveis funções de probabilidade subjetiva, sendo, por isso, de utilidade limitada.

É permitido condicionar o grau de crença obtido por meio da Def. 7-1 à evidência *anterior* ao período de avaliação da oportunidade t , mas *não* à evidência sobre eventos após t , porque isso poderia resultar em avaliações de probabilidade incoerentes. Por exemplo, para $t' > t$, a proposição $P(Ft' \mid ch_t(Ft')=0,5 \wedge Ft') = 0,5$ é incoerente, pois, de acordo com os axiomas de Komolgorov, essa probabilidade deve ser 1.

No entanto, como explicado no Cap. 5.5, as propensões singulares são empiricamente desprovidas de conteúdo, razão pela qual não se busca mais esse conceito. Fundamental para todas as aplicações empíricas de probabilidades epistêmicas, por outro lado, é o *princípio de coordenação estatístico*, ou o *princípio principal* estatístico, abreviado StK (*statistische Koordinationsprinzip*), que estabelece uma relação análoga entre probabilidades subjetivas e probabilidades estatísticas.

Esse princípio remonta (indiretamente) a De Finetti e é defendido por bayesianos dualistas que reconhecem probabilidades estatísticas, por exemplo, von Kutschera (1972, p. 82), Howson e Urbach (1996, p. 345), Strevens (2004) ou Williamson (2010, p. 40; 2013, p. 299). A Def. 7-2 resume o StK em suas três formas diferentes para predicados de um dígito. Nesse caso, Fx (ou Fa_i) significa uma fórmula potencialmente complexa em exatamente uma variável individual x (ou constante individual a_i). “ Fa_i ” também pode ter a forma “ $Ff(a_i)$ ”, onde $f(a_i)$ denota um indivíduo diferente de a_i (por exemplo, um em um momento diferente). A Def. 7-2(c) estabelece $h(F|\{a_1, \dots, a_n\})$ para a frequência de Fx em uma dada amostra de n -elementos composta por indivíduos a_1, \dots, a_n . Cada constante individual a_i refere-se ao resultado

do sorteio i° (i-ésimo) aleatório de um indivíduo de D (com substituição) em uma determinada amostragem.³⁸

(Def. 7-2) *Princípio de coordenação estatística* (“princípio principal”)

(StK):

(a) Seja H uma hipótese estatística que implica probabilisticamente $p(Gx)=r$. Então: $P(Ga_i | H \wedge E(b_1, \dots, b_n)) = r$, desde que a condição de admissibilidade “ $a_i \neq b_j$ para todo $j \in \{1, \dots, n\}$ ” seja satisfeita.

Caso especial: $P(Ga_i | p(Gx)=r \wedge E(b_1, \dots, b_n)) = r$.

Em palavras: O grau racional de crença de que um determinado indivíduo a_i possui a propriedade G , sob a suposição de que a probabilidade estatística de Gs , em um determinado intervalo de indivíduos, tem o valor r (onde, no antecedente, não há mais nada sobre a_i , mas, no máximo, sobre coisas diferentes de a_i , indivíduos b_j , ou onde outros fatos estatísticos são assumidos), é idêntico ao valor r .

(b) Seja H uma hipótese estatística que implica probabilisticamente $p(Gx|Fx)=r$. Então: $P(Ga_i | H \wedge Fa_i \wedge E(b_1, \dots, b_n)) = r$, onde a condição de admissibilidade é satisfeita tal como em (a).

Caso especial: $P(Ga_i | p(Gx|Fx)=r \wedge Fa_i \wedge E(b_1, \dots, b_n)) = r$.

Em palavras: O grau racional de crença de que um determinado indivíduo a_i possui a propriedade G , assumindo que a probabilidade estatística de Gs , na classe de Fs , tem o valor r e que a_i possui a

³⁸ As constantes a_i são, portanto, identificadas com as descrições definidas “o indivíduo sorteado no i -ésimo sorteio” para excluir o conhecimento prévio sobre esses indivíduos. As sentenças de identidade recebem, portanto, valores de verdade e probabilidade não triviais, por exemplo, $a_4 = a_{16}$ pode ser verdadeiro e $p(x_4 = x_{16})$ pode ser maior que zero.

propriedade F (onde ... nota entre parênteses como em (a)), é idêntico ao valor r.

(c) *StK para amostras aleatórias:*

$$P(h(Gx|\{a_1, \dots, a_n\}) = k/n \mid p(Gx|Fx)=r \wedge Fa_1 \wedge \dots \wedge Fa_n) = \binom{n}{k} \cdot r^k \cdot (1-r)^{n-k}$$

Em palavras: O grau racional de crença de que a frequência de Gs, nessa amostra aleatória de n Fs é k/n, assumindo uma probabilidade estatística de G dado F de valor r, é consistente com a frequência de k resultados prováveis em n repetições de um experimento aleatório binário calculado pela fórmula binomial.

O (StK) da Def. 7-2 é uma base importante da estatística bayesiana. Os teóricos da probabilidade subjetiva são chamados de “bayesianos”, porque usam o Teorema bayesiano (TB5, Teorema 3-3) para calcular a probabilidade subjetiva $P(H|E)$ de uma hipótese H, dada a evidência ou um resultado amostral E. Ao fazer isso, eles usam a probabilidade inversa $P(E|H)$ da evidência E dada a hipótese H, que também é chamada de verossimilhança. Para os bayesianos, é essencial que as probabilidades de hipóteses subjetivas convirjam contra as probabilidades intersubjetivas (ou objetivas) à medida que as evidências aumentam (ver Cap. 9.5). Isso só é possível se as semelhanças epistêmicas $P(E|H)$ forem identificadas com as probabilidades estatísticas correspondentes $p_H(E(x))$, de acordo com o StK – fato que é elaborado por Hawthorne (2005, p. 286) e por Strevens (2004), entre outros, e também é aceito pelos bayesianos personalistas (cf. Edwards *et al.*, 1963). Na Def. 7-2(a), por exemplo, H é a hipótese “ $p(Fx)=r$ ”, E é a evidência “Fa”, $p_H(E(x))$ é o valor r e $P(E|H)$ é a probabilidade $P(Fa|p(Fx)=r)$ identificada por r. Em 7-2(c), E é a evidência de amostra $h(Fx|\{a_1, \dots, a_n\}) = k/n$.

O StK para probabilidades condicionais 7-2(b) *teoricamente segue* o StK para probabilidades incondicionais (7-2)(a), conforme mencionado na nota 38.³⁹ O eventual reforço da hipótese estatística nas alíneas (a) e (b) é necessário, nomeadamente, para a derivação do StK para combinações independentes de experiências aleatórias. Por exemplo, $P(Fa \wedge Gb \mid p(Fx)=r \wedge p(Gx)=q \wedge Ec) =$ (puramente probabilístico) $P(Fa \mid p(Fx)=r \wedge p(Gx)=q \wedge Ec) \cdot P(Ga \mid p(Fx)=r \wedge p(Gx)=q \wedge Fb \wedge Ec) =$ (de acordo com 7-2) $r \cdot q = p(Fx \wedge Gy)$, onde “ $P(Ga \mid p(Fx)=r \wedge p(Gx)=q \wedge Fb \wedge Ec)$ ” “ $Fb \wedge Ec$ ” agora atua como evidência admissível. Da mesma forma, obtemos da Def. 7-2(a+b) o StK para todas as fórmulas formadas a partir de predicados de um dígito em várias constantes individuais e, em particular, o StK para amostras aleatórias em (c).⁴⁰

Em contraste com o princípio de coordenação singular, a probabilidade estatística p de um tipo de evento Gx , se for transferida como um grau de crença P para um único caso Ga_i , deve ser explicitamente relacionada a uma classe de referência específica, possivelmente vazia, Fx , cuja instanciação Fa_i expressa uma evidência *hipotética* sobre a_i , à qual P é condicionalizado. (No princípio singular de coordenação, por outro lado, “ ch_i ” está implicitamente

39 Prova: Sejam H e H_q as hipóteses estatísticas $p(Gx|Fx) = r$ e $p(Fx) = q$. Aplica-se o seguinte (*): $H_q \wedge H$ implica probabilisticamente a declaração de probabilidade estatística $p(Gx \wedge Fx) = r \cdot q$. Então $P(Ga|Fa \wedge H \wedge Eb) = P(Ga \wedge Fa | H \wedge Eb) / P(Fa | H \wedge Eb) = \sum_{q \in [0,1]} P(Ga \wedge Fa | H_q \wedge H \wedge Eb) \cdot P(H_q | H \wedge Eb) / \sum_{q \in [0,1]} P(Fa | H_q \wedge H \wedge Eb) \cdot P(H_q | H \wedge Eb) =$ (de acordo com 7-2(a) e (*)) $\sum_{q \in [0,1]} r \cdot q \cdot P(H_q | H \wedge Eb) / \sum_{q \in [0,1]} q \cdot P(H_q | H \wedge Eb) = r \cdot \sum_{q \in [0,1]} q \cdot P(H_q | H \wedge Eb) / \sum_{q \in [0,1]} q \cdot P(H_q | H \wedge Eb) = r$.

40 Prova: Porque $h(Fx|\{a_1, \dots, a_n\})$ é equivalente à disjunção exclusiva $\vee\{Fa_{i_1} \wedge \dots \wedge Fa_{i_k} \wedge \neg Fa_{i_{k+1}} \wedge \dots \wedge \neg Fa_{i_n} : \{i_1, \dots, i_k\} \in AW\}$; onde AW é o conjunto de k -sobre- n seleções possíveis de k de n índices individuais $\{1, \dots, n\}$. Com base na consideração acima, para cada uma dessas seleções possíveis o seguinte se aplica: $P(Fa_{i_1} \wedge \dots \wedge Fa_{i_k} \wedge \neg Fa_{i_{k+1}} \wedge \dots \wedge \neg Fa_{i_n} | p(Fx)=r)^\wedge = r^k \cdot (1-r)^{n-k}$.

condicionado a todo o estado do mundo antes de t .) A condicionalização a evidências adicionais (hipotéticas) $E(b_1, \dots, b_n)$ só é permitida se essas evidências não disserem nada sobre cada indivíduo a quem o princípio de coordenação é aplicado; portanto, $b_j \neq a_i$ (para $1 \leq j \leq n$) deve ser aplicada. Também chamamos as provas adicionais de $E(b_1, \dots, b_n)$ de provas admissíveis. Sem essa limitação – a da restrição a “proposições permitidas” na Def. 7-1 –, o StK pode levar a inconsistências: por exemplo, se nossa hipótese é $H = (p(Fx|Gx) = 0,5) \wedge (p(Fx|Qx) = 0,8)$, então, teríamos também $P(Fa_i|Ga_i \wedge Qa_i \wedge H) = 0,5$ e $P(Fa_i|Ga_i \wedge Qa_i \wedge H) = 0,8$, ou seja, uma contradição. De acordo com o StK, apenas $P(Fa_i|Qa_i \wedge H) = 0,8$ e $P(Fa_i|Ga_i \wedge H) = 0,5$ são corretos, ou seja, Ga_i é inadmissível em (i) e Qa_i em (ii).

A Def. 7-2 explica o StK apenas para termos de um dígito: “ Fx ” e “ Gx ” são fórmulas abertas com somente uma variável x . Explicamos agora a generalização do StK para termos (ou para fórmulas) relacionais de múltiplos lugares: aqui, o “ x ” e o “ a_i ” associado representam, respectivamente, sequências de variáveis individuais x_1, \dots, x_r e constantes individuais a_1, \dots, a_i . Nesse caso, uma pequena modificação da Def. 7-2, porque é possível que a probabilidade estatística de Gx , seja influenciada pela condicionalização nas relações de x , para indivíduos $x_2 \neq x_1$. Assim, $p(Gx_i|R_{x_1, x_2}) \neq p(Gx_i)$: um exemplo é a incidência de homens de cabelos pretos (Gx) entre os pais de todos os pares pai-filho (Rxy); isso fica claro quando se olha para a definição $p(Gx|Rxy) = p(Gx \wedge Rxy) / p(Rxy)$. Em particular, $p(Gx|Rxy) \neq p(Gx|\exists y Rxy)$ ⁴¹, ou seja, a influência estatística de Rxy sobre Gx não pode ser substituída por um predicado de um dígito.

⁴¹ Por exemplo, 30% de todos os pais que têm filhos homens podem ter cabelos pretos, mas 60% de todos os pares pai-filho podem ter um pai de cabelos pretos como

(7-1) *Generalização do StK para relações*: As fórmulas na Def. 7-2 devem ser interpretadas da seguinte maneira:

- . a fórmula aberta $Gx =_{\text{def}} G(x_1, \dots, x_r)$ contém exatamente as variáveis individuais x_1, \dots, x_r ;

- . a fórmula aberta $Fx =_{\text{def}} F(x_1, \dots, x_r, z_1, \dots, z_m)$ pode conter a variável individual adicional z_k ($1 \leq k \leq m$);⁴²

- . são definidas para a variável individual x_i ($1 \leq i \leq r$), as constantes individuais determinadas injetivamente a_{i1}, \dots, a_{ir} e, para as variáveis individuais z_k ($1 \leq k \leq m$), as constantes individuais c_{k1}, \dots, c_{km} ; e

- . a evidência admissível $E(b_1, \dots, b_n)$ só pode conter constantes individuais que sejam diferentes tanto do a_{ik} ($1 \leq k \leq r$) quanto do c_{ih} ($1 \leq h \leq m$).

Finalmente, mostramos que, sse a condição de admissibilidade da prova $E(b_1, \dots, b_n)$ for atendida, duas propriedades importantes do StK podem ser demonstradas: sua consistência e sua confiabilidade estatística (ver Apêndice 10.3.9 para prova):

(Teorema 7-1): *Coerência e confiabilidade do StK*: Seja \mathcal{L} uma linguagem de lógica de predicados de 1ª ordem (estendida, se necessário, por operadores lógicos proposicionais infinitos), AL_{F_0} e AL_{Sent} a álgebra das fórmulas ou Teoremas de \mathcal{L} , e $p:AL_{F_0} \rightarrow [0,1]$ uma função de probabilidade estatística que satisfaz a lei da independência ((3-3) ou Teorema 10-1(a)). Nesse caso, aplica-se o seguinte:

primeiro membro, desde que os pais de cabelos pretos tenham mais filhos homens do que os pais que não têm cabelos pretos.

⁴² Como restrição, pode ser exigido que as variáveis individuais z_k não sejam conjuntivamente separáveis, como no caso $F(x_1, x_2, z_1) =_{\text{def}} F^*x_1x_2 \wedge Hz_1$. O último caso pode ser reduzido à probabilidade estatística $p(Gx_1 | F^*x_1x_2)$ usando o StK estendido, porque devido à independência estatística, $p(Gx_1 | F^*x_1x_2 \wedge Hz_1) = p(Gx_1 | F^*x_1x_2)$.

(a) O StK é coerente (consistente com a teoria da probabilidade), ou seja, *há* uma função de probabilidade subjetiva $P:AL_{\text{sent}} \rightarrow [0,1]$ que satisfaz o (StK) (Def. 7-2).

(b) O StK é *confiável* no seguinte sentido: $P(A|B \wedge H \wedge E) = r$ é uma instância do StK (7-2(b)), então, uma hipótese estatística da forma $p(A^*|B^* \wedge E^*) = r$, onde A^* , B^* e E^* surgem de A , B e E , respectivamente, pela substituição bijetiva de variáveis individuais por todas as constantes individuais.

(c) Sem limitação às provas *admissíveis*, (a) e (b) são violadas.

7.2 O conteúdo empírico-indutivo das hipóteses estatísticas

O StK para amostras aleatórias será usado a fim de trazer à tona o conteúdo empírico indutivo das hipóteses estatísticas mencionado na Seção 5.3. Vamos considerar a hipótese estatística $p(Fx) = r$. Disso, segue-se a lei binomial $p(h_n(Fx)=n/k) = \binom{n}{k} \cdot r^k \cdot (1-r)^{n-k}$, que afirma que o limite de frequência daquelas amostras, cuja frequência F é k/n , é $\binom{n}{k} \cdot r^k \cdot (1-r)^{n-k}$. Disso, não se segue logicamente nenhuma afirmação empírica sobre qualquer amostra particular de n elementos com k Fs, digamos a amostra $\{a_1, \dots, a_n\}$ que acabamos de demonstrar. Agora, precisamos do StK a fim de transferir a probabilidade estatística para a amostra concreta $\{a_1, \dots, a_n\}$ e, assim, chegar exatamente ao StK para amostras aleatórias (Def. 7-2(c)), ou seja, para probabilidade de crença $P(h(Fx|\{a_1, \dots, a_n\}) = k/n \mid p(Fx)=r) = \binom{n}{k} \cdot r^k \cdot (1-r)^{n-k}$. Entendemos por conteúdo indutivo-empírico da hipótese estatística $p(Fx) = r$ o conjunto de todas as sentenças de probabilidade epistêmica desse tipo que se seguem da aceitação da hipótese com $P=1$ e do StK:

(Def. 7-3) *Conteúdo empírico-indutivo de uma hipótese estatística H:*

Todas as sentenças de probabilidade epistêmica resultantes da aplicação do StK a H, bem como as consequências probabilísticas seguintes e a suposição $P(H) = 1$ (ver Def. 3-5).

Caso especial para $H = "p(Fx)=r"$: Todas as sentenças de probabilidade epistêmica da forma " $P(h(Fx|\{a_1, \dots, a_n\}) = k/n) = \binom{n}{k} \cdot r^k \cdot (1-r)^{n-k}$ " para todos $n \in \mathbb{N}$, $1 \leq k \leq n$ e constantes individuais a_1, \dots, a_n , bem como para suas consequências probabilísticas.

Baseamos, nesse conteúdo, os procedimentos de teste de hipóteses estatísticas que serão explicados no Cap. 8. Assim, por exemplo, é possível calcular o intervalo das frequências amostrais de F, onde se encontram as 95% das frequências amostrais mais prováveis, considerando que a hipótese da população geral $p(Fx) = r$ é verdadeira. É fácil demonstrar que esse intervalo deve ser simétrico em torno do valor r, que é a frequência de F na população geral. Conforme o método de teste de Fisher, esse intervalo é chamado de intervalo de aceitação (simples), e uma hipótese é (provisoriamente) rejeitada se a frequência de amostragem observada estiver fora desse intervalo de 95% de probabilidade de aceitação. Na Seção 9.1-2, discutiremos esse método à luz de nossa abordagem dualista.

Finalmente, justificamos nosso conceito de conteúdo indutivo-empírico de uma hipótese, como definido anteriormente:

(1.) Nós deliberadamente não determinamos o conteúdo de uma hipótese estatística H como um conjunto de proposições que são tornadas suficientemente prováveis por H, ou seja, o conjunto de todos os S, com $P(S|H) \geq 1-\epsilon$, para ϵ pequenas, e de todos os modelos de probabilidade satisfazendo o StK. Isso, como veremos na Seção 9.11, levará a *inconsistências* (palavra-chave: paradoxo da loteria). Essa é a

razão pela qual expressamos as noções de conteúdo probabilístico e indutivo por conjuntos de Teoremas de probabilidade.

(2.) Para determinar suas consequências probabilísticas, precisamos transformar as proposições “A”, que não são Teoremas de probabilidade epistêmica, em sua forma de probabilidade “ $P(A)=r$ ”. Por exemplo, se $P(A|B)=r$ é um Teorema probabilístico, então, “ $P(A)=r$ ” pertencerá ao conteúdo probabilístico de B.

(3.) O StK aplica-se a P. Isso ocorre porque o conteúdo destina-se a capturar as consequências observáveis de hipóteses estatísticas que podem ser esperadas *indutivamente*, e o StK contém o princípio indutivo de inferência de especialização (Seção 4.4), que transfere as tendências estatísticas da população para casos individuais ou para amostras individuais (ver Seção 7.5 para mais detalhes).

7.3 Probabilidades iniciais independentes da experiência

A limitação mais importante do princípio de coordenação estatística é a seguinte: o StK só pode ser aplicado sem restrição se P for uma *probabilidade inicial independente da experiência que não depende de qualquer observação de indivíduos particulares*. Tal função da crença também é chamada de função de probabilidade *a priori*, sendo que “*a priori*” não significa “necessário” no sentido kantiano, mas apenas “independente da experiência”: as probabilidades “*a priori*” dependem de “vieses subjetivos” e, portanto, são subjetivamente variáveis. Para *graus reais* ou “pessoais” de crença que dependem de experiências particulares, o StK geralmente não é válido. Se, por exemplo, sabemos pela observação no tempo t que

a moeda (a) recém-lançada pousou em cara (Ga), então, para nossa função de crença real P_t , no tempo t, $P_t(Ga) = 1$, ainda que saibamos que a probabilidade estatística de caras é $1/2$. Para nossa função de crença atual P_t , segue-se então (mesmo sem condicionamento à evidência inválida "Ga") que $P_t(Ga|p(Gx)=1/2) = 1$, em contradição com o StK. Contudo, mesmo que não tenhamos 100% de certeza de nossa observação de "Ga", mas por exemplo, $P_t(Ga) = 0,95$, um *conflito* surge.

Isso porque ainda estimaríamos o valor de $P_t(Ga|p(Gx)=1/2)$ em cerca de 0,95, pelo menos maior do que o valor de 0,5 que se seguiria do StK. Se, além disso, $P_t(p(Gx)=1/2) = 1$, então, $P_t(Ga|p(Gx)=1/2) = P_t(Ga) = 0,95$ necessariamente segue, e há novamente uma contradição com o StK. Somente se ainda não observamos o resultado do lançamento da moeda (Ga ou -Ga) e somente se não sabemos nada sobre ele além de sua probabilidade estatística, faz sentido atribuir um grau de crença de $1/2$ ao resultado Ga.

A seguir, para uma distinção clara, sempre designamos funções de crença real como " P_t " (o índice de tempo refere-se ao nível atual de experiência do sujeito) e funções iniciais independentes de experiência como "P". Aplicações do StK para " P_t " podem produzir contradições não apenas quando o evento consequencial Ga é observado, mas também quando vários eventos antecedentes possíveis são observados. Schurz (2012) dá o seguinte exemplo: suponha que a hipótese estatística H implica $p(\text{Pode_Voar}(x) | \text{Pássaro}(x)) = 0,95$ e $p(\text{Pode_Voar} | \text{Vive_na_Antártida}) = 0,01$. Conforme o StK, isso significa que para um indivíduo "a": (i): $P_t(\text{Pode_Voar}(a) | H \wedge \text{Pássaro}(a)) = 0,95$ e $P_t(\text{Pode_Voar}(a) | H \wedge \text{Vive_na_Antártida}(a)) = 0,01$. Se eu tenho certeza da observação de que um é um pássaro que vive na Antártida, então, minha função de

crença real (ii) é: $Pt(Pássaro(a)) = Pt(Vive_na_Antártida(a)) = 1$. No entanto, de (i) e (ii) segue-se $Pt(Pode_Voar(a)|H) = 0,095$ e $Pt(Pode_Voar(a)|H) = 0,01$, ou seja, uma contradição. Unterhuber e Schurz (2013) mostram, em um experimento psicológico, que a maioria dos sujeitos de teste não vê nenhuma contradição na combinação de (i) e (ii). Esses autores concluem que, de acordo com a abordagem dualista, a maioria das pessoas distingue entre probabilidades iniciais independentes da experiência e graus de crença dependentes da experiência.

De modo geral, os seguintes fatos estão por trás das limitações do StK explicadas: a determinação de P pelo StK só é possível sem restrição se P já não estiver parcialmente determinado em outro lugar, por exemplo, por meio da experiência adquirida que pode entrar em conflito com StK. Nessa situação, as experiências substituem a condição de racionalidade StK, que só se aplica quando não existem dados empíricos adicionais, além da frequência de ocorrência sobre o fato cujo grau de crença é determinado pelo StK. Carnap (1971, p. 21-23) também chama de função a priori da crença a “função de credibilidade Cred”, em oposição à função de crença real, que ele chama de “função de credibilidade C”. Outra forma de expressar o caráter a priori de P é a interpretação de “*tudo o que sei*” de Pearl (1988, p. 475): conforme essa interpretação, $P(A|B)$ é o meu hipotético (ou contratual) grau de crença em A, se tudo o que eu sei é B, e $P(A)$ é o meu grau de crença em A, se eu não sei nada.

Bayesianos dualistas são céticos em relação a probabilidades a priori, uma vez que nenhum estado epistêmico é completamente independente da experiência. Eles preferem probabilidades iniciais que são parcialmente independentes da experiência e às quais o StK deve ser aplicado (por exemplo, Hawthorne, 2005, p. 305). Isso só não

é problemático sob a seguinte condição: o StK só pode ser aplicado sem restrição a essas (seqüências de) constantes individuais a_i (ou c_j em (7-1)) sobre as quais ainda não foi feita nenhuma experiência incluída na medida P . Tais experiências só podem ser obtidas por outros indivíduos, ou seja, devem ser provas admissíveis “ $E(b_1, \dots, b_n)$ ” na acepção da Def. 7-2 e da sentença 7-2.⁴³

7.4 Das probabilidades basais às crenças atuais: condicionalização à evidência geral

O fato de o StK ser irrestritamente aplicável apenas à probabilidade inicial independente da experiência “ P ” não significa que ele não possa ser aplicado a Teoremas empíricos reais como argumentos da função de probabilidade atual $P_t(-)$, em um dado ponto no tempo t – pelo contrário, esse é o propósito final da estimativa racional de expectativas. Contudo, ao aplicar o StK à experiência real, a *regra da evidência geral* deve ser respeitada: StK 7-2(b) só pode ser aplicada aos Teoremas empíricos $G(a_i)$ e $F(a_i)$ se $F(a_i)$ contiver todos os elementos de prova pertinentes. No exemplo acima, não tenho permissão para definir o StK como $P_t(\text{Pode_Voar}(a) \mid H \wedge \text{Pássaro}(a))$ e nem como $P_t(\text{Pode_Voar}(a) \mid H \wedge \text{Vive_na_Antártida}(a))$, mas apenas como $P_t(\text{Pode_Voar}(a) \mid H \wedge \text{Pássaro}(a) \wedge \text{Vive_na_Antártida}(a))$, desde que “ $\text{Pássaro}(a) \wedge \text{Vive_na_Antártida}(a)$ ” seja minha evidência total no

⁴³ O enfraquecimento desta condição para sentenças atômicas com as quais ainda não se adquiriu experiência levaria a contradições com a regra de condicionalização (Capítulo 7.4).

momento em questão. Suponhamos que H implique a hipótese estatística $p(\text{Pode_Voar}|\text{Pássaro} \wedge \text{Vive_na_Antártida}) = 0,05$ (porque quase todas as aves na Antártida são pinguins). Então, eu tenho permissão para usar o StK em $P_t(\text{Pode_Voar}(a) | H \wedge \text{Pássaro}(a) \wedge \text{Vive_na_Antártida}(a)) = 0,05$. E, considerando que $\text{Pássaro}(a) \wedge \text{Vive_na_Antártida}(a)$ é toda a evidência: $P_t(\text{Pode_Voar}(a)) = 0,05$.

De modo geral, a relação entre probabilidades iniciais e graus reais de crença é estabelecida pelo seguinte princípio de condicionalização à evidência geral (ver Carnap, 1971, p. 18; Earman, 1992, p. 34; Howson; Urbach, 1996, p. 102).

(Def. 7-4) *Condicionização à evidência geral*: Seja $P = P_o$ a probabilidade inicial (de um determinado sujeito) em um ponto inicial t_o , deixe P_t ser a probabilidade real no tempo t e deixe W_{o-t} ser o ‘conhecimento’ total singular e estatístico que essa pessoa adquire entre t_o e t (mais precisamente, W_{o-t} é a conjunção de todas as opiniões previamente adquiridas: $P_t(W_{o-t}) = 1$). Então, para cada proposição S , tem-se: $P_t(S) = P_o(S | W_{o-t})$.

Porque $P_t(W_{o-t}) = 1$, $P_t(S | W_{o-t}) = P_o(S | W_{o-t})$, ou seja, a probabilidade inicial $P_o(S | W_{o-t})$ é transferida para o grau atual de crença. A produção dos produtos referidos na Seção 7.3 é evitada, porque a regra de condicionalização segue o princípio da *evidência total*: o grau real de crença de todas as proposições não evidentes é condicionado à mesma evidência geral W_{o-t} , $P_t(-) = P_o(- | W_{o-t})$, da qual se segue (de acordo com TB₁ do Teorema 3-3) que $P_t(-)$ também é coerente se P_o é coerente.

Juntamente ao StK, a Def. 7-4 é a grafia precisa do princípio de classe de referência mais próximo, explicado na Def. 2-3: esse último pode agora ser derivado do StK e da regra de condicionalização, como segue. Suponhamos que $R(a_i)$ contenha todas as nossas evidências sobre a sequência individual a_1, \dots, a_n , ou seja, todos os conjuntos de experiências que contêm uma das constantes individuais a_1, \dots, a_n ; seja $E(b_j)$ o nosso conhecimento sobre outros indivíduos e H o nosso conhecimento estatístico total, o que implica a hipótese estatística $p(G(x_i)|E(x_i)) = r$. Com base no StK (Def. 7-2), isso implica para a probabilidade inicial:

$$(7-2) P_o(G(a_i) | R(a_i) \wedge E(b_j) \wedge H) = r$$

De acordo com o pressuposto, $W_{o-t} = R(a_i) \wedge E(b_j) \wedge H$, do qual se segue, em conformidade com a regra de condicionalização (Def. 7-4), para a função da crença real,

$$(7-3) P_t(G(a_i)) = r$$

em outras palavras, o princípio da classe de referência mais estreita. A função da crença, formada segundo o princípio da classe de referência mais estreita, é, portanto, também coerente.⁴⁴

Ao condicionalizar a apenas uma classe de referência, o princípio da classe de referência mais estreita evita contradições – mas essa não é sua única justificativa: uma justificativa mais profunda desse princípio, juntamente a regras úteis para sua prática de aplicação, é dada nas Seções 7.8-9. As regras de condicionalização da Def. 7-4

⁴⁴ Tal teorema pode ser encontrado em Baco (1990, p. 151, teorema 50), mas sem prova.

também funcionam gradualmente: seja P_{t-1} a probabilidade real de crença no tempo $t-1$ e deixe W_t ser o conhecimento recém-adquirido no tempo t . Nesse caso, aplica-se o seguinte:

$$(7-4) \text{ Condicionização passo a passo: } P_t(S) = P_{t-1}(S|W_t).$$

A condicionização nas Def. 7-4 e (7-4) também é chamada de condicionização estrita, porque a certeza ($P(E) = 1$) é assumida para a evidência à qual a condicionização é feita. Uma generalização da condicionização estrita é a *condicionização de Jeffrey*, na qual a função de crença é condicionada a evidências recém-adquiridas, mas incertas. Seja $\{E_1, \dots, E_k\}$ uma partição de evidências possíveis por meio das quais se adquire os novos graus reais de crença $P_t(E_i)$ no tempo t . Em seguida, aplica-se o seguinte a cada frase S :

$$(7-5) \text{ Condicionização de Jeffrey: } P_t(S) = \sum_{1 \leq i \leq k} P_{t-1}(S|E_i) \cdot P_t(E_i).$$

7.5 Probabilidades de suporte e o problema da evidência antiga

Graus de crença racionais orientados por estimativas de frequência são estabelecidos determinando-se o grau de crença de todas as proposições não-evidenciais condicionalmente à evidência total por meio do Princípio de Coordenação Estatística (StK). Uma vez que todas as consequências do StK estão contidas na probabilidade inicial P , independentemente da experiência, isso desempenha um papel fundamental na formação de graus racionais de crença. A seguir, discute-se algumas características dessa probabilidade inicial.

Em primeiro lugar, o StK determina apenas a probabilidade inicial para sentenças singulares cujas constantes individuais podem ser substituídas por variáveis, mas não para hipóteses gerais sem constantes individuais, como, por exemplo, “todos os Fs são Gs” ou “90% de todos os Fs são Gs”. A probabilidade inicial (“probabilidade prévia” [prior]) de hipóteses gerais é assumida como “de alguma forma dada” no contexto do bayesianismo subjetivo, baseado nos “pré-julgamentos” subjetivos. A probabilidade final (“probabilidade posterior”) da hipótese H, dada a evidência empírica E, é calculada usando o teorema de Bayes (TB5 dos Teoremas 3-3), como $P(H|E) = P(E|H) \cdot P(H)/P(E)$, onde a verossimilhança $P(E|H)$ é calculada usando o StK para amostragem aleatória.

Além disso, o StK determina a probabilidade inicial epistêmica para sentenças singulares apenas *condicionalmente* a hipóteses estatísticas, ou seja, $P(Ga|Fa \wedge H)$. A probabilidade inicial $P(Ga|Fa)$ de uma proposição singular Ga – e aplicando a regra de condicionalização ao grau real de crença $P_t(Ga)$, com Fa como evidência geral – só é assim determinada se a probabilidade estatística da fórmula atribuída (Gx) é *conhecida* ou é acreditada como $P=1$. No entanto, essa limitação pode ser sanada assumindo uma distribuição inicial subjetiva (probabilidade) sobre todas as hipóteses estatísticas possíveis (desconhecidas) de uma dada partição $\{H_1, \dots, H_n\}$. Assim, a probabilidade inicial de cada Teorema singular – incondicional às hipóteses estatísticas – é definida como o valor subjetivo esperado das probabilidades estatísticas da seguinte forma:

(Teorema 7-2) Probabilidades de suporte como valores esperados subjetivos de probabilidades estatísticas:

(a) $P(Ga_i \wedge E(b)) = \sum_{1 \leq j \leq n} P(Ga_i | H_j \wedge E(b)) \cdot P(H_j | E(b)) = \sum_{1 \leq j \leq n} r_j \cdot P(H_j | E(b))$ (de acordo com o StK). Aqui, H_j é a hipótese “ $p(Gx) = r_j$ ”, e $\{H_1, \dots, H_n\}$ é a partição de todas as hipóteses possíveis dessa forma, cuja disjunção tem o grau inicial de crença 1. Para partições contínuas ($H_r: r \in [0,1]$), a soma deve ser substituída por uma integral: $P(Ga_i) = \int_r r \cdot D(r) \cdot dr$ (com $D(r)$ como densidade de probabilidade de H_r , ver Seção 8.6). $E(b)$ é prova admissível (com $b_j \neq a_i$, conforme explicado abaixo na Def. 7-2).

(b) Para probabilidades iniciais condicionais, $P(Ga_i | Fa_i)$ é derivado de (a):

$P(Ga_i | Fa_i \wedge E(b)) = \sum_{1 \leq s \leq m} r_s \cdot P(K_s | Fa_i \wedge E(b))$, onde K_s é a hipótese estatística “ $p(Gx | Fx) = r_s$ ”, e $\{K_1, \dots, K_m\}$ é a partição de todas as hipóteses possíveis dessa forma.

O Teorema 7-2(a) segue o StK e os axiomas básicos da teoria da probabilidade. O Teorema 7-2(b) segue a 7-1(a) aplicando o StK incondicional ao numerador e denominador da definição $p(Ga | Fa \wedge E(b)) = p(Ga \wedge Fa | E(b)) / p(Fa | E(b))$; para as provas, ver o Apêndice 10.3.10.⁴⁵ Os graus atuais de crença são obtidos condicionando-se à evidência total ou à classe de referência mais estreita Fa_i , como $P_t(Ga_i) = P(Ga_i | Fa_i)$, onde a última expressão é determinada como em (b) acima. Hawthorne (2005) chama as

⁴⁵ Bacchus (1990, p. 189) escreve que o valor esperado direto das probabilidades condicionais nem sempre é idêntico aos quocientes dos valores esperados das probabilidades incondicionais. Isto só é verdade se não condicionarmos a probabilidade da hipótese subjetiva ao antecedente, ou seja, se escrevermos “ $P(K_s | E(b))$ ” em vez de “ $P(K_s | Fa_i \wedge E(b))$ ” no Teorema 7-2(b). $P(K_s | E(b))$ geralmente difere de $P(K_s | Fa_i \wedge E(b))$.

probabilidades iniciais formadas de acordo com o Teorema 7-2(a,b) de “funções de suporte”, e nós as chamamos de “probabilidades de suporte”.

As probabilidades de suporte também independem da experiência e, por isso, são denotadas pela letra “P” ou “P_o”, pois dependem apenas da distribuição inicial sobre o espaço de possíveis hipóteses gerais $\{H_1, \dots, H_n\}$, porém não de experiências particulares. Hawthorne argumenta que as medidas de confirmação bayesianas não devem ser baseadas em funções de crença reais, mas sim em probabilidades de suporte. Concordamos com essa opinião. Isso ocorre porque todas as medidas bayesianas de confirmação fazem a confirmação de uma hipótese H pela evidência E da verossimilhança $P(E|H)$ (ou numa relação entre $P(E|H)$ e $P(E)$, por exemplo, a diferença $P(E|H) - P(E)$). Tal medida, no entanto, expressa toda a influência de H na probabilidade de E somente se a probabilidade de E ainda não tiver sido determinada em outro lugar pela experiência atual.

Se, em vez disso, a probabilidade fosse identificada com a probabilidade real de crença, as probabilidades epistêmicas não poderiam ser consistentemente identificadas com as probabilidades estatísticas e, posteriormente, variariam entre sujeitos com experiências diferentes. Para a pessoa A, que não havia observado o resultado do lançamento da moeda, $P_{A,t}(\text{cara}|H) = 0,5$ e, para a pessoa B, que o havia observado, $P_{B,t}(\text{cara}|H) = 1$ ou, por exemplo, $= 0,99$. Se as probabilidades variassem subjetivamente, não apenas as probabilidades iniciais das hipóteses seriam subjetivamente variáveis, mas também as suas probabilidades finais permaneceriam subjetivamente variáveis, ou seja, a convergência intersubjetiva das probabilidades finais com as evidências empíricas comuns crescentes (discutidas na seção 9.5), que é tão importante para os bayesianos, não

poderia ocorrer. Por essa razão, a probabilidade inicial P (à qual o StK é aplicável) também deve ser usada para medidas de confirmação.

A confusão de P com P_t também está subjacente ao problema da evidência antiga, que é visto como um problema fundamental do bayesianismo (Earman, 1992, Cap. 5). Suponhamos que a pessoa A esteja em um estado epistêmico, no tempo t , no qual ela já conhece a evidência E , ou seja, $P_{A,t}(E) = 1$. Conclui-se que $P_{A,t}(E) = P_{A,t}(E|H) = 1$ e, portanto, (de acordo com TB5 da sentença 5-3) que $P_{A,t}(H|E) = P_{A,t}(H)$. Se alguém expressasse graus de confirmação usando graus atuais de crença, então, obteria o resultado absurdo de que a evidência E , favorecendo a hipótese H , não poderia mais confirmar a hipótese no momento em que ela se tornou conhecida e foi acreditada com $P_{A,t} = 1$; assim, mesmo que $P_{A,t}(E)$ seja menor que 1, ainda se obtém o resultado irracional de que o grau de confirmação de H por E dependeria de quão certo alguém está da evidência E . Com a diferença $P(H|E) - P(H)$, pode-se de fato expressar uma medida do aumento da *probabilidade* de H por E , e P não deve assumir implicitamente que a evidência E seja certa ou altamente provável. Além disso, P não deve assumir outras evidências E' como altamente prováveis se elas aumentarem a probabilidade de H . Por essa razão, a proposta de Howson e Urbach (1996, p. 404 e seg.) de usar a função de probabilidade contratada $(P_{A,t})^{-E}$ para “ P ” também é insuficiente. $(P_{A,t})^{-E}$ é a função de crença do sujeito A em um hipotético estado precursor no qual E ainda não é conhecido. De acordo com essa proposta, o grau de confirmação de H por E ainda dependeria do quanto *outra* evidência E' , apoiando H , já é implicitamente acreditada. Se, por exemplo, H é aumentado de 0,5 para 0,95 por duas evidências E e E' , ambas independentemente uma da outra e juntas, então, H não seria confirmado por E mesmo quando se usasse a função contratada $(P_{A,t})^{-E}$.

E , porque $(P_{A,t})^{-E}(E') = 1$ e, portanto, $(P_{A,t})^{-E}(H|E') = (P_{A,t})^{-E}(H) = (P_{A,t})^{-E}(H|E) = 0,95$. Isso mostra, mais uma vez, que P deve ser independente de qualquer experiência atual usada para julgamentos confirmatórios.

7.6 Permutabilidade e teorema da representação de De Finetti

As probabilidades a priori ou iniciais, formadas com a ajuda do StK e conforme o (Teorema 7-2), têm uma propriedade fundamental que é formulada, pela primeira vez, por De Finetti: satisfazem o axioma da *permutabilidade* ou o axioma equivalente da simetria de acordo com Carnap (1971, p. 117).

(Def. 7-5) Dada uma linguagem \mathcal{L} com um conjunto contável $\mathcal{K} = \{a_1, a_2, \dots\}$ de constantes individuais.

Uma função de probabilidade epistêmica P é chamada de *permutável*, em relação a um subconjunto K de \mathcal{K} , sse P é *invariante* em relação a permutações arbitrárias de constantes individuais em K – ou seja, para todas as $n \geq 1$, sentenças $A(a_1, \dots, a_n)$ e funções de permutação (funções bijetivas) $\pi: \mathcal{K} \rightarrow \mathcal{K}$ $P(A(a_1, \dots, a_n)) = P(A(\pi(a_1), \dots, \pi(a_n)))$.

P significa infinitamente permutável sse P é permutável em relação a \mathcal{K} .

A permutabilidade de uma probabilidade de suporte P é uma consequência direta das probabilidades de suporte de acordo com o Teorema 7-2, uma vez que, para quaisquer constantes individuais a_i , o seguinte é válido: $P(Ga_i) = \sum_j p_H(Gx) \cdot P(H_j)$, para todos a_i em K ou \mathcal{K} .

Se P for permutável para todas as constantes individuais em \mathcal{K} , então, é uma probabilidade inicial ou a priori completamente independente da experiência (Carnapiana); se P só pode ser trocado por um subconjunto $K \subseteq \mathcal{K}$ sobre o qual nenhuma experiência foi feita, então, P é uma probabilidade inicial parcialmente independente da experiência no sentido explicado.

Acrescente-se que o pressuposto da permutabilidade só é válido sem contradição se todos os predicados definidos forem substituídos por suas definições por meio de predicados básicos. A mesma restrição aplica-se ao StK. Isso é importante para cumprir os requisitos estabelecidos na Seção 9.7 e discutidos no paradoxo de Goodman. Nesse caso, um predicado é definido com o auxílio de constantes individuais, por exemplo, “ G^*x ” (x é verdul⁴⁶) $\leftrightarrow_{\text{def}}$ “ x é verde e pertence à amostra $\{a_1, \dots, a_{10}\}$ ou x é azul e não pertence a ela”.

A permutação de constantes individuais só funciona se for aplicada a *todas as* constantes individuais da fórmula complexa, incluindo aquelas em $\{a_1, \dots, a_{10}\}$. No entanto, na definição de “verdul”, esses indivíduos são “escondidos” no *definiens* e retirados do reino da troca. Como resultado, a aplicação simultânea do princípio da permutabilidade ao verdul e ao verde pode levar a contradições. Da mesma forma, o StK só funciona se todas as constantes individuais (incluindo aquelas em $\{a_1, \dots, a_{10}\}$) forem substituídas por variáveis individuais.

46 Nota do tradutor: quando o autor usa “*grot*”, uma combinação de “*grün*” e “*rot*”, para verde e vermelho, respectivamente, na explicação dos predicados de Goodman, prefere-se usar “verdul”, combinação de verde e azul por dois motivos: 1- no original, em inglês, de Goodman, *Facts, Fiction and Forecast*, ele usou “green” e “blue”, gerando “grue”; 2- pela facilidade de combinar partes das palavras em português.

A permutabilidade de P é mais fraca do que a independência probabilística para P : permite relações de *suporte indutivo* para P da forma $P(Fa_2|Fa_1) > P(Fa_2) > P(Fa_2|\neg Fa_1)$, desde que esses suportes sejam *independentes dos indivíduos*, ou seja, apliquem-se a todas as permutações individuais: $P(Fa_i|Fa_j) > p(Fa_i) > p(Fa_i|\neg Fa_j)$ para todos os $a_i \neq a_j$ em K e \mathcal{K} , respectivamente. Como dito na Seção 3.2, a natureza indutiva de P deve-se ao fato de a verdadeira probabilidade estatística $p(Fx)$ ser assumida como desconhecida: quanto mais vezes Fa_j já ocorreu, maior será a probabilidade esperada $p(Fx)$ e, portanto, o $P(Fa_i|Fa_j)$. Por outro lado, o pressuposto de permutabilidade para P só faz sentido se for assumido que a função de probabilidade estatística subjacente p satisfaz o princípio da independência (cf. Gillies, 2000, p. 69-83). Se se acredita que a independência estatística é violada, ou seja, que as disposições estatísticas dos eventos mudam em um determinado ponto no tempo, as probabilidades iniciais subjetivas de eventos antes e depois desse ponto no tempo não devem ser permutáveis. Spielman (1976) mostra que, assumindo σ -aditividade, a permutabilidade de P implica: as probabilidades estatísticas são independentes com probabilidade subjetiva 1.

De Finetti (1931) prova um importante teorema de representação, segundo o qual toda função de probabilidade subjetiva infinitamente permutável, em uma faixa infinita contável de indivíduos, é idêntica a um valor subjetivo esperado (ou seja, a uma média ponderada) de funções de probabilidade estatísticas independentes (“Bernoullianas”), no sentido do Teorema 7-2. Além disso, a permutabilidade é comprovadamente equivalente ao (StK), juntamente à suposição de que, com probabilidade subjetiva 1, cada evento da álgebra subjacente tem um limite de frequência $p(E)$, e p satisfaz a independência estatística (ver Apêndice 10.3.11 para a prova).

(Teorema 7-3) *Permutabilidade de P*: Seja \mathcal{L} uma linguagem lógica de predicados e P ou p uma medida de probabilidade sobre as sentenças ou fórmulas, $\text{Sent}(\mathcal{L})$ ou $\text{Form}(\mathcal{L})$, de \mathcal{L} . Em seguida, as seguintes afirmações são equivalentes – a equivalência entre (1) e (2) aplica-se apenas a uma linguagem com nomes padrão para um intervalo contável-infinito de indivíduos, e a adição entre colchetes somente se P for σ -aditivo:⁴⁷

(1) P é infinitamente permutável.

(2) (i) P sobre $\text{Sent}(\mathcal{L})$ pode ser representado como o valor esperado de P para as funções de probabilidade estatística, p sobre $\text{Form}(\mathcal{L})$, de acordo com a sentença 7-2 [(ii) onde p satisfaz a independência estatística de acordo com (3-3) ou com sentença 10-1(a)].

(3) (i) com probabilidade subjetiva $P = 1$, cada fórmula $A \in \text{Form}(\mathcal{L})$ tem um limite de frequência $p(A)$, onde (ii) P e p estão conectados pelo StK [(iii) p satisfaz a independência estatística].

7.7 Regularidade e aprendizagem indutiva

A permutabilidade de P e o StK equivalente são pressupostos de indução probabilística *fracos*. A longo prazo, o StK transfere as tendências de frequência dos tipos de eventos para quaisquer eventos individuais ou amostras. Isso só faz sentido sob a suposição indutiva de que qualquer amostra ou período de tempo específico é *prima facie*

⁴⁷ Veja de Finetti (1931, 1964); Carnap (1980), Hewitt/Savage (1955), Kingman (1978). Para domínios finitos de indivíduos ou permutabilidade restrita, o teorema da representação é apenas aproximadamente válido (Diaconis e Freedman 1980).

(ou seja, salvo se conhecido de outra forma) representativo de toda a população ou todo o curso do mundo.

A permutabilidade de P baseia-se no pressuposto de que todos os indivíduos possuem as mesmas tendências probabilísticas, independentemente de suas características particulares. Se pensarmos nos indivíduos como eventos organizados ao longo do tempo e do espaço, então, a permutabilidade significa que as tendências probabilísticas dos eventos para produzir outros eventos são as mesmas para todas as posições no espaço e no tempo – também uma suposição de uniformidade indutiva. Pode-se provar que a permutabilidade de P , juntamente à condição de regularidade (Def. 6-2(b)), implica a propriedade de aprendizagem indutiva uniforme a partir da experiência. A condição de regularidade também é chamada de não-dogmaticidade de P , uma vez que a regularidade é um pré-requisito para a aprendizagem indutiva: se a probabilidade inicial de uma hipótese H for extrema, quer dizer, 0 ou 1, isso não pode mais ser alterado por nenhuma nova experiência E , ou seja, $P(H|E)$ permanece 0 ou 1 para todas as experiências possíveis E . Sob a Def. 6-1, já mencionamos que a regularidade só faz sentido para espaços de possibilidade finitos ou contáveis. A partir do momento em que as possibilidades têm de ser descritas por conjunções infinitamente longas de sentenças atômicas, há um número incontável de possibilidades, e a regularidade não pode mais se aplicar de modo geral. Normalmente, para linguagens lógicas de predicados, só se requer regularidade para orações singulares (finitamente longas):

(Teorema 7-4) *Aprendizagem indutiva uniforme* (ou “relevância de instanciação”): Se P é infinitamente permutável para uma linguagem lógica de predicados com nomes padrão a_1, a_2, \dots , e

regularmente sobre suas sentenças singulares (ou seja, $P(S) \neq 0,1$ para todas as sentenças singulares com $\text{Mod} \supset \text{Mod}(S) \supset \emptyset$), então, a confirmação indutiva de previsões singulares aumenta continuamente com o número de instâncias que a suportam:
 $P(\text{Fa}_{n+1} | \text{Fa}_1 \wedge \dots \wedge \text{Fa}_n) > P(\text{Fa}_{n+1} | \text{Fa}_1 \wedge \dots \wedge \text{Fa}_{n-k})$ (para todos k com $0 < k < n$ e $n \in \mathbb{N}$).

A propriedade de aprendizagem uniforme da experiência também é chamada de princípio da relevância da instanciação (“relevância instancial”) (no caso de F_x , também pode ser um predicado complexo). A prova dos Teoremas 7-4 é baseada na desigualdade de Cauchy-Schwartz e pode ser encontrada, por exemplo, em Humburg (1971, p. 233, th. 5) e em Kutschera (1972, p. 128, “C11” e n. 10).⁴⁸

A natureza indutiva da permutabilidade é demonstrada por Kutschera (1972, p. 74), Earman (1992, p. 108) e Gillies (2000), entre outros. Essa percepção é significativa, porque alguns autores argumentam que a permutabilidade é uma propriedade *a priori* ou logicamente válida (por exemplo, De Finetti, 1937; Carnap, 1971; van Fraassen, 1989, Cap. 7). Contudo, como explicado anteriormente, é fácil construir situações em que a permutabilidade de P pode ser razoavelmente rejeitada, por exemplo, quando se considera provável que as tendências probabilísticas ou “leis naturais” que governam o curso do mundo mudem ao longo do tempo (ver Gillies 2000, p. 77-82).

⁴⁸ Humburg usa adicionalmente o “axioma de Reichenbach”, que é idêntico ao Teorema 9-5(b), mas que (como ali mostrado) é ele próprio derivável sob a suposição de regularidade (ver Kutschera 1972, p. 129, n. 13).

Outras consequências indutivas resultantes da permutabilidade, da regularidade ou da σ -aditividade são discutidas nas Seções 9.3-5 no contexto da estatística bayesiana. A permutabilidade de P é uma suposição indutiva mais forte do que a σ -aditividade, mas ainda é muito fraca para determinar uma distribuição de probabilidade inicial única $P(H_i)$ sobre uma partição de hipóteses possíveis $\{H_1, \dots, H_n\}$. Para tornar essa última possível, os defensores do bayesianismo intersubjetivo ou ‘objetivo’⁴⁹ propõem pressupostos ainda mais fortes para P . O pressuposto mais importante desse tipo é o princípio da indiferença, de acordo com o qual, em um estado de ignorância, todas as hipóteses possíveis *devem receber a mesma* probabilidade inicial.

A concepção “clássica” de probabilidade de Laplace (1814), o “bayesianismo objetivo” de H. Jeffrey (1939) e de Williamson (2010, p. 28) e o conceito “lógico” de probabilidade de Keynes (1921) e de Carnap (1950; 1971) baseiam-se no princípio da indiferença (cf. Gillies, 2000, Cap. 3). Como pode ser visto na Seção 9.3, no entanto, há fortes objeções ao princípio da indiferença, as quais mostram que mesmo esse princípio não garante a intersubjetividade. Por essas e outras razões, a ideia de Laplace-Carnap de que apenas as “leis lógicas” devem caracterizar uma escolha inequívoca de probabilidades razoáveis parece insustentável (cf. Kutschera, 1972, p. 144; Seção 9.3; Seção 9.4; Seção 9.7). Probabilidades “lógicas” não são um “terceiro” tipo de probabilidades, mas funções de probabilidade epistêmica limitadas por fortes axiomas adicionais que buscam aceitação intersubjetiva.

⁴⁹ Representantes são, por exemplo, H. Jeffrey (1939), Keynes (1921), Carnap (1950, 1971) ou Williamson 2010, p. 28). cf. Gillies (2000, Cap. 3).

7.8 A justificativa das classes de referência mais estreitas

Voltando à regra da classe de referência mais estreita, ainda há uma pergunta sem resposta. Para um evento previsível F_a , geralmente existem várias classes apropriadas de referência que atribuem diferentes probabilidades estatísticas a F_a (cf. o exemplo da ave que vive na Antártida na seção 7.3). Se formássemos graus subjetivos de crença usando diferentes classes de representação, teríamos graus conflitantes de crença, por exemplo, $P(\text{Pode_Voar}(a)|\text{Vive_na_Antártida}(a)) = 0,01$ e $P(\text{Pode_Voar}(a)|\text{Pássaro}(a)) = 0,95$, embora $P(\text{Pássaro}(a)) = P(\text{Vive_na_Antártida}(a)) = 1$. Nesse sentido, Hempel (1965, § 3.4) e Coffa (1974) falam da *ambiguidade* das conclusões estatísticas. Para evitar tais contradições, portanto, é necessário que um determinado indivíduo selecione uma *excelente* classe de referência, à qual se aplica o StK para graus reais de crença. No entanto, por que essa deveria ser a *classe de referência* mais estreita?

Essa pergunta é respondida por um argumento teórico da decisão de Good (1966) (ver também Rosenkrantz, 1977 e Horwich, 1982, p. 125-128). O argumento considera a *utilidade* esperada $UE(h_i)$ de possíveis ações alternativas h_1, \dots, h_m em possíveis circunstâncias alternativas u_1, \dots, u_n , expressa pelas previsões correspondentes V_1, \dots, V_n , cujas probabilidades devem ser avaliadas (V_i diz que u_i ocorre). Essa utilidade esperada é calculada pela fórmula da teoria da decisão explicada em (5-3) da seguinte forma:

(7-6) $UE(h_k) = \sum_{i: s_i \leq n} P(V_i) \cdot U(h_k, V_i)$, com $\{V_1, \dots, V_n\}$ como a partição das previsões sobre possíveis circunstâncias e $U(h_k, V_i)$ como a utilidade da ação h_k quando a predição V_i ocorre.

A utilidade esperada [UE] corresponde à utilidade média, a longo prazo, se a probabilidade epistêmica da predição V_i é obtida a partir da probabilidade estatística com a ajuda do StK. Do conjunto de ações possíveis, o decisor racional seleciona uma ação com a *máxima utilidade* esperada, a fórmula “um” permite várias ações igualmente úteis-máximas. Good (1966) prova que a condicionalização de P a novas evidências nunca pode reduzir a utilidade esperada da ação escolhida, mas pode aumentá-la, precisamente quando a condicionalização de P (isto é, a substituição de $P(-)$ por $P(-|E_i)$) fizer diferença para a utilidade escolhida. Como essa prova é filosoficamente instrutiva, nós a esboçamos aqui para o caso mais simples de um experimento com dois resultados possíveis, E e $\neg E$. A utilidade esperada da ação h_k , *antes* da execução do experimento e, é dada como

$$(7-7) UE(h_k) = \sum_{i: s_i \leq n} P(V_i) \cdot U(h_k, V_i).$$

A utilidade esperada de h_k , *após* a execução do experimento (E, $\neg E$) e a condicionalização de P sobre seu resultado, escrito como $UE(h_k|(E, \neg E))$, é calculada do seguinte modo:

$$(7-8) UE(h_k|(E, \neg E)) = P(E) \cdot \sum_{i: s_i \leq n} P(V_i|E) \cdot U(h_k, V_i) + P(\neg E) \cdot \sum_{i: s_i \leq n} P(V_i|\neg E) \cdot U(h_k, V_i).$$

Em palavras: $UE(h_k|(E, \neg E))$ é a utilidade esperada de h_k condicionalizada a E multiplicada pela probabilidade de E, mais a

utilidade esperada de h_k condicionalizada a $\neg E$ multiplicada pela probabilidade de $\neg E$.

As duas utilidades esperadas em (7-7) e em (7-8) são exatamente idênticas, porque

$$(7-9) P(E) \cdot \sum_{i \leq i \leq n} P(V_i|E) \cdot U(h_k, V_i) + P(\neg E) \cdot \sum_{i \leq i \leq n} P(V_i|\neg E) \cdot U(h_k, V_i) = \sum_{i \leq i \leq n} P(V_i \wedge E) \cdot U(h_k, V_i) + P(V_i \wedge \neg E) \cdot U(h_k, V_i) = \sum_{i \leq i \leq n} P(V_i) \cdot U(h_k, V_i) = UE(h_k).$$

Assim, a condicionalização para $(E, \neg E)$ não altera a utilidade esperada se a ação maximizadora da utilidade não se alterar como resultado da condicionalização. Contudo, assim que, sob apenas uma evidência possível, digamos sob E , a ação maximizadora da utilidade não é mais h_k , mas, digamos, h_r , o maximizador da utilidade escolhe h_r sob a condição E e h_k sob a condição $\neg E$.⁵⁰ A utilidade esperada dessa ação condicional $h^* =_{\text{def}} "h_k \text{ se } \neg E \text{ e } h_r \text{ se } E"$ aumenta em comparação com $UE(h_k) = UE(h_k|(E, \neg E))$, porque

$$(7-10) UE(h^*|(E, \neg E)) = P(E) \cdot \sum_{i \leq i \leq n} P(V_i|E) \cdot U(h_r, V_i) + P(\neg E) \cdot \sum_{i \leq i \leq n} P(V_i|\neg E) \cdot U(h_k, V_i),$$

e dado que $\sum_{i \leq i \leq n} P(V_i|E) \cdot U(h_r, V_i) > \sum_{i \leq i \leq n} P(V_i|E) \cdot U(h_k, V_i)$;
 assim, $UE(h^*|(E, \neg E)) > UE(h_k|(E, \neg E))$.

Horwich se opôs à evidência de Good de que ela seria limitada a objetivos práticos (não epistêmicos) de ação; no entanto, essa objeção não se aplica, porque as ações possíveis também podem ser ações puramente epistêmicas, por exemplo, as previsões V_i das

⁵⁰ Assumimos aqui que $\neg E$ não tem influência na utilidade esperada; caso contrário, argumenta-se analogamente para a condição $\neg E$.

circunstâncias possíveis e, no caso de sua utilidade, simplesmente o sucesso da previsão, expresso por uma medida de distância entre a previsão realmente feita e a verdadeira na partição dada V_1, \dots, V_n (cf. Schurz, 2008b). Nessa versão, o argumento de Good prova que a condicionalização para as classes de referência mais estreitas maximiza o sucesso da previsão.

7.9 Tipos de classes de referência mais estreitas e calibração

O princípio da classe de referência mais estreita envolve outras sutilezas, as quais têm a ver com a definição das classes de referência mais estreitas.

7.9.1 Predicados nomológicos

Classes de referência devem ser formadas por atributos qualitativos ou nomológicos (ou seja, capazes de expressar regras gerais) e não por atributos extensionais (ou seja, que listam indivíduos específicos). O problema dos atributos nomológicos também surge no campo problema da indução: predicados definidos por meio de constantes individuais não são projetáveis indutivamente (Goodman, 1946; Carnap, 1947, p. 146; Seção 9.7). As probabilidades $P_t(Fa) = P(Fa|Ra \wedge H) = p(Fx|Rx) = 0,9$ atuam como previsões indutivas. Isso só faz sentido se a classe de referência Rx for adequada para previsão indutiva. Se alguém permitisse classes de referência extensionais arbitrárias, então, a classe $\{a\}$ ou o predicado $x=a$ seria a classe de referência mais estreita na qual a cai e, ademais, obter-se-ia o resultado trivial de que $P_t(Fa) = p(Fx|x=a)$ só pode assumir os valores 1

ou \circ . A condição nomológica também corrige a seguinte objeção de Wójcicki (1966): suponha que queiramos prever Fa (esse pássaro pode voar) e que Va (esse animal é um pássaro) seja nossa classe de referência “natural” mais estreita aplicável a a , com $p(Fx|Vx) = \circ,98$.

Gx (x tem asas quebradas) é outra classe que não se aplica a a e que dá uma probabilidade completamente diferente de voar (perto de zero). Contudo, a disjunção “ $(Gx \vee x=a)$ ” (“tem asas quebradas ou é idêntica a a ”) aplica-se a a . Assim, “ $(Gx \vee x=a) \wedge Vx$ ” seria uma classe de referência mais estreita para a do que Vx . A probabilidade de $p(Fx|Vx \wedge (Gx \vee x=a))$ é quase idêntica à de $p(Fx|Vx \wedge Gx)$, ou seja, também próxima de \circ , já que Gx e “ $Gx \vee x=a$ ” só podem distinguir um dos muitos indivíduos. Se o predicado “ $Vx \wedge (Gx \vee x=a)$ ” fosse aceito como a classe de referência mais estreita, isso levaria à previsão absurda da falta de voo da ave voável a , simplesmente porque cumpriria o predicado disjuntivo “tem asas quebradas ou é idêntico a a ”. O predicado disjuntivo definido por meio da constante individual “ a ” também é inadequado para indução e não é nomológico, pois o indivíduo a não é uma instância representativa desse predicado (para detalhes sobre a condição de nomologicidade, ver Schurz, 2006, Seção 6.5.1).

7.9.2 Classes de referência epistêmicas versus objetivamente mais estreitas

Na Def. 2-3 e na Def. 7-4, em concordância com Reichenbach, Carnap e Hempel, definimos epistemicamente o conceito de classe de referência mais estreita como a classe de referência mais estreita a qual o sujeito dado conhece, ou acredita com certeza que está em jogo nela. Salmon (1984, p. 37), por outro lado, introduz o conceito de classe

de referência objetivamente homogênea (já mencionada em 5.4 em “contingências objetivas”): a classe de referência objetivamente mais estreita de um evento $G(a,t)$, no tempo t , é a conjunção de todas as propriedades nomológicas $F_i(a,t')$ do indivíduo a a t anteriores vezes $t' \leq t$, as quais são necessárias para que $G(a,t)$ sejam estatística e causalmente relevantes. No entanto, a probabilidade estatística de um evento, em sua classe de referência objetivamente mais estreita, só pode ser diferente de 0 ou de 1 se houver *indeterminismo físico*, ou seja, se o evento for produzido por um processo aleatório genuíno (cf. Seção recapitulativa 5.4). Como as classes de referência objetivamente mais estreitas são geralmente desconhecidas para nós (nosso conhecimento não é suficiente para isso), o conceito epistêmico é mais importante na ciência aplicada do que o conceito objetivo.

7.9.3 Classes de referência relevantes mais estreitas

Na prática, é útil limitá-lo às classes de referência relevantes mais estreitas: essas últimas não precisam necessariamente capturar todas as informações conhecidas sobre o indivíduo a dado, mas apenas as informações das quais a característica preditiva G_x é probabilisticamente *dependente*. Se R_x é uma classe de referência relevante mais estreita de a para G_x , então, no sistema de fundo epistêmico, S é conhecido ou assumido como $p(G_x|R_x) = p(G_x|R_x \wedge R^*x)$ e se aplica a qualquer outra informação conhecida R^*a . Assim, a substituição das classes de referência mais estreitas pelas classes de referência relevantes mais estreitas não altera o valor de probabilidade resultante $P_i(Ga)$. Propostas nessa direção são o conceito de Hempel sobre a classe de *referência maximamente*

determinada (1965, p. 397) e o conceito de Salmon sobre a *classe de referência homogênea mais ampla* (1984, p. 37).

7.9.4 Classes de referência mais estreitas factuais versus informacionais

Existem dois tipos distintos de classes de referência epistemicamente mais estreitas. $R_{a,x}$ é chamada de classe de referência mais estreita factual de a , no sistema epistêmico de fundo S , sse $R_{a,x}$ consiste na conjunção de todos os predicados nomológicos $F_{i,x}$, de modo que $F_{i,a}$ seja acreditado e, portanto, incluído no sistema epistêmico S . Contudo, o StK só pode ser aplicado àquelas classes de referência mais estreitas factuais $R_{a,x}$ se estiver relacionado a um traço preditivo G_a , para o qual existe conhecimento estatístico em S sobre o valor $p(G_x|R_{a,x})$. Para trabalhar com classes de referência mais estreitas factuais, devemos ter conhecimento estatístico suficientemente extenso, ou devemos estimar as probabilidades estatísticas desconhecidas usando *probabilidades iniciais subjetivas* de acordo com o Teorema 7-2. Se, por outro lado, essas últimas devem ser evitadas devido ao seu viés subjetivo, as classes de referência informacionalmente mais estreitas devem ser preferidas: $F_a =_{\text{def}} F_{1,a} \wedge \dots \wedge F_{n,a}$ são conjunções de todos os predicados nomológicos $F_{i,x}$, de modo que não só F_a , mas também um Teorema estatístico da forma $p(G_x|F_x) = r$, está contido no sistema epistêmico S .

Além da vantagem de evitar probabilidades iniciais subjetivas, as classes de referência mais estreitas informacionais têm a desvantagem de as probabilidades de crença obtidas por elas não serem mais comprovadamente coerentes. Como dito na Seção 7.4, o Teorema da coerência (Teorema 7-1(a)) é transferido da probabilidade inicial P

para a função de crença real P_t , porque, para cada indivíduo, é usada *apenas exatamente uma* classe de referência (mais estreita factual), recapitular (7-2). Se, em vez disso, usarmos as classes de referência mais estreitas informacionais, então, poderemos não mais atribuir uma, mas sim várias classes de referência ao indivíduo a , dependendo da informação estatística que tivermos para qual característica preditiva. Suponhamos que $P(Fa)$ seja determinado por $p(Fx|R_a x) = 0,2$, mas que não haja informações sobre $p(Fx \wedge Gx | Rax)$, e que $Qx \supset R_a x$ seja a classe de referência conhecida mais estreita ($P(Qa) = P(R_a a) = 1$), da qual podemos obter o valor $p(Fx \wedge Gx | Qx) = 0,4$. Então, não devemos substituir $P(Fa \wedge Ga)$ por $p(Fx \wedge Gx | Qx)$, pois isso levaria à incoerente valoração subjetiva de probabilidade $P(Fa) = 0,2$ e $P(Fa \wedge Ga) = 0,4$; com base nos axiomas básicos, deve ser $P(Fa \wedge Ga) \leq P(Fa) = 0,2$. Uma sugestão plausível de Kyburg (1974, p. 222-226), para responder a esse problema, é incluir leis intervalares da forma " $p(Fx | Rx) \in [r_1, r_2]$ " – o caso limítrofe das leis estritas $p(Fx | Rx) \in [r, r]$ está incluso. Uma classe de referência Qx para a com $p(Fx | Qx) = q$ é considerada, segundo Kyburg (1974), a classe de referência informationalmente mais estreita para Fx apenas se, para toda classe de referência mais estreita conhecida Rx , juntamente com a informação estatística $p(Fx | Rx) \in [r, r]$, for verdade que $q \in [r_1, r_2]$. Em nosso exemplo, conhecemos $p(Fx \wedge Gx | R_a x) \in [0, 0,2]$ (isso decorre de $p(Fx | R_a x) = 0,2$); por causa de $0,4 \notin [0, 0,2]$, portanto, Qx é excluído como uma classe de referência para $Fx \wedge Gx$, e temos de nos contentar com a instrução intervalar $P(Fa \wedge Ga) \in [0, 0,2]$. Isso elimina inconsistências; em vez disso, obtemos apenas uma *avaliação de probabilidade subjetiva parcial (não totalmente determinada)*.

7.9.5 Classes de referência mais estreitas factuais e calibração

Existe uma correlação interessante entre as classes de referência mais estreitas factuais e o método de calibração. A calibração é um princípio de coordenação entre probabilidades estatísticas e probabilidades subjetivas, a qual não associa graus de crença às frequências dos eventos acreditados, mas sim às frequências de verdade das previsões sobre esses eventos. O requisito de calibração surge em conexão com previsões meteorológicas probabilísticas, tais como probabilidades de chuva para o dia seguinte. Por exemplo, considerando todos os dias para os quais um meteorologista calibrado previu chuva com $r\%$ de probabilidade, ele deve estar certo em $r\%$ dos casos. Brier (1950) levanta esse problema e desenvolve uma pontuação capaz de levar o meteorologista a fazer as previsões mais calibradas possíveis. É importante fazer o requisito de calibração para cada tipo de proposição singular (não probatória) separadamente e, ademais, não misturar proposições diferentes; caso contrário, consequências indesejáveis surgem (ver van Fraassen, 1983, p. 303; Lad, 1984). A seguir, a_i ($i \in \mathbb{N}$) são nomes padrão para uma área ordenada de indivíduos ($d_i; i \in \mathbb{N}$); c_j ($j \in \mathbb{N}$), por outro lado, são metavariáveis para constantes individuais arbitrárias.

(Def. 7-6) *Calibração:*

(a) Uma função de crença real P_i é chamada de *calibrada* em relação a um tipo de Teorema singular $S(x_1, \dots, x_n)$, relativo à função de probabilidade estatística p , para cada valor r , no intervalo $[0, 1]$ (aproximado a um certo número de dígitos) sse a probabilidade

estatística de todas as instâncias-S verdadeiras $S(c_1, \dots, c_n)$, entre todas as instâncias-S acreditadas no grau r ($P(S(c_1, \dots, c_n)) = r$), é r .

Formalmente: $p(\|S(x_1, \dots, x_n)\| \mid \{(d_1, \dots, d_n) \in D^n : P_t(S(a_1, \dots, a_n)) = r\}) = r$.⁵¹

(b) P_t é chamada perfeitamente calibrada sse P_t é calibrada para cada tipo de conjunto singular $S(x_1, \dots, x_n)$.

Van Fraassen (1983) prova que uma função de crença real, formada com a ajuda das classes de referência mais estreitas, com conhecimento completo, é perfeitamente calibrada no sentido da Def. 7-6. Na formação da função de crença real, faz-se referência a uma linguagem descritiva finita \mathcal{L} , na qual todas as experiências empíricas possíveis, sobre n indivíduos variáveis x_1, \dots, x_n , são expressas na forma de uma *partição de predicados de referência de n dígitos* ($R_i(x_1, \dots, x_n) : 1 \leq i \leq m$), consistindo em conjunções de fórmulas básicas. Para todos os Teoremas singulares previsíveis da forma $S(c_1, \dots, c_n)$, a probabilidade esperada $P_t(S(c_1, \dots, c_n))$ é formada por referência a essa partição, enquanto que a $S(x_1, \dots, x_n)$, no caso não trivial, não segue a $R_i(x_1, \dots, x_n)$, mas são formadas por outras fórmulas atômicas. Por exemplo, o $R_i(x_1, \dots, x_n)$ descreve o desenvolvimento do tempo nos últimos três dias e o $S(x_1, \dots, x_n)$, o desenvolvimento do tempo amanhã.

(Teorema 7-5) *Classes de referência mais estreitas e calibração:* Seja P_t uma função de crença real, formada de acordo com o princípio das classes de referência mais estreitas, por referência à partição dos predicados de referência de n dígitos $R =_{\text{def}} \{R_i(x_1, \dots, x_n), \dots$

⁵¹ Podemos expressar isso com fórmulas em vez de extensões de fórmulas por “ $p(S(x_1, \dots, x_n) \mid P_t(S(x_1, \dots, x_n)) = r) = r$ ” se explicarmos a satisfação de expressões P_t abertas por atribuições de variáveis da seguinte forma: $I[x_1; d_1, \dots, x_n; d_n](P_t(S(x_1, \dots, x_n)) = r) =_{\text{def}} I(P_t(S(a_1, \dots, a_n)) = r)$.

$R_m(x_1, \dots, x_n)$ }, para a qual existe conhecimento completo. Assim, para qualquer tipo de Teorema singular preditivo $S(x_1, \dots, x_n)$:

$$(*) P_t(S(c_1, \dots, c_n)) = p(S(x_1, \dots, x_n) | R_{c_1-n}(x_1, \dots, x_n)),$$

com $R_{c_1-n}(x_1, \dots, x_n)$ como o predicado de referência em R que se aplica a c_1, \dots, c_n .

Então P_t é perfeitamente calibrado.

O Apêndice 10.3.12 contém uma versão simplificada da prova de van Fraassen (essa última inclui graus complicados de aproximação).

8 Teste de hipóteses estatísticas

Na Seção 7.2, explicamos como as hipóteses estatísticas geram conteúdo indutivo-empírico com a ajuda do princípio de coordenação estatística (StK), na forma de resultados amostrais esperados ou de intervalos desses resultados, que são provenientes de frequências de amostragem estatísticas calculadas e que são transferidos para o caso individual com a ajuda do StK. Todos os métodos padrão de teoria estatística textual e inferencial, que são explicados abaixo, baseiam-se nesse fato. Focalizamos aqui generalizações estatísticas simples com variáveis *binárias* (dois valores) e explicamos a generalização para variáveis contínuas na Seção 8.6. Por uma questão de simplicidade, consideremos primeiro generalizações com apenas um fator antecedente, da seguinte forma:

(8-1) 80% de todas as árvores, ao longo de rodovias (=A), estão doentes (=D)

$$p(Dx|Ax) = 0,8$$

Seja o conjunto de árvores na Europa Central, entre 2000 e 2005, o domínio ou o conjunto de base (população), e assumindo que os predicados “em autoestradas” e “doentes” são operacionalizados com precisão suficiente. *Convenção:* a seguir, omitimos a variável individual “x” para simplificar, ou seja, escrevemos “A” para “Ax”, etc.

Exigimos duas coisas de uma hipótese dessa forma, $p(D|A) = r$:

(1.) *Verdade*: supõe-se que a hipótese seja aproximadamente verdadeira, ou seja, a frequência de D, dado A, deve realmente ter aproximadamente o valor r.

(2.) *Relevância* (ou *dependência*): A deve ser estatisticamente relevante para D (ou D deve depender estatisticamente de A), isto é, $p(D|A)$ deve ser diferente (significativamente) de $p(D)$, ou seja: $p(D|A) \neq p(D)$.

Uma relação estatística relevante, entre duas características A e D, é chamada de *correlação*.

A *medida mais simples de correlação*, para características qualitativas, é a diferença de probabilidade $Kor(D,A) =_{\text{def}} p(Dx|Ax) - p(Dx)$. A é chamada de *positivamente relevante* para D se $Kor(D,A) > 0$, de *negativamente relevante* para D se $Kor(D,A) < 0$ e de *irrelevante* se $Kor(D,A) = 0$. Há uma série de medidas de correlação relacionadas (ver Seção 8.6) – por exemplo, a *covariância* que, no caso binário, é definida como $Kov(D,A) =_{\text{def}} p(A \wedge D) - p(D) \cdot p(A)$.

Existem métodos empíricos para verificar tanto a verdade quanto a relevância de uma hipótese. John Stuart Mill chama de concordância o método para verificar a verdade (presumida) e de diferença o método para verificar a relevância (presumida) (Mill, 1865, Volume 2, Livro III; Losee, 1977, p. 14). Como esses métodos aparecem em uma roupagem estatística moderna, eles serão explicados abaixo.

8.1 Teste de verificação da verdade – o método dos intervalos de aceitação

Testamos a hipótese de lei $p(D|A) = 80\%$ para a verdade presumida e aproximada, formando uma amostra aleatória de

indivíduos que possuem a característica A . Chamamos essa amostra de *amostra A*. De acordo com o método de concordância, a frequência na amostra deve corresponder, o mais próximo possível, à frequência reivindicada pela hipótese para a população, a fim de considerar a hipótese como “confirmada” pelo resultado da amostra. Assim, selecionamos 100 árvores de faixas florestais selecionadas aleatoriamente ao longo de rodovias e examinamos se elas apresentam sintomas de doenças. Supondo que, de cada 100 árvores em nossa amostra, 75 estão doentes: isso deve ser visto como confirmação ou como rejeição de nossa hipótese da lei dos 80%? De forma mais geral, como inferir, a partir da frequência da amostra coletada $h_n(D:A)$, a plausibilidade da hipótese sobre a frequência total básica $p(D|A)$? *Notação:* a partir de agora, $h_n(D)$ e $h_n(D:A)$ sempre denotam a *frequência relativa* da característica D em uma amostra aleatória de n -elementos, ou em uma amostra aleatória de n -elementos de indivíduos A .

Não é de se esperar que a frequência da amostra corresponda exatamente à frequência da população – na verdade, isso é muito improvável devido aos desvios aleatórios dos resultados amostrais. Por isso, não existe uma falsificação estrita no caso de hipóteses estatísticas (desde que as amostras não esgotem a população total, o que aqui se pressupõe). A questão de saber se há confirmação ou se há enfraquecimento não pode ser decidida qualitativamente, mas requer um cálculo quantitativo. O procedimento estatístico padrão, para isso, é o método dos *intervalos de aceitação e rejeição*, que remonta a Fisher (1956) (Hays; Winkler, 1970, p. 380; Bortz, 1985, p. 141 e seg.; Howson; Urbach, 1996, p. 171 e seg.). Aqui, calcula-se a probabilidade estatística de que a frequência de amostragem $h_n(D:A)$, de uma amostra de n -elementos, esteja dentro de um determinado intervalo de tamanho,

dado que a hipótese sobre a população, $p(D|A)=80\%$, é verdade. Esse cálculo baseia-se, no caso discreto, na distribuição binomial (Seção 3.2) e, no caso contínuo, na distribuição normal Gaussiana (Seção 8.6). Em nosso exemplo, calculamos: assumindo que 80% de todas as A s na população são D s, há 95% de probabilidade de que a frequência (absoluta) de D , em uma amostra de 100 elementos A , esteja na faixa entre 72 e 88. O valor de probabilidade de 95%, pragmaticamente escolhido, também é chamado de coeficiente de *aceitação*, e o seu valor complementar de 5% é chamado de coeficiente de significância. O intervalo entre 72 e 88 de 100 é o *intervalo de aceitação* para a hipótese dada (com um coeficiente de aceitação correspondente a 95%). A Figura 8-1 ilustra esquematicamente a distribuição de probabilidade das frequências de amostras de 100 elementos retiradas de uma população de indivíduos A , com uma probabilidade condicional $p(K|A)$ igual a 0,8, e com o intervalo de aceitação destacado (calculado utilizando MATLAB).

A frequência de amostragem mais provável coincide com a frequência na população. À esquerda e à direita dele, a probabilidade esperada cai quase simetricamente. (A simetria exata só seria dada a uma frequência populacional de 50%). Se aproximarmos os valores de frequência possíveis discretos 0, 1/100, ..., 99/100, 1 a uma variável contínua, a área total sob a curva de distribuição será 1, e a área em um dado intervalo sob a curva corresponderá à probabilidade de encontrar a amostra nesse intervalo. A área sob a curva, no intervalo de aceitação entre 72 e 88 de 100, é exatamente 95% da área total. Em ambos os lados fora do intervalo de aceitação, o intervalo de rejeição é de 5% da área total.

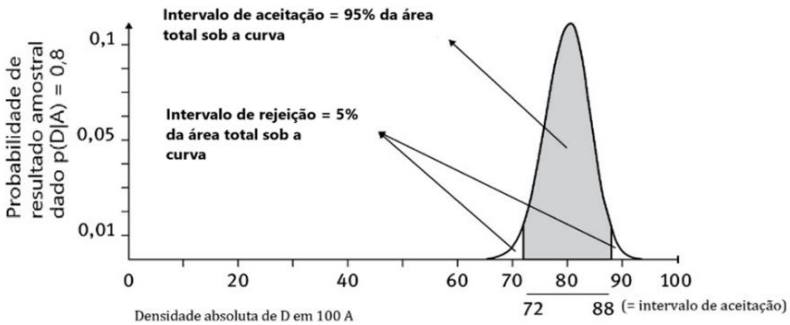


Figura 8-1: Intervalo de aceitação para $p(D|A) = 0,8$.

Naturalmente, as probabilidades estatísticas dos resultados da amostra só dizem algo sobre as tendências de amostragem no longo prazo (infinitamente). Com a ajuda do StK, as tendências de frequência dos resultados amostrais são transferidas para a amostra específica $s =_{\text{def}} \{a_1, \dots, a_{100}\}$ como sua probabilidade de expectativa epistêmica: $P = 95\%$ é a frequência na amostra A específica s, abreviada como $h(D:A|\{a_1, \dots, a_{100}\})$, na faixa de 72 a 88 em 100, dado que a hipótese populacional é verdadeira. Formalmente:

$$(8-2) P(h(D:A|\{a_1, \dots, a_{100}\}) \in [0,72, 0,88] \mid p(D|A) = 0,8 \wedge Aa_1 \wedge \dots \wedge Aa_{100}) = 95\%.$$

Essa etapa de transferência geralmente não é explicitada na teoria do texto estatístico, mas aparece apenas implicitamente: é, no entanto, fundamental para a fundamentação filosófica da teoria dos testes estatísticos. Isso porque, se as tendências de frequência de longo prazo dos resultados da amostra não fossem transferidas para o caso individual, o método de aceitação versus rejeição seria completamente injustificado.

Se a frequência de amostragem observada estiver dentro do intervalo de aceitação da hipótese, a hipótese é considerada *fracamente confirmada e mantida*. Se, por outro lado, estiver fora do intervalo de aceitação, ou seja, no intervalo de rejeição, então, a hipótese é *considerada severamente enfraquecida* ou *rejeitada*. Em nosso exemplo, o resultado da amostra é 75 de 100 no intervalo de aceitação; vemos a hipótese $p(D|A)=0,8$, com base no resultado da amostra $h_{100}(D|A) = 75$, como ainda aceita (com coeficiente de aceitação de 95%). Se o resultado tivesse sido 70 em 100, teríamos rejeitado a hipótese pelo coeficiente de significância de 5%. A definição geral do intervalo de aceitação é a seguinte:

(Def. 8-1) O *intervalo de aceitação* é o intervalo dos resultados amostrais mais prováveis, em que a frequência de amostragem tem probabilidade igual ao coeficiente de aceitação (geralmente 95%), dado que a hipótese legal a ser testada é verdadeira.

O nível do coeficiente de aceitação de 95% é pragmático, mas não é arbitrário. Se o coeficiente de aceitação for muito alto, por exemplo, 99,5%, então, o intervalo de aceitação torna-se muito amplo e, assim, poucas hipóteses são rejeitadas. Se for escolhido um muito pequeno, por exemplo, 50%, o intervalo de rejeição torna-se muito largo e, assim, o enfraquecimento não será forte o suficiente no caso negativo.

Se a distribuição binomial for aproximada pela distribuição normal, o que é legítimo para tamanhos de amostra > 30 , os intervalos de aceitação são dados pelas fórmulas em (8-3) (coluna do meio). Onde $\sigma_s = \frac{\sigma}{\sqrt{n}}$ é o desvio padrão das *frequências de amostragem*, que são (comprovadamente) calculadas como o desvio padrão σ da variável

em questão dividida pela raiz do tamanho da amostra n ; onde σ , para variáveis binárias, é dada como $\sigma = \sqrt{p \cdot (1-p)}$ com $p =_{\text{def}} p(D|A)$. As fórmulas da coluna do meio de (8-3) podem ser obtidas consultando-se os intervalos de 95% da chamada *distribuição z* (uma distribuição normal padronizada com um valor médio de zero e com um desvio padrão de um) nas tabelas e multiplicando-os pelo desvio padrão σ estimada a partir da amostra dividida por \sqrt{n} (ver Seção 8.6).

(8-3) Coef.de aceitação:	Intervalo de	Exemplo para $p=0,8, n=100$
	aceitação	
99,5%	$p \pm 2,8 \cdot \sigma_s$	[0,69 , 0,91]
95%	$p \pm 1,96 \cdot \sigma_s$	[0,72 , 0,88]
70%	$p \pm 1,03 \cdot \sigma_s$	[0,76 , 0,84]

Se o coeficiente de aceitação for fixo, o intervalo de aceitação será inversamente proporcional à raiz quadrada do tamanho da amostra. Para tamanhos de amostra cada vez maiores, o intervalo de aceitação tornar-se-á cada vez mais apertado, e nossas projeções prováveis de 95%, para o valor de amostra esperado, ficarão mais rigorosas. Ao mesmo tempo, isso resulta em uma lei de retornos decrescentes: um aumento de quatro vezes no tamanho da amostra reduz apenas pela metade o intervalo de aceitação, e assim por diante. Aqui, estão alguns intervalos de aceitação de 95% para diferentes tamanhos de amostra:

(8-4) Intervalos de aceitação para $p = 0,8$ (coeficiente de aceitação = 0,95) para variações de n :

$n = 1$: [0 , 1]	$n = 50$: [0,69 , 0,91]	$n = 1600$: [0,78 , 0,82]
$n = 10$: [0,56 , 1]	$n = 100$: [0,72 , 0,88]	$n = 10.000$: [0,79 , 0,81]
$n = 20$: [0,63 , 0,97]	$n = 400$: [0,76 , 0,84]	

No caso de estudos únicos em que a amostra não precisa ser dividida, os tamanhos de amostra maiores que 40 são habitualmente chamados de grandes, e os menores que 30 de pequenos. Em geral, os tamanhos das amostras não devem ser inferiores a 15 ou a 20. Uma restrição adicional no tamanho das amostras surge quando a probabilidade hipotética da população está próxima de 0 ou de 1. Os tamanhos amostrais devem ser escolhidos tão grandes que os limites do intervalo de aceitação realmente fiquem entre 0 e 1 (no exemplo acima, esse é o caso apenas para $n \geq 20$). Observemos que o tamanho da população total é irrelevante para questões de tamanho de amostra; a única suposição é a de que seja significativamente (pelo menos 100 vezes) maior que o tamanho da amostra (Bortz, 1985, p. 112).

8.2 Encontrando hipóteses estatísticas e intervalos de confiança

A hipótese $p(D|A)=0,8$ é apenas fracamente confirmada pelo resultado da amostra $h_{100}(D:A)=0,75$. Isso ocorre porque todas as hipóteses alternativas estatísticas, que têm um valor de $p(D|A)$ no intervalo $0,75 \pm 0,08$, são igual e fracamente confirmadas pelo resultado da amostra $h_{100}(D:A) = 0,75$, ou seriam *mantidas* se estivessem em revisão. Para todas essas hipóteses, o resultado da amostra encontra-se dentro do intervalo de probabilidade de aceitação previsto de 95%. O intervalo de todas as probabilidades populacionais hipotéticas, para as quais o resultado da amostra está apenas dentro do intervalo de aceitação de 95%, é chamado de

intervalo de confiança da probabilidade hipotética da população (em vez do coeficiente de aceitação, falamos agora do *coeficiente* de confiança de 95%).

O método dos intervalos de confiança remonta a Fisher e

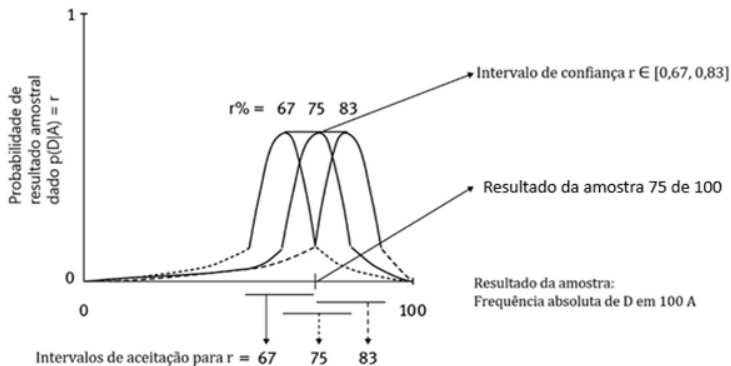


Figura 8-2: Relação entre aceitação e intervalos de confiança.

Neyman (ver Stegmüller, 1973c, p. 189; Bortz, 1985, p. 132). Se h é um resultado de amostra e se r é uma probabilidade populacional hipotética, então, por razões matemáticas, a seguinte relação simétrica entre confiança e intervalo de aceitação mantém-se:

$$h \in [r-a, r+a] \text{ sse } r \in [h-a, h+a].$$

Assim, obteremos o intervalo de confiança invertendo, simetricamente, o intervalo de aceitação por h , em vez de r . A Figura 8-2 ilustra essa conexão.

O intervalo de confiança corresponde à hipótese do intervalo estatístico $p(D|A) \in [h-a, h+a]$, que no nosso exemplo diz: *entre 67 e 83% de todas as árvores nas estradas estão doentes*. Essa hipótese do intervalo de confiança é fortemente confirmada pelo resultado da amostra $h_n(D:A)=75\%$ (com um coeficiente de confiança de 95%). Se

alguém estiver interessado em prever intervalos de confiança estritos, a amostra deve ser escolhida de modo adequadamente grande (ver Bortz, 1985, p. 138).

8.3 Teste de verificação da relevância – o método das diferenças significativas

Para verificar se a característica A também é relevante para D, compara-se a abundância de D em uma amostra A (o grupo da característica [ser árvore perto de rodovias]) com a frequência de D em uma amostra controle A (o grupo controle). No caso mais simples, a amostra A-controle consiste em um conjunto de indivíduos que *não* possuem o traço A (também poderia consistir em uma amostra aleatória, que contém uma proporção aleatória de indivíduos-A). No nosso exemplo, de cada 100 árvores nas rodovias, 75 estão doentes. Agora, formamos uma amostra de controle A com 100 árvores de faixas florestais *que não* estão perto de rodovias e descobrimos que apenas 60 árvores estão doentes. Isso significa que a proximidade de rodovias aumenta a probabilidade de as árvores ficarem doentes, ou que o desvio entre a amostra A e a amostra de controle A, que é de 15 em 100, pode ser apenas coincidência? Novamente, essa é uma questão *quantitativa* e, como dito acima, um método intervalar é usado. Com base na distribuição de probabilidade das amostras aleatórias, é possível calcular a probabilidade estatística de que o desvio entre a amostra A e a amostra controle A tenha sido puramente coincidente – ou seja, de que esse desvio tenha ocorrido sob a suposição de não haver relação estatística entre A e D na população: $p(D|A) = p(D|-A)$.

Essa hipótese de irrelevância também é chamada de *hipótese nula*. A *hipótese alternativa* a isso é a da relevância, a qual afirma existir uma relação estatística entre A e D na população: $p(D|A) \neq p(D|\neg A)$.

A distribuição de probabilidade das diferenças de frequência entre duas amostras (n_1, n_2) da mesma população assume a forma de uma distribuição binomial com média \circ e desvio padrão $\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ (ver Seção 8.6). Calcula-se o intervalo simétrico de 95% das diferenças amostrais positivas ou negativas mais prováveis. A quantidade absoluta da diferença de frequência máxima, que está dentro do intervalo de 95%, é chamada de *diferença amostral significativa*, e o coeficiente de 5% é chamado de coeficiente de *significância*.

Em outras palavras, o intervalo de 95% das diferenças amostrais mais prováveis funciona como o intervalo de aceitação da hipótese de irrelevância (hipótese nula), e o intervalo bicaudal-extremo de 5% das diferenças amostrais mais improváveis atua como um intervalo de rejeição da hipótese de irrelevância e como um intervalo de aceitação da hipótese de relevância (hipótese alternativa). Ver Fig. 8-3.

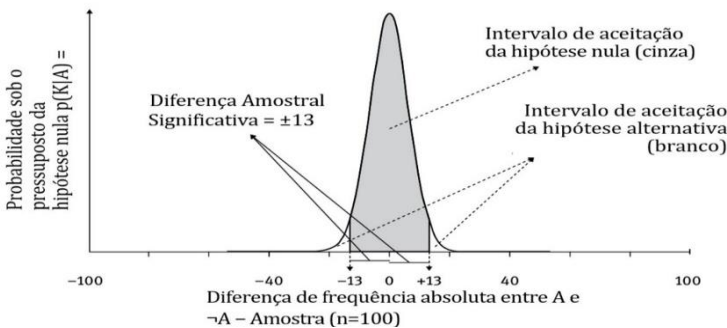
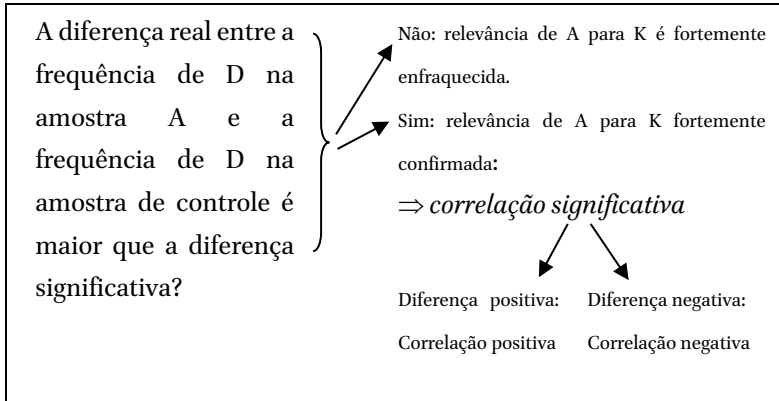


Figura 8-3: Distribuição probabilística das diferenças amostrais e diferença amostral significativa (aproximada pela distribuição normal).

(Def. 8-2) A *diferença amostral significativa* é a quantidade que a diferença entre a frequência de D em uma amostra A e em uma amostra controle A excede com uma probabilidade igual ao coeficiente de significância (geralmente 5%), dado que não há relação estatística entre A e D na população (ou seja, a diferença é puramente coincidência).

Se a diferença amostral real encontrada exceder a diferença amostral significativa, a hipótese de irrelevância será rejeitada e a hipótese de relevância será aceita. Nesse caso, supondo a hipótese de irrelevância, a probabilidade de se encontrar uma diferença amostral pelo menos tão grande quanto a efetivamente encontrada seria menor que o coeficiente de significância de 5%; a hipótese da irrelevância estaria, assim, muito enfraquecida, e a hipótese da relevância seria fortemente confirmada. Desse modo, diz-se que existe uma relação, bem como uma correlação significativa, entre A e D. Em nosso exemplo, é calculada uma diferença amostral significativa de 13 em 100 (assumindo uma distribuição aproximadamente normal). A diferença encontrada de 15 em 100 é, portanto, significativa. Se A conduz a um aumento na frequência D, a correlação considerada significativa é positiva. No caso da redução, essa correlação é negativa. Se, por outro lado, a diferença amostral encontrada for menor que a diferença significativa, a hipótese de irrelevância ainda será aceita.

O procedimento também poderá ser realizado se os coeficientes de significância forem escolhidos de forma diferente. Uma diferença amostral significativa, com um coeficiente de significância de 1%, é chamada de *altamente significativa*. Na maioria dos casos, o coeficiente de significância, no qual a diferença é *apenas significativa*, é especificado. Em nosso exemplo, a diferença de 0,15 encontrada em



No teste de relevância, a hipótese alternativa é a negação da hipótese nula. Portanto, a primeira é fortemente confirmada justamente quando a segunda está muito enfraquecida. Isso é diferente do teste de verdade, onde a hipótese dada $p(D|A)=r$ tem infinitamente muitas hipóteses alternativas da forma $p(D|A) = r^* \neq r$. O teste de relevância, mostrado acima, também é chamado de *teste bilateral*, porque diferenças positivas e negativas são consideradas. Se soubermos desde o início que o fator A só pode afetar um lado, se é que pode afetar, aplicamos o *teste de diferença unilateral*, no qual o intervalo de 5% extremo unilateral é escolhido como o intervalo de rejeição.

Tal como acontece com o método do intervalo de aceitação, o desvio padrão das diferenças relativas de frequência da amostra e, portanto, a diferença significativa diminui proporcionalmente à raiz quadrada do tamanho da amostra (n). Qualquer diferença relativa da amostra, por menor que seja, torna-se significativa se o tamanho da amostra for suficientemente grande. Alguns autores entendem esses fatos como paradoxais, mas esse não é o caso – apenas expressam a lei dos grandes números. Note-se também que esse fato só se aplica a

diferenças amostrais relativas, mas não absolutas. Essas últimas aumentam proporcionalmente à raiz do tamanho da amostra. Por exemplo, uma diferença de 1% de frequência de amostragem torna-se 5% significativa para um tamanho de amostra de $n = 14.390$ indivíduos, mas esse 1% ainda representa 144 indivíduos.

A mera afirmação de que é encontrada uma correlação significativa entre dois traços A e D seria, portanto, uma afirmação comparativamente *fraca*, sem *informações sobre o tamanho da amostra*. A afirmação simplesmente diz que há alguma correlação entre A e D, que pode ser muito baixa. Em particular, uma correlação altamente significativa não significa que o nível dessa correlação seja particularmente alto, mas apenas que há uma probabilidade muito *alta* de alguma, embora muito pequena, relação estatística entre A e D.

Isso é muitas vezes confundido nas representações populares de resultados estatísticos. Se, por exemplo, é relatado que os médicos observam uma correlação significativa entre o *consumo de salsicha extra* e a *taxa de câncer* – para citar um dos muitos exemplos possíveis –, isso será tomado como um resultado sensacional, sem considerar que essa correlação também poderia ser praticamente insignificante. Portanto, é muito importante, além da existência de uma correlação significativa, fornecer informações sobre o provável *nível* dessa correlação. Nesse sentido, uma medida adequada seria a suposta diferença de abundância na população, que é estimada a partir da diferença amostral encontrada. Se dividirmos essa diferença de frequência pelo desvio padrão σ , obteremos o chamado *força do efeito*. A força do efeito é uma medida estatística de correlação preferencialmente utilizada nos últimos tempos, que pressupõe uma variável antecedente binária A, $\neg A$ e um atributo consequente

escalonado arbitrariamente. É definida como a diferença entre a média de K na população A e na população geral, dividida pelo desvio padrão de K . Devido à sua independência de escala, a força do efeito é frequentemente usada em meta-análises, nas quais os resultados de diferentes estudos empíricos são combinados (ver Bortz; Döring, 2002, Seção 9.4).

Alguns autores recomendam a seguinte abordagem: um certo grau de diferença amostral é assumido como *praticamente significativo*, o tamanho da amostra é determinado com base nisso e essa diferença praticamente significativa seria apenas 5% significativa (ver Bortz, 1985, p. 157; Westermann; Hager, 1982). Essa abordagem é útil somente em casos especiais e é geralmente problemática. A diferença de amostragem que é significativa na prática depende de considerações práticas de custo-benefício. Se considerarmos apenas uma diferença amostral bastante grande (algo em torno de 50%, como sendo de importância prática, o que seria necessário para o sucesso de uma vacina, por exemplo), então, conforme esse procedimento, teríamos de escolher um tamanho de amostra de apenas 6 pessoas de teste, o que seria absurdo, porque amostras tão pequenas não são confiáveis.

Para chegarmos a julgamentos confiáveis, devemos escolher uma amostra suficientemente grande, *independentemente* do tamanho dessa correlação. A magnitude dessa correlação é estimada pela diferença amostral encontrada ou pela força do efeito calculado a partir dela. O mais informativo é a adição de um intervalo de confiança de 95% para o tamanho de efeito estimado (Bortz, 1985, p. 234; Westermann; Hager, 1982, p. 17). No caso qualitativo, o intervalo de confiança das diferenças amostrais é simplesmente calculado como a diferença amostral encontrada \pm a diferença amostral significativa;

em nosso exemplo $(75-55) \pm 18 = [2, 38]$. Como a diferença encontrada é apenas um pouco maior do que a diferença significativa, o intervalo de confiança fornece uma alta margem de incerteza.

O teste de relevância estatística é baseado em hipóteses legais com vários fatores antecedentes conjuntivamente $p(D|A_1 \wedge \dots \wedge A_n)$, formando uma amostra de controle A_i para cada fator antecedente A_i , composta por indivíduos que satisfazem todas as características antecedentes A_j , $j \neq i$, mas que não satisfazem A_i . A diferença de frequência entre a amostra A e a amostra controle A_i é, então, testada para diferença significativa. A diferença de frequência $h(D|A_1 \wedge \dots \wedge A_n) - h(D|A_1 \wedge \dots \wedge A_{i-1} \wedge \neg A_i \wedge A_{i+1} \wedge \dots \wedge A_n)$, na amostra, é uma medida da chamada *correlação condicional* entre D e A para as variáveis residuais retidas A_j ($j \neq i$).

8.4 Representatividade estatística e tipos de hipóteses estatísticas

Para o teste de hipóteses estatísticas comuns, o *requisito de representatividade da amostra significa* que todos os outros fatores relevantes para o predicado consequente D devem ser distribuídos da forma mais igualitária possível na amostra A , bem como na população A (cf. Bortz, 1985, p. 113). No nosso exemplo, esses seriam fatores causais, além dos gases de escape dos carros, os gases de escape industriais ou a infestação de pragas, os quais adoecem as árvores.

No caso da representatividade estatística, é importante distinguir entre *definição* e *critério*. Por *definição*, uma amostra é representativa se todas as características relevantes estiverem igualmente distribuídas nela como na população. Essa suposição é

baseada em uma *inferência indutiva* e não pode ser garantida por nenhum método. Seria *circular* considerar a representatividade de tal amostra como pré-requisito para inferências indutivas a partir dessa mesma amostra, uma vez que ela própria é resultado de generalização indutiva. Pelo contrário, são os *critérios* de representatividade que são decisivos, cujo cumprimento pode ser assegurado independentemente da etapa de generalização indutiva: apenas esses critérios podem ser considerados pré-requisitos para inferências indutivas (cf. também Campbell; Franklin, 2004, p. 84).

Os critérios de representatividade derivam dos *métodos utilizados* para produzir amostras tão representativas quanto possível. O método mais importante é o de *amostragem aleatória* – esse método é sempre recomendado quando pouco ou nada se sabe sobre a distribuição das características restantes na população. Uma amostra é uma amostra aleatória, *em sentido estrito*, quando há um procedimento de seleção aleatória que dá a cada indivíduo da população uma chance igual de ser colocado na amostra. É claro que as amostras aleatórias podem diferir *aleatoriamente* da população, mas a distribuição de probabilidade de seu desvio aleatório é estatisticamente calculável, e essa é a base dos métodos estatísticos explicados.

Um processo de seleção aleatória pressupõe que todos os indivíduos da população estejam acessíveis ao processo de seleção e, portanto, sejam registrados de alguma forma – por exemplo, por fichas indicadoras ou por listas de nomes a partir das quais a seleção é feita às cegas. Essa *definição estrita* de amostragem aleatória, frequentemente encontrada, é *desnecessariamente* estrita e *demasiadamente* estrita. Desnecessariamente estrita, porque a única coisa importante é que todos os tipos de indivíduos relevantes para o

predicado consequencial *têm a mesma chance de entrar na amostra: se esse for o caso, então, falamos de uma amostra aleatória* em sentido amplo. O procedimento de seleção não deve, portanto, distorcer as distribuições características *relevantes* (ver Mayntz *et al.* 1974, p. 69).

A definição restrita é muito estreita, pois geralmente não é viável. No nosso exemplo, é difícil numerar todas as árvores da Europa Central próximas às rodovias para, em seguida, sortear 100 números de uma grande urna. O essencial é que as áreas florestais de onde se retira a amostra aleatória de árvores a serem testadas para doenças não contenham características distorcidas (Kromrey, 2002, p. 292). Por exemplo, a infestação de pragas nessas áreas florestais não deve ser maior ou menor do que a infestação média geral de pragas (e assim por diante).

A noção de amostragem aleatória, em sentido amplo, também responde a outra objeção ao teste de hipóteses estatísticas sobre populações infinitas, apresentada por Spielman (1977), a saber: não se pode escolher amostras aleatórias de populações infinitas, porque cada operação de seleção deve ser limitada a uma parte finita da população. Esse argumento não é aplicável ao conceito amplo de amostragem aleatória. Suponhamos um experimento de lançamento de moeda no qual a hipótese $p = 1/2$ deve ser testada. A população aqui seria a sequência infinita e idealizada de todos os lances de moedas possíveis. Se eu testar essa hipótese populacional *hoje* com uma amostra finita de lançamentos, apenas lançamentos de moedas no presente ou nos futuros próximos terão chance de entrar na minha amostra. No entanto, o mero *momento* do lançamento da moeda é uma característica estatisticamente irrelevante e, portanto, ainda é uma amostra aleatória no sentido amplo.

Se a distribuição do restante dos traços relevantes na população for conhecida, pode-se formar uma amostra chamada estratificada, em vez de uma amostra aleatória para alcançar representatividade. Se, por exemplo, quisermos examinar os hábitos de consumo dos cidadãos alemães médios, podemos supor que as características da população urbana versus as da população rural, relativas às idades, aos tamanhos das famílias e aos gêneros, influenciam os comportamentos dos consumidores; assim, recomendaremos uma amostra *proporcionalmente estratificada*, na qual a porcentagem de pessoas testadas será colocada na amostra para cada característica relevante, que corresponde à frequência conhecida no estudo básico (cf. Mayntz *et al.*, 1974, p. 87 e seg.).

O conceito de amostragem estratificada pressupõe conhecimentos estatísticos já estabelecidos em outros lugares. Isso também se aplica ao conceito de amostragem aleatória em *sentido amplo*, o qual pressupõe a irrelevância estatística de algumas características residuais para a característica consequente D. Anteriormente, argumentamos que os critérios de amostragem não devem envolver circularidade. Não há aqui uma circularidade direta, pois se trata da irrelevância de *outras* características. Contudo, não há pelo menos uma circularidade indireta ou uma regressão infinita, na medida em que a irrelevância dessas outras características também deve ser confirmada, para as quais são necessárias amostras representativas?

Com efeito, o caso circular pode surgir quando o exame da relação entre D e A pressupõe a irrelevância de B e quando o exame da relação entre D e B pressupõe a irrelevância de A. Nesse caso, o método de amostragem não funciona em sentido amplo e, por isso, deve-se usar a amostragem em sentido estrito, ou seja, variando as variáveis A

e B. Muitas vezes, no entanto, não apenas propomos uma hipótese de irrelevância em relação a uma determinada população, mas também entendemos a irrelevância no sentido *causal* e, portanto, a afirmamos para distribuições arbitrárias de frequência das variáveis residuais, desde que apenas sejam preenchidas as condições capazes de definir causalmente o experimento aleatório. O teste de tais hipóteses causais e abrangentes de irrelevância não requer amostras representativas no sentido populacional definido acima, porém requer que as variáveis residuais *variem o máximo possível*, o que pode ser verificado sem uma etapa de generalização indutiva.

Assim, no exemplo do experimento de lançamento de moeda, afirmamos que o atributo 'momento' é irrelevante para quaisquer situações de lançamento de moeda, desde que sejam atendidas as condições físicas de um lançamento regular: velocidade inicial permitida e altura de queda, pressão e temperatura atmosféricas normais, e nenhuma outra força significativa além da gravidade. Ademais, não importa a “população” da qual vem o lançamento da moeda, se eu ou você a jogamos, se na Alemanha ou no Polo Norte, há 1000 anos, ou agora, etc. Para testar a completa irrelevância de uma dessas características, são necessárias apenas amostras arbitrárias que sejam tão amplamente dispersas quanto possível, mas não amostras aleatórias de base populacional.

Nessa discussão, tocamos no tema da robustez ou da *invariância* de hipóteses versus as variações nos valores de variáveis residuais (cf. Woodward, 2003, p. 250) – uma questão que também desempenha um papel nas discussões sobre o conceito de *legalidade* (“leis consistentes”) (ver Seção 7.8.2 e Seção 9.7) e sobre as chamadas condições *ceteris paribus* (cf. Reutlinger *et al.*, 2011; Schurz, 2002). Vamos considerar os seguintes exemplos de hipóteses estatísticas:

(8-5) 50% de todos os átomos de céscio-137 (de qualquer substância) decaem após 30 anos.

(8-6) O limite de frequência a que uma moeda regularmente atirada cai “cara” é de $1/2$.

(8-7) 80% de todos os pacientes com câncer de pulmão eram fumantes pesados (população: Alemanha 1980-1990).

(8-8) 80% de todas as crianças com transtornos de comportamento têm pais com distúrbios comportamentais.

(8-9) A incidência de tabagismo pesado em pacientes com câncer de pulmão é significativamente maior do que em outras pessoas.

(8-10) A incidência de crianças com transtornos comportamentais é significativamente maior em crianças que tem pais com transtornos comportamentais do que em outras crianças.

Os casos de (8-5,6,7,8) são hipóteses *estatísticas numéricas*, os casos de (8-9,10) são hipóteses meramente comparativas, ou seja, afirmações de diferença significativa de frequência. A hipótese (8-5) é uma lei física de um processo objetivamente indeterminista que se aplica a valores *arbitrários* de variáveis residuais (ou condições ambientais). Estamos falando de uma hipótese estatística *estritamente invariante*. No caso (8-6), há uma hipótese que vale para (não todos, mas pelo menos) *quase todos os* valores das variáveis residuais, a hipótese é *altamente invariante*. O exemplo (8-7), por outro lado, é uma hipótese que só faz sentido se estiver relacionada a uma *população específica* (como indicado entre parênteses). Além do tabagismo, existem outras causas de câncer de pulmão, por exemplo, a poluição média, as quais também são decisivas *para o* valor numérico de incidência; o mesmo aplica-se à hipótese (8-8). Aqui, estamos falando de uma hipótese limitada pela população. Aparentemente, o requisito habitual de representatividade, tal como o caracterizamos no início e

o encontramos nos manuais de estatística, aplica-se apenas a hipóteses limitadas pela população, mas não a hipóteses que reivindicam invariância universal ou de alto nível. Para testar essas últimas hipóteses, são imprescindíveis amostras cujas variáveis residuais variem tanto quanto possível.

Um método para passar das hipóteses estatísticas limitadas à população para as hipóteses estatísticas altamente invariantes é enfraquecer a afirmação feita numericamente para a meramente comparativa, como é feito nos exemplos (8-9,10). No caso de (8-9) – a atenuação de (8-7) –, isso conduz a uma hipótese universal – ou pelo menos altamente invariante –, porque o tabagismo pesado, independentemente dos outros fatores, aumenta a taxa de câncer de pulmão em qualquer caso (embora em um grau contextual). Schurz (2002) fala aqui de uma *hipótese comparativa ceteris paribus*.

No exemplo (8-10), a atenuação comparativa de (8-8), a invariância não é de modo algum tão clara; ao contrário, essa hipótese comparativa parece aplicar-se apenas a *certas* distribuições dos valores das variáveis residuais. Por exemplo, a hipótese é inválida para crianças que não são predominantemente criadas por seus pais; nesse sentido, Schurz (2002) fala de uma hipótese *exclusiva ceteris paribus*. Finalmente, o grau de *invariância* de uma hipótese estatística também determina sua pretensão de *regularidade*: por exemplo, consideraríamos (8-5), (8-6) e (8-9) leis consistentes; por outro lado, (8-7), (8-8) e (8-10) não são mais leis consistentes, pois dependem fortemente de circunstâncias contingentes.

8.5 Estatística de teste e estatística de inferência

Os intervalos de aceitação são *intervalos amostrais*, enquanto os intervalos de confiança são *intervalos de hipótese* (Hays; Winkler, 1970, p. 383). Os intervalos de aceitação fazem parte da chamada estatística de *teste*, a qual se preocupa com a verificação de *determinadas* hipóteses com uma plausibilidade já existente. Os intervalos de confiança, por outro lado, fazem parte da *estatística inferencial*, que se preocupa em *encontrar* as hipóteses mais plausíveis à luz de um dado resultado amostral (cf. Aron; Aron, 2002, p. 238).⁵² Na prática estatística, os dois problemas, de maneira geral, não são claramente separáveis: mesmo que já tenhamos certas hipóteses submetidas a um teste de intervalo de aceitação, ainda estaremos interessados em hipóteses que são tão bem validadas quanto possível e, portanto, em seus intervalos de confiança. Tanto a estatística de teste quanto os procedimentos estatísticos inferenciais estão sujeitos a inferências indutivas. Contudo, a diferença é a seguinte: na estatística de teste puro, apenas o princípio de indução epistêmica é usado, segundo o qual as hipóteses até então bem-sucedidas (bem-sucedidas no sentido do método do intervalo de aceitação) são mantidas, e apenas as hipóteses malsucedidas são rejeitadas. A estatística inferencial, por outro lado, é um procedimento de indução *metódico* que, dado um

⁵² A terminologia nem sempre é consistente; alguns autores chamam os intervalos de aceitação de “intervalos de confiança para resultados de amostra” (por exemplo, Lauth/Sareiter 2002, p. 276).

resultado amostral no sentido do método do intervalo de confiança, encontra o intervalo de 95% de todas as hipóteses mais plausíveis.⁵³

A teoria do teste de Fisher é, por vezes, *referida como quase-falsificacionista*, porque fornece regras metodológicas que nos dizem: em quais resultados amostrais uma dada hipótese deve ser mantida e em quais deve ser rejeitada (Howson, Urbach, 1996, p. 174). A diferença para uma *verdadeira* falsificação é que a rejeição da hipótese só é válida com uma certa probabilidade e, portanto, é fundamentalmente provisória. Por conseguinte, na minha opinião, é errado considerar tal procedimento como uma variante da falsificação de Popper. Esse último é sugerido por Popper (1935, 2002, Cap. II.68), Max Albert (1992), Gillies (2002, p. 148) e outros; tais autores entendem o método de Fisher como a “regra de falsificação metodológica” e Worrall (2006, p. 132) chama de “quasidedução” a conclusão de especialização indutiva. Mas, conforme explicado, a última conclusão é tão indutiva e incerta quanto a previsão indutiva e a conclusão da generalização (ver Seção 4.4). Popper (*ibidem*) propôs incorporar a teoria dos testes estatísticos em sua abordagem falsificacionista, considerando resultados amostrais extremamente improváveis como virtualmente impossíveis. Entretanto, como argumentam Howson e Urbach (1996, p. 174), eventos extremamente improváveis ocorrem repetidas vezes, como a distribuição exata de alelos num recém-nascido. A proposta de Popper, portanto, não parece viável.

Stegmüller critica a teoria dos testes estatísticos por restringir as decisões epistêmicas à aceitação versus rejeição (1973b, p. 142 e seg.). Essa crítica seria séria se fosse verdadeira; no entanto, não se aplica a muitas interpretações da teoria dos testes estatísticos. Hays e Winkler

⁵³ Para a diferença entre indução metódica, lógica e epistêmica, veja Schurz (2006, Capítulo 2.6.2).

(1970, p. 399), assim como Westermann e Hager (1982, p. 19), incluem na teoria do teste não apenas as opções de rejeição e de aceitação, mas também a opção de *cautela* – escolhendo o intervalo central de 66% como intervalo de aceitação, o intervalo extremo de 5% como intervalo de rejeição e o intervalo intermediário como intervalo de cautela. Cramer (1946, p. 421) vai ainda mais longe e, geralmente, interpreta os resultados dos testes estatísticos como resultados de *suporte gradual* (Howson; Urbach, 1996, p. 207).

Na minha perspectiva, a teoria de testes de Fisher fornece um método para gerar o conteúdo indutivo-empírico de uma hipótese estatística (conforme Seção 7.2) e, com base nisso, estabelecer uma *heurística* para sua confirmação/refutação e aceitação/rejeição provisória. No caso de teste, o método do intervalo de aceitação de Fisher fornece uma regra que especifica para cada tamanho de amostra n , com desvio padrão σ e intervalo $[\pm 2 \cdot \sigma]$, em torno do valor de probabilidade (ou média) r reivindicado pela hipótese, no qual a frequência de amostra observada pode se situar, para que não precise ser rejeitada como “muito implausível”. No caso da inferência, o método do intervalo de confiança de Fisher fornece uma variante do raciocínio de generalização indutiva, a qual transfere a frequência da amostra observada para a população (também chamada de “regra reta” ou de “regra percental”; cf. Salmon, 1974), juntamente à entrada de um intervalo de confiança ou de incerteza em torno do valor da amostra indutivo-generalizada.

É importante perceber que o conceito estatístico de probabilidade só pode ser usado para calcular a probabilidade dos resultados da amostra, dadas certas hipóteses populacionais, mas nunca a probabilidade das hipóteses populacionais propriamente ditas. As probabilidades estatísticas são baseadas em experimentos

aleatórios repetíveis, e a amostragem de uma população é um experimento aleatório repetível. Toda a população, ou “mundo real”, por outro lado, existe *apenas uma vez*: não tem probabilidade estatística, porque não há sequências aleatórias de mundos possíveis (cf. também Hays; Winkler, 1970, p. 328; Howson; Urbach, 1996, p. 239). *As probabilidades de hipótese* são, portanto, sempre *de natureza epistêmica* e pertencem ao domínio da teoria da probabilidade epistêmica. Por essa razão, seria confuso ler o resultado do método do intervalo de confiança dessa forma: com 95% de probabilidade estatística, a frequência da população está dentro do intervalo de confiança especificado. Ao invés disso, o método do intervalo de confiança diz, do ponto de vista estatístico, o seguinte: para todas as hipóteses no intervalo de confiança, o resultado real da amostra está no intervalo de seus 95% de resultados amostrais mais prováveis.

Em todos os testes estatísticos e procedimentos de inferência, portanto, o seguinte é importante: a magnitude da probabilidade estatística do resultado amostral E, dada uma hipótese estatística H, é usada como um *indicador* para a plausibilidade da hipótese H, dado o resultado amostral E. Na Seção 9.1, chamo essa abordagem de *intuição de verossimilhança*, porque a probabilidade estatística $p(E|H)$ também é chamada de *probabilidade* da hipótese H, dada a evidência E. Discutiremos os métodos estatísticos padrão à luz de algumas objeções bayesianas e mostraremos que a justificativa usual desses métodos (ao contrário das opiniões bayesianas) não é falsa nem arbitrária, mas meramente *incompleta*: para uma justificação completa desses métodos, é necessária, no sentido de nossa abordagem dualista, a noção de probabilidade epistêmica.

8.6 Distribuições de probabilidade e métodos estatísticos para variáveis contínuas

Uma variável matemática X é (como explicado no Apêndice 10.1) uma função $X:D \rightarrow V$ que atribui a cada indivíduo x , no domínio individual D , um valor numérico $X(x) \in V$, em uma faixa de números $V \subseteq \mathbb{R}$, a partir do conjunto de números reais \mathbb{R} . Um exemplo de X seria o peso das pessoas. X também é chamado de variável aleatória.⁵⁴ Uma distribuição de probabilidade p (ou P) sobre X é uma função de probabilidade $p:V \rightarrow [0,1]$ sobre valores numéricos variáveis da variável característica $X(x) \in V$. Se a variável característica $X(x)$ só pode assumir um número finito ou contável de valores numéricos em \mathbb{R} (por exemplo, o peso *arredondado* em kg), p é chamado de distribuição *discreta*. Se $X(x)$ pode assumir todos os valores numéricos em \mathbb{R} ou um *intervalo numérico* em \mathbb{R} , p é chamado de distribuição *contínua*. *Notação*: a seguir, $p_x(r)$ é abreviado como $p(X(x)=r)$ e $p_x([a,b])$ como $p(X(x) \in [a,b])$, onde $[a,b] = \{r \in \mathbb{R}: a \leq r \leq b\}$ denota o intervalo *fechado* de todos os números reais entre a e b , analogamente $(a,b) = \{r \in \mathbb{R}: a < r < b\}$ é o intervalo *aberto* entre a e b .

⁵⁴ Veja Bauer (1978, p. 136); Lauth/Streiter (2002, p. 255-7); Hays/Winkler (1970, p. 103); Jeffrey (1971b, p. 183).

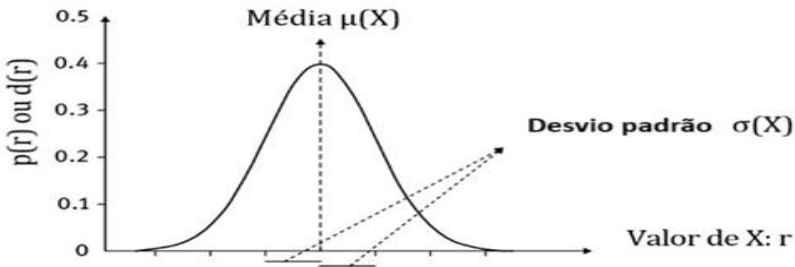


Figura 8-4: Distribuição normal gaussiana.

Na seção 3.2, aprendemos sobre a distribuição discreta mais básica das estatísticas: a distribuição binomial. Agora, vamos considerar a distribuição contínua mais básica: a distribuição normal gaussiana $p(r)$ ou $d(r)$, mostrada na Figura 8-4. A probabilidade $p(r)$ é usada na versão *discreta* da distribuição normal: por exemplo, se X é o peso, então, $p(50)$ representa a frequência de pessoas cujo peso é arredondado para 50 kg. Distribuições contínuas têm a seguinte característica especial sobre distribuições discretas: a probabilidade de que uma característica quantitativa ocupe exatamente um de um número incontável de valores numéricos reais possíveis (ou seja, um valor numérico real com precisão infinitamente precisa) é, tipicamente, sempre zero. São interessantes as probabilidades de intervalos diferentes de zero da linha numérica real, como $p(70 \pm 0,5 \text{ kg})$; essas probabilidades intervalares são tipicamente positivas. Portanto, não se pode representar a distribuição de probabilidade, sobre uma linha numérica de valor real, pelas probabilidades $p(r)$ em si, porque, então, se obteria a linha zero trivial. Em vez disso, contenta-se com a chamada densidade de *probabilidade* $d(r)$.⁵⁵ A probabilidade de encontrar o valor de X no intervalo $[a, b]$ é, portanto, dada como a

⁵⁵ Definida como a 1ª derivada da probabilidade cumulativa, $d(x) =_{\text{def}} dp([-\infty, x])/dx$.

integral da densidade de probabilidade $d(r)$ sobre esse intervalo, a qual é assim escrita como equação:

(8-11) (a) Caso discreto:

$$pX([a,b]) = \sum_{a \leq x \leq b} pX(x)$$

(b) Caso contínuo:

$$pX([a,b]) = \int_a^b d_x(r) dr$$

Matematicamente falando, a *integral* é a generalização da operação da soma para o caso contínuo. Do ponto de vista gráfico, a integral 8-11(b) corresponde à *área* sob a distribuição de densidade $d_x(r)$, na Figura 8-4, entre os valores X , a e b . A área total sob a curva de densidade, de $-\infty$ a $+\infty$, é normalizada para 1, porque a probabilidade de a variável X assumir qualquer valor é 1. A definição da função de distribuição gaussiana, para o domínio de *todos os* números reais, é uma idealização matemática, mas é inofensiva, porque as densidades gaussianas de valores- X , distantes da média, são próximas de zero.

Com base na equação (8-11), estende-se a função de probabilidade sobre V para uma medida de probabilidade sobre uma álgebra apropriada $AL(V)$ sobre V , $pX: AL(V) \rightarrow [0,1]$. No caso discreto, $AL(V)$ é tipicamente o conjunto de potência de V e, no caso contínuo, a chamada *álgebra de Borel*, $Bo(R)$ sobre R , é o conjunto de todos os intervalos de números reais completados sob concordância e complemento infinitos.⁵⁶ A função $pX: AL(V) \rightarrow [0,1]$ é redutível a uma

⁵⁶ A integral da função densidade $d(r)$ não é definida significativamente para subconjuntos arbitrários de números reais, mas apenas para subconjuntos mensuráveis. A integral ordinária (Riemanniana) é explicada para intervalos de números reais. Com base nisso, uma medida σ -aditiva pode ser definida sobre a álgebra de Borel de todas as combinações booleanas de intervalos. Esta medida também é chamada de medida de Borel-Lebesgue.

medida de probabilidade sobre uma álgebra, sobre o domínio subjacente de indivíduos ou de espaço de possibilidade D , $p:AL(D) \rightarrow [0,1]$, desde que a variável aleatória $X:D \rightarrow V$ seja *mensurável* em relação a $(D, AL(D), p)$. Isso significa que cada elemento A , em $Bo(R)$, tem um arquétipo $A_{-X} =_{\text{def}} \{x \in D: X(x) \in A\}$ em $AL(D)$, o qual é transferido para a medida $Bo(R)$, ou seja, $p_X(A) =_{\text{def}} p(A_{-X})$ (para detalhes, ver Bauer, 1996, §§4-8; Billingsley, 1995, Seção 2-4; Jeffrey, 1971).

As relações explicadas aplicam-se não apenas a distribuições normais, mas também a distribuições arbitrárias – por exemplo, *distribuições uniformes* (ou iguais), *distribuições assimétricas* ou *distribuições multimodais* (com múltiplos picos).

Os principais parâmetros estatísticos de uma distribuição $p(X)$ são sua média ou média aritmética $\mu(X)$ e seu desvio padrão $\sigma(X)$. Esses parâmetros são definidos para quaisquer distribuições. Para a distribuição normal, $\mu(X)$ e $\sigma(X)$ estão representados na Figura 8-4. O valor médio $\mu(X)$ é a média da grandeza X entre os indivíduos da população D . O desvio padrão $\sigma(X)$ de uma distribuição informa sobre o desvio médio das ocorrências individuais de X em relação à média.

Devido à compensação dos desvios direcionados $\pm(r_i - \mu(X))$ em relação à média, soma-se os quadrados dos desvios, cuja soma é chamada de variância $v(X)$, e extrai-se a raiz quadrada. Quanto maior o desvio padrão, mais achatada é a distribuição. Matematicamente, os termos são definidos da seguinte forma:

(Def. 8-3) *Média e desvio padrão na população:*

Caso Discreto: $X(x) \in \{r_1, \dots, r_n\}$

Caso Contínuo: $X(x) \in \mathbb{R}$

$$\text{Média: } \mu(X) = \sum_{i=1}^n r_i \cdot p(r_i)$$

$$\mu(X) = \int_{-\infty}^{+\infty} r \cdot d(r) \, dr$$

$$\text{Variância: } v(X) = \sum_{i=1}^n (r_i - \mu(X))^2 \cdot p(r_i)$$

$$v(X) = \int_{-\infty}^{+\infty} (r_i - \mu(X))^2 \cdot d(r) \, dr$$

Desvio padrão: Raiz quadrada da variância: $\sigma(X) = \sqrt{v(X)}$.

A definição da média também subsume características binárias XF quando Fx é codificado por 1 e $\neg Fx$ por 0: nesse caso, seguem $\mu(XF) = p(Fx)$ e $\sigma(XF) = \sqrt{p(Fx) \cdot (1 - p(Fx))}$.

Para distribuições simétricas-unimodais, a média coincide com o valor mais comum ou com o pico da distribuição, o chamado valor modal, e os demais valores são simetricamente agrupados em torno dele com frequência decrescente. Por outro lado, no caso de uma distribuição inclinada para a esquerda-unimodal (estrita e íngreme à esquerda, larga e plana à direita), o valor médio está ligeiramente à direita do valor modal: um exemplo é a distribuição χ^2 das variedades de amostra mencionadas abaixo.

Uma distribuição normal gaussiana $g(r)$ é matematicamente definida de forma inequívoca por dois parâmetros, a saber, seu valor médio μ e seu desvio padrão σ :

$$(8-12) \quad g(r) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{(r-\mu)^2}{2 \cdot \sigma^2}}$$

μ determina o centro e σ o grau de achatamento da distribuição normal. Acima do intervalo de $\mu - \sigma$ a $\mu + \sigma$ é exatamente 66% da probabilidade, acima de $[\mu - 2\sigma, \mu + 2\sigma]$ é 95,5% da probabilidade – esse é o intervalo de aceitação usual – e acima de $[\mu - 3\sigma, \mu + 3\sigma]$ é 99% da área, e assim por diante. Se $\mu = 0$ e $\sigma = 1$, obtemos a distribuição normal normalizada $g(z) = (1/\sqrt{2 \cdot \pi}) \cdot e^{- (z^2/2)}$. Qualquer distribuição normal sobre uma variável aleatória- X é normalizada, subtraindo-se a média- X dos valores numéricos- X e dividindo-se [esse resultado] pelo desvio padrão- X . Isso é chamado de “transformada z ” de uma distribuição normal:

$$(8-13) \text{ Transformada } z: z = \frac{r-\mu}{\sigma} \text{ (isto é, } \forall x \in D: ZX(x) = (X(x) - \mu(X))/\sigma(X))$$

As integrais $\int_{-\infty}^Z g(z) dz$ da distribuição normal normalizada podem ser visualizadas em tabelas de distribuição normal, as quais podem ser encontradas em livros didáticos de estatística. Essas tabelas são usadas para determinar intervalos de aceitação e confiança, bem como diferenças significativas. Para uma dada função $f(r)$ de valores X r , a expressão

$$(8-14) E(f(X)) =_{\text{def}} \sum_{i=1}^n f(r_i) \cdot p(r_i) \text{ bzw. } \int_{-\infty}^{+\infty} f(r) \cdot d(r) \text{ dr}$$

como o *valor esperado* da função $f(r)$ em relação à distribuição X $p(r_i)$ ou $d(r)$. Para $f(X) = X$, $E(X)$ corresponde à média (aritmética) de X para probabilidades estatísticas e ao valor subjetivo esperado de X para probabilidades subjetivas. A variância $v(X)$ é o valor esperado da função $(X - E(X))^2$.

Algumas leis gerais de cálculo podem ser provadas para valores esperados, que estão resumidas no (Teorema 8-2) e às quais retornaremos a seguir:

(Teorema 8-2) *Leis de cálculo para valores esperados:*

(a) *Linearidade:* $E(r_1 \cdot X_1 + \dots + r_n \cdot X_n + q) = r_1 \cdot E(X_1) + \dots + r_n \cdot E(X_n) + q$
 $(r_i, q \in \mathbb{R})$

(b) *Variâncias:* (i) $v(X) =_{\text{def}} E((X - E(X))^2) = E(X^2) - (E(X))^2$;
 (ii) $v(r_1 \cdot X + r_2) = r_1^2 \cdot v(X)$.

(c) *Variância e covariância:* $v(r \cdot X + q \cdot Y) = r^2 \cdot v(X) + q^2 \cdot v(Y) + 2 \cdot r \cdot q \cdot \text{cov}(X, Y)$.

Evidências do Teorema (8-2) podem ser encontradas em Bortz (1985, Apêndice B, p. 803) e em Hays e Winkler (1995, §§3.14, 3.21, 3.25).⁵⁷

Para descrever uma amostra individual de n elementos $s_n = \{a_1, \dots, a_n\}$, a média μ_{s_n} e o desvio padrão σ_{s_n} da variável aleatória X, em s_n , são definidas analogamente como na Def. 8-3, com a diferença de que, aqui, é mais fácil somar diretamente sobre os indivíduos da amostra (em vez de sobre seus valores):

(Def. 8-4) *Média e desvio-padrão em uma amostra* $s_n = \{a_1, \dots, a_n\}$

$$\mu_{s_n}(X) = \frac{\sum_{i=1}^n X(a_i)}{n} \quad v_{s_n}(X) = \frac{\sum_{i=1}^n (X(a_i) - \mu_{s_n})^2}{n} \quad \sigma_{s_n}(X) = \sqrt{v_{s_n}(X)}$$

⁵⁷ Por exemplo, (a) segue da linearidade das integrais o seguinte: $E(a \cdot X + b \cdot Y + c) = \int (a \cdot X(r) + b \cdot Y(r) + c) \cdot d(r) = a \cdot \int X(r) \cdot d(r) + b \cdot \int Y(r) \cdot d(r) + c \cdot \int d(r) = a \cdot E(X) + b \cdot E(Y) + c$.

Enquanto a estatística *descritiva* trata da representação das propriedades de distribuição de amostras empíricas individuais, as estatísticas de inferência e de teste estão interessadas na distribuição de probabilidade dos valores característicos de qualquer amostra aleatória – em particular, na *distribuição dos valores médios* de qualquer amostra aleatória de uma população.

A distribuição de probabilidade dos resultados amostrais de n elementos é exemplificada pela distribuição comum independente de n variáveis aleatórias idênticas $X(x_1), \dots, X(x_n)$, ou seja, a densidade de probabilidade comum do resultado $X(x_1)=r_1 \wedge \dots \wedge X(x_n)=r_n$ é considerada o produto $d(r_1) \cdot \dots \cdot d(r_n)$. Para a distribuição de probabilidades de médias de amostras com n elementos, utilizando as regras de cálculo para valores esperados, obtêm-se o seguinte valor médio $\mu_n(X)$ e o seguinte desvio padrão $\sigma_n(X)$ (para a prova, ver o Anexo 10.3.13):

(Teorema 8-3) *Média e desvio-padrão das médias amostrais:*

$$\begin{array}{lll} \text{(i)} \mu(\mu_{s_n}(X)) = \mu(X) & \text{(ii)} v(\mu_{s_n}(X)) = v(X)/n & \text{(iii)} \sigma(\mu_{s_n}(X)) \\ = \sigma(X) / \sqrt{n} \end{array}$$

A média (ou valor esperado das médias amostrais) é, portanto, idêntica à média populacional. Parâmetros amostrais cujo valor esperado coincide com o valor populacional também são chamados de *não enviesados* [*fiéis às expectativas*] (Bortz, 1985, p. 124; Hays; Winckler, 1995, p. 308). O desvio padrão da média da amostra diminui à medida que o tamanho da amostra (n) aumenta, inversamente proporcional à raiz quadrada de n . Essas relações aplicam-se não apenas

a distribuições normais, mas a quaisquer distribuições independentes e idênticas com variância finita. Desse fato, seguem *as leis dos grandes números* para qualquer variável aleatória, segundo as quais a média da amostra converge para a média da população com probabilidade 1 à medida que o tamanho da amostra se aproxima do infinito. Essas leis têm a mesma forma e são provadas analogamente, como no caso binário (Teorema 3-4), para $\mu(X)$, em vez de $p(A)$, e para $\mu_{sn}(X)$, em vez de $h_n(A)$. Igualmente importante é o *Teorema do limite central*, segundo o qual a distribuição das médias amostrais de uma distribuição arbitrária, sobre X para n crescente, converge para uma distribuição normal com média $\mu(X)$ e com desvio padrão $\sigma(X)/\sqrt{n}$ (Bauer, 1978, §51; Bortz, 1985, p. 121; Lauth; Sareiter, 2002, p. 267).

O Teorema do limite central justifica aproximar a distribuição dos valores médios da amostra e uma variável aleatória *arbitrariamente* distribuída X , para n suficientemente alto ($n \geq 30$), por uma distribuição normal.⁵⁸ Isso explica a importância central da distribuição normal para os métodos de teste e estatística inferencial: a distribuição normal gaussiana mostra-se como a forma matemática da distribuição de erros aleatórios ou como os desvios aleatórios de um parâmetro central.

Isso também explica por que a distribuição de probabilidade das características de tamanho, geralmente, segue uma distribuição normal: esse é sempre o caso quando o valor de X é o resultado de uma tendência comum para todos os indivíduos da população, a qual é sobreposta por muitos fatores de confusão espalhados aleatoriamente. *Distribuições de dois picos*, por outro lado, surgem quando a população

⁵⁸ Para amostras menores, a distribuição binomial é usada para variáveis binárias e a distribuição t é usada para variáveis quantitativas (ambas podem ser encontradas em forma de tabela em livros didáticos de estatística).

consiste em dois subgrupos heterogêneos em relação ao traço X, por exemplo, os muito leves, os muito avançados e os moderados. Como explicado, mesmo para populações tão heterogêneas, a distribuição das médias amostrais converge para uma distribuição normal.

As leis computacionais para valores esperados também podem ser usadas a fim de calcular o valor esperado da variância da amostra, o qual obedece a uma distribuição χ^2 *inclinada para a esquerda*. O valor modal dessa distribuição está à esquerda do valor esperado. O valor esperado da variância da amostra v_{sn} é calculado como

$$(8-15) E(v_{sn}(X)) = v(X) - v_{\mu_n}(X) = v(X) \cdot (n-1)/n$$

isto é, como variância menos variância da média amostral (para a prova, ver Bortz, 1985, p. 808). O resultado é que, em amostras, tanto a variância quanto a média desviam-se dos parâmetros populacionais e, em amostras com média desviante, a variância média é ligeiramente menor que a variância populacional. O desvio padrão da amostra não é, portanto, um estimador imparcial do desvio padrão da população. Em contraste, a *variação corrigida da amostra* é

$$(8-15) E(v_{sn}(X)) = v(X) - v_{\mu_n}(X) = v(X) \cdot (n-1)/n,$$

um estimador imparcial do desvio padrão da população. Esse desvio padrão amostral corrigida é usada para estimar o desvio padrão de uma população. Assim, para determinadas amostras, os critérios estabelecidos na Seção 8.1-3 explicam intervalos de aceitação, intervalos de confiança e diferenças significativas. Ilustramos isso com dois exemplos:

Exemplo 1 (característica discreta da Seção 8.1, intervalo de aceitação e confiança): a hipótese H , referente à abundância de K na população A (árvores doentes ao longo de rodovias), é $p(K|A) = \mu(K|A) = 0,8$. Extraímos uma amostra de 100 elementos. O desvio padrão das frequências amostrais $\sigma_{\mu_{100}}$ (com $p =_{\text{def}} p(K|A)$) é $\sqrt{[p \cdot (1 - p)]} / \sqrt{n} = \sqrt{(0,8 \cdot 0,2)} / \sqrt{100} = 0,04$ (arredondado). Na tabela da distribuição normal padronizada, vemos que o valor z de $-\infty$ é atingido apenas 2,5% da probabilidade cumulativa: esse é o caso de $z = -1,96$. Por razões de simetria, à direita do valor $z = +1,96$, está também 2,5% da probabilidade acumulada. O intervalo simétrico de 95% da distribuição normal padronizada estende-se, portanto, de $z = -1,96$ a $z = +1,96$. Após a inversão da transformada z em (8-13), o intervalo de aceitação de 95% está entre $\mu - 1,96 \cdot \sigma$ e $\mu + 1,96 \cdot \sigma$, mais especificamente, no intervalo $0,8 \pm 1,96 \cdot 0,040$, ou seja, entre 0,72 e 0,88, ou ainda, entre 72 e 88 K -indivíduos de 100 A -indivíduos. O resultado real da amostra é $\mu_{s,100}(K) = 0,75$, de modo que a hipótese é fracamente confirmada, e obtemos o intervalo de confiança de $0,67 \leq p(K|A) \leq 0,83$.

Exemplo 2 (característica contínua, intervalo de aceitação e confiança): Hipótese H : A idade média da primeira paixão entre adolescentes do sexo feminino é de 15 anos. Para testar essa hipótese, coletamos uma amostra de 25 meninas. O desvio padrão corrigido $\sigma_{sn}^{\text{corr}}$ dessa amostra A é 2,5; isso permite estimar o desvio padrão populacional. Analogamente ao anterior, obtemos os seguintes limites de intervalo simétricos de 95%:

$$\mu(X) \pm 1,96 \cdot \sigma_{sn}^{\text{corr}} / \sqrt{n} = 15 \pm 1,96 \cdot 2,5 / \sqrt{25} = 15 \pm 1 \text{ (arredondado)}.$$

Assim, o intervalo de aceitação da hipótese encontra-se na faixa média da amostra [14, 16]. Se nosso resultado amostral fosse $\mu_{s_{25}}(X) = 14$, nossa hipótese seria apenas fracamente confirmada ou não rejeitada, e nosso intervalo de confiança para a hipótese seria entre 13 e 15.

Nesse segundo exemplo, usamos a distribuição normal, embora a amostra seja bastante pequena. Se olharmos para a distribuição t com $n-1 = 24$ graus de liberdade, lemos o valor 2,064 em vez de 1,96; como resultado, o intervalo de aceitação é um pouco maior.

Para verificar diferenças significativas entre duas amostras *sorteadas independentemente* s_n e s_m , é utilizado o chamado *teste t* para amostras independentes (o que não significa que a distribuição t deva ser sempre usada). Uma amostra tem os antecedentes característicos A e a outra não. A diferença entre as duas médias amostrais é calculada: $\Delta = \mu_{s_n} - \mu_{s_m}$. A distribuição de probabilidade da diferença média amostral Δ é uma distribuição normal com média baseada na hipótese nula ($\mu(X_K|A) = \mu(X_K)$), com o valor médio 0 e o desvio padrão $\sigma(\Delta) = \sigma(X) \cdot \sqrt{\left(\frac{1}{n}\right) + \left(\frac{1}{m}\right)}$ (cf. a prova em Bortz, 1985, p. 166). Estimamos a variância populacional desconhecida $\sigma(X)$ pelas duas variâncias amostrais corrigidas da seguinte forma:

$$\sigma_{\text{estimado}}(X)^2 = v_{\text{estimada}}(X) = ((n-1) \cdot v_{s_n}^{\text{corr}}(X) + (m-1) \cdot v_{s_m}^{\text{corr}}(X)) / (n+m-2).$$

Os limites da diferença média significativa de 95% são obtidos multiplicando-se os limites intervalares correspondentes da distribuição z, $\pm 1,96$, pelo desvio padrão estimado da diferença amostral Δ :

$$\Delta_{\text{sign}}(X) = \pm 1,96 \cdot \sigma_{\text{estimado}}(X) \cdot \sqrt{\left[\left(\frac{1}{n}\right) + \left(\frac{1}{m}\right)\right]}.$$

Exemplo 1 (continuação, diferença significativa): nossa primeira amostra A de 100 elementos continha 75 Ks. Pegamos uma amostra de 100 elementos \bar{A} , na qual encontramos 60 Ks. Calculamos a variância populacional estimada a partir das duas variâncias amostrais dadas para a distribuição binomial, como $p \cdot (1-p)$, da seguinte forma:

$$\sigma^2(K)_{\text{estimado}} = (99 \cdot 0,75 \cdot 0,25 + 99 \cdot 0,6 \cdot 0,4) / 198 = 0,21 \text{ (arredondado)}.$$

Assim,

$$\Delta_{\text{sign}} = 1,96 \cdot \sqrt{0,21} \cdot \sqrt{\frac{2}{100}} = \pm 0,13 \text{ (arredondado)}.$$

A diferença de 15 encontrada é, portanto, significativa. O valor z, pertencente a essa diferença, é calculado como

$$z_{\text{sign}} = 0,15 / \sqrt{0,21} \cdot \sqrt{\frac{2}{100}} = \pm 2,31.$$

À esquerda e à direita de $z = \pm 2,31$, é (arredondado) 1% da distribuição z, de modo que a diferença encontrada é praticamente significativa, com um coeficiente de 2%.

Exemplo 2 (continuação, diferença significativa): desenhamos uma amostra de 30 elementos de meninos com dispersão 3 corrigida (ou seja, com variância 9). O desvio padrão corrigido da amostra é 2,5 (variância 6,25). Para o desvio padrão estimado de ambas as amostras, obtemos

$\sigma^2_{\text{estimado}} = (24 \cdot 6,25 + 29 \cdot 9) / 53 = 7,75$. Isso indica para a diferença significativa

$$\Delta_{\text{sign}} = 1,96 \cdot \sqrt{7,75} \cdot \sqrt{\frac{1}{25} + \frac{1}{30}} = 1,96 \cdot 2,78 \cdot 0,27 = \pm 1,47 \text{ (arredondado)}.$$

Uma diferença média de pelo menos 1,47 ano de primeiro amor entre meninas e meninos seria, portanto, significativa. Por exemplo, se a idade média do primeiro caso amoroso, em nossa amostra de meninos, fosse de 16,7 anos, a diferença para a amostra de meninas seria significativa. E o valor z , pertencente à diferença de 1,7 anos, seria $z_{\text{sign}} = 1,7 / 2,78 \cdot 0,27 = \pm 2,26$ que tem desempenho na distribuição z $\pm 1,2\%$. A diferença seria, portanto, significativa com um coeficiente de 2,4%.

A comparação de *amostras pareadas* ocorre quando duas medidas diferentes são tomadas nos indivíduos de uma mesma amostra; por exemplo, a medida de uma variável aleatória X antes e após o tratamento com um fator A . Nesse caso, no chamado *teste t pareado*, a variável das diferenças $\Delta_i(a) =_{\text{def}} X_1(a_i) - X_2(a_i)$ é formada para todos os membros da amostra a_i e é verificada em relação ao desvio significativo da dispersão aleatória (Bortz, 1985, p. 170). Há uma variedade de outros métodos estatísticos que funciona da mesma maneira.

Importante para o estudo das dependências é a observação da distribuição de *probabilidade comum*, $d(r,q)$, de duas variáveis aleatórias, $X:D \rightarrow R$ e $Y:D \rightarrow R$, sobre o mesmo espaço amostral D . Aqui, $d(r,q)$ é uma abreviação para $d(X(x)=r \wedge Y(x)=q)$, ou seja, para a densidade de probabilidade na qual X assume o valor r e Y o valor q . Um exemplo seria a distribuição conjunta do peso e do QI dos indivíduos em uma população. Duas dessas variáveis aleatórias X, Y são chamadas de (probabilísticas) *independentes* se não estiverem correlacionadas para todas as suas possíveis expressões de valor, de modo que para a função de densidade comum: $d(r,q) = d(r) \cdot d(q)$ para todos os $r,q \in R$ (Bauer, 1978, §31).

Por outro lado, duas variáveis X e Y são ditas *não correlacionadas* se sua chamada *covariância* for zero, a qual é definida da seguinte forma:

$$(8-17) \text{Covariância: } \text{cov}(X,Y) =_{\text{def}} E((X-E(X)) \cdot (Y-E(Y))) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (r - \mu_x) \cdot (q - \mu_y) \cdot d(r, q) dq dr .$$

A covariância $\text{cov}(X,Y)$ é uma medida do desvio conjunto das variáveis aleatórias X e Y da respectiva média, sendo maior quanto mais objetos, com um valor X acima da média, também tendem a ter um valor Y acima da média, e vice-versa. Dividindo a covariância $\text{cov}(X,Y)$ pelo produto de ambos os desvios-padrão $\sigma(X)$ e $\sigma(Y)$, obtém-se uma medida básica de correlação para variáveis intervalares, a *correlação produto-momento* $r(X,Y)$, cujos valores variam entre -1 e $+1$ (Bortz, 1985, p. 251). Notemos que $r(X,Y)$ mede a força de uma relação *linear*, se X e Y são correlacionados de forma não-linear, então, $r(X,Y)$ resulta em uma relação menor do que realmente existe (Clauß; Ebner, 1977, p. 116). Para características binárias, a covariância toma o início do Cap. 8, de forma explicada: $\text{cov}(X_F, X_G) = p(F_X \wedge G_X) - p(F_X) \cdot p(G_X)$.

A independência probabilística implica não correlação, mas não vice-versa (Bauer, 1978, p. 154, Def. 32.2; Lauth; Sareiter, 2002, p. 259). Apenas para valores de variáveis isoladas e *para variáveis binárias*, a independência probabilística e a não correlação coincidem. Para uma dada variável com mais de dois valores, sua (não) correlação é uma propriedade *média* sobre todos os valores. É possível que as variáveis X, Y estejam positivamente correlacionadas em valores altos e negativamente em valores baixos, e ambos os efeitos compensem-se mutuamente na soma a zero. Nas correlações, informações sobre dependências probabilísticas podem, assim, ser perdidas; somente se

as dependências forem lineares por natureza é que isso será impossível.

8.7 Fontes de erro nas estatísticas: representatividade, interpretação causal e caso individual

As três principais fontes de erro na aplicação dos métodos estatísticos são: (i) falta de representatividade e comparabilidade das amostras, (ii) interpretação causal apressada e (iii) supervisão de especificidades do caso de uso individual. A seguir, vamos dar uma olhada mais de perto nessas fontes de erro.

8.7.1. Representatividade

No teste de uma hipótese estatística $p(K|A) = r \neq p(K|\neg A)$, ou $\text{corr}(K,A) = r$, aplicam-se três pré-requisitos: 1.) a amostra A deve ser o mais representativa possível, 2.) deve haver uma amostra de controle \bar{A} , a qual também deve ser tão representativa quanto possível, e 3.) uma amostra A e uma amostra de controle \bar{A} devem ser tão comparáveis quanto possível em termos de características causalmente independentes de A .

Exemplos de fontes de erro de representatividade:

a) Fazemos uma enquete telefônica *flash*, utilizando somente telefones fixos, sobre uma questão política. No entanto, isso não atinge

os usuários exclusivos de celulares da geração mais jovem. A amostra não é representativa.

b) Uma socióloga envia questionários para pesquisar a carga de trabalho doméstico das mulheres. 20% dos questionários são devolvidos. Quais mulheres preenchem o questionário e o enviam de volta? Possivelmente, aquelas que são mais sofisticadas, mais comprometidas em termos de educação – e, certamente, as mais propensas a serem menos sobrecarregadas em termos de trabalho. A amostra não é representativa.

c) *Omissão de casos negativos*: um erro muito drástico de representatividade é cometido quando os casos que falam contra a própria teoria são simplesmente omitidos da amostra.

Exemplos de fontes de erro no grupo de controle:

d) *Falta de um grupo controle*: em avaliações não científicas da experiência, muitas vezes, não há comparação com um grupo controle. Se, por exemplo, são relatadas práticas de cura esotéricas que, em algumas pessoas, a doença diminui visivelmente após a aplicação da “imposição de mãos”, isso diz pouco; afinal, não sabemos (i) se a redução da doença não teria ocorrido sem a imposição das mãos, através da recuperação normal ou através de outras práticas prescritas (por exemplo, alimentação mais saudável), e (ii) se não havia um *efeito placebo* em jogo (a forte crença no poder de cura pode acelerar a recuperação). Tudo isso só pode ser descartado por comparações apropriadas de grupos controle.

e) *Falta de comparabilidade com o grupo controle*: atualmente, os grupos controle são práticas comuns para testar os efeitos de fármacos. São realizados os chamados testes *duplo-cegos*, nos quais a

metade dos participantes selecionados aleatoriamente recebe uma droga real e a outra metade uma pseudopílula, e nem os pacientes nem os experimentadores sabem quais pílulas são reais e quais pílulas são placebos. Todos os tipos de distorções na comparabilidade entre o grupo de características e o grupo de controle são, assim, excluídos, salvo erros aleatórios.

Como regra geral, deve-se notar que a comparabilidade das duas amostras só pode *ser garantida em relação às características* que são estatisticamente independentes de A, por meio de um método de amostragem simples (também conhecido como *quase-experimento*); esse é o problema no qual se baseiam as correlações não causais explicadas na próxima Seção. A comparabilidade de ambas as amostras, em relação a todas as características *causalmente independentes* de A, só pode ser alcançada por um *experimento randomizado*. Nesse caso, o recurso de antecedentes introduzido experimentalmente ainda pode introduzir falsas variáveis “ocultas” na amostra, conduzindo a uma distorção da comparabilidade. Por exemplo, um novo método de ensino deve ser comparado com um método convencional, e o grupo de alunos envolvidos com o novo método tem um desempenho melhor do que aqueles envolvidos com o método convencional em um teste de desempenho subsequente. Contudo, e se os professores aplicados ao novo método estão muito mais motivados a ensinar bem do que os professores aplicados ao método convencional? Ademais, os alunos sabem do experimento e, em caso afirmativo, são motivados de forma diferente por ele? Todas essas possibilidades conduziriam a distorções de comparabilidade e, portanto, a falsas hipóteses de significância.

8.7.2 Correlação e causalidade

Uma correlação estatística entre duas características A e B pode ser usada para *fins preditivos*. No entanto, não é imediatamente possível concluir que existe uma *relação causal* entre A e B, por duas razões: *primeiro*, porque a correlação pode ter sido causada por variáveis ocultas e, *segundo*, porque as correlações são simétricas e não indicam a direção da relação causal.

Os principais tipos de *variáveis ocultas* são *causas comuns*. A situação é mostrada graficamente na Figura 8-5.

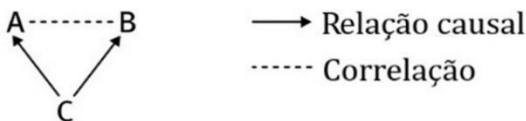


Figura 8-5: C é a causa comum de A e B.

O surgimento da correlação entre os efeitos A e B pode ser explicado, informalmente, da seguinte maneira: como as correlações são simétricas, a ocorrência do efeito A também aumenta a probabilidade de a causa C ter ocorrido; entretanto, ao mesmo tempo, C aumenta a probabilidade de ocorrência de B (independente de A). Disso, segue-se probabilisticamente que A também aumenta a probabilidade de ocorrência de B. Aqui, estão dois exemplos:

(8-18) Existe uma alta correlação entre a queda súbita do barômetro (A) e o aparecimento iminente de uma tempestade (B) (Grünbaum, 1972, p. 309).

Como o surgimento da tempestade ocorre *após* a queda do barômetro, pode-se pensar que a queda do barômetro seria a causa da

tempestade que se aproxima. Isso, é claro, é um absurdo: na verdade, ambos os eventos são o efeito comum de um terceiro evento C, ou seja, a queda repentina de pressão na atmosfera.

(8-19) Há uma alta correlação positiva entre o grau de atitude positiva em relação à empresa (A) e a saúde mental dos trabalhadores (B).

Isso significa que alguns trabalhadores constantemente incomodam a empresa, porque têm problemas psicológicos e acreditam ser a empresa a causa deles, enquanto a empresa seria inocente? A análise estatística das variáveis ocultas mostra um quadro diferente. Verifica-se que ambas as características são um efeito comum da causa oculta do *estresse no ambiente de trabalho* (C): os trabalhadores com más condições de trabalho reclamam mais da empresa e, ao mesmo tempo, a saúde mental de cada um deles sofre com o estresse constante. Esse exemplo foi elaborado por Lazarsfeld (ver Mayntz *et al.*, 1974, p. 200) e se refere a um estudo realizado na década de 1950.

Também é conhecida a piada estatística sobre o mito da cegonha que traz os bebês. Na verdade, existe uma correlação positiva entre a frequência de cegonhas e nascimentos. Essa causalidade aparente tem uma origem comum: *a variável região rural versus urbana*. Nas áreas rurais, nascem mais crianças e também há mais cegonhas.

Existe um método bem conhecido para detectar, estatisticamente, a causalidade aparente, o qual remonta a Reichenbach (1956, p. 159). Quando a correlação entre A e B é causada por uma variável comum C, a correlação entre A e B deve desaparecer ao controlar os valores da variável C. Assim, ao considerar apenas indivíduos com o atributo C, a presença adicional de B não aumenta a

probabilidade de A; e analogamente, ao considerar apenas indivíduos sem o atributo C. No Exemplo (8-19): se examinarmos apenas os trabalhadores em boas condições de trabalho, então, a correlação entre atitude perante a empresa e saúde mental deverá desaparecer; e o mesmo deveria acontecer se observássemos apenas trabalhadores em más condições de trabalho. Esse prognóstico é de fato confirmado.

Com base nesse raciocínio, Reichenbach faz a seguinte sugestão sobre como caracterizar causas comuns por meio de condições puramente estatísticas, as quais são chamadas de condições de *blindagem* (lembramos que $P(A \wedge B) \geq P(A) \cdot P(B)$ e $P(A|B) \geq P(A)$ são equivalentes se $P(B) > 0$):

(Teorema 8-4) *Condições de Reichenbach para "C é uma causa comum para A e B:*

$$\left. \begin{array}{l} (1) p(A \wedge B|C) = p(A|C) \cdot p(B|C) \\ (2) p(A \wedge B|\neg C) = p(A|\neg C) \cdot p(B|\neg C) \end{array} \right\} \begin{array}{l} C \text{ e } \neg C \text{ protegem } A \\ \text{de } B \end{array}$$

$$\left. \begin{array}{l} (3) p(A \wedge C) > p(A) \cdot p(C) \\ (4) p(B \wedge C) > p(B) \cdot p(C) \end{array} \right\} \begin{array}{l} C \text{ é positivamente relevante para} \\ A \text{ e } B \end{array}$$

De (1)-(4) segue-se (5): $p(A|B) > p(A)$, ou seja, B correlaciona-se positivamente a A.⁵⁹

59 Prova: $p(A|B) = p(A|B \wedge C) \cdot p(C|B) + p(A|B \wedge \neg C) \cdot (1 - p(C|B))$ (de acordo com TB4, Teorema 3-3) = (*): $p(A|C) \cdot p(C|B) + p(A|\neg C) \cdot (1 - p(C|B))$ devido às condições de proteção (1), (2). Das condições de relevância positiva (3), (4) segue-se que $p(A|C) > p$

As condições de Reichenbach envolvem alguns problemas (por exemplo, o problema de distinguir entre causas comuns e causas intermediárias), sobre os quais não podemos entrar em detalhes aqui (ver Schurz 2013b, Seção 4.4; Seção 6.7). No entanto, as condições de Reichenbach são extremamente úteis na prática estatística, pois podem confirmar ou refutar hipóteses causais. Essas condições também explicam por que a causalidade aparente pode ser detectada, salvo erros aleatórios em experimentos randomizados. Nesse experimento, uma amostra de indivíduos é dividida aleatoriamente em duas amostras e, só então, o fator A é realizado em um grupo (experimental) por intervenção externa, enquanto, no grupo controle, o fator A não é realizado ou só às vezes é realizado por acaso. Posteriormente, a frequência condicional de B no grupo experimental, $h_e(B) = h_e(B:A)$, é comparada com a frequência condicional $h_k(B)$ no grupo controle.

Todos os outros fatores (C, D,...) que não pertencem ao *efeito causal* de A estão agora distribuídos, estatística e uniformemente, no grupo experimental e no grupo controle, salvo erros aleatórios, porque a distribuição é aleatória. Assim, a intervenção experimental dissocia, causalmente, o evento tipo A de todas as outras variáveis que não pertencem ao *efeito causal* de A – essa intervenção também é chamada de intervenção causal (ver Pearl, 2000, p. 23; Woodward, 2004). Se a correlação entre A e B na população é devida a *alguma* causa comum (possivelmente desconhecida) C, então, não pode haver correlação entre A e B no experimento randomizado.

(A) > p(A|¬C), p(C|B) > p(C) e, portanto, (1-p(C|B)) < (1-p(C)). Isso, juntamente com (*), implica que p(A|B) deve ser maior que p(A) = p(A|C) · p(C) + p(A|¬C) · (1-C)).

Isso é mostrado a seguir. Se há uma causa comum C para A e B na população, qualquer que seja, então, o seguinte deve ser aplicado com base nas condições de Reichenbach:

$$\text{(Teorema 8-4)(3,4): (i) } h_e(B|C) \approx h_k(B|C) \text{ e } h_e(B|\neg C) \approx h_k(B|\neg C),$$

porque, se o valor C é mantido, A não exerce qualquer influência causal sobre B (com \approx para “aproximadamente o mesmo, salvo erro aleatório”). Como C não é um efeito de A (apenas o contrário), C deve ser a mesma frequência, salvo erros aleatórios no grupo experimental e no grupo controle, ou seja,

$$\text{(ii) } h_e(C) \approx h_k(C) \text{ e } h_e(\neg C) \approx h_k(\neg C).$$

Devido à Lei da Multiplicação (TB4 do Teorema 3-3), aplica-se o seguinte:

$$\text{(iii) } h_e(B) = h_e(B|C) \cdot h_e(C) + h_e(B|\neg C) \cdot h_e(\neg C)$$

$$\text{(iv) } h_k(B) = h_k(B|C) \cdot h_k(C) + h_k(B|\neg C) \cdot h_k(\neg C)$$

Substituindo as equações aproximadas (i) e (ii) nas equações (iii) e (iv), teremos como resultado $h_e(B) \approx h_k(B)$.

Assim, um experimento randomizado (em contraste com um levantamento estatístico) pode mostrar que, salvo o erro aleatório, de B para A leva a uma relação causal. No entanto, isso ainda não mostra se essa relação causal é direta ou indireta. As variáveis intervenientes, ou seja, a variável C introduzida pelo estímulo A, não podem ser excluídas por uma experiência ou só podem ser excluídas com base em pressupostos de fundo adicionais.

Mesmo que se saiba não haver outras variáveis ocultas em jogo além de A e B, não se pode derivar diretamente uma hipótese causal da existência de uma alta correlação entre A e B, porque a correlação ainda não determina a *direção causal* – ou seja, ainda não estão determinados o que deve ser considerado como causa e o que deve ser considerado como efeito. As correlações são simétricas: se há uma correlação positiva entre duas características A e B, sem que isso se deva à influência causal de terceiras variáveis, então, há três possibilidades: ou A é uma causa de B, ou B é uma causa de A, ou A e B interagem e se reforçam na forma de um *feedback* causal (ver Fig. 8-6).

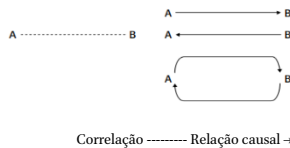


Figura 8-6: Qual é a direção causal?

Aqui estão alguns exemplos que vêm de pesquisas reais:

(8-20) O nível de QI e o nível de *status* social correlacionam-se positivamente.

A questão crucial é: qual é a causa, qual é o efeito? O QI é um produto da educação familiar e da instrução formal, que depende do *status*? Ou o defensor da correlação pensa em uma elite genética? Aparentemente, a relevância ideológica dessa correlação depende da hipótese sobre a direção causal, a qual não pode de modo algum ser lida a partir da correlação.

(8-21) Pessoas agressivas gostam de assistir a filmes agressivos.

Esse é um resultado bem conhecido. Contudo, a questão importante é: filmes agressivos tornam as pessoas agressivas? Ou as pessoas agressivas gostam de ver filmes agressivos porque desabafam de forma compensatória? Novamente, a relevância normativo-política depende da direção causal: no primeiro caso, deve-se neutralizar a exibição de filmes agressivos e, no segundo caso, deve-se tolerá-la ou mesmo incentivá-la.

Nesses exemplos, as pessoas são rapidamente inclinadas a ler uma correlação na direção correspondente às suas *próprias* visões. A atitude científica genuína só pode consistir na contenção quanto à direção causal, a menos que sejam dadas razões adicionais para a direção causal. Tais razões podem ser resultados, por exemplo, do conhecimento científico dos mecanismos causais. Em casos de relações entre eventos temporalmente separados, o evento anterior deve ser a causa do posterior, pois relações causais seguem uma direção temporal. Isso torna-se mais difícil no caso de características de disposição não localizadas no tempo, como nos exemplos (8-20,21). O objetivo aqui é descobrir, por meio de intervenções experimentais, o que é independente e qual é a variável dependente, ou se há retroalimentação.

A crítica a interpretações causais questionáveis também tem um componente *crítico-midiático*. Correlações estatísticas são frequentemente utilizadas na mídia e rapidamente interpretadas como achados causais com “valor sensacional”. Eis alguns *exemplos*: (1.) a amamentação prolongada reduz o risco de câncer de mama nas mulheres (*Osterreichische, Krone* 23.7.2002). Existe uma relação causal? Ou será que mulheres com mamas geneticamente mais

poderosas podem amamentar por mais tempo e, simultaneamente, podem ter um risco menor de câncer de mama? (2.) Há alguns anos, foi relatado que o frequente contato pele a pele correlaciona-se com um sistema imunológico mais forte. A questão é: existe uma relação causal direta entre esses fatores ou apenas as pessoas mais saudáveis e com sistemas imunológicos mais fortes, em média, são mais orientadas para o prazer? (3.) Mulheres que correm muito têm menor densidade óssea (*Salzburger Nachrichten*, 3.2.2003). Corridas frequentes devem, realmente, levar à perda de densidade óssea? Ou será que pessoas com densidade óssea geneticamente menor costumam ter melhor predisposição para andar e, portanto, para correr com mais frequência? (4.) Homens que comem muito chocolate são mais gentis e socialmente sociáveis (*Bayrisches Fernsehen*, 27.2.2003). Deve realmente haver uma relação causal e, em caso afirmativo, em que direção? Ou uma causa comum, possivelmente um tipo de característica geneticamente determinada, não deveria estar em ação? Somos tão estúpidos a ponto de deduzir dessa correlação que os homens deveriam comer mais chocolate – o que aumenta os níveis de colesterol e diminui a expectativa de vida?

A lista continua. É claro que as interpretações causais das correlações também podem ser extremamente *sérias*. Existem métodos elaborados para verificar, com mais detalhes, em que direção haveria, e se haveria, uma relação causal – por exemplo, por meio de estudos longitudinais ao longo do tempo. A maioria das pesquisas, no entanto, é baseada em estudos transversais e, muitas vezes, conclui-se injustificadamente uma relação causal, o que reiteradamente conduz os políticos a tomarem medidas que alcançam algo bem diferente do que esperam alcançar.

8.7.3 Aplicação de hipóteses estatísticas ao caso individual

A não monotonia das probabilidades condicionais (Fig. 3-2) faz uma diferença decisiva na aplicação de hipóteses estatísticas com a finalidade de prever ou de explicar casos individuais em comparação com hipóteses estritas. A conclusão K_a de uma inferência dedutiva, com premissas verdadeiras $\forall x(Ax \rightarrow Kx)$, e A_a pode ser *derivada* em qualquer momento – podemos inferir da verdade das premissas para a verdade da conclusão sem termos de nos preocupar com o que mais é verdade. No entanto, a partir das premissas $p(Kx|Ax) = 90\%$ e A_a de uma conclusão de especialização estatística indutiva, só se pode concluir que K_a é verdadeiro com uma probabilidade de crença subjetiva de 0,9 se a condição da classe de referência mais restrita, explicada na Def. 2-3, for garantida: isto é, a informação antecedente A deve incluir toda a informação sobre o indivíduo a que seja estatisticamente relevante para K . Afinal, mesmo que $p(Kx|Ax)$ seja alto, $p(Kx|Ax \wedge A^*x)$ pode ser baixo e, se A e A^* aplicam-se ao indivíduo a em questão, então, apenas o predicado mais restrito $A \wedge A^*$ pode ser usado como predicado antecedente de um prognóstico sobre o indivíduo a . Por exemplo, a penicilina (A) cura um resfriado grave (K) na maioria dos casos, mas, em pessoas alérgicas à penicilina (A^*), a terapia com penicilina teria consequências fatais.

Essa correlação tem consequências drásticas para a *aplicação prática*. Se, por exemplo, estudos estatísticos mostram que um método de ensino é benéfico em 80% de todos os casos e se aplicamos esse conhecimento ao aluno Pedro, então, devemos examinar cuidadosamente se não há informações mais específicas que comprometam o valor de probabilidade dessa regra geral estatística.

Pode ser, por exemplo, que o novo método de ensino beneficie apenas os aprendizes visuais, os quais representariam 80% de todos os alunos, porém pode ser que Pedro seja um aprendiz auditivo enquadrado nos 20% restantes. A aplicação dessa lei a Pedro seria, então, ilegítima e não lhe traria nenhum benefício, mas sim prejuízo na prática. Nesse sentido, todos os usuários devem ser advertidos a não aplicar prematuramente os resultados estatísticos aos casos individuais; antes devem examinar profundamente se existiriam outras características relevantes nos casos individuais capazes de alterar a probabilidade e de fornecer um quadro completamente diferente.

Se, por exemplo, um estudo estatístico mostrar que as crianças rurais são, em média, mais felizes do que as crianças urbanas e têm menos perturbações nervosas, ainda assim não seria fácil aconselhar o casal Meier a se mudar para o campo, mesmo que essa família tenha recursos financeiros para o deslocamento, pois seria necessário examinar cuidadosamente quais consequências adicionais essa mudança teria no caso específico individual dos filhos Meier. Pode ser, por exemplo, que tais filhos sejam arrancados de seu círculo social de amigos e que, por isso, a mudança para o campo lhes faça mais mal do que bem, apesar do ambiente idílico. Em particular, os políticos devem ser advertidos contra a conversão prematura das conclusões estatísticas em leis que sejam vinculativas para todas as pessoas.

9 Estatística bayesiana e bayesianismo

9.1 A intuição da verossimilhança

Podemos calcular a probabilidade de nossos resultados ou dados amostrais apenas se assumirmos que uma determinada hipótese estatística sobre toda a população está correta. Essa probabilidade é escrita como $p_H(E)$.⁶⁰ Por outro lado, as probabilidades de nossas hipóteses H , dadas nossas experiências E , são de natureza epistêmica (ver capítulo 8.5). Os métodos estatísticos de inferência e de teste são baseados na seguinte intuição básica, que chamo de *intuição de verossimilhança* (que é mais geral do que o "método da verossimilhança", veja abaixo). De acordo com esta intuição, a probabilidade inversa $p_H(E)$ é o critério básico para a plausibilidade e grau de confirmação da hipótese H dada E e para a seleção de uma hipótese entre várias hipóteses alternativas.

No exemplo mais simples, $p_H(E)$ é a probabilidade da frequência de amostragem $E =_{\text{def}} "h_n(Fx)=k"$ dada a hipótese populacional $H =_{\text{def}} "p(Fx) = r"$, que é calculada de acordo com a fórmula binomial:

$$p(h_n(Fx) = \frac{k}{n}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

⁶⁰ A notação " $p(E|H)$ " seria incorreta, porque H faz uma afirmação sobre $p:AL \rightarrow [0,1]$ e, portanto, não é um elemento da álgebra AL sobre a qual p é definido.

Como mencionado, $p_H(E)$ também é chamado de verossimilhança, mas a precisão conceitual é necessária aqui, porque as terminologias são diferentes: alguns autores chamam de $p_H(E)$ a probabilidade de E dado H (de acordo com esse modo de falar, verossimilhanças são probabilidades); outros autores falam da probabilidade de H dado E (de acordo com esse modo de falar, probabilidades são probabilidades *inversas*). Seguimos o último modo de falar aqui, e chamamos de $p_H(E)$ a probabilidade de H dado E. Além disso, devemos distinguir cuidadosamente entre a *probabilidade* estatística $p_H(E)$ e a *verossimilhança epistêmica* $P(E|H)$, que discutiremos na próxima seção.

Existem duas variantes diferentes da intuição de verossimilhança:

(i) *Método de maximização de verossimilhança*: De acordo com este método, que remonta a Fisher (1956) e Hacking (1965), quanto maior a probabilidade $p_H(E)$ de H dado E, maior será o suporte ou confirmação de uma hipótese H por um resultado de amostra E. Assim, se você tem que escolher entre hipóteses concorrentes à luz de uma determinada evidência, você escolhe a hipótese com maior probabilidade. Na situação da estatística inferencial, assume-se um resultado amostral θ_s sobre o valor de um parâmetro estatístico θ na amostra s encontrada – geralmente θ_s é a média amostral $\mu_s(X)$ de uma variável X , mas também pode ser a variância amostral $v_s(X)$. Supõe-se que a hipótese sobre o verdadeiro θ na população é a mais plausível, para a qual o resultado da amostra θ_s coincide com o valor modal, ou seja, o valor mais provável de θ_s , dado H (cf. Stegmüller 1973b, p. 84, III; Hays/ Winkler 1970, p. 318).

(ii) *Método da expectativa de verossimilhança*: Este método considera que o suporte de uma hipótese H por um resultado amostral

θ_s é maior quanto mais próximo o valor amostral observado θ_s do parâmetro θ chega da *média* ou do valor esperado $\mu(\theta_s)$ de θ_s , dado H. O método é utilizado na teoria de estimação estatística inferencial baseada em Fisher (1925) e Neyman (1937), que se baseia em *estimadores fiéis às expectativas* (cf. Bortz 1985, p. 124; Howson/Urbach 1996, cap. 10). Este método também é baseado na intuição de verossimilhança, em que a média de θ_s é baseada na distribuição das verossimilhanças de todos os valores possíveis de θ_s dados $\mu(\theta_s) = \int_{-\infty}^{+\infty} \theta_s \cdot d_H(\theta_s) d\theta_s$. Na situação de estatística inferencial, a hipótese para a qual o resultado da amostra θ_s coincide com a média de θ_s é assumida como a mais plausível.

Os dois métodos (i) e (ii) são frequentemente contrastados como concorrentes. Nosso ponto, por outro lado, é descobrir as semelhanças entre os dois métodos. Ambos os métodos são baseados na intuição de verossimilhança. O que é ainda mais importante: para todas as distribuições de probabilidade simétricas e unimodais e, portanto, também para todas as distribuições normais, os dois métodos fornecem o mesmo resultado, porque para estas distribuições o valor modal e o valor médio coincidem. E para todas essas distribuições, o principal resultado de ambos os métodos coincide com o resultado da inferência de generalização indutiva já mencionada, que hipoteticamente transfere a média amostral encontrada para a população. A intuição de verossimilhança acaba sendo, assim, o princípio *indutivo* fundamental da inferência estatística e da teoria dos testes.

Para distribuições assimétricas, como por exemplo a distribuição assimétrica à esquerda das variâncias amostrais, os dois métodos fornecem resultados diferentes, sobre cuja preferência intuitiva não existe consenso geral. No entanto, deve-se salientar que este problema só existe para conclusões sobre hipóteses pontuais, que

"ousadamente" afirmam um valor preciso do parâmetro da população, mas não para as hipóteses de intervalo de confiança mais cautelosas, que especificam um intervalo simétrico de hipóteses aceitáveis e que devem ser preferidas na situação da estatística inferencial. Estes intervalos de confiança são inequivocamente determinados mesmo para distribuições assimétricas. Recorde-se que o intervalo de confiança é definido como o intervalo daquelas hipóteses pontuais $H_r: \theta = r$, para as quais o valor amostral observado se encontra no seu intervalo de aceitação, que por sua vez é definido como o intervalo de 95% de resultados amostrais possíveis q_s com a maior verossimilhança média $pH(\theta_s)$ (cf. Jaynes 1976, p. 197). Isto mostra-nos que também o método dos intervalos de aceitação se baseia na intuição da verossimilhança; só que neste caso não se maximiza apenas a verossimilhança atual, mas sim a verossimilhança média. Como os intervalos de aceitação são conclusivamente determinados, os intervalos de confiança também são inequivocamente fixados, mesmo para distribuições assimétricas.

Na Fig. 9-1, pode-se ver que os intervalos de aceitação são conclusivamente determinados para a maioria das distribuições (mesmo quando assimétricas ou multimodais). Vejamos as três superfícies sob as duas curvas de distribuição na Fig. 9-1; simétrico à esquerda e assimétrico à direita. Cada um deles representa 70% da área total e, portanto, são do mesmo tamanho (para melhor visibilidade, consideramos intervalos de 70% em vez de 95%). A altura média destes intervalos é calculada dividindo a área (cada uma igual) pelo seu comprimento: a altura média máxima é, portanto, a área com o menor intervalo de 70%. Assim, o intervalo de 70% de todas as hipóteses mais prováveis coincide com o intervalo mais curto de 70%. Aparentemente, esse intervalo mais curto de 70% nas distribuições da

Figura 9-1 é conclusivamente determinado: se for deslocado para a esquerda ou para a direita, deve se tornar mais longo, já que a curva de distribuição se inclina para fora nos limites do intervalo.⁶¹

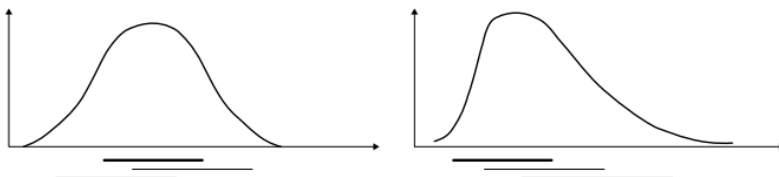


Figura 9-1: Intervalo de 70% mais curto (*negrito*) e dois intervalos mais longos de 70% [da área sob a curva] (*não negrito*) em uma distribuição de probabilidade simétrica (*esquerda*) e assimétrica (*direita*).

Os métodos intervalares de Fisher são prática estatística padrão. Várias objeções a esses métodos foram levantadas pelos bayesianos, mas há boas defesas contra eles. Por exemplo, argumenta-se que existem muitos intervalos diferentes de 95% e que é arbitrário qual deles escolhe como o intervalo de aceitação de 95% (ver Howson/Urbach 1996, p. 201). Mas isso é incorreto, porque, como acabamos de ver, o intervalo de aceitação é o intervalo de 95% das amostras com maior *probabilidade* média, e este é precisamente o intervalo mais curto de 95%. O critério do *menor* intervalo de aceitação é amplamente utilizado em estatística (cf. Hays/ Winkler 1970, p. 330). Infelizmente, a justificação dada pelos estatísticos para este critério é muitas vezes inadequada, razão pela qual Howson e Urbach duvidam que este critério tenha uma justificação sustentável. Mas, como vimos, esse critério tem a melhor justificativa imaginável,

⁶¹ Essa situação não mudaria mesmo se as distribuições na Fig. 9-1 tivessem vários picos no meio. Somente se a curva fosse uniforme ou plana, ou se tivesse vários picos pequenos na zona de rejeição de 5%, haveria mais de um intervalo mais curto de 95%.

pois a escolha do menor intervalo de aceitação maximiza a probabilidade média da hipótese em relação aos resultados de amostragem intervalar.

Outra objeção é que se obteria intervalos de aceitação diferentes se se usasse como parâmetro de teste a probabilidade de descrições amostrais *organizadas individualmente* em vez de descrições de frequência ou valores médios. No caso de uma característica binária, em vez de descrições da forma "10 de 50 indivíduos eram Fs", teríamos longas listas de descrições da forma "O indivíduo 1 era um F, o indivíduo 2 não era um F, o indivíduo 3 era um F, etc.", cuja probabilidade H seria usada como probabilidade de hipótese. Carnap falou de descrições estruturais no primeiro caso e de descrições de estado no segundo (1950, p. 71). Se a hipótese afirma uma distribuição igual, ou seja, $p_H(F) = 0,5$, então este parâmetro de amostra fornece o resultado de que cada descrição de estado tem a mesma probabilidade (ou seja, $0,5^n$ para amostras de n-elementos), o que levaria ao resultado absurdo de que a escolha de intervalos de 95% seria arbitrária e, portanto, nenhuma conclusão poderia ser tirada sobre a plausibilidade de H a partir de qualquer possível resultado amostral (cf. Howson/Urbach 1996, p. 189).

Essa observação está correta, mas a informação sobre o *arranjo dos indivíduos* em amostras com as mesmas características de frequência é estatisticamente *irrelevante* para a frequência ou média de um determinado parâmetro na população. Somente as propriedades de amostra que são *probabilisticamente dependentes do valor do parâmetro reivindicado por H* devem ser consideradas como *confirmáveis para uma hipótese*. Seidenfeld (1979) elaborou essa ideia da seguinte forma: o parâmetro teste não só deve *ser suficiente*, ou seja, conter todas as informações relevantes para H na amostra, mas

também deve ser *minimamente suficiente*, ou seja, não deve conter nenhuma informação irrelevante para H. Howson e Urbach (1996, p. 191) se opuseram ao critério de Seidenfeld de que a adição de informações irrelevantes não deveria fazer diferença - mas esse não é o caso quando se trata de intervalos de aceitação, uma vez que estes podem ser arbitrariamente distorcidos, esticados ou estendidos pela adição de informações irrelevantes.

Outra objeção de Howson e Urbach (1996, p. 186) é que os métodos estatísticos podem ser *dependentes da linguagem*: seu resultado pode depender de como e quão finamente os resultados de um experimento aleatório são classificados. Howson e Urbach descrevem um caso em que o resultado de um teste χ^2 para verificar a regularidade de um cubo muda quando a partição mais fina $\{1,2,3,4,5,6\}$ é “engrossada” para $\{1\vee 2, 3\vee 4, 5\vee 6\}$: um desvio previamente significativo da distribuição uniforme em uma amostra pode se tornar insignificante após o “engrossamento”. – Este tipo de dependência linguística é um problema geral de todos os métodos indutivos (probabilísticos ou não probabilísticos) e, portanto, não é particularmente surpreendente. Veremos na seção 9.3 que os métodos da estatística bayesiana são ainda mais dependentes da linguagem, uma vez que se baseiam no pressuposto do princípio da indiferença dependente da linguagem.

Felizmente, a dependência linguística não resulta em arbitrariedade subjetiva, porque há critérios que possibilitam, pelo menos em certos (não todos) casos, fazer uma escolha racional a partir de sistemas linguísticos concorrentes (cf. Schurz 2013b, seção 5.11.3). Por exemplo, um sistema linguístico mais fino é geralmente sempre preferível a um sistema linguístico mais grosseiro. Isso é especialmente verdadeiro para o exemplo acima de Howson e Urbach:

pode-se mostrar que diferenças significativas detectadas por engrossamento, mas não por refinamento da descrição, podem ser perdidas. No máximo, os refinamentos podem revelar novas diferenças significativas.

Uma extensão bem conhecida do método de teste de Fisher é o método de Neyman e Pearson, que, além da "hipótese nula", por exemplo, $\mu = 0,1$, também pressupõe uma "hipótese alternativa", por exemplo, $\mu = 0,2$ (Hays/Winkler 1970, p. 401; Howson/ Urbach 1996, p. 195). A falácia de rejeitar uma hipótese nula verdadeira (isto é, aceitação da falsa hipótese alternativa) também é chamada de *erro* α , e a falácia de aceitar uma falsa hipótese nula (rejeitando a hipótese alternativa verdadeira) é chamada de *erro* β . O coeficiente de rejeição de Fisher de 5% nada mais é do que a probabilidade do erro α . Neyman e Pearson sugeriram que, entre todos os intervalos de rejeição com erro α de 5%, deve-se escolher aquele que minimiza o erro β . O problema com o método de Neyman-Pearson é que ele só funciona se as hipóteses nulas e alternativas forem hipóteses pontuais, de modo que o erro β pode ser calculado usando a distribuição amostral assumindo a hipótese alternativa. Se a hipótese alternativa é um intervalo de hipótese, como no caso dos métodos de teste de Fisher, então o método de Neyman-Pearson não é aplicável. No entanto, Neyman-Pearson propuseram uma extensão de seu método na forma do chamado teste UMPU ("uniformemente mais poderoso e imparcial"; ver Howson/Urbach 1996, p. 216-8).

Um teste UMPU seleciona o intervalo de aceitação de 95% para a hipótese nula, o que minimiza a falácia β para todas as hipóteses de pontos alternativos e, ao mesmo tempo, garante a condição de racionalidade mínima de que a probabilidade de rejeição da hipótese

nula é menor no caso de sua verdade do que no caso de sua falsidade⁶². O intervalo de aceitação assim escolhido coincide comprovadamente com o intervalo de aceitação mais curto de 95 %, e a aplicação do método de Neyman-Pearson às hipóteses intervalares remete, assim, ao método de verificação de Fisher.

9.2 Justificação bayesiana da intuição de verossimilhança

Em resumo, os métodos de teste estatístico e inferência são bem fundamentados, *desde que a intuição de verossimilhança seja aceita*. O problema filosófico dos métodos estatísticos é mais profundo e diz respeito à *justificação* da intuição de verossimilhança: *Por que a probabilidade inversa $p_H(E)$ deve ser usada como medida da plausibilidade da hipótese H dada a evidência E?* Dentro da teoria estatística, não há resposta para essa pergunta. Isso porque a plausibilidade da hipótese H dada a evidência E é uma probabilidade subjetiva-epistêmica $P(H|E)$, sobre a qual a teoria estatística não faz afirmações.

No entanto, a teoria da probabilidade epistêmica tem uma resposta para a questão de como justificar a intuição da verossimilhança. A base desta resposta é fornecida pelo teorema bayesiano (teorema 3-3, TB5) e pelo princípio de coordenação estatística StK (Def. 7-2), que dá a probabilidade epistêmica $P(E|H)$ com a verossimilhança estatística $p_H(E)$. O StK juntamente com o teorema de Bayes dá:

⁶² Este é o caso sse $\text{Poder} =_{\text{def}} 1 - P(\text{erro } \beta) > P(\text{erro } \beta)$ (*ibid.*, p. 216).

$$(9-1) P(H|E) = P(E|H) \cdot P(H)/P(E) \quad (\text{de acordo com a regra de Bayes})$$

$$= p_H(E) \cdot P(H)/P(E) \quad (\text{de acordo com o StK}).$$

Em palavras, o grau de crença da hipótese estatística H, dado o resultado amostral E, é igual à probabilidade estatística de E assumindo H, multiplicada pela razão entre a probabilidade inicial de H e a probabilidade inicial de E.

Chamamos de " $p_H(E)$ " a *verossimilhança estatística* e " $P(E|H)$ " é dada a *probabilidade epistêmica* de H dado E. Estritamente falando, no caso estatístico temos que escrever $p_H(E(x))$, uma vez que $p_H(E(x))$ é o limite de abundância de uma propriedade $E(-)$ de amostras *variáveis* x (em uma sequência infinita de tais amostras), enquanto no caso epistêmico temos $P(E(a)|H)$, tratando-se da presença do evento E numa amostra individual e de fato sorteada a.

É uma característica da estatística bayesiana que certas probabilidades iniciais (ou "priors") de hipóteses devem sempre ser assumidas. Essas probabilidades iniciais constituem um elemento irreduzível e inevitavelmente subjetivo do bayesianismo, uma vez que refletem os graus de crença do sujeito em estado de ignorância. A equação (9-1) também contém a probabilidade inicial dada da experiência $P(E)$ – mas essa probabilidade inicial pode ser eliminada. A maneira mais fácil de eliminar $P(E)$ é limitar-se a avaliações comparativas de hipóteses. Se H_1, H_2 são duas hipóteses estatísticas concorrentes, então a razão de suas *probabilidades finais* em relação à evidência dada (seus "*a posteriori*") é determinada unicamente pela relação de suas probabilidades e seus "*a priori*", pois $P(E)$ é assim truncada:

$$(9-2) \frac{P(H_1|E)}{P(H_2|E)} = \frac{P(E|H_1)}{P(E|H_2)} \cdot \frac{P(H_1)}{P(H_2)}$$

No caso especial em que as hipóteses comparadas têm a mesma probabilidade inicial, $P(H_1) = P(H_2)$, a razão de suas probabilidades finais coincide com a razão de suas probabilidades, a chamada "razão de verossimilhança", $P(E|H_1)/P(E|H_2)$. De forma mais geral, a equação (9-2) implica que uma máxima verossimilhança sob determinadas hipóteses alternativas H_1, \dots, H_n é um indicador da hipótese subjetivamente mais provável se essas hipóteses tiverem a mesma probabilidade inicial. A suposição de probabilidades iniciais iguais para hipóteses concorrentes também é chamada de princípio da *indiferença*: na ausência de conhecimento adicional, as possibilidades concorrentes são assumidas como igualmente prováveis. Juntamente com a equação (9-2), o princípio da indiferença fornece assim a seguinte justificativa para a intuição de verossimilhança:

(Teorema 9-1) Justificativa bayesiana da intuição de verossimilhança: Assumindo o princípio da indiferença, o nível de probabilidade de H dado E é um indicador da probabilidade epistêmica de H dado E.

Se alguém quiser um cálculo numérico de probabilidades de hipótese, então pode-se eliminar a probabilidade inicial de E especificando uma partição (isto é, um conjunto de hipóteses mutuamente exclusivas e exaustivas). Se estamos falando sobre a probabilidade estatística $p(Fx)$ de uma característica binária Fx , então esta partição consiste em todas as hipóteses H_r da forma " $p(Fx) = r$ " com $r \in [0,1]$. Neste caso, assume-se uma distribuição subjetiva de densidade de saída $D(H_r)$ sobre todas essas hipóteses.

Alternativamente, pode-se assumir uma partição discreta de hipóteses alternativas $\{H_1, \dots, H_n\}$, por exemplo, por arredondamento ou “intervalação”, ou devido a conhecimentos prévios. Assim, a probabilidade inicial de E é calculada da seguinte forma (cf. Hays/Winkler 1970, p. 233, 461):⁶³

$$(9-3) P(E) = \int_0^1 p_{H_r}(E) \cdot D(H_r) dr. \quad \text{No caso discreto: } P(E) = \sum_{i=1}^n p_{H_i}(E) \cdot P(H_i)$$

Usando a Eq. (9-1) segue-se:

(Teorema 9-2) *Distribuição de Probabilidade Final:*

$$D(H_q|E) = p_{H_q}(E) \cdot D(H_q) / \int_0^1 p_{H_r}(E) \cdot D(H_r) dr. \quad (q \in [0,1])$$

$$\text{Caso discreto: } P(H_q|E) = p_{H_q}(E) \cdot D(H_q) / \sum_{i=1}^n p_{H_i}(E) \cdot P(H_i) \quad (1 \leq q \leq n)$$

Em palavras: A probabilidade final de uma hipótese H dada uma evidência E é dada como sua probabilidade inicial multiplicada por sua probabilidade dada E dividida pela probabilidade epistêmica de E calculada usando o StK (Teorema 7-2).

O cálculo das probabilidades finais epistêmicas das hipóteses por meio de suas probabilidades estatísticas e probabilidades iniciais é o passo central pelo qual a estatística bayesiana vai além da estatística comum. Se a distribuição inicial $D(H_r)$ é uma distribuição igual, então a distribuição final $D(H_r|E)$ seu máximo está exatamente na hipótese H_r , cuja probabilidade populacional alegada r corresponde à

⁶³ Em (9-3) está incluída a suposição mencionada no Teorema 7-3(3)(i) de que com $P=1$ as frequências convergem para limites.

frequência amostral relatada por E . Dessa forma, obtemos novamente a justificativa da intuição da verossimilhança com a ajuda do princípio da indiferença.

9.3 Bayesianismo objetivo e o princípio da indiferença: inferência indutiva I

O bayesiano objetivo difere do bayesiano subjetivo em particular na medida em que assume o princípio da indiferença e, dessa forma, busca obter probabilidades iniciais intersubjetivas ou 'objetivas' de hipóteses.⁶⁴ De acordo com este princípio, a distribuição de densidade inicial $D(H_r)$ é uniforme, o que (devido à condição de normalização $\int_0^1 D(r)dr = 1$) implica que $D(r) = 1$ se aplica para todos $r \in [0,1]$. Disto e do Teorema 9-2 segue-se imediatamente que a densidade final de H_r tem seu máximo exatamente naquele valor de r que maximiza a probabilidade $p_{H_r}(E)$ (uma vez que $D(H)$ e a integral são independentes da escolha de r). Mais importante, agora podemos finalmente calcular valores *numéricos* da probabilidade final de hipóteses, o que é uma preocupação importante para os bayesianos objetivos. Consideremos novamente todas as hipóteses possíveis relativas à probabilidade estatística de uma característica binária $F: H_r =_{\text{def}} "p(F)=r"$ (para $r \in [0,1]$). Seja $Fa_n^k =_{\text{def}} Fa_1 \wedge \dots \wedge Fa_k \wedge \neg Fa_{k+1} \wedge \dots \wedge \neg Fa_n$ ($1 \leq k \leq n$) para a descrição completa do estado de uma amostra de n elementos de k F s e $(n - k)$ $\neg F$ s. Dada a independência estatística, o limite de frequência de cada descrição de estado é calculado como

⁶⁴ Williamson (2010, p. 16, 28) fala do princípio do "equivoco".

estado de k de n Fs como antecedente. No entanto, só se obtém (c) se for uma característica binária, ou seja, a suposição de distribuição uniforme via $\{F, \neg F\}$ é assumida. Se F é uma característica de um atributo graduado mais de 2 níveis (como classes inferiores, médias e superiores), a regra c^* de Carnap, em vez de (c), (9-6) (veja abaixo), se aplica, cujo resultado depende da "largura" lógica da característica F . A regra sucessória de Laplace é, portanto, dependente da linguagem.

A afirmação (d) nos diz a distribuição de probabilidade final sobre as hipóteses possíveis da forma $p(Fx)=r$, dado que k de n indivíduos da amostra eram Fs. Em contraste com a distribuição binomial, r é agora a variável; k e n são as constantes. Esta é a chamada distribuição β (ou seja, a distribuição β_{n+2}^{k+1}). Ela tem o seu máximo onde a verossimilhança também tem o seu máximo, nomeadamente acima da frequência de amostragem k/n . A média deles é $(k+1)/(n+2)$; a distribuição é, portanto, concentrada à esquerda (ou concentrada à direita) se $p(F) < 1/2$ (ou, respectivamente $> 1/2$). A distribuição β , como a distribuição de verossimilhança na Figura 3-3, torna-se cada vez mais acentuada à medida que n aumenta. As distribuições β são uma ferramenta importante na estatística Bayesiana (Hays/Winkler 1970, p. 233 e seg.).

Esses resultados do teorema 9-3 parecem ótimos, mas infelizmente o princípio da indiferença está fundado em uma argumentação frágil. Seus resultados são *altamente* dependentes da linguagem (cf. Gillies 2000, p. 37-48). Considere, por exemplo, uma distribuição de saída uniforme sobre os valores de frequência desconhecidos (μ) de um determinado tipo de radiação eletromagnética (por exemplo, a luz emitida pelo sódio). O comprimento de onda (λ) de uma radiação é idêntico à velocidade da luz (c) dividida pela frequência da radiação, $\lambda = c/\mu$. Se agora se transforma uma distribuição uniforme sobre as frequências $\mu \in [0,$

μ_{\max}] em uma distribuição ao longo dos comprimentos de onda λ , não se obtém uma distribuição uniformemente distribuída, mas uma distribuição negativa-exponencialmente decrescente, como mostrado na Figura 9-2.

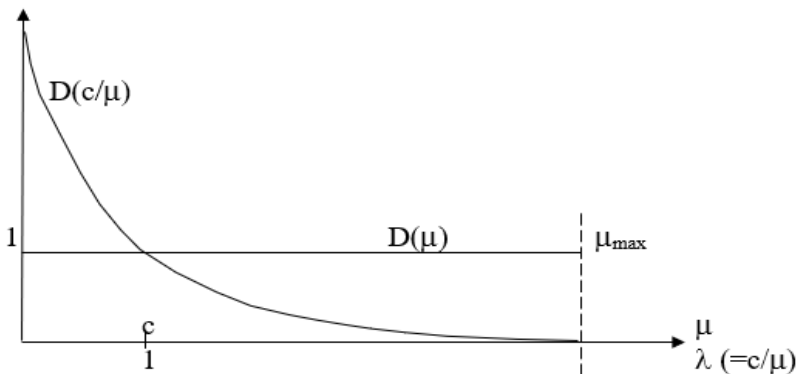


Figura 9-2: Dependência da linguagem em distribuições iguais. Uma distribuição uniforme de densidade sobre μ (frequência) leva a uma distribuição não uniforme sobre λ (comprimento de onda).

Semelhantes 'paradoxos' do princípio da indiferença já foram discutidos por Keynes (1921, seção 4). Keynes era um defensor do princípio da indiferença; propôs um aperfeiçoamento dele, o que, no entanto, não se aplica ao nosso contraexemplo (Gillies, 2000, p. 43; Howson e Urbach, 1996, p. 61). A conclusão amplamente aceita dos fatos do caso é a seguinte: *nenhuma distribuição inicial é sem preconceitos ou sem informação*, nem mesmo distribuição igualitária. Esta conclusão é ainda apoiada pelas seguintes considerações:

(1.) Consideremos (como no final da seção 9.1) em vez de todas as hipóteses de frequência possíveis todas as descrições de estado

possíveis (totalmente organizadas) da população⁶⁶, ou semanticamente o conjunto de todos os modelos possíveis. Se a distribuição de probabilidade for agora uma distribuição uniforme não sobre as hipóteses de frequência, mas sobre essas descrições de estado, então a aprendizagem indutiva através da experiência torna-se impossível e o resultado é $P(Fa_{n+1} | \pm Fa_1 \wedge \dots \wedge \pm Fa_n) = 1/2$ para qualquer descrição de estado possível $\pm Fa_1 \wedge \dots \wedge \pm Fa_n$ de uma amostra de n elementos (“±” para “não negado” ou “negado”). Porque há exatamente metade das descrições de estado que verificam Fa_{n+1} e $\pm Fa_1 \wedge \dots \wedge \pm Fa_n$ do que aquelas que apenas verificam $\pm Fa_1 \wedge \dots \wedge \pm Fa_n$. Pode-se objetar (como no Capítulo 9.1) que o *arranjo* de indivíduos nas mesmas frequências é irrelevante para a determinação indutiva de probabilidades estatísticas. Essa objeção está correta, mas ao mesmo tempo mostra que a suposição de uma distribuição igual sobre os limites de frequência em vez de sobre as descrições de estado não é incondicional, mas faz fortes suposições indutivas.

(2.) Se a distribuição inicial sobre todas as hipóteses pontuais estatísticas de uma característica binária for uma função contínua e de outra forma arbitrária (não necessariamente uniforme), então a probabilidade epistêmica inicial $P(H_r)$ de cada hipótese pontual H_r (conforme explicado em conexão com a Def. 6- 2) é zero. Portanto, para uma função P contínua, a probabilidade inicial de cada hipótese estrita, $P(\forall x Fx)$, também é zero, o que (como sabemos) torna impossível a revisão através da experiência. Uma distribuição contínua do produto implica, portanto, uma tendência extrema contra todas as hipóteses estritas. Por outro lado, a integral sobre um ponto só pode assumir um valor maior que zero se a densidade sobre

⁶⁶ No caso de domínios infinitos de indivíduos, uma descrição de estado corresponde a um conjunto infinito de sentenças básicas.

este ponto for infinitamente grande. Para obter $P(\forall xFx) > 0$, a densidade de probabilidade sobre a hipótese pontual $p(Fx)=1$ teria que ser infinitamente grande, o que significa que a distribuição de densidade conteria agora um viés extremo em relação à hipótese pontual $p(Fx)=1$ (ver Earman 1992, p. 87-94). A consequência inevitável disto é que nenhuma distribuição inicial pode estar isenta de preconceitos ou preconceitos em todos os aspectos.⁶⁷

O problema da probabilidade zero de hipóteses universais já era reconhecido por Carnap e Popper. Carnap (1950, p. 571) contornou o problema usando seu método de confirmação de instância: de acordo com isso, o grau de confirmação de $\forall xFx$ é dado por Fa_1, \dots, Fa_n com a probabilidade de uma previsão singular, $P(Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n)$. Popper, por outro lado, tirou a conclusão injustificada do problema da probabilidade zero (1935/2002, novo Apêndice vii*) de que a confirmação indutivo-probabilística era impossível. Com base em R. Jeffrey (1983), Earman (1992, p. 91) mostrou que uma probabilidade inicial $P(\forall xFx)$ maior que zero só é possível sob a seguinte suposição:

(9-4) Conteúdo Indutivo de " $P(\forall xFx) > 0$ ": A afirmação " $P(\forall xFx) > 0$ " pressupõe que as probabilidades condicionais $P(Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n)$ convergem rapidamente para 1 quando n tende ao infinito. Uma condição suficiente para isso é que P seja σ -aditiva e exista uma constante $c < 1$ e um número natural $k \in \mathbb{N}$ tal que $\forall n \geq k$ vale:

⁶⁷ Como também a seção 3.4 explicou que a não- σ -aditividade para espaços de possibilidades infinitas contáveis pode ser vista como uma instância deste problema: se a distribuição sobre os números individuais de uma urna numérica infinita for uniformemente distribuída (sem viés), então para cada número a probabilidade inicial de extrair este número é zero.

$P(\neg Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n) / P(\neg Fa_n|Fa_1 \wedge \dots \wedge Fa_{n-1}) \leq c$ (Para prova, ver Apêndice 10.3.15).

Por exemplo, se substituirmos o valor de $P(Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n)$ como $(n+1)/(n+2)$ no (Teorema 9-3c) usando a regra sequencial de Laplace, então $P(Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n)$ converge para 1 quando $n \rightarrow \infty$, mas não rápido o suficiente para evitar $P(\forall xFx) = 0$.⁶⁸ A condição $P(\forall xFx) > 0$ requer, portanto, suposições indutivas ainda mais fortes do que a regra consequencial de Laplace.

A "teoria lógica da probabilidade" de Carnap caracteriza-se pelo fato de que o princípio da indiferença é adicionado aos princípios da regularidade e da permutabilidade. Carnap (1950) desenvolveu seu campo de probabilidade sobre a álgebra de uma lógica de predicados monádicos (de um dígito) com finitamente muitos predicados básicos F_1, \dots, F_n e finitamente muitos nomes padrão para indivíduos a_1, \dots, a_N ; ele incluiu o caso de um número infinito de indivíduos considerando o valor limite $N \rightarrow \infty$. Em tal linguagem, existem muitos *predicados de estado* ou "predicados Q" $\mu =_{\text{def}} 2^n$ que têm a forma $Q_j x =_{\text{def}} \pm F_1 x \wedge \dots \wedge \pm F_n x$ ("±" para "não negado" ou "negado"). Uma descrição de estado no sentido de Carnap indica para cada indivíduo a qual predicado Q ele pertence e, portanto, tem a forma $Q_{j_1} a_1 \wedge \dots \wedge Q_{j_N} a_N$; existem $\mu^N = 2^{n \cdot N}$ tais descrições de estado. Em 1971, Carnap também permitiu mais de famílias de predicados com dois valores, que são conjuntos de

⁶⁸ Porque $\lim_{n \rightarrow \infty} P(\bigwedge_{1 \leq i \leq n} Fa_i) = \lim_{n \rightarrow \infty} \prod_{0 \leq i \leq n} P(Fa_{i+1}|Fa_1 \wedge \dots \wedge Fa_i) = \lim_{n \rightarrow \infty} (1/2) \cdot (2/3) \cdot (3/4) \cdot \dots \cdot (n+1/n+2) = \lim_{n \rightarrow \infty} (1/n+2) = 0$. Como $P(\forall xFx) \leq \lim_{n \rightarrow \infty} P(\bigwedge_{1 \leq i \leq n} Fa_i)$ é válido independente da σ -Aditividade (Schurz/Leitgeb 2008, Lemma 1), segue-se $P(\forall xFx) = 0$.

atributos disjuntos e exaustivos $\mathcal{M}_1 = \{M_{1,i}, \dots, M_{1,n}\}$, por ex. a família de predicados de cores.⁶⁹

Carnap inicialmente assumiu a distribuição uniforme sobre os predicados de estado (“simetria de P em relação aos predicados Q”): $P(Q_j a_i) = 1/\mu$ para todos j e i. Ele deixou aberta a distribuição P sobre as frequências de tais predicados Q. No entanto, ele adicionou o axioma da irrelevância da predição, segundo o qual a probabilidade condicional de uma característica (possivelmente complexa) M para um novo indivíduo a_{n+1} , dada uma descrição de estado de n indivíduos com k Ms, depende apenas da ocorrência de M para esses n indivíduos, mas não de outras características: $P(M a_{n+1} | Q_j a_1 \wedge \dots \wedge Q_j a_n) = P(M a_{n+1} | M a_n^k)$. Além disso, $P(M a_{n+1} | M a_n^k)$ não deve depender da escolha específica do predicado M, mas apenas de k e n e da amplitude lógica do traço M, w_M/μ ; onde w_M é o número de M verificando Q-predicados. Outros axiomas do sistema de Carnap revelaram-se desnecessários porque seguiram os outros axiomas, de modo que podemos limitar-nos aqui a estes dois axiomas de Carnap adicionais.⁷⁰

Carnap mostra que o sistema resultante ainda admite um continuum de possíveis sistemas indutivos segundo os quais a probabilidade condicional $P(M a_{n+1} | M a_n^k)$ é uma média ponderada entre a frequência de amostragem observada de M, $h_n(Mx) = k/n$, e a extensão lógica do atributo w_M/μ (cf. Carnap 1959, p. 218):

⁶⁹ Além disso, ele substituiu descrições de estado como argumentos da função de probabilidade por modelos (em linguagens infinitas ambos os termos são equivalentes).

⁷⁰ Carnap assumiu a relevância da instância como um axioma independente, o que, como mostrado no Teorema 7-4, decorre da permutabilidade e regularidade (ver Kutschera 1972, p. 129, n. 13).

(9-5) λ -contínuo de Carnap:

$$P(\text{Ma}_{n+1} | \text{Ma}_n^k) = \frac{k}{n} \cdot \frac{n}{n+\lambda} + \frac{w_M}{n} \cdot \frac{\lambda}{n+\lambda} = \frac{k+\lambda \cdot \left(\frac{w_M}{\mu}\right)}{n+\lambda} \quad \text{mit } \lambda \in [0, \infty).$$

(k/n) é o fator de probabilidade empírico e (w_M/μ) o fator de probabilidade 'lógico'. A razão entre o parâmetro $\lambda \in [0, \infty)$ e o tamanho da amostra n determina o peso do fator de probabilidade lógica. Se $\lambda=0$, então obtemos a regra de indução não adulterada ou “regra direta”, que, conforme explicado, está subjacente aos procedimentos estatístico-indutivos. Se $\lambda = \infty$, então a distribuição de probabilidade torna-se independente da experiência e não se afasta da distribuição uniforme. Enquanto λ for finito, o fator empírico prevalece para amostras grandes ($n \gg \lambda$), de modo que as probabilidades indutivas de Carnap no longo prazo ($n \rightarrow \infty$) levam ao mesmo resultado que a “regra direta” indutiva; mas no curto prazo podem variar muito. O sistema preferido de Carnap, o sistema c^* , define $\lambda = \mu$, o que leva a uma distribuição uniforme sobre as descrições estruturais ou hipóteses de frequência estatística, com consequências para as características binárias listadas no Teorema (9-3). Para atributos graduados ou complexos n vezes M obtém-se:

(9-6) Regra c^* de Carnap: $P(\text{Ma}_{n+1} | \text{Ma}_n^k) = (k+w_M)/(n+\mu)$.

No caso especial de características binárias, ou seja, $w_M = 1$ e $\mu = 2$, retorna-se à regra de sucessão de Laplace no Teorema 9-3(c), $(k+1)/(n+2)$.⁷¹

⁷¹ Para a prova de (12-6), veja Apêndice 10.3.14; Para a prova de (12-5) e (12-6), veja Carnap (1950), §110, Carnap (1980, cap. 19) e Kutschera (1972, p. 131-5, 532-7).

Finalmente, o problema das probabilidades 'lógicas' pode ser demonstrado usando a regra c^* de Carnap. Surge a questão de por que se deve incluir o fator "lógico" na estimativa de frequências empíricas. Suponha que observamos 10 nativos do sexo masculino de uma tribo, todos tatuados de vermelho. Em um sistema de linguagem com o predicado básico binário "(não) tatuado" e um atributo de cor assumido de 6 valores (vermelho, verde, azul, amarelo, branco, preto), isso resulta em 12 predicados de estado possíveis. De acordo com o sistema c^* de Carnap, deveríamos estimar a probabilidade indutiva de que o próximo nativo do sexo masculino também tenha uma tatuagem vermelha não como aproximadamente 1 (conforme a "regra da reta" indutiva), mas como $10+1/10+12 = 11/22 = 1/2$. Isso parece estranho. Carnap (1950, p. 227, 569) defende seu sistema apontando que, para tamanhos de amostra muito pequenos, uma mistura de frequência observada com distribuição uniforme de Laplace seria útil. A resposta para isso é que se o tamanho da amostra for muito pequeno, em vez de "adivinhar" usando suposições dependentes da linguagem e, em última análise, arbitrárias de distribuição igual, é melhor abster-se de julgar e ampliar a amostra. Outros argumentos demonstrando as vantagens da regra indutiva direta sobre misturas dela com probabilidades a priori foram apresentados por Salmon (1974) e Rescher (1987, p. 115).

Hintikka (1965) propôs um sistema indutivo no qual (ao contrário de Carnap) hipóteses universais sobre um número infinito de indivíduos são atribuídas a uma probabilidade maior que zero. Todas as hipóteses possíveis e completas estritamente quantificadas de uma lógica de predicados monádicos recebem a mesma probabilidade inicial. Essas hipóteses assumem a forma de

$$(9-7) \exists x Q_{i_1} x \wedge \dots \wedge \exists x Q_{i_k} x \wedge \forall x (Q_{i_1} x \vee \dots \vee Q_{i_k} x) \text{ for } \{i_1, \dots, i_k\} \subseteq \{1, \dots, \mu\}.$$

Em palavras: Dos μ predicados Q possíveis, tal e tal são predicados Q e apenas estes são instanciados (ou seja, a disjunção destes se aplica a todos os indivíduos).

Declarações de forma (9-7) também são chamadas de constituintes monádicos (Niiniluoto 1987, p. 51). Quanto menor k , maior o grau de regularidade no mundo descrito pelo constituinte. Hintikka (1966) ampliou esse sistema incorporando inferências estatísticas de frequência na forma de um contínuo bidimensional. Finalmente, como mostrou Jamison (1970), tanto o sistema indutivo de Carnap quanto o de Hintikka são casos especiais de uma abordagem bayesiana geral, na qual certas possibilidades de hipóteses com certas distribuições de produto são assumidas.

9.4 Probabilidades de hipóteses sem o princípio da indiferença?

A discussão até agora deixa aqueles que esperavam que o bayesianismo fosse uma justificativa funcional do princípio da verossimilhança estatística e, portanto, uma justificativa para o raciocínio indutivo em uma espécie de dilema. Por um lado, a verossimilhança aparentemente só pode ser justificada em termos bayesianos com a ajuda do princípio da indiferença. Por outro lado, o próprio princípio da indiferença está sujeito a sérios problemas.

Seria, portanto, altamente desejável justificar o princípio da verossimilhança e a inferência indutiva de uma probabilidade alta para uma probabilidade de hipótese alta por uma suposição mais

fraca, por exemplo, apenas pela permutabilidade ou pelo StK, ou mesmo por suposições ainda mais fracas. O chamado *argumento da simetria* de Williams (1947), Stove (1986) e outros (cf. Vickers 2010, § 7) – mais recentemente elegantemente elaborado por Williamson (2013) – tenta fazer exatamente isso. Apresentamos aqui uma discussão crítica desse argumento.

A probabilidade (estatística) de que a frequência h_n de um traço em uma amostra suficientemente grande seja próxima da probabilidade estatística p é muito alta de acordo com a lei dos grandes números, e isso decorre apenas dos axiomas básicos da probabilidade. É aí que entra o argumento da simetria: se h_n está próximo de p , p está próximo de h_n ; temos esse argumento de simetria analiticamente válido na seção 8.2 no cálculo dos intervalos de confiança. Logo, conclui o argumento, a probabilidade estatística (hipotética) p aproxima-se *dessa* frequência amostral h_n com o valor *atualmente* observado k/n . Como vários autores têm argumentado, há duas falhas principais nesse argumento. Antes de detalhá-las, vamos reconstruir o argumento de forma logicamente simples.

A seguir, $h(F;x^n)$ representa a frequência da característica binária F em uma *amostra* de n -elementos variáveis x^n , e $h(F;s^n)$ representa a frequência de F em uma *dada* amostra de n -elementos. Além disso, $p(F)$ representa a probabilidade estatística desconhecida de F na população. Usando os métodos de intervalo de aceitação discutidos, podemos calcular um intervalo de aceitação $\pm a_n$, que depende de n , de modo que $h(F;x^n)$ está em $[r \pm a_n]$ com uma alta, digamos 95% de probabilidade, então:

$$9-8) \text{ Probabilidade } (h(F;x^n) \in [p(F) \pm a_n]) = 0,95.$$

O argumento de Williams-Stove em sua forma original conclui disso por simetria:

$$(9-9) \text{ Probabilidade}(p(F) \in [h(F:x^n) \pm a_n]) = 0,95,$$

e ainda, com a ajuda do certo conhecimento observacional $h(F:s^n) = q$, em

$$(9-10) \text{ Probabilidade } (p(F) \in [q \pm a_n]) = 0,95.$$

Em palavras, a probabilidade de que a verdadeira probabilidade estatística de F não se desvie em mais de "a" do valor amostral encontrado q é de 95%.

Os defensores do argumento de Williams-Stove também argumentam que essa "derivação" resulta unicamente do axioma básico da probabilidade. Seria assim obtida uma hipótese indutiva "do nada"?

Não é o caso. A primeira falha no argumento é que ele não distingue entre probabilidade estatística e epistêmica. A probabilidade em (9-8) e (9-9) é uma probabilidade estatística, onde a variável x^n percorre qualquer amostra aleatória de tamanho n. Então temos que escrever:

$$(9-8^*) p(h(F:x^n) \in [p(F) \pm a_n]) = 0,95, \text{ e}$$

$$(9-9^*) p(p(F) \in [h(F:x^n) \pm a_n]) = 0,95.$$

Somente se essa probabilidade estatística for usada, (9-8*) e (9-9*) seguem os axiomas básicos para p. (9-9*) é algo diferente do que os proponentes do argumento (9-9) pretendem. (9-9*) diz a mesma

coisa que (9-8*): que em uma sequência aleatória de amostras de n -elementos, a frequência daquelas amostras x^n cuja frequência F não se desvia mais do que a de $p(F)$ é de 95%. Nesse sentido, Heys/Winkler (1970, p. 328) e Howson/Urbach (1996, p. 239) contestaram o argumento de Williams-Stove de que a probabilidade em (9-9*), apesar da inversão simétrica, ainda permanece *uma probabilidade estatística* de amostragem, mas não uma probabilidade de hipótese em relação a uma amostra individual, porque tais probabilidades de hipótese não são estatísticas, mas de natureza epistêmica.

O passo crucial que é ignorado na transição de (9-8) para (9-9) no argumento original de Williams-Stove é a transferência da probabilidade estatística para o caso individual, nossa amostra s^n atualmente observada, de acordo com o princípio de coordenação estatística (StK) para frequências amostrais (Def. 7-2)(c). Portanto, precisamos do seguinte passo intermediário:

$$(9-9a^*) P(p(F) \in [h(F:s^n) \pm a_n]) = 0,95 \text{ (usando o StK).}$$

Isso mostra que a intenção original de Williams e Stove de trazer uma justificativa "dedutiva" livre de circularidade da confiabilidade do raciocínio indutivo não funciona. Isso porque, como sabemos, o StK para amostragem ou a inferência indutiva de especialização contém o pressuposto indutivo de que a amostra atualmente observada é representativa da população, de modo que se pode inferir a amostra a partir da tendência estatística na população. É importante lembrar (Capítulo 4.4) que não apenas a generalização indutiva, mas também a inferência de especialização indutiva faz uma suposição de uniformidade indutiva.

Em sua versão revisada do argumento da simetria, Williamson (2013, p. 304-308) evita esse primeiro erro do argumento, mas deixa claro desde o início que, além dos axiomas básicos, ele se baseia em uma aplicação do StK. A questão é se o argumento está correto após esta limpeza. Se fosse esse o caso, o argumento ainda teria grande valor para o Bayesianismo objetivo. Porque então o argumento mostraria que as probabilidades de hipóteses epistêmicas podem ser obtidas apenas com a ajuda de probabilidades estatísticas e do StK, sem o princípio da indiferença problemático ou outras probabilidades de saída preferidas, esse ponto também é enfatizado pelo próprio Williamson (2013, p. 310). É aqui que entra em jogo a segunda falha principal do argumento, a meu ver. A aplicação do StK em (9-9*) está correta, na medida em que se assume que o agente *não sabe nada* sobre a amostra individual s^n , exceto que ela provém de uma população com um limiar de frequência desconhecido, mas constante, $p(F)$. Nesse caso, o agente faz uma previsão razoável sobre a amostra. Será diferente se o agente já tiver experiência com essa amostra e, principalmente, se souber sua frequência da característica. Lembremos a formulação do StK na Def. 7-2: o StK só é plenamente útil desde que ainda não se tenha nenhuma experiência dos indivíduos aos quais o StK é aplicado. Caso contrário, podem surgir inconsistências e até contradições. Mas é exatamente isso que é feito na etapa final do argumento, na qual se acrescenta o conhecimento observacional adicional certo, segundo o qual a frequência de amostragem $h(F:s^n) = q$. As etapas que faltam devem ser inseridas da seguinte forma:

$$(9-10^*) P(p(F) \in [h(F:s^n) \pm a_n] | h(F:s^n) = q) = 0,95 \text{ (de (9-9a}^*))$$

$$(9-11^*) P(h(F:s^n)=q) = 1 \text{ (suposição)}$$

(9-12*) De (9-10*)+(9-11*) segue probabilisticamente: $P(p(F) \in [q \pm a_n]) = 0,95$.

A derivação incompleta (9-8,9,10) tornou-se assim a derivação completa (9-8*,9*,9a*,10*,11*,12*). Mostramos agora que o passo (9-10*) não é geralmente permitido, mas apenas se a distribuição inicial sobre as hipóteses for uma distribuição aproximada igual. Maher (1996, p. 426) já apontava que o passo de (9-9*) para (9-10*) não é admissível se certas hipóteses forem muito mais prováveis do que outras, de modo que a amostra sorteada parece ser um caso "outlier" ou excepcional. Nesse caso, a informação $h(s^n)=q$ pode diminuir a probabilidade da afirmação $p \in [q \pm a_n]$ muito abaixo do valor de 0,95. Suponhamos que estejamos convencidos de que uma moeda é aproximadamente justa ($P(p=1/2)$ é alta). A probabilidade estatística de que em cada 100 arremessos a frequência de caras não se desvie mais de 8% da probabilidade real é de 95% (ver capítulo 8.1). Portanto, de acordo com o argumento da simetria, deveríamos acreditar, até um grau de 95%, que a probabilidade estatística de caras difere em não mais do que 8% da frequência de caras na amostra seguinte. Mas se observarmos com espanto que a moeda só caiu sobre caras 30 vezes em 100, devemos também acreditar 95% que a moeda tem um viés de cerca de 30% de probabilidade de caras? De acordo com o argumento de simetria de Williamson, seria preciso fazer o mesmo. Ou não deveríamos antes concluir que esta série de lances foi um outlier improvável e não representativo? Somente se nossa distribuição inicial for uma distribuição aproximadamente igual, ou seja, a moeda poderia muito bem ter um viés arbitrário, o resultado observado nos leva a assumir a probabilidade de cara no intervalo de confiança [0,3

$\pm 0,08]$ de acordo com o método da máxima verossimilhança.⁷² Também pode ser diretamente provado que a etapa de argumentação $(9-10^*)$, se for aplicada a dados empíricos arbitrários $h(F:s^n) =_{\text{def}} q \in [0,1]$, resulta em uma distribuição igual da distribuição de saída:

(Teorema 9-4): Supondo que a probabilidade final satisfaça a equação $(9-10^*)$, $P(p(F) \in [h(F:s^n) \pm a_n])$ para todos os resultados de amostra possíveis $h(F:s^n)=q \in [0,1] \mid h(s^n) = q = 0,95$. Então a distribuição inicial $D(p=x)$ para n crescente difere arbitrariamente pouco de uma distribuição uniforme.

Uma prova encontra-se no apêndice 10.3.16. Ilustramos o teorema com um exemplo simples. Supostamente existem apenas duas hipóteses epistemicamente possíveis $H_1: p(F) \in [q_1 \pm a_n]$ e $H_2: p(F) \in [q_2 \pm a_n]$ com $q_2 =_{\text{def}} (1 - q_1)$, $P(H_1) = h_1$ e $P(H_2) = (1 - h_1)$. De acordo com o pressuposto tem-se que: $P(H_i \mid h(F:s^n) = q_i) = 95\%$. A aplicação do teorema de Bayes produz: $P(H_i \mid h(F:s^n) = q_i) = P(h(F:s^n) = q_i \mid H_i) \cdot P(H_i) / \sum_{1 \leq j \leq 2} P(h(F:s^n) = q_i \mid H_j) \cdot P(H_j)$. Escrevemos para as probabilidades: $P(h(F:s^n) = q_i \mid H_i) =_{\text{def}} L_i$ para $i \in \{1,2\}$ e $P(h(F:s^n) = q_i \mid H_j) =_{\text{def}} L_i^*$ para $i \neq j \in \{1,2\}$. Aparentemente, $L_i > L_i^*$ se aplica. Devido à nossa

⁷² Uma objeção semelhante pode ser encontrada em Howson/Urbach (1996, p. 240), as relacionadas a uma previsão singular. Aplicado ao nosso exemplo, o argumento deles é o seguinte: O grau de crença de que a frequência de cara de uma moeda honesta na próxima série de 100 lançamentos está entre 0,42 e 0,58 é de 95%. Entretanto, suponhamos que observamos uma frequência de cara de 0,30: se agora condicionarmos o StK a essa evidência, obteremos a afirmação contraditória de que a probabilidade de 0,3 estar entre 0,42 e 0,58 é de 95%. Williamson (2013, p. 306) responde que, ao contrário das previsões amostrais, o verdadeiro valor das probabilidades de hipóteses não é conhecido. Entretanto, como mostramos, também há probabilidades iniciais sobre hipóteses que podem entrar em conflito com a condicionalização do StK.

suposição $q_2 = (1 - q_1)$, as probabilidades L_1 e L_2 (bem como L_1^* e L_2^*) são idênticas. Então definimos $L_i = L$ e $L_i^* = L^*$ e obtemos: $P(H_i | h(s^n) = q_i) = L \cdot h_i / (L \cdot h_i + L^* \cdot (1 - h_i))$, para $i \in \{1, 2\}$. No entanto, por causa de $L > L^*$, esse valor só pode concordar para $i = 1$ e $i = 2$ se $h_1 = h_2$, ou seja, se a probabilidade for distribuída uniformemente sobre as hipóteses.

Chegamos à conclusão de que a justificação do princípio da verossimilhança e o cálculo das probabilidades de hipótese com base apenas no StK não é possível sem assumir uma distribuição igual das probabilidades iniciais. Nesse sentido, Williams e Stove também replicaram Maher em uma extensão de que as probabilidades iniciais devem ser fixadas pelo princípio da indiferença (ver Vickers 2010, § 7). Isso nos leva de volta à conclusão da última seção. Entretanto, ganhamos algo, a saber, uma nova forma de justificar o princípio da indiferença. Se aceitarmos o pressuposto (9-10*), a aplicação do StK condicionalizado a resultados de amostra arbitrários, como intuição primitiva na ausência de conhecimento adicional, então o princípio da indiferença segue com a ajuda da sentença 9-4. É claro que isso não resolve o problema da dependência linguística do princípio da indiferença. Para resolvê-lo, são necessários argumentos independentes para preferir certos sistemas de conceitos básicos a outros (ver capítulo 9.7).

Se usarmos o princípio da indiferença para poder justificar o StK mesmo na presença de evidências específicas, mesmo assim, não é mais garantido que a probabilidade epistêmica contida no sentido da sentença 7-1(b) corresponda à estatística. Por exemplo, se a frequência populacional basal for $p(Fx) = 0,8$ e observarmos um valor amostral desviante de $h_{100}(Fx) = 0,2$, concluiremos com probabilidade epistêmica $P = 0,95$ que $p(Fx)$ está no intervalo $0,2 \pm 0,08$ (ver Capítulo 8.1). Assim, concluímos ainda que a frequência da próxima amostra

com $P = 95\%$ está no intervalo $0,2 \pm 0,16$. Claro que isso não é mais consistente com a frequência a longo prazo, porque de fato a frequência de futuras amostras de 100 elementos estará no intervalo de $0,8 \pm 0,08$ a $p = 95\%$.

9.5 Bayesianismo subjetivo e convergência de graus subjetivos de crença: inferência indutiva II

A maioria dos representantes contemporâneos do bayesianismo não considera o princípio da indiferença como sacrossanto por causa dos problemas descritos. Em vez disso, eles subscrevem a visão do bayesianismo subjetivo (ou não objetivo), segundo a qual não há regras objetivas para a escolha de uma distribuição inicial preferida. Em vez disso, os bayesianos subjetivos tentam mostrar que os resultados da aprendizagem indutiva a partir da experiência são *independentes* da escolha da distribuição do produto, se aprendermos por tempo suficiente. Vamos novamente dar uma distribuição inicial $D(H_r)$ sobre as possíveis hipóteses estatísticas $H_r =_{\text{def}} "p(F_x)=r"$ sobre uma característica binária F , mas desta vez não precisa ser indiferente, mas pode ser arbitrária. Em seguida, a distribuição final $D(H_r|E)$ em relação à evidência de acordo com a equação da sentença 9-2 (e o cálculo integral, que nem sempre é fácil). Esse procedimento é chamado de atualização bayesiana ou atualização de crenças. Funciona de forma simples para as distribuições β mencionadas (Hays/Winkler 1970, p. 461) ou para distribuições normais (Howson/Urbach 1996, p. 354). Também pode ser mostrado que, independentemente da forma específica da distribuição inicial, a

condicionalização desta distribuição a uma frequência de amostragem observada provoca um deslocamento na massa de probabilidade na direção da frequência de amostragem. À medida que o tamanho da amostra aumenta, a distribuição se concentra acima do resultado da amostra e produz um pico maior e mais acentuado lá (dependendo da distribuição inicial). O único pré-requisito para essa convergência de distribuição é que a distribuição inicial não *seja dogmática* em relação à verdadeira frequência da população $p(Fx) = r$. Isso significa que a função densidade de probabilidade no ponto r é positiva e contínua, implicando que a integral é positiva em cada pequeno intervalo positivo em torno de r (de Finetti 1974, Sect. XI.4.5).

Resultados de convergência desse tipo também são chamados de "diluição das crenças iniciais" (Earman 1992, p. 141 e seg.). Os bayesianos vêem isso como uma conquista central do bayesianismo, uma vez que implica ou pelo menos sugere uma intersubjetividade de probabilidades finais epistêmicas no longo prazo, mesmo sem o princípio da indiferença. A seguir, explicamos alguns resultados de convergência desse tipo, usando um domínio infinito de indivíduos com nomes padrão a_1, a_2, \dots

Em primeiro lugar, podemos distinguir entre resultados de convergência contínua, que implicam uma mudança, ainda que pequena, no grau de crença na direção da probabilidade objetiva para cada aumento de experiência ou ampliação da amostra, e resultados de convergência simples, que significam apenas algo para o limite, ou seja, para $n \rightarrow \infty$. O teorema 7-4 sobre a aprendizagem uniforme da experiência (que generalizamos na proposição 9-5(a)) bem como as relações explicadas sobre a atualização das distribuições de saída são exemplos de resultados de convergência contínua.

(Teorema 9-5) Convergência contínua para previsões indutivas (onde $[r]$ é o arredondamento inteiro do número real r):

(a) $P(F_{a_{n+1}} | h_n(F) = (k+1)/n) > P(F_{a_{n+1}} | h_n(F) = k/n)$.

Em palavras: A probabilidade de encontrar um novo F aumenta continuamente para cada tamanho de amostra n com a frequência de F s na amostra.

Pré-requisito: P é permutável e não dogmática em relação a qualquer valor $r \in [0,1]$.

(b) $\lim_{n \rightarrow \infty} P(F_{a_{n+1}} | h_n(F) = [r \cdot n]/n) = r$.

Em palavras: A probabilidade de encontrar um novo F converge para a frequência de F na amostra à medida que o tamanho da amostra aumenta. Pré-requisito: P é permutável e não dogmático em relação a r .

Para a prova da proposição (9-5)(a), ver as notas explicativas da proposição 7-4; e para a prova da frase 9-5 (b) ver apêndice 10.3.17. A prova da frase seguinte 9-6 é apresentada no apêndice 10.3.18.

(Teorema 9-6) Convergência contínua para generalizações indutivas (onde " H_r " significa " $p(F) = r$ "):

(a) Para $n > m$ $D(H_r | h_n(F) = [r \cdot n]/n) > D(H_r | h_m(F) = [r \cdot m]/m)$ para $n > m$.

Em palavras: A densidade de probabilidade da hipótese H_r a ser preferida de acordo com o método de máxima verossimilhança, dada uma determinada frequência de amostragem, aumenta continuamente à medida que n aumenta.

(b) Para $n > m$: $ED(H_x | h_n(F) = [r \cdot n]/n) - r < ED(H_x | h_m(F) = [r \cdot m]/m) - r$.

Em palavras: O valor esperado da probabilidade estatística, formado com a distribuição de densidade de probabilidade sobre o espaço de hipóteses, dada uma determinada frequência amostral, aproxima-se continuamente da frequência amostral à medida que o tamanho da amostra aumenta.

Pré-requisito para (a)+(b): P é permutável e não dogmático em relação a r.

Convergência contínua significa (tanto nos casos de previsão quanto de generalização) que a probabilidade final se aproxima contínua ou monotonicamente das probabilidades estatísticas objetivas com o aumento da experiência. Cada aumento na experiência $n \rightarrow n+1$ causa uma pequena mudança incremental no grau de crença na direção da probabilidade objetivo-estatística (isso também é declarado nas sentenças 9-5(a) e 9-6(a,b)). Os pré-requisitos centrais para a convergência contínua são a permutabilidade e o não-dogmatismo (ou "regulação"), adequadamente explicados para funções de densidade. A vantagem da convergência contínua é que tal mudança de probabilidade é registrada "durante a vida", em todos os pontos no tempo, e não apenas quando n se aproxima do infinito e "já estamos todos mortos" (como dizia Keynes). Esta é a vantagem da convergência contínua sobre a "convergência no infinito", que é imperceptível para os seres finitos.

No entanto, a convergência contínua também tem uma fraqueza: reside no fato de a mudança gradual no grau de crença poder ser arbitrariamente reduzida por uma distribuição inicial suficientemente "extrema", de modo que a mudança na probabilidade causada pela nova experiência ocorra, mas é demasiado pequeno para ocorrer "durante a vida de alguém" e provocar mudanças significativas

nos nossos preconceitos iniciais. Isso é esclarecido pelo seguinte teorema:

(Teorema 9-7) *Incorrigibilidade final de probabilidades iniciais enviesadas:*

Seja H uma hipótese verdadeira sobre um domínio infinito de indivíduos com nomes padrão $\{a_i; i \in \mathbb{N}\}$, então há para cada conjunção longa qualquer de sentenças empíricas de qualquer comprimento $E =_{\text{def}} E_1, \dots, E_n$, que juntos apoiam uma H qualquer fortemente ($1 \geq P(E|H) > P(E)$) e são logicamente consistentes com o oposto de H , uma distribuição inicial P não dogmática, mas suficientemente prejudicada (ou seja, $P(H) \notin \{0, 1\}$), de modo que $P(\neg H|E) > P(H|E)$.

Como a prova é tão simples e instrutiva, explicamos no texto principal. De acordo com o teorema de Bayes, $P(H|E) = P(E|H) \cdot P(H) / P(E)$; e também para " $\neg H$ " em vez de " H ". Com $P(H) =_{\text{def}} h$, $P(\neg H|E) > P(H|E)$ sse $P(E|\neg H) \cdot (1-h) > P(E|H) \cdot h$ sse $P(E|\neg H) > h \cdot (P(E|H) - P(E|\neg H))$. Então $P(E|H)$ é no máximo de 1, este é certamente o caso se $P(E|\neg H) > h \cdot (1 - P(E|\neg H))$ e, portanto, se (*) h é menor que o quociente $P(E|\neg H) / (1 - P(E|\neg H))$. Uma vez que E é logicamente consistente com $\neg H$, $P(E|\neg H)$ e, portanto, $P(E|\neg H) / (1 - P(E|\neg H))$ maior que zero pode ser escolhido, portanto, a condição (*) sempre pode ser satisfeita por uma probabilidade inicial suficientemente pequena h . Para dar um exemplo concreto, um bayesiano só precisa acreditar suficientemente fortemente em um Deus todo-benevolente para que nenhuma experiência de sofrimento no mundo, por mais extensa que seja, possa dissuadi-lo de sua crença original (seguindo o padrão da paródia de Leibniz, o "Dr. Pangloss" em *Cândido* de Voltaire).

O bayesianismo, que é meramente permutativo e não dogmático, não está, portanto, em posição de impedir a irracionalidade por períodos finitamente longos de tempo, a menos que haja critérios de racionalidade independentes para probabilidades iniciais. Bayesianos estão em um dilema aqui, porque o único critério parece ser o princípio da indiferença do bayesianismo objetivo, do qual os bayesianos subjetivos se desvincularam como resultado da crítica na última seção.

Ainda mais fracos do que os resultados de convergência contínua são os resultados de convergência simples. Eles permitem que a convergência ocorra somente após uma sequência arbitrariamente tardia ou ponto no tempo na sequência de indivíduos observados (a_i) . Resultados simples de convergência não garantem que seremos capazes de detectar qualquer, por menor que seja, mudança nos graus de crença na frequência objetiva ou na verdade ao longo da vida. Assim, as condições para resultados de convergência simples são mais fracas. Em vez de permutabilidade, você só precisa da σ aditividade, da medida de probabilidade P . O teorema (9-8) apresenta os resultados de convergência simples mais importantes. Seu resultado (a) tem uma gama impressionante de aplicações e foi comprovado por Gaifman e Snir em 1982. $H(\mathcal{L})$ consiste em todas as hipóteses possíveis que podem ser formuladas em uma linguagem \mathcal{L} na qual a teoria da aritmética pode ser expressa, equipada com nomes padrão para um domínio infinito e contável de indivíduos. A sequência $(\pm_w A_i; i \in \mathbb{N})$ consiste em todos os conjuntos de base (conjuntos atômicos não negados ou negados) de \mathcal{L} que são verdadeiros para \mathcal{L} no modelo possível (ou “mundo”) w . Para a prova dos Teoremas 9-8 veja o Apêndice 10.3.19.

(Conjunto 9-8) *Resultados de convergência simples:*

(a) *Convergência Gaifman/Snir:* Para todas as hipóteses H em $H(L)$, o conjunto de mundos w possíveis (L -modelos) tem em que $\lim_{n \rightarrow \infty} P(H | \pm_w A_1 \wedge \dots \wedge \pm_w A_n)$ é anulada pelo valor verdade de H em w , a probabilidade $P = 1$.

Em palavras: com P -certeza, a probabilidade final de uma hipótese em um mundo converge contra o valor verdade da hipótese nesse mundo, condicionalmente a uma sequência infinitamente crescente de dados contendo a informação completa sobre o mundo.⁷³

(b) *Caso especial de (a):* $\lim_{n \rightarrow \infty} P(p(Fx)=r | h_n(F) = [r \cdot n]/n) = 1$, se $P(p(Fx)=r) > 0$. Em palavras, a probabilidade de uma hipótese da forma " $p(Fx)=r$ " com uma probabilidade inicial positiva converge para 1, dada uma amostra infinitamente crescente com uma frequência F de aproximadamente r (arredondada para um número divisível por n).

Pré-requisito para (a)+(b): σ aditividade de P .

(c) *Convergência para previsões rigorosas:* $\lim_{n \rightarrow \infty} P(Fa_{n+1} | Fa_1 \wedge \dots \wedge Fa_n) = 1$.

Em palavras: a probabilidade de que o próximo caso seja um F é próxima de 1 diante do aumento infinito de casos anteriores que eram todos F .

Pré-requisito: $P(\forall x Fx) > 0$.

⁷³ O requisito de que $(\pm_w A_i; i \in \mathbb{N})$ contém a informação completa sobre w , ou seja, todas as sentenças básicas de \mathcal{L} que são verdadeiras em w , é uma reformulação equivalente da condição de Gaifman e Snir de que o conjunto de sentenças $\{\pm_w A_i; i \in \mathbb{N}\}$ separa o conjunto de todos os L -modelos, no sentido de que nenhum dos dois L -modelos torna as mesmas sentenças deste conjunto verdadeiras.

(d) *Convergência para generalizações estritas*: $\lim_{n \rightarrow \infty} P(\forall x Fx | Fa_1 \wedge \dots \wedge Fa_n) = 1$.

Em palavras, a probabilidade de que todos os indivíduos sejam Fs é próxima de 1 diante do aumento infinito de casos anteriores que eram todos Fs.

Pré-requisito: $P(\forall x Fx) > 0$ e σ aditividade, de P.

Além do fato de que ele só diz algo sobre convergência no infinito, o teorema de Gaifman/Snir é baseado em duas outras restrições. (1.) Não se aplica se H contém termos teóricos que não estão contidos na sequência de dados, de modo que a sequência de dados não determina a extensão dos termos de H. Neste caso, a sequência de dados não está completa ou não inclui todas as frases básicas da língua (ver n. 71). (2.) A convergência da verdade de Gaifman/Snir aplica-se apenas a um subconjunto de mundos possíveis. O conjunto de mundos em que a convergência não ocorre tem uma probabilidade de crença de 0, mas ainda pode conter um número incontável de mundos. A convergência de verdade da função de crença P é, portanto, certa apenas do ponto de vista intrínseco de P, o que Earman (1992, p. 148) chama de “sucesso autocongratatório do método Bayesiano”.

O resultado (b) é uma consequência do resultado de Gaifman-Snir. Tanto a alínea a) como (b) pressupõem a aditividade σ que é a descrita na seção 3.4 e equivale a um pressuposto indutivo fraco. Os resultados de convergência (c) e (d), por outro lado, são uma simples consequência da condição $P(\forall x Fx) > 0$, que, como mostrado em (9-4), expressa uma suposição indutiva comparativamente forte. Ao contrário de (c), a sentença (d) requer adicionalmente aditividade σ .

9.6 Evidências consistentes e independentes

Um resultado de convergência praticamente útil é obtido para situações em que uma hipótese é consistentemente apoiada por muitas evidências mutuamente independentes. Tais situações são características do progresso científico, que se caracteriza pelo fato de confiarmos nas teorias porque foram confirmadas de forma independente por um grande conjunto de evidências. Na estrutura Bayesiana, esta intuição pode receber uma base precisa.

Diferentes evidências para a mesma hipótese se comportam umas com as outras como efeitos da mesma causa: elas são *prima facie* probabilisticamente dependentes umas das outras, mas sua dependência probabilística desaparece quando são condicionadas à causa comum (ver capítulo 8.7). A independência recíproca de duas provas deve, portanto, ser entendida condicionalmente à verdade da respectiva hipótese: $P(E_1|H \wedge E_2) = P(E_1|H)$, ou seja, E_2 não contribui em nada para o aumento da probabilidade de E_1 , além da verdade da hipótese H . Por exemplo, se tivermos dois testes de gravidez incertos e independentes à nossa frente, então um resultado positivo do primeiro teste aumenta a probabilidade de gravidez e, portanto, a probabilidade de um resultado positivo do segundo teste, mas se você olhar apenas para mulheres grávidas, o primeiro resultado positivo não aumenta mais a probabilidade de um segundo resultado positivo; ou seja, a proteção contra erros do segundo teste é independente do primeiro teste. Isso nos permite definir os termos da prova independente e favorecida da seguinte forma:

(Def. 9-1) *Evidência independente e consistente:*

(a) Diz-se que a evidência E_1, \dots, E_n é mutuamente independente no que diz respeito a uma partição de hipóteses possíveis $\{H_1, \dots, H_m\}$ sse para todas $k \in \{1, \dots, n-1\}$ e $j \in \{1, \dots, m\}$ então: $P(E_{k+1} | E_1 \wedge \dots \wedge E_k \wedge H_j) = P(E_{k+1} | H_j)$.

Nota: Isto implica $P(E_1 \wedge \dots \wedge E_n | H_j) = \prod_{1 \leq i \leq n} P(E_i | H_j)$.

A evidência E_1, \dots, E_n favorece consistentemente uma hipótese H_k de uma partição de hipóteses $\{H_1, \dots, H_m\}$ na medida $\delta > 0$ sse para cada $i \in \{1, \dots, n\}$ e $r \in \{1, \dots, m\}$ com $r \neq k$ então: $P(E_i | H_k) \geq P(E_i | H_r) + \delta$.

A presença de muitos testes incertos, mas independentemente consistentes, para um achado (por exemplo, médico) é um caso de uso típico. Outro caso de uso é a evidência independente para hipóteses da teoria biológica da evolução (Schurz 2011, capítulo 3,5). Usando as duas suposições em Def. 9-1, o seguinte resultado de convergência contínua pode ser provado (Apêndice 10.3.20):

(Teorema 9-9): *Evidência independente e consistente*

$\{H_1, \dots, H_k\}$ é uma partição de hipóteses concorrentes e $\{E_1, \dots, E_n\}$ é

(i) um conjunto de provas independente da partição; que
 (ii) favorecem por unanimidade a hipótese H_k na medida em que $\delta > 0$, então se aplica o seguinte:

- (a) $P(H_k | E_1 \wedge \dots \wedge E_n) \geq \frac{h}{h + (1-h) \cdot (1-\delta)^n}$, onde h significa $P(H_k)$, e
 (b) $\lim_{n \rightarrow \infty} P(H_k | E_1 \wedge \dots \wedge E_n) = 1$.

O fator $(1-\delta)^n$ tende continuamente a zero para $n \rightarrow \infty$, razão pela qual a afirmação (a) do teorema 9-9 garante que a probabilidade final de H aumenta a cada nova evidência independente adicionada que a favorece. Se $n \rightarrow \infty$, [a probabilidade da hipótese, dadas as evidências]

tende para 1, que é a afirmação da proposição (b) do teorema. É claro que o teorema 9-9 não garantem que a hipótese favorecida H_k seja também a hipótese verdadeira da partição. Se a evidência for consistente com uma hipótese falsa, isso aumenta a probabilidade final da hipótese falsa. Mas a ocorrência de evidências consistentes para H é extremamente improvável se H é falso, e muito mais provável se H é verdadeiro – a razão dos dois é a razão de verossimilhança, que é menor quando $(1-\delta)^n$ tende para 1 (ver Apêndice 10.3.20).

Nesse caso em que a evidência E_i representa os julgamentos concordantes de muitas testemunhas independentes em relação à verdade de um fato H , o teorema 9-9 nada mais é do que uma versão *do teorema do júri condorcetiano*. A independência mútua aqui significa que as testemunhas chegaram ao seu veredicto independentemente umas das outras. Neste contexto, o Teorema 9-9 afirma que a concordância de muitos especialistas independentes com uma tendência, embora pequena, a favor da hipótese verdadeira dá a esta hipótese um grau de confirmação muito elevado (e para $n \rightarrow \infty$, próximo de 1). O teorema do júri de Condorcet tem sido demonstrado principalmente apenas para uma partição binária de hipóteses (cf. Bovens e Hartmann 2003, para. 5.2). Neste caso, a condição (b) de Def. 9-1 $P(E_i|H_k) > P(E_i|\neg H_k)$ e, portanto, $P(E_i|H_k) > 1/2$ para todos $i \in \{1, \dots, n\}$. Nós provamos o teorema 9-9 para o caso mais geral de partições de hipóteses arbitrárias. Para partições mais de 2 hipóteses, $P(E_i|H_k)$ pode ser significativamente menor que $1/2$, desde que apenas E_i favoreça H_k sobre H_j ($j \neq k$) no sentido de Def 9-1(b).

Uma vez que o teorema 9-9 se aplica independentemente de pressupostos especiais, tais como a permutabilidade (etc.), alguém se perguntará o que os pressupostos indutivos deste teorema contêm. Eles estão contidos na condição (ii), o favoritismo concorrente da

evidência para a hipótese H_k , ou seja, a hipótese H_k tem a maior probabilidade entre seus concorrentes. Lembre-se, no caso de hipóteses estatísticas, obtivemos as probabilidades de verossimilhança a partir do princípio de coordenação estatística ou da permutabilidade equivalente, ambos pressupostos indutivos. No caso de generalizações rígidas que a evidência logicamente implica, veremos na próxima seção que sem suposições indutivas usando a "maior probabilidade" não é possível distinguir entre generalizações indutivas e antiindutivas ("Goodman-like").

Assim, a condição de favoritismo (ii) não pode favorecer generalizações indutivas em detrimento de antiindutivas. Precisamos excluir hipóteses antiindutivas da partição desde o início, a fim de aplicar o teorema 9-9 a generalizações estritas – por exemplo, porque sua probabilidade inicial é baixa desde o início devido à nossa medida P indutiva. As proposições 9-9 não podem, portanto, justificar o raciocínio indutivo sem pressuposição; no entanto, é extremamente útil para aplicação prática, na qual os princípios de indução são aceitos sem problemas e uma riqueza de conhecimento prévio indutivamente adquirido está disponível.

9.7 Justificativa probabilística do raciocínio indutivo? O paradoxo de Goodman

Nas seções 9.3 e 9.5 especificamos os tipos mais importantes de raciocínio probabilístico-indutivo e explicamos seus pré-requisitos. A abordagem de probabilidade dualista tem sido excelentemente adequada para este propósito. Ao mesmo tempo, as nossas

considerações mostraram que todo tipo (mais forte ou mais fraco) de raciocínio indutivo deve fazer certas pressuposições indutivas (mais fortes ou mais fracas) sobre a função de probabilidade epistêmica P para ser válido. Conhecemos tais pressuposições indutivas – com força indutiva crescente (não necessariamente lógica):

- (a) σ Aditividade:
- (b) Existência de limites de frequência com $P=1$
- (c) Princípio da Coordenação Estatística
- (d) Permutabilidade [equivalente a (b) & (c)]
- (e) Princípio da Indiferença, bem como
- (f) Probabilidade inicial positiva de todas as hipóteses estritas.

(a) permite fácil convergência; (d) e (f) permitem a convergência contínua ((f) para teoremas estritos), e a adição de (e) permite o cálculo de probabilidades finais numéricas-indutivas de hipóteses. (Além disso, o não-dogmatismo deve ser assegurado para permitir o aprendizado com a experiência.)

Finalmente, deve-se ressaltar que, *sem* tais pressupostos indutivos sobre P , uma justificativa probabilística do raciocínio indutivo não é possível. Isso não parece surpreendente, mas precisa ser dito, já que alguns bayesianos argumentaram que certas inferências indutivas são válidas apenas com base nos axiomas básicos. Por exemplo, Howson (1997, p. 279) argumentou que os axiomas básicos implicam uma lógica indutiva fraca, uma vez que se segue que uma hipótese contingente H , que logicamente implica uma evidência, é aumentada em probabilidade por essa evidência e, portanto, confirmada bayesianamente, $P(H|E) > P(H)$ (cf. (9-1) e frases 9-10). Na seção 9.9, contudo, veremos que este aumento na

probabilidade se baseia numa mera redução do espaço de possibilidades (truncamento). Em outras palavras, E aumenta a probabilidade de H simplesmente porque E é um componente lógico de H e E aumenta sua própria probabilidade para 1 ($P(E|E) = 1$). Portanto, E não deve aumentar a probabilidade de nenhuma porção de conteúdo de H que transcenda E para aumentar a probabilidade de H. Contudo, este último [aumento] é necessário para poder falar de uma confirmação indutiva de H.

Isso ilustra, entre outras coisas, o “infame” paradoxo de Goodman. Goodman (1946) mostrou que a aplicação do raciocínio indutivo a *todos os* predicados é racionalmente *impossível* porque leva a *contradições lógicas*. Sua famosa definição (1975, p. 97 e seg.) do predicado “Verdul” (G^*x) é a seguinte:

(Def. 9-2): (Definição de “verdul” de Goodman): Dado um ponto constante no tempo t_0 no futuro até, digamos, o ano 3000, um objeto x é chamado verdul (G^*) sse x é verde sse foi observado antes de (Bxt_0) , e caso contrário é azul.

Em fórmulas: $G^*x : \Leftrightarrow ((Bxt_0 \wedge Gx) \wedge (\neg Bxt_0 \wedge Rx))$.

Dada uma amostra $\{a_i; 1 \leq i \leq n\}$ de esmeraldas verdes (S) observadas antes, então todas essas esmeraldas também são verdul. Mais especificamente, as afirmações $Sa_i \wedge Ba_i t_0 \wedge Ga_i$ e $Sa_i \wedge Ba_i t_0 \wedge G^*a_i$ são *definidamente* equivalentes. Se aplicarmos a generalização indutiva para “verde”, bem como para “verdul” em nossa amostra, as duas hipóteses $H =_{\text{def}}$ “Todas as esmeraldas são verdes” e $H^* =_{\text{def}}$ “Todas as esmeraldas são verdul” resultam. No entanto, H e H^* implicam previsões contraditórias (verde versus azul) para todas as esmeraldas não examinadas anteriormente.

Existem inúmeras formulações alternativas do predicado de Goodman (Schurz 2013b, seção 5.10.7). A variante mais simples em nosso contexto vem de Leblanc (1963) e define “verdul” em relação a uma amostra $\{a_1, \dots, a_n\}$ da seguinte forma:

(9-11) X é verdul sse X pertence à amostra observada $\{a_1, \dots, a_n\}$ e é verde, ou não pertence a ela e não é verde.

Formal: $G^*x \Leftrightarrow_{\text{def}} (x \in \{a_1, \dots, a_n\} \wedge Gx) \vee (x \notin \{a_1, \dots, a_n\} \wedge \neg Gx)$.

O predicado de Goodman foi chamado de “patológico” porque não pode ser projetado indutivamente: pois projetar “verdul” indutivamente significa fazer uma generalização antiindutiva sobre as instâncias do predicado “saudável” “verde”. Portanto, a aprendizagem indutiva uniforme não pode ser aplicada aos predicados “verde” e “verdul” ao mesmo tempo. No entanto, de acordo com a frase 7-4, a aprendizagem indutiva é uma consequência da permutabilidade e regularidade (não dogmatismo). Portanto, há a seguinte conexão entre o paradoxo de Goodman e o pressuposto de permutabilidade para P: Assumindo a regularidade de P, o princípio da permutabilidade não pode se aplicar simultaneamente a um predicado (Gx) e sua contraparte de Goodman (G^*x). Isso porque, de acordo com o teorema 7-4, decorre da permutabilidade para Gx: $P(Ga_{n+1} | Ga_1 \wedge \dots \wedge Ga_n) > P(Ga_{n+1})$. Se G^*x é definido em (9-11), então para $1 \leq i \leq n$ Ga_i é analiticamente equivalente a G^*a_i , mas Ga_{n+1} é analiticamente equivalente a $\neg G^*a_{n+1}$.

Portanto, aplicando permutabilidade a G $P(\neg G^*a_{n+1} | G^*a_1 \wedge \dots \wedge G^*a_n) > P(\neg G^*a_{n+1})$ e, por isso, (2) $P(\neg Ga_{n+1} | Ga_1 \wedge \dots \wedge Ga_n) > P(\neg Ga_{n+1})$, em contradição com o pressuposto de permutabilidade de Gx. Kutschera (1972, p. 144) concluiu que a tentativa de Carnap de

estabelecer uma lógica indutiva no sentido de um sistema de postulados analíticos deve ser considerada um fracasso.

Para o conceito bayesiano de confirmação, verifica-se que a amostra de esmeraldas verdes confirma tanto a generalização indutiva (H): “Todas as esmeraldas são verdes” quanto a generalização anti-indutiva de Goodman (H*) “Todas as esmeraldas são ‘verdul’”. Para ambas as hipóteses implicam a data experiencial E, e, portanto, a probabilidade de ambas as hipóteses é 1. A diferença entre a probabilidade final de H dado E e a de H* dado E depende *apenas* das probabilidades iniciais P(H) e P(H*). Medidas de probabilidade indutivas dão à hipótese “uniforme” H uma probabilidade inicial muito maior do que a hipótese H*, que prevê um “salto anti-indutivo”, mas medidas de probabilidade não indutivas não precisam fazer isso. Pela mesma razão, nenhuma previsão indutiva da forma $P(Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n) > P(Fa_{n+1})$ pode ser mostrada para qualquer medida de probabilidade não indutiva. Para uma medida P não indutiva, $P(Fa_{n+1}|Fa_1 \wedge \dots \wedge Fa_n) < P(Fa_{n+1})$ também é possível. Isso mostra mais uma vez que os axiomas básicos da teoria da probabilidade não são suficientes para justificar o raciocínio indutivo – uma conclusão que mais tarde foi tirada pelo próprio Howson (ver Howson 2000, p. 133).

Em última análise, esses resultados são baseados no famoso argumento cético de David Hume de que uma justificativa não circular para a confiabilidade do raciocínio indutivo é impossível (Hume 1748, § 4-6). Confiabilidade significa que inferências indutivas levam a uma conclusão verdadeira, pelo menos com uma alta probabilidade de premissas verdadeiras.

As inferências indutivas só são confiáveis se as frequências de *ocorrência observadas até agora* puderem ser transferidas para o futuro ou para os casos não observados, o que só é o caso se nossa

realidade for suficientemente indutivamente uniforme. No entanto, esta última só pode ser justificada pelo raciocínio indutivo. Um argumento não-circular em favor da confiabilidade do raciocínio indutivo não é, portanto, possível, e esse insight se repete em todas as maneiras de especificar o raciocínio indutivo na teoria da probabilidade: Cada um desses esclarecimentos nos mostrou que o tipo de raciocínio em questão só é válido se certas suposições indutivas tiverem sido feitas sobre a função de probabilidade P .

Uma maneira mais promissora de justificar o raciocínio indutivo de forma não circular é a ideia de justificar a indução como um método ótimo de cognição, que remonta a Hans Reichenbach (1935, § 80). De acordo com essa ideia, as induções são a melhor coisa que podemos fazer se quisermos alcançar nosso objetivo de obter verdadeiras proposições e, especialmente, previsões sobre o futuro. Houve também sérias objeções à ideia de Reichenbach (ver Skyrms 1999, § III.4). No entanto, foi demonstrado em Schurz (2008b) que uma justificação de otimização não circular pode de fato ser dada para um método indutivo particular, nomeadamente o método de meta-indução. Ao analisar jogos de predição, foi demonstrado que o método de meta-indução ponderada é um método de predição ideal entre todos os métodos disponíveis. Mesmo que se considere que o raciocínio indutivo é fundamentalmente justificado, ainda se coloca a questão da resolubilidade do paradoxo de Goodman. Examinando mais de perto, muitas das propostas para a sua solução revelaram-se inadequadas (cf. Kutschera 1972, p. 145-156). Em minha opinião, a sugestão mais conveniente vem de Carnap (1947, p. 146) e afirma que os predicados indutíveis *devem ser qualitativos*, ou seja, não devem se referir a constantes individuais em suas definições. Predicados que fazem isso, Carnap chamou de predicados *posicionais*.

Os predicados patológicos de Goodman são definidos por meio de constantes individuais (ou constantes de tempo) e, portanto, são posicionais. Como demonstrado na prova da frase 7-1(a) (Anexo 10.3.9), a aplicação de pressupostos de permutabilidade para predicados qualitativos não conduz a qualquer contradição. Como abaixo na Def. 7-5, a razão mais profunda para o surgimento de pressupostos contraditórios de permutabilidade para predicados posicionais reside no fato de que as constantes individuais que ocorrem na definição dos predicados são “ocultas” e, portanto, removidas da permutabilidade. Por exemplo, trocar a_2 por a_{n+1} seria $P(Ga_1 \wedge Ga_2) = P(Ga_1 \wedge Ga_{n+1})$ e $P(G^*a_1 \wedge G^*a_2) = P(G^*a_1 \wedge G^*a_{n+1})$. Agora, de acordo com a Def. (9-11) $G^*a_1 \wedge G^*a_{n+1}$ é, por definição, equivalente a $Ga_1 \wedge \neg Ga_{n+1}$, o que leva à contradição devido à natureza indutiva de P . No entanto, este não é mais o caso, mesmo que na definição de G^* , no conjunto $\{a_1, \dots, a_n\}$, a_2 seja substituído por a_{n+1} ; neste último caso, $Ga_1 \wedge Ga_{n+1}$ ainda seria equivalente a $Ga_1 \wedge Ga_{n+1}$ e uma contradição não pode ocorrer. Assim, se todos os predicados definidos são substituídos por predicados básicos, os pressupostos de permutabilidade não levam a contradições.

Mas isto ainda não resolve aquele que é talvez o problema mais difícil associado ao paradoxo de Goodman: o problema da dependência linguística. Contra a sugestão de Carnap, Goodman (1975, p. 105) argumentou que, com a ajuda de redefinições mútuas, pode-se passar da nossa linguagem comum com “verde” e “azul” como termos básicos para uma linguagem definitivamente equivalente em que “verdul” e “azuver” atuam como predicados básicos. Não podemos aprofundar este problema aqui (para detalhes ver Schurz 2013b, Capítulo 5.11.13). A consequência do problema da dependência linguística é que são necessários critérios independentes da linguagem

para distinguir entre termos de observação qualitativos e posicionais. Uma abordagem correspondente é desenvolvida em Schurz (*ibid.*), na qual se propõe distinguir entre termos de observação qualitativos e posicionais por referência à sua capacidade de aprendizagem ostensiva.

9.8 Teorias gerais bayesianas de confirmação

No capítulo 9.2-5 apresentamos métodos bayesianos de confirmação indutiva de hipóteses estatísticas. No âmbito da filosofia bayesiana da ciência, também foram desenvolvidas teorias gerais de confirmação que supostamente são aplicáveis a todos os tipos de hipóteses. O conceito bayesiano mais importante de confirmação é o seguinte (cf., por exemplo, Howson/Urbach, 1996, p. 117 e seg.):

(Def. 9-3) Confirmação Bayesiana

(a) E confirma H dado o conhecimento prévio W (em relação a uma função de crença P) sse $P(H|E \wedge W) > P(H|W)$ é válido.

Nota: (a) Em aplicações práticas, “>” deve ser interpretado como “significativamente superior”.

(b) O termo de confirmação incondicional “E confirma H (simplificado)” é obtido assumindo que W é vazio ou logicamente verdadeiro: $P(H|E) > P(H)$.

A confirmação é aqui entendida no sentido incremental, como um aumento de probabilidade, e não no sentido absoluto como uma

alta probabilidade $P(H|E)$. Note-se que as informações referidas na seção 9.11, por outro lado, exige um alto grau de probabilidade. Assumindo $P(H) = 0,95$ e $P(H|E) = 0,9$, então $\neg H$ é incrementalmente confirmado por E (porque $P(\neg H|E) = 0,1 > P(\neg H) = 0,05$), mas a hipótese H ainda é muito mais provável do que $\neg H$, dado E, sendo, portanto, racionalmente preferível a $\neg H$. Pode-se definir confirmação no sentido absoluto em vez de incremental (ver Carnap 1950, xvi; Huber 2008, p. 184). No entanto, os Bayesianos atuais utilizam quase que exclusivamente o conceito de confirmação incremental, o qual chamamos também de "confirmação Bayesiana" de forma abreviada.

Diferentes medidas quantitativas bayesianas de confirmação foram desenvolvidas, por exemplo, a medida da diferença $P(H|E \wedge W) - P(H|W)$, (b) a medida da razão $P(H|E \wedge W)/P(H|W)$, (c) a medida de razão modificada $P(H|E \wedge W) - P(H|\neg E \wedge W)$, ou (d) a dimensão do produto $P(H \wedge E|W) - P(H|W) \cdot P(E|W)$ (visão geral em Fitelson, 1999). Essas medidas têm propriedades numéricas diferentes, mas são todas ordinalmente equivalentes à medida de confirmação incremental, ou seja, fornecem confirmação positiva sse $P(H|E \wedge W) > P(H|W)$. A confirmação bayesiana tem duas propriedades bem conhecidas. Primeiro, coincide com aumento de probabilidade:

$$(9-12) P(H|E \wedge W) > P(H|W) \text{ sse } P(E|H \wedge W) > P(E|W)$$

Em segundo lugar, a confirmação hipotético-dedutiva clássica é um caso especial de confirmação bayesiana. Ou seja, sempre que uma hipótese H implica logicamente a evidência E, H é bayesianamente confirmada por E se as probabilidades iniciais de 0 e 1 forem diferentes (ver Teoremas 9-10 abaixo).

Em particular, o conceito bayesiano de confirmação está exposto aos seguintes problemas:

(1.) *O problema da evidência antiga*: já discutimos esse problema na seção 7.5. O problema tem como consequência que o conceito bayesiano de confirmação só funciona adequadamente se “P” representar uma probabilidade inicial independente de evidência ou “probabilidade de suporte”.

(2.) *A arbitrariedade das probabilidades*: As probabilidades só são determinadas objetivamente em dois casos. *Primeiro*, quando se trata de hipóteses estatísticas e $P(E|H)$ é identificado com a verossimilhança estatística $p_H(E)$ de acordo com o StK. *Segundo*, se H implica logicamente E, ou seja, a confirmação bayesiana coincide com a confirmação hipotético-dedutiva.

Em outros casos a determinação da probabilidade $P(E|H)$ parece ser cientificamente subdeterminada. Por exemplo, até que ponto devemos estimar a probabilidade da teoria geral da relatividade em relação aos dados astronómicos atuais? Ninguém pode dar uma resposta não arbitrária a isso.

(3.) *Pseudo-confirmação bayesiana*: A noção bayesiana de confirmação é muito fraca para distinguir confirmação genuína de pseudo-confirmação. Isso será mostrado na próxima seção.

9.9 Pseudo-confirmação através de redução do espaço de possibilidades (truncamento) versus confirmação genuína

O fato de que o conceito bayesiano de confirmação permite pseudo-confirmações é melhor visto no caso especial em que a hipótese implica logicamente a evidência:

(Teorema 9-10) Pseudo-confirmação Bayesiana: Toda hipótese H (não importa quão “louca”) com uma probabilidade inicial positiva é confirmada por cada evidência E , desde que E seja apenas logicamente implícito por H e $P(H) > 0$ e $P(E) < 1$ se aplica.

O teorema 9-10 segue diretamente do teorema de Bayes: $P(H|E) = P(E|H) \cdot P(H) / P(E) = 1 \cdot P(H) / P(E) > P(H)$, desde que $P(E) < 1$ e $P(H) > 0$. Portanto, E não precisa aumentar a probabilidade daquelas partes de conteúdo de H que vão além de E : para confirmar H no sentido bayesiano, é suficiente que E seja uma parte de conteúdo de H e se confirme a si mesma. Ken Gemes e John Earman (1992, p. 98, 242, n. 5) chamaram este caso de confirmação por mero “corte de conteúdo”: o espaço de possibilidades é reduzido por “ $\neg E$ ” através da condicionalização em E , em que a parcela relativa de H neste espaço de possibilidades aumenta. O fato de a confirmação bayesiana também permitir pseudo-confirmações através de meras reduções do espaço de possibilidades (truncamento) é responsável por três problemas muito discutidos:

- (a) Problemas de irrelevância lógica, como o paradoxo da colagem conjuntiva (“aderência por conjunção”);
- (b) falta de caráter indutivo e paradoxo de Goodman; bem como
- (c) Pseudo-confirmação através de especulação pós-fato.

O caso mais simples do paradoxo da cola conjuntiva é quando uma hipótese "maluca" é "colada" diretamente à evidência como um elo de conjunção. Segue-se dos teoremas 9-10 que para qualquer evidência E com $P(E) \neq 0,1$ e qualquer hipótese X com $P(E \wedge X) \neq 0$, E confirma a conjunção $E \wedge X$. Por exemplo, a observação "a grama é verde" confirma a conjunção de "a grama é verde" com o ensinamento das Testemunhas de Jeová (X). Isso, é claro, não é uma confirmação genuína, porque a única parcela do conteúdo de $E \wedge X$ que vai além de E é X , e E é probabilisticamente irrelevante para X : $P(X|E) = P(X)$. O nome "aderência por conjunção" foi introduzido por Lakatos (1970, p. 46) e Glymour (1981, p. 67). O paradoxo da cola conjuntiva tem sido discutido em particular no contexto da confirmação hipotético-dedutiva (Gemes 1993, Schurz 1994); Crupi/Trentori (2010) discutem-na no contexto da teoria bayesiana da confirmação.

Em casos mais complicados, a conjunção não está colada a E , mas a uma hipótese que implica E : Se E confirma H porque $H \parallel - E$ (e $P(E) < 1$, $P(H) > 0$), então E também confirma a conjunção $H \wedge X$ para qualquer hipótese adicional X , não importa o quão absurda, se apenas $P(H \wedge X) > 0$, porque se E segue de H , então E também segue de $H \wedge X$. Nesse sentido, os fatos biológicos confirmam não apenas a teoria da evolução, mas também sua conjunção com o ensino das Testemunhas de Jeová. Neste caso, há uma pseudoconfirmação parcial: uma conjunção E -transcendente de $H \wedge X$, ou seja, X , não é tornada mais provável por E . A seguir, descrevemos uma hipótese como

genuinamente confirmada pela evidência E somente se E também aumentar a probabilidade de frações de conteúdo E-transcendente de H. Uma fração de conteúdo transcendente de H é uma consequência de H que não é logicamente implícita por E.

No entanto, o conceito de “parte de conteúdo” ou “elemento de conteúdo” deve ser clarificado de forma adequada e lógica. Se uma hipótese fosse decomposta em cláusulas de conjunção e essas cláusulas de conjunção fossem admitidas como elementos de conteúdo, esse conceito estaria sujeito à objeção de Popper/Miller (1983) e, portanto, seria insustentável. Esta objeção é: H é logicamente equivalente a $(E \vee H) \wedge (\neg E \vee H)$, e enquanto a primeira conjunção de E está logicamente implícita, a probabilidade do segundo termo de conjunção é comprovadamente reduzida por E. No entanto, atenuações disjuntivas como $E \vee H$ e $\neg E \vee H$ não devem ser aceitas como elementos de conteúdo genuínos de H. *A seguir, entendo* que elementos de conteúdo significam *consequências relevantes conjuntivas indecomponíveis* de uma hipótese H, ou seja, consequências relevantes que não podem ser logicamente equivalentes para serem transformadas em conjunções de consequências relevantes ainda mais curtas. Uma consequência relevante de uma hipótese H é uma consequência de H na qual nenhum predicado pode ser substituído por qualquer outro predicado (igual), preservando a validade da inferência. O conceito do *elemento de conteúdo relevante* já foi tratado por mim em outros trabalhos e aplicado de diversas maneiras, de modo que não precisa ser discutido em detalhes aqui (Schurz, 1991; Schurz/Weingartner, 2010). Aqui estão alguns exemplos de elementos de conteúdo:

(9-13) Exemplos de hipóteses de elementos de conteúdo:

Hipótese: Elementos de conteúdo:

$p \wedge q$ p, q (mas não $p \vee q$, nem $p \wedge q$).

$\neg(p \vee q)$: $\neg p, \neg q$

$(p \rightarrow q) \wedge p$: p, q (mas não $P \rightarrow Q$)

$\forall x(Fx \vee Gx \rightarrow Hx \wedge Qx)$: $\forall x(Fx \rightarrow Hx), \quad \forall x(Gx \rightarrow Hx), \quad \forall x(Fx \rightarrow Qx),$

$\forall x(Gx \rightarrow Qx)$ e

todas as instanciações dos quatro teoremas universais para constantes individuais arbitrárias.

A seguir, entendo que um componente de conteúdo é uma conjunção de elementos relevantes de conteúdo. Se E é um componente do conteúdo de H e H é logicamente mais forte que E, então nem sempre precisa haver um componente de conteúdo H^* de H, de modo que H é logicamente equivalente à conjunção $E \wedge H^*$. Mas então há sempre um componente de conteúdo H^* de H que transcende E, ou seja, não está logicamente contido em E. Além disso, o conjunto de elementos de conteúdo de uma hipótese H é comprovadamente sempre equivalente a H, de modo que nenhuma informação é perdida restringindo as consequências consideradas a elementos relevantes de conteúdo (Schurz 2013b, prop. 3.12-1).

Aprendemos sobre a falta de caráter indutivo do conceito bayesiano de confirmação na seção 9.7. Seja $\{a_1, \dots, a_n\}$ o conjunto de todas as esmeraldas observadas até agora; nossa evidência E diz que eram todas verdes. Então E confirma tanto a hipótese H_1 , segundo a qual todas as esmeraldas são verdes, quanto sua contraparte de Goodman H_2 , segundo a qual apenas as esmeraldas na amostra $\{a_1, \dots, a_n\}$ são verdes, e todas as outras não são verdes. Podemos decompor ambas as hipóteses em conjunções de componentes de conteúdo da seguinte forma:

$$(9-14) \quad H_1 = E \wedge H_1^*, \text{ com } H_1^* = \forall x \notin \{a_1, \dots, a_n\} : (Ex \rightarrow Gx), \\ H_2 = E \wedge H_2^*, \text{ com } H_2^* = \forall x \notin \{a_1, \dots, a_n\} : (Ex \rightarrow \neg Gx).$$

A parcela de conteúdo da H_i é H_i^* em ambos os casos. Apenas H_1^* , mas não H_2^* , é confirmado por E indutivamente. Portanto, apenas H_1 , mas não H_2 , é genuinamente confirmado por E. É claro que o fato de E não confirmar indutivamente H_1^* não decorre dos axiomas básicos de probabilidade, mas pressupõe a permutabilidade de P para os predicados básicos de \mathcal{L} (verde e azul).

A pseudo-confirmação por meio da redução do espaço de possibilidades (truncamento) torna-se particularmente problemática no caso de hipóteses que sustentam *conceitos ou conceitos teóricos* ou que contêm variáveis *latentes* que não estão incluídas na evidência e cujos valores são arbitrariamente adaptáveis à evidência. Tais hipóteses podem ser adaptadas *post-fato*, ou seja, retrospectivamente, a observações arbitrárias e, portanto, fornecem a base para especulações pós-fato sem poder preditivo empírico ou significância (Schurz 2013b, seção 5.10.4). Por exemplo, o fato de a grama ser verde confirma a hipótese de que essa foi a vontade de Deus Todo-Poderoso, porque essa hipótese implica logicamente E. Com o esquema explicativo "Deus quis assim" você pode explicar tudo em retrospectiva sem ser capaz de prever nada. Mas isso não muda o fato de que a probabilidade da hipótese de Deus é aumentada pelo mero espaço de possibilidades (truncamento). Muitas outras hipóteses podem ser pseudo-confirmadas por analogia, por exemplo, que um diabo ou um monstro espaguete⁷⁴ fez com que a grama ficasse verde.

⁷⁴ A Igreja do "Monstro de Espaguete Voador" é uma invenção de estudantes de física para reduzir o criacionismo ao absurdo. Veja www.venganza.org/about/open-letter.

O problema das confirmações pós-fato surge não apenas quando a hipótese implica logicamente a evidência, mas também quando apenas a torna provável: por exemplo, o fato de a grama ser verde também confirma a hipótese estatística de que isso foi desejado por um monstro espaguete cujos desejos nem sempre, mas na maioria das vezes, são realizados. Com base nesses fatos, os neocriacionistas realmente procuraram justificar versões racionalizadas do criacionismo usando métodos bayesianos de validação. Assim, Swinburne (1979, cap. 13) argumenta que certas experiências aumentam a probabilidade da existência de Deus. Unwin (2003) calcula a probabilidade final da existência de Deus em 67%. A resposta padrão dos bayesianos a esse problema é que as explicações científicas são mais bem confirmadas do que as religiosas, porque têm uma probabilidade inicial maior do que as explicações religiosas (Howson/Urbach 1996, p. 141; Sóbrio, 1993, p. 31). No entanto, essa resposta é questionável, pois as probabilidades iniciais são mais subjetivas. Os criacionistas atribuem uma maior probabilidade de resultado à explicação religiosa do que à explicação científica de que a grama é verde porque contém clorofila. Além disso, a explicação religiosa parece não ser apenas "um pouco menos" do que a explicação científica, mas não é *de todo* confirmada.

Podemos fazer justiça a essa intuição com a ajuda do conceito de confirmação genuína. A característica das pseudo-explicações pós-fato acima é que elas resultam do ajuste subsequente de uma hipótese de estrutura com variáveis latentes. A hipótese de estrutura em nosso exemplo é "Há um Deus todo-poderoso cujos desejos são sempre realizados", e a variável latente é "os desejos de Deus". Uma vez que a hipótese do quadro pode ser adaptada a qualquer observação ajustando as variáveis, ou seja, substituindo valores adequados,

qualquer observação possível pode ser explicada retrospectivamente. Exatamente pela mesma razão, pseudo-explicações desse tipo nunca podem fazer previsões bem-sucedidas. Muitos teóricos científicos têm, portanto, considerado a confirmação por novas previsões como um critério central de confirmação genuína.⁷⁵

Worrall (2006) argumenta que o critério de previsão é muito estrito: para uma confirmação genuína, basta que a evidência confirmatória seja "não utilizada", ou seja, não tenha *sido usada para construir* a hipótese confirmada. Propôs, portanto, que o critério de confirmação de previsão fosse substituído pelo critério mais adequado de "novidade de uso". Uma consequência "não utilizada" de uma hipótese é uma previsão *potencial*, uma vez que poderia ter servido como uma previsão em diferentes circunstâncias (porque não foi usada para construir H).

O critério de "não-uso" de Worrall também foi objeto de uma série de objeções (cf. Howson 1990), que não pode ser discutido em detalhes aqui. Schurz (2013b, seção 5.10.4; 2013c) desenvolveu uma reconstrução *probabilística* aprimorada da ideia de "não-uso" usando o conceito explicado de confirmação genuína quando aplicado a hipóteses com variáveis ou parâmetros latentes. A seguir, $\exists xHx$ representa a hipótese geral com o parâmetro x não ajustado, cujos valores são usados para quantificar a existência; também chamamos $\exists xHx$ de *hipótese do framework*. E significa evidência e H_c para a especialização de ajuste de parâmetros de $\exists xHx$. " $\exists xHx$ " significa que existe algum valor x da variável X de modo que Hx se mantém, enquanto H_{c_E} diz que esse valor é c , onde o valor c é obtido ajustando-se à evidência E . No caso do criacionismo pós-fato, a hipótese

⁷⁵ Veja Musgrave (1974), que cita Descartes, Leibniz, Whewell e Duhem como proponentes do critério de confirmação da predição.

imprópria diz algo como " $\exists x(\text{Deus efetuou } x)$ ", e a hipótese ajustada a parâmetros H_{c_E} então diz "Deus efetuou E".

Tecnicamente, deve-se notar que "x" pode não ser apenas uma variável individual, mas também uma variável predicada, o valor variável de um parâmetro ou uma variável matemática, ou mesmo uma sequência inteira de tais variáveis. Analogamente, a constante "c" pode representar uma sequência inteira de constantes correspondentes.

É importante notar que a hipótese do framework $\exists xHx$ representa uma *fração de conteúdo E-transcendente* da hipótese ajustada H_{c_E} . Normalmente, $\exists xHx$ é um elemento de conteúdo subjuntivo indecomponível e muitas vezes até mesmo o único elemento de conteúdo de transdução eletrônica de H_{c_E} . Isso nos permite aplicar nossa abordagem de confirmação genuína e perguntar se a probabilidade desse componente de conteúdo E-transcendente é aumentada pela evidência E. Este não é o caso em nosso exemplo, precisamente porque a hipótese não ajustada $\exists xHx$ em qualquer possível evidência E' poderia ter sido igualmente bem encaixada. Para esclarecer, que $\{E_1, \dots, E_n\}$ seja uma partição de evidências possíveis, com $0 < P(E_i) < 1$ para todos $i \in \{1, \dots, n\}$. Como regra geral, estes são os possíveis resultados de atos observacionais ou experimentos realizados em um local espaço-temporal específico, por exemplo, o clima de amanhã ou o resultado da batalha. Como a hipótese de enquadramento $\exists xHx$ (pois "Deus fez algo") pode ser *igualmente* bem ajustada a cada uma dessas possíveis evidências, ela não aumenta a probabilidade de nenhuma dessas possíveis evidências acima de sua probabilidade inicial. Embora a hipótese de ajuste pós-fato $H_{c_{E_i}}$ aumente a probabilidade de E_i , $\exists xHx$ não contribui para esse aumento de probabilidade. Então $P(E_i | \exists xHx) = P(E_i)$, portanto, $P(\exists xHx | E_i) =$

$P(\exists xHx)$ para todos $i \in \{1, \dots, n\}$, ou seja, nenhuma dessas possíveis evidências aumenta a probabilidade de $\exists xHx$. Uma vez que $\exists xHx$ representa o conteúdo de Hc em excesso de E , nenhuma das evidências possíveis pode levar a uma confirmação genuína de Hc_{Ei} .

Chamamos a hipótese de ajuste Hc_E *independentemente testável* se implicar evidências não utilizadas ou predições potenciais E' que não foram usadas para ajuste. Em contraste com especulações religiosas, hipóteses científicas com variáveis latentes são geralmente testáveis independentemente. Por exemplo, a explicação científica da cor verde da grama pelo componente verde clorofila é independentemente testável porque existem métodos de identificação independentes para a clorofila. Para Deus, não existem tais métodos independentes de identificação. É claro que as hipóteses criacionistas podem ser enriquecidas de tal forma que se tornem testáveis independentemente. Mas nas “religiões racionalizadas” isto é deliberadamente evitado a fim de proteger as hipóteses religiosas de uma possível refutação empírica.

Se a hipótese ajustada Hc_E for rica o suficiente para gerar consequências não utilizadas E' contra as quais possa ser testada independentemente, e se essas consequências realmente ocorrerem, então E' não apenas aumenta ainda mais a probabilidade da hipótese Hc_E ajustada a E , mas também agora também a probabilidade da hipótese do framework $\exists xHx$. Isso porque, *depois* de encaixada em E , essa hipótese de enquadramento não cabe mais a nenhuma nova evidência E' , tal encaixe só ocorre se as duas evidências E e E' estiverem em uma relação lícita implícita por $\exists xHx$. A hipótese do quadro inadequado é então parcialmente responsável por esse sucesso de confirmação independente através da previsão potencial

E' . Neste caso, é plausível assumir um aumento de probabilidade de $\exists xHx$ por E' ; há, portanto, uma confirmação genuína de Hc_E por E' .

Na próxima seção, aplicaremos essa abordagem ao problema do *ajuste de curvas*. Antes disso, porém, o conceito de afirmação genuína é precisamente definido. Essa noção implica que o aumento da probabilidade de H por E é transferido para os componentes de conteúdo transcendente E de H (uma adaptação dessa ideia pode ser encontrada em Earman 1992, p. 106). No entanto, esta ideia permite duas explicações possíveis. Podemos exigir que E confirme bayesianamente todos os elementos de conteúdo que transcendem E de H – neste caso falamos de confirmação genuína completa.

Ou exigimos apenas que *alguns* elementos de conteúdo E -transcendente de H devem ser confirmados por E e, em seguida, totalmente genuinamente confirmados – neste caso, estamos falando de confirmação *genuína parcial*. Exigimos a confirmação completa de pelo menos alguns elementos de conteúdo E -transcendente, porque, caso contrário, o conceito de confirmação genuína parcial coincidiria com a confirmação bayesiana comum (para detalhes, ver Schurz 2013c, §§ 4.7-8).

(Definição 9-4): (a) Um elemento de conteúdo G de uma hipótese H é chamado E -transcendente se G não segue logicamente de E .

(b) H é totalmente confirmado autenticamente por E (em relação a uma função de probabilidade P) sse a probabilidade de cada elemento de conteúdo E -transcendente G de H é aumentado em E ($P(G|E) > P(G)$).

(c) H é parcialmente confirmado genuinamente por E (em relação a uma função de probabilidade P) sse existe pelo menos um

elemento de conteúdo G de H que transcende E, o que é completamente confirmado por E.

O que determina a transferência do aumento de probabilidade de uma hipótese para seus elementos de conteúdo e-transcendente? A resposta bayesiana ortodoxa para isso é: ela é determinada pela probabilidade desses elementos de conteúdo em relação a E, bem como pela probabilidade inicial de E. No entanto, como as probabilidades dos elementos de conteúdo de H são geralmente objetivamente subdeterminadas, essa resposta é inadequada. Precisamos de critérios racionais para transferir o aumento de probabilidade para os elementos de conteúdo de uma hipótese. A transferência do aumento de probabilidade para um elemento de conteúdo G deve depender da importância de G para o aumento de probabilidade de E por H. É claro que essa questão *também* depende da natureza substantiva da hipótese em relação ao nosso conhecimento prévio, razão pela qual não podemos fornecer critérios suficientes para a transferência do aumento de probabilidade. No entanto, com base no que foi dito, podemos formular as seguintes duas *condições necessárias*:

(Teorema 9-11) Condições necessárias para a transferência do aumento de probabilidade:

Se H aumenta a probabilidade de E, então o aumento resultante na probabilidade de H até E é transferido para um elemento de conteúdo G de H que transcende E ($P(G|E) > P(G)$) somente se:

(1.) G é necessário dentro de H para tornar E provável, ou seja, não há conjunção H^* de elementos de conteúdo de H que torne E pelo

menos tão provável ($P(E|H^*) \geq P(E|H)$), mas não contém logicamente G , e
 (2.) não é o caso que G tem a forma $\exists xHx$. H resultou do ajuste do parâmetro x a E ($H = H_{c_E}$) e este processo de ajuste poderia ter sido realizado com igual sucesso para qualquer outra evidência possível E' .

9.10 Ajuste de curvas

No ajuste de curvas, assume-se que os valores de duas (ou mais) variáveis de valor real X e Y são probabilisticamente dependentes uma da outra, na forma de uma função $Y = f(X)$, juntamente com um desvio padrão *aleatório* s em torno do gráfico de função. Em um circuito elétrico, por exemplo, a variável de corrente (Y) é linearmente dependente da tensão (X), mas devido a irregularidades no fio e erros de medição, essa dependência não se aplica exatamente, mas é sobreposta por variabilidade aleatória [dispersão]. Assim, $Y = f(X) + g(s)$, com $g(s)$ como função de distribuição normal com média o e dispersão s (Capítulo 8.6). Uma vez que as variáveis matemáticas são elas próprias funções da forma $X:D \rightarrow R$, " $Y = f(X) + g(s)$ " é uma abreviação para a equação " $\forall d \in D: Y(d) = f(X(d)) + g(s)$ " – ou *em palavras*: para todos os indivíduos d em D , seu valor Y é igual à função f aplicada ao seu valor X mais um desvio padrão aleatório que é em média s . A evidência E é dada por um conjunto de m pontos de dados medidos no sistema de coordenadas X - Y , $E = \{(x_1, y_1), \dots, (x_m, y_m)\}$. A hipótese do quadro inadequado $\exists xHx$ afirma um certo tipo de dependência funcional entre X e Y , por exemplo, que f é uma função linear ou quadrática.

O ajuste de curvas é geralmente realizado com funções polinomiais, uma vez que outras funções arbitrárias podem ser aproximadas com precisão arbitrária com a ajuda de polinômios adequados. Uma função polinomial (de um dígito) de grau n tem a forma $Y = c_0 + c_1 \cdot X + c_2 \cdot X^2 + \dots + c_n \cdot X^n$, com o c_i como constantes de valor real adequadamente escolhidas, os chamados "coeficientes". Para $n = 1$ isso resulta em um polinômio linear, para $n = 2$ um polinômio quadrático, e assim por diante. Assim, as hipóteses de framework correspondentes $\exists x H_n x$ têm a forma $\exists p_0, \dots, p_n, \delta \in \mathbb{R}_+ : Y = p_0 + p_1 \cdot X + \dots + p_n \cdot X^n + g(\delta)$, com o p_0, \dots, p_n como $n+1$ parâmetros livremente selecionáveis de $\exists x H_n x$, que são executados em números reais positivos. $\exists x H_n x$ afirma assim que Y depende de X polinomialmente em grau n , com coeficientes desconhecidos e dispersão aleatória dependendo dele (" x " é, portanto, uma abreviação para a sequência de parâmetros " $\langle p_0, \dots, p_n, \delta \rangle$ "). A hipótese ajustada $H_n c_E$ é a curva polinomial de grau n que (entre todas as curvas polinomiais de n graus) se aproxima de forma ideal dos pontos de dados E . O método estatístico padrão para encontrar o polinômio ótimo de um dado grau é o "Método de minimização da soma dos quadrados das diferenças", ou *minimização SAQ* para abreviar. Esse método encontra entre todas as curvas do tipo $\exists x H_n x$ a curva que minimiza a soma dos desvios quadrados dos pontos de dados observados dos pontos de dados previstos pela função $SAQ = \sum_{i=1}^m (Y - f(X_i))^2$. Esses valores são calculados formando as derivadas do SAQ de acordo com os coeficientes c_i , definindo zero, e determinando o c_i a partir disso (ver Bortz 1985, p. 219-241). Substituindo os coeficientes variáveis (p_i) da curva não ajustada pelos coeficientes (c_i) calculados desta forma, obtém-se a curva $H_n c_E$ ajustada de forma ótima, em que o valor do

desvio padrão residual aleatório (δ) é substituído pelo desvio padrão corrigido $s = \frac{\sqrt{SAQ}}{m-1}$ da curva ótima.

O método de minimização SAQ calcula a curva ótima de um determinado tipo de curva, mas não nos diz *qual* tipo de curva usar para aproximar a quantidade de dados. Encontrar o tipo certo de curva ou grau de polinômio é o principal problema filosófico do ajuste de curvas (cf. Glymour 1981, cap. VIII). O problema reside no fato de que *qualquer conjunto de pontos de dados m pode ser aproximado por qualquer função polinomial* a uma variável de espalhamento residual s. O espalhamento residual torna-se mais estrito quanto maior o grau do polinômio é escolhido, e assume o valor zero se $n+1 \geq m$ se mantém, ou seja, se o polinômio tem pelo menos tantos parâmetros livremente selecionáveis quanto existem pontos de dados. Como exemplo, vamos apenas olhar para os pontos de dados *brancos* e as duas curvas mostradas na Fig. 9-3 e ignorar os pontos cinzas. As duas curvas são o resultado do ajuste de uma curva linear versus uma curva altamente polinomial, H_{lin} versus H_{pol} , nos pontos de dados brancos. É claro que o H_{pol} aproxima melhor os pontos de dados brancos do que o H_{lin} porque o H_{pol} contém parâmetros selecionáveis mais livremente.

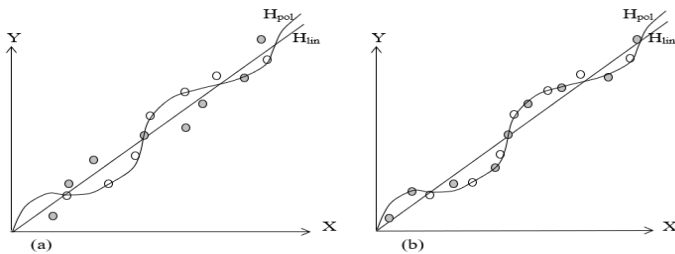


Fig. 9-3: Ajuste da curva e dados não utilizados. Os dados usados para o ajuste (E) são mostrados em branco, os novos dados (não utilizados) (E') em cinza. a): A curva linear é confirmada por E' . (b): A curva altamente polinomial é confirmada por E' .

Mas isso significa que H_{pol} é melhor confirmado por E do que H_{lin} ? Não, porque o H_{pol} também poderia ter sobreajustado os dados, ou seja, a curva H_{pol} poderia ter se adaptado à aleatoriedade da amostra e não à verdadeira dependência sistemática entre X e Y (cf. Hitchcock e Sober 2004).

O método pós-fato de minimização do SAQ não pode nos dizer qual é o tipo correto de curva – pelo menos não no contexto aqui assumido, no qual não há informações *independentes* sobre o verdadeiro valor da variação residual.⁷⁶ Como, então, podemos escolher racionalmente entre a curva linear e a curva altamente polinomial? Uma resposta tradicional para isso é que geralmente devemos preferir a curva *mais simples*, ou seja, aquela que tem menos parâmetros livremente selecionáveis (Schlesinger, 1974). No entanto, essa resposta é duvidosa, porque a simplicidade em si é um critério subjetivo: por que as leis objetivas de nosso mundo devem ser sempre as mais simples possíveis?

Precisamos de um critério objetivo, e encontramos-lo na nossa abordagem de confirmação genuína através de provas não utilizadas. Não é possível confirmar genuinamente um determinado tipo de curva polinomial ajustando post-fato a apenas um conjunto de dados (sem informações independentes sobre o desvio padrão residual). Essa confirmação genuína só pode ser alcançada testando novamente um novo conjunto de dados que não foi utilizado para ajustar a curva. A curva que é confirmada por um conjunto de dados independente é mais genuinamente confirmada e deve ser preferida.

⁷⁶ Se o verdadeiro desvio padrão σ for conhecido, a probabilidade da variância da amostra encontrada dada σ pode servir como um critério de preferência *ceteris paribus*.

A Fig. 9-3 mostra esse novo conjunto de dados na forma de pontos de dados cinzas. No caso (a) à esquerda, os novos pontos de dados constituem evidência independente para a curva linear, porque estão comparativamente distantes da curva polinomial sinuosa, mas estão dentro do desvio aleatório esperado da curva linear. No caso (b) mostrado à direita, por outro lado, os pontos de dados cinza estão muito próximos da curva polinomial sinuosa e, portanto, suportam a curva polinomial.

Note que o papel dos dois conjuntos de dados E e E' também poderia ser invertido: E' poderia ter sido usado para ajuste e E para verificação independente. Assim, a confirmação de uma hipótese por um conjunto de dados depende da informação processual sobre se essa quantidade de dados foi ou não utilizada para adequação. Temos isso na nossa ortografia " Hc_E ", que afirma que Hc foi ganho pelo ajuste de $\exists x Hx$ para E .

Em estatística, a escolha de hipóteses de estrutura com parâmetros livremente selecionáveis também é chamada de seleção de modelos. Um método bem conhecido de seleção de modelos é a *validação cruzada*. Você começa com uma grande quantidade de dados $E = \{(x_i, y_i) : 1 \leq i \leq m\}$, divide aleatoriamente E em dois subconjuntos de dados disjuntos E_1 e E_2 , ajusta a hipótese de estrutura com a ajuda de E_1 e testa a hipótese ajustada contra o conjunto de dados E_2 . Para cada hipótese concorrente, este procedimento é repetido várias vezes e calcula sua probabilidade média ($\overline{P(E_2 | Hc_{E_1})}$) em relação ao segundo conjunto de dados. O resultado é uma medida altamente confiável de confirmação genuína. Dois métodos relacionados são o critério de informação bayesiano (BIK) e o critério de informação de Akaike (AIK). Esses critérios são baseados no valor probabilístico *esperado* da medida de confirmação validada de forma cruzada, por

meio da qual são estimadas as variações necessárias para o cálculo do resultado do ajuste ao conjunto de dados (Hitchcock/Sober, 2004). De acordo com um teorema matemático (Shao, 1997), o resultado de uma validação cruzada m -de- n para n grande converge contra a medida BIK, e o resultado de uma validação cruzada 1 -de- n converge contra a medida AIK. No entanto, Paulßen (2012, cap. 8) usou simulações computacionais para mostrar que, para amostras pequenas, a medida de confirmação da validação cruzada produz resultados muito melhores do que os critérios BIK e AIK, o que demonstra novamente a superioridade do critério de evidência não utilizado no contexto do ajuste da curva.

9.11 Probabilidade e aceitação

Qual é a relação entre a probabilidade epistêmica de uma proposição S (seja uma hipótese ou uma afirmação empírica) e os critérios para sua aceitação racional, ou seja, a crença de que a proposição é verdadeira? É claro que a confirmação incremental de S através da evidência disponível E não é suficiente: para que S seja aceita, a probabilidade $P(S|E)$ não só deve ser aumentada, mas deve ser suficientemente elevada. Deve ter um valor o mais próximo possível de 1, que, no entanto, pode ser menor que 1 devido à possibilidade de erro, que nunca pode ser completamente descartado, mas deve estar acima de $\frac{1}{2}$. Além disso, de acordo com o princípio de Carnap de evidência total mencionado acima, E deve conter todas as evidências atualmente conhecidas relevantes para S (ver capítulo 7.4). Em resumo, a seguinte relação entre probabilidade epistêmica e aceitação racional ou crença racionalmente fundamentada pode ser formulada:

(9-15) *Regra de aceitação de Locke*: A aceitação de uma hipótese como verdadeira é racional, relativa a uma dada função de probabilidade epistêmica P e uma evidência total E , sse $P(S|E) \geq \alpha > 1/2$, onde α é um limiar de aceitação determinado contextualmente.

A regra (9-15) foi proposta pela primeira vez por John Locke e, portanto, também é chamada de regra de Locke (Foley 1992; Leitgeb 2013, p. 1344). Por trás desta regra inocente estão dois problemas difíceis, nomeadamente (a) a questão da determinação de α , e (b) o problema do fecho sob a formação de conjunções.

No que diz respeito à alínea a), coloca-se a questão de *saber quão elevado* ou próximo de 1 deve ser o limiar de aceitação α e *que contexto* é relevante para a resposta a esta questão. Propomos distinguir dois tipos de "contextos" e significados relacionados de "aceitação racional":

- i) *Contexto prático*: Aqui, a aceitação racional de uma proposta (ou teorema) S significa confiar na verdade de S em certas ações práticas.
- ii) *Contexto epistêmico*: Nesse contexto, a aceitação racional de S significa nossa opinião revisável (ou "crença" no sentido não religioso da palavra) de que S é verdadeiro, como parte de nossa imagem cognitiva do mundo.

No contexto prático, o limiar de aceitação depende dos valores de *utilidade* prática de nossas ações, que por sua vez dependem da verdade da proposição S em questão. Como exemplo, vamos considerar a previsão S de que *não haverá* terremoto de magnitude $> 6,5$ nos próximos dias. Se S for aceito, então as pessoas permanecerão nas suas casas durante os próximos dias, o que não implica custos se S for verdadeiro, mas implica custos K_{alto} muito elevados e possivelmente fatais se S for falso. Se, em vez disso, $\neg S$ for aceito, isso traz os custos K_{baixo} comparativamente baixos de evacuação em ambos os casos.

(9-16) *Cálculo do limite prático de aceitação:*

<i>Matriz de Utilidade:</i> Ações possíveis:	Circunstâncias possíveis:	
	S (sem terremoto)	\neg S (terremoto)
Aja conforme S (não evacue)	o	\neg k _{alto}
Aja conforme \neg S (evacue)	- k _{baixo}	- k _{baixo}

Utilidade esperada: EN (agir conforme S) = $P(S) \cdot o - (1 - P(S)) \cdot k_{\text{alto}} = k_{\text{alto}} - P(S) \cdot k_{\text{alto}}$
 EN (agir conforme \neg S) = $-P(H) \cdot k_{\text{baixo}} - (1 - P(H)) \cdot k_{\text{baixo}} = -k_{\text{baixo}}$.

Máxima de racionalidade: aceitar S racional sse $E(\text{agir de acordo com S}) > E(\text{agir de acordo com } \neg S)$.

Este é o caso se $P(H) > (k_{\text{alto}} - k_{\text{baixo}}) / k_{\text{alto}} =_{\text{def}} \alpha$.

α é o *limite prático de aceitação* para aceitar S como base para a ação.

Exemplo: Se $k_{\text{alto}} = 100 \cdot k_{\text{baixo}}$, então resulta $\alpha = 99\%$.

A utilidade *esperada* (UE) e o limiar de aceitação resultante são calculados de acordo com a fórmula da teoria da decisão explicada em (5-3), como mostrado em (9-16). O exemplo mostra que o limiar de aceitação prática pode ser muito alto quando o custo de uma crença errônea é muito alto.⁷⁷

No contexto epistêmico, o limiar racional para a aceitação de nossas hipóteses incertas depende do valor puramente *epistêmico* (epistemológico) que essas hipóteses possuem para nós no caso de sua verdade ou falsidade (cf. Levi, 1967). No entanto, Wilholt (2009)

⁷⁷ Uma aplicação prática é o terremoto de L'Aquila, que levou à condenação judicial altamente questionável de sete especialistas em terremotos (ver Schurz, 2013a, p. 326).

mostra que esses valores de utilidade epistêmica não são objetivamente fixos, mas envolvem uma certa quantidade de convenções arbitrárias.

Mesmo que assumamos que os limiares de aceitação podem ser determinados racionalmente em um determinado contexto, ainda há um segundo problema. Um postulado tradicionalmente aceito para a crença racional é que ela deve ser completada com consequências lógicas conhecidas e, portanto, pelo menos com a formação de conjunção:

(9-17) *Regra de conjunção*: A aceitação racional deve ser completada sob a formação de conjunção, ou seja, se um agente racional acredita nas proposições S_1, \dots, S_n , ele também deve acreditar na sua conjunção $S_1 \wedge \dots \wedge S_n$.

Mas a regra de conjunção colide com a regra de aceitação de Locke e produz uma contradição. Isso é demonstrado por dois paradoxos bem conhecidos, o paradoxo *da loteria* de Kyburg (1961) e o paradoxo *do prefácio* de Makinson (1965). No paradoxo da loteria, assume-se n bilhetes de loteria com chances iguais de ganhar e considera-se para cada um desses bilhetes (com o n^o i) a probabilidade de não fazer o acerto principal: $P(\neg H_i) = (n-1)/n$ ("H" para "acerto principal"). Uma vez que, para n suficientemente alto, o valor $(n-1)/n$ excede o limite de aceitação, devemos, de acordo com a regra de Locke, aceitar para cada lote a previsão " H_i " de que ele não fará o acerto principal. Com base na regra da conjunção, segue-se que também devemos aceitar a conjunção dessas previsões, ou seja, a afirmação (1) $\neg H_1 \wedge \dots \wedge \neg H_n$, segundo o qual nenhum lote fará o acerto principal. Ao mesmo tempo, no entanto, temos certeza (assim assumimos) de que a loteria funcionará de acordo com as regras e, portanto, pelo menos um bilhete fará o acerto principal, ou seja, acreditamos (2) $H_1 \vee \dots \vee H_n$, em

contradição com (1). Tomadas em conjunto, as regras de aceitação e conjunção podem levar a crenças contraditórias, o que é racionalmente inaceitável.

No paradoxo do prefácio, temos apenas um conjunto de muitas proposições altamente prováveis S_1, \dots, S_n , todas as quais cruzam o limiar da aceitação e não se apoiam (ou apenas ligeiramente) umas às outras probabilisticamente. O conhecimento de fundo, como no paradoxo da loteria, segundo o qual pelo menos uma dessas proposições está errada, não está presente aqui, e também não se supõe que essas proposições se enfraquecem negativamente (como no paradoxo da loteria, onde $P(\neg H_i)$ vale para $i \neq j$ $P(\neg H_i | \neg H_j)$). Mesmo sob esses pressupostos mais fracos, as duas regras levam a um conflito, porque a probabilidade da conjunção $S_1 \wedge \dots \wedge S_n$ fica cada vez menor quanto mais tempo essa conjunção fica. No exemplo do prefácio de Makinson, S_i é a afirmação de que o número da página i de um livro longo e cuidadosamente revisado está livre de erros. Embora o autor do livro, depois de revisar cada página dele, esteja convencido de que cada página está livre de erros, ele admite em seu prefácio que muito provavelmente ainda haverá um erro não detectado escondido em algum lugar deste livro, porque ninguém é perfeito.

Formalmente, embora para todos os $i \in \{1, \dots, n\}$ (1.) $P(S_i) > \alpha$ se mantenha, devido à falta de dependências probabilísticas positivas, obtemos (2.) $P(S_1 \wedge \dots \wedge S_n) \leq P(S_1) \cdot P(S_2) \cdot \dots \cdot P(S_n) = \alpha^n$. Agora, α^n é menor que $(1-\alpha)$ sse n for maior que $|\log(1-\alpha)| / |\log \alpha|$, que é o que queremos assumir. Assim, devido à regra de conjunção $S_1 \wedge \dots \wedge S_n$ e de acordo com a regra de aceitação de Locke $\neg(S_1 \wedge \dots \wedge S_n)$, o que novamente significa uma contradição. Por exemplo, para $\alpha = 0,95$, n deve ter pelo menos um valor de 59, e para $\alpha = 0,99$, deve ter um valor de 299 para levar à oposição.

Autores anteriores (incluindo Kyburg e Makinson) descartaram a regra de conjunção por causa desses paradoxos, e achamos que essa também é a solução correta. Autores posteriores, como Lehrer (1975, p. 303), Douven (2002, p. 396) ou Leitgeb (2013), tentaram manter a regra de conjunção e, em vez disso, restringiram a regra de aceitação de Locke. Esses autores argumentam a favor de reagir com ceticismo em situações de conflito entre as regras de aceitação e *conjunção*, ou seja, não aceitar nenhuma das afirmações altamente prováveis como racionais, por mais alta que seja sua probabilidade. No caso do paradoxo da loteria, essa solução ainda pode parecer plausível (embora não necessariamente) porque é uma questão puramente "aleatória" e porque as previsões se prejudicam negativamente. No caso do paradoxo do prefácio, no entanto, essa solução proposta leva à consequência fatal de que não se poderia mais acreditar no conhecimento científico, porque, assim como no paradoxo do prefácio, ele contém uma enorme quantidade de dados independentes e hipóteses empíricas.

No contexto epistêmico das hipóteses científicas, então, parecemos abandonar a regra da conjunção: acreditamos na verdade de muitas proposições bem fundamentadas e altamente prováveis, e temos em mente a imagem de sua conjunção como o quadro científico atual da natureza do mundo. Ao mesmo tempo, no entanto, temos certeza de que há outros erros escondidos em algum lugar desse quadro, que o desenvolvimento futuro da ciência trará à luz, ou seja, não acreditamos que a conjunção de todas essas proposições também seja verdadeira.

No entanto, o mesmo se aplica em contextos práticos. Um exemplo são as práticas de seguros. Digamos que alguém dirige para o trabalho todos os dias da semana. Como motorista consciente, ele

tinha um seguro para seu carro; sim, até mesmo um seguro abrangente. Isso significa que ele acha que teria um acidente de carro naquele dia? Claro que não: na manhã de cada dia, essa pessoa acredita que não terá um acidente de carro naquele dia. Mas ele não conclui disso (por regra de conjunção) que nunca terá um acidente de carro nos próximos 10 anos, mas permite isso como uma possibilidade epistêmica e, portanto, contrata um seguro de carro. Em suma, então, parece razoável, não apenas em contextos epistêmicos, mas também práticos, manter a regra de aceitação de Locke em caso de conflito e, em vez disso, abandonar a regra da conjunção.

10 Apêndice lógico-matemático

10.1 Fundamentos lógicos

Para uma melhor compreensão da teoria da probabilidade, é útil familiarizar-se com conceitos lógicos básicos elementares.

Na *lógica*, conceitos e sentenças são classificados de acordo com sua função lógico-semântica. A figura 10-1 apresenta uma visão geral.

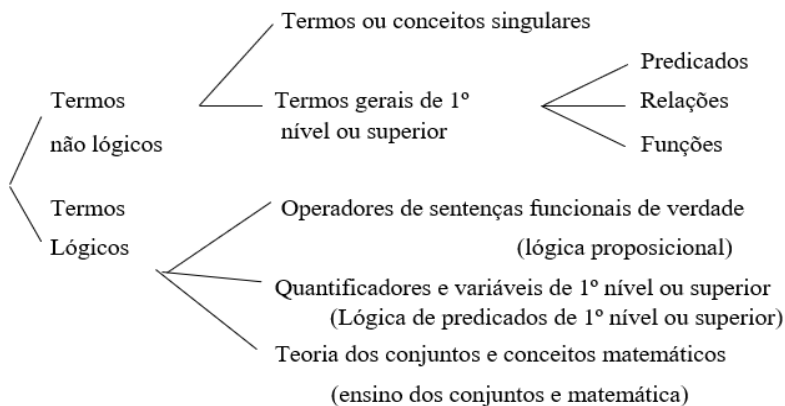


Figura 10-1: Tipos de conceitos lógicos

Termos não lógicos são usados para designar ou expressar algo no mundo. Os conceitos lógicos, por outro lado, têm apenas uma *função estrutural*.

10.1.1. Termos não lógicos

Os termos não lógicos incluem:

10.1.1.1. Termos singulares

ou termos que sempre se referem a um indivíduo específico ou a uma coisa individual. Eles incluem nomes próprios, como: Por exemplo, "Uwe Seeler", "Düsseldorf", termos ostensivos como "aquela pessoa ali" ou termos funcionais como "a mãe de Peter". Notação lógica: Termos singulares primitivos também são chamados de constantes individuais e são denotados por letras minúsculas a, b, \dots (ou indexados a_1, a_2, \dots).

10.1.1.2. Termos gerais

ou *predicados* de 1º nível são aplicados a termos singulares, criando assim uma *sentença atômica*.

10.1.1.2.1. Predicados de primeira ordem (1º nível) denotam características ou propriedades que podem ou não se aplicar a indivíduos, como "x é alto" ou "x é humano". Notação lógica: Predicados de primeira ordem são indicados por letras maiúsculas F, G, \dots (ou F_1, F_2, \dots). Eles têm apenas um argumento, que é adicionado à letra do predicado como uma variável x , por exemplo.

10.1.1.2.2. Predicados de múltiplas ordens ou *sinais de relação* denotam uma relação entre n indivíduos dados. *Notação lógica* por letras maiúsculas R, Q, \dots Por exemplo, se " Rxy " significa a relação de 2 dígitos "x é irmão de y", "a" para "Pedro" e "b" para "Paulo", então a sentença atômica " Rab " significa "Pedro é irmão de Paulo".

10.1.1.3. Símbolos de função (uma ou várias ordens)

Denotam uma função que atribui inequivocamente outro indivíduo a um ou mais indivíduos determinados. *Notação lógica* por minúsculas f, g, \dots . Por exemplo, se " $f(x)$ " para "a mãe de x ", então " $f(a)$ " denota a mãe de a , e " $b = f(a)$ " significa " b é a mãe de a ". Em matemática, as funções desempenham um papel fundamental como operações sobre números, por exemplo, adição $+$ é uma função de dois dígitos que atribui sua soma $+(x,y)$ a dois números x e y , onde a notação infixa " $x+y$ " é escrita para " $+(x,y)$ ". Todas as propriedades quantitativas (por exemplo, físicas) dos indivíduos também devem ser representadas como funções, por exemplo, $m(a)$ denota a massa do corpo a , etc.

10.1.1.4. Variáveis matemáticas e tipos de escalas

"Variáveis matemáticas", que aprendemos mais detalhadamente no Capítulo 8.6, são diferentes das variáveis no sentido lógico e são indicadas pelas letras maiúsculas X, Y, \dots

Entendem-se abaixo funções da forma $X:D \rightarrow W$ como aquelas que atribuem determinados valores de um intervalo de valores W a objetos no domínio individual D . No caso *das variáveis quantitativas*, W consiste nos números reais (\mathbb{R}) conectados a uma unidade. Por exemplo, a função Peso, representada como variável G , atribui seu peso em kg a cada objeto no intervalo selecionado, por exemplo, $G(\text{Josef}) = 75 \text{ kg}$. Entre as variáveis quantitativas, é feita uma distinção entre *escalas de razão*, nas quais apenas a escolha da unidade de escala (por exemplo, "kg") é convencionalmente determinada, e *escalas intervalares*, nas quais tanto a unidade de escala quanto a posição do ponto zero são convencionalmente determinadas (por exemplo, a escala de tempo com 1 ano como unidade e o nascimento de Cristo

como ponto zero). Enquanto as escalas de razões determinam objetivamente razões de magnitudes, as escalas intervalares apenas determinam as razões das diferenças: o ano 2000 d.C., por exemplo, não é "duas vezes mais tarde" que o ano 1000 d.C., mas duas vezes mais tempo decorrido desde o nascimento de Cristo até 2000 d.C. do que até 1000 d.C.

No caso de escalas ordinais, os valores em W consistem em números naturais que expressam simplesmente uma classificação quanto ao grau de expressão de uma característica, como no caso da variável "melhor lista": participantes $\rightarrow \{1, \dots, 20\}$. Finalmente, com escalas nominais, os valores numéricos codificam apenas categorias disjuntas (não sobrepostas) de uma subdivisão de D , por ex. B. "Afiliação de classe": $D \rightarrow \{\text{classe baixa, classe média, classe alta}\}$. Uma propriedade ordinária "F" também pode ser entendida como uma variável binária da forma $XF: D \rightarrow \{F, \neg F\}$. (para mais informações sobre teoria de escala, ver Krantz *et al.*, 2006 e Schurz, 2006, Seção 3.1.4.3).

10.1.1.5. Predicados e funções de nível superior

São aplicados a outros predicados ou teoremas, formando proposições. Um exemplo importante de uma função de nível superior aplicada a proposições é a noção epistêmica de probabilidade: " $P(A) = r$ " significa "o grau de crença da proposição A é r ". O conceito estatístico de probabilidade, por outro lado, é uma função de nível superior que é aplicada a predicados ou conjuntos.

10.1.2 Conceitos lógicos

Os conceitos lógicos podem ser classificados muito mais finamente do que na Figura 10-1. Ela contém apenas as diferenças mais importantes para nós.

10.1.2.1. Operadores de sentenças funcionais de verdade e lógica proposicional: notação lógica

A, B,... representa qualquer frase ou fórmula a seguir. Os operadores de teoremas proposicionais mais importantes são:

- . a negação \neg (leia-se: $\neg A$ – *não A*),
- . a conjunção \wedge (leia-se: $A \wedge B$ – *A e B*)
- . a disjunção (o ou *inclusivo*) \vee (leia-se: $A \vee B$ – *A ou B ou ambos*).
- . a chamada implicação “material” \rightarrow (leia-se: $A \rightarrow B$ – *se A, então B*),
- . a equivalência “material” \leftrightarrow (leia-se: $A \leftrightarrow B$ – *A exatamente se B*).

Convenção: "sse" é a seguinte abreviação para "se e somente se".

Outros operadores de sentenças podem assim ser definidos, por exemplo, o excludente ou $\dot{\vee}$ é definido como $A \dot{\vee} B : \Leftrightarrow (A \vee B) \wedge \neg (A \wedge B)$ (A ou B, mas não ambos).

As leis desses operadores de sentenças são explicadas pela *lógica proposicional (não modal)*. Os operadores da sentença são chamados *de verofuncionais* porque o valor verdade da sentença complexa é inequivocamente determinado pelos valores de verdade de suas orações parciais (argumentos), usando as *conhecidas tabelas de verdade*:

A	$\neg A$	A	B	$A \wedge B$	$A \vee B$	$A \rightarrow B$	$A \leftrightarrow B$
v	F	V	v	v	V	V	v
f	V	V	f	f	V	F	f
		F	v	f	V	V	f
		F	f	f	F	V	v

Nota: a implicação material $A \rightarrow B$ é definida como verdadeira sse A é falsa ou B é verdadeira, ou seja, $A \rightarrow B : \Leftrightarrow \neg A \vee B$. A implicação

material é *mais fraca* do que a linguagem natural se-então, uma vez que sua verdade não pressupõe qualquer conexão substantiva entre A e B.

10.1.2.2 Quantificadores, variáveis individuais e lógica de predicados

Os quantificadores de 1ª ordem quantificam sobre os indivíduos. As variáveis correspondentes de 1º ordem são chamadas *de variáveis individuais*, para as quais escrevemos x, y, \dots (ou x_1, x_2, \dots). Os dois principais tipos de quantificadores são:

- . o *quantificador universal* $\forall x$ (leia-se: $\forall x Fx$ – para todos x : Fx), e
- . o *quantificador existencial* $\exists x$: (leia-se: $\exists x Fx$ – para pelo menos um x : Fx).

O intervalo semântico de indivíduos ao qual um quantificador se refere é chamado de *seu domínio individual* D , que também é chamado de "população" ou "espaço amostral" em estatística. Por exemplo, $\forall x Fx$ diz que todos os indivíduos em D têm traço F , e $\exists x Fx$ diz que há indivíduos em D que têm traço F . Se nada é acrescentado, escolhe-se o reino universal como o reino individual.

Na expressão $\forall x Fx$, a variável individual " x " é limitada pelo quantificador " $\forall x$ ". As expressões cujas todas as variáveis individuais, se ocorrerem, estão vinculadas são chamadas de fórmulas ou sentenças fechadas. Assim, as expressões Fa , Rab , $\forall x Fx$ e $\forall x \exists y (Fx \wedge Gy)$ são fórmulas fechadas. A expressão Fx é chamada *de fórmula aberta* porque sua variável individual " x " é *livre*, ou seja, não está vinculado a nenhum quantificador. Fórmulas abertas como $Fx \wedge Gx$ (x é F e x é G) expressam características gerais complexas. Analogamente, quantificadores e variáveis do 2º (ou superior) nível quantificam via características ou fatos. Não vamos entrar nisso.

Uma relação lógica especial é a identidade: $x = y$ significa "x é idêntico a y". *Convenção*: " $\leftrightarrow_{\text{def}}$ " ou " $=_{\text{def}}$ " significa "equivalente *por definição*" ou "idênticos *por definição*".

A lógica das sentenças formadas a partir de termos singulares, termos gerais de 1º nível, operadores de sentenças extensionais e quantificadores de 1º nível é a lógica de *predicados de 1º nível* (introdução, por exemplo, Klenk 1989, Barwise/Etchemendy, 2005).

A tradução de uma sentença de linguagem natural em uma sentença de uma lógica se chama *formalização*. Veja alguns exemplos: *Traduções*: a – este animal, Rx – x é um corvo, Sx – x é preto.

(a) Este animal é um corvo preto. *Formalização*: $Ra \wedge Sa$.

(b) Todos os corvos são negros. *Formalização*: $\forall x(Rx \rightarrow Sx)$.

(c) Alguns corvos não são negros. *Formalização*: $\exists x(Rx \wedge \neg Sx)$.

Um conjunto da forma Ra, \dots (ou seja, que não contém fórmulas parciais) também é chamado de *sentença atômica*, uma sentença atômica não negada ou negada é chamada de *sentença básica*.

10.1.2.3 Símbolos da teoria dos conjuntos e teoria dos conjuntos

Conjuntos são agregações de indivíduos em um indivíduo complexo chamado "conjunto". Escrevemos letras maiúsculas para quantidades, por exemplo, M, N, ...; em determinados contextos, utilizaremos as letras A, B, ... ao mesmo tempo que quantidades ou fórmulas. Notações específicas:

. $\{x: Fx\}$ significa o conjunto de todos os indivíduos x que possuem a propriedade F.

. $\{a_1, \dots, a_n\}$ significa o conjunto constituído pelos indivíduos a_1, \dots, a_n .

. (a_1, \dots, a_n) , por outro lado, representa a sequência ordenada consistindo em a_1, \dots, a_n .

. $x \in M$ ou $x \notin M$ significa x é um (ou não é um) elemento do conjunto M .

Nota: Os conjuntos são *invariantes* sob reversões e repetições de seus elementos, por exemplo, $\{a,b\} = \{a,a,b\} = \{b,a\}$ (etc.). Em contraste, *seqüências ordenadas* requerem um certo arranjo, que também permite a repetição: $(a,b,b,c) \neq (a,b,c) \neq (b,a,c)$.

A teoria dos conjuntos é considerada *parte da* lógica informal ou metalógica, porque " \in " também é um conceito lógico em um sentido mais amplo. Na teoria dos conjuntos livres de tipos, predicacões da forma como "a é um F" (Fa) são substituídas por *relações de elementos* "a é um elemento da classe de todos os Fs" ($a \in \{x:Fx\}$). Os axiomas da teoria dos conjuntos livres de tipos (de acordo com Zermelo-Fraenkel) podem ser formulados na linguagem da lógica de predicados de 1ª ordem, no entanto, eles vão além dos princípios puramente lógicos (introdução, por exemplo, Ebbinghaus, 2003).

Outros *termos importantes da teoria dos conjuntos*:

. \emptyset (conjunto vazio)

. $M \subseteq N$ (M é um subconjunto próprio ou impróprio de N , ou seja: $\forall x(x \in M \rightarrow x \in N)$)

. $M = N$ (identidade do conjunto), definido como: $\leftrightarrow_{\text{def}} (M \subseteq N) \wedge (N \subseteq M)$

. $M \subset N$ (subconjunto verdadeiro) $\leftrightarrow_{\text{def}} (M \subseteq N) \wedge (M \neq N)$.

. $M \cup N$ (união de M e N) =_{def} $\{x: x \in M \vee x \in N\}$

. $M \cap N$ (interseção de M e N) =_{def} $\{x: x \in M \wedge x \in N\}$

. $\bigcup_{i \in N} X_i$ (ou $\bigcap_{i \in N} X_i$) para a união infinita (ou a média infinita) de todos os conjuntos X_i com i do conjunto de números naturais N .

. $M - N$ (complemento relativo) =_{def} $\{x: x \in M \wedge x \notin N\}$

. $\text{Pot}(M)$ (conjunto de potência de M) $=_{\text{def}} \{N: N \subseteq M\}$, ou seja, o conjunto de todos os subconjuntos de M . *Exemplo:* $\text{pot}(\{1,2\}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$.

. $M \times N$ (produto cartesiano) $=_{\text{def}} \{(x_1, x_2): x_1 \in M \wedge x_2 \in N\}$.

. Análogo para $M_1 \times \dots \times M_n =_{\text{def}} \{(x_1, \dots, x_n): x_i \in M_i \ (1 \leq i \leq n)\}$. Especificamente, $M^n = M \times \dots \times M$ (n vezes) é o produto cartesiano de n cópias de M , ou seja, o conjunto de todas as seqüências de n elementos de M .

. $|M|$ para a cardinalidade de M , ou seja, o número de elementos em M . R é uma *relação de n dígitos* sobre um intervalo de indivíduos D sse (sse) R é um subconjunto de D^n , ou seja, $R \subseteq D^n$. Assim, os elementos de R são seqüências de n elementos de D . R é uma relação de n -ária entre D_1, \dots, D_n sse $R \subseteq D_1 \times \dots \times D_n$.

Uma função de valor único (mapeamento) f é uma relação binária bem definida entre dois conjuntos M e N . Ou seja, para cada x em M , existe exatamente um y em N tal que $(x, y) \in f$. Isso é escrito como: $f: M \rightarrow N$. M é chamado de domínio de f , e $W(f) =_{\text{def}} \{y \in N: \exists x \in M: f(x) = y\}$ é a imagem de f . Similarmente, uma função n -ária $f: M_1 \times \dots \times M_n \rightarrow N$ é uma função de valor único de $M_1 \times \dots \times M_n$ para N . Uma função $f: M \rightarrow N$ é chamada injetiva sse for injetiva à esquerda ($\forall x, y \in M: x \neq y \rightarrow f(x) \neq f(y)$); sobrejetiva sse $W(f) = N$; e bijetiva se for sobrejetiva e injetiva.

10.1.2.4. Os termos matemáticos

Tal como os termos lógicos, os termos matemáticos designam objetos ou funções *abstrato-conceituais*. Eles podem ser caracterizados por *axiomas intrínsecos* de teorias matemáticas ou por *definições da teoria dos conjuntos*. Os conceitos matemáticos mais importantes incluem (i) os termos para diferentes tipos de números

(\mathbb{N} para o conjunto de números naturais e \mathbb{R} para o conjunto de números reais), e (ii) os termos funcionais que expressam operações sobre tais números (adição, multiplicação, etc.). Em particular, $\sum_{1 \leq i \leq n} x_i$ significa a soma de todos os números x_i e $\prod_{1 \leq i \leq n} x_i$ para o produto de todos os números x_i , para i de 1 a n .

10.1.3. Semântica lógica

Na semântica, distinguem-se dois tipos de relação semântica: toda expressão tem, por um lado, um significado ou intensão, e, por outro, uma referência a um objeto (referência) ou extensão (cf. Carnap, 1972; Runggaldier, 1990). A intensão de um conceito é o que todo falante competente deve estar ciente para entender o conceito corretamente. A extensão de um termo singular é o indivíduo que ele denota, a extensão de um predicado é o *conjunto* de indivíduos que se enquadram no predicado, e analogamente a extensão de um símbolo de relação de n dígitos é a classe de todas as n -sequências de indivíduos que satisfazem a relação. A intensão das sentenças é identificada com a proposição denotada pela sentença, e sua extensão (de acordo com a visão tradicional de Frege) com seu valor de verdade.

Na *semântica lógica*, que remonta a Alfred Tarski (1936), referências formais de objetos são atribuídas a expressões linguísticas na forma de *interpretações* extensionais (*isto é, teoria de conjuntos*). Especificamente, define-se: uma *interpretação* de uma linguagem da lógica de predicados L é um par (D, I) , composto por um domínio de indivíduos D e uma função de interpretação I , que associa a cada constante individual a e a cada variável individual x um indivíduo $I(a)$

ou $I(x)$ em D^{78} ; a cada predicado P , um subconjunto $I(P) \subseteq D$; a cada símbolo de relação de aridade n , R , uma relação n -ária $I(R) \subseteq D^n$; e a cada símbolo de função de aridade n , f , uma função n -ária $I(f): D^n \rightarrow D$. Com base nisso, as extensões e valores de verdade de termos singulares arbitrariamente complexos (t_i) e sentenças (A_i) são recursivamente definidos da seguinte forma; onde " $(D, I) \models A$ " é uma abreviação de "Interpretação (D, I) torna a proposição A verdadeira", ou – se A é uma fórmula aberta – "satisfaz a fórmula A ":

. Para termos singulares: $I(f(t_1, \dots, t_n)) = I(f)(I(t_1), \dots, I(t_n))$

. Para conjuntos atômicos: $(D, I) \models R t_1 \dots t_n$ sse $(I(t_1), \dots, I(t_n)) \in I(R)$.

. As conexões lógicas proposicionais são interpretadas de acordo com as tabelas de verdade, por exemplo, $(D, I) \models A \wedge B$ sse $(D, I) \models A$ e $(D, I) \models B$ (etc.).

. Quantificadores: $(D, I) \models \forall x A[x]$ sse para cada indivíduo $d \in D$ detém: $(D, I[x:d]) \models A[x]$; onde $A[x]$ é uma fórmula que contém livremente a variável individual x , e a função de interpretação $I[x:d]$ difere de I apenas na medida em que atribui o indivíduo d à variável individual x .

Uma interpretação (D, I) de uma língua L também é chamada de *modelo* para esta língua. Cada um desses modelos forma um "mundo possível" que pode ser descrito por meio de uma L . Um modelo (D, I) que faz um teorema A verdadeiro também é chamado de *modelo* de A . Um modelo (D, I) no qual cada indivíduo tem d_i , exatamente uma constante individual a_i como seu *nome padrão*, ou seja, $I(a_i) = d_i$, é chamado de *Modelo Padrão*. Os modelos padrão desempenham um papel importante na teoria da probabilidade epistêmica.

⁷⁸ A atribuição de indivíduos às variáveis individuais é chamada também de 'atribuição de variáveis'; ela tem apenas uma função auxiliar para fins de interpretação de fórmulas quantificadas.

Uma proposição A é chamada de logicamente verdadeira (ou L-verdadeira, ou tautológica), ou seja, torna-se verdadeira por todos os modelos possíveis; $\models A$. É chamado de logicamente errado (ou L-falso, contraditório) se estiver errado em todos os modelos possíveis. Por exemplo, $A \vee \neg A$ é uma sentença L-verdadeira e $A \wedge \neg A$ é uma sentença L-falsa. Uma sentença que não é nem L-verdadeira nem L-falsa é chamada *contingente*. A seguir, \top sempre representa uma sentença logicamente verdadeira e \perp uma sentença logicamente falsa. Uma conclusão " $P_1, \dots, P_n / C$ " (de um conjunto de premissas $\{P_1, \dots, P_n\}$ para uma conclusão C) é dita ser *logicamente válida* se todos os modelos possíveis que tornam todas as premissas verdadeiras também tornam a conclusão verdadeira.

Se uma sentença é logicamente verdadeira (ou uma conclusão é logicamente válida), então sua verdade (ou validade) resulta unicamente das supostas definições ou postulados de significado para os conceitos lógicos contidos na sentença (ou conclusão). Além disso, uma proposição é dita analiticamente verdadeira (ou uma conclusão analiticamente válida) se sua verdade (ou validade) logicamente decorre de definições pressupostas ou postulados de significado para conceitos não lógicos (para mais detalhes, ver Schurz, 2006, seção 3.3-4).

Convenção para a distinção sintática entre linguagem-objeto e metalinguagem: A rigor, teríamos que considerar os símbolos da linguagem-objeto e suas interpretações metalinguísticas, seja por meio de outros signos ($I(R) = R$), seja por meio de aspas ($I("R") = R$). Neste ensaio, usamos convenientemente os mesmos símbolos (por exemplo, "R") quando o contexto determina claramente o que se quer dizer.

10.2 Construção lógica de funções estatísticas de probabilidade sobre experimentos aleatórios combinados

Para maior transparência, tratamos primeiro os experimentos aleatórios estatísticos combinados na estrutura linguístico-semântico-estatística. A seguir, seja $o \in \mathbb{N} \cup \{\infty\}$ ou um número natural, ou o símbolo ‘ ∞ ’ para ‘infinito’, entendendo-se por este a menor ordinal infinita, ω . Para cada $o \in \mathbb{N} \cup \{\infty\}$, D^o , o o -ésimo produto cartesiano de D , funciona como o espaço de possibilidades de um experimento aleatório repetido o vezes. Em particular, D^∞ é o espaço de resultados de um experimento aleatório iterado infinitamente, consistindo em sequências infinitas de resultados. A linguagem L contém um número infinito de variáveis individuais, Iv's para abreviar, em uma indexação fixa x_1, x_2, \dots . Para cada $o \in \mathbb{N} \cup \{\infty\}$, a álgebra definida $AL(D^o)$ é uma álgebra sobre D^o , que inclui todas as extensões de fórmulas abertas da língua dada com no máximo o muitos Iv's diferentes (cf. Baco, 1990, p. 83; Adams, 1974). Para definir as extensões de fórmulas abertas de tal forma que elas possam ser recursivamente estendidas para espaços de eventos de dimensões superiores, definimos sua probabilidade estatística em relação a uma sequência de Iv's que inclui pelo menos todos os Iv's na fórmula. Para todos os $n \in \mathbb{N}$, denote no V_n a seguir uma sequência de n membros de Iv's diferentes pareados. Por exemplo, V_3 pode ser (x_1, x_2, x_7) ou (x_{12}, x_3, x_1) . Escrevemos $V_m \subseteq V_n$ quando cada Iv em V_m também ocorre em V_n (possivelmente em posições diferentes). $V(A)$ denota a sequência de todos os Iv's livres que ocorrem na fórmula

A, ordenados na ordem de sua primeira ocorrência da esquerda para a direita. Por exemplo, $V(Rx_1x_2) = (x_1, x_2)$, $V(Rx_2x_1) = (x_2, x_1)$, $V(Fx_1 \wedge Rx_2x_1) = (x_1, x_2)$, etc. Agora usamos v_1, v_2, \dots como variáveis metalinguísticas para qualquer Iv's x_i da linguagem objeto e definimos a extensão $\|A:V_n\|D$ de A em relação a uma sequência Iv $V_n =_{\text{def}} (v_1, \dots, v_n) \supseteq V(A)$ como segue (seguimos Bacchus, 1990, p. 86):

(10-1) *Extensão das fórmulas nos espaços do produto:*

Se $V_n \supseteq V(A)$: $\|A:V_n\|D =_{\text{def}} \{(d_1, \dots, d_n) \in D_n : (D, I[v_1: d_1, \dots, v_n: d_n]) \models A\}$.

Em palavras, a extensão da fórmula A em relação a uma sequência Iv V_n que inclui todos as Iv's em A é o conjunto de todas as n-tuplas de objetos D que, quando mapeadas para os Iv's da sequência V_n da esquerda para a direita, satisfazem a fórmula A.

A extensão simples de uma fórmula A é definida como $\|A\|D =_{\text{def}} \|A:V(A)\|D$

(10-2) Exemplos de (10-1):

(a) Para todos os $i, j \in N$: $\|Rx_i x_j\| = \|Rx_i x\{:(x_i, x_j)\}\| = \{(d_1, d_2) \in D_2 : I[x_i: d_1, x_j: d_2] \models Rx_i x_j\} = I(R)$.

(b) Para todos $i, j \in N$: $\|Rx_i x_j:(x_j, x_i)\| = \{(d_1, d_2) \in D_2 : I[x_i: d_2, x_j: d_1] \models Rx_i x_j\} = I(R) \text{-1}$.

(c) $\|Rx_1 x_2 \wedge Rx_2 x_1\| = \{(d_1, d_2) \in D_2 : I[x_1: d_1, x_2: d_2] \models Rx_1 x_2 \wedge Rx_2 x_1\} = \{(d_1, d_2) \in D_2 : (d_1, d_2) \in I(R) \wedge (d_2, d_1) \in I(R)\} = I(R) \cap I(R) \text{-1}$.

(d) $\|x_1 = x_1 \wedge Rx_2 x_1\| = \{(d_1, d_2) \in D_2 : I[x_1: d_1, x_2: d_2] \models x_1 = x_1 \wedge Rx_2 x_1\} = I(R) \text{-1}$.

(e) $\|Rx_2 x_3 \wedge Fx_3:(x_2, x_3, x_3)\| = \{(d_1, d_2, d_3) \in D_3 : I[x_2: d_1, x_3: d_2, x_3: d_3] \models Rx_2 x_3 \wedge Fx_3\} = \{(d_1, d_2, d_3) \in D_3 : (d_1, d_2) \in I(R) \wedge d_2 \in I(F)\} = \{(d_1, d_2) \in D_2 : (d_1, d_2) \in I(R) \wedge d_2 \in I(F)\} \times D$.

(f) $\|Rx_i x_i\| = \{d \in D : (d, d) \in I(R)\}$.

Por exemplo, é necessário relativizar a extensão das fórmulas para sequências Iv, a fim de distinguir entre $R_{x_1x_2}$ e $R_{x_2x_1}$ em combinações de ambas as fórmulas: Embora $\|R_{x_1x_2}\| = \|R_{x_2x_1}\|$, mas $\|R_{x_1x_2}:(x_1, x_2)\| \neq \|R_{x_1x_2}:(x_2, x_1)\|$ (Exemplos (a)-c)). Se "Rxy" significa a relação "x ama y", então $\|R_{x_1x_2}\|$ e $\|R_{x_2x_1}\|$ todos os casais cujo primeiro membro ama o segundo, mas $\|R_{x_1x_2} \wedge R_{x_2x_1}\|$ todos os casais cujos dois membros se amam. A definição da extensão simples de uma fórmula A está relacionada à sequência variável $V(A)$.

A concordância de que em $V(A)$ as variáveis individuais são organizadas da esquerda para a direita na ordem de sua primeira ocorrência em A é necessária para que fórmulas com n variáveis individuais livres, mas arbitrárias (por exemplo, para $n=3$ ($x_1, x_2, x_{3\circ 01}$), etc.) se refiram à álgebra sobre o espaço do produto D^n (em vez de sempre se referir a D^∞ o que seria necessário se " x_i " sempre se referisse à i.ésima sequência individual). Portanto, $\|x_1=x_1 \wedge R_{x_2x_1}\| = I(R)^{-1} \neq \|R_{x_2x_1}\| = I(R)$ (ver exemplo (a) com (d)).

A relativização para sequências Iv também é necessária para representar corretamente a atribuição entre operações lógicas e teóricas de conjuntos de acordo com (10-3) abaixo para fórmulas com múltiplas variáveis individuais. A extensão de $F_{x_1} \wedge G_{x_2}$ é apenas a média das extensões de F_{x_1} e G_{x_2} se ambas as fórmulas forem avaliadas em relação à mesma sequência Iv (x_1, x_2), ou seja, no espaço do produto D_2 (ou em espaços dimensionais superiores D^o com $o \geq 2$). Em seguida, aplica-se $\|F_{x_1} \wedge G_{x_2}:(x_1, x_2)\|D = \|F_{x_1}:(x_1, x_2)\|D \cap \|G_{x_2}:(x_1, x_2)\|D = \{(d_1, d_2): d_1 \in I(F)\} \wedge \{d_2 \in I(G)\}$. Se, por outro lado, as duas fórmulas são avaliadas em D, então a extensão de $F_{x_1} \wedge G_{x_2}$ é o produto cartesiano de ambas as extensões, ou seja $\|F_{x_1} \wedge G_{x_2}\|D = \|F_{x_1}\|D \times \|G_{x_2}\|D$. Aparentemente, as duas extensões são idênticas. Em geral, aplica-se o seguinte:

(10-3) Aditamento ao teorema (3-5): a) Para $V_n \supseteq V(A) : \|\neg A:V_n\| D = D_n - \|A:V_n\|D$. Para $V_n \supseteq V(A)$ e $V_n \supseteq V(B)$: (b) $\|(A \vee B): V_n\| D = \|A: V_n\| \|D \cup \|B: V_n\| \|D$

(c) Análogo a \wedge e \cap .

Note que cada fórmula em n Iv's livres (distintos) tem uma extensão em todos os espaços de produto pelo menos n -dimensionais. Assim, $AL(D^n)$ é a álgebra da extensão D^n de fórmulas em no máximo n Iv's livres. Assumimos agora uma medida de probabilidade $p:AL \rightarrow [0,1]$ em $AL(D^n)$, que é transferida para fórmulas abertas arbitrárias A e seqüências variáveis $V_n \supseteq V(A)$ como de costume:

$$p(A:V_n) = p(\|A:V_n\|^D).$$

A probabilidade *simples* é definida como $p(A) =_{\text{def}} p(\|A\|D) =_{\text{def}} p(\|A:V(A)\|D)$. A *lei da independência* para probabilidades estatísticas agora assume a forma geral do teorema 10-1(a), com $Iv(A)$ como o conjunto de Iv's de ocorrência livre em A . Duas outras leis que se seguem a isso são a Lei de Projeção e a Lei de Permutação.

(Teorema 10-1) *Leis para probabilidades estatísticas combinadas independentemente* (em fórmulas multivariáveis):

(a) Independência estatística (lei dos produtos):

$$\text{Se } Iv(A) \cap Iv(B) = \emptyset: p(A \wedge B) = p(A) \cdot p(B).$$

(b) Lei de projeção: Para todo $V_n \supseteq V(A)$: $p(A:V_n) = p(A)$.

(c) Lei de permutação: Para todas as permutações de variáveis individuais, $\pi: \mathbb{N} \rightarrow \mathbb{N}$: $p(A(v_1, \dots, v_n)) = p(A(v_{\pi(1)}, \dots, v_{\pi(n)}))$.

A lei de projeção decorre diretamente da lei do produto, porque $p(A(x_1): (x_1, x_2)) = p(A(x_1) \wedge (Fx_2 \vee \neg Fx_2)) = p(A(x_1)) \cdot p(Fx_2 \vee \neg Fx_2)$ (por causa da lei do produto) $= p(A(x_1))$ (porque $p(Fx_2 \vee \neg Fx_2) = 1$). Por outro lado,

$p(Rx_1x_2 \wedge Qx_2x_3) = p(Rx_1x_2) \cdot p(Qx_2x_3)$, porque Rx_1x_2 e Qx_2x_3 têm o $Iv x_2$ em comum. A lei de permutação (teorema 10-1)(c) decorre diretamente de nossa definição da extensão da fórmula simples em (10-1), porque $||A(x_1, \dots, x_n) : (x_1, \dots, x_n)|| = ||A(x_{\pi(1)}, \dots, x_{\pi(n)}) : (x_{\pi(1)}, \dots, x_{\pi(n)})||$.

Note que uma lei analítica não se aplica a *condensações variáveis*, porque $p(Rx_1x_1)$ é geralmente diferente de $p(Rx_1x_2)$, bem como de $p(Rx_1x_2 \wedge x_1 = x_2)$. Por exemplo, apenas 10% de todos os indivíduos D poderiam amar a si mesmos, mas cada indivíduo poderia amar uns aos outros. Então o seguinte se aplica a $|D| = 100$: $p(Rxx) = 0,1$, $p(Rx_1x_2 \wedge x_1 \neq x_2) = 99/100 = 0,99$, $p(Rx_1x_2 \wedge x_1 = x_2) = (1/100) \cdot 0,1 = 0,001$ e, portanto, $p(Rx_1x_2) = 0,991$.

Na estrutura matemática, o espaço de possibilidade de um experimento aleatório combinado também é caracterizado como um *espaço de produto*, que é apenas brevemente delineado aqui (cf. Jeffrey 1971b, p. 196; Bauer, 1978, p. 41, 112). Se (Ω_i, AL_i, p_i) ($i \in \{1, \dots, n\}$) são espaços de probabilidade (que não precisam necessariamente ser idênticos), então seu *espaço de produto* (Ω, AL, p) é definido da seguinte forma: (i) $\Omega =_{\text{def}} \Omega_1 \times \dots \times \Omega_n$, e (ii) $AL =_{\text{def}}$ é a menor álgebra sobre Ω , que contém o produto cartesiano $A_1 \times \dots \times A_n$ para todos os $A_1 \in AL_1, \dots, A_n \in AL_n$. Para cada $A \in \Omega$, $\pi_i(A) =_{\text{def}} \{d \in D_i : \exists (d_1, \dots, d_{i-1}, d, d_{i+1}, \dots, d_n) \in A\}$ é a *i-ésima projeção* de A ; e para cada $A_i \in AL_i$, $e(A_i) =_{\text{def}} \Omega_1 \times \dots \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \dots \times \Omega_n$ é a *extensão projetiva* de A_i . Devido a (ii), AL é a menor álgebra que contém $e(A_i)$ para cada $A_i \in AL_i$ ($1 \leq i \leq n$). A medida de probabilidade p sobre AL satisfaz a lei da independência e a lei da projeção, que agora assumem as seguintes formas⁷⁹ (cf. Bauer 1978, p. 112; Stegmüller 1973b, p. 72):

⁷⁹ A lei de permutação não é mencionada na estrutura matemática; expressaria a invariância da medida do produto em relação às reindexações das álgebras componentes.

(10-4) (a) Lei da Independência, estrutura matemática: $\forall A_i \in AL_i$ ($1 \leq i \leq n$): $p(e(A_1) \cap \dots \cap e(A_n)) = \prod_{1 \leq i \leq n} p_i(A_i)$.

(b) Lei da projeção, estrutura matemática: $\forall A \in AL_i$: $p_i(A) = p(e(A))$.

Com a ajuda de (10-1)(a) define-se a função de probabilidade sobre AL.

Note, no entanto, que a álgebra AL definida desta forma contém apenas tipos de eventos que são combinações de tipos de eventos de experimentos aleatórios simples, ou seja, linguisticamente falando, as extensões de conjunções de fórmulas monádicas. A fim de capturar a probabilidade de relações genuínas, como é possível na estrutura semântica da linguagem, uma álgebra mais abrangente do que AL deve ser escolhida.

10.3 Provas

10.3.1. Prova do teorema 3-1

O teorema (T1) segue diretamente de (A2) e (A3), porque A e $\neg A$ são disjuntos, então $p(A \vee \neg A) = p(A) + p(\neg A) = 1$. De (T1) segue imediatamente (T2) e (devido à “disjuntividade” de $A \vee \neg A$ e $A \wedge \neg A$ e (A3)) também (T3). As provas dos demais teoremas do Teorema 2 são simples na representação algébrica definida (ver Kolmogorov, 1933, § 4); nós os rastreamos na representação linguística. Em relação a (T6): (a) de $\vdash (\neg A_1 \vee A_2)$ segue-se (b) $\vdash \neg(A_1 \wedge \neg A_2)$, ou seja, A_1 e $\neg A_2$ são disjuntos, do qual de acordo com (A3) segue: (c) $P(A_1 \vee \neg A_2) = P(A_1) + P(\neg A_2) =$ (de acordo com T1) $1 + P(A_1) - P(A_2)$. Isso resulta em $P(A_1) \leq P(A_2)$ devido a (T2). Se usarmos (T6) em ambas as direções, segue-se

o teorema (T7), segundo o qual fórmulas logicamente equivalentes têm a mesma probabilidade. Isto significa que (T4) e (T5) também podem ser facilmente derivados. Para (T4) mostra-se primeiro iterando A_3 usando disjunção (d) $p(A_1 \vee \dots \vee A_n) = p(A_1) + \dots + p(A_n)$. Devido à exaustividade de $A_1 \vee \dots \vee A_n$ e (T7), (e) $p(A_1 \vee \dots \vee A_n) = p(A \vee \neg A) = 1$ e $p((B \wedge A_1) \vee \dots \vee (B \wedge A_n)) = p(B)$, ou seja, (T4). Para (T5) mostra-se primeiro $\| \text{---} A \vee B \leftrightarrow A \vee (\neg A \wedge B)$, portanto $p(A \vee B) = p(A \vee (\neg A \wedge B))$; por causa de A_3 e porque A e $\neg A \wedge B$ são disjuntos, segue (f) $p(A \vee B) = p(A) + p(\neg A \wedge B)$. Além disso, $\| \text{---} B \leftrightarrow (B \wedge A) \vee (B \wedge \neg A)$, e por causa de (A_3) portanto $p(B) = p(A \wedge B) + p(\neg A \wedge B)$, portanto (g) $p(\neg A \wedge B) = p(B) - p(A \wedge B)$. De (f) e (g) segue (T5). \square

10.3.2. Prova de (3-1) e (3-2) na seção 3.1

Em relação a (2), primeiro “sse”: (i) Se $p(B) = 0$, então $p(A \wedge B) = p(A) \cdot p(B) = 0$. (ii) Se $p(B) > 0$, então $p(A \wedge B) = p(A|B) \cdot p(B) = p(A) \cdot p(B)$ sse $p(A|B) = p(A)$. Em termos de lógica proposicional: $A \perp B$ sse $A \perp B \wedge (p(B)=0) \vee A \perp B \wedge (p(B)>0)$ (usamos os símbolos lógicos da linguagem de objeto aqui também na metalinguagem.) $A \perp B \wedge (p(B)=0)$ é equivalente a $p(B)=0$ por causa de (i), e $A \perp B \wedge (p(B)>0)$ é equivalente a $p(A|B) > p(A) \wedge (p(B)>0)$. Disto segue que $A \perp B$ sse $p(B)=0 \vee p(A|B) > p(A)$. – O segundo “sse” de (2) pode ser mostrado de forma análoga. – O primeiro “sse” de (3) é mostrado de forma análoga a (2), exceto que em vez de “ $p(B)=0$ ” existe a condição “ $\square \neg B$ ” (B não é satisfeito por nenhum modelo possível), e em vez de “ $p(B)>0$ ” “ $\diamond B$ ”. – Analogamente para o segundo “sse” de (3). \square .

10.3.3. Prova do teorema 3-3

(TB1) é mostrado provando os três axiomas básicos (Teorema 1) para a função de probabilidade condicional $p_B(-)$; veja Carnap (1971, p. 41, T1-4). – (TB2) é válido, porque de $\| \text{---} A \rightarrow B$ segue $\| \text{---} (A \wedge B) \leftrightarrow A$.

Assim $p(A \wedge B) = p(A)$ de acordo com (T7) do Teorema 3-1, do qual $p(B|A) = p(A \wedge B)/p(A) = p(A \wedge B) / p(A \wedge B) = 1$ segue. – (TB3) é uma consequência trivial da Def 3-2. – (TB4) é válido, pois, sob as suposições feitas, os eventos $A \wedge B_1, \dots, A \wedge B_n$ são disjuntos e sua disjunção é necessariamente equivalente a A , razão pela qual (de acordo com T7 do Teorema 3-1) $p(A) = p((A \wedge B_1) \vee \dots \vee (A \wedge B_n)) = \sum_{1 \leq i \leq n} p(A \wedge B_i) = \sum_{1 \leq i \leq n} p(A|B_i) \cdot p(B_i)$ (devido a T4 de Teorema 3-1 e TB3 da sentença 3-3). – (TB5) pode ser visto de $p(B|A) \cdot p(A) \stackrel{\text{def}}{=} p(A \wedge B) \cdot p(A) / p(A) = p(A \wedge B) = p(A/B) \cdot p(B)$. – (TB6) segue inserindo (T4) em (TB5).

(TB7) aplica-se sem a observação entre parênteses, desde que permitamos que uma afirmação da forma “ $p(A|B) > (=, <) \dots$ ” seja falsa se é o caso em que $p(B) = 0$ (ou com axiomatização direta $\neg \square B$).

Se $p(B) = 0$, então as afirmações à esquerda e à direita dela são falsas “sse” com base nesta convenção, e a afirmação do meio é falsa porque $p(B|A) = p(B) = 0$. Caso $p(B) = 1$, então as afirmações à esquerda e à direita do “sse” à esquerda é falsa porque então $p(A|B) = p(A)$ e $p(B|A) = p(B)$, e a afirmação à direita da direita do “sse” é falsa devido à nossa convenção porque $p(\neg B) = 0$. Isso pode ser mostrado de forma análoga para $p(A) = 0$ ou $= 1$. – Agora suponha que $1 > p(B)$, $p(A) > 0$ (ou com axiomatização direta $\diamond A, \diamond B, \diamond \neg A, \diamond \neg B$). A primeira metade de (TB7) resulta então de $p(A|B) > p(A)$ (de TB5). Vê-se que a segunda metade porque $p(A)$ é uma média ponderada de $p(A|B)$ e $p(A|\neg B)$ com pesos $p(B)$ e $p(\neg B) = 1 - p(B)$, respectivamente (TB4, caso especial). Assim, $p(A)$ deve estar (próprio ou impróprio)* entre $p(A|B)$ e $p(A|\neg B)$. De $p(A|B) > p(A)$ segue que $p(A|B) > p(A) \geq p(A|\neg B)$. A outra direção pode ser vista assim: Se $p(A|B) > p(A|\neg B)$, mas $p(A|B) = p(A)$, o peso teria que ser $p(B) = 1$; No entanto, isso é impossível devido à nossa suposição $p(B) < 1$; portanto $p(A|B) > p(A)$. \square

10.3.4. Prova do teorema (4-1)

Em relação a $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv)$: $(i) \Rightarrow (ii)$ é válido porque devido a $(i) A_1 \wedge \dots \wedge A_n \wedge B$ é logicamente equivalente a $A_1 \wedge \dots \wedge A_n$, que é porque de acordo com (T7) do Teorema 3-1 $p(B|A_1 \wedge \dots \wedge A_n) = p(A_1 \wedge \dots \wedge A_n \wedge B) / p(A_1 \wedge \dots \wedge A_n) = 1$ deve ser válido. Isto leva diretamente a (iii) e (iv) . – Em relação a $\neg(i) \Rightarrow \neg(ii) \Rightarrow \neg(iii) \Rightarrow \neg(iv)$: Se $A_1, \dots, A_n \parallel / \text{---} B$, existem modelos M que verificam $A_1 \wedge \dots \wedge A_n$ e falsificam B ; Damos ao conjunto desses modelos a probabilidade positiva x e ao conjunto de modelos que verificam $A_1 \wedge \dots \wedge A_n$ e B a probabilidade variável y . Então: $p(B|A_1 \wedge \dots \wedge A_n) = p(B \wedge A_1 \wedge \dots \wedge A_n) / (p(A_1 \wedge \dots \wedge A_n) = y / (x+y) < 1$ (porque x é positivo) ; portanto $\neg(ii)$. Isso leva diretamente a $\neg(iii)$, e assumindo adicionalmente $P(A_1 \wedge \dots \wedge A_n) = x+y = 1$, $\neg(iv)$. \square

10.3.5. Prova de que as durações mínimas dos períodos de oscilação aumentam exponencialmente sem um limite de frequência (Nota no segundo parágrafo após (5-1)):

Dada uma sequência de uns e zeros de comprimento n com frequência relativa de 1, $h_n = a$. O comprimento mínimo de uma continuação desta sequência de m unidades para conduzir a frequência 1 até $h_{n+m} = b$ ($0 \leq a < b < 1$) é dado pela equação $(a \cdot n + m) / (n + m) \geq b$. Segue-se que $m \geq n \cdot (b - a) / (1 - b)$. O comprimento mínimo de uma continuação desta sequência de p zeros, que empurra o limite de frequência de volta abaixo de a , é dado pela equação $(a \cdot n + m) / (n + m + p) \leq a$, da qual $p \geq m \cdot (1 - a) / a$ e portanto (substituído por m) $p \geq n \cdot (b - a) \cdot (1 - a) / (1 - b) \cdot a$ segue. A partir disso calcula-se $m + p \geq n \cdot (b - a) / (1 - b) \cdot a$. – Abreviamos o valor $(b - a) / (1 - b) \cdot a$ por k . $k \cdot n = m + p$ é, portanto, a duração mínima de um período de oscilação que começa no n -ésimo elemento subsequente. A duração mínima de uma

sequência de r períodos, L_r , é calculada iterativamente como $L_r = (1+k)^r - 1$. Porque $L_1 = 1$, e $L_r = L_{r-1} + L_{r-1} \cdot k = (por\ suposição\ de\ indução) (1+k)^{r-2} + k \cdot (1+k)^{r-2} = (1+k)^{r-1}$. A duração do r -ésimo período, P_r , resulta disso como $P_r = (L_r - L_{r-1}) = (1+k)^{r-1} - (1+k)^{r-2} = (1+k)^{r-2} \cdot k$. Isso significa que P_r cresce exponencialmente com $r-2$. \square

10.3.6 Prova do teorema 5-1

(Etapa 1): Seja (k_1, \dots, k_m) qualquer sequência crescente de m números naturais. Seja $f_{n,m}$ a sequência infinita de subsequências de m membros em g com números de dígitos $((a_i + k_1, \dots, a_i + k_m): i \in \mathbb{N})$. De acordo com a lei do produto (3-3), o limite de frequência de sequências de m membros de k E's e $m - k$ -E's ($k < m$) nesta sequência é $p^k \cdot (1-p)^{m-k}$, porque para cada membro $k_j, j \in \{1, \dots, m\}$, das sequências de m membros, a sequência $(a_i + k_j; i \in \mathbb{N})$ pode ser calculada independentemente do resultado e é, portanto, uma escolha permitida de posições que não altera o limite de frequência. Isso leva à lei binomial: (*) o limite de frequência de sequências de m termos em $((a_n + k_1, \dots, a_n + k_m): n \in \mathbb{N})$ com $h_m(Ex) = k/m$ é: $\binom{m}{k} \cdot p^k \cdot (1-p)^{m-k}$.

(Etapa 2): Usando o resultado (*), calcula-se para qualquer sequência ascendente de números (mesmo não computável) $(k_1, \dots, k_i, \dots: i \in \mathbb{N})$ para a sequência de sequências de resultados $((a_n + k_i; i \in \mathbb{N}): n \in \mathbb{N})$, conforme explicado abaixo Teorema 3-4, o fraco e, assumindo σ -aditividade, também a lei forte dos grandes números. \square

10.3.7. Prova do teorema 6-1

O axioma (A1) é cumprido pela definição do quociente de apostas justo para cada função de crença.

Direção \Leftarrow : Para axiomas (A2) e (A3) provamos essa direção por contraposição e mostramos que se (A2,3) não forem cumpridos, há um sistema de apostas que produz um retorno negativo em todos os

mundos possíveis. Suponha que (A_2) não seja cumprido e você aposte *contra* a proposição $p \vee \neg p$ com um quociente de apostas de $q < 1$ que você considera justo. Assim, você aposta na proposição $\neg(q \vee \neg q)$ com um quociente de apostas de $(1-q) > 0$. Então, em todos os estados possíveis do mundo, você perde $(1-q) \cdot e$, porque o caso $\neg(p \vee \neg p)$ nunca ocorre; ou seja, o sistema de apostas que consiste apenas nesta aposta é incoerente. – Suponha que você viole suas probabilidades de apostas justas (A_3) . Então você aceita o sistema de apostas $WS = \{W_A, W_B, W_{\neg(A \vee B)}\}$, consistindo em duas apostas justas W_A, W_B em duas proposições disjuntas A e B (ou seja, $A \wedge B$ é impossível) e uma aposta justa $W_{\neg(A \vee B)}$ contra a proposição $A \vee B$, onde seu quociente de apostas $q_{A \vee B}$ (em contradição com o axioma A_3) é assumido como sendo menor que $q_A + q_B$. Então seu rendimento de aposta nos três mundos possíveis, com e como a aposta constante, fica assim (porque $q_{\neg(A \vee B)} = 1 - q_{A \vee B}$):

$$A \wedge \neg B: (1 - q_A) \cdot e - q_B \cdot e - (1 - q_{A \vee B}) \cdot e = (q_{A \vee B} - (q_A + q_B)) \cdot e$$

$$\neg A \wedge B: -q_A \cdot e + (1 - q_B) \cdot e - (1 - q_{A \vee B}) \cdot e = (q_{A \vee B} - (q_A + q_B)) \cdot e$$

$$\neg A \wedge \neg B: -q_A \cdot e - q_B \cdot e + q_{A \vee B} \cdot e = (q_{A \vee B} - (q_A + q_B)) \cdot e.$$

Nos três mundos, seu rendimento total $(q_{A \vee B} - (q_A + q_B)) \cdot e$, e essa quantidade é negativa porque $q_{A \vee B} < (q_A + q_B)$. (Se o seu $q_{A \vee B}$ for maior que $q_A + q_B$, a prova é análoga ao sistema de contra-apostas $WS^* = \{W_{\neg A}, W_{\neg B}, W_{A \vee B}\}$.)

Direção \Rightarrow : Seguimos Howson e Urbach (1996, p. 86). O retorno de uma aposta simples W_i , $g(W_i)$, pode ser considerado uma variável aleatória, com os dois valores g_i se a proposição de aposta A_i for verdadeira, e $-v_i$ se A_i for falsa. Como explicado acima, o valor esperado de qualquer aposta simples justa $E(W_i) = q(A) \cdot g_i - q(\neg A) \cdot v_i$ é zero. Se q satisfaz os axiomas da probabilidade, então, de acordo com um teorema bem conhecido sobre valores esperados (teorema 8-2), o

valor esperado de uma soma de variáveis (apostas) é a soma dos valores esperados dessas variáveis (apostas). Assim, se q satisfaz os axiomas da probabilidade, o valor esperado de um sistema de apostas que consiste em apostas simples justas também deve ser zero. Mas se q fosse incoerente, então o retorno do sistema de apostas é negativo em todos os mundos possíveis, e então o valor esperado do sistema de apostas também deve ser negativo. Portanto, se q satisfaz os axiomas de probabilidade, então q deve ser coerente. \square

10.3.9. Prova do teorema 7-1

Para (a): Seja H^* uma hipótese estatística completa que especifica $p(A)$ para todas as fórmulas abertas de L (para a formulação de H^* podem ser necessários operadores lógicos proposicionais infinitos). Assumimos um mapeamento bijetivo $\pi: K=V$ de todas as constantes individuais em variáveis individuais de L . Para cada frase $A \in \text{Sent}(L)$, que $\pi(A)$ seja o resultado da substituição das constantes individuais de A por variáveis individuais. Em seguida, definimos, para todos $A \in \text{Sent}(L)$, $P(A|H^*) = p(\pi(A))$. A função resultante $PH^* =_{\text{def}} P(-|H^*)$ é coerente porque p é coerente. PH^* também satisfaz as leis específicas (a) e (c) do teorema (10-1), mas agora para constantes individuais: independência probabilística e invariância de permutação (também chamada de permutabilidade; veja abaixo). – Em seguida, mostramos que para cada hipótese estatística H que decorre de H^* e implica o mesmo valor de $p(\pi(A)) =_{\text{def}} r$ que H , $P(A|H^*) = P(A|H)$. H é L -equivalente com uma disjunção possivelmente infinita de hipóteses completas disjuntas H_i^* ($i \in I \subseteq \mathbb{N}$), todas as quais implicam $p(\pi(A)) = r$. Assim, $P(A|H) = \sum_i P(A \wedge H_i^*) / \sum_i P(H_i^*)$. Agora

$P(A \wedge H_i^*)/P(H_i^*) = r$ para todos os $i \in I$, ou seja, $P(A \wedge H_i^*) = r \cdot P(H_i^*)$, do qual segue $P(A | \bigvee_i H_i^*) = \sum_i r \cdot P(H_i^*) / \sum_i P(H_i^*) = r$.⁸⁰

Para (b): De acordo com o StK, H é seguido por uma declaração de probabilidade da forma $p(A^*|B^*) = r$. Temos que mostrar que de H também $p(A^*|B^* \wedge E^*) = q$. Se a condição de admissibilidade for atendida, então $K(E) \cap K(A, B) = \emptyset$, e assim também $V(E) \cap V(A, B) = \emptyset$ (com $K(A)$ e $V(A)$ como o conjunto de constantes individuais contidas em A, respectivamente, variáveis individuais livres), e como p satisfaz a independência estatística (teorema 10-1(a)), obtemos $p(A^*|B^* \wedge E^*) = p(A^* \wedge B^* \wedge E^*) / p(B^* \wedge E^*) =$ (devido à independência) $= p(A^* \wedge B^*) \cdot p(E^*) / p(B^*) \cdot p(E^*) = p(A^* \wedge B^*) / p(B^*) = p(A^*|B^*) = r$.

Para (c): A violação do inverso de (a) foi mostrada na seção 7-1, antes de (7-1), pelo exemplo da hipótese $H = (p(Fx|Gx) = 0,5) \wedge (p(Fx|Qx) = 0,8)$: ao violar a condição de admissibilidade, obteve-se $P(Fa_i | Ga_i \wedge H \wedge Qa_i) = 0,5$ ("Qa_i" é evidência inadmissível) e $P(Fa_i | Qa_i \wedge H \wedge Ga_i) = 0,8$ ("Ga_i" é prova inadmissível). – O inverso de (b) é violado, por exemplo, se $A = Ga$, $B = Fa$, $E = Ea$ e $p(Gx|Fx) \neq p(Gx|Fx \wedge Ex)$. \square

10.3.10. Prova do teorema 7-2(b)

Seja H_r a hipótese $p(Gx|Fx)=r$ e K_s a hipótese $p(Fx)=s$ (para a variável r, s de um conjunto de valores possíveis). Em seguida, aplica-se o seguinte (pelo que, no caso contínuo, as somas devem ser substituídas por integrais):

⁸⁰ No caso contínuo, com R^* como espaço ∞ -dimensional e $p(\pi(A))=p^* \in R^*$: $P(A|H) = \int_{p^* \in R^*} P(A \wedge H_{p^*}) \cdot dp^* / \int_{p^* \in R^*} P(H_{p^*}) \cdot dp^* = \int_{p^* \in R^*} r \cdot P(H_{p^*}) \cdot dp^* / \int_{p^* \in R^*} P(H_{p^*}) \cdot dp^* = r$.

$$\begin{aligned}
 P(Ga|Fa) &= P(Ga \wedge Fa) / P(Fa) = \\
 &= \sum_{r,s} P(Ga \wedge Fa | H_r \wedge K_s) \cdot P(H_r \wedge K_s) / \sum_{r,s} P(Fa | H_r \wedge K_s) \cdot P(H_r \wedge K_s) = \text{(de acordo com 7-1(a) e axiomas básicos)} \\
 &= \sum_{r,s} r \cdot s \cdot P(H_r) \cdot P(K_s | H_r) / \sum_r \sum_s s \cdot P(H_r) \cdot P(K_s | H_r) \\
 &= \sum_r r \cdot P(H_r) \cdot \sum_s s \cdot P(K_s | H_r) / \sum_r P(H_r) \cdot \sum_s s \cdot P(K_s | H_r) = (*). \\
 \text{De acordo com 7-2(a), } \sum_s s \cdot P(K_s | H_r) &= \sum_s P(Fa | K_s \wedge H_r) \cdot P(K_s | H_r) = P(Fa | H_r). \\
 \text{Continuamos assim: } (*) &= \sum_r r \cdot P(H_r) \cdot P(Fa | H_r) / \sum_r P(H_r) \cdot P(Fa | H_r) = \\
 &= \sum_r r \cdot P(Fa) \cdot P(H_r | Fa) / \sum_r P(Fa) \cdot P(H_r | Fa) = \\
 &= \sum_r r \cdot P(H_r | Fa) / \sum_r P(H_r | Fa) = \text{(por causa de } \sum_r P(H_r | Fa) = 1) = \sum_r r \cdot P(H_r | Fa) = \\
 &= \sum_r P(Ga | Fa \wedge H_r) \cdot P(H_r | Fa). \square
 \end{aligned}$$

10.3.11. Prova do teorema 7-3

A equivalência de (1) com (2)(i) remonta a de Finetti (1931), ver de Finetti (1964), Carnap (1980), Hewitt/Savage (1955). Spielman (1976) mostrou que se P é σ -aditiva, segue-se de (1) que com $P=1$ existe um p que satisfaz a independência estatística; Neste caso, (1) e (2i+ii) são equivalentes.

Que a implicação (2)(i) \Rightarrow (3)(i) é válida pode ser visto da seguinte forma: de acordo com (2)(i) para qualquer evento, Ea (no caso discreto), $P(Ea) = \sum_{1 \leq i \leq n} r_i \cdot P(H_i)$, com H_i como a hipótese p $p(Ex) = r_i$. Aplicando isso ao evento tautológico $Ea \vee \neg Ea$, o resultado é $P(Ea \vee \neg Ea) = 1 = \sum_{1 \leq i \leq n} r_i \cdot P(H_i)$.

Assim, $P(H_1 \vee \dots \vee H_n) = \sum_{1 \leq i \leq n} P(H_i) = 1$. A probabilidade subjetiva de que o evento Ex não tenha um limite deve, portanto, ser zero, porque $P(H_1 \vee \dots \vee H_n \vee H_{\text{sem_limite}}) = 1$ segue dos axiomas básicos. Assim, (2)(i) implica que com $P = 1$, todo evento expressável na forma (L) tem um limite de frequência. Analogamente, argumenta-se no caso contínuo, exceto que a soma é substituída por uma integral. – Além disso, decorre de (2)(i) também (3)(ii): Se aplicarmos a equação do teorema 7-2(a) de acordo com (2)(i), com $H_i = p(Ex)=r_i$, obtemos

$P(Ea|H_k) = \sum_{1 \leq i \leq n} r_i \cdot P(H_i|H_k) = r_k \cdot 1 = r_k$; ou seja, a declaração do StK de acordo com (3)(ii). – [(2)(ii) e (3)(iii) são idênticos].

Por outro lado, (3)(i+ii) \Rightarrow (2)(i): $P(Ea)$ no caso discreto é dado pelo teorema de Bayes como $P(Ea) = \sum_{1 \leq i \leq n} P(Ea|H_i) \cdot P(H_i) + P(Ea|X) \cdot P(X)$; onde X é a afirmação de que não existe limite de frequência $p(Ex)$ [que atenda à independência estatística]. De (3i) segue que $P(X) = 0$, portanto $P(H_1 \vee \dots \vee H_n) = 1$ e logo $P(Ea) = \sum_{1 \leq i \leq n} P(Ea|H_i) \cdot P(H_i)$. Disto e de (ii) obtemos $P(Ea) = \sum_{1 \leq i \leq n} r_i \cdot P(H_i)$, ou seja, (2)(i). \square

10.3.12. Prova do teorema 7-5

A seguir, usaremos " x_{1-n} " como abreviação de " x_1, \dots, x_n ". Seja $R_{r,1}, \dots, R_{r,k}$ todas as classes de referência da partição R que atribuem uma probabilidade estatística condicional de r aproximada ao tipo previsto do teorema singular $S(x_1, \dots, x_n)$. Como $S(c_1, \dots, c_n) \in Pt(S(c_1, \dots, c_n))$ foi formado pelo princípio da classe de referência estreita para todas as instanciações, $\{(d_{1i}, \dots, d_{in}) \in D^n : P_t(S(a_{1i}, \dots, a_{in})) = r\}$ – ou seja, o conjunto de n -tuplas de indivíduos para os quais se acredita que S esteja no grau r – é dado pela extensão da disjunção (exclusiva) desses predicados de referência, $R_{r,1}(x_{1-n}) \vee \dots \vee R_{r,k}(x_{1-n})$. A probabilidade estatística $p(\|S(x_1, \dots, x_n)\| \mid \{(d_{1i}, \dots, d_{in}) \in D^n : P_t(S(a_{1i}, \dots, a_{in})) = r\})$ é, portanto, idêntica a $p(S(x_1, \dots, x_n) \mid R_{r,1}(x_{1-n}) \vee \dots \vee R_{r,k}(x_{1-n}))$.

Para provar a calibração para $S(x_1, \dots, x_n)$, precisamos mostrar que essa probabilidade estatística é r . Isso é fácil, porque a disjunção $R_{r,1}(x_{1-n}) \vee \dots \vee R_{r,k}(x_{1-n})$ é exclusivo, e para cada predicado de referência $R_{r,i}$ mantém (aproximadamente): $p(S(x_1, \dots, x_n) \mid R_{r,i}(x_{1-n})) = r$. A partir disso, o que foi dito (com $S =_{\text{def}} S(x_1, \dots, x_n)$ e $R_{r,i}(x_{1-n}) =_{\text{def}} R_i$) segue o seguinte: $p(S \mid R_1 \vee \dots \vee R_n) =$ (devido à disjunção) $\sum_{1 \leq i \leq n} p(S \wedge R_i) / \sum_{1 \leq i \leq n} p(R_i) =$ (por

causa de $p(S \wedge R_i) = p(S|R_i) \cdot p(R_i) = \frac{\sum_{1 \leq i \leq n} p(S|R_i) \cdot p(R_i)}{\sum_{1 \leq i \leq n} p(R_i)}$ (por causa de $p(S|R_i) = r$ para todos i) $r \cdot \frac{\sum_{1 \leq i \leq n} p(R_i)}{\sum_{1 \leq i \leq n} p(R_i)} = r$. \square

10.3.13. Prova do teorema 8-3

Para (i): $\mu(\mu_{sn}(X)) = E((1/n) \cdot (X_1 + \dots + X_n)) = (1/n) \cdot n \cdot (E(X))$ (de acordo com o teorema 8-2(a) e porque os X_i são idênticos distribuídos, ou seja, $E(X_i) = E(X)$ para todos os $1 \leq i \leq n$) $= E(X) = \mu(X)$. – *Para (ii):* $v(sn(X)) = v((1/n) \cdot (X_1 + \dots + X_n)) = (1/n) \cdot n \cdot v(X)$ (de acordo com o teorema 8-3(c), porque os X_i são distribuídos de forma idêntica, ou seja, $v_i(X) = v(X)$, e porque $cov(X_i, X_j) = 0$ vale para $i \neq j$, já que os X_i são estatisticamente independentes um do outro) $= (1/n) \cdot n \cdot v(X)$. – (iii) segue imediatamente.

10.3.14. Prova do teorema 9-3

Para (9-3a): Usa-se o teorema da integração $\int_0^1 x^a \cdot (1-x)^b dx = a! \cdot b! / (a+b+1)! = 1 / ((a+b+1) \cdot \binom{a+b}{a})$ (Billingsley 1995, p. 279). Distto, bem como de (9-3) e da distribuição uniforme $D(x)=1$, o resultado (a) resulta:

$P(Fa_n^k) = \int_0^1 r^k (1-r)^{(n-k)} \cdot 1 dr = 1 / (\binom{n}{k} \cdot (n+1))$, do qual (conforme explicado no texto) segue diretamente (9-3b).

(9-3c) segue de (a) e a consideração de que em uma linguagem com o predicado básico binário F existem exatamente $\binom{n}{k}$ descrições de estado da forma Fa_{n+1}^{k+1} a qual a conjunção $Fa_{n+1} \wedge h_n(F) = \frac{k}{n}$ satisfaz:

$$P(Fa_{n+1} | h_n(F) = \frac{k}{n}) = \binom{n}{k} \cdot P(Fa_{n+1}^{k+1}) / \binom{n}{k} \cdot P(Fa_n^k) = P(Fa_{n+1}^{k+1}) / P(Fa_n^k) =$$

$$\frac{\binom{n}{k} \cdot (n+1)}{\binom{n+1}{k+1} \cdot (n+2)} = \frac{n! \cdot (n+1) \cdot (k+1)! \cdot (n-k)!}{(n+1)! \cdot (n+2) \cdot k! \cdot (n-k)!} = \frac{k+1}{n+2}$$

(9-3d) é obtido a partir de (9-3b), $D(p(Fx)=r) = 1$ e aplicação da regra de Bayes:

$$D(p(Fx)=r \mid h_n(F)=\frac{k}{n}) = P(h_n(F)=\frac{k}{n} \mid p(Fx)=r) \cdot D(p(Fx)=r) / P(h_n(F)=\frac{k}{n}) = \binom{n}{k} \cdot r^k \cdot (1-r)^{(n-k)} \cdot 1 \cdot (n+1).$$

10.3.15. Prova de (9-4)

O esboço de prova de Earman está completo aqui. Definimos $x_i =_{\text{def}} P(\neg Fa_{n+1} \mid Fa_1 \wedge \dots \wedge Fa_n)$. Se essa condição for atendida, então, de acordo com o teste de razão de Cauchy, a soma $\sum_{i \in \mathbb{N}} x_i$ é positiva e finita.

Se este for o caso, então o seguinte também se aplica:

(*): para cada $k \in \mathbb{N}$, a soma $\sum_{i \in \mathbb{N}} k \cdot x_i = k \cdot (\sum_{i \in \mathbb{N}} x)$ é positiva e finita.

De acordo com o critério da sequência nuclear, o produto infinito $\lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} (1+k \cdot x_i)$ é positivo e finito. Agora, por causa de $x_i < 1$ para k suficientemente grande: $(1+x_i)/(1-x_i) < 1+k \cdot x_i$, para todos i . Assim, aplica-se também o seguinte:

$$(**): \lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} ((1+x_i)/(1-x_i)) = (\lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} (1+x_i)) / (\lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} (1-x_i)) < < \lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} (1+k \cdot x_i).$$

Se $\lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} (1-x_i) = \lim_{n \rightarrow \infty} P(\bigwedge_{1 \leq i \leq n} \neg Fa_i) = 0$ fosse o caso, então a expressão média em (**) seria infinita e, portanto, $\lim_{n \rightarrow \infty} \prod_{i \in \mathbb{N}} (1+k \cdot x_i)$ também seria infinito, em contradição com (*) acima. Assim, $P(\bigwedge_{i \in \mathbb{N}} Fa_i)$ mantém > 0 e, portanto, se P é σ -aditivo, $P(\forall x Fx) > 0$. \square

10.3.16. Prova do teorema 9-4 (esboço)

Abreviamos: $p(F) =_{\text{def}} p$, $h(F:s_n) =_{\text{def}} h(s_n)$. De acordo com o pressuposto, para qualquer $q_1, q_2 \in [0,1]$:

$$(1) P(p \in [q_1 \pm a_n] \mid h(s^n)=q_1) = P(p \in [q_2 \pm a_n] \mid h(s^n)=q_2).$$

Rendimentos da transformação bayesiana (para $i \in \{1,2\}$):

$$(2) P(p \in [q_i \pm a_n] \mid h(s^n)=q_i) = P(h(s^n)=q_i \mid p \in [q_i \pm a_n]) \cdot P(p \in [q_i \pm a_n]) / P(h(s^n)=q_i), \text{ onde}$$

$$(2.1) P(h(s^n)=q_i \mid p \in [q_i \pm a_n]) = \int_{q_i-a}^{q_i+a} P(h(s^n)=q_i \mid p=x) \cdot D(x) dx,$$

$$(2.2) P(h(s^n)=q_i) = \int_0^1 P(h(s^n)=q_i \mid p=x) \cdot D(x) dx, \text{ e}$$

$$(2.3) P(p \in [q_i \pm a_n]) = \int_{q_i-a}^{q_i+a} D(x) dx.$$

(2.2) difere de (2.1) na medida em que, além disso, a probabilidade da hipótese $p=x$ está integrada em regiões x fora do intervalo de confiança de 95% da hipótese $p = q_i$, onde a probabilidade amostral $P(h(s_n)=q_i | p=x)$ é muito baixa, independentemente da densidade $D(x) =_{\text{def}} D(p=x)$. Suponha que a integral da hipótese $p=x$ seja maior no intervalo $[q_1 \pm a_n]$ do que no intervalo $[q_2 \pm a_n]$; assim, é menor fora do intervalo $[q_1 \pm a_n]$ do que fora do intervalo $[q_2 \pm a_n]$. Em seguida, a razão das duas integrais (2,1) e (2,2) para a hipótese intervalar $p \in [q_1 \pm a_n]$ com maior densidade em seu intervalo *do que* para a hipótese intervalar $p \in [q_2 \pm a_n]$ com menor densidade em seu intervalo. Além disso, sob o pressuposto $D(p=q_1) > D(p=q_2)$, a densidade inicial (2,3) no intervalo de confiança é maior em q_1 do que no intervalo de confiança em q_2 , de modo que o efeito para a probabilidade é amplificado em (2), ou seja, $P(p \in [q_1 \pm a_n] | h(s_n)=q_1) > P(p \in [q_2 \pm a_n] | h(s_n)=q_2)$. Isso significa que, se a densidade inicial é distribuída de forma desigual ao longo dos intervalos $[q_i \pm a_n]$ de diferentes hipóteses intervalares $p \in [q_i \pm a_n]$, a probabilidade final dessas hipóteses intervalares não pode ser o mesmo valor de 95 % de acordo com a equação (2). Como o intervalo $\pm a_n$ pode ser muito pequeno para n grande, a afirmação segue. \square

10.3.17. Prova do teorema 9-5(b)

Um esboço da prova do teorema (9-5)(b) pode ser encontrado em Kutschera (1972, p. 85, T2.1.6-2). A prova vem da observação de que $P(h_n(F) = k/n) = \binom{n}{k} \cdot \int_0^1 x^k \cdot (1-x)^{(n-k)} \cdot D(x) dx$, e $P(Fa_{n+1} \wedge h_n(F)=k/n) = \binom{n}{k} \cdot \int_0^1 x^{(k+1)} \cdot (1-x)^{((n+1)-(k+1))} \cdot D(x) dx$ (ver prova do teorema 9-3(c)). Conclui-se que:

$$(*) P(Fa_{n+1} | h_n(F)=k/n) = \int_0^1 x^{(k+1)} \cdot (1-x)^{(n-k)} D(x) dx / \int_0^1 x^k \cdot (1-x)^{(n-k)} \cdot D(x) dx.$$

Kutschera não inclui a condição de continuidade acima e argumenta que para n suficientemente grande, devido à lei dos

grandes números, a função de distribuição $x^k \cdot (1-x)^{(n-k)}$ está cada vez mais concentrada sobre o valor $r = k/n$, torna-se cada vez mais íngreme e converge em outros lugares para zero. Portanto $\int_0^1 x^k \cdot (1-x)^{(n-k)} D(x) dx$ difere arbitrariamente pouco de $c \cdot (k/n)^k \cdot (1-k/n)^{(n-k)}$; assim, a expressão em (*) é arbitrariamente pouco diferente de k/n . Kutschera não fornece nenhuma evidência para esse argumento. – Gillies (2000, p. 72) assume esta última afirmação como "plausível". XI.4.6) formula como uma suposição imprecisa que a densidade $D(r)$ deve ser "difusa" em torno de r . – Apresento aqui uma prova alternativa baseada no princípio da estimativa estável de Edwards, Lindman e Savage (1963) (ver Howson/Urbach 1996, p. 361). Como $D(r)$ é contínuo, a inclinação sobre r é finita. Assim, para cada $\epsilon > 0$, por menor que seja, há um intervalo positivo suficientemente pequeno $[r \pm \epsilon]$ em torno de r no qual a densidade $D(r)$ flutua por, no máximo, uma fração ϵ , ou seja, é " ϵ -aproximadamente" constante. Uma vez que $D(r)$ é positivo em toda parte em $[r \pm a]$, $\int_{r-\epsilon}^{r+\epsilon} D(r) dr$ assume uma fração não desprezível da probabilidade total. Os pré-requisitos do princípio da estimativa estável são assim preenchidos, do qual se segue que podemos fazer n tornar-se tão alto (a uma razão constante k/n) que a integral é Integral $\int_0^1 x^k \cdot (1-x)^{(n-k)}$. do que seria obtido se uma distribuição igual fosse assumida – ou seja $\int_0^1 x^k \cdot (1-x)^{(n-k)} dx = 1 / \binom{n}{k} \cdot (n+1)$ de acordo com a sentença 9-3a – apenas por um ϵ dependente monotonicamente fator ϵ^* , onde $\epsilon^* \rightarrow 0$ se $\epsilon \rightarrow 0$. Como resultado, de acordo com (*) e Teorema 9-3(c) que $P(Fa_{n+1} \wedge h_n(F)=k/n)$ difere de $(k+1)/(n+2)$ por apenas ϵ , onde para $n \rightarrow \infty$ esta expressão converge para k/n , e podemos fazer ϵ^* ir ϵ a zero escolhendo intervalos cada vez menores. \square

10.3.18. Prova da proposição 9-6

De acordo com a Proposição 9-2, $D(H_r|h_n(F)=k/n)$ é proporcional a $p_{H_r}(h_n(F)=k/n) \cdot D(H_r) = \binom{n}{k} \cdot r^k (1-r)^{(n-k)} \cdot D(H_r)$. Como na prova da sentença 9-5(b), a área total sob esta função desloca-se cada vez mais acima do valor k/n para n maior, torna-se cada vez mais íngreme e converge para outro lugar em direção a zero. Isso resulta em 9-6(a) e (b). 9-6(a) e (b) também podem ser provados usando o princípio da estimação estável, como na prova do teorema 9-5(b). Assim, a probabilidade final para n suficientemente grande pode ser aproximada por uma distribuição final que teria sido obtida a partir de uma distribuição inicial igual. Uma vez que a distribuição igual é idêntica à distribuição β da forma $\beta_{\frac{1}{2}}$, a proposição 9-6 resulta então de dois fatos conhecidos sobre as distribuições β mencionadas na seção 9.3 (Hays/Winkler 1970, p.233): (1.) O valor esperado de uma distribuição β , $E(x | \beta_n^k(x))$, é k/n e sua variância, $\text{Var}(x | \beta_n^k(x))$, é $k \cdot (n-k) / n^2 \cdot (n+1)$. (2.) Se a distribuição de saída sobre $p(Fx)$ é uma distribuição β da forma β_m^r e há um resultado de amostra $h_n(Fx)=k$, então a condicionalização a este resultado de amostra produz a distribuição final β_{m+n}^{r+k} . Assim, para todas as distribuições de saída, a média é deslocada da forma de uma distribuição β para a média da amostra com n crescente, e a distribuição torna-se mais íngreme (i.e. h., a variância é menor). \square

10.3.19. Prova do teorema 9-8

Para (a) ver Gaifman/Snir (1982). (b) segue de (a) da seguinte forma: Seja $W_{[r;n]}$ o conjunto de todos os mundos em que o aglomerado F nos primeiros n termos da sequência $(\pm_w A_1 \wedge \dots \wedge \pm_w A_n) [r;n]$.

$H_r =_{\text{def}} \{ \omega \mid p(Fx) = r \}$ representam o conjunto de mundos em que " $p(Fx)=r$ " é verdadeiro. Aplica-se o seguinte: $P(H_r \mid h_n(F) = [rn]/n) = P(H_r \cap W_{[rn]}) / P(H_r \cap W_{[rn]} + P(\neg H_r \cap W_{[rn]}))$. Assim $\lim_{n \rightarrow \infty} P(H_r \mid h_n(F) = [rn]/n) = \lim_{n \rightarrow \infty} P(H_r \cap W_{[rn]}) / (\lim_{n \rightarrow \infty} P(H_r \cap W_{[rn]}) + \lim_{n \rightarrow \infty} P(\neg H_r \cap W_{[rn]}))$. $\lim_{n \rightarrow \infty} P(H_r \cap W_{[rn]}) = P(H_r)$ é desconhecido, mas presume-se que seja maior que 0; chamamos esse valor de x . Devido a 9-8(a) (Gaifman-Snir), $\lim_{n \rightarrow \infty} P(\neg H_r \cap W_{[rn]}) = 0$. Conclui-se que $\lim_{n \rightarrow \infty} P(H_r \mid h_n(F) = [rn]/n) = x / (x+0) = 1$.

Para (c): Mesmo sem σ aditividade, $P(\forall x Fx) \leq \lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n)$. Assim, de $P(\forall x Fx) > 0$ segue $\lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n) > 0$. Por causa de $P(Fa_1 \wedge \dots \wedge Fa_n) = \prod_{1 \leq i \leq n} P(Fa_i \mid Fa_{i-1} \wedge \dots \wedge Fa_1)$ (com " Π " para o produto contínuo) pode ser $\lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n)$ para $n \rightarrow \infty$ permanecer maior que zero somente se $\lim_{n \rightarrow \infty} P(Fa_{n+1} \mid Fa_1 \wedge \dots \wedge Fa_n) = 1$. (Porque senão- se $\lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n) \leq \lim_{n \rightarrow \infty} x_n$ para um número $0 < x < 1$ e o último valor-limite é conhecido por ser zero, uma vez que diminui ainda mais para cada aumento de n , $n \Rightarrow n+1$.)

Para (d): Por causa de $P(\forall x Fx \wedge Fa_1 \wedge \dots \wedge Fa_n) = P(\forall x Fx) \cdot \lim_{n \rightarrow \infty} P(\forall x Fx \mid Fa_1 \wedge \dots \wedge Fa_n)$
 $= P(\forall x Fx \wedge Fa_1 \wedge \dots \wedge Fa_n) / P(Fa_1 \wedge \dots \wedge Fa_n) = (*) \lim_{n \rightarrow \infty} P(\forall x Fx) / P(Fa_1 \wedge \dots \wedge Fa_n)$. De $P(\forall x Fx) > 0$ segue $\lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n) > 0$ (como em c)); e, portanto, a equação (**) segue de (*): $\lim_{n \rightarrow \infty} P(\forall x Fx \mid Fa_1 \wedge \dots \wedge Fa_n) = P(\forall x Fx) / \lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n)$. No entanto, σ aditividade, agora implica o princípio de continuidade $P(\forall x Fx) = \lim_{n \rightarrow \infty} P(Fa_1 \wedge \dots \wedge Fa_n)$ (ver (3-7)). Disso e de (**) decorre a afirmação. \square

10.3.20. Prova do teorema 9-9

Abreviamos $P(E_i \mid H_k)$ por p_i , escreva $\sum \{x_1, \dots, x_n\}$ e $\Pi \{x_1, \dots, x_n\}$ para a soma e o produto dos números x_1, \dots, x_n respectivamente, e calcule da

seguinte forma: $P(H_k|E_1 \wedge \dots \wedge E_n) = P(E_1 \wedge \dots \wedge E_n|H_k) \cdot P(H_k) / \sum \{P(E_1 \wedge \dots \wedge E_n|H_r) : 1 \leq r \leq m\}$ (teorema bayesiano)

$$\begin{aligned}
 &= \frac{h \cdot \prod\{p_i : 1 \leq i \leq n\}}{h \cdot \prod\{p_i : 1 \leq i \leq n\} + \sum\{P(H_r) \cdot \prod\{P(E_i|H_r) : 1 \leq r \leq n\} : 1 \leq r \leq m, r \neq k\}} \quad (\text{com base em (i)}) \\
 &\geq \frac{h \cdot \prod\{p_i : 1 \leq i \leq n\}}{h \cdot \prod\{p_i : 1 \leq i \leq n\} + \sum\{P(H_r) : 1 \leq r \leq m, r \neq k\} \cdot \prod\{(p_i - \delta) : 1 \leq r \leq n\}} \quad (\text{por causa de (ii)}) \\
 &= \frac{h \cdot \prod\{p_i : 1 \leq i \leq n\}}{h \cdot \prod\{p_i : 1 \leq i \leq n\} + (1-h) \cdot \prod\{(p_i - \delta) : 1 \leq r \leq n\}} = \frac{1}{1 + \frac{1-h}{h} \cdot \frac{\prod\{(p_i - \delta) : 1 \leq r \leq n\}}{\prod\{p_i : 1 \leq r \leq n\}}}
 \end{aligned}$$

Por causa de $\frac{\prod\{(p_i - \delta) : 1 \leq r \leq n\}}{\prod\{p_i : 1 \leq r \leq n\}} \leq (1 - \delta)^n$ obtemos o enunciado (a) da sentença 9-9, do qual segue que $\lim_{n \rightarrow \infty} (1 - \delta)^n = 0$ (b). \square

Bibliografia

- Adams, E.W. (1974) On the Logic of “Almost All” *Journal of Philosophical Logic* v. 3, pp. 3-17.
- Albert, M. (1992) Die Falsifikation statistischer Hypothesen *Journal for General Philosophy of Science* v. 23, pp. 1-32.
- Aron, A.; Aron, E. (2002) *Statistics for Psychology*. New Jersey: Prentice-Hall.
- Bacchus, F. (1990) *Representing and Reasoning with Probabilistic Knowledge*, Cambridge: MIT Press.
- Barwise, J.; Etchemendy, J. (2005) *Sprache, Beweis und Logik*. Paderborn: Mentis.
- Bauer, H. (1978) *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*. Berlin/New York: De Gruyter.
- Bayes, T.; Price, R. (1763) An Essay Towards Solving a Problem in the Doctrine of Chances. In: E.S. Preason, M.G. Kendall (Hg.), *Studies in the History of Statistics and Probability*. New York: Griffin, pp. 134-153.
- Bernoulli, J. (1713) *Ars Conjectandi*. Basel: Thurneysen Brothers.
- Bhaskara Rao, K.P.S.; Bhaskara Rao, M. (1983) *Theory of Charges. A Study of Finitely Additive Measures*. New York: Academic Press Inc.
- Billingsley, P. (1995) *Probability and Measure* (3rd ed.). New York: John Wiley & Sons.
- Bortz, J. (1985) *Lehrbuch der Statistik*. Berlin: Springer. (Neuauf. als *Statistik für Human- u. Sozialwissenschaftler*, 6. überarb. Aufl. 2005).

- Bortz, J.; Döring, N. (2002) *Forschungsmethoden und Evaluation*. Berlin: Springer.
- Brier, G.W. (1950) Verification of Forecasts Expressed in Terms of Probability *Monthly Weather Review* v. 78, pp. 1-3.
- Campbell, S.; Franklin, J. (2004) Randomness and the Justification of Induction *Synthese* v. 138, pp. 79-99.
- Carnap, R. (1947) On the Application of Inductive Logic *Philosophy and Phenomenological Research* v. 8, pp. 133-147.
- Carnap, R. (1950) *Logical Foundations of Probability*. Chicago: Univ. of Chicago. Dt. Kurzfassung in: Carnap (1959).
- Carnap, R. (1959) *Induktive Logik und Wahrscheinlichkeit*. Bearbeitet von W. Stegmüller. Wien: Springer.
- Carnap, R. (1971) Inductive Logic and Rational Decisions e A Basic System of Inductive Logic, Part I. In: Carnap/Jeffrey (1971), Cap. 1 e 2.
- Carnap, R. (1972) *Bedeutung und Notwendigkeit*. Berlin: Springer.
- Carnap, R. (1980) A Basic System of Inductive Logic, Part 2. In: Jeffrey (1980), Cap. 6.
- Carnap, R.; Jeffrey, R. (1971) *Studies in Inductive Logic and Probability*. Berkeley: Univ. of California Press.
- Church, A. (1940) On the Concept of a Random Sequence *Journal of Symbolic Logic* v. 1, n. 40-41, pp. 101-102.
- Clauß, G.; Ebner, H. (1977) *Grundlagen der Statistik*. Thun: Harri Deutsch.
- Coffa, J. (1974) Hempel's Ambiguity *Synthese* v. 28, pp. 141-163.
- Cramér, H. (1946) *Mathematical Models of Statistics*. Princeton: Princeton Univ. Press.
- Crupi, V.; Trentori, K. (2010) Irrelevant Conjunction: Statement and Solution of a New Paradox *Philosophy of Science* v. 77, pp. 1-13.

- David, F.N. (1998) *Games, Gods, and Gambling. The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era*. Dover: Dover Publ.
- De Finetti, B.(1931) Funzione caratteristica di un fenomeno aleatorio *Atti della R. Accademia Nazionale dei Lincei. Memorie, Classe di Scienze Fisiche, Matematiche e Naturale*, v. 4, pp. 251-299.
- De Finetti, B. (1964) Foresight, its Logical Laws, its Subjective Sources. In: H. Kyburg; H. Smokler (Hg.), *Studies in Subjective Probability*. New York: John Wiley.
- De Finetti, B. (1970) *Wahrscheinlichkeitstheorie*. München: Oldenbourg, 1981 (zuerst 1970 italienisch; 1974 als *Theory of Probability* bei John Wiley, New York).
- Diaconis, P.; Freedman, D. (1980) Finite Exchangeable Sequences *Annals of Probability* v. 8, pp. 745-764.
- Douven, I. (2002) A New Solution to the Paradoxes of Rational Acceptability *British Journal for the Philosophy of Science* v. 53, pp. 391-410.
- Eagle, A. (2004) Twenty-One Arguments Against Propensity Analyses of Probability *Erkenntnis* v. 60, pp. 371-416.
- Earman, J. (1986) *A Primer on Determinism*. Dordrecht: Reidel.
- Earman, J. (1992) *Bayes or Bust?* Cambridge: MIT Press.
- Ebbinghaus, H.-D. (2003) *Einführung in die Mengenlehre*. Heidelberg: Spektrum.
- Edwards, W.; Lindman, H.; Savage, L.J. (1963) Bayesian Statistical Inference for Psychological Research *Psychological Review* v. 70, pp. 193-242.
- Fine, T. (1973) *Theories of Probability*. New York: Academic Press.
- Fisher, R.A. (1925) Theory of Statistical Estimation *Proceedings of the Cambridge Philosophical Society* v. 26, pp. 528-535.

- Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*. New York: Hafner Press (ext. ed. Oxford Univ. Press 1995).
- Fitelson, B. (1999) The Plurality of Bayesian Measures of Confirmation *Philosophy of Science* v. 66, pp. 362-378 (Proceedings).
- Foley, R. (1992) The Epistemology of Belief and the Epistemology of Degrees of Belief *American Philosophical Quarterly* v. 29, n. 2, pp. 111-121.
- Gaifman, H.; Snir, M. (1982) Probabilities Over Rich Languages *Journal of Symbolic Logic* v. 47, pp. 495-548.
- Gemes, K. (1993) Hypothetico-Deductivism, Content, and the Natural Axiomatization of Theories *Philosophy of Science* v. 54, pp. 477-487.
- Gillies, D. (2000) *Philosophical Theories of Probability*. London: Hafner Press.
- Glymour, C. (1981) *Theory and Evidence*. Princeton: Princeton Univ. Press.
- Good, I. J. (1966) On the Principle of Total Evidence *British Journal for the Philosophy of Science* v. 17, pp. 319-321.
- Good, I. J. (1983) *Good Thinking. The Foundations of Probability and Its Applications*, Univ. of Minnesota Press, Minneapolis.
- Greeno, J. (1970) Evaluating of Statistical Hypotheses Using Information Transmitted *Philosophy of Science* v. 37, pp. 279-293.
- Goodman, N. (1946) A Query on Confirmation *Journal of Philosophy* v. 44, pp. 383-385.
- Goodman, N. (1975) *Tatsache, Fiktion, Voraussage*. Frankfurt: Suhrkamp (engl. Original 1955).
- Grünbaum, A. (1972) *Philosophical Problems of Space and Time*. Dordrecht: Reidel.

- Hacking, J. (1965) *On the Logic of Statistical Inference*. Cambridge: Cambridge Univ. Press.
- Hájek, A (1999) Fifteen Arguments against Hypothetical Frequentism *Erkenntnis* v. 70, pp. 211-235.
- Halpern, J.Y. (2003) *Reasoning about Uncertainty*. Cambridge: MIT Press.
- Harman, G. (1965) The Inference to the Best Explanation *Philosophical Review* v. 74, pp. 88-95.
- Hawthorne, J. (1996) On the Logic of Non-Monotonic Conditionals and Conditional Probabilities *Journal of Philosophical Logic* v. 25, pp. 185-218.
- Hawthorne, J. (2005) *Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions* *Mind* v. 114, pp. 277-320.
- Hays, W.; Winkler, R. (1970) *Statistics: Probability, Inference, and Decision*, New York: Holt.
- Haken, H. (1983) *Synergetik*. Berlin: Springer.
- Hempel, C.G. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Macmillan: New York & London.
- Hewitt, E.; Savage, L. J. (1955) Symmetric Measures on Cartesian Products *Transactions of the American Mathematical Society* v. 80, pp. 470-501.
- Hintikka, J. (1965) Towards a Theory of Inductive Generalization. In: Bar-Hillel, Y. (Hg.), *Logic, Methodology and Philosophy of Science*, Amsterdam: North-Holland Publ. Company, pp. 274-288.
- Hintikka, J. (1966) A Two-Dimensional Continuum of Inductive Methods. In: Hintikka/Suppes (Hg.), pp. 113-132.
- Hintikka, J.; Suppes, P. (1966, Hg.), *Aspects of Inductive Logic*. Amsterdam: North-Holland Publ. Company.

- Hitchcock C.; Sober, E. (2004) Prediction Versus Accommodation and the Risk of Overfitting *British Journal for the Philosophy of Science* v. 55, pp. 1-34.
- Horwich, P. (1982) *Probability and Evidence*. Cambridge: Cambridge Univ. Press.
- Howson, C. (1990) Fitting Theory to the Facts: Probably not Such a Bad Idea After All. In: C. Savage (Hg.), *Scientific Theories*. Minneapolis: Univ. Minnesota Press, pp. 222-224.
- Howson, C.; Urbach, P. (1996) *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court.
- Huber, F. (2008). Hempel's Logic of Confirmation *Philosophical Studies* v. 139, pp. 181-189.
- Humburg, J. (1971) The Principle of Instantial Relevance. In: Carnap/Jeffrey (1971), Cap. 4.
- Hume, D. (1748) *Eine Untersuchung über den menschlichen Verstand*. Hamburg: reclam.
- Humphreys, P. (1985) Why Propensities Cannot Be Probabilities *The Philosophical Review* v. 94, pp. 557-70.
- Jamison, D. (1970) Bayesian Information Usage. In: J. Hintikka, P. Suppes (Hg.), *Information and Inference*. Dordrecht: Reidel, pp. 28-57.
- Jaynes, E.T. (1976) Confidence Intervals versus Bayesian Intervals. In: Harper, W.L., Hooker, C. (Hg., 1976), *Foundations of Probability Theory. Vol II*. Dordrecht: Reidel. pp. 175-257.
- Jeffrey, R.C. (1971) Probability Measures and Integrals. In: Carnap/Jeffrey (1971), pp. 167-224.
- Jeffrey, R.C. (1980) *Studies in Inductive Logic and Probability. Vol. II*. Berkeley: Univ. of California Press.

- Jeffrey, R.C. (1983) *The Logic of Decision*, (2nd ed.). New York: McGraw-Hill.
- Kahneman, D.; Slovic, P.; Tversky, A. (1982, Hg.) *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge Univ. Press.
- Keynes, J.M. (1921) *A Treatise on Probability*. New York: MacMillan.
- Kingman, J. F.C. (1978) Uses of Exchangeability *Annals of Probability* v. 6, pp. 183-197.
- Klenk, V. (1989) *Understanding Symbolic Logic*. Englewood Cliffs: Prentice Hall.
- Kolmogorov, A.N. (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*, (Zentralblatt der Mathematik, 2. Band). Berlin: Springer.
- Krantz, D.; Luce, R.D.; Suppes, P.; Tversky, A. (2006) *Foundations of Measurement. Vol I*. New York: Dover Publications (orig. 1971).
- Kromrey, H. (2002) *Empirische Sozialforschung*. Opladen: Leske+Budrich.
- Kutschera, F.v. (1972) *Wissenschaftstheorie, Bd. I und II*. München: Fink.
- Kyburg, H.E. (1961) *Probability and the Logic of Rational Belief*. Middletown: Wesleyan University Press.
- Lad, F. (1984) The Calibration Question *British Journal for the Philosophy of Science* v. 35, pp. 213-221.
- Lakatos, I. (1970) Falsification and the Methodology of Scientific Research Programmes. Reprinted in: Lakatos (1978), *Philosophical Papers Vol 1*. Cambridge: Cambridge Univ. Press. pp. 8-101.
- Laplace, P.S. (1814) *A Philosophical Essay on Probabilities*. New York Dover, 1951.

- Lauth, B.; Sareiter, J. (2002) *Wissenschaftliche Erkenntnis*. Paderborn: Mentis.
- Leblanc, H. (1963) A Revised Version of Goodman's Confirmation Paradox *Philosophical Studies* v. 12, pp. 49-51.
- Lehrer, K. (1975) Induction, Rational Acceptance, and Minimally Inconsistent Sets. In: G. Maxwell, R.M. Anderson (Hg.), *Induction, Probability, and Confirmation*. Minneapolis: University of Minnesota Press, pp. 295-323.
- Leitgeb, H. (2013) Reducing Beliefs Simpliciter to Degree of Belief *Annals of Pure and Applied Logic* v. 164, pp. 1338-1389.
- Lenz, J. (1974) Problems for the Practicalist's Justification of Induction. In: Swinburne (1974), pp. 98-101.
- Levi, I. (1967) *Gambling with Truth*. New York: Knopf.
- Levi, I. (1977) Direct Inference *The Journal of Philosophy* v. 74, pp. 5-29.
- Lewis, D. (1973) *Counterfactuals*. Oxford: Basil Blackwell.
- Lewis, D. (1980) A Subjectivist's Guide to Objective Chance. In: Lewis, D. (1986), *Philosophical Papers Vol II*. New York: Oxford Univ. Press, cap. 19.
- Lipton, P. (1991) *Inference to the Best Explanation*. London: Routledge.
- Losee, J. (1977) *Wissenschaftstheorie. Eine historische Einführung*. München: C.H. Beck. (engl. Orig. 1972).
- Loux, M.J. (1998) *Metaphysics*. London & New York: Routledge.
- Maher, P. (1996) The Hole in the Ground of Induction *Australasian Journal of Philosophy* v. 74, n. 3, pp. 423-432.
- Makinson, D. (1965) The Paradox of the Preface *Analysis* v. 25, pp. 205-207.
- Mayntz, R.; Holm, K.; Hübner, P. (1974) *Einführung in die Methoden der empirischen Soziologie*. Opladen: Westdeutscher Verlag.

- Mill, J. St. (1865) *System of Logic*, London; deutsch: als Bde. 2-3 der *Gesammelten Werke von J. St. Mill*, hg. von T. Gomperz, Leipzig 1872, 8. Aufl., zitiert danach.
- Miller, D. (1994) *Critical Rationalism. A Restatement and Defence*. Chicago: Open Court.
- Musgrave, A. E. (1974) Logical versus Historical Theories of Confirmation *British Journal for the Philosophy of Science*, v. 25, pp. 1-23.
- Neyman, J. (1937) Outline of a Theory of Statistical Estimation *Philosophical Transactions of the Royal Society*, v. 236A, pp. 333-380.
- Niiniluoto, I. (1987) *Truthlikeness*. Dordrecht: Reidel.
- Niiniluoto, I. (1999) Defending Abduction *Philosophy of Science* v. 66, pp. 436 – 451.
- Pearl, J. (2000) *Causality*. Cambridge: Cambridge Univ. Press. (2. Auflage 2009).
- Peirce, C.S. (1903) Lectures on Pragmatism. In: Apel, K.-O. (1976, Hg.): *Charles Sanders Peirce: Schriften zum Pragmatismus und Pragmatizismus*. Frankfurt: Suhrkamp (2. Aufl.), pp. 337-427.
- Popper, K. (1935) *Logik der Forschung*, 10. Aufl., Tübingen: J.C.B. Mohr 2005.
- Popper, K. (1959) The Propensity Interpretation of Probability *British Journal for the Philosophy of Science* v. 10, pp. 25-42.
- Popper, K. (1982) *The Open Universe*. Totowa: Rowman and Littlefield.
- Popper, K. (1990) *A World of Propensities*. London: Thoemmes.
- Popper, K.; Miller, D. (1983) A Proof of the Impossibility of Inductive Probability *Nature* v. 302, pp. 687-688.
- Raiffa, H. (1973) *Einführung in die Entscheidungstheorie*. München: Oldenbourg. (engl. Original 1968).

- Ramsey, F.P. (1926) Truth and Probability. In: ders., *Philosophical Papers*, hg. von H. D. Mellor. Cambridge: Cambridge Univ. Press, 1990.
- Reichenbach, H. (1935) *Wahrscheinlichkeitslehre*. Leiden: A.W. Sijthoff's.
- Reichenbach, H. (1938) *Experience and Prediction*. Chicago: University of Chicago Press.
- Reichenbach, H. (1949) *The Theory of Probability*. Berkeley: University of California Press (erweit. engl. Fassung von 1935).
- Rescher, N. (1987) *Induktion*. München: Philosophia Verlag. (engl. Orig. 1980).
- Reutlinger, A.; Schurz, G.; Hüttemann, A. (2011) Ceteris Paribus Laws *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <<http://plato.stanford.edu/entries/ceteris-paribus>>.
- Rosenkrantz, R. (1977) *Inference, Method and Decision*. Dordrecht: Reidel.
- Rosenthal, J. (2004) *Wahrscheinlichkeiten als Tendenzen. Eine Untersuchung objektiver Wahrscheinlichkeitsbegriffe*. Paderborn: Mentis.
- Runggaldier, E. (1990) *Analytische Sprachphilosophie*. Stuttgart: Kohlhammer.
- Ryder, J.M. (1981) Consequences of a Simple Extension of the Dutch book Argument *British Journal for the Philosophy of Science* v. 32, pp. 164-167.
- Sachs, L. (1992) *Angewandte Statistik*. Berlin: Springer. 9. Aufl. (11. überarb. Aufl. 2004).
- Salmon, W. (1974) The Pragmatic Justification of Induction. In: Swinburne (1974), pp. 85-97.

- Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton Univ. Press.
- Salmon, W. (1989) *Four Decades of Scientific Explanation*. Minneapolis: Univ. of Minnesota Press.
- Schlesinger, G. (1974) *Confirmation and Confirmability*. Oxford: Clarendon Press.
- Schurz, G. (1991) Relevant Deduction *Erkenntnis* v. 35, pp. 391-437.
- Schurz, G. (1994) Relevant Deduction and Hypothetico-Deductivism: A Reply to Gemes *Erkenntnis* v. 41.
- Schurz, G. (2002) Ceteris Paribus Laws: Classification and Deconstruction *Erkenntnis* v. 57, n. 3, pp. 351-372.
- Schurz, G. (2006) *Einführung in die Wissenschaftstheorie*. Darmstadt: Wissenschaftliche Buchgesellschaft. (3. Aufl. 2011).
- Schurz, G. (2008a) Patterns of Abduction *Synthese* v. 164, pp. 201-234.
- Schurz, G. (2008b) The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem *Philosophy of Science* v. 75, pp. 278-305.
- Schurz, G. (2011) *Evolution in Natur und Kultur*. Heidelberg: Spektrum.
- Schurz, G. (2012) Tweety, or Why Probabilism and even Bayesianism Need Objective and Evidential Probabilities. In: D. Dieks et al., *Probabilities, Laws and Structures*. New York: Springer, pp. 57-74.
- Schurz, G. (2013a) Wertneutralität und hypothetische Werturteile. In: Schurz, G., Carrier, M. (Hg.), *Werte in den Wissenschaften*. Frankfurt: Suhrkamp, pp. 305-335.
- Schurz, G. (2013b) Bayesian Pseudo-Confirmation, Use-Novelty, and Genuine Confirmation *Studies in History and Philosophy of Science* v. 45, pp. 87-96.
- Schurz, G. (2014) *Philosophy of Science: A Unified Approach*. New York: Routledge.

- Schurz, G.; Leitgeb, H. (2008) Finitistic and Frequentistic Approximations of Probability Measures with or without Sigma-Additivity *Studia Logica* v. 89, n. 2, pp. 258-283.
- Schurz, G.; Weingartner, P. (2010) Zwart and Franssen's Impossibility Theorem Holds for Possible-World-Accounts but not for Consequence-Accounts to Verisimilitude *Synthese* v. 172, pp. 415-436.
- Schuster, H. G. (1994) *Deterministisches Chaos*. Weinheim: VCH.
- Skyrms, B. (1980) *Causal Necessity*. New Haven: Yale University Press.
- Skyrms, B. (1999) *Choice and Chance – an Introduction to Inductive Logic*. Belmont: Wadsworth Publishing (orig. 1975).
- Sober, E. (1993) *Philosophy of Biology*. Boulder: Westview Press (2nd ed. 1999).
- Spielman, S. (1976) Exchangeability and the Certainty of Objective Randomness *Journal of Philosophical Logic* v. 5, pp. 399-406.
- Spielmann, S. (1977) Physical Probability and Bayesian Statistics *Synthese* v. 36, pp. 235-269.
- Stegmüller, W. (1973a,b) *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band IV: Personelle und Statistische Wahrscheinlichkeit*. Berlin: Springer.
- _____ (1973a) Erster Halbband: *Personelle Wahrscheinlichkeit*.
- _____ (1973b) Zweiter Halbband: *Statistisches Schließen*.
- Stove, D. (1986) *The Rationality of Induction*. Oxford: Clarendon Press.
- Strevens, M. (2004) Bayesian Confirmation Theory: Inductive Logic, or Mere Inductive Framework *Synthese* v. 141, pp. 365-379.
- Strevens, M. (2008) *Depth. An Account of Scientific Explanation*. Cambridge: Harvard Univ Press.

- Suppes, P. (1966) Probabilistic Inference and the Concept of Total Evidence. In: Hintikka, J., Suppes, P. (Hg.), pp. 49-65.
- Swinburne, R. (1974, Hg.) *The Justification of Induction*. Oxford: Oxford University Press.
- Tarski, A. (1936) Der Wahrheitsbegriff in den formalisierten Sprachen *Studia Philosophica* v. 1, pp. 261-405.
- Unterhuber, M.; Schurz, G. (2013) The New Tweety Puzzle: Arguments against Monistic Bayesian Approaches in Epistemology and Cognitive Science *Synthese* v. 190, pp. 1407-1435.
- Van Fraassen, B. (1980) *The Scientific Image*. Oxford: Clarendon Press (Neuausgabe 1990).
- Van Fraassen, B. (1983) Calibration: A Frequency Justification for Personal Probability. In: Cohen, S.R., Laudan, L. (Hg.), *Physics, Philosophy, and Psychoanalysis*. Dordrecht: Reidel, pp. 295-319.
- Vickers, J. (2010) The Problem of Induction *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (Hg.), URL=
<<http://plato.stanford.edu/archives/spr2010/entries/induction-problem/>>.
- Von Mises, R. (1928) *Probability, Statistics and Truth* (2nd revised ed.). London: Allen and Unwin. 1961.
- Von Mises, R. (1964) *Mathematical Theory of Probability and Statistics*. New York: Academic Press.
- Weingartner, P.; Schurz, G. (1996, Hg.), *Law and Prediction in the Light of Chaos Research*. Berlin: Springer.
- Westermann, R.; Hager, W. (1982) Entscheidung über statistische und wissenschaftliche Hypothesen *Zeitschrift für Sozialpsychologie* v. 13, pp. 13-21.

- Wilholt, T. (2009) Bias and Values in Scientific Research *Studies in History and Philosophy of Science* v. 40, pp. 92-101.
- Williams, D. (1947) *The Ground of Induction*. New York: Russell and Russell.
- Williamson, J. (2010) *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.
- Williamson, J. (2013) Why Frequentists and Bayesians Need Each Other *Erkenntnis* v. 78, pp. 293-318.
- Wójcicki, R. (1966) Filozofia Nauki W Minnesota Studies *Studia Filozoficzne*, pp. 143-154.
- Woodward, J. (2003) *Making Things Happen*. Oxford: Oxford University Press.
- Worrall, J. (2006) Theory-Confirmation and History. In: C. Cheyne, J. Worrall (Hg.), *Rationality and Reality*. New York: Springer, pp. 31-61.

Lista de figuras, definições e teoremas

Ilustrações

Fig. 2-1: Convergência de frequências relativas, Fig. 3-1 Probabilidades estatísticas condicionais, Fig. 3-2 Não monotonicidade de probabilidades condicionais, Fig. 3-3 Três distribuições binomiais, Fig. 3-4 Medidas de probabilidade σ -aditivas sobre \mathbb{N} , Fig. 5-1 Explicação da distribuição uniforme dos resultados do lançamento de dados, Fig. 8-1: Intervalo de aceitação para $p(D|A) = 0,8$, Fig. 8-2: Relação entre aceitação e intervalo de confiança, Fig. 8-3: Distribuição de probabilidade de diferenças amostrais e diferença amostral significativa, Fig. 8-4 Distribuição normal gaussiana, Fig. 8-5 Causa comum, Fig. 8-6 Independência aparente, Fig. 8-7 Direção causal, Fig. 9-1 Menor intervalo de 70%, Fig. 9-2 Dependência de linguagem de Distribuições uniformes, Fig. 9-3 Ajuste de curva e dados não utilizados, Fig. 10-1 Conceitos lógicos.

Definições

Def. 2-1 Probabilidade estatística e epistêmica, Def. 2-2 Limite de frequência, Def. 2-3 Princípio da classe de referência mais estreita, Def. 3-1 Axiomas básicos de probabilidade, Def. 3-2 Probabilidade condicional, Def. 3-3 Axiomatização direta da probabilidade condicional, Def. 3-4 Independência probabilística, Def. 3-5 Inferência

probabilística, Def. 3-6 σ -aditividade, Def. 4-1 Axiomatização de funções de Popper, Def. 5-1 Seleção de lugar permitida, Def. 5-2 Sequência estatística básica e sequência aleatória, Def. 6-1 Coerência, Def. 6-2 Coerência e regularidade estritas, Def. 7-1 Princípio de coordenação singular, Def. 7-2 Princípio de coordenação estatística, Def. 7-3 Conteúdo indutivo-empírico de uma hipótese estatística, Def. 7-4 Condicionização na evidência total, Def. 7-5 Permutabilidade, Def. 7-6 Calibração, Def. 8-1 Intervalo de aceitação, Def. 8-2 Diferença amostral significativa, Def. 8-3 Média e variância na população, Def. 8-4 Média e variância em uma amostra, Def. 9-1 Evidência consistente independente, Def. 9-2 Definição de Goodman de "verdul", Def. 9-3 Confirmação bayesiana, Def. 9-4 Confirmação genuína completa e parcial.

Teoremas

Teorema 3-1 Teorema da probabilidade incondicional, Teorema 3-2 Probabilidade condicional definida e diretamente axiomatizada, Teorema 3-3 Teoremas de probabilidade condicional, Teorema 3-4 Leis dos grandes números, Teorema 3-5 Operações lógicas e álgebra de conjuntos, Teorema 4-1, Consequência lógica e teoria da probabilidade, Teorema 4-2 Funções de Popper, Teorema 4-3 Regras da lógica de probabilidade condicional, Teorema 4-4 Soma das incertezas, Teorema 5-1 Leis de grandes números na estrutura de von Mises, Teorema 6-1 Coerência, Teorema 6-2, coerência estrita, Teorema 7-1 Coerência e confiabilidade do StK, Teorema 7-2 Probabilidades de suporte, Teorema 7-3 permutabilidade de P, Teorema 7-4 Aprendizagem indutiva uniforme, Teorema 7-5 Calibração e classe de referência mais estreita, Teorema 8-2 Leis de

cálculo para valores esperados, Teorema 8-3 Média e dispersão de médias amostrais, Teorema 8-4 Condições de Reichenbach para causas comuns, Teorema 9-1 Justificativa bayesiana da intuição de verossimilhança, Teorema 9-2 Distribuição de probabilidade final, Teorema 9-3 Consequências do princípio da indiferença, Teorema 9-4 Distribuição uniforme e argumento de Williamson, Teorema 9-5 Convergência contínua para previsões indutivas, Teorema 9-6 Convergência contínua para generalizações indutivas, Teorema 9-7 Incorrigibilidade finita de probabilidades iniciais tendenciosas, Teorema 9-8 Resultados de convergência simples, Teorema 9-9 Evidência consistente independente, Teorema 9-10 Pseudoconfirmação bayesiana, Teorema 9-11 Condições necessárias para a transferência do aumento de probabilidade, Teorema 10-1 Leis para probabilidades estatísticas combinadas independentemente.

Índice de pessoas

- Adams E.W. 46, 65, 67, 68, 307
Albert M. 189
Aron A. 188 bis
Bacchus F. 23, 46, 47, 51, 52, 143
(n. 45), 308
Barwise J. 301
Bauer H. 15, 43 bis, 44, 192 (n.
54), 195, 200, 205, 206, 304 bis
Bayes T. 12, 34 bis, 35, 68, 69,
138, 222 bis, 243, 265, 314, 315
Bernoulli J. 13, 41
Bhaskara Rao K.P.S. 57
Bhaskara Rao M. 57
Billingsley P. 15, 44, 195, 234(n.
65), 322, 334
Bortz J. 15, 167, 172, 173, 174, 180
bis, 181, 195, 198, 200, 201, 203,
205, 206, 223, 284
Campbell S. 182
Carnap R. 14, 22, 31, 32 bis, 45,
49 bis, 63, 111 (n.35), 114, 120,
133, 135, 142, 143, 145 (n. 47), 147,
148 bis, 152, 153, 219, 228, 231 bis,
232 bis, 233 bis, 234 bis, 235 bis,
236, 259, 260 bis, 261 bis, 263,
282, 297, 307, 313
Church A. 85 bis
Clauß G. 206
Coffa J. 102, 153
Cramér H. 96
Crupi V. 273
David F.N. 1 (n.1)
De Finetti B. 14, 53, 111, 113, 114,
123, 142 bis, 144, 145 (n.47), 147,
245, 313 bis
Diaconis P. 149 (n. 47)
Döring N. 185
Douven L. 293
Eagle A. 83 (n. 23), 109, 110
Earman J. 14, 99, 101, 102, 115bis,
120, 122, 135, 141, 147, 231 bis,
245, 251, 265, 274, 316
Ebbinghaus H.-D. 302
Ebner H. 206
Edwards W. 109, 325
Etchemendy J. 314
Fine T. 14, 73

- Fisher R.A. 14, 130, 162, 168, 183, 184, 185 bis, 215, 216, 218, 221 bis, 222
- Fitelson B. 270
- Foley R. 289
- Franklin J. 182
- Freedman D. 149, n. 47
- Gaifman H. 249, 250 bis, 251 bis, 319, 320
- Gemes K. 272, 273
- Gillies D. 12 (n. 1), 13 bis, 14, 71, 82 (n. 24), 93, 107, 115, 116, 119, 147 bis, 148, 183, 228, 229, 318
- Glymour C. 266, 278
- Good I.J. 149, 152
- Goodman N. 15, 143, 152, 255 bis, 257 bis, 258 bis, 259, 260, 261 bis, 266, 268
- Greeno J. 80
- Grünbaum A. 210
- Hacking J. 222
- Hager W. 175 bis, 184
- Hájek A. 80 (n. 23), 90 bis, 91, 92 bis
- Halpern J.Y. 51
- Harman G. 60
- Hawthorne J. 62 (n. 14), 125, 133, 139, 140, 162
- Hays W. 14, 41, 162, 182, 184, 185, 186 (n. 54), 192, 194, 215, 218, 221, 225, 228, 245, 319
- Hempel C.G. 95, 149, 153, 154
- Hewitt E. 145 (n. 47), 313
- Hintikka J. 235, 236 bis
- Hitchcock C. 279, 281
- Holm K. 336
- Horwich P. 153, 155
- Howson C. 14 bis, 41, 54, 73, 82 (n. 24), 88, 95, 111 (n.35), 115 bis, 118, 123, 135, 141, 162, 183, 184 bis, 185, 216, 218 bis, 219, 220 bis, 221, 222, 227 (n. 65), 229, 239, 242 (n. 72), 245, 256, 259 bis, 262, 270, 271, 311, 318
- Huber F. 133, 263
- Hübner P. 336
- Humburg J. 151 bis
- Hume D. 259 bis
- Humphreys P. 109
- Hüttemann A. 338
- Jamison D. 243
- Jaynes E.T. 224
- Jeffrey R. 121 (n. 37), 137 bis, 148, 186 (n. 54), 189, 227 (n. 65), 231, 304
- Kahneman D. 113

- Keynes J.M. 14, 148 bis, 229 bis, 247
- Kingman J.F.C 145 (n. 47)
- Klenk V. 301
- Kolmogorov A.N. 25 bis, 28, 52, 53, 64, 93, 110, 306
- Krantz D. 298
- Kromrey H. 183
- Kutschera F.v. 14, 45, 80 (n. 23), 82 (n. 24), 85, 123, 147, 148, 233 (n. 70), 234 (n. 71), 258, 260, 318 bis
- Kyburg H.E. 156 bis, 285, 287
- Lad F. 158
- Lakatos I. 273
- Laplace P.S. 12, 13 bis, 98, 100, 101, 148, 227, 232 bis, 234, 235
- Lauth B. 182 (n. 52), 186 (n. 54), 194, 200
- Leblanc H. 265
- Leitgeb H. 44, 45 (n. 8), 55, 56, 73 bis, 74 bis, 232 (n. 68), 282, 287
- Lenz J. 98
- Levi I. 23, 284
- Lewis D. 91 bis, 106 bis, 107, 122 bis
- Lindman H. 325
- Lipton P. 71 bis
- Losee J. 166
- Loux M.J. 92 (n. 29)
- Luce R.D. 335
- Maher P. 248, 250
- Makinson D. 285, 286, 287
- Mayntz R. 177, 178, 205
- Mill J. St. 166 bis
- Miller D. 106, 121 (n. 37) bis, 267
- Musgrave A.E. 271 (n. 75)
- Neyman J. 168, 216, 221 bis, 222
- Niiniluoto I. 59, 236
- Pearl J. 133, 208
- Peirce C.S. 60 bis
- Popper K. 31, 61, 62 bis, 63 bis, 88 bis, 89 bis, 95, 101 bis, 106, 183 bis, 184 bis, 231 bis, 267, 337, 338
- Price R. 13
- Ramsey F.P. 14, 111, 113, 114
- Reichenbach H. 14, 21 bis, 22 bis, 23 bis, 76, 77, 87, 95 bis, 147 (n. 48), 153, 206 bis, 207 bis, 208, 260 bis, 338
- Rescher N. 242
- Reutlinger A. 185
- Rosenkrantz R. 153
- Rosenthal J. 105
- Runggaldier E. 304
- Ryder J.M. 122 bis

- Sachs L. 37
- Salmon W. 82 (n. 24), 87, 99, 153, 155, 185, 235
- Savage L.J. 145 (n. 47), 313, 318
- Schlesinger G. 286
- Schurz G. 23, 44, 45 (n. 8), 55, 56, 59, 66, 73 bis, 74 bis, 78, 91 bis, 102, 106 (n. 34), 132, 133, 152, 153, 180, 182 bis, 183 (n. 53), 207, 220, 232 (n. 68), 253, 258, 260, 261, 262, 266, 267 bis, 268, 269, 271, 274, 284 (n. 77), 292, 300
- Schuster H.G. 104
- Skyrms B. 75 (n. 19), 80 (n. 23), 111 (n. 35), 260
- Slovic P. 335
- Snir M. 249, 250 bis, 251, 320 bis
- Sober E. 286, 288
- Spielman S. 53, 144, 178, 313
- Spielmann S. 55
- Stegmüller W. 14, 41, 80 (n. 23), 168, 184, 215, 305
- Stove D. 237, 238 bis, 239 bis, 243
- Strevens M. 103 bis, 104, 121 (n. 37), 123, 125
- Suppes P. 61, 65 bis
- Swinburne R. 277
- Tarski A. 304
- Trentori K. 273
- Tversky A. 335
- Unterhuber M. 137
- Raiffa H. 77
- Urbach P. 14 bis, 41, 54, 73, 82 (n. 24), 88, 95, 111 (n. 35), 115, 123, 135, 141, 162, 183, 184 bis, 185, 216, 218, 219, 220 bis, 221 bis, 227 (n. 65), 229, 239, 242 (n. 72), 244, 262, 270, 311, 318
- Van Fraassen B. 73, 147, 157, 158, 159
- Vickers J. 244, 250
- Von Mises R. 13, 14, 79 (n. 22), 81 bis, 82 bis, 83 bis, 84 bis, 85 bis, 86 bis, 87 bis, 88, 89, 90 bis, 92 bis, 93, 95 bis, 338, 342
- Weingartner P. 105, 274
- Westermann R. 174, 175, 184
- Wilholt T. 284
- Williams D. 237 bis, 238 bis, 239 bis, 243, 338
- Williamson J. 122 (n. 37), 123, 148 bis, 226 (n. 64), 237, 240 bis, 241, 242 (n. 72)
- Winkler R. 14, 41, 162, 182, 184, 185, 186 (n. 54), 192, 215, 218, 221, 225, 228, 239, 244, 319
- Wójcicki R. 157

Probabilidade

Woodward J. 185, 213

Worral J. 189, 278

