Inferência Indutiva com Dados Discretos: Uma Visão Genuinamente Bayesiana.

Minicurso apresentado no
XV Congreso de Matemática Capricornio,
Antofagasta, Chile,
3-6 Agosto 2005

Carlos Alberto de Bragança Pereira [1]
Julio Michael Stern [2]

[1]Carlos Alberto de Bragança Pereira, *cpereira@ime.usp.br*, é Ph.D. em Estatística pela Universidade da Florida, (Talahasse, FL, USA), Titular do Departamento de Estatística do IME-USP e Diretor Científico do Núcleo de Bioinformática.

[2]Julio Michael Stern *jstern@ime.usp.br*, é Ph.D. em Pesquisa Operacional pela Universidade de Cornell (Ithaca, NY, USA), Livre Docente do Departamento de Ciência da Computação do IME-USP, e consultor na área de Pesquisa Operacional.

Conteúdo

1 Introdução **7**

1.1 Preliminares 7

1.2 Trabalho Estatístico 9

1.2.1 Exemplo: Informação Estatística 11

1.2.2 Exemplo: Estatística Forense 14

1.2.3 Exemplo: Diagnóstico Médico 15

1.2.4 Exemplo: Paternidade 16

1.2.5 Comentários Adicionais 18

2 Análise Pré-Posteriori **19**

2.1 Introdução 19

2.2 Informação Segundo Basu 19

2.3 Informação Segundo DeGroot 22

2.4 Suficiência de Blackwell 23

2.5 Tamanho de Amostra 27

2.6 Amostras de Populações Infinitas 27

2.7 Amostras de Populações Finitas 29

3 Distribuições Derivadas do Processo de Bernoulli **35**

3.1 Preliminares 35

3.2 Notação 35

3.3 O Processo de Bernoulli 38

3.4 A Distribuição Multinomial 42

3.5 Distribuição Hipergeométrica Multivariada 44

3.6 A Distribuição Dirichlet 46
3.7 Dirichlet-Multinomial 51
3.8 Dirichlet do Segundo Tipo 54
3.9 Exemplos 55

4 Entropia **61**
4.1 Max-Ent com Restrições Lineares 63
4.2 Covergência da Posteriori 66

5 Ônus da Prova no Discurso Cientifico e Jurídico **71**
5.1 Introdução 71
5.2 O Valor de Evidência do FBST 71
5.3 Cálculo Abstrato de Crença 72
5.4 Evidência e Onus Probandi 75
5.5 Condicionalização 77
5.6 Cálculos de Crença Coexistentes no FBST 78
5.7 Comentários Finais e Agradecimentos 80

A Probabilidade **83**
A.1 Espaços de Probabilidade Discretos 83
A.2 Esperança 86
A.2.1 Propriedades de Transformação 87
A.3 Covariância 87
A.3.1 Propriedades de Transformação 88
A.3.2 Correlação 89
A.4 Espaços Contínuos 90
A.5 Exercícios 91

B Álgebra Linear Computacional **95**
B.1 Notação e Operações Básicas 95
B.2 Espaços Vetoriais com Produto Interno 96
B.3 Projetores 97

B.4 Matrizes Ortogonais 97
B.5 Fatoração QR 99
B.5.1 Mínimos Quadrados 100
B.6 Fatorações LU e Cholesky 100
B.6.1 Programação Quadrática 101
B.7 Fatoração SVD 102
B.8 Exercícios 103

Capítulo 1

Introdução

> The Statistician is the Wizard who makes "scientific" statements about invisible states and quantities. However, contrary to common wishes (or witches), he attaches uncertainties to his statements.

1.1 Preliminares

Estas notas são dirigidas às pessoas que desejam fazer afirmações sobre o desconhecido, isto é, àqueles que desejam adivinhar valores de quantidades que, no momento do desafio estão invisíveis. Evidentemente, como estamos na academia, adivinhar racionalmente será nossa tarefa. Esta tarefa é por nós chamada de inferência estatística. Usaremos indiscriminadamente as denominações estado, estado da natureza, estado do sistema, e quantidade.

Estatística é a disciplina que trata das incertezas sobre estados da natureza ou quantidades não observadas ou não observáveis pelo "cientista". As incertezas podem se apresentar em diversos níveis. A forma encontrada para medir o nível dessa incerteza - de um determinado indivíduo - sobre uma estado da natureza, é a escala das probabilidades. Entendemos por cientista, no contexto desse curso, qualquer indivíduo que está especulando sobre o valor de uma quantidade desconhecida.

Informação, no nosso contexto, é um conceito abstrato. Informação é o que ela faz: muda o nível da incerteza. Assim, tratamos de informação sobre quantidades ou estados da natureza desconhecidos.

A ferramenta básica de nosso curso é a teoria das probabilidades. Tentaremos em todas as situações descrever nossas incertezas por meio de probabilidades. As quantidades, objetos de nossas preocupações, podem ser de três tipos:

- as observadas,

- as não observadas, e
- as não observáveis.

A distinção entre as últimas categorias pode ser teórica ou essencial, ser legal ou regimental, ser meramente circunstancial, ou simplesmente irrelevante. Ademais esta distinção pode envolver conceitos da especialidade do cientista. Utilizaremos os termos não observado vs. não observável com relativa liberdade, indicando quantidades desconhecidas que podemos (queremos) vir a observar vs. quantidades desconhecidas que não podemos (queremos) vir a observar. As quantidades não observadas ou não observáveis recebem o nome de variáveis aleatórias. Em muitas ocasiões as variáveis aleatórias não observáveis são chamadas de Parâmetros.

Na literatura estatística é habitual utilizar as letras do nosso alfabeto, romanas, para representar variáveis aleatórias observáveis, e letras gregas para representar os parâmetros, isto é, para variáveis aleatórias não observáveis. Por inferência estatística entendemos o trabalho de descrever nossas incertezas sobre quantidades de interesse e fazer afirmações sobre essas quantidades.

A partir do cenário descrito acima, concluímos que as palavras-chave da estatística são incerteza e informação. Embora o cientista não saiba o verdadeiro valor de uma quantidade de interesse, θ, sua incerteza sobre esse valor pode assumir diferentes níveis. Vez por outra, ele obtém informações adicionais e, assim, seu nível de incerteza sobre θ se modifica. Uma meta utópica seria ganhar toda a informação necessária para eliminar toda a incerteza. Entretanto, na maioria dos casos, essa é uma meta impossível de ser atingida por razões práticas. No próximo parágrafo aludimos ao fato de que, em certos contextos, esta meta é teóricamente impossível. Ademais, quando a certeza é obtida, o trabalho do estatístico se torna dispensável.

Acima, utilizamos a metáfora do verdadeiro "estado da natureza". Isto é, assumimos implicitamente que, embora o valor de um parâmetro possa ser desconhecido em uma situação real, esse parâmetro sempre tem um "verdadeiro valor" θ_0.

A interpretação subjetivista assume que toda a incerteza (a repeito dos parâmetros) de um determinado modelo vem da falta ou imperfeição de nosso conhecimento. A posição subjetivista radical advoga que a interpretação subjetivista é aplicável a todos os modelos e situações possíveis. Assim, na posição subjetivista readical, o próprio termo probabilidade torna-se (sempre) interpretável como "incerteza pessoal do modelador".

A frase de Albert Einstein "Deus não joga dados" (Gott würfelt nicht) condensa a seguinte ideia: É possível que, devido a minhas limitações de simples mortal, eu não saiba o "verdadeiro estado da natureza", mas este verdadeiro estado realmente existe, pois Deus certamente o conhece. Esta frase tem pouco a ver com teologia (ao menos no sentido tradicional do termo), mas com uma série de discussões sobre fundamentos de mecanica quântica que Einstein teve com Werner Heisemberg.

Heisemberg pela primeira vez postulou a posição contrária, isto é, de que a incerteza é um elemento essencial e intransponível na medida (conhecimento) de processos físicos. Heisemberg deu uma série de exemplos onde a metáfora do "verdadeiro valor do estado de um sistema" leva a sérios paradoxos. Este tema teve muitos desdobramentos na área de fundamentos da Mecânica Quântica e Computação Quântica, e ultrapassa em muito não só o escopo deste curso, mas da própria Estatística. Para maiores esclarecimentos veja Albert (1992) e d'Espagnat (1995).

Bruno De Finetti, em parte inspirado pela postura de Einstein na Física, adotou a posição subjetivista radical, com grande influência no desenvolvimento da moderna teoria Bayesiana, veja Finetti (1000, 1000). Neste curso utilizaremos a interpretação subjetivista em diversos modelos, mas sem advogar por uma posição subjetivista radical. Voltaremos a discutir este tema ao falar de distribuições a priori e seus usos em Estatística, teste de hipótese, seleção de modelos, teoria da utilidade e seus usos, e fundamentos lógicos de sistemas de crença.

1.2 Trabalho Estatístico

O trabalho do estatístico é iniciado no momento da descrição dos níveis de incerteza de um cientista sobre as quantidades de interesse. A ferramenta usada para a descrição do nível de incerteza é a Probabilidade. A probabilidade de um estado de natureza específico, digamos θ, é um índice que indica o nível de incerteza (ou conhecimento) sobre a veracidade da afirmação "o valor do estado, θ é θ_0". Como veremos na seqüência, para este importante índice de incerteza, existem regras específicas e bem estabelecidas que devem ser obedecidas.

Probabilidades devem ser atribuídas a todos os estados de natureza possíveis. Ao conjunto de todas as afirmações probabilísticas de um problema usamos o termo modelo probabilístico ou, equivalentemente, distribuição de probabilidades.

Assim, o estatístico deve definir o modelo probabilístico que, em sua opinião, melhor representa o estado de incerteza sobre θ e procurar fontes adicionais de informação que possam diminuir esse nível de incerteza.

Na procura pela diminuição da incerteza, após definir a distribuição de probabilidades de θ, o estatístico inicia a busca por quantidades observáveis, X por exemplo, que estejam, em sua opinião, associadas a θ. O objetivo é observar o valor x de X e transformar o modelo de probabilidade de θ no modelo de probabilidade (condicional) de θ dado que $X = x$.

O mecanismo que transforma uma quantidade não observada, X, em observada, x, é aqui denominado experimento. O conjunto de quantidades obtidas após a realização de experimentos é denominado conjunto dos resultados experimentais ou banco de dados.

O modelo probabilístico de θ, definido antes (depois) da realização do experimento X, é denominado distribuição a priori (a posteriori) de θ. Quando não houver ambiguidade, utilizaremos a expressão "realizar o experimento X" para abreviar a expressão "observar a variável aleatória X"

Efetuar a operação priori/posteriori é, na verdade, construir um processo indutivo de ganho de conhecimento. O que foi chamado de distribuição a posteriori em um dado momento, pode no momento seguinte passar a ser uma distribuição a priori para um novo experimento, digamos Y.

Muitas vezes o cientista enfrenta um dilema: Que variáveis aleatórias observar para diminuir a incerteza sobre θ. Ele pode ter a opção de realizar um dentre um conjunto de experimentos, digamos $X, Y, Z, \ldots$. Pode haver custos ou restrições associados à realização dos experimentos. Assim, o estatístico é obrigado a selecionar quais serão realizados. Para uma escolha adequada, ele deve descrever o tipo de associação que, em sua opinião, existe entre θ e cada um das variáveis aleatórias dos experimentos. Essa descrição também é feita por meio de modelos probabilísticos. Por exemplo, considere que o experimento X vai ser realizado. Para cada valor θ do parâmetro, represente por $p(x \,|\, \theta)$ a função de probabilidade que avalia, para cada x, a probabilidade de observarmos $X = x$. Se existem pelo menos dois valores de θ, digamos θ_1 e θ_2, tal que as funções (em x), $p(x \,|\, \theta_1) \neq p(x \,|\, \theta_2)$, então θ e X são dependentes ou associados. Neste caso é razoável realizar-se o experimento X com o intúito de diminuir a incerteza sobre θ.

O modelo $p(x \,|\, \theta)$ é uma função de duas componentes, x e θ. Para um valor fixado de θ, p é uma função de probabilidades, isto é $p(x \,|\, \theta) = p(X = x \,|\, \theta)$. Por outro lado, após observar $X = x$, p é uma função apenas de θ. Neste caso p passa a ser denominada função de verossimilhança de θ com respeito a observação x. Para evitar mal entendidos, usaremos a notação $L(\theta \,|\, x)$ para a verossimilhança e $p(x \,|\, \theta)$ para a função de probabilidades. Antes de continuar com a descrição do trabalho estatístico, apresentamos a seguir os elementos básicos de nosso trabalho.

Seja θ o parâmetro, e X uma variável aleatória, que está relacionada a θ na opinião do cientista. O conjunto de todos os valores possíveis de θ, Θ, é denominado espaço paramétrico. O conjunto de todos os valores possíveis de X, $\mathcal{X}$, é denominado espaço amostral. Os seguintes conjuntos são os principais elementos do trabalho estatístico.

Distribuição a priori	$f(\theta) : \Theta \mapsto \mathcal{R}$
Distribuição Amostral	$p(x \,|\, \theta) : \mathcal{X} \mapsto \mathcal{R}\ ,\ \forall \theta \in \Theta$
Função de Verossimilhança	$L(\theta \,|\, x) = p(x \,|\, \theta) : \Theta \mapsto \mathcal{R}\ ,\ \forall x \in \mathcal{X}$
Distribuição a Posteriori	$f(\theta \,|\, x) : \Theta \mapsto \mathcal{R}\ ,\ \forall x \in \mathcal{X}$

As relações entre esses conjuntos serão discutidas no apêndice de probabilidade, onde apresentaremos as regras que regem a manipulação adequada das probabilidades.

Como vimos, o estatístico deve procurar realizar, dentro de suas possibilidades, os

experimentos que podem proporcionar um ganho maior de informação. Entretanto, a escolha de qual experimento deve ser realizado é feita, evidentemente, antes de sua realização. Assim, sua escolha é feita com base em suas expectativas sobre os possíveis e prováveis resultados daqueles experimentos que se apresentam como alternativas. Essa análise de expectativas, para cada um dos experimentos concorrentes, pode ser identificada como o clássico Planejamento de Experimentos. Essa área é caracterizada pela procura dos planejamentos ótimos . Em muitas situações, esses são experimentos não realizáveis devido à restrições operacionais ou de custos. No contexto deste curso, Bayesiano (devido ao reverendo Thomas Bayes), o trabalho da procura do que realizar e observar é chamado de análise pré-posteriori. Esse trabalho inicial do estatístico consiste em comparar, para cada experimento possível, priori e posteriori esperada. Para isto, define-se a distância esperada entre priori e posteriori. Com as distâncias em mão, ordenamos os experimentos passíveis de serem observados.

1.2.1 Exemplo: Informação Estatística

Nesta seção vamos considerar uma urna com 3 bolas. Originalmente a urna tinaha 4 bolas, duas brancas e duas pretas, mas uma bola, não sabemos qual, foi perdida. A tarefa aqui é adivinhar quantas bolas brancas, θ, permanecem na urna.

Questão 1. Qual a distribuição a priori natural para θ?

Resposta: Se o estatístico não possui alguma informação adicional sobre a forma pela qual a bola foi perdida, seria razoável admitir-se que os eventos {x=1 bolas brancas} e {x=2 bolas brancas} podem ser considerados equiprováveis. Isto é,

θ	$p(1 \mid \theta)$
0	1/2
1	1/2

A distribuição a priori, $f(\theta)$, é

$$f(1) = f(2) = 1/2.$$

Questão 2. Desta urna (com 3 bolas), uma bola é retirada sem ser mostrada aos interessados. Seja X o número de bolas brancas obtidas nesta retirada da urna. Qual a distribuição amostral desse experimento?

Resposta:

x	θ	$p(x \mid \theta)$
0	1	2/3
1	1	1/3
0	2	1/3
1	2	2/3

Questão 3. Qual a distribuição a posteriori de $f(\theta \,|\, x)$?

Resposta:

$$\begin{aligned} & f(1 \,|\, 0) \\ &= \frac{f(1)p(0 \,|\, 1)}{f(1)p(0 \,|\, 1) + f(2)p(0 \,|\, 2)} \\ &= \frac{(1/2)(2/3)}{(1/2)(2/3) + (1/2)(1/3)} \;=\; \frac{2}{3} \;=\; 1 - f(2 \,|\, 0) \end{aligned}$$

Analogamente temos

$$f(2 \,|\, 1) = 1 - f(1 \,|\, 1) = \frac{2}{3}.$$

Questão 4. Suponha agora que, além de X, você pode observar Y, o número de bolas brancas numa segunda retirada, após ter devolvido a primeira bola à urna. Assim, $Y = 0$ ou $Y = 1$. Qual a função de verossimilhança após as observações, $L(\theta \,|\, [x, y])$?

Resposta (as funções sem (com) linha se referem às questões 4 e 5, (6 e 7):

θ	x	y	$L(\theta \,\|\, x, y)$	$L'(\theta \,\|\, x, y)$	$f(\theta \,\|\, x, y)$	$f'(\theta \,\|\, x, y)$
1	0	0	(2/3)(2/3)	(2/3)(1/2)	4/5	1
1	0	1	(2/3)(1/3)	(2/3)(1/2)	1/2	1/2
1	1	0	(1/3)(2/3)	(1/3)1	1/2	1/2
1	1	1	(1/3)(1/3)	(1/3)0	1/5	0
2	0	0	(1/3)(1/3)	(1/3)0	1/5	0
2	0	1	(1/3)(2/3)	(1/3)1	1/2	1/2
2	1	0	(2/3)(1/3)	(2/3)(1/2)	1/2	1/2
2	1	1	(2/3)(2/3)	(2/3)(1/2)	4/5	1

Questão 5: Qual a distribuição a posteriori associada a cada verossimilhança acima, $f(\theta \,|\, [x, y])$?

Resposta:

$$
\begin{aligned}
& f(1 \mid [0,0]) \\
&= \frac{f(1)p([0,0] \mid 1)}{f(1)p([0,0] \mid 1) + f(2)p([0,0] \mid 2)} \\
&= \frac{(1/2)(2/3)^2}{(1/2)(2/3)^2 + (1/2)(1/3)^2} \;=\; \frac{4}{4+1} \;=\; \frac{4}{5} \;=\; f(2 \mid [1,1]) \\
& f(1 \mid [1,1]) \;=\; f(2 \mid [0,0]) \;=\; 1 - \frac{4}{5} = \frac{1}{5} \\
& f(1 \mid [0,1]) \\
&= \frac{f(1)p([0,1] \mid 1)}{f(1)p([0,1] \mid 1) + f(2)p([0,1] \mid 2)} \\
&= \frac{(1/2)(2/3)(1/3)}{(1/2)(2/3)(1/3) + (1/2)(2/3)(1/3)} \;=\; \frac{1}{1+1} \;=\; \frac{1}{2} \\
&= f(2 \mid [0,1]) \;=\; f(1 \mid [1,0]) \;=\; f(2 \mid [1,0])
\end{aligned}
$$

Questão 6: Vamos supor na Questão 4 que para observar Y a primeira bola não é devolvida à urna. Qual a função de verossimilhança após as observações, $L'(\theta \mid [x,y])$?

Questão 7: Qual a distribuição a posteriori associada à Questão 6, $f'(\theta \mid [x,y])$?

Resposta:

$$
\begin{aligned}
& f(1 \mid [0,0]) \\
&= \frac{f(1)p([0,0] \mid 1)}{f(1)p([0,0] \mid 1) + f(2)p([0,0] \mid 2)} \\
&= \frac{(1/2)(1/3)}{(1/2)(1/3) + (1/2)0} \;=\; 1 \;=\; f(2 \mid [1,1]) \\
& f(1 \mid [0,1]) \\
&= \frac{f(1)p([0,1] \mid 1)}{f(1)p([0,1] \mid 1) + f(2)p([0,1] \mid 2)} \\
&= \frac{(1/2)(1/3)}{(1/2)(1/3) + (1/2)(1/3)} \;=\; \frac{1}{1+1} \;=\; \frac{1}{2} \\
&= f(2 \mid [0,1]) \;=\; f(2 \mid [1,0]) \\
&= f(1 \mid [0,1]) \;=\; f(2 \mid [1,0])
\end{aligned}
$$

Conclusão 1: Ao observarmos $[X = 1, Y = 0]$ ou $[X = 0, Y = 1]$, tanto no caso com reposição quanto no caso sem reposição, a distribuição a posteriori é igual à distribuição a priori. Assim nossa opinião após a realização do experimento não se alterou. Isto quer dizer que esses pontos não são informativos.

Conclusão 2: Ao observarmos $[X = 1, Y = 1]$ ou $[X = 0, Y = 0]$ as distribuições a posteriori são diferentes das distribuições a priori. Portanto esses são os pontos informativos.

Conclusão 3: A tabela acima nos mostra que, em todos os resultados informativos, i.e. $[X, Y]$ igual a $[0, 0]$ ou $[1, 1]$, o experimento sem reposição é mais informativo que o experimento com reposição. Todavia, um resultado não infornativo, i.e. $[X, Y]$ igual a $[0, 1]$ ou $[1, 0]$, é também mais provável no caso sem reposição, $2/3$, que no caso com reposição, $4/9$.

1.2.2 Exemplo: Estatística Forense

Um roubo foi cometido e um suspeito, um jovem, está sendo julgado. O assaltante aparentemente se cortou e seu sangue pingou no local do crime. O sangue do suspeito também foi coletado. Como estamos interessados em avaliar a probabilidade do suspeito ser o criminoso, devemos usar todas as evidências disponíveis. Foi observado que o sangue do local do crime é do mesmo tipo do sangue do suspeito.

Questão 8: Como o estatístico estabelece os elementos de sua análise?

Resposta:

Parâmetro: $\theta = 1\,(= 0)$ se o suspeito é culpado (inocente).
Observável 1: $X = 1\,(= 0)$ se o sangue do suspeito é do tipo A (não A).
Observável 2: $Y = 1\,(= 0)$ se o sangue do local é do tipo A (não A).

Questão 9: Como se implementa os elementos para a operação Bayesiana?

Resposta:

Consideremos que a priori $q = \Pr(\theta = 1)$. Por outro lado como o tipo de sangue deve ser independente da culpabilidade do suspeito, X é independente de θ.

Assim, $\Pr(X = 1 \,|\, \theta) = \Pr(X = 1) = p$. Finalmente, podemos escrever a probabilidade condicional de Y dado $[\theta, X]$:

$$\Pr(Y = y \,|\, X = x, \theta) = \begin{cases} p & \text{se } \theta = 0 \wedge y = 1 \\ 1 - p & \text{se } \theta = 0 \wedge y = 0 \\ \\ 1 & \text{se } \theta = 1 \wedge y = x \\ 0 & \text{se } \theta = 1 \wedge y \neq x \end{cases}$$

Note que tanto X como Y são independentes de θ. No entanto, o vetor $[X, Y]$ é dependente de θ.

Questão 10: Como se calcula a distribuição a posteriori de θ?

Resposta: O denominador na fórmula de Bayes é:

$$\begin{aligned}
&\Pr(X=1,Y=1)\\
&= \Pr(\theta=0)\Pr(X=1\,|\,\theta=0)\Pr(Y=1\,|\,X=1,\theta=0)\\
&+ \Pr(\theta=1)\Pr(X=1\,|\,\theta=1)\Pr(Y=1\,|\,X=1,\theta=1)\\
&= \Pr(\theta=0)\Pr(X=1)\Pr(Y=1\,|\,X=1,\theta=0)\\
&+ \Pr(\theta=1)\Pr(X=1)\Pr(Y=1\,|\,X=1,\theta=1)\\
&= (1-q)p^2+qp
\end{aligned}$$

Tomando o segundo termo da direita é numerador na fórmula de Bayes,

$$t \;=\; \Pr(\theta=1\,|\,X=1,Y=1) \;=\; \frac{qp}{(1-q)p^2+qp} \;=\; \frac{q}{(1-q)p+q} > q$$

Note que, se houver concordância nos sangues da evidência e do suspeito, a probabilidade a posteriori é maior do que a priori, como esperado. Note tamém que, quanto mais raro o tipo de sangue, mais informativa será uma concordância, i.e., $p \to 0 \Rightarrow t \to 1$. Por outro lado, quanto mais comum o tipo de sangue, menos informativa será uma concordância, i.e., $p \to 1 \Rightarrow t \to q$.

1.2.3 Exemplo: Diagnóstico Médico

Aido é um médico que fez um exame clínico em sua colega Leane. Ele julga que com probabilidade 1/10 ela está infectada pelo vírus HIV. Este julgamento ele fez ao comparar Leane com o banco de dados de sua clínica e também por sua vasta experiência de colaboração com infectologistas. Além disso, na opinião de Aido, 80% (30%) das pessoas que (não) são portadores do vírus reagiriam positivamente a um teste laboratorial. Leane se submeteu ao teste e reagiu positivamente.

Questão 11: Como estabelecer os elementos e efetuar a calibração da probabilidade de Aido?

Resposta:

Parâmetro	$\theta = 1\,(=0)$ se Leane (não) é portadora do HIV
Observável	$X = 1\,(=0)$ se Leane tem resultado positivo (negativo)

Assim,

$$f(1) = \Pr(\theta=1) = \frac{1}{10},$$

$$p(1\,|\,1) = \Pr(X=1\,|\,\theta=1) = \frac{8}{10} \text{ e}$$

$$p(1\,|\,0) = \frac{3}{10}.$$

Podemos escrever a probabilidade marginal de X como:

$$\begin{aligned}
&\Pr(X=1)\\
&\quad = \quad f(1)p(1\,|\,1) + f(0)p(1\,|\,1)\\
&\quad = \quad \frac{1}{10}\frac{8}{10} + \frac{9}{10}\frac{3}{10} \;=\; 0.08 + 0.27 \;=\; 0.35,\\
&\Pr(X=0)\\
&\quad = \quad f(1)p(0\,|\,1) + f(0)p(0\,|\,0)\\
&\quad = \quad \frac{1}{10}\frac{2}{10} + \frac{9}{10}\frac{7}{10} \;=\; 0.02 + 0.63 \;=\; 0.65,
\end{aligned}$$

Usando a fórmula de Bayes obtemos:

$$\begin{aligned}
f(1\,|\,1) \;&=\; \frac{8}{35} \;\approx\; 0.229 \;>\; 0.1 = f(1),\\
f(1\,|\,0) \;&=\; \frac{2}{65} \;\approx\; 0.031 \;<\; 0.1 = f(0).
\end{aligned}$$

Evidentemente, assumimos conhecidas as freqüências populacionais de resultados positivos do teste laboratorial. Contudo, em situações práticas existe uma incerteza associada a esses valores e distribuições a priori são utilizadas. Na parte de modelagem estatísticas apresentaremos uma solução mais realista para o problema do diagnóstico médico.

1.2.4 Exemplo: Paternidade

Antônio, um jovem filho de Maria, decide seguir o conselho de sua mãe e entra na justiça para que John, um rico empresário, reconheça a paternidade de Antônio. O Juiz então decide que tanto o demandado, John, quanto os demandantes, Antônio e Maria, se submetam ao exame de DNA em amostras de seus respectivos sangues. As análises dos sangues seguiram a técnica de Microsatélites pela Reação de Cadeia da Polimerase (PCR). O resultado dos 3 primeiros locos foram os seguintes:

Loco	Mãe	Demandante	Demandado
1	11—13	13—16	12—16
2	29—33	29—33	29—35
3	19—20	20—19	19—19

Representemos as freqüências gênicas envolvidas nos locos 1, 2 e 3 por fi, gi e hi, respectivamente. O índice varia de acordo com o número do alelo. Assim, f_{13} é a freqüência gênica do alelo 13 do loco 1. Aqui será resolvido o cálculo da probabilidade de paternidade apenas para o loco 1 e os outros dois devem ser feitos para o prazer do leitor.

Questão 12: Como o estatístico estabelece os elementos de sua análise?

Resposta:

Parâmetro	$\theta = 1\,(= 0)$ se o demandado (não) é o pai
Observável 1	$X = 1\,(= 0)$ se o genotipo do demandado (não) é (12—16)
Observável 2	$Y = 1\,(= 0)$ se o genotipo do demandante (não) é (13—16)

Questão 13: Como se implementa os elementos para a operação Bayesiana?

Resposta: Consideremos que a priori $q = \Pr(\theta = 1)$. Por outro lado,

$$\Pr(X = 1 \,|\, \theta = 1) = \Pr(X = 1 \,|\, \theta = 0) = \Pr(X = 1) = 2f_{12}f_{16}.$$

Finalmente, podemos escrever a probabilidade condicional de Y dado (θ, X):

$$\Pr(Y = 1 \,|\, X = 1, \theta = 1) = \frac{1}{4}\ ,$$

$$\Pr(Y = 1 \,|\, X = 1, \theta = 0) = \Pr(Y = 1 \,|\, \theta = 0) = \frac{1}{2}f_{16}\ .$$

Note que tanto X como Y são independentes de θ. No entanto, o vetor $[X, Y]$ é dependente de θ. Utilizamos aqui o equilíbrio de Hardy-Weinberg para a definição das probabilidades.

Questão 14: Como se calcula a distribuição a posteriori de θ?

Resposta:

$$\begin{aligned}
\Pr(X = 1, Y = 1) &= \Pr(X = 1, Y = 1, \theta = 0) + \Pr(X = 1, Y = 1, \theta = 1) \\
&= \Pr(\theta = 0)\Pr(X = 1, Y = 1 \,|\, \theta = 0) + \Pr(\theta = 1)\Pr(X = 1, Y = 1 \,|\, \theta = 1) \\
&= \Pr(\theta = 0)\Pr(X = 1 \,|\, \theta = 0)\Pr(Y = 1 \,|\, X = 1, \theta = 0) \\
&+ \Pr(\theta = 1)\Pr(X = 1 \,|\, \theta = 1)\Pr(Y = 1 \,|\, X = 1, \theta = 1) \\
&= \Pr(\theta = 0)\Pr(X = 1)\Pr(Y = 1 \,|\, X = 1, \theta = 0) \\
&+ \Pr(\theta = 1)\Pr(X = 1)\Pr(Y = 1 \,|\, X = 1, \theta = 1)
\end{aligned}$$

temos então que

$$\Pr(X = 1, Y = 1) = (1 - q)\frac{(2f_{12}f_{16})f_{16}}{2} + q\frac{(2f_{12}f_{16})f_{16}}{4}$$

Consequentemente teremos, dividindo o segundo termo pelo total,

$$\Pr(\theta = 1 \,|\, X = 1, Y = 1) = (1 + 2f_{16}(1 - q)/q)^{-1}\ .$$

No caso de incerteza total a priori, q=1/2, teríamos como posteriori de paternidade $[1 + 2f_{16}]^{-1}$. Quanto mais raro for o alelo 16 na população maior será o valor da probabilidade de paternidade.

Usando a probabilidade a posteriori do loco 1 como priori para o loco 2 e a do loco 2 para o loco 3, o leitor poderá perceber que observando-se mais locos a probabilidade de paternidade aumenta com o numero de locos estudados. O leitor poderá ter dificuldades no cálculo dos outros dois locos pois terá de considerar outros aspectos não discutidos no caso do loco 1.

1.2.5 Comentários Adicionais

Nesta introdução ao método estatístico indutivo, apresentamos os elementos básicos para o trabalho de um estatístico que deseja usar informações disponíveis no ambiente especializado que definiu o problema a ser resolvido. As idéias foram colocadas de forma elementar para realçar a mecânica da operação Bayesiana de indução.

Como leitura adicional recomendamos que o leitor olhe com especial atenção o livro do Professor David Blackwell, Basic Statistics, que foi traduzido em 1972. Outra leitura obrigatória, para quem deseja entrar com o pé direito na carreira estatística, é o livro do Professor Morris DeGroot. Para quem desejar leituras mais profundas sobre os fundamentos de Estatística recomendo os livros dos Professores Dev Basu, Oscar Kempthorne, Jack Good e Bruno De Finetti. As referências completas estão na seqüência. Divirtam-se!

Referências

D.Z.Albert (1992). *Quantum Mechanics and Experience.* Harvard U. Press.

R.Barlow (1998). *Engineering Reliability.* SIAM.

D.Blackwell (1969). *Basic Statistics.* McGraw-Hill.

M.DeGroot (1986). *Probability and Statistics.* Adison-Wesley.

B.d'Espagnat (1995). *Veiled Reality.* Adison-Wesley.

I.J.Good (1983). *Good Thinking.* University of Minnesota Press.

J.K.Gosh (editor) (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by D Basu.* Springer-Verlag.

O.Kempthorne, L.Folks (1971). *Probability, Statistics and Data Analysis.* Iowa U Press.

C.A.B.Pereira, M.Viana (1982). *Elementos de Inferência Bayesiana.* 5o SINAPE.

Capítulo 2

Análise Pré-Posteriori

2.1 Introdução

Neste capítulo discutimos conceitos fundamentais para um planejamento adequado sobre coleta de informações sobre o estado da natureza ou parâmetro θ. O material aqui apresentado, diferente do apêndice sobre entropia, é mais conceitual do que operacional embora apresente soluções para alguns problemas específicos. Ressaltamos que o objetivo é a diminuição da incerteza sobre θ. Isto é, a busca por boas inferências sobre θ é parte relevante do trabalho do estatístico.

Este capítulo é dividido em cinco seções incluindo esta introdução. A Seção 3.2 apresenta o conceito de informação como visto por Basu (1975). A Seção 3.3 introduz um conceito de informação mais operacional apresentado por DeGroot (1962). A Seção 3.4 é dedicada ao conceito de experimento mais informativo, introduzido por Blackwell (1951). O conceito de informação apresentado por Blackwell tem caráter mais teórico, mas ainda assim pode ser utilizado na escolha de experimentos. Com caráter mais aplicado, a Seção 3.5 mostra o cálculo do tamanho da amostra na perspectiva Bayesiana. Duas situações são discutidas: a amostragem por atributos, tanto para o modelo binomial quanto para o hiper-geométrico.

2.2 Informação Segundo Basu

A procura por uma definição de informação nos levou inicialmente a considerar que o conceito apresentado por Basu (1975), embora não operacional, parece ser o que melhor descreve o que se entende por informação.

Informação é o que ela faz por você, muda a sua opinião.

O caráter subjetivo desse conceito está intrínseco com a inclusão da pessoa que está

tentando obter informação. Muitas vezes um conjunto de observações pode não alterar em nada o conhecimento de um indivíduo, mas pode ser bastante relevante para um outro com paradigmas diferente do primeiro.

Para operacionalizar esse conceito abstrato, devemos tentar responder as seguintes questões fundamentais para um bom trabalho estatístico:

I. Informação sobre o que?

II. Onde está a informação?

III. Quanto de informação é usada?

IV. Como a informação é extraída?

Vejamos como poderíamos responder essas perguntas.

A informação é sobre o valor de θ e é descrita pela distribuição atual de θ. Essa descrição é baseada em avaliações probabilísticas. Informações adicionais podem ser culturais ou experimentais. Isto é, o maior envolvimento na área onde a pesquisa está sendo realizada produz um ganho cultural que pode modificar a distribuição de θ. Já a informação experimental é relativa a realização de experimentos (variáveis aleatórias), X, Y, Z etc., que são passíveis de serem realizados (observadas). Na verdade, o processo de incorporação da informação sobre θ, contida nos resultados experimentais, é um processo de adestramento, diferente do cultural. Aqui temos condições de discutir apenas esse segundo tipo de informação, a experimental.

Vamos supor que o experimento X está sendo realizado e que seu resultado, x, é observado. Como conseqüência natural, passamos de $f(\theta)$ para $f(\theta \mid x)$. Recordando apenas que essa calibração da informação é obtida pela operação de Bayes: $f(\theta \mid x) \propto L(\theta \mid x) f(\theta)$

Quanto ao valor da quantidade de informação extraída, devemos definir, de acordo com os interesses do problema que está sendo resolvido, uma distância entre a distribuição a priori e a posteriori. Essa distância muitas vezes está relacionada com custos ou utilidades definidas em determinadas situações como por exemplo am estudos de qualidade de vida. Na próxima seção utilizaremos a função utilidade como medida dessa distância.

Estamos aptos agora a responder as perguntas apresentadas acima:

i. Informação sobre o que?

Ri: Sobre θ !

ii. Onde está a informação?

Rii. Está em $L(\theta \mid x)$!

iii. Quanto de informação é usada?

Riii. O valor da distância $\mathrm{D}(f(\theta \mid x), f(\theta))$.

iv. Como a informação é extraída?

Riv. Pela operação de Bayes, $f(\theta \mid x) \propto L(\theta \mid x) f(\theta)$

Voltamos a seguir ao exemplo das bolas de gude descrito no Capítulo 1:

Exemplo 3.2.1: Bolas na Caixa

Consideremos quatro bolinhas de gude, duas brancas e duas pretas. Escolho três e coloco em uma caixa. Você deve adivinhar qual o número, θ, de bolinhas brancas dentro da caixa, um ou dois. Você pode, inicialmente, retirar ao acaso uma bolinha da caixa para ganhar informação adicional. Para tentar ganhar mais informação, pode retirar uma segunda bolinha da caixa e essa retirada pode ser com ou sem a reposição, na caixa, da primeira bolinha retirada. Para cada um dos resultados experimentais, calculamos as probabilidades a posteriori que foram apresentadas no Capítulo 1. Ao escolher o experimento, esperamos produzir maior informação com a nossa escolha. Vamos considerar a distância Euclidiana, DE, entre os vetores de probabilidades. Se calcularmos a distância entre piori e posteriori para cada resultado possível e depois avaliarmos o valor esperado em cada um dos experimentos, encontraremos o mesmo valor, 0.24, independente de qual experimento, X, $[X, Y]$ ou $[X, Z]$.

Observando que a maior distância possível (maior ganho?) é $\mathrm{DE}([0.5, 0.5], [0, 1]) = 0.71$, podemos concluir o seguinte:

i. Com apenas uma retirada, ao observarmos X, o ganho é 33% do possível, tanto para $X = 0$ quanto para $X = 1$. Isto é, ganho pequeno mas garantido!

ii. Com duas retiradas com reposição, ao observarmos $[X, Y]$, o ganho pode ser 60%, resultados $[0, 0]$ ou $[1, 1]$, com probabilidade 0.56 ou pode não haver ganho, 0% com probabilidade 0.44, caso dos resultados $[0, 1]$ e $[1, 0]$. Isto é, ganho moderado com risco moderado!

iii. Com duas retiradas sem reposição, ao observarmos $[X, Z]$, o ganho pode ser 100% com probabilidade 0.33, caso dos resultados $[0, 0]$ ou $[1, 1]$ e,novamente, pode não haver ganho, 0% com probabilidade $0, 67$ se o resultado for $[0, 1]$ ou $[1, 0]$. Isto é, ganho máximo com risco alto!

O exemplo acima descreve bem a racionalidade do cotidiano na sociedade. Contudo, a distância entre distribuições, como apresentada, não nos ajuda na análise pré-posteriori, para escolher qual o experimento que desejamos realizar. A intuição nos diz que talvez o fato de não utilizarmos a topologia do espaço paramétrico nos impede de uma avaliação mais apropriada. A próxima seção apresenta uma forma de considerarmos outra nuance do problema.

2.3 Informação Segundo DeGroot

Em um artigo fundamental da teoria da decisão, DeGroot (1962) considerou uma função real que tentativamente ordena o conjunto das distribuições possíveis para o parâmetro θ. Essa função foi denominada de Função Incerteza. A idéia é a seguinte: conforme quantidades, $X, Y, \ldots$, associadas a θ são observadas, a incerteza sobre este parâmetro deve diminuir. A seguir apresentamos mais formalmente a metodologia de DeGroot.

Consideremos um experimento genérico X cujo valor observado é x. Se f é a função de probabilidade a priori, f_x a função de probabilidade a posteriori, U uma função incerteza e E o operador média, então espera-se que $U(f) \geq E(U(f_x))$. Isto é, espera-se que a incerteza a posteriori seja menor do que a incerteza a priori. Usando a conhecida desigualdade de Jensen, DeGroot mostra que uma função real é uma função de incerteza se e somente se é côncava. Podemos escrever o resultado da seguinte forma: Para quaisquer priori $f(\theta)$ e experimento X, temos o seguinte:

$$I(X, f, U) = U(f) - E(U(f_X)) \geq 0 \Leftrightarrow U \text{ é côncava.}$$

O operador I é a quantidade de informação sobre θ contida no experimento X quando a priori é f e a incerteza é U. Note que anteriormente nos referíamos a informação contida no resultado x efetivamente observado de um experimento X. Aqui estamos falando da informação que se espera de um experimento em relação ao parâmetro.

No atual contexto o objetivo é a escolha de experimentos. Consideramos os caso onde existam uma série de experimentos e nem todos podem ser realizados. Os experimentos escolhidos para serem observados devem ser aqueles com maiores valores de I.

É importante ressaltar que as escolhas de f e de U, além de subjetivas, podem estar interligadas e muitas vezes podemos ter casos onde a escolha de uma implica na caracterização da outra. Voltaremos a essa questão após discutir alguns exemplos.

Exemplo 3.3.1: Variância como Incerteza.

Consideremos um parâmetro cujo espaço paramétrico está contido na reta. Vamos considerar uma distribuição a priori com média e variância finitas. Se a variância da distribuição de θ é considerada como função incerteza, podemos provar que a variância é uma função cncava. Isto é,

$$I(X, f, U) = V(f) - E(V(f_X)) = V(E(f_X)) \geq 0$$

Um exercício necessário para o leitor é provar que $V(\theta) = E(V(\theta \,|\, X)) + V(E(\theta \,|\, X))$. Chamamos atenção para o fato de $E(\theta \,|\, X)$ ser um estimador de Bayes e $V(E(\theta \,|\, X))$ sua variância. Assim estamos considerando um estimador, $E(\theta \,|\, X)$, que certamente é viciado no sentido da estatística clássica. Ao procurarmos qual experimento deve ser realizado, escolhemos o que produz um estimador com maior variância. Completamos esse exemplo

retornando ao Exemplo 3.2.1 das bolas na caixa. Lembremos que θ assume apenas os valores 1 ou 2. Após avaliarmos $E(\theta \mid X)$, $E(\theta \mid X, Y)$ e $E(\theta \mid X, Z)$, em cada um dos resultados possíveis, calculamos as respectivas variâncias e obtivemos 0.03, 0.05 e 0.08. Assim, o experimento mais informativo no sentido de DeGroot é $[X, Z]$ que produziu a maior variância. Maior variância, maior risco de não produzir informação e possibilidade de obter toda a informação possível!

Antes de finalizar esta seção gostaríamos de ressaltar o caráter subjetivo da escolha tanto da distribuição a priori quanto da função incerteza. Por exemplo, vamos supor que na estimação da proporção de uma característica, o modelo binomial é o modelo estatístico do problema. Consideremos que a família de distribuições beta é de onde a priori será escolhida. Em conjunto com a média, que define a estimativa inicial, a fixação da variância da priori estabelece a priori do problema. Neste caso, a escolha da variância como função incerteza está diretamente relacionada com a escolha da priori.

Na próxima seção discutiremos o conceito de suficiência de Blackwell. Novamente o objetivo é escolher o experimento mais informativo para inferências sobre um parâmetro de interesse. Mostraremos que, mesmo para estatísticos que não aceitam o uso de probabilidade para parâmetros, a escolha de um experimento para inferências, em presença de alternativas, pode ser feita racionalmente. Infelizmente, ao contrário do que apresentamos na presente seção, por evitar-se o uso de medidas de probabilidade sobre os espaços paramétricos, não se pode dizer quanto de informação um experimento X tem a mais, quando comparado com um experimento menos informativo, Y. Isto é, o conceito de suficiência de Blackwell introduz uma forma de identificar-se, quando existe, o experimento mais informativo dentre um conjunto de concorrentes.

2.4 Suficiência de Blackwell

O conceito de estatística suficiente está relacionado com o objetivo de reduzir a dimensão amostral sem perda de informação. Por exemplo, a média e a variância da amostra podem conter toda a informação contida em uma amostra de uma distribuição normal com média e variância desconhecidas. Note que neste caso o espaço de observações sofre apenas uma redução, pois no lugar de precisarmos de toda a amostra, podemos trabalhar apenas com um par de variáveis, sem perder informações relevantes sobre os parâmetros populacionais. Blackwell (1951) estendeu o conceito de suficiência ao considerar diferentes espaços estatísticos cujo único elemento em comum seria exatamente o parâmetro θ de interesse e desconhecido. A melhor forma de discutir a suficiência de Blackwell é por meio do seguinte exemplo. O material apresentado nesta seção pode ser encontrado em Basu e Pereira (1990).

Exemplo 3.4.1:

Vamos supor que uma nova empresa afirma estar produzindo um produto com o dobro da qualidade de seu concorrente. Isto é, a taxa de falha de seus produtos é metade da taxa de falha, θ, de seu concorrente. Para estimar θ devemos coletar amostras da nova empresa ou da antiga? Para responder a essa questão, consideremos dois experimentos de Bernoulli: $X \mid \theta \sim \text{Br}(\theta)$ e $Y \mid \theta \sim \text{Br}(\theta/2)$. Consideremos agora um experimento adicional que é o resultado de um lançamento de uma moeda: $Z \sim \text{Br}(1/2)$. Note que Z não é relacionado a θ e pode ser realizado em qualquer lugar e em qualquer tempo, em resumo, o exercício mais simples de aleatorização. Note que se considerarmos o experimento $Y' = ZX$ temos que Y' e Y são igualmente distribuídos: $Y' \sim Y$. Assim se observarmos X e depois realizarmos um exercício simples de aleatorização, chegamos a um experimento equivalente a Y.

O exemplo acima ilustra o conceito de suficiência de Blackwell. Isto é, suponha que dois experimentos X e Y são tais que suas distribuições dependem dos parâmetros $g(\theta)$ e $h(\theta)$, funções de um mesmo parâmetro θ. Dizemos que X é suficiente para Y em relação a θ, se existir um experimento Y', obtido por uma composição de X e algum experimento Z definido por um exercício de aleatorização (distribuição conhecida após a realização de X), tal que Y e Y' são identicamente distribuídos. Neste caso escrevemos $X \gg Y$. No exemplo, a primeira questão que se apresenta é se $X \ll Y$. A resposta é negativa como veremos a seguir.

Novamente, voltamos ao conceito de informação. Na verdade continuamos a procura por um experimento mais informativo sobre um parâmetro θ. Note que como Bayesianos poderíamos comparar experimentos olhando para as possíveis posteriores. Nossa procura por definições intuitivas nos levou a considerar o conceito de experimento mais informativo na perspectiva Bayesiana. No contexto acima e considerando todas as distribuições a priori dentro da classe considerada como alternativas, vamos definir suficiência comparando as distribuições a posteriori. Na perspectiva Bayesiana, escrevemos $X \gg Y$ se para qualquer ponto amostral y de Y, existir um ponto amostral x de X, tal que $f(\theta \mid y) = f(\theta \mid x)$. Fica para o leitor mostrar que suficiência de Blackwell implica na suficiência Bayesiana. Se estivermos comparando experimentos bem comportados e pudermos falar em todas as priores possíveis, as duas definições seriam equivalentes. Para isso teríamos de ser mais cuidadosos e lembrarmos dos conceitos de dominância e modelos discretos da teoria da medida. O argumento aqui apresentado faz-nos lembrar e entender a razão de Basu considerar que a suficiência de Blackwell é um conceito Bayesiano e a denominou de suficiência Bayesiana de Blackwell. Além disso, podemos dizer, com essa equivalência, que a suficiência de Blackwell não viola o princípio da verossimilhança, pois uma inferência sobre θ com a observação $Y = y$ produz o mesmo resultado se for realizada com $X = x$, pois as verossimilhanças decorrentes são proporcionais. A seguir apresentamos um pouco da teoria de Blackwell.

Definição: Função de Transição

Consideremos dois experimentos X e Y cujos espaços amostrais são representados por $\mathcal{X}$ e $\mathcal{Y}$. Uma função de transição, F, de $\mathcal{X}$ para $\mathcal{Y}$ é uma família $F = \{f_x(y); x \in \mathcal{X}\}$ de funções (densidade) de probabilidade $f_x(y)$ definidas em $\mathcal{Y}$ e indexadas por $x \in \mathcal{X}$.

Por exemplo, a família de funções de probabilidade Hipergeométrica, $H(y; x, n, N)$, é uma função de transição de $\{0, 1, ..., N\}$ para $\{0, 1, ...n\}$, onde $n < N$ são inteiros positivos. Poderíamos aqui pensar em uma máquina produzindo peças com taxa de falha θ. Consideremos um lote de N dessas peças e uma amostra de n peças desse lote. A função de transição Hipergeométrica é na verdade a distribuição condicional de $Y \,|\, X$, onde Y é o número de defeituosas da amostra e X é o número de defeituosas do lote. Importante notar que H é a mesma função para todos os valores de θ.

Definição: Suficiência de Blackwell

Sejam X e Y dois experimentos, como pensados acima, e com funções (densidade) de probabilidade $g(x \,|\, theta)$ e $h(y \,|\, \theta)$. Dizemos que X é suficiente para Y, com respeito a θ, no sentido de Blackwell, se existir uma função de transição tal que

$$h(y \,|\, \theta) = \sum_x f_x(y) g(x \,|\, \theta).$$

(No caso de modelos dominados, devemos trocar somatórios por integrais.)

O teorema a seguir resolve o problema da comparação de experimentos de Bernoulli e mostra como o problema da amostragem em populações finitas privilegia a amostragem sem reposição.

Teorema: Comparações de Experimentos de Bernoulli

Sejam $X \,|\, \theta \sim \text{Br}(p(\theta))$ $Y \,|\, \theta \sim \text{Br}(q(\theta))$. Os experimentos X e Y são comparáveis no sentido de Blackwell se o conjunto $\{[p(\theta), q(\theta)] : \theta \in \Theta\}$ estiver contido em uma reta que corta dois lados opostos do quadrado unitário $[0, 1]^2$. Ademais, $X \gg_\theta Y$ (ou $X \ll_\theta Y$) se a reta corta os lados verticais (horizontais) do quadrado. No caso da reta ser diagonal, X e Y são equivalentes, e escrevemos $X \approx_\theta Y$.

Prova:

Quando tratamos de distribuições de Bernoully, X e Y, a existência de uma função de transição corresponde à existência de uma matriz de transição. Isto é, vamos supor que X é suficiente para Y. Isto quer dizer que $P(Y = 1) = P(X = 0 \,|\, \theta) f_0(1) + P(X = 1 \,|\, \theta) f_1(1)$. Ou seja,

$$q(\theta) = (1 - p(\theta)) f_0(1) + p(\theta) f_1(1) = f_0(1) + p(\theta)(f_1(1) - f_0(1)).$$

Por outro lado

$$1 - q(\theta) = (1 - p(\theta))(1 - f_0(1)) + p(\theta)(1 - f_1(1)).$$

Isto é:

$$\begin{bmatrix} q(\theta) & 1 - q(\theta) \end{bmatrix} = \begin{bmatrix} p(\theta) & 1 - p(\theta) \end{bmatrix} \begin{bmatrix} f_1(1) & f_1(0) \\ f_0(1) & f_0(0) \end{bmatrix}$$

Para entender que a prova está completa basta escrever $q(\theta) = a+bp(\theta)$ onde $a = f_0(1)$ e $b = f_1(1) - f_0(1)$.

Voltando ao Exêmplo 2.4.1 verificamos que $X \gg_\theta Y$, mas todavia não temos $Y \gg_\theta X$.

Finalizamos essa seção mostrando que a amostragem sem reposição é mais informativa que a com reposição quando retiramos uma amostra de peças de um lote para estimar o número de defeituosas do lote.

Exemplo 3.4.5: Amostragem

Consideremos um lote de N peças e θ o parâmetro de interesse, número de defeituosas do lote. Se vamos retirar duas peças do lote, devemos decidir se a segunda retirada deve ser com reposição ou sem reposição. Representemos por $[X, Y]$ o experimento com reposição e por $[X, Z]$ o sem reposição. Isto é:

$$X \,|\, \theta \sim \text{Br}(\theta/N) \ , \quad Y \,|\, [X, \theta] \sim Y \,|\, \theta \sim \text{Br}(\theta/N) \ \text{ e } \ Z \,|\, [X, \theta] \sim \text{Br}((\theta - X)/(N - 1)) \ .$$

De fato temos de estudar se o conjunto formado pelos pontos

$$[P_X(\theta), Q_X(\theta)] = \left[\frac{\theta - x}{N - 1}, \frac{\theta}{N}\right]$$

pertence a uma reta que corta os eixos opostos do quadrado unitário, tanto para $X = 0$ quanto para $X = 1$. Mas é evidente que

$$Q_0(\theta) = P_0(\theta)\frac{N - 1}{N} \quad \text{e} \quad Q_1(\theta) = P_1(\theta)\frac{N - 1}{N} + \frac{1}{N}$$

Isto mostra que, para estimar θ, $[X, Z]$ é mais informativo do que $[X, Y]$ no sentido de Blackwell. Falta ainda mostrarmos qual seria o exercício de aleatorização que nos permitiria construir um experimento equivalente a $[X, Y]$ a partir de $[X, Z]$. O leitor fica encarregado de mostrar que: definindo o experimento $Y^* \,|\, [X, Z]$ distribuído como Bernoulli com probabilidade $(X + (N - 1)Z)/N$, então $[X, Y^*] \,|\, \theta \sim [X, Y] \,|\, \theta$. Por outro lado, as matrizes de transição correspondentes a $X = 0$ e $X = 1$ são as seguintes:

$$T_0 = \begin{bmatrix} 1 & 0 \\ \frac{1}{N} & \frac{N-1}{N} \end{bmatrix} \ , \quad T_1 = \begin{bmatrix} \frac{N-1}{N} & \frac{1}{N} \\ 0 & 1 \end{bmatrix} \ ,$$

Vamos finalizar essa seção com um problema proposto por Blacwell (1951). Suponha que desejamos estimar o número, θ, de mulheres fumantes na população de professores da USP. Sabemos o tamanho das populações de Professores, N, de Mulheres, M, e de Fumantes F. De qual população devemos retirar a primeira pessoa a ser estudada? Da população de professores, da população de mulheres, da população de homens, da população de fumantes ou da população de não fumantes? Por simplicidade e sem perda de generalidade vamos supor que $0 < M < F < N - F < N - M < 1$. O leitor deve responder estas perguntas para melhor aproveitar esta seção.

2.5 Tamanho de Amostra

Não é raro o estatístico ter de tentar responder a pergunta "Qual o tamanho de amostra que devo selecionar?" No entanto esta é uma pergunta que pode ser difícil de ser respondida visto que depende de muitas considerações, suposições e restrições sobre o problema tratado. Recomendar a seleção da maior amostra possível é sempre uma resposta que satisfaz as qualidades estatísticas. Restrições de custo, no entanto, podem inviabilizar essa resposta. Por outro lado, restrições na precisão e na incerteza irão colaborar para um compromisso entre todos os limites estabelecidos. Neste capítulo mostraremos como podemos estabelecer o tamanho de amostra respeitando as diversas restrições do problema da estimação de proporções populacionais. Consideraremos os dois casos conhecidos, Populações Infinitas (tamanho desconhecido) e Populações Finitas (tamanho conhecido).

2.6 Amostras de Populações Infinitas

Aqui o nosso interesse é o parâmetro θ de um modelo Binomial. Isto é, estamos interessados na taxa de falha, θ, a proporção de peças defeituosas produzidas por uma máquina. Consideramos o caso em que o processo de produção está sob controle, Isto é, quando as unidades podem ser consideradas como sendo permutáveis. Equivalentemente, consideramos que as unidades são produzidas de acordo com um processo de Bernoulli com probabilidade θ. Representando o processo por $\{U_i\}_{i\geq 1}$, as variáveis U_i são, condicionalmente a θ, independentes e igualmente distribuídas segundo uma $\mathrm{Br}(\theta)$. Dessa forma temos um processo de Bernoulli, $U_1, U_2, \ldots$ definido da seguinte forma: $Pr(U_i = 1 \,|\, \theta) = 1 - Pr(U_i = 0 \,|\, \theta) = \theta$, onde a sequência $U_1, U_2, \ldots$ é composta de variáveis mutuamente independentes dado θ.

O nosso objetivo é estabelecer o número, n, de elementos da seqüência que serão observadas. Por populações infinitas entendemos aquelas que não conhecemos seu tamanho.

Os números observados de defeituosas e de boas da amostra serão denotados por X e Y, respectivamente e o tamanho da amostra por $X + Y = n$. Os parâmetros da priori beta são denotados por a e $b > 1$ e a soma por $n_0 = a + b$. Após observada a amostra, x e y são os valores assumidos por X e Y. Os parâmetros da posteriori Beta serão denotados por $A = x + a$ e $B = y + b$, e assim $A + B = n + n_0$.

O modelo estatístico é sumarizado a seguir:

Priori: $\theta \sim \mathrm{Bt}(a, b)$.

Dist. amostral: $X \,|\, [\theta, n] \sim \mathrm{Bi}(n, \theta)$.

Posteriori: $\theta \,|\, [x, n] \sim \mathrm{Bt}(A, B)$.

Média a priori: $e = E(\theta) = a/n_0$.

Média amostral: $m = E(X \mid \theta, n) = \theta n$.

Média a posteriori: $e_n = E(\theta \mid n, x) = A/(n + n_0)$.

Variância a priori: $\nu = V(\theta) = e(1 - e)/(n_0 + 1)$.

Variância amostral: $V(X \mid \theta, n) = \theta(1 - \theta)n$.

Variância a posteriori: $\nu_n = V(\theta \mid n, x) = e_n(1 - e_n)/(n + n_0 + 1)$.

Note que $\nu_n < (4(n + n_0 + 1))^{-1}$. Se considerarmos apenas valores de a e b maiores que a unidade temos que: $\nu_n < (4(n+3))^{-1}$. Assim a pior situação será obtida quando a priori é não informativa $a = b = 1$ e a posteriori for simétrica, $A = B$.

Vamos supor que para ser informativo, o intervalo que será fornecido para a inferência sobre θ não deva ser superior a 0.1. Assim o erro não pode ser superior a 0.05. Vamos imaginar que $[I_1, I_2]$ seja o intervalo final e como a pior situação é o caso simétrico, com priori não informativa teríamos que $I_1 = e_n - td$ e $I_2 = e_n + td$. Aqui, t é um multiplicador do desvio padrão, d, da posteriori. A escolha de t depende da credibilidade que se deseja para o intervalo $[I_1, I_2]$. Vamos imaginar que $\Pr(I_1 < \theta < I_2 \mid x) = 0.95$. Imitando a normal vamos considerar que $t = 1.96$. Assim, $1,96(4(n + 3))^{-0,5} < 0.05$ é a desigualdade que usaremos para encontrar o tamanho de nossa amostra. Teríamos assim, $n > 381.16$. Tomando-se então $n = 382$, uma posteriori simétrica resultaria em uma distribuição Bt$(192, 192)$. O intervalo simétrico da pior amostra possível, $[0.45, 0.55]$, teria uma credibilidade de $1 - 2\Pr(\theta < 0.45) = 1 - 2(0.025) = 0.95$.

Vamos supor agora que com a amostra de 382 observações, foram obtidos 92 sucessos e 290 fracassos. Com uma uniforme como priori a posteriori seria Bt$(93, 291)$. Com essa distribuição a posteriori, o intervalo de credibilidade teria um comprimento menor do que 0.1. De fato, o intervalo com aproximadamente 95% de credibilidade seria: $[0.20, 0.29]$ com comprimento 0.09. Lembremos que na posteriori, a média é 0.24 eo desvio padrão 0.022. Usando-se a aproximação normal para esta distribuição teríamos uma credibilidade de 0.95. Para o intervalo simétrico $[0.20, 0, 29]$ a aproximação normal daria a credibilidade de 0.95. Note que os valores calculados tanto para a média como para o desvio padrão são obtidos com a distribuição Beta.

A Tabela 3.1 apresenta tamanhos amostrais para diferentes restrições:

	Credibility (percent)							
$I_2 - I_1$	90.00	95.00	99.00	99.50	99.80	99.90	99.95	99.99
0.05	1080	1534	2650	3149	3817	4327	4844	6050
0.06	749	1065	1839	2186	2650	3004	3363	4201
0.07	550	781	1351	1605	1946	2206	2470	3086
0.08	420	598	1033	1229	1489	1689	1891	2362
0.09	331	472	816	970	1176	1334	1493	1866
0.1	268	382	660	785	952	1080	1209	1511
0.11	221	315	545	649	787	892	999	1248
0.12	185	264	458	545	660	749	839	1048
0.13	158	225	390	464	562	638	714	893
0.14	135	193	336	399	485	550	616	769
0.15	118	168	292	348	422	478	536	670
0.16	103	147	256	305	370	420	471	589
0.17	91	130	227	270	328	372	417	521
0.18	81	116	202	241	292	331	371	464
0.19	72	104	181	216	262	297	333	417
0.2	65	93	163	194	236	268	300	376
	Normal values for the corresponding tails							
	1.65	1.96	2.58	2.81	3.09	3.29	3.48	3.89

Tabela 3.1: Tamanhos de amostra vs. credibilidade

2.7 Amostras de Populações Finitas

Por populações finitas entendemos aquelas que conhecemos seu tamanho. Vamos imaginar que temos uma população com N unidades, onde N é um inteiro positivo e conhecido. O nosso interesse agora é a estimação do número θ de unidades defeituosas em um lote de N peças. Evidentemente, θ é um inteiro positivo com valor desconhecido com limite superior N. Sem perda de generalidade podemos ordenar as unidades e identifica-las por sua ordem. Isto é o conjunto de unidades será $\{1, 2, \ldots N\}$. Assim para a unidade i, estaremos associando uma variável U_i que assume o valor 1 se i for defeituosa e 0 se for uma peça boa. Dessa forma nosso vetor de valores populacionais pode ser representado por $\psi = [U_1, U_2, \ldots U_N]$ e $\theta = 1'\psi$. É natural supormos que trocarmos a ordem das unidades não altere o que se espera do vetor da distribuição de ψ. Dessa forma, consideremos que a distribuição de ψ é permutável; isto é, para qualquer permutação p,

$$\Pr(U_1 = u_1, U_2 = u_2, \ldots U_N = u_N) = \Pr(U_{p(1)} = u_1, U_{p(2)} = u_2, \ldots U_{p(N)} = u_N) \ .$$

Sob a condição de permutabilidade, pelo teorema de De Finneti, $\{U_i\}$ é um processo de Bernoulli com probabilidade de sucesso igual a π. Dessa forma é natural considerarmos

que $\theta \,|\, \pi$ tem distribuição binomial com parâmetros $[N, \pi]$.

Com o objetivo de estimar o parâmetro de interesse θ, consideremos uma amostra de unidades populacionais, digamos $\{1, 2, \ldots n\}$, sem perda de generalidades, pois estamos nos restringindo ao processo permutável. Assim, temos a independência condicional, dado π, entre a amostra $\{U_1, U_2, \ldots U_n\}$ e o complemento populacional da amostra $\{U_{n+1}, U_{n+2}, \ldots U_N\}$. Se denotarmos por $X = U_1 + U_2 + \ldots + U_n$ o total amostral, como conseqüência do teorema de De Finneti:

$$\Pr(X = x, \theta - X = k \,|\, \pi) = \Pr(X = x \,|\, \pi) Pr(\theta - X = k \,|\, \pi) \ .$$

Lembremos que π é apenas um parâmetro que entrou no desenvolvimento apenas por conveniência teórica e assim não existe algum interesse em seu valor. O nosso primeiro interesse é a obtenção do modelo estatístico, $\Pr(X = x \,|\, \theta)$.

Notemos que:

$$\Pr(X = x \,|\, \theta = t) = \Pr(X = x, \theta = t \,|\, \pi)/\Pr(\theta = t \,|\, \pi) =$$

$$\Pr(X = x, \theta - X = t - x \,|\, \pi)/\Pr(\theta = t \,|\, \pi) =$$

$$\Pr(X = x \,|\, \pi)\Pr(\theta - X = t - x \,|\, \pi)/\Pr(\theta = t \,|\, \pi) \ .$$

Por outro lado, pela permutabilidade do processo definido, sabemos que:

i. X e $\theta - X$ são condicionalmente independentes dado π,

ii. $X \,|\, [\pi, n] \sim \text{Bi}(n, \pi)$,

iii. $\theta - X \,|\, [\pi, n, N] \sim \text{Bi}(N - n, \pi)$ e

iv. $\theta \,|\, [\pi, N] \sim \text{Bi}(N, \pi)$.

Com essas propriedades podemos escrever a seguinte igualdade:

$$\Pr(X = x \,|\, \theta = t, \pi) = \Pr(X = x, \theta = t \,|\, \pi)/\Pr(\theta = t \,|\, \pi) =$$

$$\Pr(X = x, theta - X = t - x \,|\, \pi)/\Pr(\theta = t \,|\, \pi) \ .$$

Então,

$$\Pr(X = x \,|\, \theta = t, \pi) = \Pr(X = x, \theta = t \,|\, \pi)/\Pr(\theta = t \,|\, \pi) =$$

$$\Pr(X = x \,|\, \pi)\Pr(\theta - X = t - x \,|\, \pi)/\Pr(\theta = t \,|\, \pi) \ .$$

Isto é,

$$\Pr(X = x \,|\, \theta = t, \pi) = \text{Bi}(n, \pi)\text{Bi}(N - n, \pi)/\text{Bi}(N, \pi) = H(x, t; n, N) \ ,$$

onde H é a função de probabilidade hipergeométrica avaliada no ponto x, com parâmetro t e tamanhos de amostra e população dados por n e N. Como esta função não depende de π, concluímos que $\Pr(X = x \,|\, \theta = t, \pi) = \Pr(X = x \,|\, \theta = t)$.

Notamos também que ao observarmos na amostra o valor de X, o valor de $\theta - X$ passa a ser o parâmetro de interesse. Utilizando a mesma notação do caso infinito, ao considerarmos uma distribuição a priori $\mathrm{Bt}(a, b)$ para π, a distribuição a posteriori seria $\mathrm{Bt}(A, B)$. Como nosso interesse é a distribuição de $(\theta - X) \,|\, X$, utilizamos a distribuição de

$$(\theta - X) \,|\, [\pi, X] \sim (\theta - X) \,|\, \pi \sim Bin(N - n, \pi)$$

e calculamos a preditiva de $\theta - X$ ao eliminarmos π por sua posteriori $\mathrm{Bt}(A, B)$. Finalmente temos $(\theta - X) \,|\, X \sim \mathrm{Bi}(N - n; A, B)$, como detalhado no Capítulo 3. Isto é,

$$\Pr((\theta - X) = k \,|\, X = x) = \frac{\binom{N-k}{k} \Gamma(n + n_0)\Gamma(k + A)\Gamma(N - n - k + B)}{\Gamma(A)\Gamma(B)\Gamma(N + n_0)}$$

A média e a variância dessa distribuição são, respectivamente,

$$E(\theta - X \,|\, X = x) = E\left(E\left(\theta - X \,|\, X = x, \pi\right) \,|\, X = x\right) = E\left(E\left(\theta - X \,|\, \pi\right) \,|\, X = x\right) =$$

$$(N - n)E\left(\pi \,|\, X = x\right) = (N - n)A/(n + n_0) = (N - n)e_n$$

$$V(\theta - X \,|\, X = x) = \left((N - n)\left(N + n_0\right) / \left(n + n_0 + 1\right)\right) e_n \left(1 - e_n\right) \ .$$

Novamente aqui podemos notar que o pior caso é aquele onde a média a posteriori de π, e_n, é igual a 0.5. Consideremos agora um novo parâmetro definido por $\delta = \theta/N$. Vamos nos concentrar nessa proporção para estabelecer que o comprimento de um intervalo de credibilidade para δ não deve exceder a um valor prefixado, digamos $2\epsilon = 0.1$. Note que en é a média a posteriori de δ cuja variância a posteriori seria:

$$V_n = V(\delta \,|\, X = x) = \frac{(N - n)(N + n_0)e_n(1 - e_n)}{(n + n_0 + 1)} \leq \frac{(N - n)(N + n_0)}{4(n + n_0 + 1)N^2}$$

Como antes iremos procurar um intervalo $[I_1, I_2]$ tal que $\Pr(I_1 < \delta < I_2 \,|\, x) = 0.95$. Isto é, iremos encontrar o valor de t tal que $I_1 = e_n - td$ e $I_2 = e_n + td$. Aqui, t é um multiplicador do desvio padrão, d, da posteriori de δ. Neste caso, além da variável observada, também o parâmetro de interesse é discreto. Assim, os padrões da distribuição normal podem não funcionar. No entanto se conseguirmos fixar o valor do desvio padrão do parâmetro δ, encontraremos o valor de n necessário para atingir o limite fixado. Por exemplo, suponha que fixamos $td = 2d = 0.05$. Neste caso, considerando $d = 0.025$ e $a = b = 1(n_0 = 2)$, para $N = 5000$, teríamos $n = 394$. Com esse tamanho de amostra, o caso simétrico seria aquele onde $A = B = 198$ que corresponderia a $x = y = 197$. A Tabela 3.2 apresenta os tamanhos de amostra para diferentes restrições.

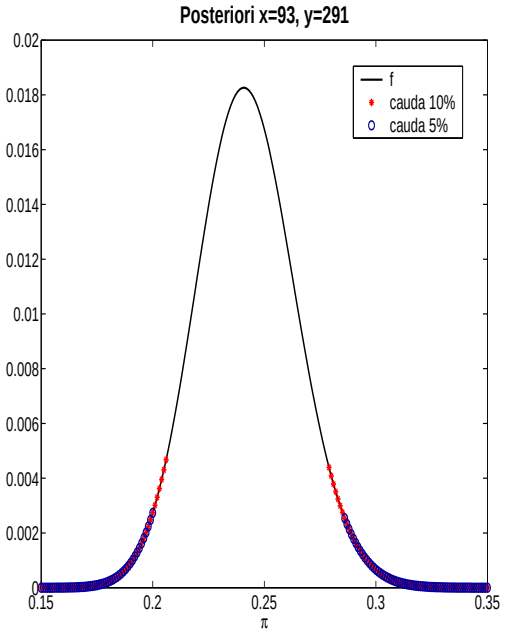
		Desvio Padrao				
Param.	Popul.	0.01	0.025	0.05	0.07	0.1
π	∞	2473	394	95	48	22
δ	5000	1655	365	93	48	22
δ	4000	1530	358	93	47	22
δ	3000	1358	348	92	47	22
δ	2000	1108	329	90	47	22
δ	1500	934	313	89	47	22
δ	1000	712	284	86	46	22
δ	750	578	261	84	45	21
δ	500	416	221	80	44	21
δ	400	345	199	78	43	20
δ	300	268	171	73	41	20
δ	200	185	133	65	39	20
δ	100	96	83	49	33	18
δ	50	49	44	33	24	15

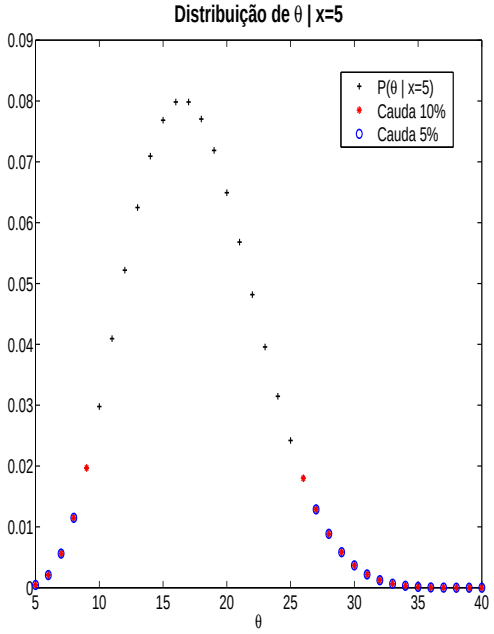
Tabela 3.2: Tamanhos de amostra vs. desvio padrão

Como exemplo, vamos supor que desejamos considerar uma amostra de um lote de $N = 50$ peças. Consideremos uma amostra de $n = 15$ peças o número X de peças defeituosas nessa amostra. Suponha que encontramos $x = 5$. Concluímos aqui que o conjunto $\Theta_0 = \{10, 11, ..., 25\}$ é um conjunto de 90.69% de credibilidade para θ. Isto é,

$$\Pr(\theta \in \Theta_0) = \Pr(0.2 \leq \delta \leq 0.5) = 0.91 \ .$$

Os gráficos das distribuições de probabilidades a posteriori de π no exemplo de populações infinitas e de θ, no caso acima são apresentados na seqüência.





Referências

D.Basu (1975). Statistical information and likelihood. *Sankhya* A 37, 1-71. Also in Lecture notes in statistics 45, Springer.

M.DeGroot (1962). Uncertainty, Information, and Sequential Experiments. *Annals of Mathematical Statistics,* 33, 404-19.

D.Blackwell (1951). Comparison of experiments. *Proceedings of the 2nd Berkeley Symposium,* 93-102.

D.Basu, C.A.B.Pereira (1990). Blackwell Sufficiency and Bernoulli Experiments. *Brazilian Journal of Probability and Statistics,* 4, 137-45.

Capítulo 3

Distribuições Derivadas do Processo de Bernoulli

3.1 Preliminares

O objetivo deste capítulo é apresentar as propriedades de algumas distribuições discretras derivadas do processo de Bernoulli, e de algumas distribuições contínuas a elas assossiadas. Estas distribuições aparecem naturalmente em processos de contagem, e são a ferramenta mais natural para tratar dados discretos ou categóricos. Uma grande variedade de problemas estatísticos admitem modelos discretizados, sendo esta uma forma possível de introduzir soluções não paramétricas.

O desenvolvimento da teoria apresentada é original, visando um tratamento unificado de uma grande variedade gama de problemas, como populações finitas e infinitas, tabelas de contengência de dimensão arbitrárias, dados deficientemente categorizados, regreções logísticas, etc. O texto utiliza uma representação singular que não é usual em textos de estatística. Todavia, a representação singular facilita a extenção dos resultados e a implementação numérica e computacional.

3.2 Notação

Inicialmente definimos algumas notações matriciais. O operador $r:s:t$, lê-se - *de* r *até* s com *passo* t, indica quer o vetor $[r, r+s, r+2s, \ldots t]$ ou o correspondente domínio de íncices. $r:s$ é uma abreviação de $r:1:s$. Usualmente escrevemos uma matriz, A, como o índice de linha subscrito, e o índice de coluna superscrito. Assim, A_i^j é o elemento na i-ésima linha e j-ésima coluna da matriz A. Vetores de índices podem ser usados para montar uma matriz extraindo de uma matriz maior um determinado sub-conjunto de linhas e colunas. Por exemplo $A_{1:m/2}^{n/2:n}$ é o bloco noroeste, i.e. o bloco com as primeiras

linhas e ultimas colunas, de A. Alternativamente, podemos escrever uma matriz com índices de linha e coluna entre parenteses, i.e. podemos escrever o bloco noroeste como $A(1:m/2, n/2:n)$.

$V > 0$ é uma matriz positiva definida. O operador diag, quando aplicado a uma matriz quadrada, extrai o vetor na diagonal principal, e quando aplicado a um vetor, produz a matriz diagonal correspondente.

$$\mathrm{diag}(A) = \begin{bmatrix} A^1_1 \\ A^2_2 \\ \vdots \\ A^n_n \end{bmatrix} \quad , \quad \mathrm{diag}(a) = \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_n \end{bmatrix}$$

Uma lista de matrizes pode ser indexada por índices subscritos ou superscritos à esquerda. No caso de matrizes blocadas, estes indices à esquerda indicam os blocos de linhas (subscritos) e colunas (superscritos), como por exemplo na matriz

$$A = \begin{bmatrix} {}^1_1A & {}^2_1A & \dots & {}^s_1A \\ {}^1_2A & {}^2_2A & \dots & {}^s_2A \\ \vdots & \vdots & \ddots & \vdots \\ {}^1_rA & {}^2_rA & \dots & {}^s_rA \end{bmatrix}$$

Assim, ${}^s_rA^j_i$ é o elemento na i-ésima linha e j-ésima coluna do bloco situado no r-ésimo bloco de linhas e s-ésimo bloco de colunas da matriz A. Alternativamente, podemos escrever os índices de bloco entre chaves, i.e. podemos escrever ${}^s_rA^j_i$ como $A\{r,s\}(i,j)$.

O operador Vec empilha as colunas da matriz argumento em um único vetor.
O produto de Kronecker (ou produto direto, ou tensorial), $\otimes$, é definido como segue:

$$\mathrm{Vec}(U^{1:n}) = \begin{bmatrix} u^1 \\ u^2 \\ \vdots \\ u^n \end{bmatrix} \quad , \quad A \otimes B = \begin{bmatrix} A^1_1B & A^2_1B & \dots & A^n_1B \\ A^1_2B & A^2_2B & \dots & A^n_2B \\ \vdots & \vdots & \ddots & \vdots \\ A^1_mB & A^2_mB & \dots & A^n_mB \end{bmatrix}$$

Introduziremos agora conceitos relativos a permutação partição de índices. Seja $1:m$ um domínio de índices, no contexto deste capítulo, índices de classe. Seja $p = \sigma(1:m)$ uma permutação destes índices de classe. A correspondente Matriz de Permutação (de Linhas) é

$$P = I_p = \begin{bmatrix} I_{p(1)} \\ \vdots \\ I_{p(m)} \end{bmatrix} \quad \text{assim} \quad P \begin{bmatrix} 1 \\ \vdots \\ m \end{bmatrix} = \begin{bmatrix} p(1) \\ \vdots \\ p(m) \end{bmatrix}$$

Com um vetor de permutação, p, e um vetor de terminação, t, definimos uma partição

das m classes originais em s super-classes:

$$\left[\begin{array}{c} p\,(1) \\ \vdots \\ p\,(t(1)) \end{array}\right] , \left[\begin{array}{c} p\,(t(1)+1) \\ \vdots \\ p\,(t(2)) \end{array}\right] \ldots \left[\begin{array}{c} p\,(t(s-1)+1) \\ \vdots \\ p\,(t(s)) \end{array}\right]$$

$$\text{onde } t(0) = 0 < t(1) < \ldots < t(s-1) < t(s) = m$$

Definimos as correspondentes matrizes de permutação, P, e partição, T, como

$$P = I_{p(1\,:\,m)} = \left[\begin{array}{c} {}_1P \\ {}_2P \\ \vdots \\ {}_sP \end{array}\right] \quad , \quad {}_rP = I_{p(t(r-1)+1\,:\,t(r))} \quad ,$$

$$T_r = \mathbf{1}'\,({}_rP) \quad \text{e} \quad T = \left[\begin{array}{c} T_1 \\ \vdots \\ T_s \end{array}\right]$$

Estas matrizes facilitam escrever funções da partição como

- Os índices (classes) na super-classe r

$${}_rP\,(1\,:m) = {}_rP \left[\begin{array}{c} 1 \\ \vdots \\ m \end{array}\right] = \left[\begin{array}{c} p\,(t(r-1)+1) \\ \vdots \\ p\,(t(r)) \end{array}\right]$$

- O número de classes na super classe r

$$T_r\,\mathbf{1} = t(r) - t(r-1)$$

- Uma sub-matriz com indices de linha na super-classe r

$${}_rP\,A = \left[\begin{array}{c} A_{p(t(r-1)+1)} \\ \vdots \\ A_{p(t(r))} \end{array}\right]$$

- A soma das linhas de uma sub-matriz com indices de linha na super-classe r

$$T_r\,A = \mathbf{1}'\,({}_rP\,A)$$

- As linhas de uma matriz, somadas por super-classe

$$T\,A = \left[\begin{array}{c} T_1\,A \\ \vdots \\ T_s\,A \end{array}\right]$$

Note que uma matriz T representa uma partição de m-classes em s-super-classes se T tem dimensão $s \times m$, $T_h^j \in \{0,1\}$ e T tem linhas ortogonais. O elemento T_h^j indica de a classe $j \in 1:m$ está na super-classe $h \in 1:s$.

Introduzimos as seguintes notações para as matrizes de observações, e seus vetores de soma:

$$U = [u^1, u^2, \ldots] \ , \ U^{1:n} = [u^1, u^2, \ldots u^n] \ , \ x^n = U^{1:n}\mathbf{1}$$

O acento til (tilde) indica alguma forma de normalização, por exemplo, $\widetilde{x} = (1/1'x)x$.

Lema: Se $u^1, \ldots u^n$ são vetores aleatórios i.i.d.,

$$x = U^{1:n}\mathbf{1} \Rightarrow \mathrm{E}(x) = n\,\mathrm{E}(u^1) \text{ e } \mathrm{Cov}(x) = n\,\mathrm{Cov}(u^1)$$

O primeiro resultado é trivial. Para o segunto resultado, basta lembrar as propriedades de transformação da esperança e da covariancia por operadores lineares,

$$E(AY) = AE(Y) \quad , \quad \mathrm{Cov}(AY) = A\,\mathrm{Cov}(Y)A'$$

e escrever

$$\begin{aligned}
\mathrm{Cov}(x) = & \ \mathrm{Cov}(U^{1:n}\mathbf{1}) \\
= & \ \mathrm{Cov}\left((\mathbf{1}' \otimes I)\,\mathrm{Vec}(U^{1:n})\right) \ = \ (\mathbf{1}' \otimes I)\left(I \otimes \mathrm{Cov}(u^1)\right)(\mathbf{1} \otimes I) \\
= & \ \left(\mathbf{1}' \otimes \mathrm{Cov}(u^1)\right)(\mathbf{1} \otimes I) \ = \ n\,\mathrm{Cov}(u^1)
\end{aligned}$$

3.3 O Processo de Bernoulli

Consideremos uma seqüência de vetores aleatórios, $u^1, u^2, ...$ onde, $\forall u^i$ pode assumir apenas dois valores

$$I^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ ou } I^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ onde } I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

representando sucesso ou fracasso. Isto é, u^i pode assumir o valor de uma coluna qualquer da matriz identidade, I. Dizemos que u^i é da classe k, $c(u^i) = k$, sse $u^i = I^k$, $k \in [1,2]$.

Suponha também que, (na sua opinião), essa seqüência seja permutável, ou seja, se $\sigma([1,2,\ldots n])$ é uma permutação de $[1,2,\ldots n]$, então, $\forall n, \sigma$,

$$\Pr\left(u^1, ...u^n\right) = \Pr\left(u^{\sigma(1)}, ...u^{\sigma(n)}\right)$$

Com apenas essa restrição de permutabilidade, que corresponde a dizer que os rótulos são não informativos, o Teorema de De Finetti estabelece que existe um vetor desconhecido

$$\theta \in \Theta = \{\mathbf{0} \leq \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \leq \mathbf{1} \mid \ \mathbf{1}'\theta = 1\}$$

tal que, condicionalmente a θ, $u^1, u^2, \ldots$ são mutuamente independentes, e a Probabilidade condicional de $\Pr(u^i = I^k \,|\, \theta)$ é θ_k, i.e.

$$(u^1 \amalg u^2 \amalg \ldots) \,|\, \theta \text{ ou } \coprod_{i=1}^{\infty} u_i \,|\, \theta \text{ , e } \Pr(u^i = I^k \,|\, \theta) = \theta_k \text{ .}$$

Condicionalmente a θ , a seqüência $u^1, u^2, \ldots$ recebe o nome de processo de Bernoulli e muitas das distribuições discretas conhecidas são obtidas de transformações desse processo.

A Esperança e covariancia (condicional a θ de um vetor qualguer da seqüência são:

- $\mathrm{E}(u^i) = \theta$
- $\mathrm{Cov}(u^i) = \mathrm{E}\,(u^i \otimes (u^i)') - \mathrm{E}\,(u^i) \otimes \mathrm{E}\,((u^i)') = \mathrm{diag}(\theta) - \theta \otimes \theta'$

Quando o domínio da soma $1:n$, estiver subentendido, relaxamos a notação usando x no lugar de x^n. Definimos ainda o operador "produtória da potencia pontual" entre dois vetores de mesma dimensão: Dados θ, e x, $n \times 1$

$$\theta \triangle x \equiv \prod_{i=1}^{n} (\theta_i)^{x_i}$$

Uma Regra de Parada, δ, estabelece, para cada $n = 1, 2, \ldots$, a decisão de observar ou não u^{n+1}, após as observações $u^1, \ldots u^n$.

Para o correto entendimento do material da seqüência, é preciso ter claro a interpretação de expressões condicionais como $x^n \,|\, n$ ou $x_2^n \,|\, x_1^n$. Em ambos os casos estamos nos referindo a um vetor desconhecido x^n, mas temos diferentes imformações parciais. No primeiro caso conhecemos n, e portanto conhecemos a soma das componentes, $x_1^n + x_2^n = n$; todavia desconhecemos tanto a componente x_1^n como x_2^n. No segundo caso conhecemos apenas a primeira componente de x^n, x_1^n, e desconhecemos a segunda componente, x_2^n, obviamente desconhecemos também a soma, $n = x_1^n + x_2^n$. Simplesmente preste atenção: Listamos o que conhecemos à direita da barra e, (salvo termos alguma informação adicional) tudo o que não puder ser deduzido desta lista é desconhecido.

A primeira distribuição que discutiremos é a Binomial. Seja $\delta(n)$ a regra que estabelece parar após n observações, onde n é um número pré-fixado.

A probabilidade (condicional) da seqüência de observações $U^{1:n}$ é

$$\Pr(U^{1:n} \,|\, \theta) = \theta \triangle x^n$$

O vetor de soma, x^n tem distribuição Binomial com parâmetros n e θ e escreve-se $x^n \,|\, [n, \theta] \sim \mathrm{Bi}(n, \theta)$. Quando n (ou $\delta(n)$) estiver implícito no contexto, escreveremos $x \,|\, \theta$

no lugar de $x^n \,|\, [n, \theta]$. A distribuição Binomial tem a seguinte expressão:

$$\Pr(x^n \,|\, n, \theta) = \left(\begin{array}{c} n \\ x^n \end{array} \right) (\theta \triangle x^n)$$

onde

$$\left(\begin{array}{c} n \\ x \end{array} \right) \equiv \frac{\Gamma(n+1)}{\Gamma(x_1+1)\,\Gamma(x_2+1)} = \frac{n!}{x_1!\,x_2!} \text{ e } n = \mathbf{1}'x \ .$$

Um bom exercício para o leitor é verificar que o vetor esperança e a matriz de covariância de $x^n \,|\, [n, \theta]$ possuem as seguintes expressões:

$$\mathrm{E}(x^n) = n\theta \ \text{ e } \ \mathrm{Cov}(x^n) = n\,(\theta \triangle \mathbf{1}) \left[\begin{array}{cc} 1 & -1 \\ -1 & 1 \end{array} \right]$$

A segunda distribuição que discutiremos é a Binomial Negativa. Seja $\delta(x_1^n)$ a regra que estabelece parar na observação de u^n ao obtermos um número pré-fixado de x_1^n sucessos.

A variável aleatória x_2^n, o número de fracassos ao obtermos todos os x_1^n sucessos requeridos, é denominada Binomial Negativa com parâmetros x_1^n e θ. Não é difícil provar que a distribuição Binomial Negativa, $x_2^n \,|\, [x_1^n, \theta] \sim \mathrm{Bn}(x_1^n, \theta)$, tem a expressão, $\forall\, x_2^n \in I\!N$,

$$\Pr(x^n \,|\, x_1^n, \theta) \ = \ \frac{x_1^n}{n} \left(\begin{array}{c} n \\ x^n \end{array} \right) (\theta \triangle x^n) \ = \ \theta_1 \Pr\left((x^n - I^1) \,|\, (n-1), \theta) \right) \ .$$

Note que, da forma como definimos a distribuição, x_1^n é um inteiro positivo. Contudo, essa restrição pode ser relaxada pois, para qualquer valor real positivo a, a função acima continua sendo uma função de probabilidade. Para tanto, usamos

$$\sum_{j=0}^{\infty} \frac{\Gamma(a+j)}{\Gamma(a) j!} (1-\pi)^j = \pi^{-a} \ , \ \forall\, a \in [0, \infty[\text{ e } \pi \in]0, 1[$$

Além dessa igualdade, um bom exercício para o leitor seria verificar que a média e a variância de x_2^n possuem as seguintes expressões:

$$\mathrm{E}\,(x_2^n \,|\, x_1^n, \theta) = \frac{x_1^n \theta_2}{\theta_1} \ \text{ e } \ \mathrm{Var}\,(x_2^n \,|\, x_1^n, \theta) = \frac{x_1^n \theta_2}{(\theta_1)^2} \ .$$

No caso particular de $\delta(x_1^n = 1)$, a distribuição Binomial Negativa toma o nome de distribuição Geométrica com parâmetro θ. Se a variáveis aleatórias são independentes e identicamente distribuidas (i.i.d.) segundo uma geométrica de parâmetro θ, então, a soma destas variáveis tem distribuição Binomial Negativa com parâmetros a e θ.

A terceira distribuição abordada é a Hipergeométrica. Voltando à seqüência inicial, $u^1, u^2, ...$, suponha que um primeiro observador conhece as N primeiras observações, enquanto um segundo observador conhece apenas uma subseqüência de $n < N$ destas observações. Como a seqüência $u^1, u^2, \ldots$ é permutável, pode-se considerar, sem perda de

generalidade, a subseqüência das n primeiras observações, $u^1, \ldots u^n$. Usando-se o teorema de De Finetti, tem-se que x^n e $x^N - x^n = U^{n+1:N}\mathbf{1}$ são condicionalmente independentes, dado θ. Isto é, $x^n \amalg (x^N - x^n) \,|\, \theta$. Além disso, podemos escrever

$$x^n \,|\, [n, \theta] \sim \text{Bi}(n, \theta) \ , \ x^N \,|\, [N, \theta] \sim \text{Bi}(N, \theta) \text{ e } (x^N - x^n) \,|\, [(N-n), \theta] \sim \text{Bi}(N-n, \theta).$$

Nosso objetivo é encontrar a função de probabilidade de $x^n \,|\, x^N$. Note que x^N é suficiente para $U^{1:N}$ dado θ, e x^n é suficiente para $U^{1:n}$. Além disto $x^n \,|\, [n, x^N]$ tem a mesma distribuição de $x^n \,|\, [n, x^N, \theta]$. Usando-se as regras do cálculo de probabilidades e as propriedades acima, temos que:

$$\begin{aligned}
&\Pr(x^n \,|\, n, x^N, \theta) \\
&= \frac{\Pr(x^n, x^N \,|\, n, N, \theta)}{\Pr(x^N \,|\, n, N, \theta)} = \frac{\Pr(x^n, (x^N - x^n) \,|\, n, N, \theta)}{\Pr(x^N \,|\, n, N, \theta)} \\
&= \frac{\Pr(x^n \,|\, n, N, \theta)\, \Pr(x^N - x^n \,|\, n, N, \theta)}{\Pr(x^N \,|\, n, N, \theta)}.
\end{aligned}$$

Assim, $x^n \,|\, [n, x^N]$ tem função de probabilidade

$$\Pr(x^n \,|\, n, x^N) = \frac{\left(\begin{array}{c} n \\ x^n \end{array} \right) \left(\begin{array}{c} N-n \\ x^N - x^n \end{array} \right)}{\left(\begin{array}{c} N \\ x^N \end{array} \right)}$$

$$\text{onde } \mathbf{0} \leq x^n \leq x^N \leq N\mathbf{1} \ , \ \mathbf{1}'x^n = n \ , \ \mathbf{1}'x^N = N$$

Esta é a representação vetorial da função de probabilidade Hipergeométrica.

$$x^n \,|\, [n, x^N] \sim \text{Hi}(n, N, x^N) \ .$$

Um bom exercício para o leitor seria verificar as seguintes expressões para a esperança e a covariância (condicional) de $x^n \,|\, [n, N, x^N]$ e covariância entre u^i e u^j, $i, j \leq n$:

$$\text{E}(x^n) = \frac{n}{N}\, x^N \quad \text{e} \quad \text{Cov}(x^n) = \frac{n(N-n)}{(N-1)}\, (x^N \triangle \mathbf{1}) \left[\begin{array}{cc} 1 & -1 \\ -1 & 1 \end{array} \right]$$

$$\text{Cov}(u^i, u^j \,|\, x^N) = \frac{1}{(N-1)N^2}\, (x^N \triangle \mathbf{1}) \left[\begin{array}{cc} -1 & 1 \\ 1 & -1 \end{array} \right]$$

A generalização para $k > 2$ categorias é natural, bastando para isso que usemos a base ortonormal $I^1, I^2, \ldots I^k$, i.e. as colunas da matriz identidade de $\mathcal{R}^k$. As distribuições multinomial e hipergeomtrica multivariada, discutidas na seqüência, são distribuições derivadas de tais generalizações.

Para complementar esta seção, apresentaremos a derivação da distribuição Beta-Binomial. Vamos supor que um primeiro observador observou um número de fracassos x_2^n até observar-se um número pré fixado de x_1^n sucessos. Um segundo observador fez mais observações observando x_2^N fracassos até completar o número pré-fixado de x_1^N sucessos, $x_1^n < x_1^N$.

Como x_1^n e x_1^N são pré-fixados, podemos escrever

$$x_2^N \,|\, \theta \sim \text{Bn}(x_1^N, \theta) \ , \quad x_2^n \,|\, \theta \sim \text{Bn}(x_1^n, \theta)$$

$$(x_2^N - x_2^n) \,|\, \theta \sim \text{Bn}(x_1^N - x_1^n, \theta) \ \text{ e } \ x_2^n \amalg (x_2^N - x_2^n) \,|\, \theta.$$

Como antes, o nosso objetivo é descrever a distribuição de $x_2^n \,|\, [x_1^n, x^N]$. Assim que o leitor perceber que $[x_1^n, x^N]$ é suficiente para $[x^n, (x^N - x^n)]$, com respeito a θ, o problema torna-se semelhante ao caso da hipergeométrica e obtemos, analogamente,

$$\Pr(x_2^n \,|\, x_1^n, x^N) = \frac{x_2^N!\,\Gamma(b)}{\Gamma(x_2^N + x_1^N)} \, \frac{\Gamma(x_2^n + x_1^n)}{x_2^n!\,\Gamma(x_1^n)} \, \frac{\Gamma(x_2^N - x_2^n + x_1^N - x_1^n)}{(x_2^N - x_2^n)!\,\Gamma(x_1^N - x_1^n)} \ , \quad x_2^n \in \{0, 1, ..., x_2^N\}.$$

Esta é a função de probabilidade de uma variável aleatória denominada Beta Binomial com parámetros x_1^n e x^N.

$$x_2^n \,|\, (x_1^n, x^N) \sim \text{BB}(x_1^n, x^N) \ .$$

As propriedades dessa distribuição serão estudadas no caso geral da Dirichlet-Multinomial, que aparece na seqüência.

3.4 A Distribuição Multinomial

Sejam u^i, $i = 1, 2, \ldots$ vetores aleatórios cujos possíveis resultados são as colunas da matriz identidade de ordem m, $I^k, k \in 1\!:\!m$. Dizemos que u^i é da classe k, $c(u^i) = k$, sse $u^i = I^k$.

Seja $\theta \in [0,1]^m$ o vetor de probabilidades de uma observação da classe k no processo de Bernoulli m-variado, i.e.,

$$\Pr(u^i = I^k \,|\, \theta) = \theta_k \ , \ \mathbf{0} \leq \theta \leq \mathbf{1} \ , \ \mathbf{1}'\theta = 1$$

Como na secção anterior, sejam U

$$U = [u^1, u^2, \ldots] \text{ e } x^n = U^{1:n}\mathbf{1}$$

Definição 2.1: Se o conhecimento de θ torna os vetores u^i independentes, então a distribuição (condicional) de x^n dado θ chama-se Distribuição Multinomial de ordem m com parâmetros n e θ, que é dada por

$$\Pr(x^n \,|\, n, \theta) = \binom{n}{x^n} (\theta \triangle x^n)$$

onde

$$\left(\begin{array}{c} n \\ x \end{array} \right) \equiv \frac{\Gamma(n+1)}{\Gamma(x_1+1)\ldots\Gamma(x_m+1)} = \frac{n!}{x_1!\ldots x_m!} \text{ e } n = \mathbf{1}'x \ .$$

Para representar a distribuição multinomial escreveremos

$$x^n \,|\, [n, \theta] \sim \mathrm{Mn}_m(n, \theta) \ .$$

Quando $m = 2$, temos o caso binomial.

A seguir, verificamos algumas propriedades da distribuição multinomial.

Lema: Se $x \,|\, \theta \sim \mathrm{Mn}_m(n, \theta)$ então a esperança e covariancia (condicional) de x são

$$\mathrm{E}(x) = n\theta \text{ e } \mathrm{Cov}(x) = n(\mathrm{diag}(\theta) - \theta \otimes \theta')$$

Demonstração: Análoga ao caso binomial.

O resultado seguinte apresenta uma caracterização da distribuição multinomial em termos da distribuição Poisson.

Lema: Propriedade Reprodutiva da Poisson.

$$x_i \sim \mathrm{Ps}(\lambda_i) \Rightarrow \ \mathbf{1}'x \,|\, \lambda \ \sim \mathrm{Ps}(\mathbf{1}'\lambda)$$

isto é, a soma de variáveis (independentes) com distribuição de Poisson é Poisson.

Theorema: Characterização da Multinomial pela Poisson.
Seja $x = [x_1, ..., x_m]'$ o vetor cujas componentes são independentes com distribuições de Poisson com parâmetros dados no vetor $\lambda = [\lambda_1, ..., \lambda_m]' > 0$, λ conhecido. Seja n um inteiro positivo. Então, dado λ,

$$x \,|\, [n = \mathbf{1}'x, \lambda] \quad \sim \mathrm{Mn}_m(n, \theta) \text{ onde } \theta = \frac{1}{\mathbf{1}'\lambda}\lambda$$

Demonstração: A distribuição conjunta de x, dado λ é

$$\Pr(x \,|\, \lambda) = \prod_{k=1}^{m} \frac{e^{-\lambda_k}\lambda_i^{x_k}}{x_k!} \ .$$

Usando a propriedade reprodutiva da Poisson,

$$\begin{aligned} &\Pr\left(x \,|\, \mathbf{1}'x = n, \lambda\right) \\ &= \ \frac{\Pr\left(\mathbf{1}'x = n \wedge x \,|\, \lambda\right)}{\Pr\left(\mathbf{1}'x = n \,|\, \lambda\right)} \ = \ \delta(n = \mathbf{1}'x)\frac{\Pr(x \,|\, \lambda)}{\Pr(\mathbf{1}'x = n \,|\, \lambda)} \end{aligned}$$

Os resultado seguintes enunciam propriedades importantes da distribuição multinomial. A demonstração destas propriedades é simples, usando a characterização da Multinomial pela Poisson, e a propriedade reprodutiva da Poisson.

Theorema: Partição de Classes da Multinomial
Seja $1:m$ o domínio dos índices que denotam as classes de uma Multinomial de ordem m. Seja T uma matriz de partição das m-classes em s-super-classes. Se $x \sim \mathrm{Mn}_m(n, \theta)$ e T é uma matriz de partição, então $y = Tx \sim \mathrm{Mn}_s(n, T\theta)$.

Teorema: Condicionamento na Soma Parcial.
Se $x \sim \mathrm{Mn}_m(n, \theta)$, então a distribuição de parte do vetor x condicionada a sua soma tem distribuição Multinomial, tendo como parâmetro a correspondente parte do parâmetro (normalizado). Para maior clareza, condicionando na soma das t primeiras componentes, temos:

$$x_{1:t} \,|\, (\mathbf{1}' x_{1:t} = j) \sim \mathrm{Mn}_t \left(j, \frac{1}{\mathbf{1}'\theta_{1:t}} \theta_{1:t} \right) \text{ onde } 0 \leq j \leq n$$

Teorema: Decomposição Multinomial–Binomial.
Usando os dois últimos teoremas, se $x \sim \mathrm{Mn}_m(n, \theta)$,

$$\begin{aligned} &\Pr(x \,|\, n, \theta) = \\ &= \sum_{j=0}^{n} \Pr\left(x_{1:t} \,|\, j, \frac{1}{\mathbf{1}'\theta_{1:t}} \theta_{1:t} \right) \\ &\qquad \Pr\left(x_{t+1:m} \,|\, (n-j), \frac{1}{\mathbf{1}'\theta_{t+1:m}} \theta_{t+1:m} \right) \\ &\qquad \Pr\left(\begin{bmatrix} j \\ (n-j) \end{bmatrix} \,|\, n, \begin{bmatrix} \mathbf{1}'\theta_{1:t} \\ \mathbf{1}'\theta_{t+1:m} \end{bmatrix} \right) \end{aligned}$$

Analogamente, poderiamos escrever a decomposição Multinomial-Trinomial para uma partição dos indices de classe em três super-classes. Poderiamos ainda escrever a decomposição m-nomial-s-nomial para uma partição dos m indices de classe em um número arbitrário de s super-classes.

3.5 Distribuição Hipergeométrica Multivariada

Na primeira seção, mostramos como a distribuição hipergeométrica pode ser gerada de um processo permutável de variáveis de Bernoulli. A generalização natural desse fato aparece ao considerarmos um processo permutável onde seus elementos são vetores aleatórios que, como na distribuição multinomial, tomam valores em I^k. Dizemos que u^i é da classe k, $c(u^i) = k$, sse $u^i = I^k$.

Uma amostra de tamanho n é tomada de uma população finita de tamanho $N(> n)$, a qual é particionada em m classes. As frequências populacionais (número de elementos

em cada categoria) são representadas por $[\psi_1, \ldots \psi_m]$, e portanto $N = \mathbf{1}'\psi$. Baseado na amostra, deseja-se inferir sobre ψ. x_k é a frequência amostral da classe k.

Uma outra forma de descrever esse problema é considerar uma caixa com N bolas de m diferentes cores que são denominadas $1, \ldots m$. O número de bolas com a k-ésima cor é denotado por ψ_k. Suponha que as N bolas são separadas em duas outras caixas de tal forma que uma caixa (caixa número 1) contém n bolas e a outra (caixa número 2) $(N-n)$ bolas. O estatístico pode observar a composição da caixa 1, representada pelo vetor x de freqüências amostrais. Agora, a quantidade de interesse para o estatístico é o vetor $\psi - x$ que representa a composição da caixa 2.

Como no caso Multinomial seleção assumimos que $U^{1:N}$ é uma secção finita de um processo permutável e, dessa forma, qualquer n-sub-seqüência extraída de $U^{1:N}$ tem a mesma distribuição de $U^{1:n}$.

Este fato implica que $x = U^{1:n}\mathbf{1}$ tem a mesma distribuição do vetor de frequências de uma amostra casual de tamanho n.

Como no caso bivariado, o nosso objetivo é encontrar a distribuição de $x \,|\, \psi$. Novamente, usando-se o teorema de De Finetti, existe um vetor $\mathbf{0} \leq \theta \leq \mathbf{1}$, $\mathbf{1}'\theta = 1$, tal que $\coprod_{j=0}^{N} u^j \,|\, \theta$ e $\Pr\left(c(u^j) = k\right) = \theta_k$. Analogamente ao caso mutinomial, esses fatos implicam nos seguintes resultados:

- $\psi \,|\, \theta \sim \mathrm{Mn}_m(N, \theta)$
- $x \,|\, \theta \sim \mathrm{Mn}_m(n, \theta)$
- $(\psi - x) \,|\, \theta \sim \mathrm{Mn}_m((N-n), \theta)$
- $(\psi - x) \amalg x \,|\, \theta$

Usando-se os resultados da seção anterior e seguindo as mesmas etapas do caso Hi_2 na primeira secção, obtemos a seguinte expressão para a função de probabilidade Hipergeométrica m-variada, $x^n \,|\, [n, N, \psi] \sim \mathrm{Hi}_m(n, N, \psi)$:

$$\Pr(x^n \,|\, n, \psi) = \frac{\begin{pmatrix} n \\ x^n \end{pmatrix} \begin{pmatrix} N-n \\ \psi - x^n \end{pmatrix}}{\begin{pmatrix} N \\ \psi \end{pmatrix}}$$

$$\text{onde } \mathbf{0} \leq x^n \leq \psi \leq N\mathbf{1} \ , \ \mathbf{1}'x^n = n \ , \ \mathbf{1}'\psi = N$$

Esta é a representação vetorial da função de probabilidade Hipergeométrica.

$$x^n \,|\, [n, x^N] \sim \mathrm{Hi}(n, N, x^N) \ .$$

Alternativamente, podemos escrever a fórmula usual

$$\Pr(x \mid \psi) = \frac{\binom{\psi_1}{x_1}\binom{\psi_2}{x_2}\cdots\binom{\psi_m}{x_m}}{\binom{N}{n}}$$

Teorema: A esperança e covariância de um vetor aleatório com distribuição Hipergeométrica, $x \sim \mathrm{Hi}_m(n, N, \psi)$, são:

$$\mathrm{E}(x) = n\widetilde{\psi}\ ,\ \ \mathrm{Cov}(x) = n\frac{N-n}{N-1}\left(\mathrm{diag}(\widetilde{\psi}) - \widetilde{\psi}\otimes\widetilde{\psi}'\right)\ \text{onde}\ \widetilde{\psi} = \frac{1}{N}\psi\ .$$

Demonstração: Use que

$$\begin{aligned}
\mathrm{Cov}(x^n) &= n\,\mathrm{Cov}(u^1) + n(n-1)\,\mathrm{Cov}(u^1, u^2)\\
\mathrm{Cov}(u^1) &= \mathrm{E}\left(u^1\otimes(u^1)'\right) - \mathrm{E}(u^1)\otimes\mathrm{E}(u^1)' = \mathrm{diag}(\widetilde{\psi}) - \widetilde{\psi}\otimes\widetilde{\psi}'\\
\mathrm{Cov}(u^1, u^2) &= \mathrm{E}\left(u^1\otimes(u^2)'\right) - \mathrm{E}(u^1)\otimes\mathrm{E}(u^2)'
\end{aligned}$$

O segundo termo da duas últimas equações são iguais, e o primeiro termo da última equação é

$$\mathrm{E}\left(u_i^1 u_j^2\right) = \begin{cases} \frac{\psi_i}{N}\frac{\psi_i - 1}{N-1} \text{ se } i = j \\ \frac{\psi_i}{N}\frac{\psi_j}{N-1} \text{ se } i \neq j \end{cases}$$

Obtemos agora o resultado por manipulação algébrica.

Note que, como no caso de ordem 2, os elementos diagonais de $\mathrm{Cov}(u^1)$ são positivos, enquanto os elementos diagonais de $\mathrm{Cov}(u^1, u^2)$ são negativos. Nos elementos fora da diagonal os sinais se invertem.

3.6 A Distribuição Dirichlet

Na segunda seção apresentamos a distribuição Multinominal $\mathrm{Mn}_m(n, \theta)$. Nesta Seção apresentaremos a distribuição de Dirichlet para θ. Recordemos primeiramente as distribuições univariadas de Poisson e Gama.

Uma variável aleatória tem distribuição Gama, $x \mid [a, b] \sim G(a, b), a, b > 0$, se a distribuição é contínua com densidade

$$f(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1}\exp(-bx)\ ,\ x > 0$$

A esperança e a variância desta variável são

$$E(x) = \frac{a}{b} \text{ e } \mathrm{Var}(x) = \frac{a}{b^2}$$

Lema: Propriedade Reprodutiva da Gama.
Se n variáveis independentes $x_i \,|\, a_i, b \sim G(a_i, b)$, então

$$\mathbf{1}'x \sim G(\mathbf{1}'a, b)$$

Lema: A distribuição Gama é conjugada à distribuição de poisson.

Demonstração:
Se $y \,|\, \lambda \sim \text{Ps}(\lambda)$ e λ tem priori $\lambda \,|\, a, b \sim G(a, b)$, então

$$\begin{aligned} f(\lambda \,|\, y, a, b) &\propto L(\lambda \,|\, y) f(\lambda) \\ &= \exp(-\lambda)\frac{\lambda^y}{y!}\,\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda) \;\propto\; \lambda^{y+a-1} exp(-(b+1)\lambda) \end{aligned}$$

Isto é, a distribuição a posteriori de λ é gama com parâmetros $[a + y, b + 1]$.

Definição: Distribuição de Dirichlet.
Um vetor aleatório

$$y \in \mathcal{S}_{m-1} \equiv \{y \in \mathcal{R}^m \,|\, \mathbf{0} \leq y \leq \mathbf{1} \wedge \mathbf{1}'y = 1\}$$

tem distribuição de Dirichlet de ordem m com parâmetro positivo $a \in \mathcal{R}^m$ se sua densidade for

$$\Pr(y \,|\, a) = \frac{y \triangle (a - \mathbf{1})}{B(a)}$$

Note que $\mathcal{S}_{m-1}$, o Simplex de dimensão $m-1$, é uma região de $\mathcal{R}^m$ sujeita a 1 "vínculo", $\mathbf{1}'y = 1$, tendo portando $m - 1$ "graus de liberdade". Neste sentido, dizemos que a distribuição de Dirichlet tem uma "representação singular". É possivel dar uma representação não singular a distribuição de $[y_1, \ldots y_{m-1}]'$, conhecida como distribuição Beta Multivariada, mas ao custo de complicar bastante a álgebra envolvida, e perder muito da interpretação geométrica da representação singular.

O fator de normalização é simplesmente

$$B(a) \equiv \int_{y \in \mathcal{S}_{m-1}} (y \triangle (a - 1))\, dy$$

Lema: Função Beta.
O fator de Normalização da distribuição de Dirichlet acima é a Função Beta, definida como

$$B(a) = \frac{\prod_{k=1}^m \Gamma(a_k)}{\Gamma(\mathbf{1}'a)}$$

Theorema: Dirichlet Conjugada da Multinomial:
Se $\theta \sim \mathrm{Di}_m(a)$ e $x \,|\, \theta \sim \mathrm{Mn}_m(n, \theta)$ então

$$\theta \,|\, x \sim \mathrm{Di}_m(a + x) \ .$$

Demonstração:

Basta lembrar que a verossimilhança da Multinomial é proporcional a $\theta \triangle x$, e que a priori Dirichlet de θ é proporcional a $\theta \triangle (a - \mathbf{1})$. Portanto a posteriori é proporcional a $\theta \triangle (x + a - 1)$. Por outro lado, $B(a+x)$ é o fator de normalização, i.e., é igual a integral sob θ de $\theta \triangle (x + a - 1)$, e assim temos uma função de densidade Dirichlet como acima definida.

Lema: Momentos da Dirichlet.
Se $\theta \sim \mathrm{Di}_m(a)$ e $p \in I\!N^m$, então

$$\mathrm{E}\left(\theta \triangle p\right) \ = \ \frac{B(a+p)}{B(a)}$$

Demonstração:

$$\begin{aligned}
&\int_\Theta (\theta \triangle p) f(\theta \,|\, a) d\theta \\
&= \ \frac{1}{B(a)} \int_\Theta \left(\theta \triangle p\right) \left(\theta \triangle (a-1)\right) d\theta \ = \ \frac{1}{B(a)} \int_\Theta \left(\theta \triangle (a+p-1)\right) d\theta \ = \ \frac{B(a+p)}{B(a)}
\end{aligned}$$

Escolhendo aproprianamente os expoentes, p, temos

Corolário: Se $\theta \sim \mathrm{Di}_m(a)$, então

$$\begin{aligned}
\mathrm{E}(\theta) \ &= \ \widetilde{a} \equiv \frac{1}{\mathbf{1}'a} a \\
\mathrm{Cov}(\theta) \ &= \ \frac{1}{\mathbf{1}'a + 1} \left(\mathrm{diag}(\widetilde{a}) - \widetilde{a} \otimes \widetilde{a}'\right)
\end{aligned}$$

Theorema: Caracterização da Dirichlet pela Gama.
Sejam as componentes do vetor aleatório $x \in \mathcal{R}^m$ independentes e com distribuição $G(a_k, b)$. Então o vetor normalizado

$$y = \frac{1}{\mathbf{1}'x} x \sim \mathrm{Di}_m(a) \ , \ \mathbf{1}'x \sim \mathrm{Ga}(\mathbf{1}'a) \text{ e } y \amalg \mathbf{1}'x$$

Demonstração:

Considere a normalização,

$$y \;=\; \frac{1}{t}x \;,\; t \;=\; \mathbf{1}'x \;,\; x \;=\; ty$$

como uma transformação de variáveis. Note que uma das novas variáveis, digamos $y_m \equiv t(1 - y_1 \ldots - y_{m-1})$, torna-se redundante.

A matriz Jacobiana da transformação

$$J \;=\; \frac{\partial\,(x_1, x_2, \ldots x_{m-1}, x_m)}{\partial\,(y_1, y_2, \ldots y_{m-1}, t)} \;=\; \begin{bmatrix} t & 0 & \cdots & 0 & y_1 \\ 0 & t & \cdots & 0 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & t & y_{m-1} \\ -t & -t & \cdots & -t & 1 - y_1 \cdots - y_{m-1} \end{bmatrix}$$

Realizando operações elementares que somam todas (exceto a última) as linhas à última linha, obtemos a fatoração da matriz Jacobiana $J = LU$, onde

$$L = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ -1 & -1 & \cdots & -1 & 1 \end{bmatrix} \;\text{ e }\; U = \begin{bmatrix} t & 0 & \cdots & 0 & y_1 \\ 0 & t & \cdots & 0 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & t & y_{m-1} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

As matrizes triangulares tem determinante igual a produtória dos elementos na diagonal principal, portanto $|J| = |L|\,|U| = 1\,t^{m-1}$.

Por outro lado, a distribuição conjunta de x é

$$f(x) \;=\; \prod_{k=1}^{m} \mathrm{Ga}(x_k \,|\, a_k, b) \;=\; \prod_{k=1}^{m} \frac{b^{a_k}}{\Gamma(a_k)} e^{-bx_k} (x_k)^{a_k - 1}$$

e a distribuição conjunta no novo sistema de coordenadas é

$$\begin{aligned} g([y,t]) &= |J|\, f\left(x^{-1}([y,t])\right) \\ &= t^{m-1} \prod_{k=1}^{m} \frac{b^{a_k}}{\Gamma(a_k)} e^{-bx_k} (x_k)^{a_k - 1} \;=\; t^{m-1} \prod_{k=1}^{m} \frac{b^{a_k}}{\Gamma(a_k)} e^{-bty_k} (ty_k)^{a_k - 1} \\ &= \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k - 1}}{\Gamma(a_k)} \right) b^{1'a} e^{-bt} t^{1'a - m} t^{m-1} \;=\; \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k - 1}}{\Gamma(a_k)} \right) b^{1'a} e^{-bt} t^{1'a - 1} \end{aligned}$$

Logo, a distribuição marginal de $y = [y_1, \ldots y_k]'$ é

$$\begin{aligned} g(y) &= \int_{t=0}^{\infty} g([y,t])dt \\ &= \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k-1}}{\Gamma(a_k)}\right) \int_{t=0}^{\infty} b^{1'a} e^{-bt} t^{1'a-1} dt \\ &= \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k-1}}{\Gamma(a_k)}\right) \Gamma(\mathbf{1}'a) \;=\; \frac{y \triangle (a-1)}{B(a)} \end{aligned}$$

Na penúltima passagem, substituimos a integral pelo fator de normalização de uma densidade Gama, $\mathrm{Ga}(\mathbf{1}'a, b)$. Assim, obtemos uma densidade proporcional a $y \triangle (a-1)$, i.e., uma Dirichlet, Q.E.D.

Na última passagem obtemos também o fator de normalização da Dirichlet, demonstrando o lema da função Beta.

Lema: Bipartição dos indices da Dirichlet.
Seja $1:t$, $t+1:m$ uma bipartição do domínio dos índices das classes de uma Dirichlet de ordem m, $1:m$, em duas super-classes. Seja $y \sim \mathrm{Di}_m(a)$, e

$$z^1 = \frac{1}{\mathbf{1}'y_{1:t}} y_{1:t} \;,\; z^2 = \frac{1}{\mathbf{1}'y_{t+1:m}} y_{t+1:m} \;,\; w = \begin{bmatrix} \mathbf{1}'y_{1:t} \\ \mathbf{1}'y_{t+1:m} \end{bmatrix}$$

Temos então que, $z^1 \amalg z^2 \amalg w$ e

$$z^1 \sim \mathrm{Di}_t(a_{1:t}) \;,\; z^2 \sim \mathrm{Di}_{m-t}(a_{t+1:m}) \text{ e } w \sim \mathrm{Di}_2\left(\begin{bmatrix} \mathbf{1}'a_{1:t} \\ \mathbf{1}'a_{t+1:m} \end{bmatrix}\right)$$

Demonstração:
Pelo Teorema da Caracterização da Dirichlet pela Gama podemos imaginar que o vetor y é originado pela normalização de um vetor x de m variáveis

$$y = \frac{1}{\mathbf{1}'x} x \;,\; x_k \sim \mathrm{Ga}(a_k, b) \;,\; \coprod_{k=1}^{m} x_k$$

Considerando apenas cada super-classe isoladamente, construimos os vetores z^1 e z^2 que são distribuidos como

$$\begin{aligned} z^1 &= \frac{1}{\mathbf{1}'y_{1:t}} y_{1:t} \;=\; \frac{1}{\mathbf{1}'x_{1:t}} x_{1:t} \;\sim\; \mathrm{Di}_t(a_{1:t}) \\ z^2 &= \frac{1}{\mathbf{1}'y_{t+1:m}} y_{t+1:m} \;=\; \frac{1}{\mathbf{1}'x_{t+1:m}} x_{t+1:m} \;\sim\; \mathrm{Di}_{m-t}(a_{t+1:m}) \end{aligned}$$

$z^1 \amalg z^2$, que são por sua vez independentes das somas parciais

$$\mathbf{1}'x_{1:t} \sim \text{Ga}(\mathbf{1}'a_{1:t}, b) \text{ e } \mathbf{1}'x_{t+1:m} \sim \text{Ga}(\mathbf{1}'a_{t+1:m}, b)$$

Usando novamente o teorema da Caracterização da Dirichlet pela Gama para estas duas variáveis Gama, obtemos o resultado, Q.E.D.

Podemos generalizar este resultado para uma partição qualquer do conjunto de classes, como segue. Se $y \sim \text{Di}_m(a)$ e T é uma s-partição de suas classes, então, as distribuições intra extra super-classes são Dirichlets independentes, como segue

$$\begin{aligned} z^r &= \frac{1}{T_r y}\, {}_rPy \;\sim\; \text{Di}_{T_r 1}\,({}_rPa) \\ w &= Ty \;\sim\; \text{Di}_s(Ta) \end{aligned}$$

3.7 Dirichlet-Multinomial

Dizemos que um vetor aleatório $x \in I\!N^n \,|\, \mathbf{1}'x = n$ tem distribuição Dirichlet-Multinomial com vetor de parâmetros n e $a \in \mathcal{R}^m$, sse

$$\Pr(x \,|\, n, a) \;=\; \frac{B(a+x)}{B(a)} \left(\begin{array}{c} n \\ x \end{array} \right) \;=\; \frac{B(a+x)}{B(a)\, B(x)} \frac{1}{x \triangle \mathbf{1}}$$

Teorema: Caracterizaçã da DM por mixtura de Multinomial por Dirichlet.

$$\text{Se } \theta \sim \text{Di}_m(a) \text{ e } x \,|\, \theta \sim \text{Mn}(n, \theta) \text{ então } x \,|\, [n, a] \sim \text{DM}_m(n, a)$$

Demonstracao: A distribuição conjunta de θ, x é proporcional a $\theta \triangle (a + x - 1)$, cuja integral em θ é $B(a+x)$. Assim, multipicando pelas constantes da distribuicao conjunta, temos a marginal de x. Q.E.D. Assim, demostramos também que a função DM integra um, isto é

$$\begin{aligned} \Pr(x) \;&=\; \int_{\theta \in \mathcal{S}_{m-1}} \left(\begin{array}{c} n \\ x \end{array} \right) (\theta \triangle x)\, \frac{1}{B(a)} \theta \triangle (a - \mathbf{1})\, d\theta \\ &=\; \frac{1}{B(a)} \left(\begin{array}{c} n \\ x \end{array} \right) \int_{\theta \in \mathcal{S}_{m-1}} (\theta \triangle (x + a - \mathbf{1}))\; d\theta \;=\; \frac{B(x+a)}{B(a)} \left(\begin{array}{c} n \\ x \end{array} \right) \end{aligned}$$

Teorema: Caracterização da D-M por m Binomiais Negativas.
Seja $a \in I\!N_+^m$, e $x \in I\!N_m$, um vetor cujas componentes são variáveis aleatórias independentes, $a_k \sim \text{Bn}(a_k, \theta)$. Então

$$x \,|\, [\mathbf{1}'x = n, a] \sim \text{DM}_m(n, a)$$

Demonstracao:

$$\Pr(x \,|\, \theta, a) \;=\; \prod_{k=1}^{m} \left(\begin{array}{c} a_k + x_k - 1 \\ x_k \end{array} \right) \theta^{a_k} (1-\theta)^{x_k}$$

$$\Pr(\mathbf{1}'x \,|\, \theta, a) \;=\; \left(\begin{array}{c} \mathbf{1}'a + \mathbf{1}'x - 1 \\ \mathbf{1}'x \end{array} \right) \theta^{\mathbf{1}'a} (1-\theta)^{\mathbf{1}'a}$$

Então

$$\Pr(x \,|\, \mathbf{1}'x = n, \theta, a) \;=\; \frac{\Pr(x \,|\, a, \theta)}{\Pr(\mathbf{1}'x = n \,|\, \theta)} \;=\; \frac{\prod_{k=1}^{m} \left(\begin{array}{c} a_k + x_k - 1 \\ x_k \end{array} \right)}{\left(\begin{array}{c} \mathbf{1}'a + \mathbf{1}'x - 1 \\ \mathbf{1}'x \end{array} \right)}$$

Portanto

$$\begin{aligned} \Pr(x \,|\, \mathbf{1}'x = n, \theta, a) \;&=\; \Pr(x \,|\, \mathbf{1}'x = n, a) \\ &=\; \prod_{k=1}^{m} \frac{\Gamma(a_k + x_k)}{x! \Gamma(a_k)} \;/\; \frac{\Gamma(\mathbf{1}'a + n)}{\Gamma(\mathbf{1}'a) n!} \;=\; \frac{B(a+x)}{B(a)} \left(\begin{array}{c} n \\ x \end{array} \right) \end{aligned}$$

Teorema: A DM como Pseudo-Conjugada da Hipergeométrica

$$\text{Se } x \sim \text{Hi}_m(n, N, \psi) \text{ e } \psi \sim \text{DM}_m(N, a) \text{ então } (\psi - x) \,|\, x \sim \text{DM}_m(N - n, a)$$

Demonstração: Usando as propriedades da Hipergeométrica já apresentadas, temos a independência, $(\psi - x) \amalg x \,|\, \theta$. Podemos portanto usar a amostra Multinomial $x \,|\, \theta$ para atualizar a priori para obter a posteriori

$$\theta \,|\, x \sim \text{Di}_m(a + x)$$

Assim, a distribuição parte da população não amostrada, $\psi - x$, dada a amostra x, é a mistura de $(\psi - x)\theta$ pela posteriori de θ. Pela caracterização da DM como mistua de Multinomial por Dirichlet, segue o teorema, i.e.,

$$\left. \begin{array}{c} (\psi - x) \,|\, [\theta, x] \sim (\psi - x) \,|\, \theta \sim \text{Mn}_m(N - n, \theta) \\ \theta \,|\, x \sim \text{Di}_m(a + x) \end{array} \right\} \Rightarrow (\psi - x) \,|\, x \sim \text{Di}_m(N - n, a + x)$$

Lema: Esperança e Covariancia da DM.
Se $x \sim \text{DM}_m(n, a)$ então

$$\begin{aligned} \text{E}(x) \;&=\; n\widetilde{a} \equiv \frac{1}{\mathbf{1}'a} a \\ \text{Cov}(x) \;&=\; \frac{n(n + \mathbf{1}'a)}{\mathbf{1}'a + 1} \left(\text{diag}(\widetilde{a}) - \widetilde{a} \otimes \widetilde{a}' \right) \end{aligned}$$

Demonstração:

$$
\begin{aligned}
\mathrm{E}(x) &= \mathrm{E}_\theta\left(\mathrm{E}_x(x\,|\,\theta)\right) = \mathrm{E}_\theta(n\theta) = n\widetilde{a} \\
\mathrm{E}(x\otimes x') &= \mathrm{E}_\theta\left(\mathrm{E}_x(x\otimes x'\,|\,\theta)\right) \\
&= \mathrm{E}_\theta\left(\mathrm{E}(x\,|\,\theta)\otimes \mathrm{E}(x\,|\,\theta)' + \mathrm{Cov}(x\,|\,\theta)\right) \\
&= \mathrm{E}_\theta\left(n\left(\mathrm{diag}(\theta) - \theta\otimes\theta'\right) + n^2\theta\otimes\theta'\right) \\
&= n\,\mathrm{E}_\theta\left(\mathrm{diag}(\theta)\right) + n(n-1)\,\mathrm{E}_\theta(\theta\otimes\theta') \\
&= n\mathrm{diag}(\widetilde{a}) + n(n-1)\left(\mathrm{E}(\theta)\otimes \mathrm{E}(\theta)' + \mathrm{Cov}(\theta)\right) \\
&= n\mathrm{diag}(\widetilde{a}) + n(n-1)\left(\widetilde{a}\otimes\widetilde{a}' + \frac{1}{\mathbf{1}'a+1}\left(\mathrm{diag}(\widetilde{a}) - \widetilde{a}\otimes\widetilde{a}'\right)\right) \\
&= n\mathrm{diag}(\widetilde{a}) + n(n-1)\left(\frac{1}{\mathbf{1}'a+1}\mathrm{diag}(\widetilde{a}) + \frac{\mathbf{1}'a}{\mathbf{1}'a+1}\widetilde{a}\otimes\widetilde{a}'\right) \\
\mathrm{Cov}(x) &= \mathrm{E}(x\otimes x') - \mathrm{E}(x)\otimes \mathrm{E}(x)' = \mathrm{E}(x\otimes x') - n^2\widetilde{a}\otimes\widetilde{a}' \\
&= \left(n + \frac{n(n-1)}{\mathbf{1}'a+1}\right)\mathrm{diag}(\widetilde{a}) + \left(n(n-1)\frac{\mathbf{1}'a}{\mathbf{1}'a+1} - n^2\right)\widetilde{a}\otimes\widetilde{a}' \\
&= \frac{n(n+\mathbf{1}'a)}{\mathbf{1}'a+1}\left(\mathrm{diag}(\widetilde{a}) - \widetilde{a}\otimes\widetilde{a}'\right) \quad \text{Q.E.D.}
\end{aligned}
$$

Teorema: Bipartição das classes na DM
Seja $1:t$, $t+1:m$ uma bipartição do domínio das classes (índices) de uma DM de ordem m, $1:m$, em duas super-classes. Entã são equivalentes a condição iv e o conjunto de condições i a iii abaixo:

i: $\quad x_{1:t}\amalg x_{t+1:m}\,|\,n_1 = \mathbf{1}'x_{1:t}$

ii-1: $\quad x_{1:t}\,|\,n_1 = \mathbf{1}'x_{1;t} \sim \mathrm{DM}_t(n_1, a_{1:t})$

ii-2: $\quad x_{t+1:m}\,|\,n_2 = \mathbf{1}'x_{t+1:m} \sim \mathrm{DM}_{m-t}(n_2, a_{t+1:m})$

iii: $\quad \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \sim \mathrm{DM}_2\left(n, \begin{bmatrix} \mathbf{1}'a_{1:t} \\ \mathbf{1}'a_{t+1:m} \end{bmatrix}\right)$

iv: $\quad x \sim \mathrm{DM}_m(n, a)$

Demonstração: Basta mostrar que adistribuição conjunta pode ser fatorada desta forma. Pela caracterizaçã da DM como mistura, podemos escreve-la como mistura de de uma multinomial por uma Dirichlet. Pelos Teoremas de bipartição podemos fatorar tanto a Multinomial quanto a Dirichlet. O teorema segue imediatamente.

3.8 Dirichlet do Segundo Tipo

Considere $y \sim \mathrm{Di}_{m+1}(a)$. O vetor $z = (1/y_{m+1})y_{1:m}$ tem distribuição Dirichlet do segundo tipo.

Teorema: Caracterização pela Gama da Dirichlet do segundo tipo.
Usando a caracterização da Dirichlet pela Gama, Podemos escrever a variável Dirichlet do segundo tipo em função de $m+1$ variáveis Gama independentes:

$$z_{1:m} \sim (1/x_{m+1})x_{1:m} \quad \text{onde} \quad x_k \sim Ga(a_k, b)$$

De forma semelhante ao que fizemos com a Dirichlet do primeiro tipo, podemos escrever a função densidade a Dirichlet do segundo tipo e seus momentos como:

$$f(z \mid a) = \frac{z \triangle (a_{1:m} - 1)}{(1 + 1'z)^{1'a}\ B(a)}$$

$$E(z) = e = (1/a_{m+1})a_{1:m}$$

$$\mathrm{Cov}(z) = \frac{1}{a_{m+1} - 2}\left(\mathrm{diag}(e) + e \otimes e'\right)$$

O logaritmo de uma variável Gama é bem aproximado por uma variável Normal, veja Aitichson e Shen (1980), Biometrika 67, 261-272. Esta aproximação é a chave para varios procedimentos computacionais eficientes, e motiva o cálculo dos dois primeiros momentos da distribuição log-Dirichlet do segundo tipo. Para tanto usaremos as funções Digama, $\psi(\ )$, e Trigama, $\psi'(\ )$, definidas como:

$$\psi(a) = \frac{d}{da}\ln\Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)} \quad , \quad \psi'(a) = \frac{d}{da}\psi(a)$$

Lema: A esperança e covariancia de uma variável log-Dirichlet do segundo tipo são:

$$E(\log(z)) = \psi(a_{1:m}) - \psi(a_{m+1})\mathbf{1} \quad , \quad \mathrm{Cov}(\log(z)) = \mathrm{diag}(\psi'(a_{1:m}) + \psi'(a_{m+1})1 \otimes 1'$$

Demonstração: Consideremos uma variável Gama, $x \sim G(a, 1)$:

$$1 = \int_0^\infty f(x)dx = \int_0^\infty \frac{1}{\Gamma(a)}x^{a-1}\exp(-x)dx$$

Derivando em relação ao parâmetro a, temos

$$0 = \int_0^\infty \ln(x)x^{a-1}\frac{\exp(-x)}{\Gamma(a)}dx - \frac{\Gamma'(a)}{\Gamma^2(a)}\Gamma(a) = E(\ln(x)) - \psi(a)$$

Derivando novamente em relação ao parâmetro a,

$$\begin{aligned}\psi'(a) &= \frac{d}{da}E(\ln(x)) = \frac{d}{da}\int_0^\infty \frac{\ln(x)}{\Gamma(a)}x^{a-1}\exp(-x)dx \\ &= \int_0^\infty (\ln(x)^2 x^{a-1}\frac{\exp(-x)}{\Gamma(a)}dx - \frac{\Gamma'(a)}{\Gamma(a)}E(\ln(x)) \\ &= E(\ln(x)^2) - E(\ln(x))^2 = \mathrm{Var}(\ln(x))\end{aligned}$$

O lema segue da caracterização pela Gama da Dirichlet do segundo tipo.

3.9 Exemplos

Exemplo 2.1: Sejam A, B dois atributos, cada um deles presente ou ausente numa população. Então, cada elemento desta população pode ser classificado em exatamente uma das $2^2 = 4$ categorias

A	B	k	I^k
presente	presente	1	$[1,0,0,0]'$
presente	ausente	2	$[0,1,0,0]'$
ausente	presente	3	$[0,0,1,0]'$
ausente	ausente	4	$[0,0,0,1]'$

De acordo com a notação acima, podemos escrever $x \,|\, n, \theta \sim \mathrm{Mn}_4(n, \theta)$.

Se $\theta = [0.35,\ 0.20,\ 0.30,\ 0.15]$ e $n = 10$, então

$$\Pr(x^{10} \,|\, n, \theta) = \begin{pmatrix} 10 \\ x^{10} \end{pmatrix} (\theta \triangle x^{10})$$

Assim, para calcularmos a probabilidade de $x = [1,2,3,4]'$ dado θ, utilizamos a expressão acima e obtemos

$$\Pr\left(\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \;|\; \begin{bmatrix} 0.35 \\ 0.20 \\ 0.30 \\ 0.15 \end{bmatrix}\right) = 0.000888$$

Exemplo 2.2: Se $X \,|\, \theta \sim \mathrm{Mn}_3(10, \theta)$, como no Exemplo 2.1, então, conclui-se a partir do resultado acima que

$$\mathrm{E}(X) = (2, 3, 1.5),$$

enquanto que a matriz Σ de variâncias e covariâncias resulta em

$$\Sigma = \begin{bmatrix} 1.6 & -0.6 & -0.3 \\ -0.6 & 2.1 & -0.45 \\ -0.3 & -0.45 & 1.28 \end{bmatrix}$$

Exemplo 2.3: Suponha que $X \mid \theta \sim \text{Mn}_3(10, 0)$, com $\theta = (0.20, 0.30, 0.15)$, como no Exemplo 2.1. Tomemos $A_0 = \{0, 1\}$, $A_1 = \{2, 3\}$. Então

$$\sum_{A_1} X_i \mid \theta = X_2 + X_3 \mid \theta \sim \text{Mn}_1(10, \theta_2 + \theta_3) \ ,$$

ou

$$X_2 + X_3 \mid \theta \sim \text{Mn}_1(10, 0.45).$$

Analogamente,

$$\begin{aligned} X_0 + X_1 \mid \theta &\sim \text{Mn}_1(10, 0.55) \ , \\ X_1 + X_3 \mid \theta &\sim \text{Mn}_1(10, 0.35) \ , \\ X_2 \mid \theta &\sim \text{Mn}_1(10, 0.30) \ . \end{aligned}$$

Note que, em geral, se $X \mid \theta \sim \text{Mn}_k(n, \theta)$ então $X_i \mid \theta \sim \text{Mn}_1(n, \theta_i)$, $i = 1, ..., k$.

Exemplo 2.4: Tabelas de Contingência 3x3.
Suponha que $X \mid \theta \sim \text{Mn}_8(n, \theta)$, como numa tabela de contingências 3x3:

x_{11}	x_{12}	x_{13}	$x_{1\cdot}$
x_{21}	x_{22}	x_{23}	$x_{2\cdot}$
x_{31}	x_{32}	x_{33}	$x_{3\cdot}$
$x_{\cdot 1}$	$x_{\cdot 2}$	$x_{\cdot 3}$	n

Aplicando o Resultado 2.4 obtemos

$$(X_{1\cdot}, X_{2\cdot}) \mid \theta \sim \text{Mn}_2(n, \theta'), \theta' = (\theta_1, \theta_{2\cdot}), \theta'_0 = \theta_3 \ .$$

Este mesmo resultado nos diz que

$$(X_{i1}, X_{i2}, X_{i3}) \mid \theta \sim \text{Mn}_3(n, \theta'_i) \ ,$$

com

$$\theta'_i = (\theta_{i1}, \theta_{i2}, \theta_{i3}) \ , \ \theta'_{0i} = 1 - \theta_{i\cdot} \ , \quad i = 1, 2, 3 \ .$$

Podemos então aplicar o Resultado 2.5 para obter a distribuição de probabilidade de cada linha da Tabela de Contingências, condicional na respectiva soma, ou na soma das outras duas linhas.
Temos então

$$(X_{i1}, X_{i2}) \mid x_{i\cdot} \ , \theta \sim \text{Mn}_2(x_{i\cdot}, \theta'_i)$$

com

$$\theta'_i = \frac{(\theta_{il}, \theta_{i2})}{\theta_{i\cdot}} \ , \ \theta'_{0i} = \frac{\theta_{i3}}{\theta_{i\cdot}} \ .$$

O seguinte resultado expressa a distribuição de $X \mid \theta$ em termos das distribuições condicionais, de cada linha da tabela, na sua respectiva soma, e em termos das distribuições destas somas.

Resultado 2.6: Se $X \mid \theta \sim \text{Mn}_{r^2-1}(n, \theta)$, como numa Tabela de contingências $r \times r$, então $P(X \mid \theta)$ pode ser escrita como

$$P(X \mid \theta) = \left[\prod_{i=1}^{r} P(X_{i1}, ..., X_{i,r-1} \mid x_{i\cdot}, \theta) \right] P(X_{1\cdot}, ..., X_{r-1\cdot} \mid \theta) \ .$$

Demonstração: Temos:

$$\begin{aligned} P(X \mid \theta) &= n! \prod_{i=1}^{r} \frac{\theta_i^{x_i}}{x_i!} = n! \ \frac{\theta_{11}^{x_{11}} \ ... \ \theta_{rr}^{x_{rr}}}{x_{11}! \ ... \ x_{rr}!} \\ &= \left[\prod_{i=1}^{r} \frac{x_{i\cdot}!}{x_{i1}! \ ... \ x_{ir}!} \left(\frac{\theta_{i1}}{\theta_{i\cdot}} \right)^{x_{i1}} ... \ \left(\frac{\theta_{ir}}{\theta_{i\cdot}} \right)^{x_{ir}} \right] \ \frac{n!}{x_{i\cdot}! \ ... \ x_{r\cdot}!} \ \theta_{1\cdot}^{x_{1\cdot}} \ ... \ \theta_{r\cdot}^{x_{r\cdot}} \end{aligned}$$

Dos Resultados 2.4 e 2.5, como no exemplo anterior, reconhecemos cada um dos primeiros r fatores acima como as probabilidades de cada linha da Tabela, condicional na sua respectiva soma, e reconhecemos o último fator como sendo a distribuição de probabilidade conjunta para as somas destas r linhas.

Corolário 2.1: Se $X \mid \theta \sim \text{Mn}_{r^2-1}(n, \theta)$, como no Resultado 2.6, então

$$P(X \mid x_{1\cdot}, ..., x_{r-1\cdot}, \theta) = \prod_{i=1}^{r} P(X_{i1}, ..., X_{i,r-1} \mid x_{i\cdot}, \theta)$$

e portanto, conhecidos $\theta, x_{1\cdot}, ..., x_{r-1\cdot}$,

$$(X_{11}, ..., X_{1,r-1}) \amalg ... \amalg (X_{r1}, ..., X_{r,r-1}) \ .$$

Demonstração: Como

$$P(X \mid \theta) = P(X \mid x_{1\cdot}, ..., x_{r-1\cdot}, \theta) P(X_{1\cdot}, X_{2\cdot}, ..., X_{r-1\cdot} \mid \theta) \ ,$$

do Resultado 2.6 obtemos a igualdade proposta.

O seguinte resultado será utilizado mais à frente e confere ao Resultado 2.6 uma forma de representação canônica para $P(X \mid \theta)$.

Resultado 2.7: Se $X \mid \theta \sim \text{Mn}_{r^2-1}(n, \theta)$, como no Resultado 2.6, então a transformação

$$T : (\theta_{11}, ..., \theta_{1r}, ..., \theta_{r1}, ..., \theta_{r,r-1}) \rightarrow (\lambda_{11}, ..., \lambda_{1,r-1}, ..., \lambda_{r1}, ..., \lambda_{r,r-1}, \eta_1, ..., \eta_{r-1})$$

dada por

$$\begin{aligned} \lambda_{11} = \tfrac{\theta_{11}}{\theta_{1\cdot}} \quad , \quad ... \quad , \quad \lambda_{1,r-1} = \tfrac{\theta_{1,r-1}}{\theta_{1\cdot}} \\ \vdots \\ \lambda_{r1} = \tfrac{\theta_{r1}}{\theta_{r\cdot}} \quad , \quad ... \quad , \quad \lambda_{r,r-1} = \tfrac{\theta_{r,r-1}}{\theta_{r\cdot}} \\ \eta_1 = \theta_{1\cdot}, \ \eta_2 = \theta_{2\cdot}, ..., \eta_{r-1} = \theta_{(r-1)\cdot} \end{aligned}$$

é uma transformação biunívoca definida em $\{0 < \theta_{11} + ... + \theta_{r,r-1} < 1 \ ; \ \ 0 < \theta ij < 1\}$ sobre o cubo unitário com dimensão $r^2 - 1$. Além disso, o Jacobiano da transformação T é

$$J = \eta^{r-1} \ \eta_1^{r-1} \ ... \ \eta_{r-1}^{r-1} \ (1 - \eta_1 - ... - \eta_{r-1})^{r-1} \ .$$

A demonstração é recomendada ao leitor, como exercício.

Exemplo 2.5: Vejamos o caso de uma tabela de contingência 2×2:

x_{11}	x_{12}
x_{21}	x_{22}

n

θ_{11}	θ_{12}
θ_{21}	θ_{22}

1

Para obtermos a representação canônica de $P(X \mid \theta)$ usamos a transformação T no caso $r = 2$:

$$\begin{aligned} \lambda_{11} &= \frac{\theta_{11}}{\theta_{11} + \theta_{12}} \ , \\ \lambda_{21} &= \frac{\theta_{11}}{\theta_{21} + \theta_{22}} \ , \\ \eta_1 &= \theta_{11} + \theta_{12} \ , \end{aligned}$$

de modo que

$$P(X \mid \theta) = \begin{pmatrix} x_{1\cdot} \\ x_{11} \end{pmatrix} \lambda_{11}^{x_{11}} (1 - \lambda_{11})^{x_{12}} \begin{pmatrix} x_{2\cdot} \\ x_{21} \end{pmatrix} \lambda_{21}^{x_{21}} (1 - \lambda_{21})^{x_{22}} \begin{pmatrix} n \\ x_{1\cdot} \end{pmatrix} \eta_1^{x_{1\cdot}} (1 - \eta_1)^{x_{2\cdot}} \ ,$$

$$0 < \theta_{11} < 1 \ , \ \ 0 < \theta_{21} < 1 \ , \ \ 0 < \eta_1 < 1 \ .$$

Referências

J.J.Martin (1975). *Bayesian decision and probelms and Markov Chains.*

C.A.B.Pereira, M.A.G.Viana (1982). *Elementos de Inferência Bayesiana.* 5o Sinape, São Paulo.

S.S.Wilks (1962). *Mathematical Statistics.* NY: Wiley.

R.L.Winkler (1975). *Statistics: Probability, Inference, and Decision.* Harcourt School.

Capítulo 4

Entropia

O conceito de entropia surgiu inicialmente a partir da mecânica estatística, e sua aplicação tem se estendido para diversos outros fenômenos (físicos ou não). A entropia de uma distribuição, $H(p(x))$, mede a incerteza, (ou impureza, confusão) de um sistema cujos estados $x \in \mathcal{X}$ tem esta distribuição. Seguiremos de perto a apresentação dada nas referências.

Se $H(p(x))$ é uma medida de incerteza, é razoavel que satisfaça os requisitos listados a seguir. Por simplicidade, apresentaremos aqui a teoria para espaços enumeráveis.

1) Se o sistema tem n estados possíveis, $x_1, \ldots x_n$, a entropia do sistema com uma dada distribuição $p_i \equiv p(x_i)$, é uma função,

$$H = H_n(p_1, \ldots, p_n)$$

2) H é uma função contínua.

3) H é uma função simétrica em seus argumentos.

4) A entropia de um sistema não se altera mediante a adição de um estado impossível, i.e.,

$$H_n(p_1, \ldots p_n) = H_{n+1}(p_1, \ldots p_n, 0)$$

5) A entropia de um sistema é mínima e nula, quando o sistema é determinado, i.e.,

$$H_n(0, \ldots, 0, 1, 0, \ldots 0) = 0$$

6) A entropia de um sistema é máxima, quando todos seus estados são igualmente prováveis, i.e.,

$$\frac{1}{n}\mathbf{1} = \arg\max H_n$$

7) A entropia maxima de um sistema aumenta com o número de estados, i.e.

$$H_{n+1}\left(\frac{1}{n+1}\mathbf{1}\right) > H_n\left(\frac{1}{n}\mathbf{1}\right)$$

8) Dados dois sistemas independentes, com distribuições p e q, a entropia do sistema composto é a soma das entropias dos sistemas individuais, i.e.,

$$H_{nm}(r) = H_n(p) + H_m(q) \quad , \; r_{i,j} = p_i \, q_j$$

A medida de entropia de Boltzmann-Gibbs-Shannon,

$$H_n(p) = -I_n(p) = -\sum_{i=1}^{n} p_i \log(p_i) = -\,\mathrm{E}_i \log(p_i) \quad , \quad 0 \, \log(0) \equiv 0$$

atende aos requisitos (1) a (8), e é a mais utilizada medida de entropia. Em Física e Estatística costumamos tomar o logaritmo na base de Neper, enquanto em Computação e Engenharia, na base 2. O oposto da medida de entropia, $I(p) = -H(p)$, a Neguentropia, é uma medida da Informação disponível sobre o sistema.

Com respeito ao requisito 8, podemos calcular a Negentropia de Boltzmann-Gibbs-Shannon do sistema composto no caso não termos independência:

$$\begin{aligned} I_{nm}(r) &= \sum_{i=1,j=1}^{n,m} r_{i,j} \, \log(r_{i,j}) = \sum_{i=1,j=1}^{n,m} p_i \mathrm{Pr}(j \,|\, i) \, \log\left(p_i \mathrm{Pr}(j \,|\, i)\right) \\ &= \sum_{i=1}^{n} p_i \log(p_i) \sum_{j=1}^{m} \mathrm{Pr}(j \,|\, i) + \sum_{i=1}^{n} p_i \sum_{j=1}^{m} \mathrm{Pr}(j \,|\, i) \, \log\left(\mathrm{Pr}(j \,|\, i)\right) \\ &= I_n(p) + \sum_{i=1}^{n} p_i \, I_m(q^i) \quad \text{onde} \quad q^i_j = \mathrm{Pr}(j \,|\, i) \end{aligned}$$

Se adicionarmos esta última identidade como um nono requisito na nossa lista, teremos uma caracterização da entropia de Boltzmann-Gibbs-Shannon, vide Kinchine (1957) e Renyi (1961).

Como todo conceito importante, esta medida de entropia foi re-descoberta inúmeras vezes em diferentes contextos, sendo que por vezes a unicidade do conceito não foi imediatamente (ou até passados muitos anos) percebida. É famoso o comentário de von Neumann, respondendo a pergunta de Shannon sobre como chamar o novo conceito recem-descoberto em teoria de sistemas de comunicação, "Chame o de entropia, pois esta é uma palavra que empregada em um argumento que o fará ganhar qualquer debate, já que muito poucas pessoas realmente compreendem seu significado".

Para verificar o requisito (6) é obedecido podemos usar (com $q \propto 1$) o seguinte lema:

Lema: Desigualdade de Shannon
Se p e q são duas distribuições sobre um sistema com n estados possíveis, e $q_i \neq 0$, então a Informação Relativa de p em relação a q, $I_n(p, q)$, é não negativa, anulando-se sse $p = q$, onde,

$$I_n(p, q) \equiv \sum_{i=1}^{n} p_i \, \log\left(\frac{p_i}{q_i}\right) \quad , \; I_n(p, q) \geq 0 \quad , \; I_n(p, q) = 0 \Rightarrow p = q$$

Demonstração: Pela desigualdade de Jensen, se φ é uma função convexa,

$$\mathrm{E}\left(\varphi(x)\right) \geq \varphi\left(\mathrm{E}(X)\right)$$

Tomando

$$\begin{aligned} \varphi(t) &= t\ln(t) \text{ e } t_i = \frac{p_i}{q_i} \\ \mathrm{E}_q(t) &= \sum_{i=1}^{n} q_i \frac{p_i}{q_i} = 1 \\ I_n(p,q) &= \sum q_i\, t_i \log t_i \geq 1\, \log(1) = 0 \end{aligned}$$

A desigualdade de Shannon motiva usarmos a Informação Relativa como uma medida de "distância" (não simétrica) de uma distribuição em relação a outra. Em Estatística esta medida é conhecida como Medida de Discriminação de Informação de Kullback-Leibler. As denominações Divergência Dirigida ou Informação Cruzada são habituais em Engenharia. A demonstração da desigualdade de Shannon motiva a seguinte generalização de divergência:

Definição: φ-divergência de Csiszar.
Dada uma função convexa φ,

$$\begin{aligned} d_\varphi(p,q) &= \sum_{i=1}^{n} q_i\, \varphi\left(\frac{p_i}{q_i}\right) \\ 0\,\varphi\left(\frac{0}{0}\right) &= 0 \ , \ 0\,\varphi\left(\frac{c}{0}\right) = c \lim_{t\to\infty} \frac{\varphi(t)}{t} \end{aligned}$$

Por exemplo, podemos definir a divergência quadrática e absoluta como

$$\begin{aligned} \chi^2(p,q) &= \sum \frac{(p_i - q_i)^2}{q_i} \quad , \quad \text{para } \varphi(t) = (t-1)^2 \\ Ab(p,q) &= \sum \frac{|p_i - q_i|}{q_i} \quad , \quad \text{para } \varphi(t) = |t-1| \end{aligned}$$

4.1 Max-Ent com Restrições Lineares

Dada uma densidade a priori, q, gostariamos de encontrar o vetor p que minimiza a Informação relativa $I_n(p,q)$, onde p está sujeito às restrições de ser uma distribuição, podendo haver restrições adicionais sobre a esperança de funções dos estados, i.e., queremos

$$p^* \in \arg\min I_n(p,q) \ \ , \ p \geq 0 \ \mid \mathbf{1}'p = 1 \text{ e } Ap = b \ \ , \ A\ (m-1) \times n$$

p^* é a distribuição de Mínima Informação, ou de Máxima Entropia, em relação a q, com as restrições A, b. Vemos ainda que podemos escrever a restrição de normalização como

uma restrição linear genérica, incluindo como 0-ésima linha da matriz A a linha $\mathbf{1}'$. Assim procederemos, não mais distinguindo a restrição de normalização das demais. Como de hábito, as operações $\odot$ e $\oslash$ indicam o produto e a divisão ponto-a-ponto entre matrizes de mesma dimensão.

A Lagrangiana deste problema e suas derivadas são:

$$\begin{aligned} L(p,w) &= p' \log(p \oslash q) + w'(b - Ap) \\ \frac{\partial L}{\partial p_i} &= \log(p_i/q_i) + 1 - w'A^i \\ \frac{\partial L}{\partial w_k} &= b_k - A_k p \end{aligned}$$

Igualando a zero as $n+m$ derivadas acima, temos um sistema de $n+m$ incognitas e equações, dando as condições de viabilidade e otimalidade do problema:

$$\begin{aligned} p_i &= q_i \exp\left(w'A^i - 1\right) \quad \text{ou} \quad p = q \odot \exp\left((w'\,A)' - \mathbf{1}\right) \\ A_k p &= b_k \quad , \; p \geq 0 \end{aligned}$$

Substituindo as probabilidades incognitas p_i, podemos escrever as condições de viabilidade e otimalidade (EVO) em função apenas das variáveis duais,

$$h_k(w) \equiv A_k \left(q \odot \exp\left((w'\,A)' - \mathbf{1}\right)\right) - b_k = 0$$

Esta forma das EVO motiva usarmos a lógica de algoritmos iterativos tipo Gauss-Seidel para sistemas lineares onde, ciclicamente, "acertamos" uma das equações do sistema. Para uma analise neste espírito deste tipo de algoritmo veja Elfving (1980), Censor and Zenios (1994, 1997), Iusem and Pierro (1987).

Algoritmo de Balanceamento de Bregman:

Inicialização:

Tome $t = 0$, $w^t \in \mathcal{R}^m$, e

$$p_i^t = q_i \exp\left(w^{t\prime} A^i \; - 1\right)$$

Passo de iteração: para $t = 1, 2, \ldots$

Tome

$$k = (t \bmod m) \quad \text{e} \quad \nu \,|\, \varphi(\nu) = 0 \;, \quad \text{onde}$$

$$
\begin{aligned}
w^{t+1} &= \begin{bmatrix} w_1^t \\ \vdots \\ w_k^t + \nu \\ \vdots \\ w_m^t \end{bmatrix} \\
p_i^{t+1} &= q_i \exp(w^{t+1\prime} A^i - 1) \quad = \quad p_i^t \exp(\nu A_k^i) \\
\varphi(\nu) &= A_k p^{t+1} - b_k
\end{aligned}
$$

De nossa discussão sobre otimização de Entropia com restrições lineares, fica claro que a distribuição que maximiza a entropia relativa de um sistema com restrições sobre a esperança de funções dos estados, $E_{p(x)} a_k(x) = \int a_k(x) p(x) dx = b_k$, (incluindo a restrição de normalização, $a_0 = 1, b_0 = 1$) tem forma

$$p(x) = q(x) \exp\left(-\theta_0 - \theta_1\, a_1(x) - \theta_2\, a_2(x) \ldots\right)$$

Note que tomamos $\theta_0 = -(w_0 - 1)$, $\theta_k = -w_k$ e indexamos o estado i pela variavél x, para estarmos de acordo com a forma padrão em textos de estatística.

Muitas das distribuições usuais em estatística podem ser interpretadas como distribuições de maxima entropia (em relação á distribuição uniforme), dadas algumas estrições sobre algumas esperanças de funções dos estados. Por exemplo:

A distribuição de Wishart

$$f(S \,|\, \nu, V) \equiv c(\nu, V) \exp\left(\frac{\nu - d - 1}{2} \log(\det(S)) - \sum_{i,j} V_{i,j} S_{i,j}\right)$$

é caracterizada como a distribuição de minima entropia no suporte $S > 0$, dadas as esperanças dos elementos e do log-determinante da matriz S.

A distribuição de Dirichlet

$$f(x \,|\, \theta) = c(\theta) \exp\left(\sum_{k=1}^{m} (\theta_k - 1) \log(x_k)\right)$$

é caracterizada como a distribuição de minima entropia no suporte $x \geq 0 \,|\, \mathbf{1}'x = 1$, dadas as esperanças dos logaritmos das coordenadas, $\mathrm{E}(\log(x_k))$.

Exercícios:
1) Implemente o algoritmo de Bregmann. Note que pode ser mais conveniente numerar as linhas de A de 1 a m, e tomar $k = (t \bmod m) + 1$.
2) Ganhei um dado, que cria a priori ser honesto. Um amigo emprestou o dado e relatou te-lo lançado 60 vezes, obtendo 4 i's, 8 ii's, 11 iii's, 14 iv's, 13 v's e 10 vi's.
A) Qual minha posteriori Bayesiana?

Bi) Qual é a média amostral do valor da face? (3.9).
Bii) Qual a esperança a posteriori desta estatística?
C) Liguei para o fabricante do dado, e este me garantiu que a esperança desta estatística, para este modelo de dado, é exatamente 4.0. Use o algoritmo de Bregman para obter a posteriori entrópica, que é a distribuição mais proxima da priori que obedece às restrições dadas. Use como priori: i) a uniforme, ii) a posteriori Bayesiana.
D) Discuta a diferença entre uma atualização Bayesiana e uma atualização entrópica. Qual é o dado em cada caso? Observações ou restrições?
E) Discuta como fazer testes hieárquicos para hipóteses compostas, de uma maneira coerente com a filosofia do FBST.
3) De a caracterização de máxima entropia de todas as distribuições vistas no curso.

4.2 Covergência da Posteriori

A Informação relativa, $I(p,q)$, pode ser usada na demonstração de uma série de resultados assimtóticos que fundamentam toda a teoria estatística Bayesiana. Delinearemos aqui os agumentos utilizados na obtenção deste tipo de resultado, seguindo Gelman (1995).

Theorema da Consistência da Posteriori para Parâmetros Discretos:

Considere um modelo onde $f(\theta)$ é a priori em um espaço paramétrico discreto, $\Theta = \{\theta^1, \theta^2, \ldots\}$, $X = [x^1, \ldots x^n]$ uma série de observações, e a posteriori, em função da priori e da distribuição amostral,

$$f(\theta^k \mid X) \propto f(\theta^k)\, p(X \mid \theta^k) = f(\theta^k) \prod_{i=1}^{n} p(x^i \mid \theta^k)$$

Considere ainda que neste modelo um único valor do vetor de parâmetros, θ^0, melhor aproxima a "verdadeira" distribuição amostral $g(x)$, no sentido de minimizar a informação relativa

$$\begin{aligned} \{\theta^0\} &= \arg\min_k I\left(g(x), p(x \mid \theta^k)\right) \\ I\left(g(x), p(x \mid \theta^k)\right) &= \int_{\mathcal{X}} g(x) \log\left(\frac{g(x)}{p(x \mid \theta^k)}\right) dx = \mathrm{E}_{\mathcal{X}} \log\left(\frac{g(x)}{p(x \mid \theta^k)}\right) \end{aligned}$$

Então,

$$\lim_{n\to\infty} f(\theta^k \mid X) = \delta(\theta^k, \theta^0)$$

Argumentação: Considere o coeficiente logarítmico

$$\log\left(\frac{f(\theta^k \mid X)}{f(\theta^0 \mid X)}\right) = \log\left(\frac{f(\theta^k)}{f(\theta^0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(x^i \mid \theta^k)}{p(x^i \mid \theta^0)}\right)$$

O primeiro termo é uma constante, e o segundo termo é uma somatória cujos termos tem valor esperado (em relação a x, para $k \neq 0$) negativo pois, por hipótese θ^0 é o único argumento que minimiza $I(g(x), p(x \mid \theta^k))$. Assim (para $k \neq 0$), o lado direito tende a menos infinito a medida que n aumenta. Portando, do lado esquerdo, temos que $f(\theta^k \mid X)$ tende a zero. Como a probabilidade total soma um, $f(\theta^0 \mid X)$ tende a um, QED.

Podemos estender o resultado acima para espaços paramétricos contínuos, assumindo diversas condições de "regularidade", como continuidade, diferenciabilidade e ter o argumento θ^0 como ponto interior de Θ. Neste contexto podemos afirmar que, fixada uma pequena vizinhança em torno de θ^0, como $C(\theta^0, \epsilon)$ o cubo de lado ϵ de centro θ^0, esta vizinhança concentra toda a massa de $f(\theta \mid X)$, a medida que o número de observações vai a infinito. Sob as mesmas condições de regularidade, temos também que o máximo da posteriori é um estimador consistente, i.e., $\widehat{\theta} \to \theta^0$.

Os próximos resultados referem-se a convergência em distribuição para uma distribuição normal. Para tanto precisamos do seguinte lema:

Lema da Identidade da Matriz de Informação: A matriz de informação de Fisher, $J(\theta)$, definida como menos a esperança da Hessiana de uma log-verossimilhança, pode também ser escrita como a matriz de covariancia do gradiente desta mesma log-verossimilhança, i.e.,

$$J(\theta) \equiv -\operatorname{E}_{\mathcal{X}} \frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^2} = \operatorname{E}_{\mathcal{X}} \left(\frac{\partial \log p(x \mid \theta)}{\partial \theta} \, \frac{\partial \log p(x \mid \theta)}{\partial \theta} \right)$$

Demonstração:

$$\int_{\mathcal{X}} p(x \mid \theta) dx = 1 \Rightarrow \int_{\mathcal{X}} \frac{\partial p(x \mid \theta)}{\partial \theta} dx = 0 \Rightarrow$$

$$\int_{\mathcal{X}} \frac{\partial p(x \mid \theta)}{\partial \theta} \frac{p(x \mid \theta)}{p(x \mid \theta)} dx = \frac{\partial \log p(x \mid \theta)}{\partial \theta} p(x \mid \theta) dx = 0$$

diferenciando novamente em relação ao parâmetro,

$$\int_{\mathcal{X}} \left(\frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^2} p(x \mid \theta) + \frac{\partial \log p(x \mid \theta)}{\partial \theta} \, \frac{\partial p(x \mid \theta)}{\partial \theta} \right) dx = 0$$

observando que o segundo termo da integral pode ser reescrito como

$$\int_{\mathcal{X}} \frac{\partial \log p(x \mid \theta)}{\partial \theta} \, \frac{\partial p(x \mid \theta)}{\partial \theta} \, \frac{p(x \mid \theta)}{p(x \mid \theta)} dx = \int_{\mathcal{X}} \frac{\partial \log p(x \mid \theta)}{\partial \theta} \, \frac{\partial \log p(x \mid \theta)}{\partial \theta} \, p(x \mid \theta) dx$$

obtemos o lema.

Teorema da Aproximação Normal da Posteriori: A distribuição da posteriori tende para uma distribuição Normal de média θ^0 e precisão $nJ(\theta^0)$.

Demonstração (esquemática): Basta escrever a expanção da log-posteriori como uma

série de Taylor centrada em $\widehat{\theta}$, até segunda ordem,

$$\begin{aligned}\log f(\theta \,|\, X) &= \log f(\widehat{\theta} \,|\, X) + \frac{\partial \log f(\widehat{\theta} \,|\, X)}{\partial \theta}(\theta - \widehat{\theta}) \\ &+ \frac{1}{2}(\theta - \widehat{\theta})' \frac{\partial^2 \log f(\widehat{\theta} \,|\, X)}{\partial \theta^2}(\theta - \widehat{\theta}) + \mathcal{O}(\theta - \widehat{\theta})^3\end{aligned}$$

O termo de ordem zero é uma constante. O termo linear é nulo, pois $\widehat{\theta}$ é o máximo da posteriori no interior de Θ. A Hessiana no termo quadrático é

$$H(\widehat{\theta}) = \frac{\partial^2 \log f(\widehat{\theta} \,|\, X)}{\partial \theta^2} = \frac{\partial^2 \log f(\widehat{\theta})}{\partial \theta^2} + \sum_{i=1}^{n} \frac{\partial^2 \log p(x^i \,|\, \widehat{\theta})}{\partial \theta^2}$$

A Hessiana é negativa definida, pelas condições de regularidade e por $\widehat{\theta}$ ser o máximo da posteriori. O primeiro termo é constante e o segundo a soma de n variáveis i.i.d.. Por outro lado, já obtivemos a consistência do estimador $\widehat{\theta}$, e também que θ^0 é o ponto que tende a concentrar toda a massa da posteriori. Vemos ainda que a Hessiana tende a crescer linearmente com n, e que os termos de ordem superior não podem crescer super-linearmente. Por outro lado, para um dado n e $\theta \to \widehat{\theta}$, o termo quadrático domina os termos de ordem superior, de modo que a aproximação quadrática da log-posteriori é progressivamente mais precisa, QED.

Teorema da Aproximação Normal do MLE: O estimador de máxima verossimilhança (MLE), é asintoticamente normal, com média θ^0 e precisão $nJ(\theta_0)$.

Demonstração (esquemática): Assumindo todas as condições de regularidade necessárias,

Pela condição de otimalidade de primeira ordem,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log p(x^i \,|\, \widehat{\theta})}{\partial \theta} = 0$$

assim, pelo teorema do valor médio, há um $\widetilde{\theta}$ tal que

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log p(x^i \,|\, \theta^0)}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log p(x^i \,|\, \widetilde{\theta})}{\partial \theta^2}(\theta^0 - \widehat{\theta})$$

ou, equivalentemente,

$$\sqrt{n}(\widehat{\theta} - \theta^0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log p(x^i \,|\, \widetilde{\theta})}{\partial \theta^2}\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial \log p(x^i \,|\, \theta^0)}{\partial \theta}$$

Assumiremos condições de regularidade suficientes para garantir que:

$$-\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log p(x^i \,|\, \widetilde{\theta})}{\partial \theta^2}\right]^{-1} \to J(\theta^0)^{-1}$$

pois o MLE é consistente, $\widehat{\theta} \to \theta^0$, e portanto também o valor médio, $\widetilde{\theta} \to \theta^0$; e

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log p(x^i \mid \theta^0)}{\partial \theta} \to N(0, J(\theta^0))$$

pois temos a soma de n vetores i.i.d. com média 0 e, pelo lema da identidade da matriz de informação, covariância $J(\theta^0)$.

Assim, temos finalmente que

$$\sqrt{n}(\widehat{\theta} - \theta^0) \to N\left(0, J(\theta^0)^{-1} J(\theta^0) J(\theta^0)^{-1}\right) = N\left(0, J(\theta^0)^{-1}\right)$$

Referências:

L.M.Bregman (1967). The Relaxation Method for Finding the Common Point Convex Sets and its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Mathematical Physics,* 7, 200-217.

Y.Censor, S.Zenios (1994). *Introduction to Methods of Parallel Optimization.* IMPA, Rio de janeiro.

C.Cercignani (1998). *Ludwig Boltzmann, The Man who Trusted Atoms.* Oxford Univ.

G.C.Chow (1983). *Econometrics.* Singapore: McGraw-Hill.

I.Csiszar (1974). Information Measures. *7th Prage Conf.of Information Theory,* 2, 73-86.

T.Elfving (1980). On Some Mrthods for Entropy maximization and Matrix Scaling. *Linear algebra and its applications,* 34, 321-339.

S.C.Fang, J.R.Rajasekera, H.S.J.Tsao (1997). *Entropy Optimization and Mathematical Programming.* Kluwer, Dordrecht.

A.Gelman, J.B.Carlin, H.S.Stern, D.B.Rubin (1995). *Bayesian Data Analysis.* London, Chapman and Hall.

D.V.Gokhale (1975). Maximum Entropy Characterization of some Distributions. In Patil,G.P., Kotz,G.P., Ord,J.K. *Statistical Distributions in Scientific Work.* V-3, 299-304.

A.N.Iusem, A.R.De Pierro (1986). Convergence Results for an Accelerated Nonlinear Cimmino Algorithm. *Numerische Matematik,* 46, 367-378.

J.N.Kapur (1989). *Maximum Entropy Models in Science and Engineering.* New Delhi: John Wiley.

A.I.Khinchin (1953). *Mathematical Fundadtions of Information Theory.* NY: Dover.

S.D.Pietra, V.Pietra, J.Lafferty (2001). *Duality and Auxiliary Functions for Bregman Distances.* Tec.Rep. CMU-CS-01-109R, Carnegie Mellon.

A.Renyi (1961). On Measures of Entropy and Information. *Proc. 4-th Berkeley Symp. on Math Sats. and Prob.* V-I, 547-561.

J.M.Schervich (1995). *Theory of Statistics.* Berlin, Springer.

M.Teboulle (1992). Entropic Proximal Mappings with Applications to Nonlinear Programming. *Mathematics of operations Reasearch,* 17, 670-690.

J.Uffink (1995). The Constraint Rule of the Maximum Entropy Principle. *Studies in the History and Philosophy of Modern Physics,* 26B, 223-261.

J.Uffink (1996). Can the Maximum Entropy principle be Explained as a Consistency Requirement?. *Studies in the History and Philosophy of Modern Physics,* 27, 47-79.

Capítulo 5

Ônus da Prova no Discurso Cientifico e Jurídico

5.1 Introdução

O Teste de Significância Totalmente Bayesiano (Full Bayesian Significance Test ou FBST) foi apresentado por Pereira e Stern, [16], como um teste Bayesiano coerente para hipóteses precisas. O FBST foi motivado por casos concretos de consultoria estatística no contexto legal, visando a certificação e verificação de software através de simulação de caixa preta, isto é, sem acesso ao código fonte. Neste contexto os testes de significância deveriam obedecer aos requisitos lógicos do principio jurídico do Benefício da Dúvida, ou Onus Probandi.

Estes requisitos lógicos, embora já anteriormente mencionados em outros trabalhos do autor, nunca foram formalmente analisados. Esta análise é o maior objetivo do presente artigo, e é feita nas seções 4, 6 e 7, com auxílio do formalismo de cálculo abstrato de crença (ABC), como definido por Darwiche e Ginsberg em [2] e [3]. O FBST é definido na seção 2, e o formalismo ABC é apresentado nas seções 3 e 5.

5.2 O Valor de Evidência do FBST

Seja $\theta \in \Theta \subseteq \mathcal{R}^p$ um vetor de parâmetros de interesse, e seja $L(\theta \mid x)$ a verossimilhança associada aos dados observados, x, como é padrão em modelagem estatística. No paradigma Bayesiano a densidade a posteriori, $p_x(\theta)$, é proporcional ao produto da verossimilhança e de uma densidade a priori,

$$p_x(\theta) \propto L(\theta \mid x)\, p(\theta).$$

A hipótese (nula) H, enuncia que o parâmetro está em um conjunto nulo, definido

por restrições de desigualdade e de igualdade, dadas por funções vetoriais g e h no espaço paramétrico,

$$\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq \mathbf{0} \wedge h(\theta) = \mathbf{0}\}$$

Estamos particularmente interessados em hipóteses precisas, i.e., com $\dim(\Theta_H) < \dim(\Theta)$.

O valor de evidência contra a hipóteses definido pelo FBST é

$$\begin{aligned} \mathrm{Ev}(H) &= \int_{T_H} p_x(\theta)\, d\theta \ , \text{ onde } \quad , \quad T_H = \{\theta \in \Theta \,|\, s(\theta) > s_H\} \\ s_H &= \sup_{\theta \in \Theta_H} s(\theta) \quad , \quad s(\theta) = \left(\frac{p_x(\theta)}{r(\theta)}\right) \end{aligned}$$

A função $s(\theta)$ é conhecida como surpresa a posteriori em relação a uma dada densidade de referência, $r(\theta)$. A função surpesa foi utilizada por vários estatísticos, como Good, Evans, e Royall, [4], [7], [20]. Seu papel no FBST é tornar $\mathrm{Ev}(H)$ explicitamente invariante por transformações apropriadas do sistema de coordenadas do espaço paramétrico.

O conjunto tangente T_H é um conjunto de maior surpresa relativa, (HRSS, ou Highest Relative Surprise Set). Ele contém os pontos do espaço paramétrico com surpresa, em relação a densidade de referência, maior que qualquer ponto do conjunto nulo, Θ_H. Quando $r(\theta) \propto 1$, T_H é o conjunto de maior densidade de Probabilidade a posteriori, ou HDPS Highest Density Probability Set, Tangente ao conjunto nulo Θ_H.

A probabilidade a posteriori de T_H dá uma indicação da inconsistência entre a posteriori e a hipótese: Valores "pequenos" de $Ev(H)$ indicam que (a variedade representando) a hipótese atravessa regiões de alta densidade, dando parca evidência contra a hipótese. Por outro lado, se a probabilidade a posteriori de T_H é "grande", o conjunto nulo situa-se em uma região de baixa densidade a posteriori, e os dados fornecem forte evidência contra a hipótese. O valor de evidência definido acima tem uma caracterização geométrica simples e intuitiva.

Diversas aplicações do FBST (incluindo certificação e verificação de software através de simulação de caixa preta), detalhes para implementação numérica, alguns comentários sobre implicações epistemológicas, e extensa bibliografia, podem ser encontrados nos artigos do autor previamente publicados.

5.3 Cálculo Abstrato de Crença

O FBST foi originalmente motivado por alguns requisitos sobre o que constitui um valor de evidência válido contra um enunciado hipotético. Sob circunstância adequadas, estes requisitos são bom senso de raciocínio jurídico. Estes requisitos serão enunciados precisamente na próxima seção, usando o formalismo ABC apresentado abaixo.

O Cálculo Abstrato de Crença, ou ABC Abstract Belief Calculus, é definido em [2] e [3] como uma generalização simbólica do cálculo de probabilidades. ABC é uma ferramenta poderosa. Além de servir para computação tanto numérica quanto simbólica, o formalismo também estabelece os fundamentos para algoritmos computacionais de propagação para crença abstrata. O ABC também unifica vários casos particulares de cálculo de incerteza propostos na literatura. É neste contexto que utilizaremos o ABC para analisar o conceito de valor de evidência na abordagem do FBST.

O primeiro conceito no ABC é o de Função de Suporte Abstrata, Φ, que atribui valores abstratos de suporte a enunciados em um universo $\mathcal{U}$, fechado por disjunção, negação e conjunção. Usamos a notação usual de teoria dos conjuntos para denotar a imagem dos valores de suporte aos enunciados, $\Phi(\mathcal{U})$. Os axiomas A1 a A5 abaixo impõem condições de coerência aos estados de suporte.

A1: Em qualquer função de suporte, enunciados equivalentes têm suporte equivalente,

$$(A \Leftrightarrow B) \Rightarrow \Phi(A) = \Phi(B)$$

A2: Existe uma soma de suportes, $\oplus$, tal que, para qualquer função de suporte, o valor do suporte da disjunção de dois enunciados logicamente disjuntos é uma função dos seus valores de suporte individuais.

$$\oplus : \Phi(\mathcal{U}) \times \Phi(\mathcal{U}) \mapsto \Phi(\mathcal{U}) \quad , \quad \neg(A \wedge B) \Rightarrow \Phi(A \vee B) = \Phi(A) \oplus \Phi(B)$$

A3: Para qualquer função de suporte, se o enunciado A implica o enunciado B, que por seu lado implica o enunciado C, e os enunciados A e C têm o mesmo valor de suporte, então todos os três enunciados têm o mesmo valor de suporte,

$$((A \Rightarrow B \Rightarrow C) \wedge (\Phi(A) = \Phi(C))) \Rightarrow \Phi(B) = \Phi(A)$$

A4: Para qualquer função de suporte, enunciados falsos têm valor de suporte zero,

$$A \text{ falso } \Rightarrow \Phi(A) = 0$$

A5: Para qualquer função de suporte, enunciados tautológicos têm valor de suporte pleno,

$$A \text{ verdadeiro } \Rightarrow \Phi(A) = 1$$

Pode ser mostrado que, vide [3], sob os axiomas A1 a A5 a soma de suportes é uma função parcial definida para quaisquer $a, b \in \Phi(\mathcal{U})$ que sejam valores de suporte de enunciados logicamente disjuntos. Mais precisamente, para quaisquer $a, b \in \Phi(\mathcal{U})$ tais que haja enunciados $A, B \in \mathcal{U}$ para os quais $a = \Phi(A)$, $b = \Phi(B)$ e $\neg(A \wedge B)$. Ademais, a soma de suportes tem as seguinte propriedades algébricas:

X0: Simetria,

$$a \oplus b = b \oplus a$$

X1: Transitividade,

$$(a \oplus b) \oplus c = a \oplus (b \oplus c)$$

X2: Convexidade,

$$\text{se } a \oplus b \oplus c = a \text{ então } a \oplus b = a$$

X3: Existe um único elemento $0 \in \Phi(\mathcal{U})$ tal que

$$\forall a \in \Phi(\mathcal{U}), a \oplus 0 = a$$

X4: Existe um único elemento $1 \in \Phi(\mathcal{U})$ tal que $1 \neq 0$ e

$$\forall a \in \Phi(\mathcal{U}), \exists! \, b \in \Phi(\mathcal{U}) \mid a \oplus b = 1$$

O par função de suporte e soma de suporte, $\langle \Phi, \oplus \rangle$ é denominado Estrutura Parcial de Suporte. Estruturas parciais de suporte para alguns cálculos de incerteza, a saber, lógica clássica, e os cálculos de probabilidade, possibilidade, e descrença, são dadas na tabela 1.

O valor de suporte de um enunciado não determina, em geral, o valor de suporte de sua negação. Todavia, para qualquer função de suporte, Φ, o formalismo ABC define a função de crença, $\ddot{\Phi}$, para a qual o valor de crença de um enunciado determina o valor de crença da sua negação.

$$\ddot{\Phi}(A) = \langle \Phi(A), \Phi(\neg A) \rangle$$

Tabela 1: Exemplos de estruturas parciais de suporte

$\Phi(\mathcal{U})$	$a \oplus b$	0	1	$a \preceq b$	Calculo
$\{0,1\}$	$\max(a,b)$	0	1	$a \leq b$	Logica classica
$[0,1]$	$a+b$	0	1	$a \leq b$	Probabilidade
$[0,1]$	$\max(a,b)$	0	1	$a \leq b$	Possibilidade
$\{0..\infty\}$	$\min(a,b)$	∞	0	$b \leq a$	Descrença

A estrutura parcial de suporte pode também ser usada para definir ordens parciais $\Phi(\mathcal{U})$ e em $\ddot{\Phi}(\mathcal{U})$. O símbolo $\preceq$ é usado para a ordem de suporte, e o símbolo $\sqsubseteq$ é usado para a ordem de crença.

$$a \preceq b \Leftrightarrow \exists \, c \,|\, a \oplus c = b \;\text{ e }\; \langle a, b \rangle \sqsubseteq \langle c, d \rangle \Leftrightarrow a \preceq c \text{ e } d \preceq b$$

Os estados extremos, mínimo e máximo, de suporte e crença, com respeito a estas ordens são, respectivamente, 0 and 1 para a ordem de suporte, e $\langle 0, 1 \rangle$ e $\langle 1, 0 \rangle$ para a ordem de crença. Enunciados com mínima e máxima crença são ditos ser, respectivamente, Rejeitados e Aceitos.

5.4 Evidência e Onus Probandi

O valor de evidência contra uma hipótese do FBST foi motivado por aplicações de estatística Bayesiana a questões legais, onde as hipóteses precisas testadas eram enunciados de réus, a serem julgados de acordo com o princípio jurídico do benefício da dúvida, ou Onus Probandi, [16]. Neste contexto, nossa interpretação do princípio do Onus Probandi na teoria estatística Bayesiana estabelece alguns requisitos para o valor de suporte, $\Phi(H) = \overline{\mathrm{Ev}}(H) = 1 - \mathrm{Ev}(H)$, a favor da hipótese $H : \theta \in \Theta_H \subseteq \Theta$. A saber:

R1: Valor de Evidência como uma Probabilidade: O valor de evidência contra uma hipótese, H, deve ser a probabilidade a posteriori de um subconjunto (mensurável) Γ_H do espaço paramétrico,

$$\mathrm{Ev}(H) = \int_{\Gamma_H} p_x(\theta) d\theta$$

Se um ponto do espaço paramétrico $\theta \in \Theta$ está no conjunto de evidência Γ_H dizemos que θ constitui evidência contra a hipótese H. Se θ está no conjunto nulo Θ_H dizemos que θ é compatível com (ou admissível, legal ou válido pela) hipótese H.

R2: Surpresa Relativa: Se um ponto θ constitui ou não evidência contra H, depende apenas da ordem no espaço paramétrico estabelecida pelo valor da surpresa relativa a uma dada densidade de referência, $s(\theta) = p_x(\theta)/r(\theta)$.

R3: Auto Incriminação Inválida: Um ponto paramétrico compatível com uma hipótese não pode constituir evidência contra a mesma hipótese,

$$\Theta_H \cap \Gamma_H = \emptyset$$

R4: Lei de De Morgan: Um ponto paramétrico constitui evidência contra um hipótese composta sse constitui evidência contra todos os seus termos,

$$\text{se } H = A \vee B \text{ então } \Gamma_H = \Gamma_A \cap \Gamma_B$$

R5: Interpretação mais Favorável: A evidência a favor de uma hipótese composta é a evidência mais favorável a favor de seus termos,

$$\text{se } \; H = A \vee B \quad \text{então } \; \overline{\mathrm{Ev}}(H) = \max(\overline{\mathrm{Ev}}(A), \overline{\mathrm{Ev}}(B))$$

R6: Coerência do Suporte: $\langle \overline{\mathrm{Ev}}, \max \rangle$ deve ser uma estrutura parcial de suporte.

R7: Continuidade: Se a densidade a posteriori $p_x(\theta)$ e as restrições definindo o conjunto nulo, são funções suaves (contínuas, diferenciáveis, etc.) de seus argumentos, então também o é o valor da evidência $\mathrm{Ev}(H)$.

R8: Invariância: $\mathrm{Ev}(H)$ deve ser invariante por qualquer reparametrização bijetiva e suave, i.e., transformações de coordenadas do espaço paramétrico.

R9: Consistência: $\mathrm{Ev}(H)$ deve ser um indicador consistente para aceitar/rejeitar a hipótese sendo testada, no sentido que $\mathrm{Ev}(H)$ converge para 0 ou 1, conforme H seja verdadeira ou falsa, com o aumento da informação proveniente dos dados.

Definir o valor de evidência por meio de uma medida de probabilidade é comum à maioria das teorias estatísticas de teste de significância. Na estatística freqüentista, por exemplo, um p-valor é definido como a probabilidade de que, sob a hipótese H, um ponto seja "mais extremo" que os dados observados. Esta é, todavia, uma probabilidade no espaço amostral. O conceito de p-valor também requer uma ordem no espaço amostral para definir quão extremo um ponto é. Para uma análise crítica de p-valores, [10], [18].

Em estatística Bayesiana, um valor de evidência é usualmente definido como uma probabilidade no espaço paramétrico, como requerido em R1. De acordo com Basu, [1], Good, [7], e outros, exigir que $\mathrm{Ev}(H)$ dependa dos dados observados apenas através da função de verossimilhança, é a essência do Princípio da Verossimilhança. Este requisito é estabelecido em R2.

Os requisitos R3, R4 e R5 tentam capturar o princípio do Onus Probandi, assim como este se apresentou na prática de assessoria e consultoria e na pesquisa do autor, como reportado em artigos já e a serem publicados. Um exemplo simples e sua discussão é apresentado na seção 6. O Onus Probandi é um princípio básico do raciocínio legal, também conhecido como Ônus da Prova, [5], [11]. Este princípio dode ser enunciado como:

Não existe culpa conquanto haja uma base razoável para crença, efetivamente imputando o ônus da prova (Onus Probandi) ao acusador, a quem, em um processo legal, cabe provar falsas as alegações do réu, sem fazer qualquer suposição que não tenha sido explicitamente enunciada pelo réu, ou decorrente da legislação existente.

O princípio da Interpretação mais Favorável, que de acordo com o contexto também é conhecido como Benefício da Dúvida, In Dubito Pro Reo, ou Presunção de Inocência, é uma conseqüência do princípio do Onus Probandi, e requer que a corte considere a evidência do ponto de vista mais favorável ao réu, [22]:

"Moreover, the party against whom the motion is directed is entitled to have the trial court construe the evidence in support of its claim as truthful, giving it its most favorable interpretation, as well as having the benefit of all reasonable inferences drawn from that evidence."

R6 requer que $\langle \overline{\mathrm{Ev}}(H), \max \rangle$ seja uma estrutura parcial de suporte. Darwiche, [3], faz uma análise exaustiva de porque R6 estabelece as condições lógicas mínimas para uma função de suporte.

R7, R8, e R9 são propriedades desejáveis padrão na teoria estatística de teste de hipótese. R9 é um corolário da teoria de convergência da posteriori, vide [6], cap.10.

Não é difícil verificar que os requisitos R1 a R9 são satisfeitos no caso do FBST, i.e.,

tomando $\Gamma_H = T_H$. Mais interpretações do FBST e sua estrutura parcial de suporte são dadas nas seções 6 e 7. Antes disto, todavia, temos que introduzir alguns fatos adicionais sobre o formalismo ABC.

5.5 Condicionalização

O formalismo ABC também estabelece um conjunto de axiomas para condicionalização, i.e., para como atualizar a função de suporte, Φ, para uma função de suporte, Φ_A, "a posteriori" da aceitação de um enunciado não rejeitado, A. Darwiche e Ginsberg, [2], [3], definem como Condicionalizações Plausíveis aquelas dadas por uma função (parcial),

$$\oslash : \Phi(\mathcal{U}) \times \Phi(\mathcal{U}) \mapsto \Phi(\mathcal{U})$$

obedecendo os Axiomas A6 a A11 abaixo. Para facilidade de notação, nos referiremos a $\Phi(B)$ e $\Phi_A(B)$, respectivamente, como o valor de suporte incondicional de B, e o valor de suporte condicional de B dado A (ou dada a aceitação de A). $\oslash$ é denominada Função de Escala do Suporte.

A6: O valor de suporte condicional de B dado $A \vee B$ é uma função dos valores de suporte incondicionais B e $A \vee B$,

$$\Phi_{A\vee B}(B) = \Phi(B) \oslash \Phi(A \vee B)$$

Pode ser demonstrado que o axioma 6 é equivalente a

$$\Phi_A(B) = \Phi(A \wedge B) \oslash \Phi(A)$$

A7: Aceitar um enunciado não rejeitado mantem todos os enunciados rejeitados.

$$(\Phi(A) \neq 0 \ \wedge \ \Phi(B) = 0) \ \Rightarrow \ \Phi_A(B) = 0$$

A8: Aceitar um enunciado aceito não modifica a função de suporte condicional,

$$\Phi(A) = 1 \ \Rightarrow \ \Phi_A = \Phi$$

A9: Quando $A \vee B$ é igualmente suportado por duas funções de suporte, condicionar em $A \vee B$ não introduz igualdade ou ordem entre os suportes incondicionais de A, se Φ e Ψ são funções de suporte e $\Phi(A \vee B) = \Psi(A \vee B)$, então

$$\Phi_{A\vee B}(A) \preceq (=)\Psi_{A\vee B}(A) \Rightarrow \Phi(A) \preceq (=)\Psi(A)$$

A10: Após aceitar uma conseqüência lógica do enunciado A, o suporte condicional de A, ou aumenta, ou não se altera,

$$\Phi(A \vee B) \neq 0 \ \Rightarrow \ \Phi(A) \preceq \Phi_{A\vee B}(A)$$

A11: Se o suporte condicional de A dado C é igual seu suporte condicional dado $B \wedge C$, então o suporte condicional de B dado C é igual a seu suporte condicional dado $A \wedge C$,

$$(\Phi(A \wedge B \wedge C) \neq 0 \wedge \Phi_C(A) = \Phi_{B\wedge C}(A)) \quad \Rightarrow \quad \Phi_C(B) = \Phi_{A\wedge C}(B)$$

$\langle \Phi(\mathcal{U}), \oplus, \oslash \rangle$ é denominado uma Estrutura de Suporte.

Para os exemplos na tabela 1, as funções de escala são, respectivamente para a lógica clássica, para os cálculos de probabilidade e possibilidade, e para o cálculo de descrença:

$$\Phi_A(B) = \min\left(\Phi(A \wedge B), \Phi(A)\right) \quad , \quad \Phi_A(B) = \frac{\Phi(A \wedge B)}{\Phi(A)} \quad , \quad \Phi_A(B) = \Phi(A \wedge B) - \Phi(A)$$

5.6 Cálculos de Crença Coexistentes no FBST

Uma análise crítica do valor de evidência do FBST, no contexto colocado pelas últimas seções, pode ajudar a desvendar os benefícios bem como alguns aparentes paradoxos resultantes do uso do FBST para teste de hipóteses em estatística.

No FBST, os valores de supporte, $\overline{\mathrm{Ev}}(H)$, são computados usando o cálculo de probabilidade padrão em Θ, que tem um operador de condicionalização intrínseco. As evidências computadas, por outro lado, formam uma estrutura parcial de suporte possibilística, o cálculo de evidência. Não é possível todavia definir uma função de escala para o cálculo de evidência que seja compatível com a função de suporte do FBST, $\overline{\mathrm{Ev}}$, assim com esta foi definida. Assim, dois cálculos de crença estão em uso simultâneo no Contexto do FBST: os cálculos de probabilidade e de possibilidade.

A maioria das teorias estatísticas padrão de teste de hipótese usam um único cálculo de crença, a saber, o cálculo de probabilidade. Para tanto, elas tentam usar a probabilidade no conjunto nulo como um valor de suporte para a hipótese. Isto pode também tomar uma forma indireta, como pela integração de uma função de utilidade ou de perda. Em muitas aplicações legais, onde se coloca uma hipótese H composta e precisa, nem uma "probabilidade" do conjunto nulo, $\Pr(\Theta_H)$, nem uma razão de chances, $\Pr(\Theta_H)/\Pr(\Theta_{\overline{H}})$, são fornecidas pelo réu, ou tacitamente implicadas pela legislação vigente. De acordo com o requisito R2, se tais probabilidades não são dadas, então tais probabilidades não podem ser utilizadas. Esta afirmação contradiz muitas práticas estatísticas bem estabelecidas para teste de hipóteses precisas, incluindo testes baseados em fatores de Bayes, [7].

Como subterfúgio para obter uma probabilidade artificial para o conjunto de medida nula Θ_H, no caso de uma hipótese precisa, muitos testes de hipótese padrão em estatística Bayesiana usam uma parametrização particular da hipótese, bem como uma medida de probabilidade definida na sub-variedade representando a hipótese derivada desta parametrização, por vezes em conjunção com uma massa a priori definida para a hipótese precisa, [6], [8].

Outro artifício freqüentemente utilizado em testes padrão para hipóteses precisas é o procedimento, por vezes trabalhoso, de eliminação de parâmetros estorvo (nuisance), [1], [21]. O FBST não segue o paradigma de eliminação de parametros estorvo; de fato, permanecer no espaço paramétrico original, em sua plena dimensão, explica a propriedade de "regularização intrínseca" do FBST, quando este é usado para seleção de modelos, [17].

O FBST é baseado na probabilidade do conjunto tangente, e não diretamente na probabilidade do conjunto nulo. Desta forma o FBST pode sobrepujar várias dificuldades práticas e conceituais de outros testes estatísticos para hipóteses precisas, relacionadas ao uso direto ou indireto de $\Pr(\Theta_H)$, vide [1], [7], [10], e os artigos anteriores do autor.

Examinemos alguns aspectos da estrutura parcial de suporte do FBST. O requisito da interpretação mais favorável implica que o cálculo de evidência no FBST tenha uma estrutura possibilística, e não probabilística. Novamente, este requisito contraria várias abordagens estabelecidas para teste de hipóteses precisas que utilizam diretamente a estrutura de suporte probabilística.

Darwiche, [3], faz alguns comentários interessantes a respeito das ordens de suporte e crença. A saber:

1- Se dois enunciados têm igual crença, então eles têm igual suporte; mas não a conversa.

2- Enunciados rejeitados são sempre minimamente suportados, e enunciados aceitos são sempre maximalmente suportados. Mas se todavia enunciados minimamente suportados são rejeitados, enunciados maximalmente suportados não são necessariamente aceitos.

3- Um enunciado e sua negação podem ser maximalmente suportados simultaneamente, enquanto nenhum deles é aceito.

Considere como um exemplo ilustrativo, as hipóteses

$$A : \theta \in \Theta \quad \text{e} \quad B : \theta \in \{\widehat{\theta}\}$$

onde $\widehat{\theta}$ é o único argumento que maximiza uma posteriori (própria e suave) no espaço paramé-trico $\Theta = \mathcal{R}^p$, $\{\widehat{\theta}\} = \arg\max_{\theta\in\Theta} p_x(\theta)$. Assuma uma referência uniforme, $r(\theta) \propto 1$. Temos portanto, $\overline{\text{Ev}}(A) = \overline{\text{Ev}}(B) = \overline{\text{Ev}}(\neg B) = 1$ and $\overline{\text{Ev}}(\neg A) = 0$. Assim, A e B têm ambas pleno suporte, mas A é aceita, enquanto B não o é. Este exemplo, ou variantes dele, foram dados ao autor tanto como exemplo de como uma função de suporte deveria funcionar no contexto jurídico, quanto como um paradoxo, no contexto de testes de significância tradicionais.

No contexto jurídico a interpretação é a seguinte: Um réu descreve um sistema (máquina, software, código genético, etc.) por um parâmetro θ, e afirma que θ foi fixado em um valor pertencente a um conjunto legal ou válido, o conjunto nulo, Θ_H. O parâmetro não pode ser observado diretamente, mas podemos observar uma variável aleatória cuja distribuição é uma função $f(x;\theta)$. O parâmetro θ foi fixado em um, e apenas um valor.

Afirmar que θ foi fixado no valor mais verossimilhante, $\theta = \widehat{\theta}$, (dadas n observações) deve dar ao réu pleno suporte, pois ser absolutamente vago, i.e., afirmar apenas que $\theta \in \Theta$, não deve colocá-lo em posição melhor.

Na maioria dos testes de significância tradicionais em estatística, $\Phi(\Theta_H)$ depende da medida de probabilidade do conjunto nulo, $\Pr(\Theta_H)$. Se Θ_H é um conjunto unitário em $\mathcal{R}^p$, com uma posteriori suave, então ele deveria ter suporte nulo. Realmente, a refutação de qualquer hipótese precisa é um preço que muitos filósofos, como Popper [19], e vários estatísticos estão dispostos a pagar, como explicitamente afirmado por I.J.Good:

"If by the truth of Newtonian mechanics we mean that it is approximately true in some appropriate well defined sense we could obtain strong evidence that it is true; but if we mean by its truth that it is exactly true then it has already been refuted. ... Very often the statistician doesn't bother to make it quite clear whether his null hypothesis is intended to be sharp or only approximately sharp. ... It is hardly surprising then that many Fisherians (and Popperians) say that - you can't get (much) evidence in favor of the null hypothesis but can only refute it."

5.7 Comentários Finais e Agradecimentos

Para discutir conceitos como: Testar uma hipótese e aceitá-la ou rejeitá-la em um dado nível; o poder do teste; níveis ótimos, etc. a teoria do FBST exige novos desenvolvimentos. Isto é feito em [23], e outros artigos do autor. Para uma visão alternativa do FBST, sob a ótica da teoria de decisão, vide [13].

A interpretação literal do princípio da dúvida sugere tomar como priori a densidade (possivelmente imprópria) uniforme como densidade de referência, no espaço paramétrico natural. No contexto Bayesiano, este é normalmente o espaço onde o cientista acessa sua priori. Podemos generalizar o procedimento usando outras densidades de referência, como por exemplo uma densidade não informativa, caso alguma esteja disponível. Esta possibilidade é sugerida pelo artigo de Evans, [4], em conjunção com as regras de Jeffreys para obter prioris não infromativas, [24], cap.2.

O autor é grato à FAPESP pela bolsa 2001-03484-1, ao CNPq pela bolsa de produtividade em pesquisa, ao BIOINFO e ao Departamento de Ciência da Computação da Universidade de São Paulo, bem como ao Departamento de ciência Matemáticas da Universidade Estadual de Nova York, SUNY-Binghamton, USA. O autor é grato a vários de seus colegas, especialmente a Wagner Borges, Carlos Alberto de Bragança Pereira, Sergio Wechsler, e Shelemyahu Zacks. O autor pode ser contatado em *jstern@ime.usp.br*.

Referências

D.Basu (1988). Statistical Information and Likelihood. Edit J.K.Ghosh. *Lect. Notes in Statistics*, 45.

A.Y.Darwiche e M.L.Ginsberg (1992). A symbolic Generalization of Probability Theory. *AAAI-92, Tenth National Conference on Artificial Inteligence.*

A.Y.Darwiche (1993). A symbolic Generalization of Probability Theory. Ph.D. Thesis, Stanford Univ.

M.Evans (1997). Bayesian Inference Procedures Derived via the Concept of Relative Surprise. *Communications in Statistics*, 26, 1125–1143.

R.H.Gaskins (1992). *Burdens of Proof in Modern Discourse.* New Haven: Yale Univ. Press.

M.H.DeGroot (1970). *Optimal Statistical Decisions.* NY: McGraw-Hill.

I.J.Good (1983). *Good Thinking.* Univ. of Minesota.

I.Hacking (1965). *Logic of Statistical Inference.* Cambridge Univ. Press.

T.Z.Irony, M.Lauretto, C.A.B.Pereira, e J.M.Stern (2002). A Weibull Wearout Test: Full Bayesian Approach. In: Y.Hayakawa, T.Irony, M.Xie. *Systems and Bayesian Reliability*, 287–300. Singapore: World Scientific.

O.Kempthorne (1980). Foundations of Statistical Thinking and Reasoning. *Australian CSIRO-DMS Newsletter*, 68, 1–5; 69, 3–7.

J.Kokott (1998). *The Burden of Proof in Comparative and International Human Rights Law.* The Hague: Kluwer.

M.S.Lauretto, F. Nakano, C.O.Ribeiro, J.M.Stern (1998). REAL: Real Attribute Learning Algorithm for Strategic Market Operations. ISAS-SCI-98, 2, 315–321.

M.R.Madruga, L.G.Esteves e S.Wechsler (2001). The Bayesianity of Pereira-Stern Tests. *Test*, 10, 291–299.

M.R.Madruga, C.A.B.Pereira e J.M.Stern (2003). Bayesian Evidence Test for Precise Hypotheses. *Journal of Statistical Planning and Inference.* In press, doi:10.1016/S0378-3758(02)00368-3.

C.A.B.Pereira e J.M.Stern (1999). A Dynamic Software Certification and Verification Procedure. ISAS-SCI-99, 2, 426–435.

C.A.B.Pereira e J.M.Stern (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal*, 1, 69–80.

C.A.B.Pereira e J.M.Stern (2001). Model Selection: Full Bayesian Approach. *Environmetrics*, 12, 559–568.

C.A.B.Pereira e S.Wechsler (1993). On the Concept of *p-value*. *Brazilian Journal of Probability and Statistics*, 7, 159–177.

K.R.Popper (1989). *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge.

R.Royall (1997). *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.

D.B.Rubin (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12, 1151–1172.

Ruta v. Breckenridge-Remy Co., USA, 1982.

J.M.Stern e S.Zacks (2002). Testing the Independence of Poisson Variates under the Holgate Bivariate Distribution The Power of a new Evidence Test. *Statistical and Probability Letters*, 60, 313–320.

A.Zellner (1971). *An Introduction to Bayesian Inference in Econometrics.* NY: Wiley.

Appendix A

Probabilidade

A.1 Espaços de Probabilidade Discretos

Em muitas circunstâncias estamos interessados em estudar situações que produzem resultados imprevisíveis. Chamamos estas situações de *experimentos*. Por exemplo, ao lançarmos um dado, consideramos seis resultados possíveis, cada resultado correspondendo a uma determinada face do dado voltada para cima. Chamamos de *Espaço Amostral*, $\mathcal{A}$, ao conjunto de todos os possíveis resultados. Espaços amostrais *discretos* contêm um número finito (ou enumerável) de resultados. Sempre que possível apresentaremos a teoria em espaços discretos, onde ela é muito mais simples que em espaços contínuos. Chamamos *evento* um subconjunto do espaço amostral, $E \subseteq \mathcal{A}$. Assim, no exemplo do dado, uma representação do espaço amostral é o conjunto $\{F1, F2, F3, F4, F5, F6\}$; nesta representação, o evento "obter uma face ímpar" corresponde ao subconjunto $\{F1, F3, F5\}$.

Nos experimentos que estudaremos é útil atribuirmos a cada um dos eventos um valor numérico. Esta atribuição, ou função, chama-se *variável aleatória*. No exemplo do dado é usual atribuir valores, entre 1 e 6, a cada um dos eventos, i.e. atribuir $X(Fk) = k$.

Consideremos outro exemplo: lançamos dois dados, um verde e outro amarelo, cada qual tendo suas faces numeradas de 1 a 6. No exemplo dos dados verde e amarelo, entre muitas outras, poderíamos considerar as seguintes variáveis aleatórias:

1. O número decimal de dois dígitos cujo primeiro dígito corresponde a face de cima do dado verde, e o segundo dígito corresponde a face de cima do dado amarelo.

2. A soma dos números na face de cima dos dois dados.

Note que construímos duas variáveis aleatórias distintas, sobre um mesmo experimento: o lançamento dos dados verde e amarelo. Note também que pelo valor da primeira

variável aleatória podemos saber exatamente o resultado obtido no experimento; o mesmo já não é verdade para a segunda variável aleatória. Em geral, quando lidamos com um experimento, temos uma dada variável aleatória em mente e, não havendo ambigüidade, quando mencionamos o experimento já subentendemos a variável aleatória apropriada.

Definimos a imagem de um evento $E \subseteq \mathcal{A}$ por X, $X(E)$, como o conjunto de todos os valores assumidos por X dentro de E, i.e., $X(E) \equiv \{x = X(e), e \in E\}$. A imagem do espaço amostral é $\mathcal{X} = X(\mathcal{A})$. Analogamente definimos a pré-imagem de C por X, $X^{-1}(C)$, como o evento formado pelos resultados do experimento onde a variável X assume valores em C, i.e., $X^{-1}(C) = \{a \in A \,|\, X(a) \in C\}$.

Para não sobrecarregar a notação, não havendo ambiguidade, falamos abreviadamente do evento $A \subseteq \mathcal{X}$, no lugar da pré-imagem $X^{-1}(A)$. Ainda para não sobrecarregar a notação, freqüentemente abreviaremos a descrição completa de um conjunto pela condição lógica que o define, como por exemplo $X < x$ ao invés de $\{X \in \mathcal{X} \,|\, X < x\}$. Assim, geralmente um evento $A = X^{-1}(\{X < x\})$, será abreviada por $X < x$.

Interpretamos a *probabilidade* de um evento como a freqüência com que o obtemos o evento como resultado do experimento, ou como o grau de certeza que temos que o evento será observado. A probabilidade de que uma variável aleatória X assuma um valor dentro de um conjunto C, é simplesmente a probabilidade da pré-imagem de C por X. A *distribuição de probabilidade* de uma variável aleatória (discreta) é uma tabela especificando a probabilidade de que a variável assuma cada um dos seus possíveis valores.

Exemplos:

- Um dado é dito honesto se a freqüência com que gera cada um dos valores $\{1, \ldots, 6\}$ é a mesma. A probabilidade de obtermos um dado valor com um dado honesto é portanto de $1/6$.
- No experimento descrito no exemplo dos dados verde e amarelo, assumindo que ambos os dados são honestos, a probabilidade de cada variável aleatória assumir cada um dos valores possíveis é dada pelas tabelas seguintes:

  1. $\Pr(X = x) = 1/36$, $\forall x \in \{11, 12, \ldots, 16, 21, \ldots, 26, \ldots, 61, \ldots, 66\}$.
  2. $X^{-1}(2) = \{(1,1)\} \Rightarrow \Pr(X = 2) = 1/36$, $X^{-1}(3) = \{(1,2),(2,1)\} \Rightarrow \Pr(X = 2) = 2/36$, $\ldots$, $X^{-1}(12) = \{(6,6)\} \Rightarrow \Pr(X = 12) = 1/36$.

A probabilidade condicional de A dado H, $\Pr(A \,|\, H)$, representa a probabilidade de ocorrencia do evento A, quando temos certeza do evento H. Para não sobrecarregar a notação, quando um determinado conhecimento, H, estiver implícito no contexto, abreviamos a notação, escrevendo $\Pr(A)$ ao invés de $\Pr(A \,|\, H)$.

Para assegurar a possibilidade de interpretar probabilidade de um evento como a de sua freqüência relativa, ou o grau de certeza da sua ocorrencia, exigiremos que uma (medida

de) probabilidade sobre o espaço amostral de um experimento aleatório, satisfaça os três Axiomas, ou Leis de Coerência seguintes.

L1 Convexidade: $0 \leq \Pr(A \,|\, H) \leq 1$ e $\Pr(A \,|\, A) = 1$.
L2 Adição finita: $\Pr(A \vee B \,|\, H) = \Pr(A \,|\, H) + \Pr(B \,|\, H)$ se $A \cap B = \emptyset$.
L3 Multiplicação: $\Pr(A \wedge B \,|\, H) = \Pr(A \,|\, H)\Pr(B \,|\, A \wedge H)$.

De L3 obtemos imediatamente a relação

$$\Pr(B \,|\, A) = \Pr(A \wedge B)/\Pr(A)$$

que é por vezes apresentada como a definição de probabilidade condicional.

Uma extensão da lei da Adição finita é apresentada abaixo. Entretanto não é conseqüência de conceitos simples. Note que nenhuma das 3 leis pode ser deduzida das outras.

L2' Adição enumerável: Seja $(A_1, ..., A_n, ...)$ um conjunto enumerável de eventos mutuamente exclusivos. Então,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i \,|\, H\right) = \sum_{i=1}^{\infty} \Pr(A_i \,|\, H) \ , \ A_i \cap A_j = \emptyset, i \neq j$$

De L3 obtemos imediatamente a relação

Se em L3 for razoável escrever $\Pr(A \wedge B \,|\, H) = \Pr(A \,|\, H)\Pr(B \,|\, H)$, dizemos que A e B são (condicionalmente a H) independentes. No caso de $A = \{X = x\}$ e B=$\{$Y=y$\}$ e esta relação valer para todo $x \in \mathcal{X}$ e $y \in \mathcal{Y}$ então X e Y são variáveis aleatórias (condicionalmente a H) independentes. Isto é,

$$\Pr(X = x \,|\, Y = y \ \wedge H) = \Pr(X = x \,|\, H), \forall [x, y] \in \mathcal{X} \times \mathcal{Y}$$

Se X e Y são independentes, notamos X II Y. Em palavras, se X II Y, saber o valor de Y não altera a função de probabilidade (condicional) de X, ou ainda

$$X \text{ II } Y \Leftrightarrow \Pr(X = x \wedge Y = y) = \Pr(X = x)\Pr(Y = y) \ , \ \forall [x, y] \in \mathcal{X} \times \mathcal{Y}$$

Introduziremos a fórmula de inversão do condicionamento usando apenas as leis de probabilidade listadas acima. Considere conhecidas as funções de probabilidade $\Pr(X = x \,|\, H)$ e $\Pr(Y = y \,|\, X = x \wedge H)$. Usando L3 podemos escrever

$$\begin{aligned} \Pr(X = x \wedge Y = y \,|\, H) &= \Pr(X = x \,|\, H)\Pr(Y = y \,|\, X = x \wedge H) \\ &= \Pr(Y = y \,|\, H)\Pr(X = x \,|\, Y = y \wedge H) \end{aligned}$$

e assim,

$$\begin{aligned} \Pr(X = x \,|\, Y = y \wedge H) &= \frac{\Pr(X = x \wedge Y = y \,|\, H)}{\Pr(Y = y \,|\, H)} \\ &= \frac{\Pr(X = x \,|\, H)}{\Pr(Y = y \,|\, H)}\Pr(Y = y \,|\, X = x \wedge H) \end{aligned}$$

Note que também necessitamos da função de probabilidade marginal de Y para conhecermos a função condicional de X dado Y. Por outro lado, usando L2 obtemos a teorema da probabilidade total. Isto é,

$$\begin{aligned}\Pr(Y = y \,|\, H) &= \sum_{\mathcal{X}} \Pr(X = x \wedge Y = y \,|\, H) \\ &= \sum_{\mathcal{X}} \Pr(X = x \,|\, H)\Pr(Y = y \,|\, X = x \wedge H)\end{aligned}$$

onde $\sum_{\mathcal{X}}$ representa soma sobre todo o espaço $\mathcal{X}$. Usando essa expressão no denominador da fórmula anterior obtemos a famosa Fórmula de Bayes:

$$\Pr(X = x \,|\, Y = y \wedge H) = \frac{\Pr(X = x \,|\, H)\Pr(Y = y \,|\, X = x \wedge H)}{\sum_{\mathcal{X}} \Pr(X = x \,|\, H)\Pr(Y = y \,|\, X = x \wedge H)}$$

Exercício: Prove, a partir dos axiomas, as seguintes propriedades da probabilidade:

4. $\Pr(\emptyset) = 0$,
5. $\Pr(\bar{B}) = 1 - \Pr(B)$,
6. $\Pr(B \cup C) = \Pr(B) + \Pr(C) - \Pr(B \cap C)$

A.2 Esperança

Dado um experimento, o valor esperado de uma variável aleatória X, ou sua *esperança*, é a média aritmética sobre o conjunto dos valores possíveis, $x \in \mathcal{A}$, ponderados pela distribuição de probabilidade:

$$E(X) = \sum_{x \in X(A)} x\Pr(X = x) \quad .$$

Dada uma amostra, i.e. uma seqüência de valores,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad ,$$

definimos a sua **média** aritmética como sendo

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i \quad .$$

A.2.1 Propriedades de Transformação

Estaremos sempre interessados em estudar transformações lineares de variáveis aleatórias, $Z = \alpha X + \beta Y + \gamma$ e como estas transformações se refletem nas quantidades estatísticas a serem definidas, por exemplo:

$$E(\alpha X + \beta Y + \gamma) = \alpha E(X) + \beta E(Y) + \gamma$$

A prova pode ser dividida em duas partes:

$$\begin{aligned} E(\alpha X + \gamma) &= \sum_x (\alpha x + \gamma)\Pr(X = x) \\ &= \gamma + \alpha \sum_x x\Pr(X = x) \\ &= \gamma + \alpha E(X) \end{aligned}$$

$$\begin{aligned} E(X + Y) &= \sum_{x,y} (x + y)\Pr(X = x \wedge Y = y) \\ &= \sum_{x,y} x\Pr(X = x \wedge Y = y) \quad + \sum_{x,y} y\Pr(X = x \wedge Y = y) \\ &= \sum_x x\Pr\left(X = x \wedge Y \in Y(A)\right) + \sum_y y\Pr\left(Y = y \wedge X \in X(A)\right) \\ &= E(X) + E(Y) \end{aligned}$$

A.3 Covariância

A variância de uma variável aleatória é uma medida de erro ou dispersão desta variável em torno da sua esperança:

$$\mathrm{Var}(X) = E\left((X - E(X))^2\right) \quad .$$

É fácil ver que também podemos calcular a variância como

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - E(X)^2 \end{aligned}$$

O desvio padrão, $\sigma_x = \sqrt{\mathrm{Var}(x)}$ tem a mesma dimensão, ou unidade de medida, que $E(x)$ ou x, i.e., é uma medida de desvio comensurável com a média ou os valores assumidos pela variável aleatória, sendo portanto de interpretação mais natural.

A covariância entre duas variáveis aleatórias, X e Y, é definida como

$$\mathrm{Cov}(X,Y) = E\left((X - E(X))(Y - E(Y))\right) \quad .$$

A covariância é uma medida da "interdependência" dos desvios de ambas as variáveis em relação a média. Adiaremos uma interpretação intuitiva mais detalhada para o conceito de correlação, discutido adiante. Por hora, podemos verificar as seguintes propriedades:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\left(XY - YE(X) - XE(Y) + E(X)E(Y)\right) \\ &= E(XY) - E(Y)E(X)\end{aligned}$$

donde

$$\mathrm{Cov}(X,X) = E(X^2) - E(X)^2 = \mathrm{Var}(X)$$

A.3.1 Propriedades de Transformação

Lema :

$$\begin{aligned}&\mathrm{Cov}(\alpha X + \beta Y + \gamma, Z) \\ &= E((\alpha X + \beta Y + \gamma)Z) - E(\alpha X + \beta Y + \gamma)E(Z) \\ &= \alpha E(XZ) + \beta E(YZ) + \gamma E(Z) - \alpha E(X)E(Z) - \beta E(Y)E(Z) - \gamma E(Z) \\ &= \alpha\,\mathrm{Cov}(X,Z) + \beta\,\mathrm{Cov}(Y,Z)\end{aligned}$$

Do lema segue que:

$$\mathrm{Var}(\alpha X + \beta Y + \gamma) = \alpha^2\,\mathrm{Var}(X) + \beta^2\,\mathrm{Var}(y) + 2\alpha\beta\,\mathrm{Cov}(X,Y)$$

Dados X e Y, vetores de variáveis aleatórias, sua matriz de correlação é definida por

$$\mathrm{Cov}(X,Y)_{i,j} \equiv \mathrm{Cov}(X_i, Y_j)$$

$$\mathrm{Cov}(X) \equiv \mathrm{Cov}(X,X)$$

Dada A uma matriz real, os resultados precedentes implicam na forma genérica da esperança e variância de uma transformação linear, que é dada por:

$$E(AX + b) = AE(X) + b$$

$$\mathrm{Cov}(AX + b, Y) = A\ \mathrm{Cov}(X,Y)\ ,\ \ \mathrm{Cov}(X, AY + b) = \mathrm{Cov}(X,Y)\,A'$$

$$\mathrm{Cov}(AX + b) = A\ \mathrm{Cov}(X)\,A'$$

Em estatística é usual a notação $\mathrm{Cov}(X)_{i,j} = \sigma_{i,j}$. Nesta notação o desvio padrão é $\sigma_i = \sqrt{\sigma_{i,i}}$.

Consideremos n variáveis aleatórias independentes, e identiacmente distribuidas, i.i.d., $X = [X_1, \ldots, X_n]$. A variância da média é dada por

$$E(\bar{X}) = E(\ (1/n)\ \mathbf{1}'X\ ) = (1/n)\ \mathbf{1}'E(X) = E(X_1)$$

na ultima equação denotamos a somatória $\sum X$ por $\mathbf{1}'X$, onde $\mathbf{1}' = [1, 1, \ldots 1]$.

$$\begin{aligned} \mathrm{Var}(\bar{X}) &\equiv Var(\ (1/n)\ \mathbf{1}'X\ ) \\ &= (1/n^2)\ \mathbf{1}'\ \mathrm{diag}([\sigma_{1,1}, \ldots \sigma_{n,n}])\ \mathbf{1} \\ &= (1/n^2)(\ \sigma_{1,1} + \ldots + \sigma_{n,n}\ ) \\ &= (1/n)\sigma_{1,1} \end{aligned}$$

Consideremos n vetores aleatórios $m \times 1$ e i.i.d. (independentes, e identiacmente distribuidos), $X = [X^1, \ldots, X^n]$. A soma destes vetores,

$$Y \equiv X^1 + X^2 \ldots + x^n \ = \ \left[\ I \ \ \cdots \ \ I\ \right] \mathrm{Vec}(X)$$

$$\text{onde } \mathrm{Vec}(X) \equiv \begin{bmatrix} X^1 \\ \vdots \\ X^n \end{bmatrix}$$

A variância da média é dada por

$$E(\bar{X}) = E(\ (1/n)\ \mathbf{1}'X\ ) = (1/n)\ \mathbf{1}'E(X) = E(X_1)$$

na ultima equação denotamos a somatória $\sum X$ por $\mathbf{1}'X$, onde $\mathbf{1}' = [1, 1, \ldots 1]$.

A.3.2 Correlação

A correlação entre duas variáveis aleatórias, $Cor(X, Y)$ ou $\rho_{i,j} \equiv\ Cor(X_i, X_j)$, é a covariância "normalizada" pelo desvio padrão:

$$Cor(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \ \text{ ou } \ \rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$$

A correlação várias propriedades que permitem sua interpretação geométrica como ângulos entre variáveis aleatórias:

1) Se duas variáveis aleatórias (de variância limitada), X e Y, são independentes, então $\mathrm{Cov}(X, Y) = 0$, pois $E(XY) = E(X)E(Y)$.

2) Todavia, correlação nula não garante independência! Considere duas variáveis aleatórias definidas sobre os resultados de um dado honesto: X assumindo valor -1

em $F1$, 1 em $F6$, e valor 0 em todas as outras faces; Y assumindo valor 1 em $\{F1, F6\}$, e valor 0 em todas as outras faces. As variáveis aleatórias X e Y não são independentes, embora tenham correlação nula (verifique).

3) No caso de uma dependência linear, $Y = \alpha X + \gamma$, temos que:

$$Cor(Y, X) = \frac{\mathrm{Cov}(\alpha X + \gamma, X)}{\sigma(\alpha X + \gamma)\sigma(X)} = \frac{\alpha \,\mathrm{Var}(X)}{\sqrt{\alpha^2}\sigma_X \sigma_X} = sign(\alpha)$$

onde $sign(x) \equiv \;\; 1$ se $x > 0$, $\;\; -1$ se $x < 0$, e $\;\; 0$ se $x = 0$.

4) A correlação é limitada ao intervalo $-1 \leq \rho_{i,j} \leq 1$:
Tomemos $X = [X_1, X_2]'$, e $Y = [a_1, a_2]X$.

$$\begin{aligned}
\mathrm{Var}(Y) &= \left[\begin{array}{cc} a_1 & a_2 \end{array}\right] \left[\begin{array}{cc} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{array}\right] \left[\begin{array}{c} a_1 \\ a_2 \end{array}\right] \\
&= \left[\begin{array}{cc} a_1 & a_2 \end{array}\right] \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \end{array}\right] \left[\begin{array}{cc} 1 & \rho_{1,2} \\ \rho_{2,1} & 1 \end{array}\right] \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \end{array}\right] \left[\begin{array}{c} a_1 \\ a_2 \end{array}\right] \\
&= \left[\begin{array}{cc} b_1 & b_2 \end{array}\right] \left[\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array}\right] \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right] \\
&= b_1^2 + 2\rho b_1 b_2 + b_2^2
\end{aligned}$$

Observemos agora que, pela definição de variância, $\mathrm{Var}(Y) \geq 0$. Observemos ainda que se $-1 \leq \rho \leq 1$, podemos sempre "completar os quadrados" de modo a reescrever $\mathrm{Var}(Y) = (b_1 \pm b_2)^2 + \mathrm{abs}((1 \pm \rho)b_1 b_2)$, quantidade obviamente positiva. Todavia se $\mathrm{abs}(\rho) > 1$, podemos sempre escolher valores para b_1 e b_2 em $\{-1, 1\}$ que tornam $b_1^2 + 2\rho b_1 b_2 + b_2^2 < 0$, uma contradição.

A.4 Espaços Contínuos

A exposição das seções precedentes é válida para espaços amostrais finitos. Para espaços não finitos, especialmente os não enumeráveis, uma estrutura mais mais complexa é necessária (vide [Billingsley]). Um espaço de probabilidade, $(\Omega, \mathcal{A}, P)$, é um espaço amostral, Ω, uma σ-algebra, $\mathcal{A}$ e uma medida de probabilidade, $P : \mathcal{A} \mapsto [0, 1]$. Uma variável aleatória é uma função $x : \Omega \mapsto \mathcal{R}$, tal que $x^{-1}(t) \in \mathcal{A}$, $\forall t \in \mathcal{R}$.

A distribuição (cumulativa) de uma variável aleatória x, $F : \mathcal{R} \mapsto [0, 1]$, é definida por
$F(t) \equiv$
$Pr(\{\omega \,|\, x(\omega) \leq t\})$.

A esperança de uma variável aleatória, $E(x)$, é definida por

$$E(x) \equiv \int_{-\infty}^{\infty} t dF(t).$$

No caso de uma distribuição diferenciável, ou discreta, temos, respectivamente

$$E(x) = \int tf(t)dt \quad \text{ou} \quad E(x) = \sum tf(t) \ .$$

O k-ésimo **momento** de uma variável aleatória é a esperança de sua k-ésima potência (omitiremos o índice k para $k = 1$), $\mu_k(x) \equiv E(x^k)$. O k-ésimo **momento central** de uma variável aleatória é a esperança da k-ésima potência do desvio em relação a sua esperança, $\mu_k^c(x) \equiv E((x - \mu_1)^k)$. A variância corresponde ao segundo momento central.

Podemos agora considerar um espaço vetorial sobre as variáveis aleatórias (neste Espaço de probabilidade) com segundo momento limitado, $L^2(\Omega, \mathcal{A}, P)$, de elementos $\{x \,|\, E(x^2) < \infty\}$. A origem de L^2 é a variável aleatória identicamente nula, $x \equiv 0$, e o oposto de uma variável aleatória x é $-x = (-1)x$ (explique).

As operações usuais de soma e produto por escalar de variáveis aleatórias está bem definida neste espaço, pois

1. $E((\alpha x)^2) = \alpha^2 E(x^2) < \infty$.
2. $E((x + y)^2) \leq E(2x^2 + 2y^2) \leq 2E(x^2) + 2E(y^2) < \infty$.

Em L^2 adotamos a seguinte definição de produto interno:

$$< X \,|\, Y > \equiv E(XY) \ ,$$

que satisfaz trivialmente as propriedades de simetria, linearidade, e semi-positividade (prove). Algumas tecnicalidades são necessárias para assegurar a positividade deste produto interno.

A.5 Exercícios

1. Formule o problema de mínimos quadrados como um problema de programação quadrática.
   (a) Assuma dada uma base N de $N(A)$.
   (b) Calcule diretamente o resíduo $z = b - y$ em função de A.
2. Prove que a distribuição F é
   (a) Não decrescente, com $F(-\infty) = 0$ e $F(\infty) = 1$.
   (b) Sempre contínua à esquerda e continua à direita exceto num numero enumerável de pontos.

3. Se as variáveis aleatórias x e y têm, respectivamente, distribuições F e G, determine a distribuição de

   (a) αx.

   (b) $x + y$.

   (c) xy.

4. Dadas x e y, variáveis aleatórias, mostre que:

   (a) $E(\alpha x + \beta y) = \alpha E(x) + \beta E(y)$.

   (b) $\operatorname{Var}(\alpha x + \beta y) = \alpha^2 \operatorname{Var}(x) + \beta^2 \operatorname{Var}(y) + 2\alpha\beta \operatorname{Cov}(x, y)$.

5. Dado x, um vetor de variáveis aleatórias, mostre que:

   (a) $E(Ax) = AE(x)$.

   (b) $\operatorname{Cov}(Ax) = A \operatorname{Cov}(x)A'$.

6. O **traço** de uma matriz A é definido por $tr(A) \equiv \sum A_i^i$. Mostre que

   (a) $tr(A + B) = tr(A) + tr(B)$.

   (b) $tr(AB) = tr(BA)$.

   (c) $x'Ay = tr(Ayx') = tr(yx'A)$.

   (d) Se A, $m \times n$, tem posto pleno, $\rho(A) = n$, então $tr(P_A) = n$.

   (e) Nas condições do ítem anterior, definindo $R_A = (I - P_A)$, temos que $tr(R_A) = m - n$.

7. Dado x um vetor de variáveis aleatórias, $E(x) = \mu$, $\operatorname{Cov}(x) = V$, e S uma matriz simétrica, temos que
$$E(x'Sx) = tr(SV) + \mu'S\mu.$$
   Sugestão: Use que $E(x'Sx) = tr\left(SE(xx')\right)$.

8. Num modelo linear de posto pleno, com $\hat{p} = (A'A)^{-1}A'y$ estimando o parâmetro p, mostre que

   (a) $\operatorname{Cov}(\hat{p}) = \sigma^2(A'A)^{-1}$.

   (b) O erro quadrático médio, $MSE \equiv \|y - P_A y\|^2/(m - n)$, é um estimador não tendencioso de σ^2. Sugestão: Use $MSE = y'R_A y/(m-n)$, onde $R_A = (I-P_A)$.

Justificativa da Norma Quadrática:

Existem outras alternativas para medir o tamanho do erro, x, além da norma $L_2 = (x'x)^{1/2}$; Por exemplo as normas $L_1 = \mathbf{1}'\text{abs}(x)$, ou $L_\infty = \max(\,\text{abs}(x)\,)$. Tanto L_1 como L_∞ são usadas na estatística em certas situações especiais, todavia, L_2 é usada na maioria das situações. Além de ser computacionalmente mais simples, é a única que tem a propriedade de ser invariante por uma (pelo grupo de) rotação.

Appendix B

Álgebra Linear Computacional

B.1 Notação e Operações Básicas

Este parágrafo define algumas notações matriciais. Indicamos por $(1:n)$ a lista $[1,\dots n]$, e $j \in (1:n)$ indica que o índice j está neste domínio. Uma lista de matrizes tem um (ou mais) índices superscritos, $S^1 \dots S^m$. Assim $S^k_{h,i}$ é o elemento na linha h e coluna i da matriz S^k. A matriz

$$A = \begin{bmatrix} A^{1,1} & \cdots & A_{1,s} \\ \vdots & \ddots & \vdots \\ A^{r,1} & \cdots & A_{r,s} \end{bmatrix}$$

é uma matriz blocada, onde $A^{p,q}$ é o $p-q$-ésimo bloco, ou sub-matriz.

Quando estamos falando de apenas uma matriz, X, costumamos esceve-la com o índice de linha subscrito, e o índice de coluna superscrito Assim x_i, x^j, e x^j_i são a linha i, a coluna j, e o elemento (i,j) da matriz X. ESta notação é mais compacta, e resalta o fato de vermos a matriz X como uma matriz blocada por vetores coluna. A matriz $X^{j:k}_{h:i}$ é um bloco extraido da matriz X, fazendo os índices de linha e coluna percorrer os domínios indicados. $\mathbf{0}$ e $\mathbf{1}$ são matrizes de zeros e uns, geralmente vetores coluna , $n \times 1$. Quando a dimensão não está indicada, ela pode ser deduzida do contexto. $V > 0$ é uma matriz positiva definida. Definimos a p-norma de um vetor x por $||x||_p = (\sum |x_i|^p)^{-p}$. Assim, se x para um vetor não negativo, podemos escrever sua 1-norma como $||x||_1 = \mathbf{1}'x$.

O produto de Kroneker de duas matrizes é uma matriz blocada onde o bloco (i,j) é a segunda matriz multiblicada pelo elemento $(1,j)$ da primeira matriz:

$$A \otimes B \ = \ \begin{bmatrix} A^1_1 B & A^2_1 B & \cdots \\ A^1_2 B & A^2_2 B & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

As seguintes propriedades podem ser facilmente verificadas:

- $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
- $(A \otimes B)' = A' \otimes B'$
- $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

O operador Vec "empilha" as colunas de uma matriz em um único vetor coluna, i.e., se A é $m \times n$,

$$\mathrm{Vec}(A) = \begin{bmatrix} A^1 \\ \vdots \\ A^n \end{bmatrix}$$

As seguintes propriedades podem ser facilmente verificadas:

- $\mathrm{Vec}(A + B) = \mathrm{Vec}(A) + \mathrm{Vec}(B)$
- $\mathrm{Vec}(AB) = \begin{bmatrix} AB^1 \\ \vdots \\ AB^n \end{bmatrix} = (I \otimes A)\,\mathrm{Vec}(B)$

B.2 Espaços Vetoriais com Produto Interno

Dados dois vetores $x, y \in \mathcal{R}^n$, o seu **produto escalar** é definido como

$$< x \mid y > \equiv x'y = \sum_{i=1}^{n} x_i y^i \ .$$

Com esta definição vê-se que o produto escalar é um operador que satisfaz as propriedades fundamentais de **produto interno**, a saber:

1. $< x \mid y > = < y \mid x >$, simetria.
2. $< \alpha x + \beta y \mid z > = \alpha < x \mid z > + \beta < y \mid z >$, linearidade.
3. $< x \mid x > \geq 0$, semi-positividade.
4. $< x \mid x > = 0 \Leftrightarrow x = 0$, positividade.

Através do produto interno, definimos a norma:

$$\|x\| \equiv < x \mid x >^{1/2} \ ;$$

e definimos também o ângulo entre dois vetores:

$$\Theta(x, y) \equiv \arccos(< x \mid y > / \|x\|\|y\|) \ .$$

B.3 Projetores

Consideremos o subespaço linear gerado pelas colunas de uma matriz A, m por n, $m \geq n$:

$$C(A) = \{y = Ax, x \in \mathcal{R}^n\} \ .$$

Denominamos $C(A)$ de imagem de A, e o complemento de $C(A)$, $N(A)$, de espaço nulo de A,

$$N(A) = \{y \mid A'y = 0\} \ .$$

Definimos a projeção de de um vetor $b \in \mathcal{R}^m$ no espaço das colunas de A, pelas relações:

$$y = P_{C(A)}b \leftrightarrow y \in C(A) \wedge (b - y) \perp C(A)$$

ou, equivalentemente,

$$y = P_{C(A)}b \leftrightarrow y = Ax \wedge A'(b - y) = 0 \ .$$

No que se segue suporemos que A tem posto pleno, i.e. que suas colunas são linearmente independentes. Provemos que o projetor de b em $C(A)$ é dado pela aplicação linear

$$P_A = A(A'A)^{-1}A' \ .$$

Se $y = A((A'A)^{-1}A'b)$, então obviamente $y \in C(A)$. Por outro lado,

$$A'(b - y) = A'(I - A(A'A)^{-1}A')b = (A' - IA')b = 0 \ .$$

B.4 Matrizes Ortogonais

Dizemos que uma matriz quadrada e real é **ortogonal** sse sua inversa é igual a sua transposta.Dada Q uma matriz ortogonal, suas colunas formam uma base ortonormal de $\mathcal{R}^n$, como pode ser visto da identidade $Q'Q = I$. A norma quadrática de um vetor v, ou seu quadrado

$$\|v\|^2 \equiv \sum (v_i)^2 = v'v$$

permanece inalterada por uma transformação ortogonal, pois

$$(Qv)'(Qv) = v'Q'Qv = v'Iv = v'v \ .$$

Dado um vetor em $\mathcal{R}^2$, $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, a rotação deste vetor por um ângulo θ é dada pela transformação linear

$$G(\theta)x = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} .$$

Notemos que uma rotação é uma transformação ortogonal, pois

$$G(\theta)'G(\theta) = \begin{bmatrix} \cos(\theta)^2 + \sin(\theta)^2 & 0 \\ 0 & \cos(\theta)^2 + \sin(\theta)^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} .$$

A rotação de Givens é um operador linear cuja matriz coincide com a identidade, exceto num par de linhas onde imergimos uma matriz de rotação bidimensional:

$$G(i,j,\theta) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \cos(\theta) & & \sin(\theta) & & \\ & & & \ddots & & & \\ & & -\sin(\theta) & & \cos(\theta) & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} .$$

Dizemos que a aplicação deste operador numa matriz A, $G'A$, roda as linhas i e j de A de um ângulo θ (no sentido ant-horário).

Como o produto de transformações ortogonais continua ortogonal (prove), podemos usar uma seqüência de rotações para construir transformações ortogonais.

Listamos agora algumas rotações em $\mathcal{R}^2$, construidas a partir de simples relações trigonométricas, que serão utilizadas como blocos de construção para diversos agoritmos.

Consideremos, em $\mathcal{R}^2$, v um vetor, S uma matriz simétrica, e A uma matriz assimétrica,

$$v = \begin{bmatrix} x \\ y \end{bmatrix} , \quad S = \begin{bmatrix} p & q \\ q & r \end{bmatrix} , \quad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Para anular a segunda componente do vetor v por uma rotação á esquerda, $G(\theta_v)'\, v$, basta tomar um ângulo

$$\theta_v = \arctan\left(\frac{y}{x}\right) .$$

Para diagonalizar a matriz simétrica por uma rotação simétrica, $G(\theta_{diag})'\, S\, G(\theta_{diag})$, basta tomar um ângulo

$$\theta_{diag} = \frac{1}{2}\arctan\left(\frac{2q}{r-p}\right) .$$

Para simetrizar a matriz assimétrica por uma por uma rotação à esquerda, $G(\theta_{sym})'\, A$, basta tomar um ângulo

$$\theta_{sym} = \arctan\left(\frac{b-c}{a+d}\right) .$$

Assim, para diagonalizar a matriz assimétrica basta simetriza-la, e em seguida diagonaliza-la. Alternativamente, podemos diagonalizar a matriz assimétrica por um par de "rotações de Jacobi", à esquerda e à direita, $J(\theta_r)'\, A\, J(\theta_l)$; bastando tomar os ângulos

$$\theta_{sum} = \theta_r + \theta_l = \arctan\left(\frac{c+b}{d-a}\right) \ , \quad \theta_{dif} = \theta_r - \theta_l = \arctan\left(\frac{c-b}{d+a}\right) \quad \text{ou}$$

$$J(\theta_r)' = G(\theta_{sum}/2)'\, G(-\theta_{dif}/2)' \ , \quad J(\theta_l) = G(\theta_{dif}/2)\, G(\theta_{dif}/2) \ .$$

No cálculo das rotações, as funções trigonométricas, Seno, Coseno e Arco-Tangente não são realmente utilizadas, já que nunca utilizamos os ângulos propriamente ditos, mas apenas $c = \sin(\theta)$ e $s = \sin\theta$, que podemos computar diretamente como

$$c = \frac{x}{\sqrt{x^2+y^2}} \ , \quad s = \frac{-y}{\sqrt{x^2+y^2}} \ .$$

Para prevenir overflow podemos utilizar o cálculo:

- Se $y == 0$, então $c = 1$, $s = 0$.
- Se $y \geq x$, então $t = -x/y$, $s = 1/\sqrt{1+t^2}$, $c = st$.
- Se $y < x$, então $t = -y/x$, $c = 1/\sqrt{1+t^2}$, $s = ct$.

B.5 Fatoração QR

Dada A uma matriz real de posto pleno $m \times n$, $m \geq n$, podemos sempre encontrar uma matriz ortogonal Q tal que $A = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$, onde R é uma matriz quadrada e triangular superior. Esta decomposição é dita uma fatoração "QR", ou fatoração ortogonal, da matriz A. O fator ortogonal $Q = [C \mid N]$ nos da uma base ortonormal de $\mathcal{R}^m$ onde as n primeiras colunas são uma base ortonormal de $C(A)$, e as $m-n$ últimas colunas são uma base de $N(A)$, como pode ser visto diretamente da identidade $Q'A = \begin{bmatrix} R \\ 0 \end{bmatrix}$. Construiremos a seguir um método para fatoração ortogonal.

Abaixo ilustramos uma seqüência de rotações de linhas necessárias que leva uma matriz 5×3 à forma triangular superior. Cada par de índices, (i, j), indica que rodamos estas linhas do ângulo apropriado para zerar a posição na linha i, coluna j. Supomos que inicialmente a matriz é densa, i.e. todos os seus elementos são diferentes de zero, e ilustramos o padrão de esparsidade da matriz nos estágios assinalados com um asterisco na seqüência de rotações.

$$(1,5) * (1,4)(1,3)(1,2) * (2,5)(2,4)(2,3) * (3,5)(3,4) *$$

$$\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ 0 & x & x \end{bmatrix} \quad \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \end{bmatrix} \quad \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & x \\ 0 & 0 & x \end{bmatrix} \quad \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

B.5.1 Mínimos Quadrados

Dado um sistema superdeterminado, $Ax = b$ onde a matriz A $m \times n$ tem $m > n$, dizemos que x^* "resolve" o sistema no sentido dos mínimos quadrados, ou que x^* é a "solução" de mínimos quadrados, sse x^* minimiza a norma quadrática do resíduo,

$$x^* = Arg \min_{x \in \mathcal{R}^n} \|Ax - b\| \ ,$$

Dizemos também que $y = Ax^*$ é a melhor aproximação, no sentido dos mínimos quadrados de b em $C(A)$.

Como a multiplicação por uma matriz ortogonal deixa inalterada a norma quadrática de um vetor, podemos procurar a solução deste sistema (no sentido dos mínimos quadrados) minimizando a transformação ortogonal do resíduo usada na fatoração QR de A,

$$\|Q'(Ax - b)\|^2 = \| \begin{bmatrix} R \\ 0 \end{bmatrix} x - \begin{bmatrix} c \\ d \end{bmatrix} \|^2 = \|Rx - c\|^2 + \|0x - d\|^2.$$

Da última expressão vê-se que a solução, a aproximação e o resíduo do problema original são dados por, respectivamente

$$x^* = R^{-1}c \ , \quad y = Ax^* \ \text{ e } \ z = Q \begin{bmatrix} 0 \\ d \end{bmatrix}.$$

Como já havíamos observado, as $m - n$ últimas colunas de Q formam uma base ortonormal de $N(A)$, logo $z \perp C(A)$, de modo que concluímos que $y = P_A b$!

B.6 Fatorações LU e Cholesky

Dada uma Matriz A, a **operação elementar** determinada pelo **multiplicador** m_j^i, é subtrair da linha j a linha i multiplicada por m_j^i. A operação elementar aplicada a matriz

identidade gera a correspondente **matriz elementar**,

$$M(i,j) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & i \\ & & \vdots & \ddots & & & \\ & & -m_j^i & & 1 & & j \\ & & \vdots & & & \ddots & \\ & & & & & & 1 \end{bmatrix} .$$

Aplicar uma operação elementar a matriz A equivale a multiplicá-la a esquerda pela correspondente matriz elementar.

Na **Fatoração de Gauss**, ou fatoração LU, usamos uma seqüência de operações elementares para levar A a forma triangular superior,

$$MA = M(n-1,n)\cdots M(2,3)\cdots M(2,n)M(1,1)\cdots M(1,n-1)M(1,n)A = U \ .$$

A inversa da produtória desta seqüência de matrizes elementares tem a forma triangular inferior (verifique)

$$L = M^{-1} = \begin{bmatrix} 1 & & & & \\ m_2^1 & 1 & & & \\ \vdots & \vdots & \ddots & & \\ m_{n-1}^1 & m_{n-1}^2 & & 1 & \\ m_n^1 & m_n^2 & \cdots & m_n^{n-1} & 1 \end{bmatrix} .$$

No caso de fatorarmos uma matriz simétrica, $S = LU$, podemos por em evidencia os elementos diagonais de U obtendo $S = LDL'$. Se S for positiva definida assim será D, de modo que podemos escrever $D = D^{1/2}D^{1/2}$, $D^{1/2}$ a matriz diagonal contendo a raíz dos elementos em D. Definindo $C = LD^{1/2}$, temos $S = CC'$, a **fatoração de Choleski** de S.

B.6.1 Programação Quadrática

O problema de **programação quadrática** consiste em minimizar a função

$$f(y) \equiv (1/2)y'Wy + c'y \ , \quad W = W'$$

sujeitos às **restrições**

$$g_i(y) \equiv N_i'y = d_i.$$

Os gradientes de f e g_i são dados, respectivamente, por

$$\nabla_y f = y'W + c' \ , \text{ e } \nabla_y g_i = N_i' \ .$$

As **condições de otimalidade** de primeira ordem (condições de Lagrange) estabelecem que as restrições sejam obedecidas, e que o gradiente da função sendo minimizada seja uma combinação linear dos gradientes das restrições. Assim a solução pode ser obtida em função do **multiplicador de Lagrange**, i.e. do vetor l de coeficientes desta combinação linear, como

$$N'y = d \ \wedge \ y'W + c' = l'N' \ ,$$

ou em forma matricial,

$$\begin{bmatrix} N' & 0 \\ W & N \end{bmatrix} \begin{bmatrix} y \\ l \end{bmatrix} = \begin{bmatrix} d \\ c \end{bmatrix} \ .$$

Este sistema de equações é conhecido como o **sistema normal**. O sistema normal tem por matriz de coeficientes uma matriz simétrica. Se a forma quadrática W for **positiva definida**, i.e.se $\forall x \ x'Wx \geq 0 \ \wedge \ x'Wx = 0 \Leftrightarrow x = 0$, e as restrições N forem lineramente independentes, a matriz de coeficientes do sistema normal será também positiva definida.

B.7 Fatoração SVD

A fatoração SVD decompõem uma matriz real A, $m \times n$, $m \geq n$, em um produto $D = U'AV$, onde D é diagonal, e U, V são matrizes ortogonais. Consideremos primeiramente o caso $m = n$, i.e. uma matriz quadrada.

O algoritmo de Jacobi é um algoritmo iterativo que, a cada iteração, "concentra a matriz na diagonal", através de rotações de Jacobi.

$$J(i,j,\theta_r)' \, A^k \, J(i,j,\theta_l) = A^{k+1} = \begin{bmatrix} A^{k+1}_{1,1} & \cdots & A^{k+1}_{1,i} & \cdots & A^{k+1}_{1,j} & \cdots & A^{k+1}_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A^{k+1}_{i,1} & \cdots & A^{k+1}_{i,i} & \cdots & 0 & \cdots & A^{k+1}_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A^{k+1}_{j,1} & \cdots & 0 & \cdots & A^{k+1}_{j,j} & \cdots & A^{k+1}_{j,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A^{k+1}_{n,1} & \cdots & A^{k+1}_{n,i} & \cdots & A^{k+1}_{n,j} & \cdots & A^{k+1}_{n,n} \end{bmatrix}$$

Consideremos a soma dos quadrados dos elementos fora da diagonal na matriz A $\mathrm{Off}_2(A)$. Vemos que

$$\mathrm{Off}_2(A^{k+1}) = \mathrm{Off}_2(A^k) - (A^k_{i,j})^2 - (A^k_{j,i})^2$$

Assim, escolhendo a cada iteração o par de índices que maximiza a soma dos quadrados do par fora da diagonal a ser anulado, temos um algoritmo que converge linearmente para uma matriz diagonal.

O algoritmo de Jacobi nos dá uma prova construtiva da existência da fatoração SVD, e é a báse para vários algoritmos mais eficientes de fatoração SVD.

Se A é uma matriz retangular, basta inicialmente fatorar $A = QR$, e aplicar o algoritmo de Jacobi ao bloco quadrado superior de R. Se A é quadrada e simétrica, a fatoração obtida é denominada decomposição de autovalores de A.

As matrizes U e V podem ser interpretadas como bases ortogonais dos respectivos espaços de dimensão m e n. Os valores na diagonal de S são denominados valores singulares da matriz A, e podem ser interpretados geométricamente como fatores multiplicadores do mapa $A = UDV'$, que leva cada versor da base V para um múltiplo de um versor da báse U.

B.8 Exercícios

1. Use as propriedades fundamentais do produto interno para provar:

   (a) A desigualdade de Cauchy-Scwartz: $| < x \mid y > | \leq \|x\|\|y\|$. Sugestão: Calcule $\|x - \alpha y\|^2$ para $\alpha =< x \mid y >^2 / \|y\|$.

   (b) A Desigualdade Triagular: $\|x + y\| \leq \|x\| + \|y\|$.

   (c) Em que caso temos igualdade na desigualdade de Cauchy-Schwartz? Relacione sua resposta com a definição de ângulo entre vetores.

2. Use a definição do produto interno em $\mathcal{R}^n$ para provar a Lei do Paralelogramo: $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$.

3. Uma matriz é idempotente,ou um projetor não ortogonal sse $P^2 = P$. Prove que:

   (a) $R = (I - P)$ é idempotente.

   (b) $\mathcal{R}^n = C(P) + C(R)$.

   (c) Todos os autovalores de P são 0 ou +1. Sugestão: Mostre que se 0 é uma raiz do polinômio característico de P, $\varphi_P(\lambda) \equiv \det(P - \lambda I)$, então $(1 - \lambda) = 1$ é raiz de $\varphi_R(\lambda)$.

4. Prove que $\forall P$ idempotente e simétrico, $P = P_{C(P)}$. Sugestão: Mostre que $P'(I - P) = 0$.

5. Prove que o operador de projeção num dado sub-espaço vetorial V, P_V, é único e simétrico.

6. Prove o theorema de Pitágoras: $\forall b \in \mathcal{R}^m, u \in V$ temos que $\|b - u\|^2 = \|b - P_V b\|^2 + \|P_V b - u\|^2$.

7. Suponha termos a fatoração QR de uma matriz A. Considere uma nova matriz $\tilde{A}$ obtida de A pela substituição de uma única coluna. Como podemos atualizar nossa

fatoração ortogonal usando apenas $3n$ rotações de linha? Sugestão: (a) Remova a coluna alterada de A e atualize a fatoração usando no máximo n rotações. (b) Compute a nova coluna alterada pelo fator ortogonal corrente, $\tilde{a} = Q'a = R^{-t}A'a$. ($c$) Adicione $\tilde{a}$ como a ultima coluna de $\tilde{A}$, e torne a atualizar a fatoração com $2n$ rotações.

8. Compute as fatorações LDL e Cholesky da matriz

$$\begin{bmatrix} 4 & 12 & 8 & 12 \\ 12 & 37 & 29 & 38 \\ 8 & 29 & 45 & 50 \\ 12 & 38 & 50 & 113 \end{bmatrix}.$$

9. Prove que

   (a) $(AB)' = B'A'$.

   (b) $(AB)^{-1} = B^{-1}A^{-1}$.

   (c) $A^{-t} \equiv (A^{-1})' = (A')^{-1}$.

10. Descreva quatro algoritmos, para computar $L^{-1}x$ e $L^{-t}x$, acessando a matriz L, triangular inferior de diagonal unitária, por linha ou por coluna.