Diagramas de influência e teoria estatística

Rafael Bassi Stern

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística
Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq e da FAPESP

São Paulo, dezembro de 2008

Diagramas de influência e Teoria Estatística

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Rafael Bassi Stern
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof. Dr. Carlos Alberto de Bragança Pereira (orientador) - IME-USP.
- Prof. Dr. Sérgio Wechsler - IME-USP.
- Prof. Dr. Fábio Gagliardi Cozman - EP-USP.

"La théorie des probabilités n'est, au fond,
que le bon sens réduit au calcul;
elle fait apprécier avec exactitude
ce que les esprits justes
sentent par une sorte d'instinct,
sans qu'ils puissent souvent s'en rendre compte."
— Pierre-Simon Laplace

Agradecimentos

Agradeço ao meu orientador, Professor Carlinhos, por todos os ensinamentos que recebi e conversas que tivemos. Foi uma honra aprender com pessoa que tem tamanha intuição para a Estatística e a vida.

Agradeço aos meus pais e às minhas irmãs que sempre me apoiaram de todas as formas possíveis durante o desenvolvimento deste trabalho.

Agradeço ao falecido Professor Flávio Wagner Rodrigues que, ainda nos tempos de colégio, teve a paciência para me ensinar Análise Combinatória e Probabilidade. Lembro dessas aulas com extrema afeição até hoje.

Agradeço aos Professores que tive durante a Graduação e o Mestrado e, em especial, a Adilson Simonis, Antonio Galves, Elizabeti Kira, Fabio Gagliardi Cozman, Luis Gustavo Esteves, Pablo Augusto Ferrari, Sérgio Wechsler e Vladimir Belitsky, cujos ensinamentos foram essenciais para a conclusão desta pesquisa.

Agradeço a todos os meus amigos que contribuíram direta ou indiretamente na elaboração desta dissertação. Em especial, agradeço a Fernando Vieira Bonassi pelas profundas discussões a respeito dos mais diversos temas e por ter auxiliado na correção e formatação do texto. Agradeço a Estéfano Alves de Souza que auxiliou na revisão dos teoremas apresentados. Agradeço a Alexandre Galvão Patriota e Paulo Cilas Marques Filho pelas discussões a respeito do Princípio da Verossimilhança. Agradeço a João Vinicius França Carvalho pelas discussões que tivemos e pelo apoio durante todo o Mestrado.

Agradeço aos inestimáveis comentários e sugestões da Banca Examinadora que em muito auxiliaram no aperfeiçoamento da dissertação.

Resumo

O objetivo principal deste trabalho foi analisar o controverso conceito de informação em estatística. Para tal, primeiramente foi estudado o conceito de informação dado por Basu. A seguir, a análise foi dividida em três partes: informação nos dados, informação no experimento e diagramas de influência. Nas duas primeiras etapas, sempre se tentou definir propriedades que uma função de informação deveria satisfazer para se enquadrar ao conceito.

Na primeira etapa, foi estudado como o princípio da verossimilhança é uma classe de equivalência decorrente de acreditar que experimentos triviais não trazem informação. Também foram apresentadas métricas que satisfazem o princípio da verossimilhança e estas foram usadas para avaliar um exemplo intuitivo.

Na segunda etapa, passamos para o problema da informação de um experimento. Foi apresentada a relação da suficiência de Blackwell com experimentos triviais e o conceito usual de suficiência. Também foi analisada a equivalência de Blackwell e a sua relação com o Princípio da Verossimilhança anteriormente estudado. Finalmente, as métricas apresentadas para medir a informação de conjuntos de dados foram adaptadas para também medir a informação de um experimento.

Finalmente, observou-se que nas etapas anteriores uma série de simetrias mostraram-se como elementos essenciais do conceito de informação. Para ganhar intuição sobre elas, estas foram reescritas através da ferramenta gráfica dos diagramas de influência. Assim, definições como suficiência, suficiência de Blackwell, suficiência mínima e completude foram reapresentadas apenas usando essa ferramenta.

Palavras-chave: Estatística Bayesiana, Informação, Diagramas de Influência, Suficiência de Blackwell, Análise Pré-Posteriori.

Abstract

The main objective of this work is to analyze the controversial concept of information in Statistics. To do so, firstly the concept of information according to Basu was presented. Next, the analysis was divided in three parts: information in a data set, information in an experiment and influence diagrams. In the two first parts, we always tried to define properties an information function should satisfy in order to be in accordance to the concept of Basu.

In the first part, it was studied how the likelihood principle is an equivalence class which follows from believing that trivial experiments do not bring information. Metrics which satisfy the likelihood principle were also presented and used to analyze an intuitive example.

In the second part, the problem became that of determining information of a particular experiment. The relation between Blackwell's sufficiency, trivial experiments and classical sufficiency was presented. Blackwell's equivalence was also and analyzed and its relationship with the Likelihood Principle was exposed. Finally, the metrics presented to evaluate the information in a data set were adapted to do so with experiments.

Finally, in the last parts a number of symmetries were shown as essencial elements of the concept of information. To gain more intuition about these elements, we tried to rewrite them using the graphic tool of influence diagrams. Therefore, definitions as sufficiency, Blackwell's sufficiency, minimal sufficiency and completeness were shown again, only using influence diagrams.

Keywords: Bayesian Statistics, Information, Influence Diagrams, Blackwell Sufficiency, Pre-Posterior Analysis.

Sumário

1 Introdução **1**
1.1 O Conceito de Informação 1
1.2 Definições 3

2 Informação num Conjunto de Dados **5**
2.1 Simetrias na Informação 5
2.2 Funções de Informação 8

3 Informação num Experimento **11**
3.1 Suficiência de Blackwell 11
3.2 Simetrias na Informação de um Experimento 14
3.3 Funções de Informação num Experimento 18

4 Diagramas de Influência **21**
4.1 Definição e Propriedades 21
4.2 Diagramas de Influência aplicados à Teoria Estatística 22

5 Conclusões **27**

Referências Bibliográficas **29**

Capítulo 1

Introdução

1.1 O Conceito de Informação

O conceito de informação é talvez um dos mais controversos na Estatística. Uma pessoa pode encontrar inúmeras diferentes medidas de quanta informação um experimento ou um conjunto de dados traz, cada uma dela tendo propriedades que a aproximam do conceito de informação. Parece importante explorar essas abordagens pois, no final das contas, extrair informação parece ser um dos maiores objetivos da Estatística. No entanto, antes de prosseguir, pretendemos limitar o objeto de nosso estudo e também estudar mais a fundo o que seria o conceito de informação.

Para fazê-lo, tentaremos seguir a definição intuitiva de informação dada por Basu [2]: "Informação é aquilo que ela faz por você, ela muda sua opinião". Como conseqüência natural dessa afirmação parecem seguir quatro perguntas (Pereira [8]):

- Informação sobre o quê?
- Aonde está a informação?
- Como a informação é extraída?
- Quanta informação é usada?

Estamos interessados em definir informação a respeito de um parâmetro de interesse, θ, cujos possíveis valores pertencem ao conjunto Θ. O parâmetro representa algo sobre o qual temos incerteza. Por exemplo, alguém pode estar interessado no número de dias em que irá chover em um determinado ano. Nesse exemplo, θ poderia ser esse número e Θ todos os números naturais menores que 366.

A seguir, tentamos responder onde está a informação sobre o parâmetro de interesse. Primeiramente, é importante notar que definindo o conjunto Θ já estamos demonstrando ter conhecimento a respeito de θ. No examplo, informamos que todo ano tem, ao máximo, 366 dias e portanto θ deve ser menor que esse número. No entanto, se soubéssemos que o ano não era bissexto, de Θ poderia ser excluído o valor 366. Além de determinar Θ, o observador pode acreditar que alguns valores nesse conjunto são mais prováveis que outros antes de realizar o experimento. Essa informação é dada pela distribuição a priori para θ. Finalmente, o estatístico geralmente está interessado na realização de um experimento relacionado ao parâmetro de interesse. Dessa forma, também existe informação a respeito de θ nos dados coletados. A pergunta natural parece ser: Como se extrai a informação dos dados?

Adotaremos como forma de extrair a informação o Teorema de Bayes. Utilizando-o, podemos determinar a distribuição a posteriori para θ quando observamos algum conjunto de dados. Finalmente, é possível perguntar: Quanto de informação é extraída?

No segundo capítulo, tentamos responder quão informativo é um conjunto de dados. Para fazê-lo apresentamos duas simetrias: O Princípio da Condicionalidade e o Princípio da Suficiência. A seguir, com base no conceito de informação apresentado, mostramos que essas simetrias são razoáveis para uma função de informação. Como decorrência dessa análise chegamos ao Princípio da Verossimilhança. A seguir, apresentamos algumas métricas compatíveis com esse princípio e estudamos sua aplicação a um exemplo intuitivo.

No terceiro capítulo, tendo estudado a questão da informação de um conjunto de dados, passamos a uma pergunta relacionada: Como obter o máximo possível de informação? Isso é, como determinar o experimento que trará maior quantidade de informação? Para tal, primeiramente definimos o que é a suficiência de Blackwell e apresentamos alguns exemplos desta. A seguir, definimos a Equivalência de Blackwell e mostramos como essa se relaciona ao Princípio da Verossimilhança. Finalmente, voltamos às métricas definidas para conjunto de dados e as adaptamos para experimentos pelo uso da Teoria da Decisão. Para tal, consideramos como a informação de um experimento a sua informação esperada.

Por último, no quarto capítulo, observamos que existem de definições essenciais ao conceito de informação que foram apresentadas anteriormente. Para ganhar intuição adicional a respeito dessas definições, estudamo-as novamente por um novo enfoque: o dos diagramas de influência. Para tal,

primeiramente esses diagramas são definidos e suas principais operações apresentadas. A seguir, definições equivalentes às de suficiência, suficiência de Blackwell, suficiência mínima e completude são apresentadas apenas usando diagramas de influência. Como conseqüência, esperamos ganhar maior intuição a respeito dessas propriedades.

1.2 Definições

Nos capítulos seguintes, estudaremos definições de informação relativas a conjuntos de dados e também a experimentos estatísticos. No entanto, antes de fazê-lo é importante apresentar algumas definições como, por exemplo, a de experimento estatístico. Essas definições estarão em conformidade com aquelas dadas em Basu [3].

Um experimento estatístico ϵ é definido como uma tripla (χ, Θ, P). χ corresponde ao espaço amostral, isto é, o conjunto de todos os possíveis dados que podem ser observados como resultado de uma performance do experimento. Θ corresponde ao conjunto de todos possíveis valores que o parâmetro de interesse, θ, pode assumir. Consideraremos apenas experimentos estatísticos nos quais χ e Θ são enumeráveis. Dessa forma, associaremos a cada um deles a σ-álgebra discreta, isto é, aquela aos quais pertencem todos os subconjuntos de χ e de Θ, respectivamente. Finalmente, P é a função de probabilidade, um mapa $P : \chi \times \Theta \mapsto [0, 1]$ que satisfaz a propriedade $\sum_{x \in \chi} P(x|\theta) = 1$ para todo $\theta \in \Theta$.

Definimos também que não existe dado $x \in \chi$ tal que $P(x|\theta) = 0, \forall \theta \in \Theta$. Parece razoável fazer essa suposição uma vez que o espaço amostral é discreto e, portanto, pode-se imaginar pontos desse tipo como, por exemplo, tirar um 7 na rolagem de um dado de 6 faces.

Definimos uma estatística, T, como um mapa $T : \chi \mapsto \tau$, sendo τ um conjunto qualquer. Isso nos leva à definição do experimento estatístico induzido por T, $\epsilon^T = (\tau, \Theta, P^T)$. A função de probabilidade é dada por $P^T(t|\theta) = P(x : T(x) = t|\theta)$. Finalmente, também caracterizamos o experimento residual de ϵ ao observar que $T = t$, $\epsilon_t^T = (\chi_t, \Theta, P_t^T)$. χ_t corresponde ao conjunto de dados $x \in \chi$ tais que $T(x) = t$. A função de probabilidade é dada por $P_t^T(x|\theta) = P(x|\theta, T(x) = t)$.

É simples demonstrar que para toda estatística T, $P(x|\theta) = P^T(T(x)|\theta) P_{T(x)}^T(x|\theta)$. Portanto, é possível imaginar que a realização de um experimento ϵ consiste em observar primeiramente ϵ^T e, a seguir, observar ϵ_t^T, onde t é o resultado do primeiro experimento e o resultado do experi-

mento combinado é aquele dado por ϵ_t^T. Pela afirmação inicial, observamos que esse experimento combinado tem mesma função de probabilidade que ϵ.

Finalmente, um particular tipo de experimento estatístico é denominado trivial. Um experimento trivial é aquele no qual, para todo $x \in \chi$, a função de verossimilhança, $L_x(\theta) : \Theta \mapsto [0, 1]$, $L_x(\theta) = P(x|\theta)$, é constante. Por exemplo, é razoável supor que o resultado do lançamento de uma moeda não tem relação com o número de dias que irá chover durante um ano. Isso leva à conclusão de que a probabilidade de uma pessoa obter coroa é a mesma, não importando o número de dias que chova durante esse ano. Portanto, o lançamento de moeda é trivial para o parâmetro de interesse número de dias em que chove durante o ano. A idéia de que a probabilidade de obter coroa é igual para qualquer número de dias que chova durante o ano é dada pela função de verossimilhança constante.

Capítulo 2

Informação num Conjunto de Dados

Nesse capítulo, o objeto do estudo será a informação que se ganha observando um particular conjunto de dados num experimento estatístico. Para responder a essa pergunta, sempre se volta aos conceitos apresentados no primeiro capítulo. Em um primeiro momento, apresentamos simetrias na informação que culminam no Princípio da Verossimilhança. A seguir, estudamos funções de informação que satisfazem esse Princípio aplicadas a um exemplo intuitivo.

2.1 Simetrias na Informação

Nessa seção estudaremos funções que têm como parâmetros o experimento estatístico realizado e o conjunto de dados observado. Mais especificamente, apresentaremos propriedades que uma função desse tipo deveria ter para que esteja de acordo com o conceito de informação. Uma vez que não haverá necessidade para definir o contra-domínio dessa função, a princípio, deixaremo-lo indefinido.

Contra essa abordagem, poder-se-ia argumentar que tal função deveria depender da distribuição a priori para se adeqüar ao conceito de informação. Essa crítica é pertinente, no entanto, uma vez que as simetrias estudadas nessa seção não tem relação com a distribuição a priori, consideraremos que os resultados obtidos valem para qualquer particular distribuição a priori fixada.

Para nos referirmos a uma função genérica do tipo estudado utilizaremos a notação $Inf(\epsilon, x)$, onde $\epsilon = (\chi, \Theta, P)$ é um experimento estatístico e x é um resultado possível desse experimento. Seguiremos com uma abordagem semelhante àquela dada por Basu [3] e Birbaum [5].

Uma estatística $T : \chi \mapsto \tau$ é definida como sendo suficiente se $\forall t \in \tau$, ϵ_t^T é um experimento

trivial. Como foi visto anteriormente, um experimento ϵ pode ser encarado como o resultado de ϵ_t^T, onde t é o resultado de ϵ^T. Anteriormente, também foi discutido que há razões para para que se acredite que um experimento trivial não traz informação sobre o parâmetro de interesse. Portanto, uma vez que ϵ_t^T é um experimento trivial para todo t, parece razoável que toda a informação a respeito de ϵ é ganha pela observação inicial de ϵ^T. Essa idéia corresponde ao princípio da suficiência. De acordo com ele, se Inf é uma função de acordo com o conceito de informação, então para toda estatística suficiente T, $\forall x \in \chi, \forall y \in \chi$, se $T(x) = T(y)$ então $Inf(\epsilon, x) = Inf(\epsilon, y)$. Ainda que esse princípio às vezes não seja diretamente mencionado, ele é seguido pela maioria dos estatísticos. Isso é, responder a questões de inferência sobre o parâmetro de interesse apenas observando uma estatística suficiente ou o conjunto de dados leva a maioria dos estatísticos aos mesmos resultados.

Um outro princípio bastante freqüente é o da condicionalidade. Esse princípio pode ser visto como recíproco ao da suficiência. Enquanto que no último dizemos que um experimento trivial realizado após ϵ^T não nos traz informação, no primeiro dizemos que um experimento trivial realizado para terminar um dentre vários possíveis experimentos posteriores também não traz informação a respeito. Para descrever esse princípio nós chamaremos de $\epsilon_0 = (\{1, 2\}, \Theta, P_0)$ algum experimento trivial. Por outro lado, consideraremos $\epsilon_1 = (\chi_1, \Theta, P_1)$, e $\epsilon_2 = (\chi_2, \Theta, P_2)$ dois outros experimentos estatísticos quaisquer. O experimento $\epsilon_{1,2}$ corresponde à mistura entre ϵ_1 e ϵ_2 e é realizada da seguinte forma. Se o resultado de ϵ_0 for 1 então observamos o resultado de ϵ_1, se for 2 então observamos o resultado de ϵ_2. Portanto, o espaço amostral de $\epsilon_{1,2}$ é $\{1, 2\} \times (\chi_1 \cup \chi_2)$. O Princípio da Condicionalidade determina que para qualquer experimento trivial ϵ_0 e dois quaisquer outros experimentos ϵ_1 e ϵ_2 temos que se Inf é uma função de acordo com conceito de informação então $Inf(\epsilon_{1,2}, (i, x)) = Inf(\epsilon_i, x)$, $\forall i \in \{1, 2\}$. Esse princípio é mais controverso que o da suficiência, como veremos até o fim da seção.

O último princípio que será apresentado é o Princípio da Verossimilhança. De acordo com ele, se quaisquer dois possíveis conjuntos de dados tem verossimilhança proporcional então eles trazem mesma informação sobre o parâmetro. Portanto, esse princípio especifica que se Inf está de acordo com o conceito de informação, então para todo $\epsilon_1 = (\chi_1, \Theta, P_1)$, para todo $\epsilon_2 = (\chi_2, \Theta, P_2)$, qualquer $x \in \chi_1$ e qualquer $y \in \chi_2$, se $L_x(\theta) \propto L_y(\theta)$ então $Inf(\epsilon_1, x) = Inf(\epsilon_2, y)$. É fácil demonstrar que o Princípio da Verossimilhança é mais forte que o Princípio da Suficiência e do que o da

Condicionalidade. No entanto, pode parecer que os Princípios da Suficiência e Condicionalidade juntos ainda são mais fracos que o da Verossimilhança uma vez que nesse último não há qualquer referência explícita a experimentos triviais. Birbaum [5] apresentou a primeira demonstração de que os Princípios da Suficiência e Condicionalidade implicam o Princípio da Verossimilhança. Nessa seção seguiremos estreitamente o método empregado em Basu [3] para demonstrar esse resultado.

TEOREMA: Os Princípios da Suficiência e Condicionalidade implicam o Princípio da Verossimilhança.

Demonstração: Consideremos x e y possíveis resultados de, respectivamente, experimentos ϵ_1 com espaço amostral χ_1 e ϵ_2 com espaço amostral χ_2, tal que $L_x(\theta) \propto L_y(\theta)$. Definimos uma estatística $T : \{1,2\}\times(\chi_1\cup\chi_2) \mapsto (\{1,2\}\times(\chi_1\cup\chi_2))\cup\{0\}$ tal que: 1)$T(i,z) = (i,z)$, se $(i,z) \neq (1,x)$ e $(i,z) \neq (2,y)$, 2)$T(i,z) = 0$, se $(i,z) = (1,x)$ ou $(i,z) = (2,y)$. Definimos $\epsilon_0 = (\{1,2\},\Theta,P)$ tal que $P(1) = P(2) = 1/2$. ϵ_0 é um experimento trivial e pode ser utilizado para definir o experimento misto $\epsilon_{1,2}$ como definido acima. É simples demonstrar que T é uma estatística suficiente em $\epsilon_{1,2}$. Portanto, pelo Princípio da Suficiência, $Inf(\epsilon_{1,2},(1,x)) = Inf(\epsilon_{1,2},(2,y))$. A seguir, por condicionalidade, $Inf(\epsilon_1,x) = Inf(\epsilon_{1,2},(1,x))$ e $Inf(\epsilon_2,y) = Inf(\epsilon_{1,2},(2,y))$. Como x, y, ϵ_1 e ϵ_2 eram arbitrários, segue o Princípio da Verossimilhança.

Portanto, se um pesquisador acredita no Princípio da Verossimilhança, ele pode realizar inferência sobre parâmetro de interesse apenas considerando a função de verossimilhança do conjunto de dados observado. Esse Princípio, no entanto, não é seguido por todos os Estatísticos. Um exemplo interessante pode ser encontrado em Wechsler et al. [9]. Nesse exemplo, uma pessoa observa 3 fracassos numa seqüência de 10 testes realizados em dois experimentos distintos. No primeiro experimento a pessoa fixou que realizaria 10 testes e acredita que a distribuição do número de sucesso é binomial. No segundo experimento a pessoa fixou que realizaria testes até que ocorressem 3 fracassos e acredita que a distribuição do número de testes realizados é binomial negativa. Ainda que

o p-valor para cada uma dessas observações em geral seja diferente, eles levam à mesma distribuição a posteriori. Dessa forma, uma vez que os Estatísticos clássicos seguem o Príncipio da Suficiência e não seguem o Princípio da Verossimilhança, é possível concluir que eles não seguem o Princípio da Condicionalidade. Por outro lado, como foi verificado em Wechsler et al. [9] Estatísticos Bayesianos geralmente seguem todos os três princípios.

Acreditamos que é razoável assumir que um experimentro trivial não deveria trazer informação adicional a respeito de um parâmetro de interesse. Nessa seção, esperamos ter mostrado que o Princípio da Verossimilhança é bastante relacionado a essa crença. Portanto, ao procurar por definições mais específicas de funções informação na próxima seção, o Princípio da Verossimilhança será assumido. Assim, fixado todo mais, a informação dada pela observação de diferentes conjuntos de dados não fornecerá valores diferentes quando estes tiverem a mesma função de verossimilhança

2.2 Funções de Informação

Nessa seção estamos interessados em estudar algumas funções freqüentemente consideradas como representantes da informação ganha observando um particular conjunto de dados. Na seção anterior foi visto que é razoável que tal função não forneça valores diferentes para pontos com a mesma verossimilhança. No entanto, essa propriedade apenas dá uma idéia vaga sobre como uma função adeqüada ao conceito de informação deveria ser. Portanto, nós voltamos à definição dada na Introdução em busca de intuição: "Informação é aquilo que ela faz por você, ela muda sua opinião".

Como foi visto anteriormente, a opinião sobre o parâmetro de interesse antes que alguem realize um experimento é dada pela distribuição a priori sobre ele. Por outro lado, a opinião que se tem após a realização do experimento se torna, pelo uso do Teorema de Bayes, a distribuição a posteriori. Portanto, uma vez que a informação deveria representar a mudança de opinião, parece razoável que ela seja uma comparação entre a distribuição a priori e a distribuição a posteriori.

Quando a distribuição a priori é igual à distribuição à posteriori, uma vez que informação é uma mudança de opinião, parece razoável supor que não há informação. Portanto, iremos definir que a informação ganha quando a distribuição a priori é igual à distribuição a posteriori é 0. Uma vez que para qualquer distribuição a posteriori diferente da distribuição a priori existe alguma mudança de

opinião, parece razoável assumir que a informação é sempre maior ou igual a 0.

Também foi discutido que um experimento trivial não deveria trazer qualquer informação sobre o parâmetro de interesse. Assim, um experimento trivial deveria ser igual ou menos informativo que qualquer outro. Por outro lado, em qualquer experimento trivial a distribuição a priori é igual à distribuição a posteriori. Portanto, as definições dadas no parágrafo acima são consistentes com as propriedades esperadas de experimentos triviais.

A seguir, analizaremos um exemplo simples de acordo com algumas medidas famosas de informação para conjunto de dados. Nesse caso, alguem escolhe 3 bolas dentre 2 que são pretas e outras 2 que são brancas. Essas bolas são postas em uma urna. A seguir, a uma outra pessoa é dada a possiblidade de realizar um dentre três experimentos. O primeiro consiste em retirar apenas uma bola da urna. O segundo consiste em retirar duas bolas, com reposição, da urna. O último consiste em retirar duas bolas, sem reposição, da urna. O objetivo do experimento é determinar o número de bolas brancas na urna, 1 ou 2. Assumimos que a pessoa, a priori, não acredita que qualquer combinação possível de bolas é mais provável que outra. Por último, assumimos que a pessoa acredita que, a cada momento, cada uma das bolas da urna tem igual probabilidade de serem retiradas.

Seja θ o número de bolas brancas na urna e X_i o número de bolas brancas observadas no i-ésimo experimento, dizemos que $P(\theta|X_i = x_i) = (a, b)$ se após observar x_i bolas brancas no i-ésimo experimento, $P(\theta = 1|X_i = x_i) = a$ e $P(\theta = 2|X_i = x_i) = b$. Agora calculamos a distribuição a posteriori para cada experimento:

1. $P(\theta|X_1 = 0) = (2/3, 1/3)$, $P(\theta|X_1 = 1) = (1/3, 2/3)$
2. $P(\theta|X_1 = 0) = (4/5, 1/5)$, $P(\theta|X_1 = 1) = (1/2, 1/2)$, $P(\theta|X_1 = 2) = (1/5, 4/5)$
3. $P(\theta|X_1 = 0) = (1, 0)$, $P(\theta|X_1 = 1) = (1/2, 1/2)$, $P(\theta|X_1 = 2) = (0, 1)$

Algumas funções de informação comumente aplicadas a esses exemplos são:

1. Distância Euclidiana: $Inf_E(\epsilon_i, x_i) = \sqrt{\sum_{i\in\{1,2\}} (P(\theta = i) - P(\theta = i|X_i = x_i))^2}$
2. $Inf_V(\epsilon_i, x_i) = (E(\theta|X_i = x_i) - E(\theta))^2$
3. Divergência de Kullback-Leibler: $Inf_{KL}(\epsilon_i, x_i) =$
   $\sum_i P(\theta = i|X_i = x_i) log(P(\theta = i|X_i = x_i)/P(\theta = i))$

Portanto, no primeiro experimento, a informação dada por essas funções, tendo como índice o número de bolas brancas observadas, será respectivamente: $(1/3\sqrt{2}, 1/3\sqrt{2})$, $(1/36, 1/36)$ e $(0.0246, 0.0246)$. No segundo experimento elas serão, respectivamente, $(1/2, 0, 1/2)$, $(9/100, 0, 9/100)$ e $(0.0837, 0, 0.0837)$. Por último, no terceiro experimento, temos $(1/\sqrt{2}, 0, 1/\sqrt{2})$, $(1/4, 0, 1/4)$ e $(0.3010, 0, 0.3010)$.

Contudo, essas medidas não apresentam uma forma evidente de comparar experimentos. Portanto, no próximo capítulo, essa questão será estudada. Finalmente, na terceira seção do próxima capítulo apresentaremos adaptações dessas métricas para comparar experimentos.

Capítulo 3

Informação num Experimento

Nesse capítulo estaremos interessados na comparação entre experimentos em relação ao grau de informação que estes proporcionam. Para tal, na primeira seção será apresentada a definição de Suficiência de Blackwell e sua relação com o Princípio da Suficiência. A seguir, será apresentada a Equivalência de Blackwell e a sua relação com o Princípio da Verossimilhança antes estudado. Por último, as métricas apresentadas na última seção do segundo capítulo serão adaptadas para a comparação de experimentos usando-se a Teoria da Decisão.

3.1 Suficiência de Blackwell

Uma definição clássica na Inferência Estatística é a de Suficiência. De acordo com o Princípio da Suficiência exposto no capítulo passado, a inferência baseada nos dados ou em uma estatística suficiente é a mesma. No entanto, para utilizar esse princípio aquilo que se deseja comparar deve estar construído dentro do mesmo experimento estatístico. Por outro lado, geralmente, quando se comparar dois diferentes experimentos apenas a sua distribuição marginal é dada e, portanto, não é possível determinar se um é suficiente para o outro ou não. Nessa seção nós apresentamos a definição da Suficiência de Blackwell e tentamos mostrar que ela é uma generalização natural da idéia de Estatística Suficiente para a comparação entre experimentos.

Como foi apresentado no capítulo passado, uma estatística T, com contra-domínio τ, é suficiente para um experimento ϵ se o experimento residual ϵ_t^T é trivial para qualquer $t \in \tau$. Utilizando essa definição, obtemos que se T é suficiente, então $P(x|\theta) = P(T(x)|\theta)P(x|T(x))$. Por outro lado, foi mostrado na Introdução que a função de probabilidade do experimento ϵ_t^T, onde t é o resultado de

ϵ^T, é igual à de ϵ. Portanto, se T é uma estatística suficiente, é possível obter a distribuição dos dados observando T e, em seguida, realizando um experimento que não depende de θ.

Sejam $\epsilon_X = (\chi_X, \Theta, P_X)$ e $\epsilon_Y = (\chi_Y, \Theta, P_Y)$ dois experimentos estatísticos. Dizemos que ϵ_X é suficiente no sentido de Blackwell para ϵ_Y se existe um mapa $F : \chi_X \times \chi_Y \mapsto [0,1]$ satisfazendo duas propriedades. Primeiramente, $\sum_{y \in \chi_Y} F(x,y) = 1, \forall x \in \chi_X$. Em segundo, $P_Y(y|\theta) = \sum_{x \in \chi_X} P_X(x|\theta) F(x,y), \forall y \in \chi_Y$.

Definimos para todo $x \in \chi_X$ o experimento $\epsilon_x^X = (\chi_Y, \Theta, P_x^X)$, onde $P_x^X(y|\theta) = F(x,y), \forall \theta \in \Theta$. Portanto, para todo $x \in \chi_X$, ϵ_x^X é um experimento trivial. Além disso, pela definição de suficiência de Blackwell é fácil ver que o resultado do experimento, ϵ_x^X, onde x é o resultado do experimento ϵ_X, tem mesma distribuição de probabilidade que ϵ_Y. Portanto, ϵ_X é suficiente no sentido de Blackwell para ϵ_Y se e somente é possível obter dados com a mesma distribuição de probabilidade que ao observar ϵ_Y observando, ao contrário, ϵ_X e, em seguida, realizando um experimento trivial. Esse resultado é bastante semelhante a uma das definições de Estatística Suficiente e, portanto, parece razoável que é melhor observar ϵ_X do que ϵ_Y. Também é possível concluir que uma estatística suficiente é suficiente no sentido de Blackwell para o conjunto de dados e vice-versa.

Tendo apresentado os aspectos gerais da suficiência de Blackwell agora prosseguimos com alguns exemplos desse conceito. Esses estão relacionados com aqueles encontrados em Basu & Pereira [4] e parecem dar apoio a algumas idéias bem conhecidas utilizando apenas ferramentas simples. Nesses exemplos, para evitar excesso de notação, constantemente utilizaremos a variável aleatória que representa o resultado de um experimento estatístico ao invés de diretamente descrevê-lo.

Exemplo 1: Consideramos as variáveis aleatórias X e Y, ambas tendo parâmetro de interesse $\theta \in [0,1]$. Também definimos que $X \sim \text{Bernoulli}(\theta)$ e $Y \sim \text{Bernoulli}(\alpha\theta + (1-\alpha)p)$, onde α e p são duas constantes conhecidas em $[0,1]$. É fácil demonstrar que se $A \sim \text{Bernoulli}(\alpha)$ e $B \sim \text{Bernoulli}(p)$, ambos independentes de todas as outras variáveis aleatórias, então $AX + (1-A)B$ tem a mesma distribuição de Y. Portanto, como $\forall x \in \{0,1\}$, $AX + (1-A)B|X = x$ não depende de θ, X é suficiente no sentido de Blackwell para Y.

Exemplo 2: Sejam X e Y duas variáveis aleatórias com parâmetro de interesse $\theta \in [0,1]$, tal que $X \sim \text{Binomial}(n, \theta)$ e $Y \sim \text{Binomial}(n, \theta/k)$, $k \geq 1$. Sabemos que X tem a mesma distribuição de uma estatística suficiente para a seqüência $X_1, ..., X_n$ condicionalmente independentes dado θ e tal que $X_1 \sim \text{Bernoulli}(\theta)$. Portanto, X é suficiente no sentido de Blackwell para $X_1, ...X_n$. pelo exemplo 1, tomando $\alpha = 1/k$ e $p = 0$ sabemos que a seqüência anteriormente considerada é suficiente no sentido de Blackwell para a seqüência $Y_1, ..., Y_n$ condicionalmente independentes dado θ e tal que $Y_1 \sim \text{Bernoulli}(\theta)$. Essa seqüência é trivialmente suficiente no sentido de Blackwell para Y. Dessa forma, uma vez que a suficiência de Blackwell é uma relação transitiva, X é suficiente no sentido de Blackwell para Y.

Exemplo 3: A seguir, estudaremos uma generalização do exemplo abordado na parte 2.2. Consideramos uma urna com N bolas. θ dessas bolas são pretas e $N - \theta$ são brancas. n dessa bolas, $n \leq N$, são coletadas.

Definimos que uma seqüência $(X_1, ..., X_n)$ corresponde a uma amostragem com reposição se:

1. $X_1 \sim \text{Bernoulli}(\theta/n)$;
2. as variáveis aleatórias são identicamente distribuídas;
3. X_{i+1} é condicionalmente independente dado θ de $(X_i, ...X_1)$, $\forall i \in \{1, ...n-1\}$.

Por outro lado, definimos que $(Y_1, ..., Y_n)$ corresponde a uma amostragem sem reposição se:

1. $Y_1 \sim \text{Bernoulli}(\theta/n)$
2. $Y_{i+1}|(Y_i = y_i, ...Y_1 = y_1) \sim \text{Bernoulli}((\theta - \sum_{j=1}^{i} y_j)/(N-i))$,
   $\forall i \in \{1, ..., n-1\}, \forall (y_i, ..., y_1) \in \{0,1\}^i$.

Iremos demonstrar que $(Y_1, ...Y_n)$ é suficiente no sentido de Blackwell para $(X_1, ...X_n)$.

Primeiramente, definimos $X_1^* = Y_1$. A seguir, consideramos $\forall i \in \{1, ..., n-1\}$ duas variáveis aleatórias A_{i+1} e B_{i+1}. Definimos que $A_{i+1}|(Y_i = y_i, ..., Y_1 = y_1) \sim \text{Bernoulli}(\sum_{j=1}^{i} y_j/N)$ e é de todas as outras variáveis aleatórias consideradas. Também $B_{i+1}|(Y_i = y_i, ...Y_1 = y_1) \sim \text{Bernoulli}((N - i - 1 + \sum_{j=1}^{i} y_j)/N)$ e é independente de todas as outras variáveis aleatórias consideradas.

É fácil ver que se definirmos $X^*_{i+1} = A_{i+1}(1 - Y_{i+1}) + B_{i+1}Y_{i+1}$, então $X^*_{i+1}|(Y_i = y_i, ..., Y_1 = y_1) \sim$ Bernoulli(θ/N), $\forall i \in \{1, ..., n-1\}$, $\forall (y_i, ..., y_1) \in \{0,1\}^i$. Assim, $X^*_{i+1} \sim$ Bernoulli(θ/N) e é condicionalmente independente de $(Y_i, ..., Y_1)$. Finalmente, como $(X^*_i, ..., X^*_1)$ é uma função de $(Y_i, ...Y_1)$, $(A_i, ..., A_2)$ e $(B_i, ..., B_2)$, concluímos que X^*_{i+1} é independente de $(X^*_i, ..., X^*_1)$.

Pelas conclusões anteriores sabemos que $(X^*_1, ..., X^*_n)$ tem mesma distribuição de $(X_1, ..., X_n)$. Por construção, também é sabido que $(X^*_1, ..., X^*_n)|(Y_1 = y_1, ..., Y_n = y_n)$ é trivial, $\forall (y_i, ..., y_1) \in \{0,1\}^i$. Portanto, está demonstrado que a amostragem sem reposição é suficiente no sentido de Blackwell para a amostragem com reposição.

Exemplo 4: Mostramos um corolário simples do último resultado. Em primeiro lugar, sabemos que $\sum_{i=1}^n Y_i \sim$ Hipergeométrica(N, n, θ) e é uma estatística suficiente para $(Y_1, ..., Y_n)$. Portanto, $\sum_{i=1}^n Y_i$ é suficiente no sentido de Blackwell para $(Y_1, ..., Y_n)$. Pelo resultado prévio, sabemos que $(Y_1, ..., Y_n)$ é suficiente no sentido de Blackwell para $(X_1, ..., X_n)$. É trivial que $(X_1, ..., X_n)$ é suficiente no sentido de Blackwell para $\sum_{i=1}^n X_i$ e também é conhecido que $\sum_{i=1}^n X_i \sim$ Binomial$(n, \theta/N)$. Dessa forma, uma vez que a suficiência de Blackwell é uma relação transitiva, está provado que a distribuição hipergeométrica com parâmetros N, n e θ é suficiente no sentido de Blackwell para a distribuição binomial com parâmetros n e θ/n.

3.2 Simetrias na Informação de um Experimento

Na última seção a suficiência de Blackwell foi apresentada. Utilizando-a, é possível definir uma relação de equivalência entre dois experimentos. Dizemos que um experimento ϵ_1 é equivalente no sentido de Blackwell a ϵ_2 se ϵ_1 é suficiente no sentido de Blackwell para ϵ_2 e vice-versa. Nessa seção estudaremos essa relação de equivalência e tentaremos mostrar que ela é relacionada ao Princípio da Verossimilhança apresentado na parte 2.1. Para fazê-lo, provaremos o seguinte teorema:

TEOREMA: Sejam $\epsilon_1 = (\chi_1, \Theta, P_1^*)$ e $\epsilon_2 = (\chi_2, \Theta, P_2^*)$. ϵ_1 é equivalente no sentido de Blackwell a ϵ_2 se e somente se $P_1^*(x : P_1^*(x|\theta) \propto L(\theta)) = P_2^*(y : P_2^*(y|\theta) \propto L(\theta))$, $\forall\theta \in \Theta$ e toda função de verossimilhança $L(\theta)$ possível em ϵ_1 ou ϵ_2.

Para provar esse teorema vamos definir uma notação especial que reduz a álgebra envolvida. Em primeiro lugar, sempre que um conjunto finito é considerado tomaremos alguma ordem sobre ele. A seguir, seja $P(x|\theta)$ uma função de probabilidade, definimos que $\forall\theta_0 \in \Theta$, $p(.|\theta_0)$ é um vetor tal que em sua i-ésima posição o valor assumido é $P(x_i|\theta_0)$, onde x_i é o i-ésimo elemento da ordem assumida no conjunto de possíveis valores par x. A seguir, seja F um mapa de $\xi_1 \times \xi_2$ em $[0, 1]$. Usaremos o símbolo F também para designar a matriz que tem em sua j-ésima linha e i-ésima coluna o número $F(x_i, y_j)$, onde x_i é o i-ésimo elemento da ordem assumida em ξ_1 e y_j é o j-ésimo elemento da ordem assumida em ξ_2. Finalmente, uma matriz de transição é definida como qualquer uma em que a soma dos elementos de qualquer coluna é 1.

Demonstração:

($\Leftarrow$) Seja $S : \chi_1 \mapsto [0, 1]^\Theta$, tal que $S(x) = P_1^*(x|\theta)/(\sum_{i\in\Theta} P_1^*(x|i))$. De forma similar, consideramos $T : \chi_2 \mapsto [0, 1]^\Theta$ tal que $T(x) = P_2^*(x|\theta)/(\sum_{i\in\Theta} P_2^*(x|i))$. É fácil demonstrar que S é uma estatística suficiente para ϵ_1 e T para ϵ_2. Portanto, ϵ_1^S é equivalente no sentido de Blackwell a ϵ_1 e ϵ_2^T a ϵ_2. Tomando a afirmação à direita no teorema, ϵ_1^S tem a mesma função de probabilidade que ϵ_2^T e, portanto, eles são equivalentes no sentido de Blackwell. Concluímos pela propriedade transitiva da equivalência de Blackwell que ϵ_1 é equivalente no sentido de Blackwell a ϵ_2.

($\Rightarrow$)

Primeiramente, as mesmas estatísticas S e T definidas acima. Definimos $\eta_1 = (\xi_1, \theta, P_1(x|\theta))$ como ϵ_1^S. De forma análoga, definimos $\eta_2 = (\xi_2, \theta, P_2(x|\theta))$ como ϵ_1^T. É fácil demonstrar que, em qualquer um desses experimentos, se dois pontos do espaço amostral tem verossimilhança proporcional então eles são o mesmo ponto. Finalmente, como S e T são estatísticas suficientes, pelas suposições é possível concluir que η_1 é equivalente no sentido de Blackwell a ϵ_1 e η_2 é equivalente no sentido de Blackwell a ϵ_2.

Como η_1 é suficiente no sentido de Blackwell para η_2, existe um mapa $A : \xi_1 \times \xi_2 \mapsto [0, 1]$ tal que A é uma matriz de transição e:

$$Ap_1(.|\theta_0) = p_2(.|\theta_0), \forall \theta_0 \in \Theta$$

Por outro lado, η_2 também é suficiente no sentido de Blackwell para η_1 e, de forma similar, existe um mapa $B : \xi_2 \times \xi_1 \mapsto [0, 1]$ tal que B é uma matriz de transição e:

$$Bp_2(.|\theta_0) = p_1(.|\theta_0), \forall \theta_0 \in \Theta$$

Usando ambas essas relações nós sabemos que existem duas matrizes de transição $M = BA$ e $N = AB$, tais que:

$$Mp_1(.|\theta_0) = p_1(.|\theta_0), \forall \theta_0 \in \Theta$$

$$Np_2(.|\theta_0) = p_2(.|\theta_0), \forall \theta_0 \in \Theta$$

Adotaremos como estratégia de demonstração imaginar que M é a matriz de transição de uma Cadeia de Markov em ξ_1. Dessa forma, temos que $p_1(.|\theta_0)$ é uma medida invariante de M, $\forall \theta_0 \in \Theta$.

Em primeiro lugar, é fácil demonstrar que não existem estados transientes em M. Se existisse algum estado transiente x em M então $P(x|\theta_0) = 0$, $\forall \theta_0 \in \Theta$. Tal estado é impossível pelas definições dadas no primeiro capítulo. Assim, concluímos que não há estados transientes em M.

A seguir, apresentaremos um lema muito útil que é estudado em Cadeias de Markov. Esse lema será aplicado à matriz M para auxiliar a demonstrar o Teorema.

LEMA : Para qualquer Cadeia de Markov em um espaço finito χ com matriz de transição A e nenhum estado transiente. Consideramos em A os componentes irredutíveis $C(1)$, ..., $C(n)$. Existe um conjunto único de funções de probabilidade $p_1(i)$, $i \in \{1, \ldots, |C(1)|\}$, ... , $p_n(i)$, $i \in \{1, \ldots, |C(n)|\}$, tais que toda medida invariante (μ) de A pode escrita da seguinte forma:

Se $c_{k,i}$ é o elemento de ordem i de $C(k)$ então $\mu(c_{k,i}) = p_k(i).q_k$, para alguma q função de probabilidade em {1,...,n}.

Esse resultado pode ser encontrado em Ferrari & Galves [7]. Uma forma de interpretá-lo é considerando a sub-matriz A_k associada a $C(k)$. Uma vez que essa matriz é irredutível, ela apenas

tem uma medida invariante que é p_k. Agora tomamos que na posição inicial (x_0) da cadeia cada componente $C(k)$ tem probabilidade q_k de ser escolhido. Quando n vai para infinito a lei de x_n converge para aquela dada pelo lema.

Agora, usando o lema, sabemos que se $C(1), \ldots, C(n)$ são componentes irredutíveis de M e $c(k,i)$ é o elemento de ordem i de $C(k)$ então $p_1(c(k,i)|\theta_0) = q_{k,\theta_0}.p_k(i)$. Como conseqüência, $p_1(c(k,i)|\theta_0) = p_1(c(k,j)|\theta).(p_k(i)/p_k(j))$. Finalmente, temos que se dois estados estão no mesmo componente irredutível de M então suas verossimilhanças são proporcionais. A mesma prova se aplica à matriz N.

Dizemos que o i-ézimos elemento de ξ_1 está conectado ao j-ésimo elemento de ξ_2 se $A(i,j) > 0$. De forma semelhante, o i-ésimo elemento de ξ_2 está conectado ao j-ésimo elemento de ξ_1 se $B(i,j) > 0$.

Nota-se que todo estado em ξ_1 está conectado a pelo menos um estado em ξ_2 e vice-versa. Isso acontece pois A e B são matrizes de transição.

Após isso, é fácil ver que para todo $x_1 \in \xi_1$, se x_1 está conectado a $x_2 \in \xi_2$ então x_2 está apenas conectado a x_1. Se existisse um estado $y \in \xi_1$ tal que x_2 estivesse conectado a y, então x_1 e y estariam no mesmo componente irredutível de M. Portanto, x_1 e y teriam verossimilhanças proporcionais e, pela definição de S, $x_1 = y$. De forma similar, se um estado $x_2 \in \xi_2$ está conectado a um outro $x_1 \in \xi_1$ então x_1 está apenas conectado a x_2.

Pela conclusão acima está provado que todo estado $x_1 \in \xi_1$ está conectado a um e a apenas um estado $x_2 \in \xi_2$ e este está conectado apenas a x_1. Isso implica que se x está conectado a y, então $P_1(x|\theta_0) = P_2(y|\theta_0)$, $\forall \theta_0 \in \Theta$. Como S e T são estatísticas suficientes, o teorema está provado.

Por uma simples aplicação do Teorema aprovado e do Princípio da Verossimilhança é possível notar que: se ϵ_1 é equivalente no sentido de Blackwell a ϵ_2 então $P_1^*(x : Inf(\epsilon_1, x) = k|\theta) = P_2^*(y : Inf(\epsilon_2, y) = k|\theta)$, $\forall \theta \in \Theta$, $\forall k$ possível resultado da função de informação Inf que respeita o Princípio da Verossimilhança. Portanto, a equivalência de Blackwell é uma conseqüência do Princípio da Verossimilhança e a suposição de que se a informação dada por dois experimentos tem

a mesma distribuição $\forall\theta \in \Theta$ então eles são equivalentes.

3.3 Funções de Informação num Experimento

Nessa seção, estaremos interessados em funções capazes de descrever a informação de um experimento estatístico. Uma possível abordagem é considerar que a informação ganha é uma função de utilidade (DeGroot [6]) que o Estatístico deseja maximizar. Dessa forma, pelos resultados da Teoria da Decisão, $Inf(\epsilon, P^*(\theta)) = E(Inf(\epsilon, X))$, onde X é o resultado do experimento ϵ, $Inf(\epsilon)$ é a informação ganha com ϵ e $P^*(\theta)$ a distribuição a priori para o parâmetro.

Manteremos essa abordagem para comparar as diferentes funções de informação apresentadas na parte 2.2. Para fazê-lo, vamos lembrar que para qualquer uma das funções apresentadas o maior resultado possível é obtido quando a distribuição a posteriori é tal que $P(\theta = 0|x) = 0$ ou $P(\theta = 0|x) = 1$. Portanto, vamos dividir todas as funções de informação por esses máximos de forma a colocá-los em uma forma comparável.

Como foi visto, nos três experimentos possíveis, os vetores de informação usando a distância euclidiana são, respectivamente, $(1/3\sqrt{2}, 1/3\sqrt{2})$, $(1/2, 0, 1/2)$ e $(1/\sqrt{2}, 0, 1/\sqrt{2})$. Portanto, no primeiro experimento, com probabilidade 1 o ganho de informação é 33% do maximo, isto é, há pouco ganho e pouco risco. No segundo experimento, com probabilidade 56% o ganho é 60% do máximo e com probabilidade 44% ele é 0% do máximo, ganho moderado com risco moderado. No terceiro experimento é possível obter 100% da máxima informação possível com probabilidade 33% e é possível obter 0% da máxima possível informação com probabiliade 67%, alto ganho com alto risco. No entanto, em todos esses experimentos o ganho esperado de informação é de 24% e não haveria preferência por qualquer um deles. Isso poderia soar estranho, uma vez que o terceiro é suficiente no sentido de Blackwell para os outros dois e o segundo também o é para o primeiro.

Por outro lado, agora consideramos a segunda função de informação apresentada, $Inf(\epsilon, x) = (E(\theta) - E(\theta|X = x))^2$. Nos três possíveis experimentos, os vetores de informação usando essa função são, respectivamente, $(1/36, 1/36)$, $(9/100, 0, 9/100)$ e $(1/4, 0, 1/4)$. Dividindo pelo máximo, obtemos os seguintes vetores: $(1/9, 1/9)$, $(9/25, 0, 9/25)$ e $(1, 0, 1)$. Portanto, os ganhos esperados de informação são, respectivamente, 11%, 20% e 33%. Portanto, o terceiro experimento é mais informativo que os outros dois e o segundo experimento é mais informativo que o primeiro. É

interessante notar que a medida de informação utilizando essa métrica é dada por: $Inf(\epsilon) = V(E(\theta|X))$.

Finalmente, os vetores de informação dados pela terceira métrica são, respectivamente, $(0.0246, 0.0246)$, $(0.0837, 0, 0.0837)$ e $(0.3010, 0, 0.3010)$. Dividindo pelo máximo, obtemos os seguintes vetores: $(0.081, 0.081)$, $(0.2780, 0, 0.2780)$ e $(1, 0, 1)$. Assim, os ganhos esperados de informação são respectivamente, $2,4\%$, $4,6\%$ and 33%. Novamente, o terceiro experimento é mais informativo que os outros dois e o segundo experimento é mais informativo que o primeiro.

Poderia parecer coincidência que a segunda e a terceira definição de função de informação respeitassem a suficiência de Blackwell. No entanto, esse resultado pode ser provado notando-se que ambas as métricas são côncavas e, em seguida, utilizando a desigualdade de Jensen na esperança da utilidade esperada de um experimento.

Capítulo 4

Diagramas de Influência

Nos capítulos precedentes as simetrias em relação à informação, em geral, foram apresentadas condicionalmente dado um valor arbitrário para o parâmetro de interesse. Dessa forma, para definir suficiência, por exemplo, não foi necessário especificar se o parâmetro de interesse era uma variável aleatória ou não. Ainda que essa abordagem seja rigorosa, acreditamos que parte da intuição dos conceitos é perdida por não utilizar o parâmetro como uma variável aleatória.

Nesse capítulo, muitas das definições vistas enquanto definindo simetrias em informação são revistas pelo uso de diagramas de influências. Além disso, algumas novas simetrias são apresentadas e relacionadas ao conceito de informação. Essa abordagem somente é possível considerando-se que o parâmetro de interesse é uma variável aleatória. Acreditamos que pela própria natureza gráfica dos diagramas de influência as definições ganhem nova intuição.

A apresentação das definições básicas relativas a diagramas de influência será feita de acordo com Barlow [1].

4.1 Definição e Propriedades

Definimos um grafo orientado como sendo um conjunto de vértices, V, e um mapa $A : V \times V \mapsto \{0, 1\}$ que indica o conjunto de arestas do grafo. Dizemos que existe uma aresta ligando um vértice, v_1, a outro v_2 se $F(v_1, v_2) = 1$. Dizemos que não existe uma aresta com essa propriedade se $F(v_1, v_2) = 0$. Mais suscintamente, dizemos se v_1 está conectado ou não está conectado a v_2. Na representação gráfica de um grafo cada nodo é representado por um círculo e a existência de uma aresta conectando v_1 a v_2 é representada por uma flecha ligando esses vértices.

Definimos que existe um caminho ligando $a \in V$ a $b \in V$ se existe uma seqüência finita de elementos de V, $(v_1, ...v_n)$, tal que $A(a, v_1) = 1$, $A(v_n, b) = 1$ e $A(v_i, v_{i+1}) = 1$, $\forall i \in \{1, n-2\}$. Quando necessário, dizemos que a é a origem do caminho e b é o destino do caminho. Finalmente, um grafo é definido como sendo acíclico se para qualquer $v \in V$ não existe caminho ligando v a v. Como conseqüência de um gráfico ser acíclico sempre há pelo menos um vértice ao qual nenhum outro se conecta. Qualquer vértice com essa propriedade é denominado raiz. Também é conseqüência de um grafo ser acíclico a existência de pelo menos um vértice que não se conecta a qualquer outro. Esse vértice é denominado folha.

Um diagrama de influência é um grafo acíclico acrescido de uma família de funções de probabilidade. A cada vértice do grafo acíclico está associada uma variável aleatória. A partir desse ponto não faremos mais distinção entre a variável aleatória e o próprio vértice. A essa variável aleatória está associada a sua probabilidade condicional dados todos os valores possíveis dos vértices que se conectam a ela. É simples provar que essa caracterização gera uma única a probabilidade conjunta de todos os vértices. No entanto, pode existir mais de um diagrama de influência para representar uma determinada probabilidade conjunta dos vértices. Por último, desenhamos dois círculos em torno de um vértice se ele é função determinística dos vértices que se conectam a ele.

Como conseqüências dessas definições, temos uma importante resultado:

Resultado: Se dois vértices, v_1 e v_2, são tais que v_1 não está conectado a v_2 e v_2 não está conectado a v_1, então condicionalmente a todos os vértices que se conectam a v_1 ou a v_2 eles são independentes.

Essa propriedade da independência condicional será utilizada na seção seguinte na definição de todos os conceitos relativos às simetrias na informação.

4.2 Diagramas de Influência aplicados à Teoria Estatística

Nessa seção estudaremos os conceitos relativos a simetrias na informação aplicando neles a idéia de independência condicional abordada na última seção.

A primeira definição apresentada relativa a simetrias na informação foi o de estatística suficiente. Conforme presente na seção 2.1, uma estatística $T(X)$ é suficiente para um conjunto de dados X

em relação a um parâmetro de interesse θ se $P_t^T(x|T = t, \theta = \theta_0)$ é constante no parâmetro de interesse para qualquer valor t assumido pela estatística. Em particular, podemos escrever $P(\theta = \theta_0, T = t, X = x) = P(\theta = \theta_0)P(T = t|\theta = \theta_0)P(X = x|T = t)$. Portanto, temos que θ e X são condicionalmente independentes dado T. Como foi visto, essa relação pode ser representada com θ sendo um vértice raiz cuja única ligação é com T que, por sua vez, tem como única ligação X, um vértice folha.

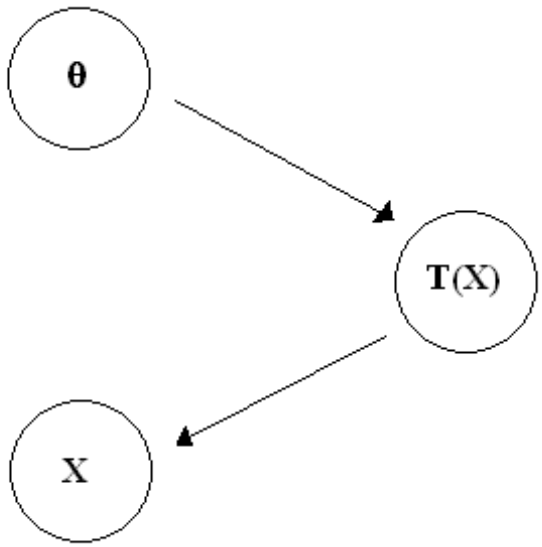

Figura 4.1: $T(X)$ é suficiente para θ.

Uma outra definição apresentada foi a de suficiência de Blackwell. Dizemos que X é suficiente no sentido de Blackwell para Y se existe uma variável aleatória Z independente de θ e uma função $F(X, Z)$ mensurável na σ-álgebra gerada pela união das σ-álgebras de X e Z tal que $F(X, Z|\theta = \theta_0)$ tem mesma distribuição de Y condicionalmente a θ. Utilizando diagramas de influência, dizemos que X é suficiente no sentido de Blackwell para Y quando existem dois nós, Z e Y^* tal que, θ é raiz e se liga apenas a X, X se liga apenas a Y^*, Z é raiz e se liga apenas a Y^* e Y^* é um vértice folha com distribuição condicional a θ igual à de Y.

Uma definição que não foi apresentada nos capítulos anteriores é o de Estatística Suficiente Mínima. No entanto, uma vez que essa tem relação com informação e pode ser explicada pelo uso de diagramas de influência, iremos introduzí-la. Usualmente, é dito que uma estatística $S(X)$ é suficiente mínima se para toda estatística suficiente $T(X)$ existe uma função F tal que $F(T(X)) =$

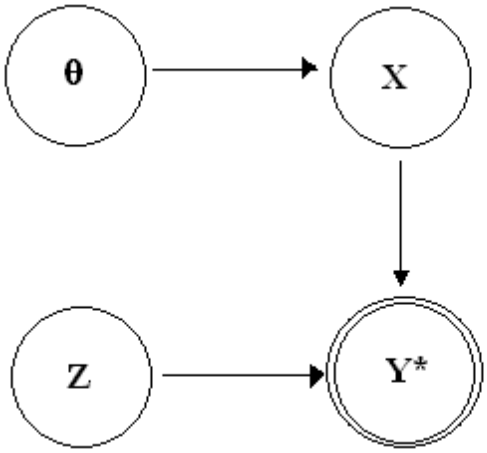

Figura 4.2: X é suficiente no sentido de Blackwell para Y. Y tem mesma distribuição de Y^*.

$S(X)$. Essa propriedade pode ser traduzida em diagramas de influência da seguinte forma: Para toda estatística suficiente $T(X)$ é possível introduzir um novo vértice $S(X)$ entre θ e $T(X)$.

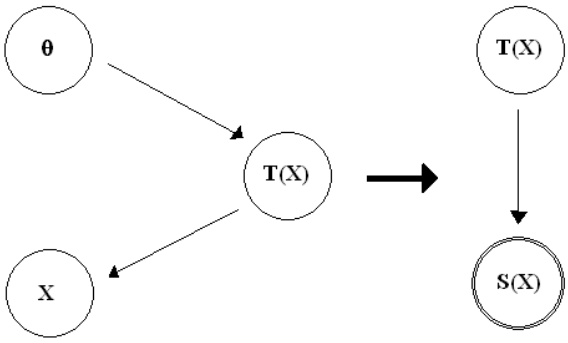

Figura 4.3: S(X) é suficiente mínima.

Uma definição que também não foi apresentada nos capítulos anteriores é o de estatística completa. Geralmente, define-se que uma estatística $T(X)$ é completa para um conjunto de dados X e um parâmetro de interesse θ se, dada uma função com domínio nos possíveis valores de T e contra-domínio na reta F, $E(F(T(X)))$ implica que $F(t) = 0$, para qualquer possível valor t de T. Demonstraremos que as seguintes afirmações são equivalentes:

1. T é completa

2. o conjunto de funções de verossimilhança do experimento ϵ^T é linearmente independente

3. Para qualquer aleatorização Y de T, se Y é suficiente mínima considerando T como parâmetro (e Y como os dados) então Y é suficiente mínima considerando θ como o parâmetro (e Y como os dados).

Primeiramente demonstraremos que T é completa se e somente se o conjunto de funções de verossimilhança do experimento ϵ^T é linearmente independente. Se o conjunto de funções de verossimilhança é linearmente independente então $\sum_{t\in\tau} a_t P^T(t|\theta) = 0, \forall\theta \in \Theta$ somente se $a_t = 0$, $\forall t \in \tau$. Portanto, $E(F(T)) = \sum_{t\in\tau} F(t)P^T(t|\theta) = 0, \forall\theta \in \Theta$, se somente se F(t) = 0, $\forall t \in \tau$ e, portanto, T é completa. A volta decorre percebendo que E(F(T)) corresponde a uma combinação linear das verossimilhanças com fatores $F(t)$.

A seguir, demonstramos que as funções de verossimilhança de um experimento, ϵ^T, são linearmente independentes se e somente se toda aleatorização, Y, de T que é suficiente mínima tomando-se T como parâmetro é também suficiente mínima tomando θ como o parâmetro.

Se o conjunto de verossimilhanças de ϵ^T é linearmente dependente então existe $a_{t\in\tau} \in [0,1]^\tau$ tal que $\exists t \in \tau$ com $a_t > 0$ e existe uma partição de τ, $\{\tau_1, \tau_2\}$, tal que $\sum_{t\in\tau_1} a_t P^T(t|\theta) = \sum_{t\in\tau_2} a_t P^T(t|\theta)$, $\forall\theta \in \Theta$. Definimos U como uma variável aleatória uniforme em $[0,1]$ e independente de T. Também definimos $F : \tau \times [0,1] \mapsto \{1,2,3\}$ da seguinte forma: $F(t,u) = 3$, se $u > a_t$, $F(t,u) = 1$, se $u \leq a_t$ e $t \in \tau_1$ e $F(t,u) = 2$, caso contrário. A aleatorização $F(T,U)$ é suficiente mínima tomando T como o parâmetro mas não é suficiente mínima para θ pois $L_1(\theta) = L_2(\theta)$. Dessa forma, provamos a volta.

Suponhamos que exista uma aleatorização de T, $Y : \tau \mapsto \chi_Y$, que seja suficiente mínima tomando T como parâmetro, mas o não seja tomando θ como parâmetro. Portanto, existem $x, y \in \chi_Y$ tais que $L_x(\theta) \propto L_y(\theta)$. Como Y é uma aleatorização de T existem $a_{t\in\tau}, b_{t\in\tau} \in [0,1]^\tau$ tais que $L_x(\theta) = \sum_{t\in\tau} a_t P^T(t|\theta), \forall\theta \in \Theta$, e $L_y(\theta) = \sum_{t\in\tau} b_t P^T(t|\theta), \forall\theta \in \Theta$. Assim, $\exists K \in R$ tal que $\sum_{t\in\tau} (a_t - Kb_t)P^T(t|\theta) = 0, \forall\theta \in \Theta$. Como Y é suficiente mínima tomando T como parâmetro, temos que $a_{t\in\tau}$ e $b_{t\in\tau}$ não são proporcionais. Portanto, as verossimilhanças do experimento ϵ^T são linearmente dependentes, concluindo a demonstração.

Como já descrevemos a definição de suficiência mínima usando diagramas de influência, o teorema demonstrado permite que a definição de estatística completa também seja apresentada usando diagramas de influência.

Capítulo 5

Conclusões

Durante o decorrer do trabalho, estivemos interessados no conceito de informação. Especialmente, estivemos interessados em determinar propriedades uma função de informação deveria ter. Essa pesquisa foi orientada pela noção dada em [2] de que "Informação é aquilo que ela faz por você, ela muda sua opinião".

Para fazê-lo, inicialmente definimos quão informativo é um conjunto de dados. Primeiramente, a idéia de um experimento trivial foi definida. Depois disso, foi mostrado que o Princípio da Verossimilhança é uma conseqüência de acreditar que um experimento trivial não deve trazer informação a respeitodo parâmetro de interesse. Seguindo esse resultado, foi estabelecido que funções de informação deveriam trazer mesmo resultado para dados com verossimilhanças proporcionais. Finalmente, foram apresentadas diferentes possíveis funções de informação e essas foram aplicadas a um exemplo intuitivo.

A seguir, focamos no problema do quão informativo é um determinado experimento. Para fazê-lo, primeramente definimos o conceito de suficiência de Blackwell. Esse conceito também parece ser conseqüência da idéia de que um experimento trivial não deve trazer informação a respeito do parâmetro. A seguir, mostramos que a suficiência de Blackwell induz uma classe de equivalência relacionada ao Princípio da Verossimilhança. Finalmente, usando o mesmo exemplo anteriormente apresentado, estudamos a aplicação de algumas funções de informação a diferentes experimentos. Isso foi feito considerando que o ganho de informação de um experimento é o ganho esperado de informação nos possíveis resultados do experimento.

Finalmente, estudamos as simetrias relacionadas ao conceito de informação utilizando a ferramenta gráfica de diagramas de influência. Fazendo-o, esperamos ter conseguido apresentar uma

idéia intuitiva de definições como suficiência, suficiência de Blackwell, suficiência mínima e completude.

Referências Bibliográficas

[1] R. E. Barlow, *Engineering reliability*, American Statistical Association, Virginia, 1998.

[2] D. Basu, *Statistical information and likelihood*, Sankhya: The Indian Journal of Statistics **37** (1975), no. Series A, 1–71.

[3] _____, *Statistical information & likelihood: A collection of critical essays*, Gosh, J.K., Editor. Springer-Verlag, 1988.

[4] D. Basu and C.A.B Pereira, *Blackwell sufficiency and bernoulli experiments*, Brazilian Journal of Probability and Statistics **4** (1990), 137–145.

[5] A. Birnbaum, *On the foundations of statistical inference*, Journal of the AmericanStatistical Association **57** (1962), 269–326.

[6] M. H. DeGroot, *Optimal statistical decisions*, McGraw-Hill, New York, 1970.

[7] P. A. Ferrari and A. Galves, *Coupling and regeneration for stochastic processes*, Sociedad Venezoelana de Matematicas, Caracas, 2000.

[8] C.A.B Pereira, *Bayesian solutions to some classical problems of statistics*, Ph.D. thesis, Florida State University, 1980.

[9] S. Wechsler, C.A.B. Pereira, and P.C. Marques, *Birnbaum's theorem redux*, Proceedings of 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (2008).