Mecânica Estatística de Sistemas de Processamento de Informação

Nestor Caticha

11 de março de 2011

Sumário

Capítulo 1

Probabilidades e Informação

Estas notas são a versão preliminar de um livro intitulado provisoriamente de "Mecânica Estatística de Sistemas de Processamento de Informação". Eventuais críticas, correções ou sugestões serão bemvindas.

1.1 Teoremas de Cox

Há muitas definições matemáticas possíveis que poderiam ser usadas na tentativa de formalizar o conceito coloquial de informação. Uma forma de avançar, que é bastante comum em ciência, começa por definir matematicamente algo e depois tentar interpretar as fórmulas matemáticas para mostrar que esta interpretação esta de acordo com algumas das características que podemos atribuir ao conceito coloquial de informação que temos . Não haverá forma que satisfaça a todos pois cada um terá um exemplo onde este conceito falha.

Em lugar de começar por uma estrutura matemática pré-escolhida para servir de ferramenta de análise, começamos por uma interpretação e depois encontramos a estrutura matemática que se adapte à interpretação. A interpretação passa por estabelecer em alguns casos particulares suficientemente simples, tais que haja algum tipo de consenso, o quê deveria resultar da teoria. É possível que este procedimento pareça novo ao leitor e será surpreendente quantos resultados serão extraidos deste método e do rigor matemático que a teoria se vestirá. Como este procedimento permite saber exatamente do que estamos falando e do que não estamos, achamos que esta é atualmente a melhor maneira de introduzir a teoria de informação. [1]

Queremos analisar uma asserção, isto é, uma frase A que em princípio é uma proposição que se apresenta como verdadeira. Uma frase pode ser julgada cor-

[1] Também ocorrerá que os resultados não serão universalmente satisfatórios, pois há lugar a discussões sobre o tipo de interpretação *a priori* que será imposta. Ver [**?**] possíveis extensões e críticas leves que talvez não sejam tão relevantes

reta ou não de várias maneiras. Podemos pensar se é correta do ponto de vista da sua estrutura gramatical ou sintática. No entanto, nenhuma asserção sozinha pode ser analisada, no que diz respeito a se é verdadeira ou não, de forma independente do resto do universo conceitual. Ela será julgada verdadeira ou não quando analisada dentro de um contexto. A informação trazida por uma asserção C, será usada para atribuir um grau de verdade à asserção A, ou seja dentro do contexto C. Poderiamos chamar esse grau de, por exemplo, probabilidade de que A seja verdade se C for dada. Mas fazendo isto estariamos definindo de antemão que a ferramenta matemática apropriada para descrever informação é a teoria de probabilidades. Isto parece bem razoável mas não escapa às críticas acima e permite que outra ferramenta matemática seja usada por simplesmente expressar o gosto de outras pessoas ou a facilidade de uso em determinados problemas práticos com a mesma justificativa: *parece razoável , eu gosto, funciona, é prático* . Não descartamos o uso de outras ferramentas matemáticas, mas queremos deixar claro que estas poderão ser vistas como aproximações mais ou menos adequadas de uma estrutura que unifica e tem um posição diferente. O **objetivo** deste capítulo é mostrar que a escolha da teoria de probabilidades como a ferramenta matemática adequada para tratar informação é muito mais do que simplesmente conveniente. Isto nos levará à teoria de inferência, baseada na teoria de probabilidades, que tem **exatamente** a estrutura da Mecânica Estatística dos pioneiros Boltzmann e Gibbs. Os "exatamentes"não são coincidências. Somos levados a repensar a Mecânica Estatística como uma teoria de inferência, mas muito mais sobre isto será dito adiante. Antes disso há muito o que fazer.

Se a informação em C não permite a certeza sobre a verdade de A então diremos que a crença que temos sobre A esta baseada em informação incompleta. Em casos particulares poderá ocorrer que dado C possa ser concluido, com certeza que a asserção A é verdadeira ou ainda em outros casos que é falsa. Quando não há alternativa para a conclusão, quando ela segue por força da informação disponível, dizemos que a conclusão é racional ou lógica. Dizemos que estamos frente a casos de raciocínio dedutivo. Nestes casos a informação disponível é *completa* pois nada falta para ter certeza. A análise destes casos remonta a Aristóteles.

Exemplos de informação completa são dados pelos silogismos Aristotélicos: suponha que recebemos a informação contida em $C =$ "$A \rightarrow B''$, isto é, A implica B. Traduzindo, isto significa "se souber que A é certamente verdade, segue que a proposição B também o é." Dado isso, o que podemos dizer sobre B? Nada com certeza, mas se também recebemos a informação adicional A, isto é, que A é Verdade, então segue B, ou seja B é Verdade.

Outro caso de informação completa é $\overline{B}$ ou seja B é Falso, então segue $\overline{A}$, isto é, que A é Falso.

Nas condições que $C =$ "$A \rightarrow B''$ e A é Falso , o quê pode ser concluido? Do ponto de vista lógico clássico nada podemos concluir sobre B. Da mesma forma se for dada a informação B é Verdade, nada podemos concluir sobre A. Estamos frente a casos de informação incompleta e a lógica clássica não serve para chegar a uma conclusão. Não é possível deduzir nada. A indução, o que quer que isto

seja, e que será discutido mais à frente, será necessária para avançar [2].

A forma dedutiva da lógica permite somente tres tipos de respostas, *sim, não* e *não segue*. A indução nos força ou permite dividir esta última em várias possibilidades e os casos extremos nesse espectro são aqueles onde havendo certeza absoluta, haverá portanto a força da dedução. Podemos falar então sobre quais das alternativas intermediárias é mais razoável acreditar com base no que sabemos. Nota-se então a necessidade de estender a lógica para poder tratar de forma racional casos de informação incompleta. Richard T. Cox, ao se defrontar com este problema por volta da década de 1940, decidiu, como dito acima, estabelecer um conjunto de desejos (*desiderata* [**?**]) que a teoria deveria satisfazer, e estes serão então os axiomas da extensão da lógica. Aqui podemos discordar, propor outros axiomas, mas uma vez aceitos serão provados os teoremas de reparametrização de Cox que mostram que a teoria de probabilidade é a ferramenta para o tratamento de forma racional de situações de informação incompleta. O surpreendente disto é que surge a teoria das probabilidades como a forma para lidar de forma *racional* com a informação e que corremos riscos de ser inconsistentes caso a regras de manipulação de probabilidades não sejam seguidas. Segue que não há probabilidades que não sejam condicionais embora às vezes simplesmente a linguagem esqueca de deixar explícitas as relações de condicionalidade. A amplidão da aplicabilidade da teoria que emerge é impressionante e por exemplo, quando o tipo de asserção for limitado àqueles entendidos em teoria de conjuntos as regras de manipulação serão não mais nem menos que aquelas ditadas pelos axiomas de Kolmogorov. Veremos que emerge uma relação natural entre probabilidade e freqüência e ficará claro de que forma estes conceitos estão ligados e mais importante, de que forma são distintos.

1.1.1 Axiomas de Cox-Jaynes

É interessante notar que os axiomas de Cox descritos por Jaynes não são exatamente iguais aos que Cox apresenta no seu livro *The algebra of probable inference* . A exposição de Jaynes é muito mais simples e por isso nos referimos a Cox-Jaynes enquanto que Jaynes só se refere a Cox. Cox, por sua vez, esclarece sua dívida com J. M.Keynes e seu livro *A treatise on Probability*, que deve muito a Laplace e Bernuolli.

A maneira de construir a teoria está baseada na seguinte forma de pensar bastante simples. Queremos construir uma teoria geral para a extensão da

[2]Segundo Harold Jeffreys em seu livro *Theory of Probability*, Bertrand Russell disse que "induction is either disguised deduction or a mere method of making plausible guesses". Jeffreys diz que "é muito melhor trocar a ordem dos dois termos e que muito do que normalmente passa por dedução é indução disfarçada, e que até alguns dos postulados de *Principia Mathematica* foram adotados por motivações indutivas" (e adiciona , são falsos). Com o tempo o próprio Russell mudou de posição, dobrado pela evidência (?) e diz no fim da sua autobiografia: "I was troubled by scepticism and unwillingly forced to the conclusion that most of what passes for knowledge is open to reasonable doubt". Sobre indução disse ainda: "The general principles of science, such as the belief of the reign of law, and the belief that every event must have a cause, are as completeley dependent on the inductive principle as are the beliefs of daily life."(On Induction)

lógica nos casos de informação incompleta. Se ela for geral, deverá ser válida em casos particulares. Se o caso for suficientemente simples, então podemos saber qual é o resultado esperado. Poderia ocorrer que ao analisar um número de casos particulares sejam reveladas as inconsistências entre eles, nesse caso não poderemos chegar a uma teoria geral. Mas pode ser que os casos particulares sirvam para restringir e determinar a teora geral [3].Isto é o que mostraremos a seguir.

Em primeiro lugar queremos falar sobre uma asserção A no caso de informação incompleta. Nos referimos então à plausibilidade de A ser verdade dado B e a denotamos pelo símbolo $A|B$. Por que não mais provável? Porque já existe uma teoria matemática de probabilidade e não sabemos se esta será a estrutura matemática que emergirá desta análise. Poderiamos usar outras palavras

Queremos analisar o primeiro caso simples que lida com o conceito de *mais plausível*. Se A é mais plausível dada informação B do que A dada C, e esta é ainda mais plausível que A dado D então A dado B deveria ser mais plausível que A dado D. Temos assim nosso primeiro desejo, a plausiblidade deverá satisfazer alguma forma de transitividade. Isto é fácil se:

- D_1: A plausibilidade $A|B$ deverá ser representada por um número real.

Dados

$$A|B > A|C$$

e

$$A|C > A|D,$$

segue imediatamente, uma vez que são números reais, que

$$A|B > A|D,$$

de acordo com o axioma 1. Note que dizer que alguma coisa é um número real nos dá imediatamente a transitividade, mas não diz nada sobre que número deve ser atribuido, nem sobre como mudá-lo se a informação passa de B para C.

Através de certas operações e de diferentes asserções podemos criar asserções compostas. Exemplos de operadores são a negação, o produto e a soma lógicos. A negação de A é denotada por $\overline{A}$. O produto ou conjunção de duas asserções é uma terceira asserção: $C = AB$ ou, mudando a notação $C = A$ **e** B. A soma de duas asserções é uma terceira asserção $D = A + B$, ou $D = A$ **ou** B.

A tabela 1.1 mostra a tabela verdade para as operações de soma e produto lógico, onde 1 = Verdade e 0 = Falso. Note que as últimas duas colu-

[3]Este comentário parece trivial, mas o uso que será dado a seguir é totalmente não trivial. Neste contexto de probabilidades foi colocado primeiro por J.Skilling, mas não de forma explicita. O destaque a este procedimento apareceu por priemeira vez no livro de A. Caticha. Usaremos novamente este estilo de fazer teoria ao introduzir o conceito de entropia.

nas,colocadas aqui para futura referência, mostram que $\overline{A+B}$ e $\overline{A}\,\overline{B}$ são iguais.

A	B	$A+B$	AB	$\overline{A+B}$	$\overline{A}\,\overline{B}$
1	1	1	1	0	0
1	0	1	0	0	0
0	1	1	0	0	0
0	0	0	0	1	1

Tabela 1.1

O próximo caso simples lida com informação neutra. Suponha que

$$A|C \geq A|C',$$

ou seja a plausibilidade de A diminui quando a informação disponível passa de C para C'. Suponha que para B isso não aconteça. Pensemos no caso que B é indiferente ante a mudança de C para C'. Isto é

$$B|C = B|C'.$$

Parece razoável que se a asserção conjunta AB for considerada, isto é a conjunção A **e** B, então seria desejável que

- D_2: $A|C \geq A|C'$ e $B|C = B|C'$ implicam que $AB|C \geq AB|C'$

Jaynes defende que este desejo está de acordo com o *bom senso*. Talvez seja difícil definir o que é bom senso, mas sería mais difícil negar que isto seja razoável.

O leitor talvez possa se convencer através de um simples exemplo. Seja A='Há vida em Marte', C= 'Há água em Marte', $C' = \overline{C}$, a negação de C. Suponhamos óbvio que $A|C \geq A|C'$. Suponha que B='Hoje é segunda feira'. Certamente $B|C = B|C'$. e também é razoável que a plausibilidade de que haja vida em Marte e hoje seja segunda feira' dado que 'há agua em Marte' é maior ou igual a plausibilidade de que 'haja vida em Marte e hoje seja segunda' dado que 'não há agua em Marte'.

Suponha que tenhamos um método, usando a teoria geral que procuramos e ainda não temos, de analizar a plausibilidade de uma asserção composta por várias asserções através de conjunções ou disjunções. Esperamos que a plausibilidade possa ser expressa em termos da plausibilidade de asserções mais simples. Talvez haja mais de uma forma de realizar essa análise. Queremos então que:

- D_3: Se a plausibilidade de uma asserção puder ser representada de mais de uma maneira, pela plausibilidade de outras asserções, todas as formas deverão dar o mesmo resultado.

Podemos resumir os três axiomas da seguinte forma

- D_1: Transitividade da plausibilidade: Representação por números reais
- D_2: Bom senso: Monotonicidade
- D_3: Diferentes análises devem dar o mesmo resultado: consistência

Há várias formas de usar a a palavra *consistência*. Aqui a usamos da seguinte forma. Impor que duas formas de análise devam dar o mesmo resultado não garante a consistência da teoria geral, no entanto uma teoria onde isso não ocorra será manifestamente inconsistente.

Diferentes autores param neste ponto de definir axiomas. Mas há mais dois pontos que precisam ser deixados explícitos. Talvez uma análise por parte de um filósofo ou lógico profissional permita distinguir seu status.

Todo operador na álgebra Booleana pode ser representado pelas operações conjunção (e) e negação ($\bar{\cdot}$) [4], isto é, o produto e a negação lógicas. A soma lógica pode ser obtida usando $A + B = \overline{\overline{A}\,\overline{B}}$. Precisamos então analisar a plausibilidade de asserções compostas usando esses operadores em termos das plausibilidade de asserções mais simples. Já que este conjunto de operadores é completo, esperamos que só tenhamos que analisar estes dois operadores.

Primeiro olhamos para o produto lógico. Novamente C se refere à informação subjacente e estamos interessados na plausibilidade $y = A_1A_2|C$. Há 4 plausibilidades que serão interessantes para esta análise:

$$x_1 = A_1|C,\ x_2 = A_2|C,\ x_3 = A_2|A_1C,\ x_4 = A_1|A_2C$$

. Notamos que deve haver uma dependência entre $A_1A_2|C$ e algum subconjunto de $\{x_i\}$, então

- D_4': Deve existir uma função F_p que relaciona $A_1A_2|C$ e algum subconjunto de $\{x_i\}$.

Porque um subconjunto? Qual subconjunto? Todos? Como decidir? Há 11 subconjuntos de dois ou mais membros: Seis ($\frac{4!}{2!2!}$) pares (x_i, x_j), quatro ($\frac{4!}{3!1!}$) triplas (x_i, x_j, x_k) e o conjunto inteiro $\{x_i\}$

Tribus analisou as 11 possibilidades e verificou que só há duas que sobrevivem a casos extremos (ver Apêndice 1 para uma prova). Os dois conjuntos são (x_1, x_3) e (x_2, x_4). Note que se o primeiro deles fosse um dos sobreviventes, o segundo também deveria ser pela simetria trazida pela comutatividade do produto lógico.

Porque não serve o subconjunto mais óbvio (x_1, x_2)? Seja A_1=' Helena usa um sapato esquerdo marrom' enquanto que A_2=' Helena usa um sapato direito preto' . A plausibilidade dessas duas asserções será julgada dada a seguinte informação C='Helena gosta de sapatos pretos e de sapatos marrons', e talvez seja possível concluir que as duas asserções são bastante plausíveis. Mas se tivessemos $y = F_p(x_1, x_2)$ poderiamos ser levados a pensar que 'Helena usa um sapato esquerdo marrom e um sapato direito preto' é bastante plausível.

Cox coloca o axioma na segunte forma:

- D_4: Deve existir uma função F_p que relaciona $A_1A_2|C$ e ($A_1|C$, $A_2|A_1C$ (que é (x_1, x_3)) .

[4] Este conjunto não é mínimo, mas é útil e claro.

Ou seja $y = F_p(x_1; x_3)$. Por comutatividade do produto lógico devemos ter $y = F_p(x_2; x_4)$

Note que agora será possível concluir que 'Helena usa um sapato esquerdo marrom e um sapato direito preto' pode ser pouco plausível por que precisamos saber a plausibilidade de 'Helena usa um sapato esquerdo marrom' dado que 'Helena usa um sapato direito preto' e isto pode ser pouco plausível. Porém, conhecendo a Helena...

Se uma asserção A dada informação C tiver a sua plausibilidade determinada então saberemos algo da plausibilidade da negação de A na mesma informação C. Em particular, suponha que C pertença a uma familia de asserções onde algum tipo de continuidade puder ser identificada. Suponha então que há uma 'pequena' mudança de informação C para C' o que acarreta uma também 'pequena' mudança na plausibilidade de A: $A|C$ passa a $A|C'$. O quê pode ser dito sobre $\overline{A}|C$ e $\overline{A}|C'$? O mais simples é que deve haver uma relação entre $A|C$ e $\overline{A}|C$, assim $A|C = F_S(\overline{A}|C)$. Para satisfazer o bom senso em casos que há continuidade, F_S deve ser uma função contínua nos reais:

- D_5: Deve existir uma função F_S que relaciona $A|C$ e $\overline{A}|C$

Esperamos ainda que F_S seja monotónica decrescente. Se a informação recebida torna mais plausível que 'choverá hoje', esperamos que torne menos plausível que ' não choverá hoje'.

A seguir mostraremos que estes desejos podem ser transformados em equações funcionais e assim determinar a teoria geral.

Antes um comentário e um desafio. Vamos encontrar, daqui a pouco,que a ferramenta adequada para lidar com informação incompleta é a teoria da probabilidade. Fazendo uso do mesmo princípio, que uma teoria geral deve ser válida em casos particulares, agora para o problema de atribuição e atualização de probabilidades chegaremos ao método de Máxima Entropia no capítulo 2. Esta é a entropia de Boltzmann-Gibbs-Shannon e terá um lugar central nos processos de inferência. Sistemas de processamento de informação evoluidos por seleção natural não necessariamente usam métodos de inferência que atingem os limites impostos por estas teorias. Certamente passaram por estágios no seu desenvolvimento onde a informação não era processada de forma ótima. Aqui fica a pergunta como desafio: Há formas de detectar falhas nos métodos de inferência de sistemas de inferência bioógicos? Poderiam ser inconsistências, quebra na transitividade ou no bom senso. Há várias áreas da ciência que lidam com estas perguntas. Você pode pensar em exemplos?

1.1.2 Equações funcionais

Os cinco axiomas estão naturalmente divididos em dois grupos. O segundo grupo (D_4 e D_5) especifíca a existência de certas funções. Para avançar devemos examinar o que os três primeiros axiomas impoem sobre as funções F_p e F_s ligadas respectivamente ao produto e soma lógicas

A regra do Produto

Queremos analisar produtos do tipo $y = A_1A_2|C$. Temos por D_3 e D_4 que

$$y = F_P(x_1; x_3) = F_P(x_2; x_4) \tag{1.1}$$

Queremos olhar para asserções mais complexas como $B_1B_2B_3|C$. Podemos identificar $A_1 = B_1B_2$ e $A_2 = B_3$, a eq. 1.1 nos dá

$$B_1B_2B_3|C = F_P(B_1B_2|C; B_3|B_1B_2C) \tag{1.2}$$

e podemos separar asserções compostas usando novamente a eq. 1.1

$$B_1B_2B_3|C = F_P(B_1B_2|C; B_3|B_1B_2C) = F_P(F_P(B_1|C; B_2|B_1C); B_3|B_1B_2C) \tag{1.3}$$

Note que a associatividade permite escrever $(B_1B_2)B_3|C = B_1(B_2B_3)|C$ e usando agora $A'_1 = B_1$ e $A'_2 = B_2B_3$ separar a asserção complexa em duas de outra forma. Equivalente à eq. 1.1 teremos

$$B_1B_2B_3|C = F_P(B_1|C; B_2B_3|B_1C) = F_P(B_1|C; F_P(B_2|B_1C; B_3|B_1B_2C)) \tag{1.4}$$

Consistência (D_3) impõe que F_P deve satisfazer:

$$F_p(F_P(x; y); z) = F_P(x; F_P(y; z)) \tag{1.5}$$

Esta equação, primeiramente estudada por Abel é chamada de "equação da associatividade".

Há mais de uma forma de provar a estrutura das suas soluções. Algumas são construtivas, mas a que escolhemos é a simples verificação que a solução proposta resolve a equação da associatividade [5]. Suponha w uma função monotónica e inversível , mas fora isso, arbitrária sobre os reais. Então , qualquer função F_P da forma

$$F_p(x; y) = w^{-1}[w(x)w(y)] \tag{1.6}$$

satisfaz a eq. 1.5. Para prová-lo basta substituir a eq. 1.6 do lado direito da eq. 1.5

$$\begin{aligned} F_p(F_P(x; y); z) &= w^{-1}[w(F_P(x; y))w(z)] \\ &= w^{-1}[w(w^{-1}[w(x)w(y)])w(z)] \\ &= w^{-1}[w(x)w(y)w(z)], \end{aligned} \tag{1.7}$$

e do lado esquerdo

$$\begin{aligned} F_p(x; F_P(y; z)) &= w^{-1}[w(x)w(F_P(y; z))] \\ &= w^{-1}[w(x)w(w^{-1}[w(y)w(z)] \\ &= w^{-1}[w(x)w(y)w(z)] \end{aligned} \tag{1.8}$$

[5] **D2** é necessário para mostrar que esta é a única forma possível

provando o que queriamos. Usamos a identidade $w^{-1}(w(x)) = x$.

Temos um resultado extremamente importante:

$$w(F_p(x;y)) = w(x)w(y) \tag{1.9}$$

ou em termos das plausibilidades

$$\begin{aligned} w(A_1A_2|C) &= w(A_1|C)w(A_2|A_1C) \\ w(A_1A_2|C) &= w(A_2|C)w(A_1|A_2C) \end{aligned} \tag{1.10}$$

Regra do Produto e informação completa

A análise de casos simples ainda vai nos dar muita informação sobre estas estruturas. Podemos olhar para o caso simples onde $A_1 = A_2$.Logo, o produto $A_1A_2 = A_1A_1 = A_1$ e temos, neste caso especial que a eq. 1.10 se reduz a

$$w(A_1|C) = w(A_1|C)w(A_1|A_1C). \tag{1.11}$$

Se A_1 dado C não for zero, temos que $w(A_1|A_1C) = 1$. Note que neste caso simples temos certeza de que A_1 será verdadeiro pois, isto é considerado sob a luz da informação que é verdadeiro. Certeza da verdade impõe que a função desconhecida w do número desconhecido (plausibilidade) $A_1|A_1C$ tenha valor 1.

Outro caso especial permite ver o que acontece quando temos certeza que uma asserção é falsa. Para tal tomemos $C \Rightarrow \overline{A_1}$, ou seja dizer dado C significa que A_1 é falsa. Segue que

$$w(A_1A_2|C) = w(A_1|C) \tag{1.12}$$

já que a plausibilidade de $A_1A_2|C$ independe de A_2 pois se A_1 é impossível em C , A_1A_2 também o será. Da mesma forma $A_1|A_2C = A_1|C$ para qualquer B que não seja incompatível com C. Assim

$$\begin{aligned} w(A_1A_2|C) &= w(A_2|C)w(A_1|A_2C) \\ w(A_1|C) &= w(A_2|C)w(A_1|C) \end{aligned} \tag{1.13}$$

e temos que a certeza da falsidade deverá ser representada por $w(A_1|C) = 0$ ou ∞. O convencional será tomar o zero para representar a total certeza da falsidade. A monotonicidade requerida de w nos leva então ao resultado

$$0 \leq w(x) \leq 1 \tag{1.14}$$

Note que ao aceitar o primeiro axioma, a estrutura dos números reais foi usada para fornecer uma classificação para as asserções. Agora encontramos que há uma transformação monotónica desse sistema de classificação que está entre zero e um. A monotonicidade não muda a ordem das classificações, e ainda obtivemos a eq. 1.10 que é chamada de regra do produto, e serve para analisar uma conjunção de asserções.

São estes números as probabilidades? Ainda não, temos que analisar as propriedades de asserções compostas por disjunções, isto é, somas. Para isso precisamos lidar com a negação.

A regra da Soma

Para somas usaremos o resultado (ver acima tabela 1.1) $\overline{A+B} = \overline{A}\,\overline{B}$. Se soubermos lidar com a negação, as propriedades de asserções obtidas por somas poderão ser analisadas.

A negativa da negativa de uma asserção é a própria asserção, logo D_5 leva, para $x = w(A|B)$

$$x = F_S(F_S(x)) \tag{1.15}$$

portanto

$$F_S(x) = F_S^{-1}(x) \tag{1.16}$$

ou

$$F_S^2(x) = x \tag{1.17}$$

A solução geral não é suficientemente restritiva. Imagine uma função par $v = h(u) = h(-u)$, com $|h'(u)| < 1$ e agora gire os eixos $(u, v) \to (x, y)$ no sentido antihorário por $\pi/2$. A função é agora simétrica em relação ao raio $y = x$ e portanto igual à sua inversa. Para restringir mais as soluções precisamos olhar o que acontece quando olhamos a negação junto com a regra do produto.

$$\begin{aligned}
w(A_1A_2|C) &= w(A_1|C)w(A_2|A_1C) = w(A_1|C)F_S(w(\overline{A_2}|A_1C)) \\
w(A_1A_2|C) &= w(A_1|C)F_S(\frac{w(A_1\overline{A_2}|C)}{w(A_1|C)}) \\
w(A_1A_2|C) &= w(A_2|C)F_S(\frac{w(A_2\overline{A_1}|C)}{w(A_2|C)})
\end{aligned}$$

onde a última linha decorre da simetria de troca de A_1 por A_2, logo, para quaisquer três asserções

$$w(A_1|C)F_S(\frac{w(A_1\overline{A_2}|C)}{w(A_1|C)}) = w(A_2|C)F_S(\frac{w(A_2\overline{A_1}|C)}{w(A_2|C)}). \tag{1.18}$$

Esta é uma relação entre quatro variáveis e não nos ajuda muito, devemos buscar um caso particular para poder impor restrições.

A_1	D	$\overline{A_2}$	$A_1\overline{A_2}$	$\overline{A_1}A_2$
1	1	1	1	0
1	0	0	0	0
0	1	0	0	1
0	0	0	0	1

$\overline{A_2} = A_1D$, D qualquer. Tabela 1.2

Por exemplo, no caso que $\overline{A_2} = A_1D$ para uma asserção D qualquer teremos $A_1\overline{A_2} = \overline{A_2}$ e $\overline{A_1}A_2 = \overline{A_1}$, ver tabela 1.2 e chamando

$$x = w(A_1|C)\ \ , y = w(A_2|C)$$

teremos

$$w(A_1|C)F_S(\frac{w(\overline{A_2}|C)}{w(A_1|C)}) = w(A_2|C)F_S(\frac{w(\overline{A_1}|C)}{w(A_2|C)}). \tag{1.19}$$

reduzida a

$$xF_S(\frac{F_S(y)}{x}) = yF_S(\frac{F_S(x)}{y}). \tag{1.20}$$

Defina agora $u = F_S(y)/x$ e $v = F_S(x)/y$. Isto e o resultado de derivar com respeito a x, a y e a x e y, obtemos, respectivamente

$$\begin{aligned} xF_S(u) &= yF_S(v) \\ F'_S(v)F'_S(x) &= F_S(u) - uF'_S(u) \\ F'_S(u)F'_S(y) &= F_S(v) - vF'_S(v) \\ uF''_S(u)F'_S(y)/x &= vF''_S(v)F'_S(x)/y \end{aligned} \tag{1.21}$$

Multiplicando a primeira e a última equação termo a termo, eliminamos x e y:

$$uF''_S(u)F'_S(y)F_S(u) = vF''_S(v)F'_S(x)F_S(v) \tag{1.22}$$

Usando agora a segunda e a terceira, eliminamos a dependência em x e y e após algumas manipulações, podemos separar as variáveis u e v:

$$\frac{uF''_S(u)F_S(u)}{(uF'_S(u) - F_S(u))F'_s(u)} = \frac{vF''_S(v)F_S(v)}{(vF'_S(v) - F_S(v))F'_S(v)} \tag{1.23}$$

e cada termo deve ser igual a uma constante c arbitrária, já que u e v podem ser independentes:

$$\frac{F''_S(u)}{F'_S(u)} = c(\frac{F'_S(u)}{F_s(u)} - \frac{1}{u}) \tag{1.24}$$

$$dF_S(x)'/F'_S = c(dF_S/F_S - dx/x) \tag{1.25}$$

que pode ser integrada para obter

$$F_S(x)' = A(F_S/x)^c \tag{1.26}$$

que pode ser integrada novamente

$$F_S(x)^{1-c} = Ax^{1-c} + K \tag{1.27}$$

Dado que F_S é sua própria inversa podemos determinar A e K

$$\begin{aligned} F_S(F_S(x)) &= (AF_S(x)^{1-c} + K)^{\frac{1}{1-c}} \\ = x &= (A(Ax^{1-c} + K) + K)^{\frac{1}{1-c}} \end{aligned} \tag{1.28}$$

de onde tiramos que $AK + K = 0$ e $A^2 = 1$, logo, com $m = 1 - c$ uma constante arbitrária

$$F_S(x) = (1 - x^m)^{1/m} \tag{1.29}$$

$$F_S^m(x) + x^m = 1 \tag{1.30}$$

Assim, a uma asserção e sua negação, sob a mesma informação B satisfazem

$$w^m(A|B) + w^m(\overline{A}|B) = 1 \tag{1.31}$$

Temos que uma função monotônica arbitrária w elevada a um número m arbitrário, aplicada ao número $A|B$ desconhecido que representa a plausibilidade e o equivalente da negação, satisfazem uma equação com valores numéricos definidos! Assim introduzimos uma novo número para descrever a crença na veracidade da asserção de que A é verdadeiro sob informação B e usamos a notação $p(A|B) = w^m(A|B)$ e agora

$$0 \leq p(A|B) \leq 1 \tag{1.32}$$

$$p(A|B) + p(\overline{A}|B) = 1 \tag{1.33}$$

A eq. 1.10 passa agora a ser:

$$\begin{aligned} p(A_1A_2|C) &= p(A_1|C)p(A_2|A_1C) \\ &= p(A_2|C)p(A_1|A_2C) \end{aligned} \tag{1.34}$$

Isto é o famoso teorema de Bayes. Como veremos mais adiante este teorema esta por trás da regra de Bayes para atualização de probabilidades em face a nova informação.

1.1.3 Disjunção

A soma e o produto lógicos formam um conjunto completo de operadores booleanos. Sabemos como proceder na análise de asserções compostas por conjunções e suas negações. Deve ser possível determinar como proceder na soma lógica ou disjunção. Usamos $\overline{A+B} = \overline{A}\,\overline{B}$, assim

$$\begin{aligned} P(A+B|C) &= 1 - P(\overline{A+B}) \\ &= 1 - P(\overline{A}\,\overline{B}|C) \\ &= 1 - P(\overline{A}|C)P(\overline{B}|\overline{A}C) \\ &= 1 - (1 - P(A|C))(1 - P(B|\overline{A}C) \\ &= P(A|C) + P(B|\overline{A}C)(1 - P(A|C) \\ &= P(A|C) + P(\overline{A}B|C) \\ &= P(A|C) + P(B|C)P(\overline{A}|BC) \\ &= P(A|C) + P(B|C)(1 - P(A|BC)) \\ &= P(A|C) + P(B|C) - P(AB|C) \end{aligned} \tag{1.35}$$

A e B são mutuamente exclusivos se $P(AB|C) = 0$, neste caso

$$P(A+B|C) = P(A|C) + P(B|C) \tag{1.36}$$

1.1.4 Probabilidades

A monotonicidade de $w^m(x)$ é fundamental. Se achamos que $A|C > B|C$ e tivermos $w^m(A|C) > w^m(B|C)$ a transitividade discutida anteriormente não é alterada. Isto é, a ordem de preferência inicialmente fornecido pelas plausibilidades não muda frente às *reparametrizações* introduzidas por $w^m(x)$. Este é

o conteúdo dos teoremas de Cox: uma atribuição de números para descrever as crenças em asserções, dada a informação, que satisfaça os casos particulares, pode ser mudada de forma a não alterar a ordem de preferências e a satisfazer as regras da probabilidade. Tem cheiro e cor de probabilidade e tem todas as propriedades das probabilidades. Não falaremos mais sobre plausibilidade. Não sabiamos o que era, e a abandonamos como a um andaime, após ter contruido o edifício da teoria de probabilidades. Obviamente este exercício não forneceu os valores das probabilidades. Que bom, senão fechariam os institutos dedicados ao estudo e às aplicações das probabilidades. Mais sérios, podemos dizer que a nossa grande preocupação agora será dirigida à busca de técnicas que baseadas na informação disponível permitam atribuições ou talvez o problema associado mas diferente, de atualização dos números associados a probabilidades dos eventos ou asserções de interesse quando recebemos nova informação. Esta é a preocupação central da inferência e da teoria de aprendizado e nos levará à introdução da idéia de entropia. A entropia no sentido de teoria de informação está intimamente ligada à idéia de entropia termodinâmica e mais ainda à de Mecânica Estatística como veremos mais tarde. Poderemos afirmar que a Mecânica Estatística foi a primeira teoria de informação, embora não seja costumeiro colocá-la nessa luz.

1.1.5 Informação Incompleta

Vejamos agora alguns exemplos da utilização destes resultados em casos simples onde há informação incompleta.

Voltemos agora aos silogismos iniciais. Suponha que

- A="Está chovendo"
- B="Há nuvens"
- $C =$ "$A \rightarrow B''$

Note que a implicação lógica não segue da causalidade física. Chove porque há nuvens do ponto de vista de causalidade, mas do ponto de vista lógico saber que chove obriga à conclusão que deve haver nuvens. Suponha que seja dada a informação B, ou seja é dado que há nuvens. Dentro da lógica aristotélica nada podemos dizer. Devemos com base nisso desprezar por ilógicos quem nos aconselha a levar um guarda-chuva porque há nuvens? Vejamos o que nos diz a teoria das probabilidades. Neste caso o teorema de Bayes começa a mostrar a sua força. A probabilidade $P(A|CI)$ representa a crença que esteja chovendo, sob a informação C, mas não levando em conta se há ou não nuvens. Também leva em conta I, tudo o que é sabido sobre o clima nesta estação do ano, podendo ser muita informação ou nenhuma. Não importa efetivamente que número $P(A|CI)$ seja, estará entre zero e um. Esta probabilidade é dita *a priori* em relação a B. Uma vez que se recebe e incorpora a informação que efetivamente há nuvens,

ou seja B, então passaremos a $P(A|BCI)$, outro número, que é chamada a probabilidade *a posteriori* ou simplesmente posterior. Aplicando Bayes

$$P(A|BCI) = \frac{P(A|CI)P(B|ACI)}{P(B|CI)}, \tag{1.37}$$

que relaciona a probabilidade *a priori* e a posterior. Cortando e deixando para depois uma discussão longa sobre inferência, podemos dizer que é razoável que usemos a posterior para decidir se levaremos ou não o guarda-chuvas. A probabilidade $P(B|ACI)$ recebe o nome de verossimilhança (*likelihoood* e poderia ser calculada se tivessemos um modelo sobre a influência de A em B, mas é isso o que temos, este é um caso de informação completa! Temos certeza da veracidade de B se AC for dado. Assim

$$P(B|ACI) = 1. \tag{1.38}$$

O quê pode ser dito sobre o denominador $P(B|CI)$? O mínimo que pode ser dito é que

$$P(B|CI) \leq 1. \tag{1.39}$$

Substituindo estes resultados obtemos

$$P(A|BCI) \geq P(A|CI), \tag{1.40}$$

a probabilidade que atribuiremos a que A seja verdade é maior ou igual se levarmos em conta o fato que há nuvens, que aquela que atribuimos sem saber se há nuvens ou não. Finalmente nos diz que a pessoa que percebe que há nuvens e leva o quarda-chuvas está agindo de forma lógica, não dentro da lógica aristotélica, mas segunda a extensão da lógica para casos de informação incompleta, representada pela teoria das probabilidades. Vemos que o bom senso diário desta situação pode ser deduzido dos desejos impostos por Cox.

Suponha outro caso de informação imcompleta. Agora A é dado como falso. Continuaremos a insistir que não podemos dizer nada sobre B do ponto de vista da lógica? O teorema de Bayes, nos diz

$$P(B|\overline{A}CI) = \frac{P(B|CI)P(\overline{A}|BCI)}{P(\overline{A}|CI)}, \tag{1.41}$$

e também sabemos que $P(A|BCI) \geq P(A|CI)$ da análise anterior. Ainda mais, temos que $P(A|BCI) = 1 - P(\overline{A}|BCI)$ e $P(A|CI) = 1 - P(\overline{A}|CI)$, portanto

$$P(B|\overline{A}CI) \leq P(B|CI) \tag{1.42}$$

levando à conclusão que se não está chovendo, devemos atribuir uma probabilidade menor a que haja nuvens. Quem está mais disposto a carregar um chapeú porque recebeu informação que não está chovendo, age de forma lógica.

Exemplo

Consideremos um exemplo clássico de testes médicos. Um teste médico serve para ajudar a determinar se um paciente está doente, mas ele não é perfeito e há evidência, baseado na história que há falsos positivos e falsos negativos Consideremos as asserções

- D="paciente está doente"
- A="teste deu positivo"

junto com os dados sobre

- especificidade: $P(A|D) = .90$, a probabilidade de dar positivo no teste na condição de estar doente
- sensibilidade: $P(A|\overline{D}) = .2$, a probabilidade de teste dar positivo no caso em que o paciente não está doente,

Vemos que o teste é bastante específico (90%) e bastante sensível ($(80 = 100 - 20)\%$).

Suponha que seu resultado no teste deu positivo, A é verdade. Isto significa que está doente? Há possibilidade de erros portanto não temos informação completa. Qual é a pergunta que devemos fazer? Pode não ser o mais óbvio a se fazer quando se recebe uma notícia ruim, mas em geral devemos aplicar o teorema de Bayes. Assim poderemos calcular $P(D|AI)$ que é o que realmente interessa, a probabilidade de ter a doença,

$$P(D|AI) = \frac{P(D|I)P(A|DI)}{P(A|I)}, \tag{1.43}$$

e também

$$P(\overline{D}|AI) = \frac{P(\overline{D}|I)P(A|\overline{D}I)}{P(A|I)}, \tag{1.44}$$

os denominadores são inconvenientes e os eliminamos olhando para a razão

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)}. \tag{1.45}$$

Após considerar a equação acima percebemos que não temos dados suficientes para entrar em pânico. A razão entre as probabilidades que nos interessa é $P(D|AI)/P(\overline{D}|AI)$ depende de dados que temos, sobre a especificidade e sensibilidade do teste e de dados que não temos sobre a distribuição da doença na população. A teoria que não pode nesta altura nos dar a resposta que buscamos, faz a segunda melhor coisa, indicando que informaçao adicional devemos procurar. Após esta análise voltamos ao médico e perguntamos se ele tem informação sobre a distribuição *a priori* da doença na população caracterizada

por I. Suponha que recebamos informação que $\frac{P(D|I)}{P(\overline{D}|I)} = .99/.01$, só 1% da população tem a doença. Segue que

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)} = \frac{.01 \times .90}{.99 \times .20} = 0.045. \tag{1.46}$$

ou seja a probabilidade de não ter a doença é aproximadamente .95. Não que isto seja uma boa notícia, afinal a probabilidade que era de 1% de ter a doença passou para 4.5% : aumentou quase cinco vezes. Mas não devemos ainda entrar em pânico nem jogar fora a informação que ganhamos com o teste.

Capítulo 2

Exemplos de algumas distribuições

2.1 Distribuição de Bernoulli ou Binomial

Seja I a seguinte asserção: "Uma moeda é jogada de forma que cai com a cara para cima ou para baixo." Escrevemos $s_i = 1$ se a $i-$ésima jogada cai cara para cima e $s_i = -1$ se não. Supomos que os resultados de jogadas anteriores , ou posteriores, não influenciam os resultados:

$$P(s_i|\{s_j\}_{j\neq i}I) = P(s_i|I)$$

para qualquer conjunto $\{s_j\}_{j\neq i}$ de jogadas que não incluam i, então

$$P(s_i s_k|I) = P(s_i|I)P(s_k|s_i I) = P(s_i|I)P(s_k|I)$$

e

$$P(s_1 s_2....s_n|I) = \prod_{i=1}^{n} P(s_i|I)$$

Suporemos que as condoções em que é realizado o experimento não muda com o tempo e chamaremos $P(s_i = 1|I) = h$, (independente de i), e $P(s_i = -1|I) = 1 - h$.

Assim numa particular sequência que inclua k caras e $n - k$ coroas temos

$$P(s_1 s_2....s_n|I) = h^k(1-h)^{n-k}. \tag{2.1}$$

Mas queremos calcular a probabilidade $P(k|nI)$ de obter k caras independentemente da ordem. Como as sequências são mutuamente exclusivas

$$P(k|nI) = \sum_{Seq_n} P(s_1 s_2....s_n|I) = C_k^n h^k(1-h)^{n-k}, \tag{2.2}$$

onde C_k^n é o número de sequências de n tentativas e k caras (ou sucessos).

Dada a indepenência das jogadas podemos calcular $P(k|nI)$ considerando que a última jogada pode ter saido cara o coroa. Mostre que

$$P(k|nI) = P(k|n-1I)(1-h) + P(k-1|n-1I)h \tag{2.3}$$

Assim deduzimos a relação de recorrência

$$C_k^n = C_{k-1}^{n-1} + C_k^{n-1} \tag{2.4}$$

que junto com as condições iniciais $C_1^1 = 1$ e $C_0^1 = 1$ e o extremo $C_n^n = 1$, permite obter os seus valores.

Exercício Mostre que $C_k^n = \frac{n!}{k!(n-k)!}$

Exercício Mostre que $P(k|nI)$ esta normalizada: $P\sum_{k=1}^{n}(k|nI) = 1$.

NEstas notas de forma provisória este exercício continua no proximo capitulo

2.2 Distribuiçao de Poisson

(Baseado em Jaynes) Considere um detector de reocs que faz clic quando detecta o reoc. Assim, chamando A:"ocorre clic", A asserção I' é dada por

I': "Existe um número λ, real positivo tal que no intervalo suficientemente pequeno dt, entre t e $t+dt$ a probabilidade de que A é proporcional a λ e ao tamanho do intervalo. "

$$P(A|\lambda I') = \lambda dt \tag{2.5}$$

A asserção I é dada por

I: "I' e para qualquer Q que não implique A, conhecimento de λ torna Q desnecessário. "

$$P(A|\lambda I) = \lambda dt, \quad P(A|Q\lambda I) = P(A|\lambda I) \tag{2.6}$$

O que acontece para intervalos nao microscópicos?

Seja $h(t)$ a probabilidade de que não ocorrem clics no intervalo $(0,t)$. E seja R o evento que não há clics em $(0, t+dt)$. Vemos que esta asserção é formada pela conjunção

R: (Não há clics no intervalo $(0,t)$) e (Não há clics no intervalo $(t, t+dt)$).

Segue que, pela regra do produto

$$h(t+dt) = h(t)(1-\lambda dt) \tag{2.7}$$

$$\frac{dh}{dt} = -h(t)\lambda, \tag{2.8}$$

Dado que $h(0) = 1$, a probabilidade de clics no intervalo nulo, temos que $h(t) = exp(-\lambda t)$

Considere o evento B, dado por

B:"no intervalo (0,t) há exatamente n clics, nos instantes $(t_1, t_2, ...t_n)$, medidos com tolerância $(dt_1, dt_2....dt_3)$, com $(t_1 < t_2 < ... < t_n$.

B é uma conjunção de $2n+1$ asserções:

B: "(não há clics em $(0,t_1)$) e (clic em dt_1) e (não há clics em (t_1,t_2) ....e (clic em dt_n) e (não há clics em (t_n,t)"

$$P(B|\lambda I) = e^{-\lambda t_1}(\lambda dt_1)e^{-\lambda(t_2-t_1)}(\lambda dt_2)..(\lambda dt_n).e^{-\lambda(t-t_n)}, \tag{2.9}$$

simplificando temos

$$P(B|\lambda I) = e^{-\lambda t}\lambda^n dt_1 dt_2...dt_n. \tag{2.10}$$

Qual é a probabilidade $P(n|\lambda tI)$ de que ocorram exatamente n clics em quaisquer instante dentro do intervalo $(0,t)$?

Como os B's são mutuamente exclusivos

$$P(n|\lambda tI) = \int_0^t dt_n \int_0^{t_n} dt_{n-1}... \int_0^{t_2} dt_1 e^{-\lambda t}\lambda^n dt_1 dt_2...dt_n. \tag{2.11}$$

O resultado é $P(n|\lambda tI) = \frac{e^{-\lambda t}}{n!}(\lambda t)^n$, que depende conjuntamente do produto $\theta = \lambda t$. A distribuçao $P(n|\theta) = \frac{e^{-\theta}}{n!}(\theta)^n$ é chamada de Poisson.

Exercício Mostre que $E[n] = var[n] = \theta$

Suponha agora que n_1 e n_2 sejam as contagens nos intervalos $(0,t_1)$ e $(0,t_2)$. Queremos responder as seguintes perguntas

quanto é $P(n_2|n_1,t_1,t_2,\lambda,I)$?

quanto é $P(n_1|n_2,t_1,t_2,\lambda,I)$?

Ou seja, ocorrem n_1 clics no primeiro intervalo e depois ocorrem $n_2 - n_1$ clics no intervalo (t_1,t_2).

Assim

$$\begin{aligned} P(n_1n_2|\lambda t_1t_2I) &= P(n_1|\lambda t_1 I)P(n_2|n_1\lambda t_1t_2I) && (2.12)\\ &= [\frac{e^{-\lambda t_1}}{n_1!}(\lambda t_1)^{n_1}][\frac{e^{-\lambda(t_2-t_1}}{(n_2-n_!)!}(\lambda(t_2-t_1))^{(n_2-n_1)}] && (2.13)\end{aligned}$$

e rearranjando

$$\begin{aligned} P(n_1n_2|\lambda t_1t_2I) &= [\frac{e^{-\lambda t_2}}{n_2!}(\lambda t_2)_2^n]\ [C_{n_1}^{n_2}(\frac{t_1}{t_2})^{n_1}(1-\frac{t_1}{t_2})^{(n_2-n_1)}] && (2.14)\\ &= P(n_2|\lambda t_2 I)P(n_1|n_2\lambda t_1t_2I) && (2.15)\end{aligned}$$

$$\begin{aligned} P(n_1|n_2\lambda t_1t_2I) &= [C_{n_1}^{n_2}(\frac{t_1}{t_2})^{n_1}(1-\frac{t_1}{t_2})^{(n_2-n_1)}] && (2.16)\\ &= P(Bin(n_2,\frac{t_1}{t_2}) = n_1) && (2.17)\end{aligned}$$

note que nao depende de λ!

Capítulo 3

Probabilidades e o Teorema do Limite Central

3.1 Introdução: Kolmogorov e as probabilidades

Kolmogorov introduziu na décadada dos trinta [1] os seus famosos axiomas para a teoria das probabilidades. No seu livro ele declara que não vai entrar no debate filosófico sobre o significado de probabilidades e depois dá uma pequena justificativa dos axiomas com base na interpretação freqüentista de von Mises. No capítulo anterior descrevemos os motivos que nos levam a achar tal posição, isto é freqüentista, incompleta e até, como mostraremos abaixo, errada. Pelo contrário, os axiomas de Kolmogorov, que codificam o bom senso da área já existente no trabalho de Laplace, podem ser vistos equivalentes aos resultados obtidos no capítulo 1.

Kolmogorov começa for considerar E uma coleção de elementos $A, B, C....$ que são eventos elementares e em nossa discussão anterior chamamos de asserções. $\mathcal{F}$ é o conjunto de todos os subconjuntos de E. Os axiomas são

- AK1) E pertence a $\mathcal{F}$
- AK2) $\mathcal{F}$ é um σ-campo,

isto é, se $A \in \mathcal{F}$ e $\overline{A} = E - A$, então $\overline{A} \in \mathcal{F}$

- AK3) $\mathcal{F}$ é fechado ante uniões contáveis,

ou seja se $A_i, A_j \in \mathcal{F}$ então $A_i + A_j \in \mathcal{F}$

- AK4) Existe uma função $P(A)$ que é uma medida de integração tal que
- (4.1) $P(E) = 1$

[1] e.g. ver http://www.mathematik.com/kolmogorov

- (4.2) Para todo $A \in \mathcal{F}$, $P(A) \geq 0$
- (4.3) Se A_i e A_j são disjuntas, isto é $A_i A_j = 0$, ou ainda não tem elementos em comum então $P(A_i + A_j) = P(A_i) + P(A_j)$

Vemos que estes axiomas estão de acordo com os resultados do capítulo anterior. A probabilidade da *certeza* é 1 por AK4.1; a probabilidade está entre zero e um por AK4.2; e a probabilidade da disjunção de asserções que não tem elementos em comum é a soma das probabilidades. Notamos porém a falta de uma regra para o produto. Kolmogorov também e na página 6 ele introduz as probabilidades condicionais através de

$$p(A|B) = \frac{P(AB)}{P(B)} \tag{3.1}$$

de onde segue para a prova do teorema de Bayes.

Exercício

O que falta para poder obter o teorema de Bayes?

Se uma vez estabelecidos os axiomas, partirmos para as aplicações matemáticas, não haverá nenhuma diferença de resultados. Enfatizamos que as diferenças que temos são sobre a motivação dos axiomas e com a interpretação da idéia de probabilidades.

3.2

A partir de agora introduziremos alguns resultados matemáticos que serão úteis no desenrolar do curso. Em particular estamos interessados em grandezas físicas descritas por variáveis que tomam valores em intervalos dos reais, que chamaremos I.

No que segue lidaremos com asserções do tipo "a variável X toma valores entre x e $x + dx$". Não interessa ainda como, mas suponha que atribuimos um número a esta probabilidade. Introduziremos a densidade $P(x)$ tal que a probabilidade de que "a variável X toma valores entre x e $x + dx$" é dada por $P(x)dx$. A função $P(x)$ não é uma probabilidade mas é chamada de densidade de probabilidade [2]. Teremos então que

- $P(x) \geq 0$
- $\int_I P(x)dx = 1$

Suponha que queremos ao falar de X dar um número que nos dé alguma informação sobre os seu valor. A informação disponível será equivalente à densidade para todo x. Isto talvez seja muito. Queremos poder comunicar o valor de x com um número, isto é um estimador ou estimatica de X. Há várias possibilidades e cada uma tem utilidade

[2] Usamos a letra P por motivos históricos e evntualmente a chamaremos de probabilidade, por preguiça.

- (1) $x_M = maxargP(x)$
- (2) $< x >= E[x] = \int_I xP(x)dx$
- (3) x_m tal que $\int_{x \leq x_m} P(x)dx = \int_{x \geq x_m} P(x)dx$

estes números recebem os nomes de (1) moda, (2) valor esperado ou média e (3) mediana. Podemos pensar em outros.

Pode ser muito útil caracaterizar a distribuição pelas *flutuações* em torno da média: quanto se afasta x da su média, $\Delta x = x - < x >$. Novamente podemos olhar para a média, só que agora das flutuações e vemos que $< \Delta x >= 0$, isto não significa que a flutuação não é útil, só que por construção a sua média é nula. A média do seu quadrado é muito útil:

$$\sigma_x^2 =< (x - < x >)^2 > . \tag{3.2}$$

σ_X recebe o nome de variância. Mostre que $\sigma_x^2 =< x^2 > - < x >^2$.

Exercício

Pense no significado de cada um dos estimadores e da variância σ_X e proponha outros estimadores.

O valor esperado será muito usado no que segue. Podemos generalizar a sua definição e introduzir os momentos de uma distribuição:

- $< x^n >= E[x^n] = \int_I x^n P(x)dx$

para valores inteiros de n (claro que caso a integral exista). A notação que usamos de alguma forma deixa esquecida a idéia que a probabilidade depende da informação disponível portanto usaremos [3] a notação

- $< x^n >_{|C} = E[x^n|C] = \int_I x^n P(x|C)dx$

para identificar claramente que estes são os momentos de X sob a informação C

Os momentos centrais são definidos da mesma forma mas para a variável deslocada para que a média seja nula:

- $< (x - < x >)^n >_{|C} = \int_I (x - < x >)^n P(x|C)dx$

As grandezas de interesse em Mecânica Estatística serão tipicamente originadas por somas de grande número de outras variáveis, por exemplo a energia de um gás tera contribuições das energias cinéticas de cada molécula mais as interações enter elas. Suponha que $Y = X_1 + X_2$. O que Y significa? Do ponto de vista de aritmética não insultaremos o leitor. Significa o óbvio. Do ponto de vista de asserções, temos um conjunto de asserções simples do tipo "a variável X_i toma valores entre x_i e $x_i + dx_i$" para $i = 1, 2$ e suponha que de alguma

[3] usaremos esta notação às vezes, pois usaremos o direito de ser inconsistentes na notação, se isso não confundir o leitor

forma atribuimos números a suas probabilidades. Queremos analisar, sob essa informação a asserção "a variável Y toma valores entre y e $y+dy$". Notemos que a asserções compostas A_1 ="$x_1 = .17$ e $x_2 = .25$" e A_2 ="$x_1 = .42$ e $x_2 = 0$." levam à mesma conclusão sobre o valor de Y. Mas elas são disjuntas no sentido que A_1A_2 como produto lógico não pode ser verdade. As duas nào podem ser simultaneamente verdadeiras. A probabilidade da soma lógica $A_1 + A_2$ é então a soma das probabilidades. Mas há outros casos de conjunções que dão o mesmo resultado para Y e devem ser levadas em conta: devemos somar sobre todas elas. Olharemos para somas deste tipo, $Y = X_1 + X_2 + X_3 + ...X_N$, quando o número de termos na soma é muito grande. Lembrem que o número de átomos em algums poucos gramas é da ordem de 10^{23}.

3.3 Convoluções e Cumulantes

Considere variáveis idénticas X_i que tomam valores reais x e consideremos *amostras* $\{x_1, x_2, ...\}$ tal que $P(X_i)$ é o mesmo para todo i. Também introduzimos a noção de independência entre duas variáveis, se X_i e X_j são tais que $P(X_i|X_j) = P(X_i)$ então segue que $P(X_iX_j) = P(X_i)P(X_j)$ as distribuições conjuntas são o produto das distribuições de cada uma. Consideraremos o caso em que para qualquer $i \neq j$, os X_i são independentes entre si e são igualmente distribuidos [4]. Estamos interessados em somas de $Y = \sum_{i=1..n} x_i$. Em particular, qual é a distribuição de $P(Y|N = n)$? Comecemos com $N = 2$, a probabilidade que Y tenha um valor entre y e $y + dy$ é obtida a partir de todas as formas que $y \leq x_1 + x_2 \leq y + dy$, com pesos iguais à probabilidade de ocorrência de x_1 e x_2. Ver a figura 3.1. Para ser específicos chamaremos $P(X_i)$ a distribuição de valores de x_i, embora estejamos considerando que independe de i. A asserção que o "valor de Y esta entre y e $y + dy$" é a soma lógica de todas as asserções do tipo "X_1 tem valor x_1 e X_2 tem valor x_2", restritas ao caso em que $y \leq x_1 + x_2 \leq y + dy$ e portanto tem probabilidade

$$P(y|N = 2)dy = \int_{y \leq x_1 + x_2 \leq y+dy} dx_1 dx_2 P(x_1)P(x_2), \tag{3.3}$$

pois cada par de valores temos uma asserção disjunta. O vínculo $y \leq x_1 + x_2 \leq y + dy$ pode ser removido introduzindo a função χ_A que é 1 se a condição A for satisfeita e zero se não [5].

$$P(y|N = 2)dy = \int \chi_{y \leq x_1 + x_2 \leq y+dy} dx_1 dx_2 P(x_1)P(x_2), \tag{3.4}$$

[4] Independentes e igualmente distribuidos: usualmente abreviado por i.i.d.

[5] χ_A é chamada a função característica do intervalo ou conjunto A, não confunda com a função característica da distribuição de probabilidades definida abaixo.

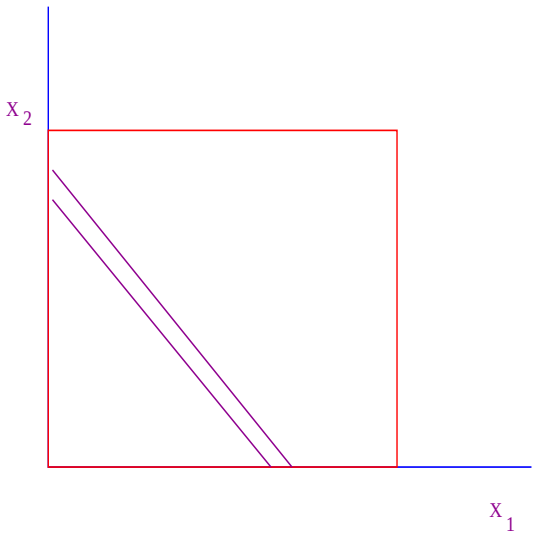

Figura 3.1: No plano X_1X_2 temos a região onde o valor de Y está entre y e $y + dy$. Todos os pares x_1 e x_2 nela contribuem para a probabilidade de Y

onde agora a integração é sobre todo o domínio de (x_1, x_2). Introduzimos uma representação para χ em termos da integral de uma seqüência de funções $\delta_n(A)$:

$$\begin{aligned} \delta_n(y \leq x_1 + x_2 \leq y + \Delta y_n) &= \frac{1}{\Delta y_n}, se \;\; y \leq x_1 + x_2 \leq y + \Delta y \\ &= 0, se \; não \end{aligned} \tag{3.5}$$

e obtemos, tomando o limite para $n \leftarrow \infty$, tal que Δy_n va para zero,

$$P(y|N = 2) = \int dx_1 dx_2 P(x_1)P(x_2)\delta(y - x_1 + x_2), \tag{3.6}$$

$$P(y|N = 2) = \int dx P(x)P(y - x), \tag{3.7}$$

isto é, a convolução de $P(x_1)$ e $P(x_2)$ denotada por $(P * P)(y)$.

Exercício

Discuta a diferença entre $P(Y)$ a distribuição da soma e $P(X_1X_2)$, para o produto lógico que às vezes denotaremos por $P(x_1, x_2)$ a distribuição conjunta de X_1 e X_2. Considere também a variável Z que toma valores iguais ao produto dos valores de X_1 e X_2 [6]

Suponha que $P(x)$ satisfaz as seguintes condições:

[6] Se você achar este problema ofensivo continue em frente. Se achar difícil, volte atrás e pense. Este tipo de dúvida é comum.

- $\int P(x)dx = 1$, $P(x) \geq 0$ para todo x
- $< x >= \int_{-\infty}^{\infty} xP(x)dx < \infty$,
- $< x^2 >= \int_{-\infty}^{\infty} x^2 P(x)dx < \infty$,

podemos introduzir a transformada de Fourier (TF) [7] e a inversa

$$\hat{P}(k) = \int_{-\infty}^{\infty} e^{-ikx} P(x)dx \tag{3.8}$$

$$P(x) = \int_{-\infty}^{\infty} e^{ikx} \hat{P}(k)\frac{dk}{2\pi} \tag{3.9}$$

A TF de uma distribuição de probabilidades é chamada de função característica. Tomemos a TF dos termos da equação 3.7, e usando :

$$\delta(k) = \int \frac{dx}{2\pi} e^{ikx} \tag{3.10}$$

obtemos

$$\begin{aligned} \hat{P}(k|N=2) &= \int dydx e^{-iky} P(x)P(y-x), \\ &= \int \frac{dxdydk_1dk_2}{(2\pi)^2} \hat{P}(k_1|1)\hat{P}(k_2|1)e^{-iky+ik_1x+ik_2(y-x)}. \end{aligned} \tag{3.11}$$

Integrando sobre x e usando a representação da delta:

$$\begin{aligned} &= \int \frac{dydk_1dk_2}{2\pi} \hat{P}(k_1|1)\hat{P}(k_2|1)e^{-iky+ik_2y}\delta(k_1-k_2), \\ &= \hat{P}(k|1)\hat{P}(k|1) = \hat{P}^2(k|N=1) \end{aligned} \tag{3.12}$$

Para a soma de $N = n$ variáveis x_i

$$P(y|N=n) = \int \prod_{i=1\ldots n} dx_i P(x_1)P(x_2)...P(y - \sum_{i=1}^{n-1} x_i), \tag{3.13}$$

ou, introduzindo uma integral mais

$$P(y|N=n) = \int \prod_{i=1\ldots n} dx P(x_1)P(x_2)...P(x_n)\delta(y - \sum_{i=1}^{n} x_i), \tag{3.14}$$

obtemos

$$\hat{P}(k|N=n) = \hat{P}^n(k|N=1) \tag{3.15}$$

e a inversão da transformada nos dá a distribuição de $P(y|N=n)$. No espaço de Fourier a convolução é simples produto, ou seja vamos para o espaço de Fourier,

[7] Para que exista é suficiente ainda que P seja seccionalmente contínua em cada intervalo $[-M, N]$ e definir $\hat{P} = \lim_{N,M\to\infty} \int_{-M}^{N} e^{-ikx} P(x)dx$

multiplicamos e depois voltamos ao espaço original fazendo a transformação inversa.

Caso as funções características sejam positivas podemos tomar o logarítmo de cada lado da equação 3.15 e dado que produtos, ao tomar logaritmos, viram somas, temos

$$\log \hat{P}(k|N=n) = \sum_{i}^{n} \log \hat{P}(k|N=1) = n \log \hat{P}(k|N=1), \tag{3.16}$$

que nos leva a discutir os cumulantes $\{C_s(n)\}$ através da expansão em série de potências de ik

$$log\hat{P}(k|N=n) = \sum_{s=0}^{\infty} C_s(n) \frac{(-ik)^s}{s!} \tag{3.17}$$

e a equação 3.16 nos indica o motivo do nome dos cumulantes: a aditividade (ou acúmulo) ante convoluções

$$C_s(n) = nC_s(1), \tag{3.18}$$

mas as convoluções vem das somas de variáveis aleatórias. Quando variáveis aleatórias indepndentes se somam, os cumulantes da distribuição da soma é a soma dos cumulantes das distribuições.

Mas qual é a interpretação dos cumulantes? Pela definição através da série de potências, vemos que em termos da função característica

$$C_s = \frac{1}{(-i)^s} \frac{d^s \log \hat{P}}{dk^s}|_{k=0} \tag{3.19}$$

Podemos calcular alguns dos primeiros,

$$\begin{aligned}
\log \quad & \hat{P}(k|N=1) = \log \int e^{-ikx} P(x) dx \\
= \quad & \log \int \sum_{s=0}^{\infty} \frac{(-ikx)^s}{s!} P(x) dx \\
= \quad & \log(1 + \sum_{s=1}^{\infty} \frac{(-ik)^s}{s!} < x^s >) \\
= \quad & \sum_{s_1=1}^{\infty} \frac{(-ik)^{s_1}}{s_1!} < x^{s_1} > -\frac{1}{2} \sum_{s_1,s_2=1}^{\infty} \frac{(-ik)^{s_1+s_2}}{s_1! s_2!} < x^{s_1} >< x^{s_2} > \\
+ \quad & \frac{1}{3} \sum_{s_1,s_2,s_3=1}^{\infty} \frac{(-ik)^{s_1+s_2+s_3}}{s_1! s_2! s_3!} < x^{s_1} >< x^{s_2} >< x^{s_3} > +...
\end{aligned} \tag{3.20}$$

onde usamos $\log(1+u) = -\sum_{l=1}^{\infty} (-u)^l / l$. Juntando os termos com a mesma potência de k obtemos os cumulantes em função dos momentos $< x^s >$:

$$\begin{aligned} C_0 &= 0, \\ C_1 &= < x >, \\ C_2 &= < x^2 > - < x >^2, \\ C_3 &= < x^3 > -3 < x^2 >< x > +2 < x >^3, \\ C_4 &= < x^4 > -4 < x^3 >< x > -3 < x^2 >^2 +12 < x^2 >< x >^2 \\ &- 6 < x >^4, \end{aligned} \tag{3.21}$$

O cumulante para $s = 0$ é nulo, devido à normalização da distribuição. Para $s = 1$ é a média e para $s = 2$ é a variância, ficando mais complicados para valores maiores de s .

Fica mais interessante se olharmos além da soma Y, para $Z = \frac{Y}{\sqrt{n}}$ e para $W = \frac{Y}{n}$. Colocamos um índice para indicar a que variável se refere o cumulante e obtemos a propriedade que é chamada de homgeneidade:

$$nC_1^x = C_1^Y(n) = \sqrt{n}C_1^Z(n) = nC_1^W(n), \tag{3.22}$$

Portanto $C_1^W(n) = C_1^x$ independe de n, o que é óbvio. Mas para valores de s maiores

$$nC_s^x = C_s^Y(n) = n^{s/2}C_s^Z(n) = n^s C_s^W(n), \tag{3.23}$$

Portanto

$$\begin{aligned} C_s^Y(n) &= nC_s^x \\ C_s^Z(n) &= \frac{1}{n^{\frac{s}{2}-1}}C_s^x \\ C_s^W(n) &= \frac{1}{n^{s-1}}C_s^x, \end{aligned} \tag{3.24}$$

que mostram o decaimento dos cumulantes como função de n. O expoente de n tem duas contribuições; o 1, que vem do acúmulo, e o $s/2$ ou s que vem do fator de escala de Z ou Y respectivamente. É mais interessante olhar para quantidades adimensionais para poder entender o sgnificado relativo desses decaimentos. Podemos olhar para $(C_2^x)^{1/2}$ como a escala típica das flutuações de x em torno da média. A razão $u_s^x = C_s^x/(C_2^x)^{s/2}$ é adimensional e

$$u_s^Y(n) = \frac{C_s^Y(n)}{(C_2^Y(n))^{\frac{1}{2}}} = n^{1-\frac{s}{2}}u_s^x, \tag{3.25}$$

Este decaimento mostra que para s fixo, $s \geq 3$ a contribuição relativa dos cumulantes superiores fica cada vez menor com o aumento de n. Já que independe da escala, isso vale para Z e W também (verifique).

Exercício

Calcule os cumulantes para a distribuição normal $\mathcal{N}(\mu, \sigma)$, ou seja $P(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Calcule a função característica. É óbvio que $C_1 = \mu$ e $C_2 = \sigma^2$.

Mostre que $C_s = 0$ para $s \geq 3$. Segue que as quantidades adimensionais u_s são nulas para $s \geq 3$.

O que significa, frente a este resultado para a gaussiana, o decaimento de $u_s^Y(n) = n^{1-\frac{s}{2}} u_s^x$? De forma pedestre isto mostra que a distribuições de Y, Z e W estão ficando mais perto de uma gaussiana para n grande. E de forma não pedestre? Este é o tema da próxima secção.

3.4 O Teorema do Limite Central I

3.4.1 Aquecimento

Este teorema tem um papel central em qualquer discussão de probabilidades. Isto no entanto não é o que *central* quer dizer aqui. O termo se refere a que, nas condições da secção anterior, as distribuições de Y, Z ou W são aproximadamente gaussianas na parte *central* ou seja perto do valor máximo. Para aquecer vemos que pela equação 3.17, a função característica de Y é

$$\hat{P}(k|n) = \exp(\sum_{s=0}^{\infty} C_s^Y(n) \frac{(-ik)^s}{s!}) \tag{3.26}$$

e a distribuição e'

$$P(Y|N=n) = \int_{-\infty}^{\infty} \exp(iky - \sum_{s=0}^{\infty} C_s^Y(n) \frac{(-ik)^s}{s!}) \frac{dk}{\sqrt{2\pi}}. \tag{3.27}$$

Até aqui é exato. Desprezando os cumulantes de ordem superior a segunda, pela eq. 3.25

$$P(Y|N=n) = \int_{-\infty}^{\infty} \exp(iky + ikC_1^Y(n) - k^2 C_2^Y(n)/2) \frac{dk}{\sqrt{2\pi}} \tag{3.28}$$

Chamemos $< x >= \mu$ e $\sigma_x^2 =< x^2 > - < x >^2$. Realizando a transformada de Fourier

$$P(Y|N=n) = \frac{1}{\sqrt{2\pi n \sigma_x^2}} e^{-\frac{(y-n\mu)^2}{2n\sigma_x^2}} \tag{3.29}$$

Da mesma forma, e com o mesmo grau de rigor ou falta dele:

$$\begin{aligned} P(Z|N=n) &= \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(z-\sqrt{n}\mu)^2}{2\sigma_x^2}} \\ P(W|N=n) &= \frac{1}{\sqrt{2\pi\frac{\sigma_x^2}{n}}} e^{-\frac{(w-\mu)^2}{2\frac{\sigma_x^2}{n}}} \end{aligned} \tag{3.30}$$

Vemos que as distribuições são gaussianas e escrevemos as tres para mostrar que as diferentes formas de ajustar a escala da soma leva a que diferentes quantidades tenham um valor limite fixo ou que mude com alguma potência de n.

O leitor, neste ponto, deve se perguntar qual é a operação matemática de desprezar que nunca antes viu definida. Analisaremos isto nas próximas secções. Mas antes um exemplo onde o cálculo é exato.

Exercício

Mostre que os resultados acima (eqs. 3.29 e 3.30) são exatos no caso particular que a distribuição $P(x)$ é gaussiana:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}}$$

Solução: O exercício anterior mostra que os cumulantes com $s \geq 3$ são nulos. Logo, não é necessário *desprezá-los*. Temos o resultado importante que somas de variáveis gaussianas tem distribuição gaussiana. Este é um exemplo de uma distribuição dita estável sob adições. Somas de variaveis gaussianas são gaussianas.

Exercício

Mostre que a distribuição de Cauchy $P(x) = \frac{1}{\pi}\frac{b}{x^2+b^2}$ é estável. Note que portanto a soma de variáveis de Cauchy não é gaussiana. Discuta primeiro a variância de x para ver onde os argumentos acima falham.

A média e a concentração em torno da média

A distribuição de W dada pela eq. 3.29 mostra que a média de W é igual à média de x. Isto não deve causar nenhuma surpresa, devido à linearidade da integral. Se os diferentes x_i forem considerados como diferentes medidas de X, então W pode ser entendido como a média empírica de X. Isto é o conteúdo da lei fraca dos grandes números. Quanto se afasta a média empírica da média? Ou de outra forma, diferentes experiências levam a diferentes médias empíricas, qual é a probabilidade de que hajam flutuações grandes? Usemos a desiguladade de Chebyshev que pode ser obtida desta forma:

Considere $\epsilon > 0$ e pela equação 3.24, temos que $C_2^W(n) = \sigma_x^2/n$, satisfaz

$$\begin{aligned}
C_2^W(n) &= \int_{-\infty}^{\infty} dw(w^2 - < w >^2) P(W = w|N = n) \\
&= \int_{-\infty}^{\infty} dw(w - < w >)^2 P(W = w|N = n) \\
&\geq \int_{|w-<w>|\geq\epsilon} dw(w - < w >)^2 P(W = w|N = n) \\
&\geq \int_{|w-<w>|\geq\epsilon} dw\epsilon^2 P(W = w|N = n) \\
&\geq \epsilon^2 Prob(|w - < w > | \geq \epsilon). \qquad (3.31)
\end{aligned}$$

onde usamos $Prob(|w- < w > | \geq \epsilon = \int_{|w-<w>|\geq\epsilon} dwP(W = w|N = n)$ e chegamos à desigualdade de Chebyshev, que dá uma cota do decaimento com ϵ da probabilidade de ter flutuações maiores que ϵ:

$$Prob(|w- < w > | \geq \epsilon) \leq \frac{C_2^W(n)}{\epsilon^2} \tag{3.32}$$

Mas $C_2^W(n)$ depende de maneira simples de n. Extraindo esta dependência temos que a "probabilidade de que uma amostra de n valores $\{x_i\}$ que tenha uma média empírica $< w >$ e que este valor se afaste do valor médio por mais que ϵ", isto é, , $Prob(|w- < w > | \geq \epsilon)$ está limitada por:

$$Prob(|w- < w > | \geq \epsilon) \leq \frac{C_2^x}{n\epsilon^2} \tag{3.33}$$

As flutuações de w de tamanho maior que ϵ fixo, ficam mais improváveis quando n cresce.

O próximo exercício mostra de que forma a frequência de um evento esta relacionada com a probabilidade.

Exercício: frequência e probabilidade

Considere a seguinte informação I= "Uma moeda é jogada para cima, bate no teto, no ventilador do teto, e cai no chão plano". Há vários motivos para atribuir $p = 1/2$ à probabilidade que caia a *cara* para cima, isto é $p = P(s = 1|I) = 1/2$ e $q = P(s = -1|I) = 1/2$. Poderiamos considerar outra experiência I'[8] onde p, q tem outro valores (entre zero e um). Consideremos as jogadas independentes, para duas jogadas i e j quaisquer $P(s_i|s_jI') = P(s_i|I')$. Chame m o número de caras para cima, quando a moeda é jogada n vezes. A frequência de caras é definida por $f = m/M$

- **(A)** Mostre que a distribuição de m, é a distribuição binomial:

$$P(m|N = nI') = \frac{n!}{m!(n-m)!}p^m q^{n-m} \tag{3.34}$$

- **(B)** Calcule $< m >$, $< m^2 >$. [Dica: Use a expansão binomial de (i) $(p+q)^n$, (ii) $p\frac{\partial}{\partial p}p^m = mp^m$ e (iii) a normalização $p + q = 1$; resposta: $< m >= np$, $< m^2 >= n^2p^2 + np(1-p)$]

- **(C)** Refaça a dedução da desigualdade de Chebyshev para distribuições de variáveis que tomam valores discretos e mostre que para ϵ fixo, a probabilidade que a frequência f se afaste do *valor esperado* $< f >= p$ por mais que ϵ, cai com $1/n$.

[8] por exemplo I'= "Deixe a moeda, inicialmente de cara para cima e num plano horizontal, cair até a mesa, a partir de uma altura h, sem girar". Considere $h = 1$ mm, $h = 1$ cm e $h = 1$ m.

- **(D)** Discuta e pense: Então de que forma a frequência está ligada à probabilidade? A frequência *converge*, quando n cresce, para a probabilidade p. Toda convergência precisa ser definida em termos de uma distância, que vai para zero quando se toma algum limite. É fundamental entender que a distância aqui não é ϵ, mas é a **probabilidade** que f se afaste de p por mais de ϵ. Assim, a frequência f converge **em probabilidade** à probabilidade p.

A conclusão do exercício acima é fundamental. Como poderiamos definir probabilidades em termos de frequência, se para mostrar que a frequência está associada à probabilidade usamos o conceito de convergência em probabilidade? Discuta se é errado ou não definir um conceito usando esse conceito na definição.

Mas o exercício acima mostra porque pode parecer sedutor usar a frequência em lugar da probabilidade. Se tivermos informação I' sobre uma experiência e dados sobre uma sequência de experimentos nas condições I' podemos atribuir valor à probabilidade de forma mais segura.

3.5 O Teorema do Limite Central II

Não há uma prova só, mais muitas, que refletem os objetivos em estudar este problema. Podemos olhar para diferentes condições sobre $P(x)$ e com isso mudar os resultados sobre a região central que é gaussiana e sobre quão grandes são os erros nas caudas das distribuições. Dependendo das condições, a região central vai depender de forma diferente do valor de n.

Esperamos pela eq. que a variável $Z - < Z >= \frac{Y - N\mu}{\sqrt{N}\sigma}$ tenha distribuição normal de média nula e variância 1, pelo menos na região *central*.

Podemos transladar a origem de x e tornar $\mu = 0$.

Teorema LC

(Kinchin) Suponhamos que existam A, a, b , c e d constantes positivas tal que

- $dP(x)/dx$ é contínua
- $\int |dP(x)/dx|\, dx < A$
- $a < < x^2 >= \sigma^2 < b$
- $< |x^3| >< b$
- $< x^4 >< b$
- $< |x^5| >< b$
- $|\hat{P}(k)| > d$ para $|k| < c$
- Para cada intervalo (k_1, k_2), com $k_1 k_2 > 0$, existe um número $\rho(k_1, k_2) < 1$, tal que para $k_1 < k < k_2$ temos

$$|\hat{P}(k)| < \rho.$$

Então

- Na região central, definida por $|x| < 2\log^2 n$

$$P(Y|N=n) = \frac{1}{\sqrt{2\pi n\sigma^2}} e^{-\frac{y^2}{2n\sigma^2}} + \frac{S_n + yT_n}{(n\sigma^2)^{5/2}} + O(\frac{1+|x|^3}{n^2})$$

  onde S_n e T_n são independentes de y e não crescem mais rápido que n.

- Para y arbitrário

$$P(Y|N=n) = \frac{1}{\sqrt{2\pi n\sigma^2}} e^{-\frac{y^2}{2n\sigma^2}} + O(\frac{1}{n})$$

A prova é razoávelmente simples e pode ser encontrada no Apêndice de [Kinchin]. O leitor poderá ver que a essência da prova está no controle dos termos superiores da expansão (ver eq. 3.27) *desprezados* anteriormente para chegar até a eq. 3.28.

3.6 O Teorema do Limite Central III

Apresentamos alguns exemplos para distribuições $P(x)$ simples.

3.6.1 A distribuição uniforme

$P(x) = 1/L$ para $-L/2 < x < L/2$ e 0 para outros valores de x. A função característica

$$\hat{P}(k|1) = \frac{1}{L}\int_{-L/2}^{L/2} e^{-ikx}dx = \frac{2}{kL}\sin(\frac{kL}{2}) \tag{3.35}$$

$$P(Y=y|N=n) = \int_{-\infty}^{\infty} [\hat{P}(k|1)]^n e^{iky}\frac{dk}{2\pi} = \int_{-\infty}^{\infty} [\frac{2}{kL}\sin(\frac{kL}{2})]^n e^{iky}\frac{dk}{2\pi} \tag{3.36}$$

A figura 3.2 mostra que a função caracterísitica fica mais parecida com uma gaussiana e na figura 3.3 vemos que efetivamente o $log(|\hat{P}(\sqrt{|u|}|N=n)|)$ com $u = k^2$ fica cada vez mais perto de $-c|u|$ (gaussiana).

3.6.2 A distribuição exponencial

$P(x) = \Theta(x)ae^{-xa}$, portanto $\mu = \sigma = a^{-1}$ A função característica

$$\hat{P}(k|1) = \int_0^{\infty} ae^{-xa}e^{-ikx}dx = \frac{a}{a+ik} \tag{3.37}$$

$$P(Y=y|N=n) = \int_{-\infty}^{\infty} (\frac{a}{a+ik})^n e^{iky}\frac{dk}{2\pi} \tag{3.38}$$

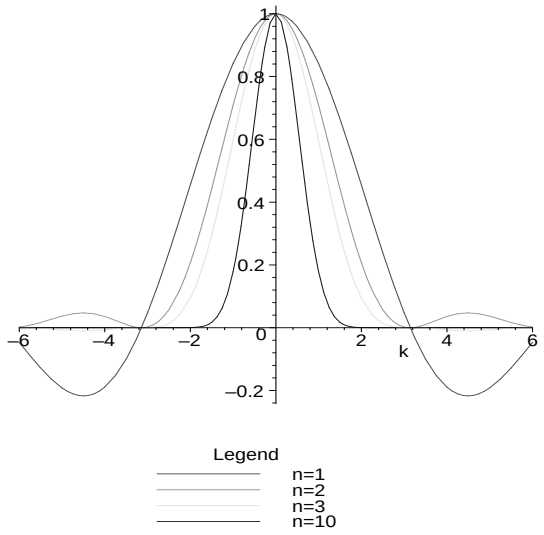

Figura 3.2: A função característica $\hat{P}(k|N = n)$ para a soma de n varíaveis uniformemente distribuidas.

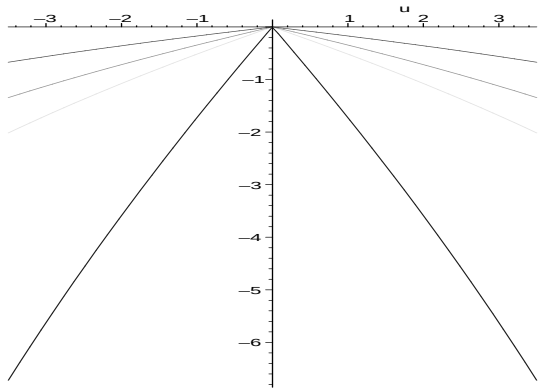

Figura 3.3: A função característica $log(|\hat{P}(\sqrt{|u|}|N = n)|)$ contra $u = k^2$ para a soma de n varíaveis uniformemente distribuidas. Nesta representação gaussianas aparecem como retas. A legenda é a mesma da figura 3.2

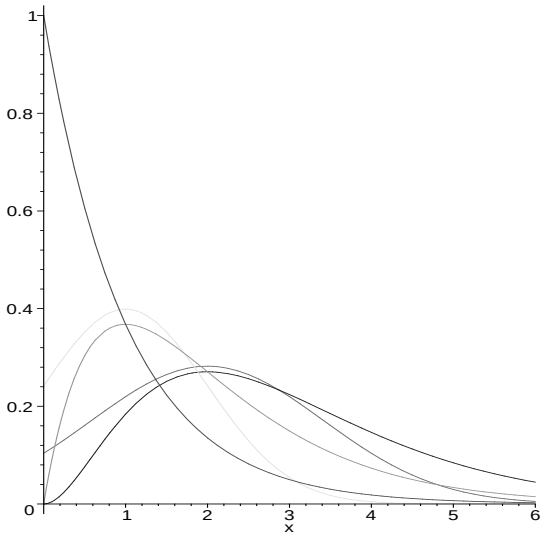

Figura 3.4: A densidade de probabilidade $P(y|N=n, a=1)$ para a soma de n varíaveis exeponencialmente distribuidas, $n = 1, 2, 3$. Para as duas últimas mostramos as gaussianas com $\mu = \sigma^2 = n - 1$

Integrando por partes ($u = e^{iky}, dv = (\frac{a}{a+ik})^l dk$, para $l = n, n-1, ...1$) obtemos:

$$P(Y = y|N = n) = \Theta(y) a^n \frac{y^{n-1} e^{-ax}}{(n-1)!} \tag{3.39}$$

que não é uma gaussiana. No entanto, a região central sim, se parece com uma gaussiana.

A figura 3.4 mostra que a distribuição para n baixo não se parece em nada com uma gaussiana , mas à medida que n aumenta fica mais parecida com uma gaussiana, figura 3.5. Note que as distribuições, nessa figura são claramente assimétricas. Pense no que significa que a distribuição resultante seja gaussiana se as variáveis somadas são sempre positivas e portanto $Y > 0$ sempre. Esse é o significado de *central*, nas caudas não dizemos nada.

3.6.3 A distribuição binomial revisitada

A distribuição de Bernoulli é dada por $P(x) = p\delta(x-1) + q\delta(x+1)$. O número de aplicações que usaram esta distribuição é enorme. Só para ter uma ilustraçào em mente, podemos pensar em jogadas de uma moeda, ou um passo dado por um bêbado numa caminhada unidimensional. Se há N repetições ($i = 1...N$) e $P(x_i)$ é a mesma para todo i e $P(x_i|x_j) = P(x_i)$ para qualquer $i \neq j$, e queremos $P(Y|N)$ para $Y = \sum_{i=1..n} x_i$. Este é exatamente nosso exemplo acima sobre a distribuição binomial onde estudamos a relação entre frequência e probabilidade.

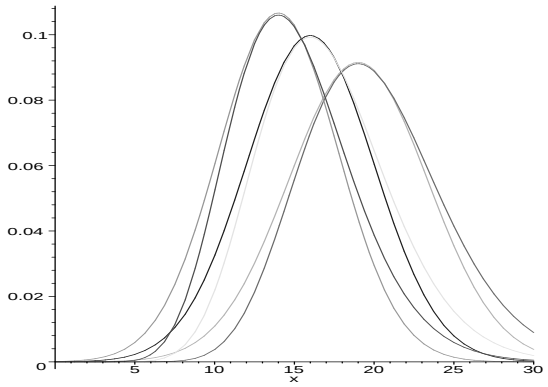

Figura 3.5: A densidade de probabilidade $P(y|N = n, a = 1)$ para a soma de n varíaveis exeponencialmente distribuidas, $n = 15, 17, 20$. Junto estão mostradas as gaussianas com $\mu = \sigma^2 = n - 1$.

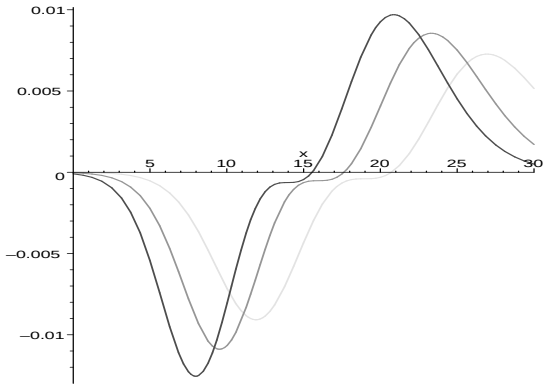

Figura 3.6: A diferença entre a densidade de probabilidade $P(y|N = n, a = 1)$ para a soma de n varíaveis exeponencialmente distribuidas, $n = 15, 17, 20$ e as gaussianas com $\mu = \sigma^2 = n - 1$. Os mesmos parâmetros da figura 3.5. Note que a região central é bem aproximada. Há uma região de transição, ao afastar-se para as caudas, e finalmente as caudas vão rapidamente para zero, assim como a sua diferença

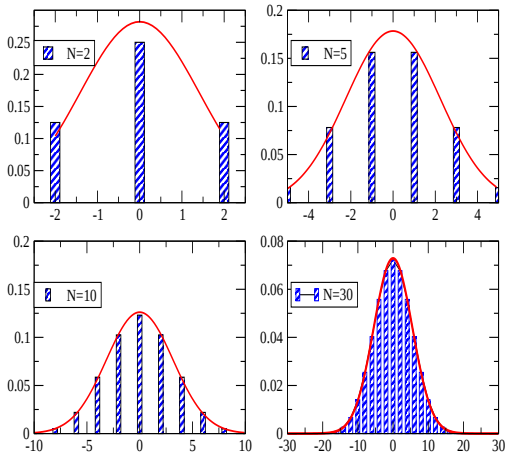

Figura 3.7: A binomial (dividida por $\Delta Y = 2$, barras) e a densidade gaussiana correspondente (linha contínua), para $N = 2, 5, 10$ e 30

Aqui há um pequeno problema. A distribuição de probabilidades binomial deve ser comparada com a densidade de probabilidade gaussiana. Note que se N é par a probabilidade de que ocorra um valor de Y impar é zero, ou seja $\Delta Y = 2$.. Ao apresentar os gráficos da figura 3.7 a binomial foi dividida por ΔY. De outra forma: a probabilidade da binomial que Y tenha um dado valor num intervalo $(y, y + 2)$ é aproximado pela integral da gaussiana entre y e $y + 2$.

3.6.4 Caminho Aleatório

Novamente olhamos para a distribuição binomial. Olhe para a figura 3.8. Definimos o caminho aleatório através de

Difusão: $K(= 10000$ na figura 3.8) seqüências de N passos de um processo binomial, definidos por $y_k(n_t) = y_k(n_{t-1} + x_k(n_t))$.

$$y_n = y_{n-1} + x_n \tag{3.40}$$

onde $P(x) = 1$ para $-0.5 \leq x \leq 0.5$. O índice n pode ser interpretado como tempo numa dinâmica discreta, a cada intervalo de tempo Δt uma partícula se desloca uma quantidade x. O deslocamento total

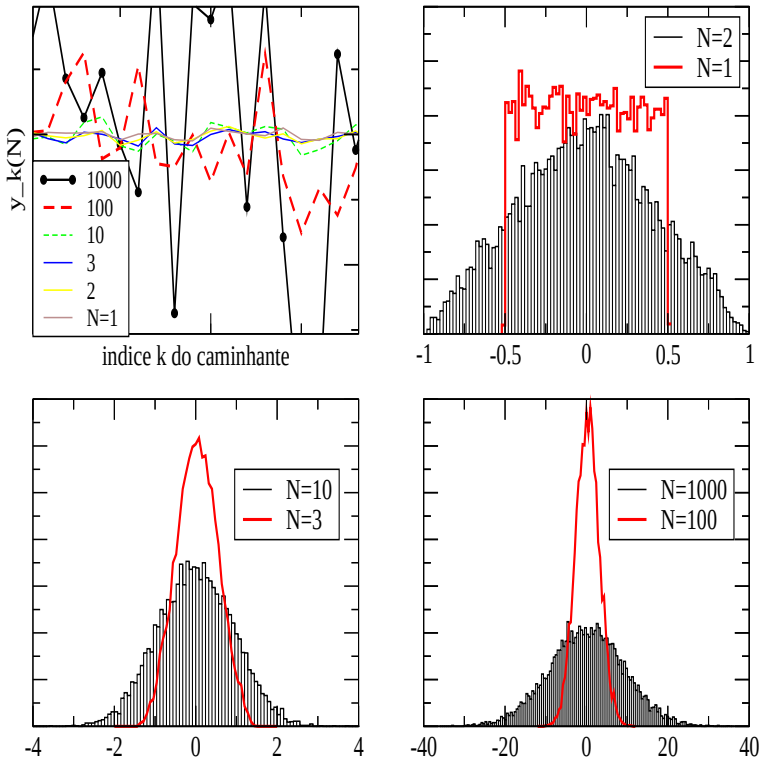
y_k(N)
1000
100
10
3
2
N=1
indice k do caminhante
N=2
N=1
-1
-0.5
0
0.5
1
N=10
N=3
-4
-2
0
2
4
N=1000
N=100
-40
-20
0
20
40

Figura 3.8: