Teoria: Sistemas de Processamento de Informação

Nestor Caticha

28 de agosto de 2015

Capítulo 1

Teoremas de Regraduação de Cox

1.1 Probabilidades e Informação

Estas notas para o curso "Mecânica Estatística de Sistemas de Processamento de Informação" estão e parece sempre estarão em estado embrionário e preliminar. Eventuais críticas, correções ou sugestões serão bemvindas. Algumas partes são baseadas nos livros de E. Jaynes [1] e de A. Caticha [2] que estas notas não substituem. Outras não, grande parte são baseadas em conjunto de notas de uma versão anterior deste curso.

A primeira parte discute o uso de probabilidades. O leitor, aluno deste curso não deverá ler isto como uma validação das suas crenças. Afinal, se está estudando Mecânica Estatística já parece natural o uso de probabilidades. Deve procurar falhas no raciocínio. Procurar exceções.

A idéia de apresentar uma forma de pensar que tem aplicações em uma vasta gama de assuntos, pode levar o leitor a pensar que está na presença de alguém que com um martelo, pensa que todos os problemas são pregos. Ou que estamos apresentado dogmas, dos quais não abriremos mão. No fim talvez não saiba como me defender de tais acusações, exceto alegando que o único ponto sobre o qual serei inflexível será que só podemos acreditar naquilo que a informação e evidência permitem, e só enquanto não surgir informação contraditória. Não faz sentido acreditar em algo que não é respaldado por informação. [3]

Estas notas servem dois propósitos. Um como material de suporte para um curso de Mecânica Estatística, outro para Física de processamento de informação.

[1] E. T. Jaynes, Probability Theory: the Logic of Science

[2] Entropic Inference and the Foundations of Physics

[3] Há outras formas de pensar, por exemplo acreditar em algo porque isso me deixa mais feliz. Mas eu não saberia dar um curso sobre isso. "I have a lot of beliefs, and I live by none of them - that's just the way I am... they make me feel good about who I am." –Louis CK

1.2 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.

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 mais claramente do que estamos falando e do que não estamos, achamos que esta é atualmente a melhor maneira de introduzir a teoria de informação. [4]

Pode parecer estranho para o estudante Física que o elemento principal a seguir seja a idéia de asserção, isto é, uma frase que em princípio é uma proposição que se apresenta como verdadeira. Uma frase pode ser julgada correta ou não de várias maneiras. Podemos pensar se é correta do ponto de vista da sua estrutura gramatical ou sintática. Não é isto que queremos fazer e consideraremos as asserções a seguir suficientemente bem formadas. Queremos analisar seu conteúdo informacional, se realmente a podemos creer verdadeira. Mas quando se diz " a massa de Saturno está entre m_1 e m_2" ou "... entre m_3 e m_4" estamos usando asserções diferentes e a tarefa é determinar quanto acreditamos que uma ou a outra sejam verdade e aqui o estudante reconhece a linguagem científica.

Consideremos a asserção "Existem zumbies". Isto é verdade? Se o contexto for o de filmes gravados em Pittsburgh na década de oitenta, a resposta será uma. Se for no mundo real, outra. 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

[4] 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

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. O "exatamente" não é coincidência. 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 $C =$ "$A \to B''$, 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 $C =$ "$A \to B''$ 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, [5] o que quer que isto seja, e que será discutido mais à frente, será necessária para avançar. 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

[5] 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)

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*[6] 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 esqueça 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.

1.2.1 Axiomas de Cox

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. 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 exposição de Jaynes teve uma grande influencia, mas ainda recebeu críticas e complementos [**?**]. Eu seguirei a apresentação de A. Caticha, que é mais completa [7].

A maneira de construir a teoria está baseada na seguinte forma de pensar bastante simples. Queremos construir uma teoria geral para a extensão da lógica nos casos de informação incompleta. Se ela for suficientemente geral, deverá ser válida em casos particulares. Se o caso for suficientemente simples, então podemos saber qual é o resultado esperado que não viole expectativas razoáveis. 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 teoria geral. [8] Isto é o que mostraremos a seguir.

[6] Ao leitor que demande uma definição de racional, podemos dizer que pelo menos não queremos se manifestamente irracionais e que terá que esperar por uma resposta.

[7] Notem que há lugar ainda para avanços nestes primeiros passos. Tentem encontrar defeitos, generalizações, melhorias nos argumentos

[8] 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

Em primeiro lugar queremos falar sobre uma asserção A no caso de informação incompleta. Nos referimos então à crença ou plausibilidade de A ser verdade dado B e a denotamos pelo símbolo $A|B$ que lemos "a plausibilidade A dado B" ou ainda de "... de A condicionada a B". Por que não à probabilidade de A dado B? Porque já existe uma teoria matemática de probabilidade e não sabemos se será a estrutura matemática que emergirá desta análise. Poderiamos usar outras palavras, mas crença ou plausibilidade são conhecidas o suficiente para serem úteis neste contexto e não tem por agora o problema de ser definidas formalmente .

Queremos analisar o primeiro caso simples que lida com o conceito de *mais plausível*. Se A dado B é mais plausível do que A dado C escrevemos $A|B \succ A|C$. Suponha ainda que é ainda que $A|C \succ A|D$. Queremos, para seguir o uso cotidiano da linguagem impor que A dado B seja mais plausível que A dado D

Temos assim nosso primeiro desejo, a plausiblidade deverá satisfazer alguma forma de transitividade:

- D_1: Se $A|B \succ A|C$ e $A|C \succ A|D$ então deve ser o caso que $A|B \succ A|D$

Isto é fácil se impusermos:

- 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: $C = AB$, $C = A \wedge B$ ou ainda $C = A$ **e** B.
- A **soma** ou disjunção de duas asserções é uma terceira asserção, que constuma ser denotada por $D = A + B$ ou $D = A \vee B$, ou ainda $D = A$ **ou** B.

explicita. O destaque a este procedimento apareceu por primeira vez no livro de A. Caticha. Usaremos novamente este estilo de fazer teoria ao introduzir o conceito de entropia.

A tabela 1.1 mostra a tabela verdade para as operações de soma e produto lógico, onde V = Verdade e F = Falso. Note que as últimas duas colunas,colocadas aqui para futura referência, mostram que $\overline{A+B}$ e $\overline{A}\,\overline{B}$ são iguais.

A	B	$A+B$	AB	$\overline{A+B}$	$\overline{A}\,\overline{B}$
V	V	V	V	F	F
V	F	V	F	F	F
F	V	V	F	F	F
F	F	F	F	V	V

Tabela 1.1

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_2: 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.

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 inconsistente. Usamos consistência no sentido de não manifestamente inconsistente, que é o que D_2 acima declara.

Agora olhamos para o caso simples em que a e b são mutuamente exclusivos na condição c e em qualquer outra condição. Então $a|b$ e $b|a$ representam a plausibilidade de algo que sabemos ser falso. Assim como $a|a$ e $b|b$ são a plausibilidade de algo que sabemos ser verdade. Poderia ser que hajam falsidades mais falsas que outras, ou verdades mais verdadeiras que outras, mas achamos razoável impor

- D_3 : Para todo a, $a|a = v_v$ e para a e b mutuamente exclusivos $a|b = v_f$.

Não sabemos que valores dar para v_v ou v_f, mas supomos o mesmo valor em todos os casos que tenhamos certeza de verdade ou falsidade

Todo operador na álgebra Booleana pode ser representado pelas operações conjunção (e) e negação ($\bar{\cdot}$) [9], 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.

Agora olhamos para a soma lógica. Novamente c se refere à informação subjacente e estamos interessados na plausibilidade $y = a \vee b|c$. Há 4 plausibilidades que serão interessantes para esta análise:

$$x_1 = a|c,\ x_2 = b|c,\ x_3 = a|bc,\ x_4 = b|ac$$

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

. Notamos que deve haver uma dependência entre $a \vee b|c$ e algum subconjunto de $\{x_i\}$, então

- D_4: Deve existir uma função F que relaciona $a \vee b|c$ e algum subconjunto de $\{x_i\}$.

É claro que trocando soma por produto parece razoável desejar:

- D_5: Deve existir uma função G que relaciona $ab|c$ e algum subconjunto de $\{x_i\}$.

Não impomos nada além da existência dessas funções, além de que dependam em algumas, se não todas, as variáveis $\{x_i\}$.

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_1, x_2, x_3, x_4)

1.2.2 A regra da soma

Começamos com a função F e consideramos a e b mutuamente exclusivos

$$a \vee b|c = F(a|c, b|c, a|bc, b|ac) = F(a|c, b|c, v_f, v_f)$$

mas esta é uma função de apenas duas variáveis, e da constante desconhecida v_f:

$$a \vee b|c = f(a|c, b|c)$$

Agora consideremos tres asserções a, b e c mutuamente excludentes nas condições d. Duas maneiras equivalentes de escrever a disjunção das tres são $(a \vee b) \vee c|d = a \vee (b \vee c)|d$ o que permite usar a função f

$$\begin{aligned} a \vee b \vee c|d &= f(f(a|d, b|d), c|d) \\ &= f(a|d, f(b|d, c|d)) \end{aligned}$$

ou em notação óbvia

$$f(f(x, y), z) = f(x, f(y, z)) \tag{1.1}$$

chamada equação da associatividade, primeiramente estudada por Abel no contexto de teoria de grupos. Pode se provar [10] que existe um bijeção ϕ, dos reais nos reais, que tomaremos como monotonicamente crescente, tal que

$$f(x, y) = \phi^{-1}(\phi(x) + \phi(y)). \tag{1.2}$$

O leitor deverá neste ponto mostrar que a expressão 1.2 é uma solução da equação 1.1. Agora um ponto importante: podemos *regraduar* as atribuições de plausibilidade e não mais falar dos números do tipo $a|d$ mas de números $\phi(a|d)$. Note que o fato de ser uma bijeição resulta que a ordem de preferências

[10] Aequationes mathematicae 1989, Volume 37, Issue 2-3, pp 306-312 The associativity equation revisited R. Craigen, Z. Páles, ou o livro do Axcel

não se altera, se antes as crenças sobre as asserções tinham uma certa ordem, depois da regraduação, a representação numérica das crenças não se altera. [11] Continuamos sem saber que números são esses, mas avançamos a ponto de poder dizer que para quaisquer eventos mutuamente exclusivos

$$\phi(a \vee b|d) = \phi(a|d) + \phi(b|d). \tag{1.3}$$

No caso particular que $d = \overline{a}$, isto significa

$$\begin{aligned} \phi(a \vee b|\overline{a}) &= \phi(a|\overline{a}) + \phi(b|\overline{a}) && (1.4)\\ \phi(b|\overline{a}) &= \phi(a|\overline{a}) + \phi(b|\overline{a}) && (1.5)\\ & && (1.6) \end{aligned}$$

pois a crença $\phi(a \vee b|\overline{a})$ é equivalente à crença $\phi(b|\overline{a})$. Segue que

$$\phi(a|\overline{a}) = \phi(v_f) = \phi_f = 0 \tag{1.7}$$

Embora modesto, eis o primeiro resultado númerico. **O valor regraduado da certeza da falsidade é zero.**

Mas e se não forem mutuamente exclusivos? O interessante é que o resultado anterior serve para o caso geral, mas precisamos usar o truque de escrever

$$a = (a \wedge b) \vee (a \wedge \overline{b}) \;\; e \;\; b = (b \wedge a) \vee (b \wedge \overline{a})$$

Podemos escrever $a \vee b$ como uma disjunção de asserções mutuamente exclusivas:

$$\begin{aligned} a \vee b &= (a \wedge b) \vee (a \wedge \overline{b}) \vee (b \wedge a) \vee (b \wedge \overline{a})\\ &= (a \wedge b) \vee (a \wedge \overline{b}) \vee (b \wedge \overline{a}) \end{aligned}$$

assim a equação 1.3 pode ser usada, levando a

$$\begin{aligned} \phi(a \vee b|d) &= \phi(a \wedge b|d) + \phi(a \wedge \overline{b}|d) + \phi(b \wedge \overline{a}|d)\\ &= \phi(a \wedge b|d) + \phi(a \wedge \overline{b}|d) + \phi(b \wedge \overline{a}|d) + \phi(a \wedge b|d) - \phi(a \wedge b|d) \end{aligned}$$

onde, na última linha adicionamos e subtraimos o mesmo número. Usando novamente a equação 1.3 para asserções mutuamente exclusivas

$$\begin{aligned} \phi(a \vee b|d) &= \phi(a \wedge b \vee a \wedge \overline{b}|d) + \phi(b \wedge \overline{a} \vee a \wedge b|d) - \phi(a \wedge b|d)\\ &= \phi(a|d) + \phi(b|d) - \phi(a \wedge b|d), && (1.8) \end{aligned}$$

o que será óbvio uma vez que o leitor provar que $a \wedge b \vee a \wedge \overline{b} = a$ e $b \wedge \overline{a} \vee a \wedge b|d = b$.

Exercício Desenhe o diagrama de Venn adequado a esta situação.

[11] Note que tomaremos a função ϕ como estritamente monotonica: se $x > y$ segue que $\phi(x) > \phi(y)$, sem poder haver igualdade. Isto é uma decisão pragmática, pois não queremos que asserções com crenças diferentes sejam mapeadas em valores ϕ iguais, pois antes da regraduação tinhamos uma separação de preferências e depois da regraduação poderiamos ter confusão entre asserções mapeadas no mesmo valor de ϕ. Mas de qualquer forma veja A. Patriota referencia

1.2.3 Regra do produto: quais as variáveis relevantes?

Queremos expressar $y = \phi(ab|c)$ em termos da função ainda por determinar G e de algum dos subconjuntos de $\{x_i\}$ Tribus sugeriu a análise das 11 possibilidades para verificar que só há duas que sobrevivem a casos extremos. 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.

Vejamos como chegar a esta conclusão (novamente seguimos AC)

1. $y = G(\phi(a|c), \phi(b|c))$ (1 possibilidade)
2. $y = G(\phi(a|c), \phi(a|bc))$ (2 possibilidades $a \leftrightarrow b$)
3. $y = G(\phi(a|c), \phi(b|ac))$ (2 possibilidades $a \leftrightarrow b$)
4. $y = G(\phi(a|bc), \phi(b|ac))$ (1 possibilidade)
5. $y = G(\phi(a|c), \phi(b|c), \phi(a|bc))$ (2 possibilidades $a \leftrightarrow b$)
6. $y = G(\phi(a|c), \phi(a|bc), \phi(b|ac))$ (2 possibilidades $a \leftrightarrow b$)
7. $y = G(\phi(a|c), \phi(b|c), \phi(a|bc), \phi(b|ac))$ (1 possibilidade)

Caso 1 Mostraremos que $y = a \wedge b|c = G(\phi(a|c), \phi(b|c)) = G(x_1, x_2)$ não funciona pois nao satisfaz o esperado em um caso simples. Porque não serve o subconjunto mais óbvio (x_1, x_2)? Seja a=' Helena usa um tenis esquerdo vermelho' enquanto que b=' Helena usa um tenis direito preto' . A plausibilidade dessas duas asserções será julgada dada a seguinte informação c='Helena gosta de tenis pretos e de tenis vermelhos', e talvez seja possível concluir que as duas asserções são bastante plausíveis. Mas se tivessemos $y = G(x_1, x_2)$ poderiamos ser levados a pensar que 'Helena usa um tenis esquerdo vermelho e um tenis direito preto' é bastante plausível. Posso acreditar bastante nas duas asseções, mas não que use um tenis de cada cor. Devemos rejeitar esta forma para G.

Para convencer os incrédulos no exposto acima, um argumento mais formal: Suponha que $a|d = a'|d$ e $b|d = b'|d$, mas que embora a e b sejam mutuamente exclusivos, a' e b' não o sejam. Neste caso teriamos que

$$\phi(a'b'|d) = G(\phi(a'|d), \phi(b'|d)) = G(\phi(a|d), \phi(b|d)) = \phi(ab|d) = \phi_F = 0.$$

E isto ocorreria para qualquer par de asserções não mutuamente exclusivas (a', b'), pois sempre poderiamos supor um caso auxiliar (a, b) adequado.

Caso 2 Se $y = G(\phi(a|c), \phi(a|bc))$ em geral, consideramos o caso particular em que $b = ad$ para qualquer d não seja mutuamente exclusivo a a. Logo

$$\begin{aligned} G(\phi(a|c), \phi(a|bc)) &= G(\phi(a|c), \phi(a|adc)) \\ &= G(\phi(a|c), \phi_v) = g(\phi(a|c)) \end{aligned}$$

segue que

$$y = \phi(ab|c) = \phi(b|c)$$

$$\phi(b|c) = g(\phi(a|c))$$

onde o lado esquerdo depende de d mas o lado esquerdo não. Esperamos que isso não ocorra em geral, e portanto eliminamos este candidato.

Caso 3 Para o caso $y = G(a|c, b|ac)$ e a alternativa $G(b|c, a|bc)$ ninguém tem encontrado casos que se oponham ao bom senso. Este será o único candidato a sobreviver e será a pedra de sustentação a toda a teoria que segue.

Caso 4 Se $y = G(\phi(a|bc), \phi(b|ac))$ somos levados a algo inaceitável considerando $a = b$, pois seguiria que

$$\phi(ab|c) = \phi(a|c) = G(\phi(a|ac), \phi(a|ac)) = G(\phi_v, \phi_v)$$

e $\phi(a|c)$ seria constante independente de a.

Caso 5 $y = G(\phi(a|c), \phi(b|c), \phi(a|bc))$. Este caso é mais complicado de analisar. Mostraremos, no entanto que se reduz a algum dos casos anteriores, sob a hipótese razoável de diferenciabilidade de G com respeito a qualquer um dos seus argumentos. Ainda consideraremos a conjunçao de mais de duas asserções , $abc|d$, que pode ser escrito de duas formas diferentes $(ab)c|d = a(bc)|d$, portanto, considerando a primeira forma obtemos

$$\begin{aligned}
\phi((ab)c|d) &= G(\phi(ab|d), \phi(c|d), \phi(ab|cd)) \\
&= G(G(\phi(a|d), \phi(b|d), \phi(a|bd)), \phi(c|d), G(\phi(a|cd), \phi(b|cd), \phi(a|bcd)) \\
&= G(G(x, y, z), u, G(v, w, s)) \qquad (1.9)
\end{aligned}$$

$$\begin{aligned}
\phi(a(bc)|d) &= G(\phi(a|d), \phi(bc|d), \phi(a|bcd)) \\
&= G(\phi(a|d), G(\phi(b|d), \phi(c|d), \phi(b|cd), \phi(a|bcd)) \\
&= G(x, G(y, u, w), s) \qquad (1.10)
\end{aligned}$$

Notamos duas maneiras de escrever a mesma coisa, por D_2 que declarava que não queremos ser manifestamente inconsistentes, devemos ter

$$G(G(x, y, z), u, G(v, w, s)) = G(x, G(y, u, w), s).$$

Ainda notamos que embora éstas variáveis possam ter quaisquer valores, não ocorre o mesmo conjunto dos dois lados: Lado esquerdo $\{x, y, z, u, v, w, s\}$, lado direito $\{x, y, u, w, s\}$. Portanto o lado esquerdo nao deve depender de $z = \phi(a|bd))$ nem de $v = \phi(a|cd)$ explicitamente. As derivadas parciais com respeito a z ou v devem dar zero:

$$\begin{aligned}
0 &= \frac{\partial}{\partial z} G(G(x, y, z), u, G(v, w, s)) \\
&= \frac{\partial}{\partial r} G(r, u, G(v, w, s))_{r=G(x,y,z)} \frac{\partial}{\partial z} G(x, y, z) \qquad (1.11)
\end{aligned}$$

Se um produto é zero, pelo menos um dos fatores é zero, de onde concluimos que ou G não depende do primeiro argumento ou não depende do terceiro. Se não depende do primeiro

$$y = G(\phi(a|c), \phi(b|c), \phi(a|bc)) = G(\phi(b|c), \phi(a|bc)),$$

voltamos ao **Caso 3**. Se não depende do terceiro

$$y = G(\phi(a|c), \phi(b|c), \phi(a|bc)) = G(\phi(a|c), \phi(b|c))$$

e voltamos ao **Caso 1.**

Fica como

Exercício mostrar que o **Caso 6** pode ser reduzido ao **Caso 3** ou ao **Caso 4** e que o **Caso 7** aos **Caso 5** ou **Caso 6**

Concluimos portanto que

$$\begin{aligned} \phi(ab|c) &= G(a|c, b|ac) \\ &= G(b|c, a|bc) \end{aligned} \tag{1.12}$$

Cox coloca isto como um axioma, mas não precisamos fazer isto, basta dizer que existe uma função G mas que não sabemos *a priori* quais seus argumentos. A eliminação dos casos que contradizem o bom senso em casos suficientemente simples, mostra de forma satisfatória (o leitor pode pular e reclamar, mas terá que encontrar argumentos) que as equações 1.12 refletem a única opção. Uma das queixas pode ser sobre a diferenciabilidade de G. Mas estamos interessados em situações onde a informação pode mudar e não alterar significativamente as crenças e esperamos ao menos continuidade de G.

Note que agora será possível concluir que 'Helena usa um tenis esquerdo vermelho e um tenis direito preto' pode ser pouco plausível por que precisamos saber a plausibilidade de 'Helena usa um tenis esquerdo vermelho dado que Helena usa um tenis direito preto' e isto pode ser pouco plausível.

Mas ainda não acabamos. Precisamos determinar a função especifica G.

1.2.4 Regra do produto: qual é a função G?

Novamente olhamos para um caso simples, onde podemos escrever o resultado de duas maneiras. Considere a, b, c e d com $b|d$ e $c|d$ mutuamente exclusivos, e a asserção $a(b \vee c)$ uma conjunção que pode ser escrita como uma disjunção:

$$a(b \vee c) = (ab) \vee (ac). \tag{1.13}$$

Podemos usar o resultado para a soma para estudar o produto $\phi(a(b \vee c)|d)$:

$$\begin{aligned} \phi(a(b \vee c)|d) &= G(\phi(a|d), \phi(b \vee c)|ad)) \\ &= G(\phi(a|d), \phi(b|ad) + \phi(c|ad)) \end{aligned} \tag{1.14}$$

$$\begin{aligned} \phi((ab) \vee (ac))|d) &= \phi(ab|d) + \phi(ac)|d)) \\ &= G(\phi(a|d), \phi(b|ad))) + G(\phi(a|d), \phi(c|ad)) \end{aligned} \tag{1.15}$$

onde a equação 1.14 usa primeiro que $a(b \vee c)$ é um produto e em segundo lugar a regra da soma para asserções mutuamente exclusivas $b|d$ e $c|d$. A equação 2.1

mostra o resultado de considerar a soma $(ab) \vee (ac)$. Mas devido à equação 1.13 e D_2, estas duas formas devem dar o mesmo resultado:

$$G(x, y + z) = G(x, y) + G(x, z). \tag{1.16}$$

Novamente requerindo a diferenciabilidade, desta vez duas vezes, e definindo $w = y + z$ obtemos a equação diferencial

$$\frac{\partial^2 G(x, w)}{\partial w^2} = 0 \tag{1.17}$$

que tem solução geral $G(x, w) = A(x)w + B(x)$ em termos de duas funções desconhecidas, mas fáceis de determinar. Substituindo esta forma em 1.16 obtemos

$$A(x)(y + z) + B(x) = A(x)(y + z) + 2B(x), \tag{1.18}$$

portanto $B(x) = 0$, ou seja $G(x, w) = A(x)w$. Agora olhamos para $a|d$ e usamos $a|d = ad|d$ para a e d quaisquer.

$$\begin{aligned} \phi(a|d) &= \phi(ad|d) = G(\phi(a|d), \phi(d|ad)) \\ &= G(\phi(a|d), \phi_v) = A(\phi(a|d))\phi_v \end{aligned} \tag{1.19}$$

onde $\phi(d|ad) = \phi_v$ pois, obviamente d é informação completa para d. Ou seja $x = A(x)\phi_v$, logo

$$G(x, w) = \frac{xw}{\phi_v} \tag{1.20}$$

isto significa que, para $e = b \vee c$, b e c mutuamente exclusivos

$$\phi(ae|d) = \frac{\phi(a|d)\phi(e|ad)}{\phi_v} \tag{1.21}$$

o que permite regraduar mais uma vez os números associados as crenças sem mudar a ordem.

Mas resta um problema: e se retirarmos a restrição de b e c mutuamente exclusivos? Precisamos usar a equação 1.8 para obter:

$$\phi(a \vee b|d) = \phi(a|d) + \phi(b|d) - \phi(ab|d) \tag{1.22}$$

$$\begin{aligned} \phi(a(b \vee c)|d) &= G(\phi(a|d), \phi(b \vee c)|ad)) \\ &= G(\phi(a|d), \phi(b|ad) + \phi(c|ad) - \phi(bc|ad)) \end{aligned} \tag{1.23}$$

$$\begin{aligned} \phi((ab) \vee (ac)|d) &= \phi(ab|d) + \phi(ac|d) - \phi(abc|d) \\ &= G(\phi(a|d), \phi(b|ad))) + G(\phi(a|d), \phi(c|ad)) - G(\phi(a|d), \phi(bc|ad)) \\ &= G(\phi(a|d), \phi(b|ad))) + G(\phi(a|d), \phi(c|ad)) - G(\phi(a|d), G(\phi(b|ad), \phi(c|abd)) \end{aligned} \tag{1.24}$$

igualando os lados direitos das equações 1.23 e 1.24, obtemos uma nova equação funcional. Substituindo a forma para G da equação 1.20, vemos que o produto também funciona neste caso geral.

Exercício Mostre que a forma produto (eq. 1.20) é solução da equação funcional. Mostre que esta é a única forma se G for diferenciável duas vezes em cada argumento.

Da equação 1.21 obtemos

$$\frac{\phi(ae|d)}{\phi_v} = \frac{\phi(a|d)}{\phi_v}\frac{\phi(e|ad)}{\phi_v} \tag{1.25}$$

o que permite regraduar mais uma vez os números associados as crenças sem mudar sua ordem. Crenças regraduadas, de forma bijetora representam o mesmo ordenamento e portanto podem ser ainda chamados de crenças. Definimos os novos números

$$p(a|b) = \frac{\phi(a|b)}{\phi_v} \tag{1.26}$$

e reescrevemos os resultados

$$\begin{aligned} p(a|a) &= p_v = 1 \\ p(a|\overline{a}) &= p_f = 0 \\ p(a \vee b|c) &= p(a|c) + p(b|c) - p(ab|c) \\ p(ab|c) &= p(a|c)p(b|ac) \\ &= p(b|c)p(a|bc) \end{aligned} \tag{1.27}$$

Começamos a reconhecer as fórmulas que descrevem as probabilidades da soma e do produto. Mas ainda não acabamos. Precisamos determinar o que acontece com a negação. Começamos com $a \vee \overline{a}|d$ que deve ser sempre verdade e com $a\overline{a}|d$ que deve ser sempre falso:

$$\begin{aligned} 1 &= p(a \vee \overline{a}|d) \\ &= p(a|d) + p(\overline{a}|d) - p(a\overline{a}|d) \\ &= p(a|d) + p(\overline{a}|d), \end{aligned} \tag{1.28}$$

ou a soma das crenças regraduadas de uma asserção e da sua negação é um.

Isso completa a identificação das crenças ou plausibilidade regraduadas em números que satisfazem as regras da probabilidade. Concluimos que a estrutura matemática adequada, e que usaremos nestas notas, para descrever situações de informação incompleta é a teoria de probabilidades.

O que foi obtido pode ser comparado com os axiomas de Kolmogorov [12]. Vemos uma diferença importante. Na formulação da teoria de probabilidade como um capítulo da teoria da medida, as probabilidades são medidas e não há menção a condicionais. Rao adicionou mais tarde a complementação introduzindo, como uma idéia tardia, a probabilidade condicional definida a partir

[12] Kolmogorov

do teorema de Bayes, que Cox obteve como uma consequência direta da consistência. A partir de

$$\begin{aligned} p(ab|c) &= p(a|c)p(b|ac) \\ &= p(b|c)p(a|bc) \end{aligned} \tag{1.29}$$

obtemos o teorema de Bayes [13]

$$p(a|bc) = \frac{p(a|c)p(b|ac)}{p(b|c)} \tag{1.30}$$

que esconde, atrás de sua grande simplicidade, uma importancia enorme, que deriva em parte do número ilimitado que tem encontrada em várias áreas da ciência.

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 o ordenamento das crenças e 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 construido 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.

1.2.5 O teorema de Bayes e 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 = "$A \to B$"

[13] T. Bayes formulou a parte da inversão: $p(a|bc) \propto p(b|ac)$, Laplace o escreveu pela primeira vez e deu-lhe a devida importância

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

$$P(A|BCI) = \frac{P(A|CI)P(B|ACI)}{P(B|CI)}, \tag{1.31}$$

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

$$P(B|ACI) = 1. \tag{1.32}$$

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

$$P(B|CI) \leq 1. \tag{1.33}$$

Substituindo estes resultados obtemos

$$P(A|BCI) \geq P(A|CI), \tag{1.34}$$

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

$$P(B|\overline{A}CI) = \frac{P(B|CI)P(\overline{A}|BCI)}{P(\overline{A}|CI)}, \tag{1.35}$$

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

$$P(B|\overline{A}CI) \leq P(B|CI) \tag{1.36}$$

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 de sol 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. O que significa um resultado positivo? Para proceder, o mais importante é esclarecer quais são as asserções relevantes.

Consideremos as asserções

- D="paciente está doente"
- A="resultado do teste é positivo"

junto com os dados sobre

- especificidade: $P(A|D) = .90$, a probabilidade de dar positivo no teste na condição de estar doente
- sensibilidade: $1 - P(A|\overline{D}) = 1. - .2 = .8$, 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,

$$P(D|AI) = \frac{P(D|I)P(A|DI)}{P(A|I)}, \tag{1.37}$$

e também

$$P(\overline{D}|AI) = \frac{P(\overline{D}|I)P(A|\overline{D}I)}{P(A|I)}, \tag{1.38}$$

os denominadores são inconvenientes e os eliminamos olhando para a razão

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)}. \tag{1.39}$$

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

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)} = \frac{.01 \times .90}{.99 \times .20} = 0.045. \tag{1.40}$$

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.

1.2.6 Jaynes e o bom senso

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, esta seria mais plausível nas condições C que C'; isto é seria desejável que a teoria satisfizesse

- $A|C \geq A|C'$ e $B|DC = B|DC'$, para qualquer D, 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. Jaynes coloca isto como um dos axiomas para chegar à teoria de probabilidades.

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'.

Pelo regra do produto, podemos provar isto

$$P(AB|C) = P(A|C)P(B|AC) = P(A|C)P(B|AC')$$
$$P(AB|C) \geq P(A|C')P(B|AC') = P(AB|C').$$

1.2.7 Exemplo do Teorema de Bayes e Ajuste de funções

Uma das primeiras lições que os estudantes de física tem ao entrarem num laboratório é sobre ajuste de curvas usando conjuntos de medidas empíricas.

Um objeto cai e medimos as posições ou velocidades como função do tempo. Estão de acordo com o que se espera de um objeto que cai na presença de um campo gravitacional? Qual é o valor de g, a aceleração da gravidade? Só para deixar isto claro, não faltarão exemplos compicados mais adiante nestas notas, olharemos para o caso em que obtemos um conjunto de dados

$$D = \{v_1, v_2, ...v_N\} \tag{1.41}$$

para as velocidades medidas em

$$T = \{t_1, t_2, ...t_N\}. \tag{1.42}$$

O modelo que queremos avaliar, refutar ou aceitar (pelo menos até ter mais dados) é

$$\mathcal{M} : v = v_0 + gt \tag{1.43}$$

A pergunta que quer ser respondida diz respeito a asserções do tipo $H(g)$:"O valor da aceleração da gravidade é g". Para cada valor de g que for inserido nessa frase, teremos uma asserção diferente. O que queremos é comparar o mérito de cada asserção, qual é a probabilidade de cada uma delas, para todos os valores que possam ser inseridos.

O teorema de Bayes nos permite escrever

$$P(H|DI) = \frac{P(H|I)P(D|HI)}{P(D|I)}. \tag{1.44}$$

O que será discutido a seguir é fundamental para este curso. Será discutido em contextos mais complícados e portanto vale a pena o esforço de entender cada passo. É tão importante que cada termo recebe um nome.

Em primeiro lugar temos que definir as asserções relevantes ao problema. A parte que parece menos importante, mas que na realidade é fundamental é I, que define várias coisas que de tão importantes são consideradas desnecessárias pois, para que falar o óbvio? I denota toda a informação sobre a experiência. Qual é a teoria que queremos confrontar com os dados? Quais são as características do aparelho de medida? Em que instantes de tempo t_i fazemos as medidas, quais as incertezas que estas medidas têm? Em que planeta estamos? e muito mais que ficará tácitamente escondida, mas ainda relevante.

D é o conjunto de dados, H é a hipótese que quer ser testada.

Agora o significado das probabilidades que aparecem na equação 1.44. Começamos pelo conhecimento que temos sobre o contexto experimental mas sem levar em consideração os dados. A distribuição de probabilidades *a priori* $P(H|I)$ codifica tudo o que sabemos sobre a gravitação antes de entrar no laboratório. Se não soubermos o planeta onde a experência é realizada, fica difícil esperar um valor e não outro. Todas as gerações de estudantes que fizeram esta experiência,

dos quais temos noticia, o fizeram na terra. O resultado deu algo que se parece com 9.8 ms^{-2}. Se o resultado final fosse 9.8 kms^{-2} o aluno ficaria tentado a mudar seu resultado, mudaria de forma ad hoc seu valor no relatório, o que seria desonesto, ou faria novamente as contas. Se ainda persistir o problema, jogaria fora os dados. Isto é desonesto? Não se estiver de acordo com a sua probabilidade a priori. Qual é a probabilidade que a aceleração da gravidade seja 9.8 kms^{-2} em São Paulo? qual é a probabilidade que você atribuiria antes de entrar no laboratório? Quanto voce estaria disposto a apostar contra a veracidade dessa asserção? A priori, o estudante sabe que o valor estará por volta de dez, e pode ser constante entre 7 e 15. Muito mais que isso ou muito menos, deve ser erro, e é melhor jogar o que o estudante chama de ponto fora da curva. Isso é perfeitamente lógico e deve ser feito a não ser que em I haja a possibilidade de que algo possa mudar o valor esperado. Por exemplo a experiência esta sendo feita em cima de uma cratéra aberta por um meteorito composto do elemento X. Então podemos permitir a suposição que novos valores sejam encontrados. Seriamos cegos se considerassemos a probabilidade a priori de encontrar valores muito diferentes, nula e se assim for feito, certamente não os encontraremos.

A probabilidade $P(D|HI)$ descreve quão verossímel seria encontrar esse conjunto de dados se além de I, o valor particular de g representado por H fosse o correto. Esta é a famosa contribuição do reverendo Thomas Bayes[14]: a inversão. Queriamos saber a probabilidade de g ter um certo valor nas condições que os dados foram observados, mas estamos olhando para a probabilidade dos dados no caso que a teoria (contida em I) e um valor particular do parâmetro g sejam verdade. Este termo recebe o nome verossimilhança (likelihood em inglés).

O denominador $P(D|I)$ será interessante em outros contextos. Em geral é chamado de evidência. Pode ser obtido usando o fato que g não pode ter dois valores diferentes. As asserções para valores de de g diferentes são mutuamente exclusivas. Portando a soma sobre todas as possibilidades é um. Neste caso em que g toma valores reais, é interessante considerar que as asserções tem o significado que o valor da aceleração da gravidade está entre g e $g + dg$, somas são substituidas por integrais.

O resultado de toda a análise será a obtenção de $P(H|DI)$ que se chama a distribuição de (densidade de) probabilidade posterior, ou simplesmente a posterior.

A crítica mais comum é que a *realidade objetiva* é única e portanto não é possível que haja uma probabilidade para o valor de g. Mas não é isso o que esta probabilidade significa. g pode ter um valor único objetivo [15]. O que a posterior, ou a a priori significam é que não temos informação completa e que só podemos atribuir probabilidades às diferentes asserções sobre o valor de g. Mais dados, ou seja mais informação, permitirão novas estimativas. O que estas

[14] referencia de bayes

[15] Sabemos que g uniforme, constante é só uma aproximação válida para quedas em distâncias pequenas em comparação ao raio da terra dentro da teoria de Newton. Mas também sabemos que essa teoria não é final, tendo sido substituida pela de Einstein, e certamente não sabemos por qual teoria vai ser substituida em anos futuros. Não sobra muito do conceito de um g que descreve uma realidade objetiva

probabilidades codificam não é o valor de g, mas a crença que esse seja o valor correto.

1.2.8 Obtendo a posterior

Há vários exemplos que mostram a importância de determinar a distribuição a priori com muito cuidado. Podemos dizer que a probabilidade que $g < 0$ deve ser zero. Os objetos mais densos que o ar não caem para cima. Também podemos limitar os valores superiores. Poderiamos dizer que $P(H|I) = c$ se $g_{min} < g < g_{max}$ e zero fora desse intervalo. A constante c é tal que $\int_{g_{min}}^{g_{max}} P(H|I)dg = 1$ ou $c^{-1} = g_{max} - g_{min}$.

A verossimilhança $P(D|HI)$ leva em conta que as medidas são sujeitas a erros. Poderiamos dizer, por exemplo, que o modelo teórico e o modelo sobre o aparelho de medidas, juntos nos levam a esperar, que para os valor de tempo t_i, onde é feita a medida,

$$v_i = v_0 + gt_i + \eta_i. \tag{1.45}$$

O resultado esperado puramente pelo modelo teórico (eq. 1.43) é corrompido por algo que chamamos ruído. Isto esconde uma grande quantidade de ignorância sobre o processo de medida. Se pudessemos aumentar o contrôle sobre o aparelho de medida (e.g. temperatura, vento, correntes elétricas, valores das resistências, ...etc.) a amplitude de η_i poderia ser menor. Mas sempre há uma incerteza sobre o valor medido. Temos que fazer algumas hipotéses sobre η_i. Estas, supostas verdadeiras, serão incluidas na asserção I. Como não temos informação completa, devemos descrever o conjunto de ηs por uma distribuição de probabilidade $P(\eta_1....\eta_N|I_{exp})$. É razoável supor que as diferentes medidas são independentes, e usando a regra do produto lógico

$$\begin{aligned} P(\eta_1\eta_2....\eta_N|I_{exp}) &= P(\eta_1|I_{exp})P(\eta_2....\eta_N|\eta_1 I_{exp}) \\ &= P(\eta_1|I_{exp})P(\eta_2\eta_3....\eta_N|I_{exp}) \\ &= P(\eta_1|I_{exp})P(\eta_2|I_{exp})P(\eta_3....\eta_N|\eta_2 I_{exp}) \\ &... \\ &= \prod_i^N P(\eta_i|I_{exp}), \end{aligned} \tag{1.46}$$

onde usamos na primeira e terceira linha a regra do produto e na segunda a independência dos valores de $\eta_2, \eta_3....\eta_N$ e o de η_1. Temos que a distribuição conjunta é o produto das distribuições inddividuais.

Qual é a distribuição $P(\eta|I_{exp})$ a ser usada. Ainda devemos supor algo mais, por exemplo média nula e variância finita σ^2. No capítulo sobre entropia justificaremos porque isto nos leva a uma distribuição gaussiana

$$P(\eta_1, \eta_2....\eta_N|I_{\exp}) = \frac{e^{-\sum_{i=1}^N \frac{\eta_i^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}$$

mas $\eta_i = v_i - v_0 - gt_i$, portanto

$$P(\eta_1, \eta_2 .... \eta_N | I_{\text{exp}}) = \frac{e^{-\sum_{i=1}^{N} \frac{(v_i - v_0 - gt_i)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}$$

e a posterior

$$P(H|DI) = \frac{P(H|I)}{PD|I)} \frac{e^{-\sum_{i=1}^{N} \frac{(v_i - v_0 - gt_i)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}. \tag{1.47}$$

O problema de inferência está pronto. Mas qual é a resposta a ser dada? Há várias quantidades que podem ser extraidas da posterior. Por simplicidade nos contentamos com o valor de g que é mais provável. Se a a priori é constante na região que a gaussiana é relevante, podemos esquecer o prefator. Teremos a estimativa conhecida como máxima verossimilhança. A resposta é simplesmente o valor que torna o argumento da exponencial máximo,

$$g_{MV} = \arg\min_g \sum_{i=1}^{N} \frac{(v_i - v_0 - gt_i)^2}{2\sigma^2} \tag{1.48}$$

que é o velho método de mínimos quadrados. Se $P(H|I)$ fosse relevante, teriamos o máximo aposteriori g_{MAP}.

Mas escolher um valor sobre os outros esconde que não temos certeza absoluta. A largura da gaussiana $\sigma/\sqrt{N}$ nos dá uma medida da incerteza.

Ainda podemos levar em conta que valores vizinhos de g_{MAP} tem probabilidade não desprezível e apresentar o valor esperado

$$g^* = \int gP(H|DI)dg. \tag{1.49}$$

O que ganhamos em apresentar assim o método dos mínimos quadrados que os estudantes devem ter visto há muito tempo? Suponha por exemplo, que voce colha mais informação sobre o aparelho de medida e chegue à conclusão que a distribuição dos η não é gaussiana. Ainda assim usaria o método dos mínimos quadrados? Podemos ver quais as suposições necessárias e tentar verificar se cada uma delas é razoável ou não. Isto não é pouco, a apresentação cuidadosa pode evitar suposições que não gostariamos de fazer ao analisar os dados de uma experiência. Tão importante quanto usar a informação disponível é não usar a que não o é. O próximo capítulo levará esta idéia adiante.

Capítulo 2

Probabilidades e o Teorema do Limite Central

2.1 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 subconjuntos de E. Um sistema de conjuntos é chamado um campo se a soma, produto, interseção de dois elementos quaisquer pertencem ao sistema. Os axiomas são

- AK1) $\mathcal{F}$ é um campo de conjuntos fechado ante um número de uniões e interseções enumeráveis e se $A \in \mathcal{F}$ e $\overline{A} = E - A$, então $\overline{A} \in \mathcal{F}$

ou seja $\mathcal{F}$ é um σ-campo,

- AK2) $\mathcal{F}$ contém o conjunto E.

- AK3) A cada conjunto $A \in E$ é atribuido um número real não negativo, chamado de probabilidade do evento A.

[1] e.g. ver http://www.mathematik.com/Kolmogorov/index.html

- AK4) $P(E) = 1$
- AK5) Se $A \vee B = \emptyset$, então $P(A + B) = P(A) + P(B)$

Vemos que estes axiomas estão de acordo com os resultados do capítulo anterior. A probabilidade da *certeza* é 1 por AK4; a probabilidade está entre zero e um 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

$$p(A|B) = \frac{P(AB)}{P(B)} \tag{2.1}$$

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.

2.2 Momentos...

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 estimativa de X. Há várias possibilidades e cada uma tem utilidade

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

- (1) $x_M = maxargP(x)$
- (2) $< x >= I\!E[x] = \int_I xP(x)dx$
- (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 idéia flutuação não seja útil, só que por construção a sua média é nula. A média do seu quadrado é muito útil:

$$\sigma_x^2 =< (x - < x >)^2 > . \tag{2.2}$$

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:

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

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

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

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 entre 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, esperando que isso não confunda o leitor, mas torne imune às várias notações na literatura.

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}.

2.2.1 Marginais, Independência

Considere a distribuição conjunta de duas variáveis X, e Y. Para cada valor x que X pode tomar, podemos *marginalizar*, isto é calcular a soma sobre todos os possíveis valores de Y

$$p(x) = \sum_y p(x, y) \tag{2.3}$$

é chamada de distribuição de probabilidades marginal, que não é nada além da distribuição de probabilidades de X. O processo de soma (ou integração no caso de valores no contínuo). É claro que também podemos marginalizar somando sobre os valores que X pode tomar:

$$p(y) = \sum_x p(x, y). \tag{2.4}$$

Também podemos usar a regra do produto $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$. No caso especial em que conhecimento de X não diz nada sobre Y, ou seja quando $p(y|x)$ não depende de x segue que $p(y|x) = p(y)$ e assim

$$p(x, y) = p(y)p(x|y) = p(x)p(y), \tag{2.5}$$

e as variáveis X e Y são ditas *independentes*.

Exercício

Prove que se $p(x|y)$ não depende de y, então $p(y|x)$ não depende de x.

2.3 Convoluções e Cumulantes

Considere variáveis idénticas X_i que tomam valores reais $\{x_1, x_2, ...\}$ tal que $P(x_i)$ é o mesmo para todo i. Consideraremos o caso em que para qualquer $i \neq j$, os X_i são independentes entre si e são igualmente distribuidos [4]. Estamos

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

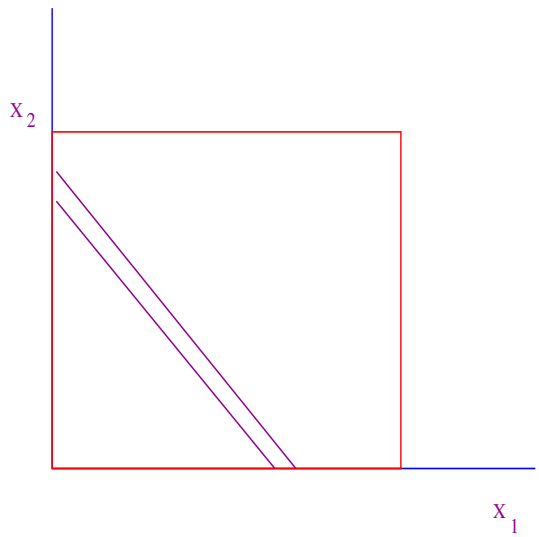

Figura 2.1: No plano X_1X_2 temos a região onde o valor de Y está entre y e $y + dy$. Todos os pares x_1 e x_2 nela contribuem para a probabilidade de Y

interessados na variável Y que toma valores em $y = \sum_{i=1..n} x_i$. 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 2.1. 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

$$P(y|N = 2)dy = \int_{y \leq x_1 + x_2 \leq y + dy} dx_1 dx_2 P(x_1)P(x_2), \tag{2.6}$$

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].

$$P(y|N = 2)dy = \int \chi_{y \leq x_1 + x_2 \leq y + dy} dx_1 dx_2 P(x_1)P(x_2), \tag{2.7}$$

[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.

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 $\delta_n(A)$:

$$\begin{aligned} \delta_n(y \leq x_1 + x_2 \leq y + \Delta y_n) &= \frac{1}{\Delta y_n}, \quad \text{se } y \leq x_1 + x_2 \leq y + \Delta y_n \\ &= 0, \quad \text{se não} \end{aligned} \tag{2.8}$$

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

$$P(y|N = 2) = \int dx_1 dx_2 P(x_1) P(x_2) \delta(y - x_1 + x_2), \tag{2.9}$$

$$P(y|N = 2) = \int dx P(x) P(y - x), \tag{2.10}$$

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

Outra forma: marginalização

Podemos ver como o resultado acima decorre das regras da probabilidade de outra forma: marginalizando. Começamos com a distribuição conjunta das variáveis Y, X_1 e X_2 e integramos sobre todos os valores de X_1 e X_2:

$$P(y) = \int dx_1 dx_2 P(y, x_1, x_2). \tag{2.11}$$

Da regra do produto

$$P(y) = \int dx_1 dx_2 P(y|x_1, x_2) P(x_1, x_2). \tag{2.12}$$

Mas Y esta totalmente determinado se X_1 e X_2 forem conhecidos, portanto $P(y|x_1, x_2) = \delta(y - x_1 - x_2)$. Se novamente considerarmos X_1 e X_2 independentes: $P(x_1, x_2) = P(x_1)P(x_2)$, obtemos novamente a equação 2.10. A vantagem disto é que podemos facilmente obter expressões para funções gerais. Se $y = f(x_1, x_2)$, então

$$P(y) = \int dx_1 dx_2 \delta(y - f(x_1, x_2)) P(x_1, x_2) \tag{2.13}$$

Exercício

Discuta a diferença entre $P(Y)$ a distribuição da soma e $P(X_1 X_2)$, para o produto lógico, que denotaremos por $P(x_1, x_2)$ a chamamos de 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: obtenha uma expressão para $P(z)$ quando $z = x_1 x_2$.

Distribuição da soma de variáveis e a função característica

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

- $\int P(x)dx = 1$, $P(x) \geq 0$ 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) [6] e a inversa

$$\hat{P}(k) = \int_{-\infty}^{\infty} e^{-ikx} P(x)dx \tag{2.14}$$

$$P(x) = \int_{-\infty}^{\infty} e^{ikx} \hat{P}(k) \frac{dk}{2\pi} \tag{2.15}$$

A TF de uma distribuição de probabilidades é chamada de função característica. Ela também é chamada de função geradora dos momentos, pois uma expansão formal em série de potências da exponencial na equação 2.14 e a troca da ordem de integração e somatória nos mostra que os coeficientes estão relacionados aos momentos:

$$\begin{aligned} \hat{P}(k) &= \int_{-\infty}^{\infty} \sum_{s=0}^{\infty} \frac{(-ikx)^s}{s!} P(x) \\ &= \sum_{s=0}^{\infty} \frac{(-ik)^s}{s!} < x^s > \end{aligned} \tag{2.16}$$

Tomemos a TF dos termos da equação 2.10, e usando :

$$\delta(k) = \int \frac{dx}{2\pi} e^{ikx} \tag{2.17}$$

obtemos

$$\begin{aligned} \hat{P}(k|N=2) &= \int dydx e^{-iky} P(x)P(y-x), \\ &= \int \frac{dxdydk_1dk_2}{(2\pi)^2} \hat{P}(k_1|1)\hat{P}(k_2|1) e^{-iky+ik_1x+ik_2(y-x)}. \end{aligned} \tag{2.18}$$

Integrando sobre x e usando a representação da delta:

$$\begin{aligned} &= \int \frac{dydk_1dk_2}{2\pi} \hat{P}(k_1|1)\hat{P}(k_2|1) e^{-iky+ik_2y} \delta(k_1-k_2), \\ &= \hat{P}(k|1)\hat{P}(k|1) = \hat{P}^2(k|N=1) \end{aligned} \tag{2.19}$$

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

Para a soma de $N = n$ variáveis x_i

$$P(y|N = n) = \int \prod_{i=1...n} dx_i P(x_1)P(x_2)...P(y - \sum_{i=1}^{n-1} x_i), \tag{2.20}$$

ou, introduzindo uma integral mais

$$P(y|N = n) = \int \prod_{i=1...n} dx P(x_1)P(x_2)...P(x_n)\delta(y - \sum_{i=1}^{n} x_i), \tag{2.21}$$

obtemos

$$\hat{P}(k|N = n) = \hat{P}^n(k|N = 1) \tag{2.22}$$

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, 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 2.22 e dado que produtos, ao tomar logaritmos, viram somas, temos

$$\log \hat{P}(k|N = n) = \sum_{i}^{n} \log \hat{P}(k|N = 1) = n \log \hat{P}(k|N = 1). \tag{2.23}$$

Isto nos leva a discutir os cumulantes $\{C_s(n)\}$ de uma distribuição, definidos através da expansão em série de potências de ik da sua função caracterísitca:

$$\log \hat{P}(k|N = n) = \sum_{s=0}^{\infty} C_s(n) \frac{(-ik)^s}{s!}. \tag{2.24}$$

A equação 2.23 nos indica o motivo do nome dos cumulantes: a aditividade (ou acúmulo) ante convoluções

$$\begin{aligned} C_s(\sum_i x_i) &= \sum_i C_s(x_i), \\ C_s(n) &= nC_s(1), \end{aligned} \tag{2.25}$$

onde a equação 2.25 segue porque as $\{x_i\}$ são identicamente distribuidas. Concluimos que quando variáveis aleatórias independentes se somam, os cumulantes da distribuição da soma são 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

$$C_s = \frac{1}{(-i)^s} \frac{d^s \log \hat{P}}{dk^s}|_{k=0} \tag{2.26}$$

Podemos calcular alguns dos primeiros,

$$\begin{aligned}
\log \quad & \hat{P}(k|N=1) = \log \int e^{-ikx} P(x)dx \\
= \quad & \log \int \sum_{s=0}^{\infty} \frac{(-ikx)^s}{s!} P(x)dx \\
= \quad & \log(1 + \sum_{s=1}^{\infty} \frac{(-ik)^s}{s!} < x^s >) \\
= \quad & \sum_{s_1=1}^{\infty} \frac{(-ik)^{s_1}}{s_1!} < x^{s_1} > - \frac{1}{2} \sum_{s_1,s_2=1}^{\infty} \frac{(-ik)^{s_1+s_2}}{s_1! s_2!} < x^{s_1} >< x^{s_2} > \\
+ \quad & \frac{1}{3} \sum_{s_1,s_2,s_3=1}^{\infty} \frac{(-ik)^{s_1+s_2+s_3}}{s_1! s_2! s_3!} < x^{s_1} >< x^{s_2} >< x^{s_3} > + ... \qquad (2.27)
\end{aligned}$$

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 $< x^s >$:

$$\begin{aligned}
C_0 &= 0, \\
C_1 &= < x >, \\
C_2 &= < x^2 > - < x >^2, \\
C_3 &= < x^3 > -3 < x^2 >< x > +2 < x >^3, \\
C_4 &= < x^4 > -4 < x^3 >< x > -3 < x^2 >^2 +12 < x^2 >< x >^2 \\
&- 6 < x >^4, \qquad (2.28)
\end{aligned}$$

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:

$$nC_1^x = C_1^Y(n) = \sqrt{n} C_1^Z(n) = nC_1^W(n), \qquad (2.29)$$

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

$$nC_s^x = C_s^Y(n) = n^{s/2} C_s^Z(n) = n^s C_s^W(n), \qquad (2.30)$$

Portanto

$$\begin{aligned}
C_s^Y(n) &= nC_s^x \\
C_s^Z(n) &= \frac{1}{n^{\frac{s}{2}-1}} C_s^x \\
C_s^W(n) &= \frac{1}{n^{s-1}} C_s^x, \qquad (2.31)
\end{aligned}$$

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

$$u_s^Y(n) = \frac{C_s^Y(n)}{(C_2^Y(n))^{\frac{s}{2}}} = n^{1-\frac{s}{2}} u_s^x, \tag{2.32}$$

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.

2.4 O Teorema do Limite Central I

2.4.1 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 2.24, a função característica de Y é

$$\hat{P}(k|n) = \exp(\sum_{s=0}^{\infty} C_s^Y(n) \frac{(-ik)^s}{s!}) \tag{2.33}$$

e a distribuição e'

$$P(Y|N=n) = \int_{-\infty}^{\infty} \exp(iky - \sum_{s=0}^{\infty} C_s^Y(n) \frac{(-ik)^s}{s!}) \frac{dk}{\sqrt{2\pi}}. \tag{2.34}$$

Até aqui é exato. Desprezando os cumulantes de ordem superior a segunda, pela eq. 2.32

$$P(Y|N=n) = \int_{-\infty}^{\infty} \exp(iky + ikC_1^Y(n) - k^2 C_2^Y(n)/2) \frac{dk}{\sqrt{2\pi}} \tag{2.35}$$

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

$$P(Y|N=n) = \frac{1}{\sqrt{2\pi n \sigma_x^2}} e^{-\frac{(y-n\mu)^2}{2n\sigma_x^2}} \tag{2.36}$$

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

$$\begin{aligned} P(Z|N=n) &= \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(z-\sqrt{n}\mu)^2}{2\sigma_x^2}} \\ P(W|N=n) &= \frac{1}{\sqrt{2\pi\frac{\sigma_x^2}{n}}} e^{-\frac{(w-\mu)^2}{2\frac{\sigma_x^2}{n}}} \end{aligned} \tag{2.37}$$

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. 2.36 e 2.37) 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.

2.5 A média e a concentração em torno da média

A distribuição de W dada pela eq. 2.36 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 2.31, temos que $C_2^W(n) = \sigma_x^2/n$, satisfaz

$$\begin{aligned} C_2^W(n) &= \int_{-\infty}^{\infty} dw(w^2 - < w >^2)P(W = w|N = n) \\ &= \int_{-\infty}^{\infty} dw(w - < w >)^2 P(W = w|N = n) \\ &\geq \int_{|w-<w>|\geq\epsilon} dw(w - < w >)^2 P(W = w|N = n) \\ &\geq \int_{|w-<w>|\geq\epsilon} dw\epsilon^2 P(W = w|N = n) \\ &\geq \epsilon^2 Prob(|w - < w > | \geq \epsilon). \end{aligned} \tag{2.38}$$

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:

$$Prob(|w - < w > | \geq \epsilon) \leq \frac{C_2^W(n)}{\epsilon^2} \tag{2.39}$$

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:

$$Prob(|w - < w > | \geq \epsilon) \leq \frac{C_2^x}{n\epsilon^2} \tag{2.40}$$

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'[7] onde p, q tem outro valores (entre zero e um). Consideremos as jogadas independentes,

[7] 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.

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 $f = m/n$

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

$$P(m|N = nI') = \frac{n!}{m!(n-m)!}p^m q^{n-m} \tag{2.41}$$

- **(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 $1/n$.

- **(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. A frequência é informação que pode ser usado para atribuir um número à probabilidade, mas não é o único tipo de informação para fazer isso.

2.6 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 de Taylor da equação 2.34 que foram *desprezados* anteriormente para chegar até a equação 2.35. Pense na diferença entre *desprezar* e *controlar*. Muitas vezes, em Física *desprezamos* sem *controlar*, pois tentar *controlar* pode ser tão díficl que evitaria a possibilidade de avanço. Uma vez que se encontra algo interessante, sempre podemos voltar atrás e tentar *controlar* termos antes *desprezados* usando a nova maneira de enxergar um problema, que o avanço menos cuidadoso permitiu.

2.7 O Teorema do Limite Central III

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

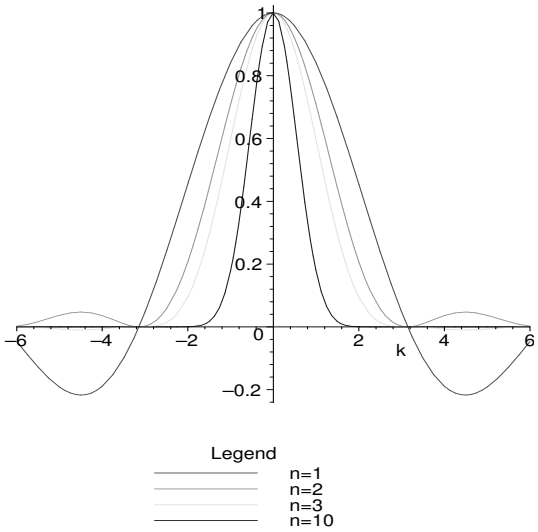

Figura 2.2: A função característica $\hat{P}(k|N = n)$ para a soma de n varíaveis uniformemente distribuidas.

2.7.1 A distribuição uniforme

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

$$\hat{P}(k|1) = \frac{1}{L}\int_{-L/2}^{L/2} e^{-ikx}dx = \frac{2}{kL}\sin(\frac{kL}{2}) \tag{2.42}$$

$$P(Y = y|N = n) = \int_{-\infty}^{\infty}[\hat{P}(k|1)]^n e^{iky}\frac{dk}{2\pi} = \int_{-\infty}^{\infty}[\frac{2}{kL}\sin(\frac{kL}{2})]^n e^{iky}\frac{dk}{2\pi} \tag{2.43}$$

A figura 2.2 mostra que a função caracterísitica fica mais parecida com uma gaussiana e na figura 2.3 vemos que efetivamente o $\log(|\hat{P}(\sqrt{|u|}|N = n)|)$ com $u = \pm k^2$ fica cada vez mais perto de $-c|u|$ (gaussiana).

2.7.2 A distribuição exponencial

A distribuição $P(x) = \Theta(x)ae^{-xa}$ é chamada de exponencial. Mostre que $\mu = \sigma^2 = a^{-1}$. A função característica

$$\hat{P}(k|1) = \int_0^{\infty} ae^{-xa}e^{-ikx}dx = \frac{a}{a+ik} \tag{2.44}$$

$$P(Y = y|N = n) = \int_{-\infty}^{\infty}(\frac{a}{a+ik})^n e^{iky}\frac{dk}{2\pi} \tag{2.45}$$

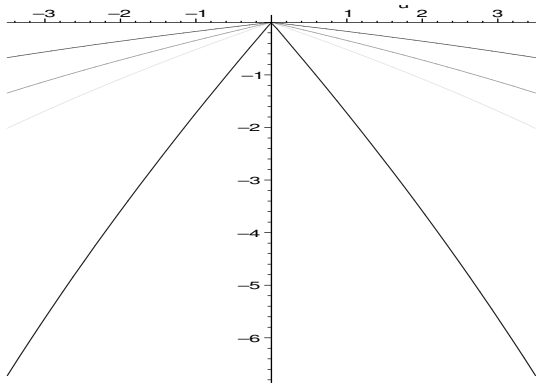

Figura 2.3: A função característica $\log(|\hat{P}(\sqrt{|u|}|N=n)|)$ contra $u = \pm k^2$ para a soma de n varíaveis uniformemente distribuidas. Nesta representação gaussianas aparecem como retas. A legenda é a mesma da figura 2.2

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

$$P(Y = y|N = n) = \Theta(y)a^n \frac{y^{n-1}e^{-ay}}{(n-1)!}. \tag{2.46}$$

Faça a conta. Obviamente não é uma gaussiana, mas uma distribuição gamma. No entanto, a região central sim, se parece com uma gaussiana.

A figura 2.4 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 2.5. 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.

2.7.3 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. 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 2.7 a binomial foi dividida por ΔY. De outra

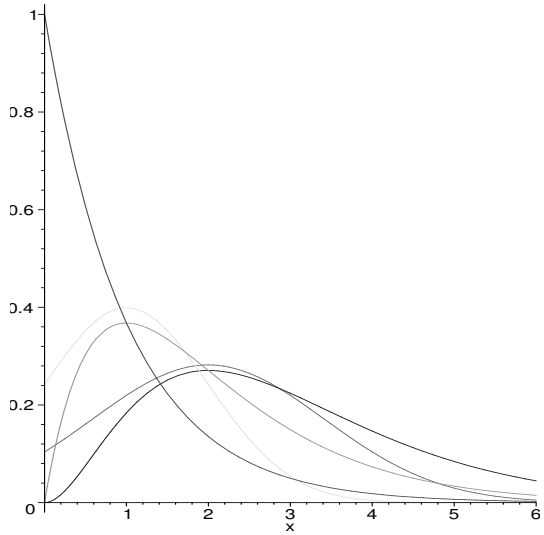

Figura 2.4: A densidade de probabilidade $P(y|N = n, a = 1)$ para a soma de n varíaveis exponencialmente distribuidas, $n = 1, 2, 3$. Para as duas últimas mostramos as gaussianas com $\mu = \sigma^2 = n - 1$

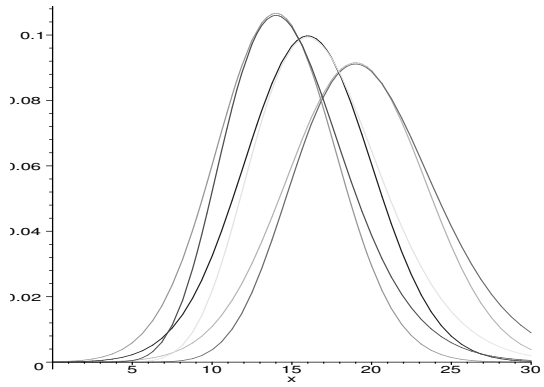

Figura 2.5: A densidade de probabilidade $P(y|N = n, a = 1)$ para a soma de n varíaveis exponencialmente distribuidas, $n = 15, 17, 20$. Junto estão mostradas as gaussianas com $\mu = \sigma^2 = n - 1$.

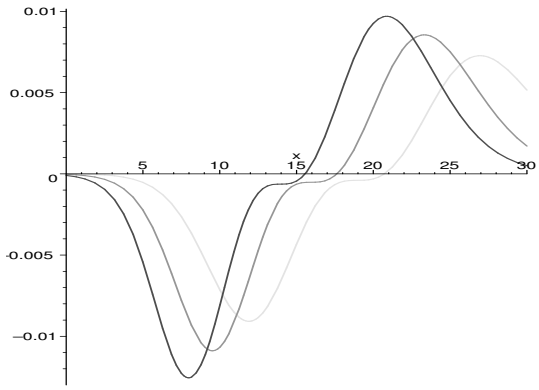

Figura 2.6: A diferença entre a densidade de probabilidade $P(y|N = n, a = 1)$ para a soma de n varíaveis exponencialmente distribuidas, $n = 15, 17, 20$ e as gaussianas com $\mu = \sigma^2 = n - 1$. Os mesmos parâmetros da figura 2.5. Note que a região central é bem aproximada. Há uma região de transição, ao afastar-se para as caudas, e finalmente as caudas vão rapidamente para zero, assim como a sua diferença

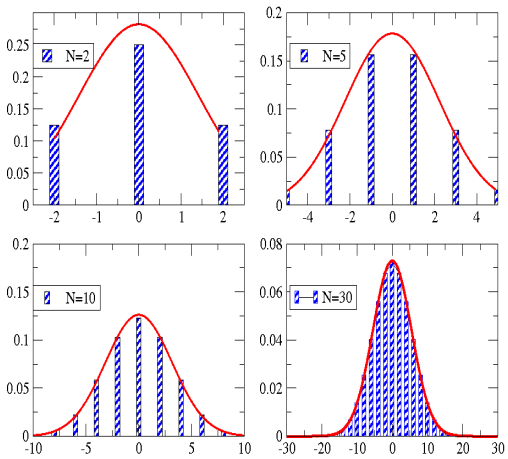

Figura 2.7: A binomial (dividida por $\Delta Y = 2$, barras) e a densidade gaussiana correspondente (linha contínua), para $N = 2, 5, 10$ e 30

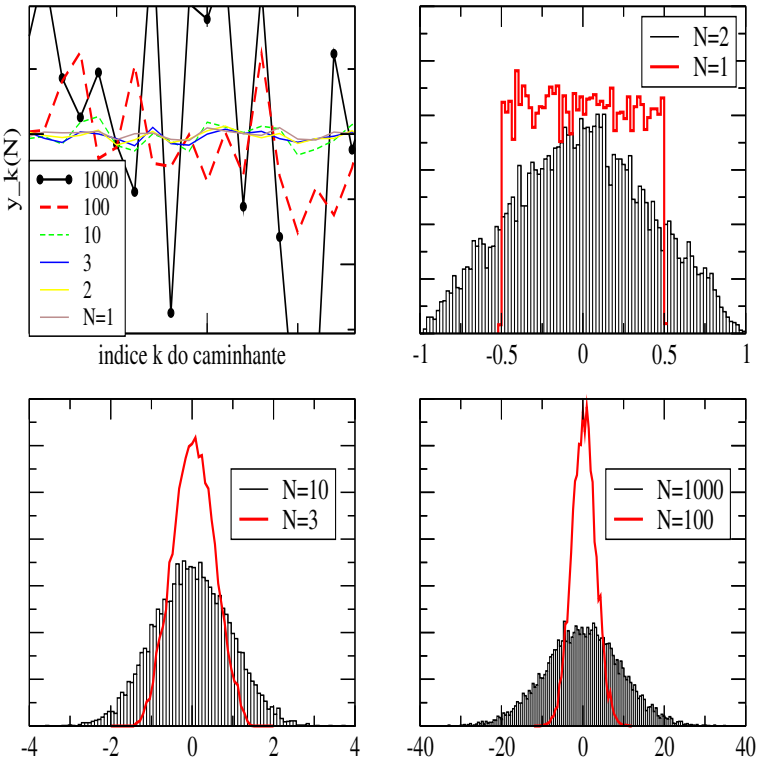

Figura 2.8:

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 $y+2$.

2.7.4 Caminho Aleatório

Novamente olhamos para a distribuição binomial. Olhe para a figura 2.8. Definimos o caminho aleatório através de

Difusão: $K(= 10000$ na figura 2.8) seqüências de N passos de um processo binomial, definidos por

$$y_n = y_{n-1} + x_n \tag{2.47}$$

onde $P(x = 1) = p$ e $P(x = -1) = q = 1 - p$. 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 tem probabilidade dada pela binomial. Mostramos agora que a para valores altos de n a binomial se parece com uma gaussiana.

O resultado necessário é a aproximação de Stirling para o fatorial,

$$\log N! = (N + \frac{1}{2}) \log N - N + \frac{1}{2} \log 2\pi + \log \left(1 + \frac{1}{12N} + \frac{1}{288N^2} + \mathcal{O}(\frac{1}{N^3}) \right) \tag{2.48}$$

que é uma expansão assintótica, isto é melhora quando N aumenta [8]. Não precisamos todos esses termos, basta $\log N \approx N \log N - N$ onde desprezamos $\mathcal{O}(\log N)$

$$\begin{aligned} \log P(m|N = nI') &= \log n! - \log m! - \log(n-m)! + m \log p + (n-m) \log q \\ &= n \log n - n - m \log m + m - (n-m) \log(n-m) \\ &\quad + n - m + m \log p + (n-m) \log q \end{aligned} \tag{2.49}$$

Podemos tratar m como uma variável real e encontrar onde a probabilidade$P(m|N = nI')$ atinge o valor máximo. Tomamos a primeira e segunda derivadas

$$\begin{aligned} \frac{\partial \log P(m|N = nI')}{\partial m} &= -\log m + \log(n-m) + \log p - \log q \\ \frac{\partial^2 \log P(m|N = nI')}{\partial m^2} &= -\frac{1}{m} - \frac{1}{n-m} \end{aligned}$$

Temos que em $m^* = np$ a probabilidade máxima e nesse ponto a segunda derivada vale $= \frac{1}{npq}$. A expansão de Taylor até segunda ordem do $\log P(m|N = nI')$ nos leva a uma gaussiana para $P(m|N = nI')$ de forma óbvia (qualquer expansão até segunda ordem é quadrática). O que deve ser verificado é se essa expansão faz sentido. Esperamos, pela equação 2.32, que os termos superiores decaiam com n. Verifiquemos isso explicitamente para o termo cúbico (proporcional ao terceiro cumulante) e superiores:

$$\frac{\partial^3 \log P(m|N = nI')}{\partial m^3} = \frac{1}{m^2} - \frac{1}{(n-m)^2}$$

e notamos que as derivadas de ordem superior aumentarão o decaimento dos cumulantes.

Vemos que a distribuição binomial é bem aproximada por

$$P(m|N = nI) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(m - m^*)^2 \tag{2.50}$$

onde repetindo $m^* = np$ e $\sigma^2 = npq$. Repetimos que um dos pontos importantes é a dependencia $\sigma \propto n^{\frac{1}{2}}$.

2.8 Apêndice: Stirling

Considere a função Gama, definida para nossos propósitos imediatos sobre os números reais positivos:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt, \tag{2.51}$$

[8] Jeffreys, note quão boa é a approximação $1! = 0.9221$ sem o termo $1/12N$ e $1! = 1.002$ com esse termo. Para $2! = 1.9190$ e 2.0006 respectivamente.

que podemos estender por continuação analítica para o plano complexo ($x \neq$ inteiro menor que 1). Integre por partes e mostre que $\Gamma(x+1) = x\Gamma(x)$. Mostre que $\Gamma(1) = 1$ e portanto $\Gamma(n+1) = n!$. Mostre também para uso futuro que $\Gamma(1/2) = \sqrt{\pi}$ fazendo a mudança de variável $t = u^2/2$.

Podemos escrever

$$\Gamma(x+1) = \int_0^\infty e^{-t} t^x dt = \int_0^\infty e^{f(t;x)} dt \tag{2.52}$$

onde $f(t;x) = x \log t - t$. Para x fixo, o valor máximo de f ocorre em $t_{max} = x$. A derivada segunda com respeito a t, nesse ponto é $f''(t;x) = -1/x$ e as derivadas superiores, de ordem k caem com x^{1-k}. Para valores de $t >> x$ o integrando morre rapidamente e para valores grandes de x, f é bem aproximado por

$$f(t;x) = x \log x - x - \frac{1}{2x}(x-t)^2. \tag{2.53}$$

Supondo que não é uma aproximação muito ruim desprezar as derivadas superiores,

$$\begin{aligned} \Gamma(x+1) &\approx e^{x \log x - x} \int_{-x}^\infty e^{-\frac{1}{2x}(x-t)^2} dt \\ &\approx e^{x \log x - x} \int_{-\infty}^\infty e^{-\frac{1}{2x}(x-t)^2} dt \\ &= \sqrt{2\pi x} e^{x \log x - x} \end{aligned} \tag{2.54}$$

onde estendemos o limite de integração até $-\infty$ porque devido ao decaimento gaussiano não há contribuições importantes nessa região. Assim temos

$$\log n! = \log \Gamma(n+1) = (n + \frac{1}{2}) \log n - n + \frac{1}{2} \log 2\pi, \tag{2.55}$$

mostrando de forma intuitiva os primeiros termos da equação 2.48. Também é intuitivamente razoável que a aproximação deva ser melhor para valores maiores de x, dado que as derivadas superiores desprezadas são menores: $f'''(t_{max})/f''(t_{max}) = 2/t_{max} = 2/x$. Estas considerações podem ser provadas, assim como as correções da equação 2.9, mas o leitor deverá procurar outros textos (e.g. [**?**])

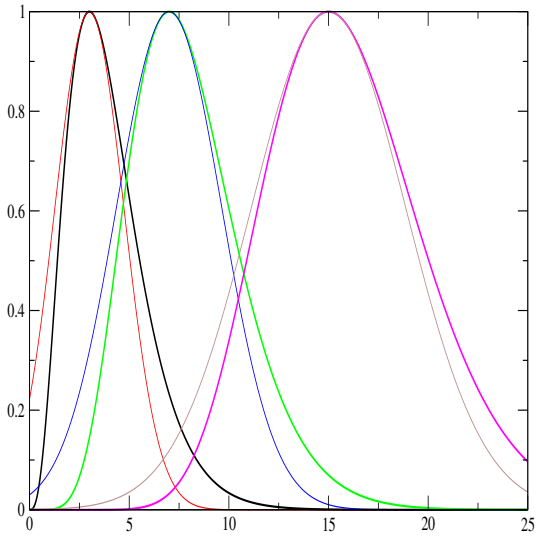

Figura 2.9: A figura mostra três pares de gráficos, para $x = 3, 7$ e 15 respectivamente. Cada par é formado pelas funções $e^{-t}t^x/(x^x e^{-x})$ e $\exp(-(x-t)^2/(2x))$

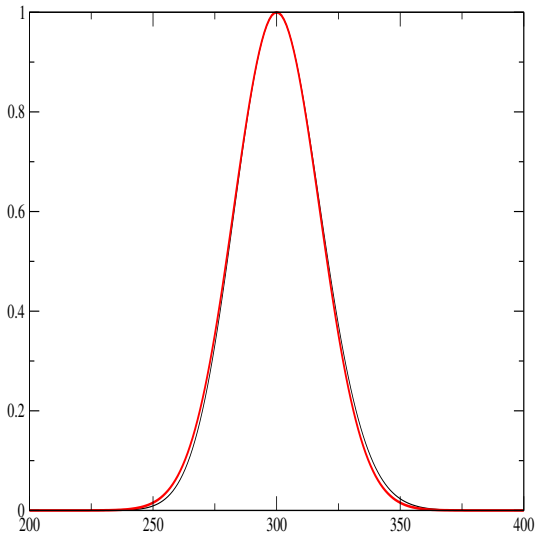

Figura 2.10: O mesmo que a figura 2.9 para x=300. Note aqui, assim como na figura anterior, que ao usar a gaussiana que é simétrica em relação ao máximo, o erro na integral cometido à esquerda do máximo tem sinal oposto ao cometido à direita.