NESTOR CATICHA

PROBABILIDADES (NOTAS INCOMPLETAS)

Prólogo

"...and probability, from being a despised and generally avoided subject, becomes the most fundamental and general guiding principle of the whole of science."

Harold Jeffreys in Scientific Inference

O estudo da Teoria de Probabilidade não tem recebido a importância que merece nos currículos típicos de graduação em Física. Isto é surpeendente dada sua utilidade em quase qualquer campo da Física. Alguns resultados e técnicas são usados sem jutificativa além de que todos sabemos o quê é probabilidade. É um fato empírico que os alunos, em geral, não sabem os rudimentos. Os prérequisitos matemáticos não são muitos. Cálculo Diferencial e Integral serão a lingua franca. No nível do curso, Integração significa integral de Riemann. Não precisaremos ir além da idéia de integrar em $\mathbb{R}^n$. A simplicidade matemática não implica que os conceitos serão simples. A interpretação pode ser bastante sutil e é esse, de forma geral, o objetivo do curso: fazer o aluno pensar e talvez modificar suas idéias sobre o que significa probabilidade. Rigor matemático não substitui rigor intelectual. No entanto, nesta área rigor intelectual não pode ser atingido sem algum rigor matemático.

O principal objetivo do curso é que o aluno entre em contato com a idéia de probabilidade como expressão da informação disponível sobre uma possibilidade. Introdução à Teoria de Informação poderia ser um título deste livro, mas se não do ponto de vista de engenharia, de um ponto de vista bem mais geral que encontrará aplicações em uma variedade muito grande de áreas da ciência. Estas incluem questões fundamentais em Física, mas também é claro em Estatística e portanto em tratamento de dados empíricos. Outras áreas como Ciência de Computação, Ciência Cognitiva e Neurociência tem tido uma grande influência desta forma de pensar sobre informação.

A primeira parte discute a definição de probabilidades. Todos os estudantes já foram expostos a probabilidades, tanto na linguagem coloquial quanto no curso secundário. Começaremos de forma diferente de outros cursos. Uma forma de proceder consiste em primeiro expor os princípios matemáticos e a partir deles calcular as consequência em aplicações interessantes. Faremos de outra forma. Não sabemos qual é a estrutura matemática que deve ser usada em geral, mas talvez possamos investigar se há casos simples em que podemos concordar com pessoas razoáveis como proceder. Isso dará uma lista de desejos de requisitos que a teoria deve satisfazer. Todas as estruturas que não estejam de acordo com a lista de desejos serão eliminadas. O último candidato em pé será a estrutura desejada. O aluno será convidado a procurar falhas no raciocínio, procurar exceções. Deste tipo de exercício decorrerá a confiança na estrutura final. Em ciência não devemos ser a favor de uma teoria ou sua interpretação, a não ser pelos motivos

que decorrem do respeito gerado por ter resistido a todos os embates em que se tentou derrubá-la.

Há outras formas de introduzir probabilidades e aqui me refiro às idéias frequentistas. O leitor não deve esperar uma exposição neutra, onde todos tem mérito e direito a ser ouvidos. O século XX ficou para trás e somente poucos frequentistas restarão no futuro. Acredito que há um falso embate entre os chamados Bayesianos e os frequentistas. O que é importante é o uso de probabilidades e a grande separação entre estas escolas é sobre que tipo de variável pode ser tratado de forma probabilistica. Os Bayesianos incluem qualquer variável no conjunto daquelas assim tratadas. Os frequentistas fazem em geral uma distinção entre variáveis que podem ser descritas de forma probabilistica e as que por sua natureza ontologica tem um valor definido e portanto não probabilistica. Os Bayesianos retrucam dizendo que não é probabilidade desse valor, mas do que pode ser acreditado sobre essa quantidade ter valor num certo intervalo. O subjetivismo tem uma má reputação em certos meios, acredito (de forma subjetiva) que ésta repulsa está mal colocada. As análises derivadas do uso de probabilidade envolvem uma mistura de ingredientes subjetivos e objetivos. Isso depende da natureza da informação disponível. Os métodos de processamento de informação, no entanto, são objetivos, pois diferentes operadores, com as mesmas informações devem chegar às mesmas conclusões.

Estas notas apresentam aos estudantes de Física uma visão que tem se mostrado frutífera e tem conquistado cada vez mais adeptos. Aqueles que estão interessados em aplicações e análise de dados terão acesso aos métodos atuais. O uso de técnicas numéricas e do computador não podem ser deixadas de lado e mesmo que não seja o objetivo principal, um pouco desse universo será explorado. O nível do curso é introdutório e um dos seus aspectos mais bonitos, a parte formal de Probabilidades como uma parte da Matemática, um ramo da análise funcional e teoria de medida não será explorada.

Se a primeira parte introduz a estrutura matemática e a linguagem apropriada, a segunda parte tem como tema a Entropia, o método principal para traduzir informação em probabilidades. A Mecânica Estatística é a aplicação natural e considerável espaço é dedicado á sua apresentação. Isto é material para um segundo semestre. Diferentes exemplos de processamento de dados, em particular aprendizagem em Redes Neurais, também são tratados de um ponto de vista entópico. Uma terceira parte, ainda a ser escrita lidará com aplicações a sistemas biologicos e sociais. Cabe um pedido de desculpas e uma explicação. A estrutura da teoria e suas aplicações não se ajusta bem a um espaço unidimensional. Senão, seguindo o conselho dado à Alice, deveriamos começar pelo começo e de ai proceder. Alguma desordem na apresentação pode incomodar o leitor sistemático, mas ao que consulte algum que outro capítulo de aplicações essa desordem pode não ser óbvia e repetições, incômodas para o primeiro, permitirão ao segundo maior facilidade de navegação.

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. Há ou-

tras formas de pensar, por exemplo acreditar em algo porque isso me deixa mais feliz. Mas eu não saberia dar um curso sobre isso.Não faz sentido acreditar em algo que não seja respaldado por informação.

Ariel Caticha, com que tenho meio século de discussões cientícas, tem sido uma grande influência ao longo dos anos. Ele escreveu um texto que vai muito além destas notas sobre a Física da Informação. O. Kinouchi, R. Vicente, F. Alves, S. Rabbani, M.V. Baldo com quem tenho colaborado ao longo dos anos ajudaram a modelar como penso sobre estas questões. Os livros de Jaynes e Jeffreys foram fundamentais e são referências constantes ao longo de quase todas estas notas.

São Paulo, 5 de junho de 2020

Sumário

1 *Teoremas de Regraduação de Cox* 9

2 *Outras definições de probabilidade* 31

3 *Uso elementar de Probabilidades* 37

4 *Frequência e Probabilidade* 53

5 *A distribuição Normal* 77

6 *Aplicações da regra de Bayes* 99

7 *Teorema do Limite Central* 123

8 *Seleção de Modelos* 139

9 *Monte Carlo* 147

10 *Entropia* 167

1

Teoremas de Regraduação de Cox

Alea jacta est
Júlio César

Introdução: Determinismo Newtoniano ou aleatório?

Júlio César ao cruzar com seu exército o Rio Rubicom quebrou uma regra na República Romana. O exército devia ser mantido longe de Roma. Não havia volta. Ou conseguia o poder ou perdia tudo. Qual seria o desenlace da sua ação? Nem ele sabia e segundo Suetônio teria dito: *Alea jacta est.* A sorte está lançada. Saber estimar as consequências de uma ação é aconselhável para poder decidir que curso tomar. César talvez tenha procedido da seguinte forma. Primeiro fez uma lista das possibilidades à sua frente. Uma decisão é tomada e uma das possibilidades seguidas. Estas poderiam incluir: (Ação I) Continuar na Gália. (Ação II) Fazer uma aliança com Pompeu , (Ação III) Fugir de Pompeu, (Ação IV) Se aposentar, (Ação V) Voltar a Roma com seu exército e lutar contra Pompeu. Historiadores certamente poderiam incluir outras. Como decidir? Supomos que uma escolha foi feita. Quais as consequências? Para cada curso de ação ele deve ter feito uma lista de possibilidades. Suponha que considere tomada a Ação V. Então as consequências poderiam ser (Consequência 1 da Ação V) Vitória total, com a formação do Império e ele como Imperador. (Consequência 2 da Ação V) Derrota total levando à sua morte. (Consequência 3 da Ação V) Guerra Civíl interminável ...etc. Mas não devia acreditar que cada uma das possiblidades teria a mesma chance de ocorrer. A cada consequência de cada Ação, César poderia ter associado um valor numérico indicando sua crença na chance de ocorrer. Veremos que isto será codificado em probabilidades de ocorrência. Mas também poderia ter associado um valor numérico de quão feliz ele seria se efetivamente essa consequência ocorresse. Estes números descrevem o que se chama de utilidade, de cada possibilidade, para o agente Júlio Cesar. Parece óbvio que as utilidades dependem do agente, mas talvez não seja óbvio que as probabilidades também dependam do agente, ou melhor do que este sabe. Resumindo, Júlio César decidiu o seu curso de ação após identificar as possibilidades de ação, das consequências de cada ação, das chances de cada consequência ocorrer, e da utilidade ou felicidade que cada consequência teria. Neste curso não falaremos sobre decisão a partir das utilidades. Atualmente, em geral, este tópico não cabe em cursos de Física. Faremos um estudo sistemático sobre a chances de algo ocorrer sem importar quão feliz você fique com cada possibilidade. O ponto central será definir com

cuidado o que queremos dizer com *chances*, como atribuir números e como mudá-los quando recebemos informação.

Teria Júlio César dúvidas sobre sua sorte ou saberia mais que os outros atores do drama? Se soubesse mais talvez estaria jogando um jogo de cartas marcadas enquanto os outros jogariam a cegas. A frase também implica num certo determinismo. Não há nada a fazer. O curso natural das coisas conduzirá os atores. Como observadores verão simplesmente o desenrolar da história.

Há alguma inconsistência em pensar que a as consequências são inevitáveis por um lado, e por outro ficar torcendo para ter sorte? Seria como torcer ao ver a gravação de um jogo de futebol que já foi jogado, mas não sabemos o resultado. Talvez seja um exercício interessante ver grandes jogos do passado sem saber qual jogo é, torcendo para seu time ganhar com direito a ficar tão feliz como quando o jogo é assistido ao vivo.

Todas estas situações são complexas. Comecemos por algo mais simples. Uma das maiores revoluções intelectuais da história da humanidade foi a introdução da Mecânica por Newton. Sabemos que caso fosse necessário temos o formalismo da Mecânica para poder calcular a trajetória de uma moeda. O determinismo Newtoniano permite fazer predições sobre o futuro a partir do estado atual. Por outro lado, os casos mais associados à sorte são o jogo de dados ou um jogo de cara ou coroa com uma moeda. Não é por acaso que a frase de César que teria sido dita em grego menciona $\kappa\upsilon\beta o\sigma$, o cubo ou dado. Estes jogos deram origem a o estudo matemático das probabilidades.

Como podemos associar a uma moeda simultaneamente as propriedades de ser um sistema determinista, governado pelas leis de Newton e a condição de exemplo mais usado ao falar de sistemas aleatórios? É necessário ter cuidado com as palavras. O que significa *aleatório*? Teremos todo este curso para atribuir-lhe significado. Em geral, ao ser usado coloquialmente, significa que não é totalmente determinado *a priori* por eventos passados.

As possibilidades do estado da moeda são determinados ao especificar 12 números. 3 dizem respeito à sua posição, por exemplo do centro de massa. Sua orientação é determinada por 3 ângulos.Veja, num livro de Mecânica a definição de ângulos de Euler. Ou senão, simplesmente considere 2 eixos no plano da moeda e um terceiro perpendicular ao plano e as rotações em torno deles. Esse número é duplicado ao levar em conta as suas derivadas temporais (velocidades). A dinâmica em 12 dimensões é dada pelas equações de Newton [1]. É óbvio que as equações não são suficientes para determinar como cairá a moeda. Há muitas maneiras de jogar a moeda, mas só um conjunto de equações. As mesmas equações devem ser complementadas com diferentes conjuntos de condições iniciais que parametrizam cada trajetória possível.

As figuras 1.1 e 1.2 mostram porque não há incompatibilidade nessas duas caracterizações da moeda, como paradigma do aleatório e sistema dinâmico Newtoniano. Por simplicidade fixamos 10 parâmetros e olhamos o que ocorre quando dois parâmetros são mudados numa certa região. As figuras foram construidas de forma totalmente determinística integrando as equações diferenciais. Cada ponto é colorido de acordo com a face mostrada pela moeda. Azul se cara, branco se coroa. Vemos que a aleatoriedade não está na evolução dinâmica descrita por Newton, mas na ignôrancia que poderiamos ter sobre a condições iniciais. Se ao jogarmos a moeda não tivermos conhecimento muito preciso das condições iniciais, não teremos como prever

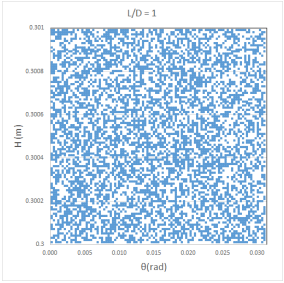


Figura 1.1: Integração numérica das equações de movimento de um modelo Newtoniano de uma "moeda"feita de massas (m) e molas (k). A figura mostra um espaço restrito de condições iniciais. H é a altura da moeda ao ser lançada e θ o ângulo com a horizontal, a moeda é solta do repouso. Nesta figura a altura é "grande"(em relação a mg/k). A estrutura é formada por quatro massas nos vértices do que seria em repouso um retângulo, ligadas por seis molas nas arestas e diagonais . O sistema está restrito a duas dimensões e a cada batida mesa há dissipação de energia. É um modelo de uma moeda ou um cubo simplificado. As simulações foram feitas por Guilherme Galanti e Osame Kinouchi, que gentilmente autorizaram o uso destas figuras.

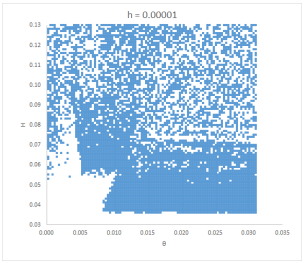


Figura 1.2: Igual à anterior, mas a moeda é solta de uma altura menor, para diferentes ângulos.

[1] Nem a dedução destas equações e muito menos a sua solução, serão necessárias aqui, mas cabem num curso de Mecânica.

se o ponto final será azul ou branco. Este é um indício que o conhecimento pode influenciar as probabilidades (que ainda não sabemos o que são) de que caia cara ou cora. Dois agentes apostando neste jogo terão chances diferentes de ganhar se tiverem informações diferentes sobre o modo como a moeda será jogada. Note que para alturas muito pequenas, o poder de predição fica mais forte, pois há regiões grandes com a mesma cor. Faça a experiência. Segure uma moeda com os dedos na posição horizontal. Solte a moeda, sem girá-la, de uma altura de 1 metro, 10 cm, 1 cm , 1mm. Seu poder de prever o que vai ocorrer aumenta. O determinismo é igualmente descrito pelas equações de Newton em todas as condições. A incerteza na previsão tem a ver com a forma como se solta a moeda; na falta de informação sobre o valor exato das condições iniciais.

Ainda isto é coloquial e não sabemos o que é probabilidade, informação ou aleatório. O objetivo do que segue é vestir isto com uma estrutura matemática. A história do desenvolvimento das ideias é complexo e não é o interesse destas notas. Porém elas estarão salpicadas de referências a grandes figuras do passado. A história contada, não é certamente como ocorreu, pois isso não sabemos. A seguir discutiremos as idéias que vem de Jakob Bernoulli, Laplace, Maxwell, Kolmogorov, Ramsey, Keynes, Pólya, Cox, Jeffreys, Jaynes entre outros. Começaremos a história no meio contando como R. Cox tentou criar uma extensão da lógica Booleana, com origens na Grécia antiga, para situações de informação incompleta. Ele poderia ter suposto de início que a estrutura matemática era a de probabilidades, mas se recusou a isso. Tentou encontrar essa estrutura e ao descobrir que era ou a teoria de probabilidade ou uma regraduação monotónica trivial, primeiro se convenceu da impossibilidade de escapar dessa estrutura e segundo forneceu uma sólida interpretação para o que queremos dizer com informação e como molda nossas crenças e para o que queremos dizer com probabilidades.

Há vários exemplos de tentativas de axiomatizar extensões da lógica a situações de informação incompleta. Savage e Lindley são exemplos importantes, mas seu objetivo é descrever o processo de tomada de decisão e isso leva a considerar utilidades. O caminho que escolhemos leva à mesma estrutura de probabilidades deixando claro que decisões é um capítulo a parte. O objetivo de um físico é descrever a natureza fazendo previsões e não tomando decisões.

Informação completa ou incompleta

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 com 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.

Pode parecer estranho para o estudante de 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. Mas a matemática é um tipo de linguagem que tem a vantagem de permir a quem a usa ser muito cuidadoso com o que diz. Denotaremos asserções por letras $A, B, C...a, b, c...$ 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[2]. Queremos analisar seu conteúdo informacional, se realmente a podemos crer 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.

[2] Embora o formalismo a ser introduzido também possa ser usado nesta direção, mas não agora.

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 ser verdadeira ou não, de forma independente do resto do universo conceitual. Ela será julgada verdadeira ou não quando analisada dentro de um contexto. A informação trazida por uma asserção C, será usada para atribuir um grau de verdade à asserção A, ou seja dentro do contexto C. Poderiamos chamar esse grau de, por exemplo, probabilidade de que A seja verdade se C for dada. Mas fazendo isto estariamos definindo de antemão que a ferramenta matemática apropriada para descrever informação é a teoria de probabilidades. Isto parece bem razoável mas não escapa às críticas acima e permite que outra ferramenta matemática seja usada por simplesmente expressar o gosto de outras pessoas ou a facilidade de uso em determinados problemas práticos com a mesma justificativa: *parece razoável, eu gosto, funciona, é prático.* Não descartamos o uso de outras ferramentas matemáticas, mas queremos deixar claro que estas poderão ser vistas como aproximações mais ou menos adequadas de uma estrutura que unifica e tem um posição diferente. O **objetivo** deste capítulo é mostrar que a escolha da teoria de probabilidades como a ferramenta matemática adequada para tratar informação é muito mais do que simplesmente conveniente. A teoria de probabilidades segue porque é a extensão da lógica a situações de informação incompleta. Mas até aqui não sabemos o que é *lógica, informação* nem *incompleta.*

A análise da lógica remonta a Aristóteles e passa por Boole no século XIX, que contribuiu para que a lógica pudesse ser representada em linguagem matemática [3]. Uma lógica envolve (i) um conjunto de proposições supostas verdadeiras, (ii) um método de dedução para estabelecer a validade de argumentos e (iii) um método para estabelecer invalidades.

[3] Veja para uma comparação: Aristotle's Prior Analytics and Boole's Laws of Thought, John Corcoran, History and Philosophy of Logics 2003.

Um argumento lógico é composto por duas partes. Um conjunto de asserções, chamadas as premissas e uma única asserção chamada de conclusão. Um argumento é válido se a conclusão pode ser obtida aplicando as regras (ii) e (iii).

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 como verdade, 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.

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

Outro caso de informação completa, novamente no contexto C, ocorre quando é dado como verdade a negação $\overline{B}$ ou seja "B é Falso". Segue $\overline{A}$, isto é, que "A é Falso" como conclusão inescapável. Note que se A não fosse falso, B não poderia sé-lo.

Nas condições que $C =$ "$A \rightarrow B$" e "A é Falso", o quê pode ser concluido? Do ponto de vista lógico clássico nada podemos concluir sobre B. Da mesma forma se for dada a informação "B é Verdade", nada podemos concluir sobre A. Estamos frente a casos de informação incompleta e a lógica clássica não serve para chegar a uma conclusão. Não é possível deduzir nada. A indução, o que quer que isto seja, e que será discutido mais à frente, será necessária para avançar. [4]

A forma dedutiva da lógica permite somente tres tipos de respostas, *sim, não* e *não segue*[5]. A indução nos força ou permite dividir esta última em várias possibilidades e os casos extremos nesse espectro são aqueles onde havendo certeza absoluta, haverá portanto a força da dedução. Podemos falar então sobre quais das alternativas intermediárias é mais razoável acreditar com base no que sabemos. Nota-se então a necessidade de estender a lógica para poder tratar de forma racional casos de informação incompleta. Seguindo uma tradição de desenvolvimento teórico onde talvez os antecessores imediatos sejam Ramsey e Keynes, Richard T. Cox, ao se defrontar com este problema por volta da década de 1940, decidiu, como dito acima, estabelecer um conjunto de desejos ou *desiderata*[6] que a teoria deveria satisfazer, e estes serão então os axiomas da extensão da lógica. Aqui podemos discordar, propor outros desejos ou axiomas, mas uma vez aceitos serão provados os teoremas de reparametrização de Cox que mostram que a teoria de probabilidade é a ferramenta para o tratamento de forma racional de situações de informação incompleta. O surpreendente disto é que surge a teoria das probabilidades como a forma para lidar de forma *racional* com a informação e que corremos riscos de ser inconsistentes caso a regras de manipulação de probabilidades não sejam seguidas. Ao leitor que demande uma definição de racional, podemos dizer que pelo menos não queremos ser manifestamente irracionais. Não acredito que haja uma definição de consenso sobre o que é ser racional. Há consenso porém em apontar alguns casos de irracionalidade. Segue que **não há probabilidades que não sejam condicionais**, embora às vezes simplesmente a linguagem esqueça de deixar explícitas as relações de condicionalidade. A maior fonte de erros será devida a falhas na especificação cuidadosa das asserções condicionantes. Aparentemente a notação $a|b$ com a a a asserção a ser analisada e b a asserção condicionante é devida a John Maynard Keynes, no seu Tratado.

A amplidão da aplicabilidade da teoria que emerge é impressionante

[4] 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)

[5] Nem o leitor nem o autor destas notas deve neste momento ceder à tentação de discutir lógicas de um ponto de vista mais geral. Precisamos um subconjunto de Lógica proposicional, não muito mais que lógica Aristotélica, como exposta por George Boole. Talvez caiba aqui a desculpa "I have not worked out the mathematical logic of this in detail, because this would, I think, be rather like working out to seven places of decimals a result only valid to two. My logic cannot be regarded as giving more than the sort of way it might work". Frank P. Ramsey (1926) "Truth and Probability", in Ramsey, 1931, The Foundations of Mathematics and other Logical Essays, Ch. VII, p.156-198, editado por R.B. Braithwaite, 1999 electronic edition.

[6] *Desiderata*: as coisas desejadas, em Latim. Termo usado em filosofia para denotar um conjunto de propriedades essenciais de alguma estrutura. Alguns ficam tentados a chamar *axiomas*.

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. Também 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.

Desiderata à la 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 Bernoulli, a Frank P. Ramsey e George Pólya. A exposição de Jaynes teve uma grande influência, mas ainda recebeu críticas e complementos [7]. Eu seguirei a apresentação de A. Caticha, que é mais completa e clara, mas farei algumas pequenas mudanças[8].

[7] Tribus, A. Caticha

[8] Notem que há lugar ainda para avanços nestes primeiros passos. Tentem encontrar defeitos, generalizações, melhores argumentos.

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 [9]. Isto é o que mostraremos a seguir.

[9] Este comentário parece trivial, mas o uso que será dado a seguir é totalmente não trivial. Neste contexto de probabilidades o destaque a este procedimento apareceu por primeira vez no livro de A. Caticha que o chamou de indução eliminativa e o atribuiu a J. Skilling, que tendo usado-o de forma não explicita no seus trabalhos sobre entropia, se declarou surpreso com a atribuição . Usaremos novamente este estilo de fazer teoria ao introduzir o conceito de entropia.

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 de 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 . A Desiderata que segue tem cinco desejos denotados $DP_1 ... DP_5$ e é um bom exercício tentar mostrar que não fazem sentido. Se você conseguir e convencer outros terá feito uma grande contribuição. Se não conseguir, terá mais respeito pelas conclusões que seguem.

DP_1 *Representação de crenças e transitividade*

Queremos analisar o primeiro caso simples que lida com o conceito de *mais plausível*. Suponha que A dado B é mais plausível do que A dado C, então escrevemos $A|B \succ A|C$. Suponha 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 a transitividade:

- DP_1: Se $A|B \succ A|C$ e $A|C \succ A|D$ então deve ser o caso que $A|B \succ A|D$

Além disso, dadas duas crenças podemos imaginar que há outra asserção intermediária.

Isto é fácil de satisfazer se impusermos:

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

Podemos satisfazer este tipo de ordenamento representado crenças com números racionais. A escolha de números reais permite usar integrais, o que não é pouco, pois fazer somas é difícil. Note que sempre usamos integrais em fisica, mesmo que o espaço tenha uma estrutura subjacente (desconhecida atualmente mas que poderia ser na escala de e.g 10^{-31} m). Não sabemos se tem, mas nos modelos para o *mundo* usados em Mecânica, os pontos do espaço e tempo vivem numa variedade 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 desejo DP_1. 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 condicionante passa de B para C. Também não diz que a representação das crenças seja única. Uma mudança dos números estritamente monotonica crescente não mudará a ordem. Isto levará a que há familias de atribuições de números que representam a ordem da mesma forma.

DP_2 Asserções compostas:

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, há diferentes notações equivalentes possíveis: AB, $A \wedge B$ ou ainda A **e** B.
- A **soma** ou disjunção de duas asserções é uma terceira asserção, que constuma ser denotada por $A + B$ ou $A \vee B$, ou ainda A **ou** B.

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

Tabela 1.1
Tabela verdade para a negação e algumas asserções compostas.

Isso significa que $A+B=\overline{\overline{A}\,\overline{B}}$ portanto o conjunto de operações negação e conjunção permite construir a disjunção de asserções.

Ao falar de silogísmos introduzimos a operação $\Rightarrow$ que significa implicação. Se é verdade que $A \Rightarrow B$, significa que se A é verdade segue B. Isto não é um novo operador pois é equivalente dizer que C é verdade para $C=\overline{A\wedge\overline{B}}$.

A	B	$C=\overline{A\wedge\overline{B}}$
V	V	V
V	F	F
F	V	V
F	F	V

Tabela 1.2
Tabela verdade para a implicação C.

Suponha que haja um método, usando a teoria geral que procuramos e ainda não temos, de analisar a plausibilidade de uma asserção composta por várias asserções através de conjunções ou disjunções ou negações. 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:

- DP_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 DP_2 acima declara.

DP_3 *Informação completa*

Um tratamento geral de situações de informação incompleta deve abarcar os casos particulares de informação completa. Então olhemos para casos simples em que há informação completa.

O mais simples é $a|a$ que é a plausibilidade de algo que sabemos ser verdade, para qualquer a.

Se $a|bc$ e $b|ac$ representam a plausibilidade de algo que sabemos ser falso com certeza, chamamos a e b de mutuamente exclusivos na condição c. Poderia ser que hajam falsidades absolutas mais falsas que outras falsidades absolutas; ou verdades absolutas mais verdadeiras que outras verdades absolutas, mas achamos razoável impor

- DP_3 : Existem dois números v_v e v_f tal que 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. Este desejo inclui também a negação de uma asserção, pois a asserção e sua negação são mutuamente exclusivos, e estamos dizendo que $\overline{a}|a = v_f$ para qualquer a.

Usaremos frequentemente a propriedade de um conjunto de asserções $\{a_i\}_{i=1....K}$ serem mutuamente exclusivos sob condições c, que vale se para qualquer i, j diferentes $a_i|a_jc = a_j|a_ic = v_f$

DP_4 *Soma e* DP_5 *Produto*

Como sugerido na tabela 1, todo operador na álgebra Booleana pode ser representado pelas operações conjunção a *e* b (denotada ab ou $a \wedge b$) e negação de a (denotada por $\overline{a}$) [10], isto é, o produto e a negação lógicas. A soma lógica pode ser obtida usando $a \vee b = \overline{\overline{a}\,\overline{b}}$. 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, conjunção e negação. Mas é mais fácil, olhar para a conjunção e a disjunção, e junto com DP_3 obteremos a forma geral de tratar a negação.

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

Agora olhamos para a disjunção ou 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:

$$x_1 = a|c,\ x_2 = b|c,\ x_3 = b|ac,\ x_4 = a|bc. \tag{1.1}$$

É importante notar que todas estas plausibilidades são condicionadas a c, a informação que por hipótese é suposta verdadeira. Além disso podem ser condicionadas a outras asserções relevantes e as únicas disponíveis são a e b por separado. Não tem sentido considerar ab como parte do condicionante. Deve haver uma dependência entre $a \vee b|c$ e algum subconjunto de $\{x_i\} = \{x_1, x_2, x_3, x_4\}$, então

- DP_4: Regra da Soma: Deve existir uma função F que relaciona $a \vee b|c$ e algum subconjunto de $\{x_i\}$ e não deve tomar um valor constante, independente dos valores de $\{x_i\}$.

É claro que trocando soma por produto parece razoável desejar:

- DP_5: Regra do Produto. Deve existir uma função G que relaciona $ab|c$ e algum subconjunto de $\{x_i\}$ e não deve tomar um valor constante, independente dos valores de $\{x_i\}$.

Como F e G representam a plausibilidade de asserções (compostas), também devem tomar valores reais. Além disso não impomos nada além de que dependam em algumas, se não todas, as variáveis $\{x_i\}$. Parece natural exigir que não tenham valores constantes, pois senão a todas as asserções compostas lhes seria atribuido o mesmo número.

Para facilitar as deduções também imporemos diferenciabilidade até segunda ordem com repeito a quaisquer dois argumentos. Isto não é necessário, mas as provas ficam mais longas e no fim o resultado vem na forma de funções diferenciáveis.

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). Analisaremos casos particulares em que é fácil ver que alguns subconjuntos levam a resultados absurdos. Do ponto de vista axiomático poderiamos adicionar estes casos particulares à lista de desejos.

Consequências da Lista de Desejos

Parece difícil que desta lista DP_1...DP_5 surja uma estrutura matemática, quanto mais única. Ou como veremos, essencialmente única a menos de regraduações montônicas que não alteram a ordem das crenças. Talvez o que será surpreendente para o leitor, é que seja a teoria de probabilidades. A estrutura matemática aparecerá analisando as restrições nas funções F e G impostas pelos desejos.

A regra da soma

Começamos com a disjunção $a \vee b|c$ e a função F. Primeiro consideramos a e b mutuamente exclusivos, mas depois veremos que isto permitirá analisar o caso geral. Sob esta restrição $a|bc = b|ac = v_f$ para qualquer c por DP_3. Logo

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

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

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

Para avançar olhamos para asserções compostas mais complexas, que podem ser analisadas de mais de uma maneira, que pelo desejo DP_2, devem dar o mesmo resultado. Para três asserções a, b e c mutuamente excludentes nas condições d, duas maneiras equivalentes de escrever a disjunção das três são $(a \vee b) \vee c|d = a \vee (b \vee c)|d$ o que permite usar a função f duas vezes

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

ou em notação óbvia, f satisfaz

$$f(x, f(y, z)) = f(f(x, y), z) \tag{1.2}$$

chamada equação da associatividade, primeiramente estudada por Abel no contexto de teoria de grupos. Pode se provar [11] que para toda solução de 1.2, existe um bijeção ϕ, dos reais nos reais, que tomaremos como crescente, e portanto será estritamente monotonica crescente, tal que

$$f(x, y) = \phi^{-1}(\phi(x) + \phi(y)). \tag{1.3}$$

Para o leitor bastará mostrar neste ponto, que a expressão 1.3 é uma solução da equação 1.2.

Agora um ponto central: podemos *regraduar*, usando ϕ, as atribuições de plausibilidade e não mais falar dos números do tipo $a|d$ mas de

[11] Para condições em f ver Aequationes mathematicae 1989, Volume 37, Issue 2-3, pp 306-312 *The associativity equation revisited* R. Craigen, Z. Páles, ou o livro Aczél, J. (1966), *Lectures on functional equations and their applications, Mathematics in Science and Engineering* 19, New York: Academic Press,

números $\phi(a|d)$. Por ser uma bijeção, resulta que a ordem de preferências não se altera, se antes as crenças sobre as asserções tinham uma certa ordem, depois da regraduação, o ordenamento da representação numérica das crenças é o mesmo. É importante ver que a função ϕ é estritamente monotonica: se $x > y$ segue que $\phi(x) > \phi(y)$, sem poder haver igualdade. Isto significa que asserções com crenças diferentes são mapeadas em valores ϕ diferentes. Caso ocorresse a possibilidade de igualdade, antes da regraduação teriamos uma separação de preferências e depois da regraduação poderiamos ter confusão entre asserções mapeadas no mesmo valor de ϕ. [12] Continuamos sem saber que números são esses, mas avançamos a ponto de poder dizer que para quaisquer eventos mutuamente exclusivos a crença da disjunção, uma asserção composta pode ser expressa em termos das crenças nas asserções mais simples:

[12] Veja A. Patriota onde as condições sobre f são relaxadas e as consequências de aceitar soluções não estritamente monotonicas são consideradas.

$$\phi(a \vee b|d) = \phi(a|d) + \phi(b|d). \tag{1.4}$$

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

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

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

$$a = (a \wedge b) \vee (a \wedge \overline{b}) \text{ e } b = (b \wedge a) \vee (b \wedge \overline{a}). \tag{1.8}$$

O leitor deve mostrar que as relações acima são verdadeiras, no estilo da tabela 1. Podemos escrever $a \vee b$ como uma disjunção de asserções mutuamente exclusivas:

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

assim a equação 1.4, que descreve a soma de asserções mutuamente exclusivas, 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. Chamamos pela ordem os termos do lado direito da equação acima de 1, 2,...5. Usando novamente a equação 1.4 para asserções mutuamente exclusivas, juntando 1 com 2 , e 3 com 4:

$$\begin{aligned} \phi(a \vee b|d) &= \phi\left((a \wedge b) \vee (a \wedge \overline{b})|d\right) + \phi\left((b \wedge \overline{a}) \vee (a \wedge b)|d\right) - \phi(a \wedge b|d) \\ &= \phi(a|d) + \phi(b|d) - \phi(a \wedge b|d), \end{aligned} \tag{1.9}$$

que segue das relações da equação 1.8. Temos um dos resultados principais para lidar com asserções compostas por somas de asserções

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

Mas ainda não acabamos pois não sabemos o que fazer com $\phi(ab|d)$, que olharemos a seguir.

Regra do produto: quais as variáveis relevantes?

Queremos expressar

$$y = \phi(ab|c) \tag{1.10}$$

em termos da função ainda por determinar G e de algum dos subconjuntos de $\{x_i\}$. Lembramos a notação $x_1 = a|c$, $x_2 = b|c$, $x_3 = b|ac$, $x_4 = a|bc$. Tribus sugeriu a análise das 11 possibilidades para verificar que só há duas que sobrevivem a casos extremos. Seguimos A. Caticha, pois corrige vários erros anteriores. Os dois conjuntos sobreviventes 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. Cox já parte da conclusão de que estes dois subconjuntos são os adequados. O exercício que segue mostra que ele tinha razão, mas retira a arbitrariedade aparente, de fazer a escolha sem analisar outros candidatos.

Vejamos como chegar a esta conclusão (novamente seguimos AC) Os 11 casos são reduzidos a 7 por simetria:

1. $y = G(\phi(a|I), \phi(b|I))$ (1 possibilidade)
2. $y = G(\phi(a|I), \phi(a|bI))$ (2 possibilidades $a \leftrightarrow b$)
3. $y = G(\phi(a|I), \phi(b|aI))$ (2 possibilidades $a \leftrightarrow b$)
4. $y = G(\phi(a|bI), \phi(b|aI))$ (1 possibilidade)
5. $y = G(\phi(a|I), \phi(b|I), \phi(a|bI))$ (2 possibilidades $a \leftrightarrow b$)
6. $y = G(\phi(a|I), \phi(a|bI), \phi(b|aI))$ (2 possibilidades $a \leftrightarrow b$)
7. $y = G(\phi(a|I), \phi(b|I), \phi(a|bI), \phi(b|aI))$ (1 possibilidade)

Caso 1 Mostraremos que $y = a \wedge b|I = G(\phi(a|I), \phi(b|c)) = G(x_1, x_2)$ não funciona pois nao satisfaz o esperado em um caso simples. Porque não serve o subconjunto mais óbvio (x_1, x_2)? Primeiro vejamos que não segue o bom senso. 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 asserções, mas não que seja muito plausível que use um tenis de cada cor ao mesmo tempo. 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 e portanto teria um valor constante, independende das asserções sob consideração. Insistindo, suponha que Bruno joga uma moeda contra o teto, bate no ventilador e cai. A Helena pega outra moeda e faz o mesmo. Temos a mesma crença que saia cara ou coroa nas duas situações. Chamamos c_B a asserção que saiu cara no primeiro experimento e c_H no segundo. Achamos razoável escrever

$$\phi(c_B|I) = \phi(c_H|I) \quad \text{e} \quad \phi(\overline{c}_B|I) = \phi(\overline{c}_H|I)$$

E também achamos impossível que $c_B\overline{c}_B|I$ seja verdade, não pode ser verdade que Bruno obteve cara e coroa nessa única jogada. Mas seriamos levados a pensar que

$$\begin{aligned} \phi(c_B\overline{c}_H|I) &= G(\phi(c_B|I), \phi(\overline{c}_H|I)) \\ &= G(\phi(c_B|I), \phi(\overline{c}_B|I)) = \phi(c_B\overline{c}_B|I) = 0 \end{aligned} \tag{1.11}$$

que significaria que se Bruno obteve cara, Helena não poderia ter obtido coroa.

Caso 2

Para qualquer asserção $b|I$, sob quaisquer condições teriamos

$$\phi(b|I) = \phi(Ib|I) = G(\phi(I|I), \phi(I|bI)) = G(\phi_V, \phi_V) = \text{constante.}$$

Um método que atribui o mesmo valor numérico a todas as asserções não pode ser aceitável.

Caso 3 Para o caso $y = G(\phi(a|I), \phi(b|aI))$ e a alternativa $G(\phi(b|I), \phi(a|bI))$ 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. Não analisaremos as consequências disto agora. Ainda falta eliminar os outros candidatos e posteriormente encontrar a forma específica de G.

Caso 4 Se $y = G(\phi(a|bI), \phi(b|aI))$ somos levados a algo inaceitável considerando que para qualquer asserção b teriamos

$$\phi(b|I) = \phi(bb|I) = G(\phi(b|bI), \phi(b|bI)) = G(\phi_v, \phi_v) = \text{constante}$$

independente de b. Novamente a crença sobre a plausibilidade de uma asserção seria independente da asserção.

Caso 5 $y = G(\phi(a|I), \phi(b|I), \phi(a|bI))$. Este caso é mais complicado de analisar. Mostraremos, no entanto que se reduz a algum dos casos anteriores. Ainda consideraremos a conjunção de mais de duas asserções , $abc|I$, que pode ser escrito de duas formas diferentes $(ab)c|I = a(bc)|I$, portanto, considerando a primeira forma obtemos

$$\begin{aligned} \phi((ab)c|I) &= G(\phi(ab|I), \phi(c|I), \phi(ab|cI)) \\ &= G(G(\phi(a|I), \phi(b|I), \phi(a|bI)), \phi(c|I), G(\phi(a|cI), \phi(b|cI), \phi(a|bcI)) \\ &= G(G(x, y, z), u, G(v, w, s)). \end{aligned} \tag{1.12}$$

Para a segunda, com as mesmas definições das variáveis $x, y...$, obtemos

$$\begin{aligned} \phi(a(bc)|I) &= G(\phi(a|I), \phi(bc|I), \phi(a|bcI)) \\ &= G(\phi(a|I), G(\phi(b|I), \phi(c|I), \phi(b|cI), \phi(a|bcI)) \\ &= G(x, G(y, u, w), s) \end{aligned} \tag{1.13}$$

Notamos as duas maneiras de escrever a mesma coisa. Repetimos que por DP_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 $z = \phi(a|bI))$ nem de $v = \phi(a|cI)$ explicitamente. Para que essa expressão não dependa de z nem v , podemos impor que G não dependa do terceiro argumento o que levaria a eliminar o que foi riscado na equação abaixo:

$$G(G(x,y,\cancel{z}),u,\cancel{G(v,w,s)}) = G(x,G(y,u,\cancel{w}),\cancel{s})$$

levando a que G tem só dois argumentos e uma expressão sem z nem v:

$$G(G(x,y),u) = G(x,G(y,u))$$

e portanto somem todas as variáveis exceto x, y e u. Lembrando suas definições

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

que equivale ao **Caso 1** e portanto já foi eliminado.

Mas também podemos dizer que não depende do primeiro argumento, que também elimina z e v:

$$G(\cancel{G(x,y,z)},u,G(\cancel{v},w,s)) = G(\cancel{x},G(\cancel{y},u,w),s)$$

que leva à expressão

$$G(u,G(w,s)) = G(G(u,w),s)$$

que voltando às variáveis originais toma a forma

$$G(\phi(c|I),G(\phi(b|cI),\phi(a|bcI))) = G(G(\phi(c|I),\phi(b|cI)),\phi(a|bcI))$$

e mostra ser equivalente ao que teriamos obtido se partissemos do **Caso 3** e portanto aceitável.

Fica como exercício mostrar que

1. o **Caso 6** pode ser reduzido ao **Caso 2,** ao **Caso 3** ou ao **Caso 4**
2. o **Caso 7** pode ser reduzido aos **Caso 5** ou **Caso 6.**

Concluimos portanto que

$$\begin{array}{|lcl|}\hline \phi(ab|c) & = & G(\phi(a|c),\phi(b|ac)) \\ & = & G(\phi(b|c),\phi(a|bc)) \\ \hline\end{array}$$

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 1.3.2 refletem a única opção. Outra queixa e que introduzimos casos simples onde os casos diferentes do 3 se mostraram contrários ao bom senso. Isto significa que o DP_5 é mais complexo do que parecia inicialmente.

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, com a vantagem que pelo menos sabemos seus argumentos.

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:

$$a(b \vee c) = (ab) \vee (ac). \tag{1.14}$$

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

$$\begin{aligned} \phi(a(b \vee c)|d) &= G\left(\phi(a|d), \phi(b \vee c)|ad)\right) \\ &= G\left(\phi(a|d), \phi(b|ad) + \phi(c|ad)\right) \qquad (1.15) \\ \phi((ab) \vee (ac))|d) &= \phi(ab|d) + \phi(ac|d) \\ &= G\left(\phi(a|d), \phi(b|ad)\right) + G\left(\phi(a|d), \phi(c|ad)\right) \qquad (1.16) \end{aligned}$$

onde a equação 1.15 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 1.16 mostra o resultado de considerar a soma $(ab) \vee (ac)$. Mas devido à equação 1.14 e DP_2, estas duas formas devem dar o mesmo resultado:

$$G(x, y + z) = G(x, y) + G(x, z). \tag{1.17}$$

Para obter a solução geral desta equação notemos que o primeiro argumento é o mesmo nos três termos, é portanto um parâmetro que podemos manter fixo em qualquer valor arbitrário. Não é necessário supor diferenciabilidade, mas requerindo que G seja duas vezes diferencíavel, e definindo $w = y + z$ obtemos a equação diferencial

$$\frac{\partial^2 G(x, w)}{\partial w^2} = 0 \tag{1.18}$$

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 1.17 obtemos

$$A(x)(y + z) + B(x) = A(x)y + B(x) + A(x)z + B(x), \tag{1.19}$$

portanto $B(x) = 0$, ou seja $G(x, w) = A(x)w = G(x, 1)w$ [13]. Agora olhamos para $a|d$ e usamos $a|d = ad|d$ para a e d quaisquer.

$$\begin{aligned} \phi(a|d) &= \phi(ad|d) = G(\phi(a|d), \phi(d|ad)) \\ &= G(\phi(a|d), \phi_v) = A(\phi(a|d))\phi_v \end{aligned} \tag{1.20}$$

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

$$G(x, w) = \frac{xw}{\phi_v} \tag{1.21}$$

[13] Suponha a equação $h(x + y) = h(x) + h(y)$, para qualquer x, y. Em particular, para $n \neq 0$ e m inteiros, vale que $h(nx) = h((n-1)x + x) = h((n-1)x) + h(x) = h((n-2)x + x) + h(x) = h((n-2)x) + 2h(x) = \cdots = nh(x)$. Considere $x = x'/n$. Segue que $h(x') = nh(x'/n)$. Tome $x' = m$, portanto $h(x') = h(m) = mh(1) = nh(m/n)$. Logo $h(m/n) = (m/n)h(1)$. Basta supor continuidade que podemos passar dos racionais para os reais e obter $h(x) = xh(1)$.

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

$$\phi(ae|d) \quad = \quad \frac{\phi(a|d)\phi(e|ad)}{\phi_v}. \tag{1.22}$$

Mas resta um problema: e se retirarmos a restrição de b e c mutuamente exclusivos? É simples de considerar pois novamente usamos a equação 1.8

$$e = (e \wedge h) \vee (e \wedge \overline{h}), \tag{1.23}$$

agora para qualquer asserção h, de tal forma que $b = e \wedge h$ e $c = e \wedge \overline{h}$ [14]. Portanto não ha restrições para o resultado que obtivemos.

[14] Agradeço a ...e a ... por me lembrar deste truque.

Se não usarmos esse atalho deveriamos usar a equação 1.9 para obter:

$$G(x, y + z - G(y, w)) = G(x, y) + G(x, z) - G(x, G(x, w))$$

e sabemos que a solução é dada pela equação 1.21. Sem usar esse atalho é mais difícil mostrar que esta é a única forma se G for diferenciável duas vezes em cada argumento. O leitor interessado deverá consultar Áczel. Temos assim uma possibilidade de uma prova muito mais simples.

Da equação 1.22, dividindo por ϕ_v obtemos

$$\frac{\phi(ae|d)}{\phi_v} \quad = \quad \frac{\phi(a|d)}{\phi_v}\frac{\phi(e|ad)}{\phi_v} \tag{1.24}$$

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

$$p(a|b) = \frac{\phi(a|b)}{\phi_v} \tag{1.25}$$

que serão daqui a pouco chamados de probabilidade de a dado b. E a regra do produto em termos destes novos números regraduados é

$$\boxed{p(ab|c) = p(b|c)p(a|bc) = p(a|c)p(b|ac)}$$

Temos uma regra para o produto e para soma lógicas de asserções. Como fica a negação? Apesar de não ter introduzido nada específico sobre ela veremos que com os desejos impostos podemos deduzir a plausibilidade regraduada ou probabilidade da negação de uma asserção a partir da probabilidade de sua afirmação.

A regra do produto e a consistência permitem escrever

$$p(a|bc) = \frac{p(a|c)p(b|ac)}{p(b|c)} \tag{1.26}$$

que é chamado de Teorema de Bayes, mas que foi escrito pela prmeira vez por Laplace. A contribuição de Bayes foi apontar a relação chamada de inversão

$$p(a|bc) \propto p(b|ac) \tag{1.27}$$

onde a probabilidade de uma asserção a condicionada a outra b é proporcional à probabilidade de b condicionada a a. Não podemos exagerar a importância desta afirmação que ficara clara à luz da variedade de aplicações tanto teóricas quanto experimentais que veremos adiante.

Negação

A lista de desejos inclui a menção de algo sobre a negação. A crença em asserções condicionadas à sua negação constituem casos de informação completa: $\phi(a|\overline{a}) = p(a|\overline{a}) = 0$. Também sabemos que $a \vee \overline{a}$ deve ser verdade, pois não resta alternativa. Portanto

$$\begin{aligned} \phi(a|\overline{a}d) &= p(a|\overline{a}d) = 0 \\ p(a \vee \overline{a}|d) &= \frac{\phi(a \vee \overline{a}|d)}{\phi_v} = 1 \end{aligned} \tag{1.28}$$

$$\begin{aligned} 1 &= p(a \vee \overline{a}|d) \\ &= p(a|d) + p(\overline{a}|d) - p(a\overline{a}|d) \\ &= p(a|d) + p(\overline{a}|d) - p(\overline{a}|d)p(a|\overline{a}d) \\ &= p(a|d) + p(\overline{a}|d), \end{aligned} \tag{1.29}$$

$$\boxed{p(\overline{a}|d) = 1 - p(a|d)}$$

ou a soma das crenças regraduadas de uma asserção e da sua negação é 1. Essencialmente chegamos ao fim do começo.

Estrutura matemática sobrevivente

Em termos destes números, reescrevemos os resultados até aqui obtidos:

$p(a\|a)$	$=$	$p_v = 1$	Certeza da veracidade
$p(a\|\overline{a})$	$=$	$p_f = 0$	Certeza da falsidade
$p(a \vee b\|c)$	$=$	$p(a\|c) + p(b\|c) - p(ab\|c)$	regra da soma
$p(ab\|c)$	$=$	$p(a\|c)p(b\|ac)$	regra do produto
$p(ab\|c)$	$=$	$p(b\|c)p(a\|bc)$	regra do produto
$p(\overline{a}\|d)$	$=$	$1 - p(a\|d)$	regra da negação

Tabela 1.2
Probabilidades

Não falaremos mais em números $a|b$, nem na sua regraduação $\phi(a|b)$ mas somente na última transformação $p(a|b)$ que chamaremos a probabilidade de a dado b, ou a probabilidade de a condicionada à informação que b é verdadeira. O motivo disto é que ao longo de séculos estas regras from destiladas pelo bom senso de vários matemáticos e cientistas. Por volta de 1930, Kolmogorov formalizou, sem incluir a regra do produto nem condicionantes, usando linguagem de teoria de medida ou integração, mas já eram conhecidas desde Laplace. O que não estava claro é porque essas e não outras. Está 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 leitor, caso deseje usar

outras regras para manipular informação deverá responder quais dos desejos considerados acima não é razoável e portanto ao ser evitado, justificar essas outras regras.

O que foi obtido vai ser comparado com os axiomas de Kolmogorov na próxima secção. 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, razoável mas *ad hoc*, a probabilidade condicional definida a partir da regra do produto e portanto colocando com a mão o teorema de Bayes, que Cox obteve como uma consequência direta da consistência em particular e dos outros membros da desiderata.

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.

Exercícios

Mostre, construindo a tabela verdade as seguines propriedades da Algebra Booleana a partir da tabela verdade para a soma, produto e negação

- Idempotência do produto $AA = A$
- Idempotência da soma $A + A = A$
- Comutatividade do produto $AB = BA$
- Comutatividade da soma $A + B = B + A$
- Associatividade do produto $A(BC) = (AB)C$
- Associatividade da soma $(A + B) + C = A + (B + C)$
- Distributividade $A(B + C) = AB + AC$
- Dualidade $C = AB \Rightarrow \overline{C} = \overline{A} + \overline{B}$ e
  $C = A + B \Rightarrow \overline{C} = \overline{AB}$

Mostre que $(A + B)A = A$ e portanto $A + BC = (A + B)(A + C)$

Exercícios Propostos

- Mostre que a conjunção e a disjunção não formam um conjunto de operadores completo para a álgebra booleana. Por exemplo mostre que não há combinação de estes operadores que permitam obter a negação. Mas nos propusemos uma função F e uma G e obtivemos uma forma de lidar com a negação. Como isso é possível? A resposta será achada ao ver que que o desejo DP_3 sobre informação completa introduz a noção de negação mas só parcialmente ao dizer que a e sua negação são mutuamente exclusivos e que $a|\overline{a} = v_f$ como o mesmo v_f para todo a. Outra forma de proceder poderia ser introduzir um desejo do tipo: Deve existir uma função H, desconhecida tal que $a|c = H(\overline{a}|c)$. Isto codifica o desejo de encontrar uma teoria em que conhecimento sobre a implica conhecimento sobre $\overline{a}$. Claro que nesta altura sabemos que $H(x) = 1 - x$. Tente deduzir a as consequências ao trocar disjunção F por negação H no Desiderata para lidar com informação incompleta.

- Mostre a relação da equação 1.8. Desenhe o diagrama de Venn.
- A equação 1.9 relaciona a crença da disjunção às crenças nas asserções primitivas, mas inclui a subtração da crença na conjunção. Desenhe o diagrama de Venn adequado a esta situação. Discuta a origem do term subtraido.

- Voltemos ao **Caso 5** e suponhamos que G seja diferencíavel com respeito a qualquer argumento. As derivadas parciais com respeito a z ou v devem dar zero. Use a regra da cadeia para mostrar que

$$\begin{aligned} 0 &= \frac{\partial}{\partial z} G(G(x,y,z),u,G(v,w,s)) \\ &= \frac{\partial}{\partial r} G(r,u,G(v,w,s))_{r=G(x,y,z)} \frac{\partial}{\partial z} G(x,y,z) \end{aligned} \tag{1.30}$$

  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 mostre que voltamos ao **Caso 3** . Se não depende do terceiro mostre que voltamos ao **Caso 1.**

- **2** Para a função G da regra do produto 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.**

- Mostre que a forma produto (eq. 1.21) é solução da equação funcional. Mostre que esta é a única forma se G for diferenciável duas vezes em cada argumento.

- Escreva a regra do produto $P(AB|I)$, da soma $P(A+B|I)$ e da negação de $A|I$, de A no contexto I em termos das Chances, percentagem e Logprob definidos abaixo. Mostre que cada uma dessas é uma transformação monotónica ϕ das probabilidades e portanto uma regraduação possível da representação numérica das crenças.

1. Chances: Defina as chances (odds em inglés) como $O(A|I) = \frac{P(A|I)}{P(\overline{A}}$).
2. Percentagem é o que chamariamos a probabilidade se em lugar de estar confinada ao intervalo $[0,1]$ estivesse no intervalo$[0,100]$.

3. Logprob $L_P(A|I) = \log P(A|I)$.
4. Logit ou log-odds: $\text{Logit}(P(A|I)) = \log(\frac{P(A|I)}{P(\overline{A})})$.
5. Exprob $\text{Exp}_P(A|I) = \exp P(A|I)$ (Essa acabei de inventar).
6. Sineprob $\text{Sen}_P(A|I) = \sin \frac{P(A|I)\pi}{2}$ (Posso continuar.)

Em algum caso as regras escritas em termos das regraduações são mais simples do que a regraduação que leva às probabilidades? [15]

[15] Não é verdade que neste caso "What's in a name? that which we call a rose By any other name would smell as sweet"

Exercício Problema de Linda 1. Amos Tversky and Daniel Kahneman colocaram a questão a seguir, chamada de Problema de Linda, sobre probabilidades. Considere as asserções a seguir:

- I : Linda tem 31 anos, é solteira, assertiva, e muito inteligente. Ela se formou em filosofia. Quando estudante, estava profundamente preocupada com questões de discriminação e justiça social, e também participou de manifestações anti-nucleares.
- A : Linda é bancária .
- B : Linda é bancária e participa do movimento feminista .

Responda rapidamente qual das duas asserções é mais provável?

Exercício Problema de Linda 2. Não continue lendo até ter respondido à pergunta anterior.

Responda após pensar.O problema é atribuir números a $P(A|I)$ e $P(B|I)$. Qual é maior? Responda usando a regra do produto e use o fato que qualquer probabilidade tem uma cota superior 1. Este problema também é chamado de Falácia da conjunção. [16] Introduza a asserção

[16] These long-term studies have provided our finest insight into "natural reasoning" and its curious departure from logical truth. Stephen Jay Gould, sobre Tversky and Kahneman

- C : Linda é bancária e não participa do movimento feminista .

Qual seria o ordenamento das três probabilidades $P(A|I)$, $P(B|I)$ e $P(C|I)$? Procure alguém feminista e faça a pergunta, faça o mesmo com alguém machista. Divirta-se com a percepção que as pessoas são irracionais. O que você acha que as pessoas acham que respondem quando tem que ser rápidas? Note que muitas vezes ao fazer uma pergunta, quem responde está respondendo a uma pergunta parecida mas não exatamente aquela demandada.

Exercício Problema de Linda 3. Mostre usando a regra do produto que $P(A|I) \geq P(B|I)$. Tente inferir o que as pessoas fazem quando acham que está certo que $P(A|I) \leq P(B|I)$. Encontre asserções $A'|I'$ e $B'|I'$ *parecidas* com $A|I$ e $B|I$ tal que seja razoável supor mais provável supor o ordenamento contrário.

Exercício

- I : O preço do petróleo cai a 10 dolares o barril
- A : A Russia invade a Ucrânia
- B : A Russia invade a Ucrânia e os Estados Unidos corta relações diplomáticas com a Russia

Dado I qual é mais provável, A ou B? Note que as pessoas que cometem o erro de Falácia da Conjunção agem aparentemente como se estivessem comparando $P(A|I)$ com $P(C|AI)$, onde $B = AC$. Se você fosse presidente, manteria como assesor em política internacional alguém que ache $A|I$ menos provável que $B|I$?

Exercício

- I : Sou estudante da USP;
- A : Não estudei probabilidades
- B : Não estudei probabilidades e cometo a Falácia da conjunção

Dado I qual é mais provável, A ou B?

2

Outras definições de probabilidade

Kolmogorov e as probabilidades

Kolmogorov introduziu na década 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. Já descreveremos alguns dos motivos que nos levam a achar a posição freqüentista, incompleta e até, como mostraremos abaixo, insuficiente e errada. Pelo contrário, os axiomas de Kolmogorov, que codificam o bom senso da área já existente no trabalho de Laplace, podem ser vistos como não antagonicos aos resultados obtidos no capítulo 1. Interessante ler Kolmogorov. Ele não tem outro objetivo que

[1] Foundations of the Theory of Probability http://www.mathematik.com/Kolmogorov/index.html

> "... colocar no seu lugar natural, entre as noções gerais de matemática moderna, os conceitos básicos da teoria de probabilidade - conceitos que até recentemente eram considerados bastante peculiares.
>
> Esta tarefa não teria tido esperança de sucesso antes da introdução das teorias de medida e integração de Lebesgue..."
>
> A. N. Kolmogorov

Ele está organizando uma área após ficar claro como fazé-lo graças ao trabalho de Lebesgue e também Fréchet e admite que este ponto de vista era comum entre certos matemáticos mas merecia uma exposição concisa e livre de algumas complicações desnecessárias. Kolmogorov começa for considerar E [2] 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 tal sistema de conjuntos é chamado um $\sigma-$campo se $E \in \mathcal{F}$ e se a soma, produto, interseção de dois elementos quaisquer pertencem ao sistema. Os axiomas de Kolmogorov para a teoria de Probabilidades são

[2] Em física E é conhecido como espaço de fases

- AK1) $\mathcal{F}$ é um campo de conjuntos com $E \in \mathcal{F}$ e fechado ante um número de uniões (disjunções) e interseções (conjunções) 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 denotado por $P(A)$.

- AK4) $P(E) = 1$

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

Vejamos se estes axiomas estão de acordo com os resultados da seção anterior. Em primeiro lugar uma definição que não será necessária neste curso, a de σ-campo. É uma coleção de subconjuntos fechado ante um número contável de operações de conjunto, tais como disjunção, conjunção, negação. Esta noção só é necessária ao falar de conjuntos com infinitos elementos. Vimos que a coleção de asserções também permite tais operações. Portanto estamos lidando com o mesmo tipo de coleção de eventos que Kolmogorov. [3] Um exemplo de um σ-campo é o conjunto de conjuntos abertos nos reais. Neste curso usaremos asserções do tipo: "A variável X tem valor no aberto $(x, x + dx)$" e extensões a $\mathbb{R}^N$. A ideia de σ-campo é essencial na teoria de integração de Lebesgue e aparecerá em tratamentos matematicamente mais sofisticados de probabilidade. Neste curso não iremos além de integrais de Riemann e somas infinitas.

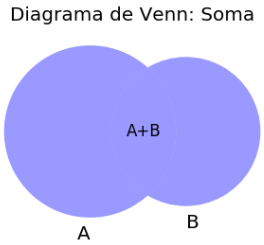
[3] Talvez a queixa é que as provas do capítulo 1 são para número finito de conjunções e disjunções. Isto porém não deve ser motivo de preocupação agora pois não é um empecilho irremovível.

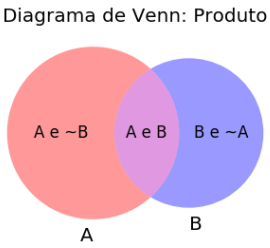
A probabilidade da *certeza* é 1 por AK4; a probabilidade está entre zero e um e a probabilidade da disjunção de asserções que não tem elementos em comum é a soma das probabilidades. Notamos, na introdução aos axiomas no livro de Kolmogorov, porém a falta de uma regra para o produto lógico ou disjunção de asserções ou ainda de intersecções de conjuntos. Kolmogorov não a introduziu incialmente e só em trabalhos posteriores foi incluida por sugestão de Rao. No livro (página 6) ele introduz, como um adendo aos axiomas, 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, usando a comutação de A e B, portanto $P(AB) = P(BA)$ e a simetria ante troca $A \leftrightarrow B$.

Se uma vez estabelecidos os axiomas e dados valores numéricos para as probabilidades partirmos para as aplicações matemáticas, não haverá nenhuma diferença de resultados pois será a mesma estrutura matemática. Enfatizamos que as diferenças que temos são sobre a motivação dos axiomas e com a interpretação da idéia de probabilidades. Isso tem importância em inferência e portanto em aplicações. Em muitos livros o estudante encontrará uma diferença entre probabilidades e probabilidades condicionais. Deve ficar claro que no ponto de vista destas notas, não há probabilidade que não seja condicional.





Ainda outras definições de Probabilidade

Outra proposta de definição de probabilidades é a frequentista, que tem mais chances de ser a que o leitor já viu. A definição parece muito simples: é o limite da razão entre o número de vezes que um evento é verdade e o número de tentativas, quando este último vai para infinito.

Esta definição veio no esteio de uma colocada por Jacob Bernoulli e Laplace. Para eles é às vezes conveniente definir a teoria de chances pela redução de eventos do mesmo tipo a certo número de casos, igualmente possíveis e a

> "...probabilidade, que é então simplesmente a fração cujo numerador é o número de casos favoráveis e cujo denominador é o número de todos os casos possíveis."[4]

O que significa "do mesmo tipo"? O físico verá aqui a uso da ideia de simetria. Se diferentes estados são tais que somos indiferentes ou incapazes de distingui-los então os colocamos na mesma categoria. Idéias de simetria são extremamente frutíferas. Mas quando não há simetria ou simplesmente não temos informação sobre ela é preciso estender a definição. Na época de Laplace as coisas não estavam muito claras, embora este tipo de regra seja útil e como veremos adiante não é uma regra nova a ser adicionada à *Desiderata* mas a ser deduzida do que já obtivemos. Além disso Laplace e Bernoulli deixaram claro em outros lugares que a probabilidade era uma manifestação numérica de crenças a partir de informação, portanto foram predecessores do exposto aqui. Considere, como Laplace há mais de duzentos anos, M_S a massa de Saturno. Ele fez asserções do tipo: "A probabilidade que $M_S < M_0$ ou $M_S > M_0 + \Delta m$ é menor que

[4] The theory of chances consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probabilty is sought. The ratio of this number to that of all the cases possible is the measure of this probability which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all te cases possible

A Philosophical Essay on Probabilities, Pierre Simon, Marquis de Laplace. 6a ed. F.W.Truscott e F.L. Emory trads.

10^{-4}", que ele colocou em linguagem de apostas. Em linguagem atual é algo como $P(M_0 < M_S < M_0 + \Delta m|I) > 1 - 10^{-4}$. A informação de fundo condicionante I representa a teoria de Newton e os dados experimentais [5]. Ele não está dizendo que a massa de Saturno é uma grandeza que apresenta variações e se for medida exatamente apresentará diferentes valores. Esqueça meteoritos, que poderiam mudar sua massa. Por exemplo, ao jogar um dado, se medirmos qual é o número de pontos na face que está para cima, este terá variações para diferentes jogadas. Alguns autores acham que só este tipo de variável merece ser descrito por probabilidades. Mas não a massa de Saturno à qual se atribui a propriedade de ter um valor *real*[6]. O que Laplace quer dizer é sobre o valor que atribuimos, com base nos dados e teoria, à crença que a massa está em um ou em outro intervalo. Quem acredita na definição de probabilidades como frequência, no pode falar da massa de Saturno em termos de probabilidade, pois não há um conjunto de Saturnos com diferentes massas. Falam em lugar disto, da probabilidade de que o conjunto de medidas seja observado para o caso em que a massa seja M_0. Em alguns casos isto dará o mesmo resultado, mas em outros não. Se você for acuado a definir a maior diferença entre alguém que define probabilidades através de frequências e quem as usa para expressar graus de crença, poderá responder de forma simplificada que este último não hesita em falar da distribuição de probabilidades de um parâmetro, como a massa de Saturno, enquanto o primeiro não admite tal linguagem [7].

[5] A *incerteza* Δm que Laplace tem é da ordem de 1% de M_0. O erro da estimativa de Laplace em relação ao valor estimado moderno é de aproximadamente 0.6%. Ou seja, ele teria ganho a aposta. O valor numérico $P(M_0 < M_S < M_0 + \Delta m|I)$ representa a crença que M_s esteja dentro do intervalo $(M_0, M_0 + \Delta m)$

[6] Real no sentido de ter existência independente do observador. Procure o significado de ontologico e de epistemico.

[7] Em linguagem mais técnica, ao espaço de parâmetros também é atribuido um $\sigma-$campo.

Algumas definições

Nesta altura podemos identificar os elementos formais principais para falar de probabilidades na linguagem de Kolmogorov. Primeiro é necessário deixar claro sobre o que se está falando:

- E a coleção de elementos A, B, C.... eventos elementares ou de asserções. Em alguns meios é chamado de espaço amostral.
- $\mathcal{F}$ o campo: o sistema de conjuntos de asserções. Espaço de eventos
- $P(A)$ a atribuição de um número positivo a cada elemento de $\mathcal{F}$;

Desta forma é costumeiro chamar a trinca (Espaço amostral, Espaço de eventos, Probabilidade de cada evento).

$$\boxed{(E, \mathcal{F}, \mathcal{P})}$$

de Espaço de probabilidades.

A apresentação do capítulo 1 não discorda disto, a não ser pelo ponto essencial que as probabilidades serão sempre condicionais e esquecer isso será a maior fonte de erros nas aplicações. Quando alguém se refere a uma probabilidade tipicamente tem em mente detalhes que se recusa a deixar explícitos pois esse exercício pode parecer cansativo. Outras vezes, e isso é mais perigoso, age como se tivesse em mente certos detalhes de informação, mas ao não perceber pode achar que não há alternativas. Além disso quando a teoria tem parâmetros, como será discutido em mais detalhes no próximo

capítulo, queremos poder falar das probabilidades de que os parâmetros tenham valores em uma dada região. Isto não está em desacordo com a posição de quem adota os axiomas de Kolmogorov. Basta aumentar o espaço amostral e o σ-campo e atribuir probabilidades aos elementos do novo campo. Isto porém não está de acordo com uma visão frequentista pois a massa de Saturno, ou qualquer outro parâmetro de uma teoria tem uma natureza ontológica que não lhes permite ser descrito em termos de frequência.

3
Uso elementar de Probabilidades

Este capitulo é muito mais simples que os anteriores, pois agora é uma questão de começar a desenvolver a estrutura matemática para poder lidar com aplicações simples.

Exemplos de sistemas onde a informação é incompleta

A escolha das variáveis e identificação de suas características para descrever um problema é o passo mais importante em todo o processo que iremos descrever. Em geral estamos interessados em identificar antes de tudo os graus de liberdade relevantes do problema e o espaço em que essas variáveis vivem.

Agora introduziremos alguns exemplos de sistemas que permitirão justificar o interesse do estudante no desenvolvimento futuro da teoria:

Moeda: Vamos apostar num jogo que envolve jogar uma moeda a vários metros de altura e deixar cair no chão Uma moeda é feita de níquel e ferro, tem propriedades magnéticas. No desenho na parte central, de um lado, aparece em relevo o 1 real e o ano que foi cunhada. Do outro, a imagem do rostro da República. Na parte externa um disco de bronze. A massa é aproximadamente 7g. A espessura 1.95 mm... Posso continuar dando informação irrelevante. Neste caso é fácil reconhecer que é irrelevante. O que voce quer saber é que posso descrever o estado final por uma variável que toma um de dois valores: $s = +1$ ou $s = -1$. Uma asserção sobre a que podemos pensar é "A moeda caiu com a cara para cima". É claro que neste caso foi muito fácil identificar a irrelevância da maior parte do que foi dito, mas isso nem sempre é óbvio e devemos ter cuidado.

Radioatividade: Um contador Geiger detecta partículas ionizantes. A asserção sobre aqual não temos informação completa é: "O intervalo de tempo entre detecções T", que pode tomar valores t, com $0 < t < \infty$. Ou no mesmo problema: "Qual é o número n de partículas detectadas num intervalo Δt."

Partícula: As coordenadas de uma partícula $\boldsymbol{R} = (X, Y, Z)$ tomam valores $\boldsymbol{r} = (x, y, z)$ dentro de uma caixa cúbica de lado L. Assim e.g. $0 < x < L$. Podemos atribuir probabilidade a asserções do tipo "A partícula tem coordenadas $\boldsymbol{R}$ dentro uma caixa de volume dV centrada em $\boldsymbol{r}$". Por preguiça diremos a mesma frase de forma simplificada "A partícula tem coordenadas $\boldsymbol{R} = \boldsymbol{r}$". Ou "A velocidade tem valores numa vizinhança de $\boldsymbol{v}$", onde a vizinhança tem tamanho

dado $v_x \in (v_x, v_x + \delta v_x)$, com expressões similares para a v_y e v_z. Mais sobre isto daqui a pouco.

Isto será interessante para descrever um gás de moléculas numa caixa:

Duas partículas. Na caixa descrita acima temos duas partículas idénticas mas distinguíveis. As coordenadas de cada uma são respectivamente $\boldsymbol{R}_1$ e $\boldsymbol{R}_2$. O espaço de fases é o produto cartesiano dos dois espaços. Como exemplos de asserções em que podemos estar interessados: "(A partícula 1 tem coordenadas $\boldsymbol{R}_1 = \boldsymbol{r}_1$) **e** (A partícula 2 tem coordenadas $\boldsymbol{R}_2 = \boldsymbol{r}_2$). Note que ao falar de $P(\boldsymbol{R}_1 = \boldsymbol{r}_1 \text{ e } \boldsymbol{R}_2 = \boldsymbol{r}_2|I)$ estamos falando do produto lógico das asserções individuais. Em geral e por preguiça a escreveremos $P(\boldsymbol{r}_1, \boldsymbol{r}_2|I)$ ou simplesmente $P(\boldsymbol{r}_1, \boldsymbol{r}_2)$

N partículas. Igual que acima mas agora N partículas. Falaremos da probabilidade $P(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N|I)$. Isto e variações sobre tema serão os tópicos principais do curso de Mecânica Estatística. O significado de I é de extrema importância, pois as probabilidades dependerão de que tipo de partícula estamos falando, das suas interações e das condições experimentais do sistema. A influência das partículas vizinhas sobre a partícula 1 pode ser descrita por probabilidades $P(\boldsymbol{r}_1|\boldsymbol{r}_2, ..., \boldsymbol{r}_N, I)$.

Medida da carga do elétron: Um conjunto $D = (d_1, d_2, ..., d_K)$ de medidas é feito no laboratório. A teoria nos fornece um modelo para a experiência que relaciona o parâmetro de interesse, neste caso a carga do elétron e, com a quantidade que medimos: $d = F(e)$. Mas sabemos que o dado d_i não é livre de erro de medida, ou seja não temos informação completa sobre d. Podemos então tentar codificar o que sabemos sobre d através de uma distribuição $P(d|D)$. Finalmente podemos falar sobre o conhecimento incompleto que temos sobre a carga e através de $P(e|D, I)$. Este tipo de análise é básico para a extração de informação a partir de medidas experimentais.

Cognição Um modelo de cognição de um animal pode ser feito considerando as variáveis relevantes. Os estados de neurônios de um sistema sensorial são descritos conjuntamente por uma variável X que toma valores x em algum espaço bastante complicado que não vem ao caso agora. Os estados de outras partes do cérebro são descritos por uma variável Z que toma valores z. O meio ambiente onde se encontra o animal é modelado por um conjunto de variáveis Y que tomam valores y, que certamente é um subconjunto das variáveis que poderiam ser usadas para descrever o *mundo lá fora*. O problema de cognição pode ser atacado considerando probabilidades $P(y|x, z, I)$. Neste caso I representa o conhecimento de Neurociência que tenhamos incluindo anatomia, dinâmica dos neurônios e dinâmica das sinápses. O mundo está em algum estado, mas o modelo só pode atribuir probabilidades às diferentes possibilidades, pois o animal tem informação incompleta. Pense sobre a modelagem de ilusões visuais, onde algo *parece mas não é verdade*. Substitua a palavra animal por máquina nesta modelagem e teremos a possibilidade de descrever modelos artificiais de cognição que são básicos na área de aprendizagem de máquinas (*machine learning*).

Agentes Econômicos e Sociais: Daremos alguns exemplos no decorrer das aulas, mas é interessante notar que o uso de estatística em ciências sociais precede o seu uso em física.

Esportes Um jogador de basquete arremessa com uma probabilidade $P(C|I)$ de converter uma cesta. Há dias em que tem uma mão quente?

Como vemos, tanto o teórico quanto o experimental poderão usar as ferramentas da teoria de probabilidades para tratar situações de informação incompleta.

Continue, olhe em volta e identifique sistemas que possam ser interessantes e descreva as varíaveis de interesse. Exemplos: Um dado cúbico, jogo de Bingo, condições de vida em um planeta, epidemia de Zika, bolsa de valores, uma amostra de ferro, e muito mais.

A partir de agora introduziremos alguns resultados matemáticos que serão úteis no desenrolar do curso.

Tipos de Váriáveis

Variáveis discretas

Uma variável S toma valores no conjunto $E = (s_1, s_2, ...s_N)$. Por exemplo para um dado de cúbico $E_{\text{dado}} = (1, 2, 3, 4, 5, 6)$. Mas pode ser muito mais rico que isto. As asserções que faremos serão do tipo A_i ="S vale s_i". Ou talvez B_{13} ="S toma valores no conjunto (s_1, s_3)."

Por preguiça, ou melhor para simplificar a notação, confundiremos as notações de tal forma que sob condições I a probabilidade $P(A_i|I)$ pode ser escrita simplesmente por $P(s_i|I)$. Ainda cometeremos a notação $P(s_i)$ sem especificar que há um condicionante I, talvez tacitamente suposto presente, mas às vezes esquecido de forma a levar a confusão e até a erros grosseiros. I será chamado de informação de fundo e envolve tudo o que sabemos sobre o problema. Chamaremos o conjunto de valores $P(s_i|I)$ de distribuição de probabilidades da variável S.

As asserções A_i são mutuamente exclusivas se o valor de S não pode ter simultaneamente dois valores quaisquer. Neste caso $A_i \wedge A_j = \varnothing$, para $i \neq j$ e portanto $P(A_i \wedge A_j|I) = 0$. Também são exaustivos de forma que não há possibilidade de que S tenha valores fora desse conjunto. Assim temos que

$$A_1 \vee A_2 \vee .... \vee A_N = E$$

e temos certeza que E é verdadeiro. Segue que

$$\begin{aligned} 1 &= P(A_1 \vee A_2 \vee .... \vee A_N|I) \\ 1 &= \sum_{i=1}^{N} P(A_i|I). \end{aligned} \tag{3.1}$$

Esta última expressão indica que a soma sobre todas os valores possíveis de S é um e será satisfeita por toda distribuição de probabilidades. É chamada condição de *normalização*.

Váriáveis reais: densidades de probabilidades

Em particular estamos interessados em grandezas físicas descritas por variáveis que tomam valores em intervalos dos reais, que chamaremos L.

No que segue lidaremos com asserções do tipo "a variável X toma valores entre x e $x + dx$". Não sabemos ainda como, mas suponha que atribuimos um número a esta probabilidade. Como seria se lidassemos com a probabilidade de "X toma valor x"? Escolha um número entre 0 e 1. Se todos os números forem igualmente prováveis,

a probabilidade de cada um deles seria zero, pois a soma deve dar um. Vemos que rapidamente chegamos a bobagens. Em geral e porque ainda não temos a matemática para lidar como esse tipo de problema, iremos falar somente de probabilidade de intervalos. Isso nos permite introduzir a densidade $P(x|I)$ tal que a probabilidade de que "a variável X toma valores entre x e $x+dx$" é dada por $P(x|I)dx$. $P(x|I)$ não é uma probabilidade mas é chamada de densidade de probabilidade [1]. Teremos então que

[1] Usamos a letra P por motivos históricos e eventualmente a chamaremos de probabilidade, por preguiça. Também esqueceremos de apontar os condicionantes e escreveremos muitas vezes simplesmente $P(x)$.

- $P(x|I) \geq 0$
- $\int_L P(x|I)dx = 1$

Aqui reconhecemos a generalização da condição de normalização da equação 3.1, pois o intervalo L engloba todas as possibilidades de valores de X. Mas para qualquer intervalo $D : \{x|x \in [x_1, x_2]\}$, a probabilidade de X estar em D ou $x_1 \leq x \leq x_2$ é

$$P(x \in D|I) = \int_D P(x|I)dx$$

Distribuição cumulativa de probabilidade

Se uma variável X toma valores x no eixo real, e é descrita por uma densidade $P(x|I)$, a distribuição cumulativa é definida por

$$\Phi(x|I) = \int_{-\infty}^{x} P(x'|I)dx'. \tag{3.2}$$

seque que $\Phi(x|I)$ é a probabilidade de X tomar valores menores que x e a densidade de probabilidade é

$$P(x|I) = \frac{d}{dx}\Phi(x) \tag{3.3}$$

] A probabilidade de que X tenha valores num intervalo é

$$P(x_i < X < x_2|I) = \Phi(x_2) - \Phi(x_1)$$

Caracterização de distribuições e densidades de Probabilidade

A informação disponível ao falar de X será equivalente à densidade de probabilidade para todo x. Mas isto talvez seja muito. É comum que seja necessário caracterizar, pelo menos parcialmente, o valor de X com um número, isto é um estimador ou estimativa de X. Há várias possibilidades e cada uma tem utilidade

- (1) $x_M = \text{maxarg } P(x|I)$
- (2) $\langle x \rangle = \mathbb{E}[x] = \int_L xP(x|I)dx$
- (3) x_m tal que $\int_{x \leq x_m} P(x|I)dx = \int_{x \geq x_m} P(x|I)dx$

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

A moda é o valor mais provável. Não quer dizer que se fizermos uma medida de X o obteremos, mas é o valor que terá mais probabilidade de ser encontrado. Podem haver vários valores que satisfazem o critério. A média leva em consideração todos os valores possíveis, cada um com voto proporcional à sua probabilidade. A mediana é o

valor tal que a probabilidade de ser menor ou maior é igual. Cada uma é útil ou não em diferentes cirscunstâncias. Veja os exercícios. Cada uma resume a informação de forma a contar uma história. Devemos ter cuidado pois o contador da história pode ter um motivo para contar a história de forma resumida da maneira que é mais ou menos favorável a uma idéia que quer ver defendida. Podemos pensar em outras formas generalizando as idéias acima.

O valor esperado ou esperança de uma função $f(x)$, denotado por $\mathbb{E}_x(f)$ ou $\mathbb{E}(f)$, ou ainda alternativamente por $\langle f(x)\rangle$, é definido por

$$\mathbb{E}_x(f) = \langle f(x)\rangle = \int_L f(x)P(x)dx \tag{3.4}$$

Usaremos tanto a notação $\mathbb{E}_x(f)$ ou $\mathbb{E}(f)$, preferida em textos de Matemática quanto $\langle f(x)\rangle$ mais usada em textos de Física. A notação que usamos de alguma forma deixa esquecida a idéia que a probabilidade depende da informação disponível. Quando for necessário deixar explícita a informação condicionante usaremos $\mathbb{E}_x(f|C)$ ou $\langle f(x)\rangle|_C$.[2]

[2] Usaremos esta notação às vezes, pois usaremos o direito de ser inconsistentes na notação, esperando que isso não confunda o leitor, mas o torne imune às várias notações na literatura. Isso

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

$$\mathrm{Var}(X) := \mathbb{E}((x - \mathbb{E}(x))^2) = \langle (x - \langle x\rangle)^2\rangle. \tag{3.5}$$

Obviamente $\mathrm{Var}(X) \geq 0$. É fácil mostrar que $\mathrm{Var}(X) = \langle x^2\rangle - \langle x\rangle^2$. Algumas vezes nos referiremos à variância por σ_X^2, por preguiça que veremos justificada algumas vezes, mas outras não.

O valor esperado será muito usado no que segue, podemos generalizar a ideia e introduzir os momentos de uma distribuição:

- $m_n := \langle x^n\rangle = \mathbb{E}[x^n] = \int_L x^n P(x)dx$

para valores inteiros de n (claro que caso a integral exista). Em notação mais carregada

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

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 sua média seja nula:

- $M_{n|C} := \langle (x - \langle x\rangle)^n\rangle_{|C} = \int_L (x - \langle x\rangle)^n P(x|C)dx$

e note que $\mathrm{Var}(X) = M_2$.

Marginais e Independência

As idéias de Marginalização e independência são de grande importância em toda a teoria e as aplicações que seguem.

Marginalização

Considere as asserções $a, b, \bar{b}, c$ e os produtos $ab|c$ e $a\bar{b}|c$. Um resultado extremamente útil, que já usamos no capítulo 1, é

$$p(a|c) = p(ab|c) + p(a\bar{b}|c)$$

A prova é simples e a intuição também. Por exemplo a ='uma pessoa tem altura entre h e $h + \Delta h$ ', b ='uma pessoa tem peso menor que w'. Assim temos que a probabilidade de a, ter altura no intervalo é a soma das probabilidades de ter altura nesse intervalo e ter peso menor que w somada à probabilidade de ter altura nesse intervalo e ter peso maior ou igual a w.

A prova usa a regra do produto duas vezes, e a da negação uma:

$$\begin{aligned} p(ab|c) + p(a\bar{b}|c) &= p(a|c)p(b|ac) + p(a|c)p(\bar{b}|ac) \\ &= p(a|c)\left(p(b|ac) + p(\bar{b}|ac)\right) \\ &= p(a|c) \end{aligned} \tag{3.6}$$

Claro que se tivermos b que toma valores sobre um conjunto de asserções $\{b_i\}_{i=1,...N}$ mutuamente exclusivas e exaustivas teremos

$$p(a|c) = \sum_{i=1}^{N} p(ab_i|c)$$

e dizemos ao marginalizar $p(ab|c)$ sobre a variável b obtemos a distribuição $p(a|c)$.

Voltando às alturas e pesos olhe uma tabela das probabilidades conjuntas onde cada entrada descreve o conhecimento para uma certa faixa de peso e de altura. Somamos as entradas para cada linha e as escrevemos na margem direita. Estas são simplesmente as probabilidades para a faixa de altura sem levar em conta o peso. Essa é a origem do termo marginal, pois era anotado à margem da tabela conjunta quando o papel era o meio usado para aumentar a memória do usuário. Somando as entradas ao longo das colunas temos a probabilidade do peso independente de altura.

	w_1	w_2	...	w_N	$\sum_{i=1...N} P(h_j, w_i)$
h_1	$P(h_1, w_1)$	$P(h_1, w_2)$	...	$P(h_1, w_N)$	$P(h_1)$
h_2	$P(h_2, w_1)$	$P(h_2, w_2)$	...	$P(h_2, w_N)$	$P(h_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h_M	$P(h_M, w_1)$	$P(h_M, w_2)$	...	$P(h_M, w_N)$	$P(h_M)$
$\sum_{j=1...M} P(h_j, w_i)$	$P(w_1)$	$P(w_2)$	...	$P(w_N)$	

Tabela 1.3
As marginais são escritas na margem!

O outro conceito de extrema importância é o de

Independência

A regra do produto em geral

$$p(ab|c) = p(a|c)p(b|ac)$$

se reduz ao produto das marginais

$$p(ab|c) = p(a|c)p(b|c) \tag{3.7}$$

quando $p(b|ac)$ não depende de a. Se informação da veracidade de a não altera crenças sobre b, dizemos que nas condições de que c seja verdadeiro, b é independente de a. É óbvio que a independência é reflexiva, pois também podemos escrever

$$p(ab|c) = p(b|c)p(a|bc)$$

o que significa, dada a equação 10.35 que $p(a|bc) = p(a|c)$. Assim temos que distribuições conjuntas de variáveis independentes se reduzem a produtos. As interações físicas entre partículas serão descritas por distribuições que não se fatorizam nas marginais, i.e nas probabilidades das variáveis de cada partícula.

Independência aos pares , mútua e condicional

Suponha que tenhamos um conjunto de asserções sob consideração $S = \{A_1, A_2, ...A_K\}$. Dizemos que os A_i são independentes aos pares na condição C, se para todo i, j com $1 \leq i \leq K$ e $1 \leq j \leq K$, $i \neq j$ tivermos

$$P(A_i|A_jC) = P(A_i|C).$$

Dizemos que os membros de S são mutuamente independentes na condição C se

$$P(A_1, A_2, A_3...A_K|C) = P(A_1|C)P(A_2|C)....P(A_K|C).$$

Mas é claro que em geral, a distribuição conjunta pode ser manipulada usando a regra do produto. Para $K = 3$, supondo independência aos pares

$$\begin{aligned} P(A_1, A_2, A_3|C) &= P(A_1|C)P(A_2A_3|A_1C) = P(A_1|C)P(A_2|A_1C)P(A_3|A_1A_2C) \\ &= P(A_1|C)P(A_2|C)P(A_3|A_1A_2C) \end{aligned} \tag{3.8}$$

e para a chegar ao produto $\prod_{i=1,2,3} P(A_i|C)$, deveriamos ainda impor que $P(A_3|A_1A_2C) = P(A_3|C)$ que é mais restritiva que indepêndencia aos pares. Mas isto é sutil e merece um exemplo específico para ficar mais claro.

Vamos imaginar uma moeda sendo jogada. A_1 :"cara na primeira jogada", A_2 :"cara na segunda jogada", A_3 :"as duas jogadas tiveram o mesmo resultado", que é equivalente a escrever $A_3 = A_1A_2 + \overline{A_1}\,\overline{A_2}$. Dada a independência das duas jogadas temos

$$\begin{aligned} &P(A_1|A_2C) = P(A_1|C), \qquad P(A_2|A_1C) = P(A_2|C) \\ &P(A_2|A_3C) = P(A_2|C), \qquad P(A_3|A_2C) = P(A_3|C) \\ &P(A_3|A_1C) = P(A_3|C), \qquad P(A_1|A_3C) = P(A_1|C) \end{aligned} \tag{3.9}$$

mas

$$P(A_1|A_2A_3C) = 1 \neq P(A_1|C)$$

Completamos a definição de mutuamente independente se $P(A_i|B_iC) = P(A_i|C)$ onde B_i é uma conjunção de qualquer subconjunto de S que não inclua A_i.

Como toda probabilidade é condicional, a independência também depende do contexto. Podemos ter $P(X|YZ_1) = P(X|Z_1)$ mas $P(X|YZ_2) \neq P(X|Z_2)$. Por exemplo no caso das moedas Z_1 e Z_2 poderiam diferir nas condições iniciais do lançamento e.g. altura, energia, velocidade angular, etc. Vamos supor que X, Y e Z tomem valores reais. Se X e Y forem independentes na condição Z, ou seja

$P(XY|Z) = P(X|Z)P(Y|Z)$, então a como funções dos valores destas variáveis teremos $P(XY|Z) = P(X = x, Y = y|Z = z) = P(x, y|z)$ deve satisfazer

$$P(x, y|z) = f(x, z)g(y, z).$$

Por outro lado se $P(x, y|z) = f(x, z)g(y, z)$ é possível mostrar que X e Y são independentes na condição Z.

Para concluir estas definições, o estudante deve notar que a idéia de independência não deve ser confundida com a de mutuamente exclusivo. Independência leva a que a regra do produto é

$$P(ab|c) = P(a|c)P(b|ac) = P(a|c)P(b|c).$$

Mutuamente exclusivo implica em

$$P(a+b|c) = P(a|c) + P(b|c) - P(ab|c) = P(a|c) + P(b|c) - P(a|c)P(b|ac) = P(a|c) + P(b|c).$$

Exemplos de Famílias de Distribuições de probabilidade

No contexto deste curso, uma variável aleatória é simplesmente alguma variável para a qual não temos informação completa e portanto, o que soubermos será usado para construir uma distribuição de probabilidades. É comum que a distribuição seja escolhida dentro de uma família. Uma função de pelo menos duas variáveis $f(x;\Theta)$, não negativas e integráveis no primeiro argumento, pode ser considerada uma família paramétrica de funções de x com Θ como parâmetro. Tanto x quanto Θ podem ser multidimensionais. Apresentaremos a seguir exemplos de familias onde x pode ser discreto ou contínuo, unidimensional ou multidimensional. Algumas famílias das distribuições aparecem de forma recorrente em muitas aplicações e vale a pena ter certa familiaridade. Podemos ter diferentes motivos que levem ao uso de uma família. Por exemplo, desde o mais simples como informação sobre o domínio de valores de uma variável, a motivos teóricos sobre a dependência entre as variáveis relevantes. Os motivos teóricos podem ser toda a área da Mecânica Estatística e as dependências terem relação com forcas entre partículas.

O que segue não pode ser considerado uma exposição completa das propriedades das distribuições. Algumas, como a binomial e a gaussiana, serão tratadas com muito mais detalhe em capítulos posteriores. A notação usual em estatística ao dizer que a variável X tem distribuição do tipo Blablabla com parâmetros Θ é

$$X \sim Bla(\Theta)$$

usando algumas letras do nome da distribuição que pode ser o nome de alguma pessoa, indicando também os valores ou nomes dos parâmetros.

A utilidade varia de motivações teóricas que forçam um dado tipo de modelo a simplemente a possibilidade de fazer algum avanço analítico. De qualquer forma é sempre útil ter um poste onde possamos procurar a chave perdida.

Bernoulli

Esta distribuição é uma das mais simples. Se uma variável está distribuida de acordo com a distribuição (ou equivalentemente é uma

variável) de Bernoulli escrevemos $S \sim \text{Ber}(p)$. Neste caso, S tem dois valores possíveis. Por exemplo o espaço de valores possíveis de S é $E = \{-1, +1\}$ ou $\{\text{cara, coroa}\}$, ou $\{0, 1\}$. A distribuição de Bernoulli é em termos de um parâmetro p, $0 \leq p \leq 1$

$$P(S|p) = \begin{cases} p & \text{se } S = +1 \\ 1-p & \text{se } S = -1. \end{cases}$$

Esta forma de escrever a probabilidade dá a impressão que p é uma probabilidade, mas isso é errado e leva a grande confusão; p é um parâmetro (que talvez fosse melhor chamar θ). Uma maneira muito melhor de escrever a probabilidade é usando a delta de Kronecker: para a e b tomando valores em um conjunto discreto $\delta ab = 1$ se $a = b$ e zero se forem diferentes. Assim

$$P(S|p) = p\delta_{S,1} + (1-p)\delta_{S,-1},$$

onde agora fica claro o que é variável aleatória e o que é parâmetro. Também pode ser escrita, usando o parâmetro $m = 2p - 1$ como

$$P(S|m) = \begin{cases} \frac{1+m}{2} & \text{se } S = +1 \\ \frac{1-m}{2} & \text{se } S = -1. \end{cases}$$

O valor esperado de S é

$$\mathbb{E}(s|p) = \langle S \rangle = \sum_{s=-1,1} sP(S = s) = m = 2p - 1,$$

que dá a interpretação de m e o motivo por que é interessante usá-lo como parâmetro da distribuição. O segundo momento é simples pois $S^2 = 1$ portanto

$$\mathbb{E}(s^2|p) = \langle S^2 \rangle = 1$$

a para a variância σ_S temos

$$\sigma_S^2 = \langle S^2 \rangle - \langle S \rangle^2 = 1 - m^2 = 4p(1-p)$$

Sob o risco de ser maçante introduzimos a variável $R = \frac{1+S}{2}$ e agora temos

$$P(R|p) = \begin{cases} p & \text{se } R = 1 \\ 1-p & \text{se } R = 0. \end{cases}$$

portanto o valor esperado $\langle R \rangle = p$ e a variância $\sigma_R^2 = \langle R^2 \rangle - \langle R \rangle^2 = p(1-p)$. Estas variáveis sozinhas podem parecer muito simples, mas ao juntar várias partículas cujos estados são descritos por variáveis deste tipo vamos poder modelar fenômenos bem complexos. Por exemplo S pode representar classicamente o spin de um íon numa rede cristalina ou R pode indicar a presença ou ausência de uma partícula num modelo do que se chama um gás de rede.

A variância, o valor esperado do quadrado da flutuação, vai a zero quando $p = 0$ ou $p = 1$ que são os casos em que a informação é completa: $S = -1$ sempre no primeiro caso e $S = 1$ sempre no segundo. A variância traz informação sobre a largura da distribuição e isso não se restringe a esta distribuição.

Podemos introduzir uma terceira maneira de representar a distribuição de Bernoulli

$$P(s|\beta) = \frac{e^{-\beta s}}{\zeta(\beta)} \tag{3.10}$$

que é muito usada em Mecânica Estatística. A normalização

$$1 = \sum_{s=\pm 1} P(s|\beta) = \frac{\sum_{s=\pm 1} e^{-\beta s}}{\zeta(\beta)} \tag{3.11}$$

portanto

$$\zeta(\beta) = \frac{1}{2}(e^{\beta} + e^{-\beta}) = 2\cosh\beta.$$

A ligação com as representações anteriores

$$\mathbb{E}(s|p) = m = 2p - 1 = \frac{e^{\beta} - e^{-\beta}}{e^{\beta} + e^{-\beta}} = \tanh\beta$$

que pode ser estranha neste momento mas se mostrará útil.

Uniforme

Uma variável $X \sim U(0, L)$ toma valores no intervalo do eixo real $\mathcal{L} : 0 < x < L$ e sua probabilidade é uma constante dentro do intervalo e zero fora:

$$P(X|L) = \begin{cases} \frac{1}{L} & \text{se } X \in \mathcal{L} \\ 0 & \text{se não.} \end{cases}$$

Os valores esperados e variância são

$$\begin{aligned} \langle X \rangle &= \int_{\mathcal{L}} xP(x)dx = \frac{L}{2} \\ \langle X^2 \rangle &= \int_{\mathcal{L}} x^2 P(x)dx = \frac{L^2}{3} \\ \sigma_X &= \frac{L}{2\sqrt{3}} \end{aligned}$$

Obviamente podemos fazer translações $Y = aX + B$ e teremos $Y \sim U(B, aL + B)$ com probabilidade $1/aL$ dentro e 0 fora do intervalo.

Binomial

Uma variável de Bernoulli toma valores $s = +1$ ou $s = -1$ e é amostrada N vezes de forma mutuamente independente. Ou seja temos um conjunto de dados escritos como uma lista $(s_1, s_2, ...s_N)$. A variável binomial é m o número de vezes que aparece o $+1$ nessa lista. Assim $m \sim \text{Bin}(p; N)$. Obviamente a distribuição de Bernoulli é Ber(p)=Bin$(p; 1)$. Mostraremos no próximo capítulo que

$$\begin{aligned} P(m|pN) &= \binom{N}{m} p^m (1-p)^{N-m} \\ &= \frac{N!}{m!(N-m)!} p^m (1-p)^{N-m} \end{aligned} \tag{3.12}$$

Você encontrará frequentemente que isto é descrito como a probabilidade de n sucessos em N tentativas, quando p é a probabilidade de sucesso em cada tentativa. Voltaremos a falar desta distribuição várias vezes. Em particular fica faltando aqui discutir mais indepêndencia entre as tentativas.

Binomial Negativa

Esta é uma variação sutil sobre o tema de Bernoulli com respeito à distribuição anterior. Se obter $s_i = 1$ foi chamado de sucesso, então é natural chamar $s_i = -1$ de fracasso, Agora fixamos o número de fracassos k e pedimos a probabilidade do numero de sucessos n até obter k fracassos.

$$P(n|pk) = \binom{n+k-1}{n} p^n (1-p)^k$$

Nas primeiras $n+k-1$ tentativas a ordem pode ser qualquer e o número destas seqüencias é $\binom{n+k-1}{n}$. A última tentativa, a $n+k$ deve ser um fracasso. A média é $\mathbb{E}(n) = pk/(1-p)$ e a variância $pk/(1-p)^2$

Para verificar que a normalização é correta precisamos alguns truques. Primeiro usamos a soma da progressão geometrica

$$\frac{1}{1-p} = \sum_{s=0}^{\infty} p^s = 1 + p + p^2 \cdots + p^{k-1} + p^k + ...$$

e a derivada de ordem $k-1$

$$\frac{d^{k-1}}{dp^{k-1}}\left(\frac{1}{1-p}\right) = \frac{(k-1)!}{(1-p)^k}$$

que elimina os primeiros $k-1$ termos da soma da PG. Deixamos os detalhes para o leitor.

Poisson

Para descrever a estatística de contagens de um detetor é útil introduzir a distribuição de Poisson. Veremos adiante que esta distribuição está relacionada com a binomial. A probabilidade de n, número de contagens em um certo intervalo de tempo, dado o parâmetro λ que caracteriza o processo, é

$$P(n|\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$$

O valor médio

$$\begin{aligned}
\langle n \rangle &= \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \lambda \frac{d}{d\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \lambda \frac{de^{\lambda}}{d\lambda} \\
&= \lambda, \qquad (3.13)
\end{aligned}$$

e o segundo momento

$$\begin{aligned}
\langle n^2 \rangle &= \sum_{n=0}^{\infty} n^2 \frac{\lambda^n}{n!} e^{-\lambda} \\
&= e^{-\lambda} \left(\lambda \frac{d}{d\lambda}\right) \left(\lambda \frac{d}{d\lambda}\right) e^{\lambda} \\
&= \lambda + \lambda^2. \qquad (3.14)
\end{aligned}$$

Portanto a variância

$$\sigma^2_{Poisson} = \lambda \qquad (3.15)$$

Beta

Uma variável X toma valores x no intervalo $0 \leq x \leq 1$ e tem dois parâmetros

$$P(x|a;b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \tag{3.16}$$

Note que se a e b forem números inteiros, podemos escrever

$$\begin{aligned} P(x|a;b) &= \frac{(a+b-1)!}{((a-1)!(b-1)!} x^{a-1}(1-x)^{b-1} \\ P(x|n=a-1;N=b+m-1) &= \frac{(N+1)!}{n!(N-n)!} x^{n}(1-x)^{N-n} \end{aligned} \tag{3.17}$$

onde a parametrização da última linha mostra uma certa semelhança com a binomial. Uma pequena diferença é que em lugar de N temos $N+1$ no numerador. A diferença fundamental é que na binomial falamos da probabilidade de n e aqui de x. As duas distribuições estão relacionadas pelo resultado de Bayes: $P(n|x) \propto P(x|n)$. Voltaremos a falar nesta relação ao falar de distribuições conjugadas.

Para a normalização usamos um resultado devido a Euler

$$E^n_{N-n} = \int_0^1 p^n(1-p)^{N-n}dp = \frac{n!(N-n)!}{(N+1)!} \tag{3.18}$$

Suponha que $n < N-n$. Integramos por partes com $dv = (1-p)^k dp$ e $u = p^r$ que leva a $v = -\frac{1}{k+1}(1-p)^{k+1}$ e $du = rp^{r-1}$, assim

$$\begin{aligned} E^r_k &= \int_0^1 p^r(1-p)^k dp \\ &= \frac{r}{k+1}\int_0^1 p^{r-1}(1-p)^{k+1}dp \\ &= \frac{r}{k+1}E^{r-1}_{k+1}. \end{aligned}$$

Começando com $r = n$, Após n passos temos uma integral $\propto \int_0^1 (1-p)^N dp = 1/(N+1)$. Iterando

$$E^n_{N-n} = \frac{n}{N-n+1}\frac{n-1}{N-n+2}\cdots\frac{n-(n-1)}{N-n+n}\frac{1}{N+1} \tag{3.19}$$

Multiplicando e dividindo por $(N-n)!$ obtemos o resultado 3.18. Se $n > N-n$, mude variáveis de integração $p \to 1-p$ e proceda da mesma forma. Podemos calcular momentos da Beta da mesma forma, pois $\mathbb{E}(p^r|\text{Beta}(n,N)) \propto E^{n+r}_{N-n}$

Gamma

O nome desta distribuição é devido a que a função Gama é definida pela integral

$$\Gamma(u) = \int_0^\infty e^{-t}t^{u-1}dt, \tag{3.20}$$

que voltaremos a ver várias vezes, em particular no capítulo 5. Uma variável X toma valores x no intervalo $0 \leq x < \infty$ e tem dois parâmetros a, conhecido com o parâmetro de escala e b o parâmetro de forma:

$$P(x|a;b) = \frac{1}{a\Gamma(b)}(\frac{x}{a})^{b-1}e^{-\frac{x}{a}} \tag{3.21}$$

O valor esperado $\langle x\rangle = \mathbb{E}(x) = ab$ e a variância $\mathbb{E}(x^2) - \mathbb{E}(x)^2 = a^2b$.

Gaussiana ou Normal

Dedicaremos o capítulo 5 ao estudo desta distribuição. Uma variável X toma valores x no intervalo $-\infty < x < \infty$ e tem dois parâmetros: μ a média e σ^2 a variância:

$$P(x|\mu;\sigma) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.22}$$

uma váriável que tem esta ditribuição é dita normal ou gaussiana e tipicamente na literatura estatística se escreve

$$X \sim \mathcal{N}(\mu, \sigma)$$

Distribuição Exponencial

X toma valores reais não negativos, $x \geq 0$. Um único parâmetro, $a > 0$ dá a escala e a variância. A distribuição é

$$P(x|a) = \frac{1}{a} e^{-\frac{x}{a}} \tag{3.23}$$

O valor médio e a variância são respectivamente

$$\mathbb{E}(x) = a, \qquad \mathbb{E}(x^2) - \mathbb{E}(x)^2 = a^2$$

Laplace

Semelhante à exponencial, mas com x podendo ser qualquer valor real, portanto também conhecida como dupla exponencial,

$$P(x|\mu a) = \frac{1}{2a} e^{-\frac{|x-\mu|}{a}} \tag{3.24}$$

onde μ é um parâmetro de localização e a de escala. Note o fator 2 para garantir a normalização.

Cauchy

a distribuição de Cauchy tem vários nomes associados, Lorentz, Cauchy-Lorentz, Breit-Wigner.

$$P(x|x_0, a) = \frac{1}{a\pi} \frac{1}{1 + \frac{(x-x_0)^2}{a^2}} \tag{3.25}$$

O valor médio não é definido da forma convencional, mas usando uma definição da integração em intervalos infinitos devida a Cauchy, o valor principal de Cauchy

$$\begin{aligned} \mathbb{E}(x - x_0) \quad &= \quad \mathbb{E}(x) - x_0 = \mathcal{P} \int_{-\infty}^{\infty} (x - x_0) P(x|x_0, a) dx \\ &= \quad \lim_{L \to \infty} \int_{-L}^{L} (x - x_0) P(x|x_0, a) dx = 0 \end{aligned} \tag{3.26}$$

por simetria, logo

$$\mathbb{E}(x) = x_0,$$

que coincide com a moda e a mediana. O interessante é que $\mathbb{E}((x - x_0)^2)) = \infty$ e portanto a variância é infinita. Assintoticamente as contribuições para a integral vão como $x^2/x^2 \sim$ constante. Mas ainda podemos definir a largura a meia altura, que é $2a$, a a separação entre os pontos $x_0 + a$ e $x_0 - a$, onde a probabilidade é $1/2\pi a$.

Mudança de Variáveis

Ao analisar um sistema em física, o problema mais importante e imediato é o de identificar as variáveis relevantes para representar seus estados. Estudantes inexperientes podem achar que essa parte é fácil. O motivo é que foi dito que o espaço tem esta e aquela característica, que o tempo é esse parâmetro t que todos sabem o que é (menos eu). O que talvez não fique claro é que milhares de anos de tentativas levaram a atribuir certos modelos matemáticos a sistemas físicos e ficam escondidas as várias tentativas que acabaram em becos sem saida, ou que se verificou posteriormente, podiam ser significativamente simplificados.

Suponha que voce tenha informação I sobre uma variável X que toma valores x reais e codifique esse conhecimento numa densidade de probabilidade $P(x|I)$. Por algum motivo, fica claro que seria útil introduzir Y que esta relacionada com X por uma função f conhecida

$$y = f(x).$$

A pergunta que se coloca é o que podemos dizer sobre a densidade de Y sob as mesmas condições de informação I?

A resposta é fácil se pensarmos sobre o significado de densidade de probabilidade. Vamos começar com $f(x)$ uma função monotónica, que permite uma inversão $x = f^{-1}(y)$. Consideremos $y_i = f(x_i)$ para $i = 1, 2$ e f crescente. A asserção

"O valor de X toma valores x, tal que $x_1 < x < x_2$"

deve ser equivalente à asserção

"O valor de Y toma valores y, tal que $y_1 < y < y_2$"

Equivalente no sentido de que a mesma probabilidade deve ser atribuida a cada uma delas se o contexto for o mesmo

$$Prob(y_1 < y < y_2|I) = Prob(x_1 < x < x_2|I)$$

A relação entre as densidades de probabilidades deve ser

$$\int_{y_1}^{y_2} P(y|I)dy = \int_{x_1}^{x_2} P(x|I)dx$$

Se os intervalos de integração forem suficientemente pequenos podemos escrever

$$P(y|I)\Delta y = P(x|I)\Delta x$$

e no limite

$$P(y|I) = P(x|I)\frac{dx}{dy}$$

isto não é mais do que simplesmente tomar a derivada com respeito ao limite superior (no ponto $y_2 = y$) e usar a regra da cadeia. As regras de mudança de variáveis não são mais que as regras de mudança de variável na teoria de integração ou de medida. É difícil exagerar a importância deste resultado.

O leitor poderá agora estender os resultados para o caso em que f for decrescente. Agora $dx/dy = df^{-1}(y)/dy$ deverá ser substituida por $-dx/dy$. Também deve poder encontrar as regras quando f não for monotónica, ou ainda quando x e y forem generalizadas para mais dimensões.

Se a função $f(x)$ não for monotónica precisamos ter cuidado. Olhemos para um exemplo simples. Seja $U = X^2$, portanto um valor

u de U está associado a um valor x de X por $u = x^2$. A asserção que U é menor que um dado valor u, $U < u$ é idéntica à asserção que $-\sqrt{u} < X < \sqrt{u}$, portanto, em termos da cumulativa

$$\begin{aligned}\Phi(u|I) = Prob(U < u|I) &= Prob(-\sqrt{u} < X < \sqrt{u}|I) \\ &= Prob(X < \sqrt{u}|I) - Prob(X < -\sqrt{u}|I) \\ &= Prob(X < x|I) - Prob(X < -x|I) \\ &= \Phi(x|I) - \Phi(-x|I) \qquad (3.27)\end{aligned}$$

derivando com respeito a u temos a densidade de probabilidade

$$\begin{aligned}P(u|I) &= \frac{d}{du} Prob(U < u|I) \\ &= (\frac{d}{dx} Prob(X < x|I) - \frac{d}{dx} Prob(X < -x|I)) \frac{dx}{du} \text{ onde } x = \sqrt{u} \\ &= \left(P(X = \sqrt{u}|I) + P(X = -\sqrt{u})\right) \frac{1}{2\sqrt{u}}.\end{aligned} \qquad (3.28)$$

A transformação neste caso não é invertível e precisamos levar em conta os dois ramos da inversa, tanto $+\sqrt{u}$ quanto $-\sqrt{u}$.

A integração especialmente em espaços de alta dimensionalidade é uma das tarefas mais comuns nas aplicações e consumirá a maior parte dos esforços computacionais. No capítulo sobre integração Monte Carlo veremos como mudanças de variáveis serão elevadas a uma forma de arte.

Covariância e correlações

Introduziremos de forma rápida mas voltaremos a usar muitas vezes a idéia de correlações que é central nas aplicações. Duas variáveis X e Y tem distribuição conjunta $P(x, y|I)$ sob informação I. O valor esperado do produto é

$$\mathbb{E}(xy) = \langle xy \rangle = \int xy P(xy|I) dxdy.$$

e o valor esperado do produto das variáveis truncado, isto é, substraido o valor médio de cada variável, é a covariância

$$\begin{aligned}\mathrm{Cov}_{xy} = \mathbb{E}\left((x - \mathbb{E}(x))(y - \mathbb{E}(y))\right) &= \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle \\ &= \langle xy \rangle - \langle x \rangle \langle y \rangle \qquad (3.29)\end{aligned}$$

que é o valor esperado do produto das flutuações em torno da média. Dadas as propriedades de X, o maior valor quea covariância pode ter é quando X e Y são iguais, pois integral só tem contribuições positivas. Nesse caso $\mathrm{Cov}_{xx} = \mathrm{Var}(x)$. Isso sugere introduzir a correlação r, que aparentemente foi introduzida por Pearson

$$r = \frac{\mathrm{Cov}_{xy}}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} \qquad (3.30)$$

e que satisfaz $-1 \leq r \leq 1$.

No caso de n variáveis X_i, $i = 1...n$, a matriz de correlações C_{ij} tem elementos $C_{ij} = \mathrm{Cov}_{x_i x_j}$. Quando i é um índice temporal o estudo das correlações temporais é de grande utilidade em Física para caracterisar a dinâmica de um sistema.

Exercício Pense no significado de cada um dos estimadores de X e da variância Var_X e proponha outros estimadores. Mostre casos em que a moda, a média e a mediana não são iguais.

Exercício Um herói luta com inimigos iguais e sempre com as mesma armas. Cada luta é independente de todas as outras e o heroi tem probabilidade $q = 1 - p$ de ganhar cada luta, Obervamos que ele se aposenta após lutar N vezes, quando é derrotado pela n-ésima vez. Ou seja temos $n = N - k$ derrotas e k vitórias. O problema é estimar p supondo

- (1) Ele pode lutar um número indefinido de lutas, mas só pode perder n vezes até sua aposentadoria. Portanto N é uma variável aleatória.
- (2) Ele só pode lutar um número N de lutas e o número de derrotas n é aleatório.

4
Frequência e Probabilidade

> Professors of probability have been often and justly derided for arguing as if nature were an urn containing black and white balls in fixed proportions. Quetelet once declared in so many words—"l'urne que nous interrogeons, c'est la nature." John Maynard Keynes, Treatise on Probability

Considere as duas frases abaixo

> 1) Acredito que o estudante que chega a este ponto já estudou algo sobre probabilidade.
>
> 2) Amiúde o estudante que chega a este ponto já estudou algo sobre probabilidade.

Parece que ambas dizem essencialmente a mesma coisa. Uma expressa uma crença sobre a história dos estudantes, a outra revela que se verifica algo para os alunos que aqui chegaram. Mas não dizem exatamente a mesma coisa. Poderia ser que o conhecimento da primeira deriva de ter estudado o currículo do secundário e mesmo sem nunca ter visto um estudante, nem uma aproximação, poderiamos ter informação sobre o que estudou. A segunda revela que é frequente encontrar estudantes que já fizeram algo.

A linguagem comum pode ser muito rigorosa e sutil. No entanto outras interpretações poderiam ser dadas às frases. Como essencialmente as frases acima não são verdadeiras tentaremos, dentro do formalismo descrito nos capítulos anteriores, deixar mais claro de que forma a intuição de que são equivalentes é justificada e de que forma não o é.

Até agora nos preocupamos com as regras de manipular probabilidades, mas não lhe atribuimos valores numéricos. Vamos começar por estudar de que forma a informação sobre simetria permite essa atribuição.

Simetria

Um experimento é descrito pela informação contida em I_1= "Suponha que temos uma moeda com duas faces, que descrevemos pela variável $\sigma = \{\pm 1\}$. O valor $\sigma = 1$ está associado à cara e $\sigma = -1$ à coroa. Jogo a moeda para cima, bate no ventilador do teto, e cai num lugar onde não podemos no momento ver o resultado."

Suponha que você, o jogador J_1, jogue contra o jogador J_2. Esta pessoa, por exemplo a Linda, não fala muito bem português e chama

os resultados de Karra e Korroa. Consideremos o seguinte jogo, se $\sigma = 1$ você ganha e ela perde. Do contrário, ela ganha. Ela aposta um feijão. Quanto você estaria disposta a apostar?[1] A resposta tem relação, para pessoas racionais, que não dependem do feijão para sobreviver, com as probabilidades $P(\sigma = 1|IJ_1)$ e $P(\sigma = -1|IJ_1)$ que você atribui com base na informação I que inclui todo o que se sabe sobre a moeda e a forma como foi jogada [2].

É natural supor que voces concordem que

$$\begin{aligned} P(\sigma = 1|IJ_2) &= P(\sigma = 1|IJ_1) \\ P(\sigma = -1|IJ_2) &= P(\sigma = -1|IJ_1). \end{aligned} \tag{4.1}$$

Mas agora descobrimos uma falha enorme de comunicação, o que Linda chama de Karra, você chama de coroa. Voces pensam um pouco e atribuem probabilidades

$$\begin{aligned} P(\sigma = 1|I'J_2) &= P(\sigma = -1|I'J_1) \\ P(\sigma = -1|I'J_2) &= P(\sigma = 1|I'J_1). \end{aligned} \tag{4.2}$$

onde I' descreve o novo estado de informação. Se os jogadores acharem que a nova informação não leva a mudar suas expectativas com respeito à atribuição de probabilidades, ou seja são indiferentes, dirão que os conjuntos de equações 4.1 e 4.2 continuam válidos, mas agora podem ser escritos

$$\begin{aligned} P(\sigma = 1|I''J_2) &= P(\sigma = 1|I''J_1) \\ P(\sigma = -1|I''J_2) &= P(\sigma = -1|I''J_1) \\ P(\sigma = 1|I''J_2) &= P(\sigma = -1|I''J_1) \\ P(\sigma = -1|I''J_2) &= P(\sigma = 1|I''J_1). \end{aligned} \tag{4.3}$$

onde I'' declara que I e I' são equivalentes.

Dado que $P(\sigma = 1|I''J_1) + P(\sigma = -1|I''J_1) = 1$ e que ambos termos são iguais a $P(\sigma = -1|I''J_2)$, devemos concluir que $P(\sigma = 1|I''J_1) = 1/2$ e $P(\sigma = -1|I''J_1) = 1/2$.

Porque tantas voltas para chegar ao óbvio? Por vários motivos. Em primeiro lugar notamos que este não é o único exemplo onde usaremos simetria ou indiferença. A história da Física mostra muitas generalizações do uso de simetria para atribuir probabilidades ou definir a dinâmica, o que não é totalmente diferente, pois dinâmica vem das interações e as interações estão relacionadas, como veremos adiante, com probabilidades condicionais e dependência. A idéia de analisar este caso simples deve-se a que as coisas vão ficar mais difíceis e é interessante se apoiar em casos simples.

Se tivessemos um dado de n faces, com σ tomando valores de 1 a n, teriamos chegado a $P(\sigma = i|I) = 1/n$, a distribuição uniforme. Note que esta atribuição tem a ver com a simetria da nossa informação sobre o experimento do dado e não é postulada *a priori*. Não tem a ver com a simetria do dado. Representar o dado através de um modelo matemático para o cubo perfeito, de densidade uniforme, não passa de uma aproximação. Não é que será difícil, mas é impossível de aproximar na prática. Portanto $1/n$ é devido à simetria de informação e não a simetria física do cubo.

Este método de atribuição de probabilidades parece ter sido usado pela primeira vez por J. Bernoulli e posteriormente por Laplace. Recebe nomes como princípio da razão insuficiente ou da indiferença.

[1] Jaynes não gosta de basear os fundamentos da teoria em algo tão vulgar como apostas por dinheiro. No entanto esperamos que qualquer noção *a priori* sobre apostas tenha evoluido por seleção natural onde as apostas amiúde não são por dinheiro mas sim pela própria vida.

[2] Este problema é talvez muito mais complicado pois não sabemos o que seja uma pessoa racional, mas simplesmente consideremos alguém que quer jogar e quer ganhar, mesmo que isso acabe com objetivos de longo prazo. Definir racionalidade deve passar por estipular uma escala de tempo em que o agente deve maximizar algo que pode ser chamado de *utilidade* ou *felicidade*, mas às vezes na ausência de boas definições, são comumente substituidas por *dinheiro*. Em ciência e em geral nas atividades humanas, perguntas difícies costumam ser substituidas por outras mais simples, à primeira vista parecidas, mas que não necessariamente o são. Veja o livro de D. Kahneman, Thinking fast and slow.

Moedas, Dados, Baralhos, Urnas

Ao longo dos estudos o estudante encontrará sistemas que são simples e portanto estudados muitas vezes. Em dinâmica estudará a partícula livre e o oscilador harmônico, posteriormente o átomo de hidrogênio e o spin de Ising. Em termodinâmica usará caixas rígidas de paredes termicamente isolantes. Nada será tão simples na vida real. Uma partícula nunca está isolada. Nem mesmo o átomo de hidrogênio é um proton e um életron e nada mais. E mesmo assim é desta forma que aprendemos. Aqui a urna, estudada por Bernoulli e Laplace é o sistema simples. Um baralho de cartas ou uma moeda também são sistemas simples e recorrentes, embora nunca sejam de interesse final nas aplicações que nos motivam a estes estudos. Não obstante Quetelet, a urna ideal não tem nada a ver com a natureza. Isto é um exercício e se não soubermos como agir em condições simples não teremos nenhuma chance contra os problemas reais. É um erro grosseiro olhar para um recorte do mundo, achar que é uma urna e depois criticar a teoria de probabilidades por resultados que contradigam o bom senso.

Uma urna ideal é uma bolsa opaca com bolas iguais ao tato. Alguém com uma luva de box fará a extração de uma bola por vez. Há vários jogos que podem ser jogados. O conjunto de bolas pode ter número conhecido ou não. As bolas podem ter cores diferentes e poderemos saber ou não quantas bolas de cada cor estão dentro. Podemos retirar bolas e repó-las ou não, podemos tirar uma bola sem ver que tipo é e proceder a retirar outras. Voce pode retirar a bola de uma urna que eu preparei, ou voce pode ver um mago retirar a bola de uma urna que voce viu enquanto ele a preparava. Há uma fauna enorme de jogos que podem ser feitos e essencialmente em todos, o objetivo é fazer previsões sobre o que pode ocorrer a seguir, ou o que pode ter ocorrido antes. [3]

[3] *Predictions are risky, specially about the future.* Vários autores, alguns sérios outros não. Já a vi atribuida a Bertrand Russell e Niels Bohr mas também a Dan Quayle e Yogi Berra. Não sei se estas atribuições são verdadeiras. O significado de uma frase é condicionado a quem a enunciou.

Urnas

O caso mais simples talvez seja I_1 ="uma urna com N bolas numeradas de $i = 1...N$". Qual é a probabilidade de extrair a bola j? Por simetria de informação é natural associar a mesma probabilidade a cada uma delas. Como são exclusivas e mutuamente exaustivas além de iguais, temos que $P(\text{bola} = j|I) = P(j|I) = 1/N$. Isso é óbvio. Parece até uma imposição da qual não podemos escapar. Mais ainda, uma lei da física. Mas certamente não é.

Suponha que voce jogue contra um mafioso e a bola será extraida por um mágico profissional cuja filha foi raptada pelo mafioso. É claro que você deve suspeitar que as probabilidades das diferentes bolas não devem ser iguais para o mágico nem para o mafioso. Mas e para você? A simetria de sua informação não permite distinguir entre as bolas e não pode ir além de atribuir a mesma probabilidade. Agora você escuta que o mágico sugeriu ao mafioso apostar na bola 17. A informação não é mais simétrica. Tudo isso ocorreu antes de extrair uma bola sequer. A frequência ainda não pode ser definida.

Voltemos ao caso simétrico. I_2 = "Das N bolas M são vermelhas (V) e $K = N - M$ são azuis (A)". Por simplicidade para $1 \leq i \leq M$ as bolas são vermelhas e para $M + 1 \leq i \leq N$ são azuis. Esqueça o mágico, agora acreditamos que a pessoa que realiza extração não é influenciada pela cor da bola. Portanto a probabilidade de extração

de cada bola é igual a $1/N$. Qual é a probabilidade que a bola extraida seja vermelha? O evento "a bola é V"é verdadeiro se a bola extraida tem o número i com $1 \leq i \leq M$. Os eventos "a bola é i"são mutuamente exclusivos, portanto
$V = (i = 1) \vee (i = 2) \vee \cdots \vee (i = M)$ que a bola seja V é a união ou soma de que tenha índice $1 \leq i \leq M$. A regra da soma nos dá

$$\begin{aligned} P(V|I_2) &= \sum_{i=1}^{M} P(i|I_2) = M \times \frac{1}{N} \\ &= \frac{M}{N} \end{aligned} \tag{4.4}$$

Este é um resultado obtido a partir da regra da soma e da simetria de informação sobre as bolas antes de extrair uma única bola. Na seção 2.1.1 vimos que em algum ponto da história isto foi usado como definição de probabilidade por Bernoulli e Laplace [4]. A probabilidade de extração de uma bola V é simplesmente a razão entre os casos "favoráveis"ou vermelhos e o total de casos. O estudante pode achar que já sabia isto e portanto é uma perda de tempo. Deve entender que o objetivo aqui era o de identificar as hipóteses por trás deste resultado trivial e intuitivo. Deve ficar claro que isto não é nenhuma frequência porque ainda não foi retirada uma única bola da urna. Aprender a identificar as hipóteses subjacentes é um dos objetivos do curso. Quando é fácil, quando é inituitivo, quando lembramos de ter escutado falar deste problema no curso primário, parece desnecessário percorrer um caminho longo. Quando o estudante tiver que resolver problemas nunca antes vistos, ou mais interessante ainda, formular novos problemas, o exercício de identificar as hipóteses subjacentes será amiude a única ferramenta disponível. Vemos que a regra M/N é muito retritiva pois se aplica ao caso I_2 e não permite levar em conta a existência de mafiosos nem outra variantes que podem ocorrer na natureza. Portanto não deveria ser tomada como a definição de probabilidades mas simplesmente um resultado obtido a partir das regras de manipulação dos números que representam nossas crenças, obtidas no capítulo 1, para uma experiência realizada sob um conjunto de restrições determinado.

[4] Repetimos: "...probabilidade, que é então simplesmente a fração cujo numerador é o número de casos favoráveis e cujo denominador é o número de todos os casos possíveis."No contexto:

"The theory of chances consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probabilty is sought. The ratio of this number to that of all the cases possible is the measure of this probability which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible."

A Philosophical Essay on Probabilities, Pierre Simon, Marquis de Laplace. 6a ed. F.W.Truscott e F.L. Emory trads.

Urnas: extrações repetidas com reposição.

Extraimos uma bola, que chamamos a primeira, anotamos sua cor e chamamos x_1 que pode ser V ou A. Colocamos a bola novamente, isto é chamado de Reposição. Chacoalhamos a urna. Fazemos isso R vezes e obtemos assim a série $D_R = \{x_1, x_2, ...x_R\}$, que chamamos os Dados (dado=*datum* e dados= *data* em inglés.) Chamaremos R de tamanho da sequência.

Pense e discuta o que significa chacoalhar a urna. Para cada extração estamos nas condições do caso anterior: M, K e N tem o mesmo significado que antes. O resultado de uma extração independe de quais foram as bolas extraidas antes:

$$P(x_n = V|x_1, x_2, ..., x_{n-1} I_2) = P(x_n = V|I_2) = \frac{M}{N} \tag{4.5}$$

Para uma dada sequência usamos a regra do produto

$$\begin{aligned} P(x_1, x_2, ..., x_n|I_2) &= P(x_n|x_1, x_2, ..., x_{n-1}I_2)P(x_1, x_2, ..., x_{n-1}|I_2) \\ &= P(x_n|I_2)P(x_{n-1}|x_1, x_2, ..., x_{n-2}I_2)P(x_1, x_2, ..., x_{n-2}|I_2) \\ &= ... \\ &= P(x_n|I_2)P(x_{n-1}|I_2)...P(x_1|I_2) \\ &= \prod_{i=1}^{n} P(x_i|I_2) \end{aligned} \tag{4.6}$$

Se a sequência for e.g. $VVAAAVV$ teremos

$$P(VVAAAVV|I_2) = ppqqqpp = p^4q^3 \tag{4.7}$$

onde usamos a notação $p = M/N$ e $q = K/N$, com $p + q = 1$. Devido à independência entre os resultados de cada extração, a ordem temporal das ocorrências de vermelho e azul é irrelevante, portanto a única coisa que importa é o número de vezes que na sequência apareceu o vermelho ou que apareceu o azul.

A distribuição binomial

Agora fazemos outra pergunta: independentemente da ordem, qual é a probabilidade de ter m vermelhas e $k = R - m$ azuis (numa extração com reposição de R repetições da extração de uma bola, quando M e K são os números conhecidos de bolas vermelhas e azuis, respectivamente)? É comum dizer de forma equivalente que queremos a distribuição de m sucessos em R tentativas, quando a probabilidade de sucesso é $p = M/N$.

Novamente usaremos as regras da probabilidade. Primeiro as sequências diferentes de R extrações são eventos mutuamente exclusivos. Ou aconteceu uma, ou aconteceu outra, alguma aconteceu e não podem ser duas simultaneamente verdadeiras. Dado R, a probabilidade de obter m bolas vermelhas (e portanto obrigatoriamente k azuis) é obtida da regra da soma, como a soma das probabilidades sobre todas as sequências com m, k. Mas cada sequência tem a mesma probabilidade $p^m q^k$, buscamos portanto o número de sequências com m e k.

O resultado deve ser familiar. Chame o número de sequências de tamanho R com m, k de C_R^m. Considere que já resolvemos o problema para sequências de tamanho $R - 1$, para qualquer $0 \leq m \leq R - 1$. Portanto C_{R-1}^{m-1} e C_{R-1}^m são consideradas conhecidas. Suponha que extraimos $R - 1$ bolas. Há somente duas formas de obter m e k após a retirada da última bola. Isto só pode ocorrer se após $R - 1$ extrações

- (i) tivermos obtido $m - 1$ vermelhas após $R - 1$ extrações e na R-ésima, for extraida uma bola vermelha, o que pode ter ocorrido de C_{R-1}^{m-1} formas ,
- (ii) tivermos obtido m vermelhas após $R - 1$ extrações e na R-ésima, for extraida uma bola azul que pode ter ocorrido de C_{R-1}^m formas diferentes.

Portanto, temos, para $R > 1$ a relação de recorrência para o número de sequências

$$C_R^m = C_{R-1}^{m-1} + C_{R-1}^m, \tag{4.8}$$

que é a famosa relação de recorrência devida a Pascal. Isto é uma máquina de gerar os coeficientes binomiais, que precisa ser

alimentada com valores iniciais. Para $R = 1$ é óbvio que $C_1^0 = C_1^1 = 1$, pois se olharmos sequências de tamanho 1, só há duas possibilidades, a primeira bola foi azul ($C_1^0 = 1$), ou alternativamente foi vermelha ($C_1^1 = 1$).

Usando a notação do fatorial, que é definida pela recursão $n! = n(n-1)!$, para $n = 1, 2, ...$ inteiros positivos, com condições iniciais $0! = 1$ e portanto $n! = 1.2.3....n$, os coeficientes são dados por

$$C_R^m = \frac{R!}{m!(R-m)!}, \tag{4.9}$$

pois satisfazem as relações de recorrência e às condições iniciais. Basta provar unicidade da solução, que é fácil. Note que simplesmente, usando o resultado 4.9, e

$$\frac{R!}{m!(R-m)!} = \frac{(R-1)!}{(m-1)!(R-1-m+1)!} + \frac{(R-1)!}{m!(R-1-m)!} \tag{4.10}$$

temos que a relação 4.8 é satisfeita. Estes coeficientes são chamados os coeficientes binomiais. O motivo disto é que

$$(a+b)^R = \sum_{m=0}^{R} C_R^m a^m b^{R-m}, \tag{4.11}$$

que é amplamente conhecida desde Newton. Mas é instrutivo provar este resultado, supondo-o válido para $R-1$, calculando $(a+b)^{R-1}(a+b)$ e usando a relação de recorrência. A notação $C_R^m = \binom{R}{m}$ também é muito popular e é dito que representa o número de maneiras de escolher m elementos de um total de R ou o número de combinações de R, m a m.

Temos o resultado desejado,

$$P(m|p, R, I_2) = \binom{R}{m} p^m q^{R-m} \tag{4.12}$$

que é a distribuição binomial. Também poderiamos ter escrito $P(m|M, R, I_2)$. Obviamente a distribuição está normalizada, pois

$$\sum_{m=0}^{R} P(m|p, R, I_2) = \sum_{m=0}^{R} \binom{R}{m} p^m q^{R-m} = (p+q)^R = 1 \tag{4.13}$$

Momentos da Binomial

É interessante calcular os valores esperados da distribuição binomial. A expressão da expansão binomial 4.11 escrita com p e q arbitrário é útil para calcular os valores esperados $\langle m \rangle$, $\langle m^2 \rangle$. Usamos a expansão binomial para valores p e q quaisquer, derivamos com respeito a p e multiplicamos por p para obter, usando o truque que $p\frac{\partial}{\partial p} p^m = m p^m$:

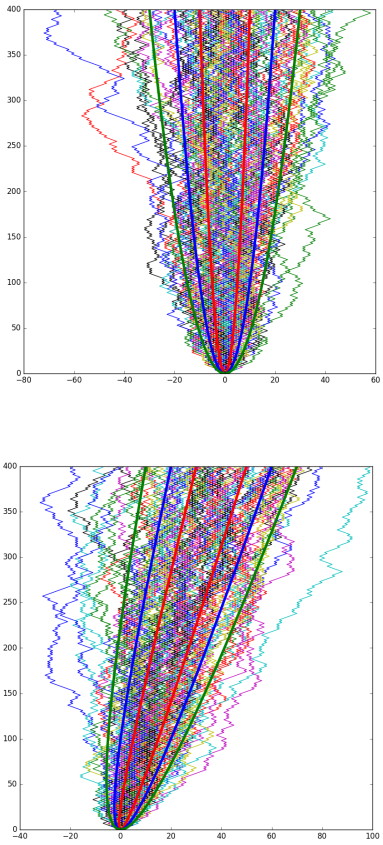


Figura 4.1: Modelo de difusão I: Caminhos aleatórios binomiais, $N = 100$ corridas de $T_{\max} = 400$ passos cada uma. Acima: A cada instante de uma dinâmica discreta, um caminhante dá um passo à direita com $p = 1/2$ ou à esquerda com $q = 1/2$, independentemente de qualquer outra coisa. As parábolas mostram os valores σ, 2σ e 2σ respectivamente, como função do tempo, onde $\sigma = \sqrt{Tp(1-p)}$. Abaixo: $p = 0.55$.

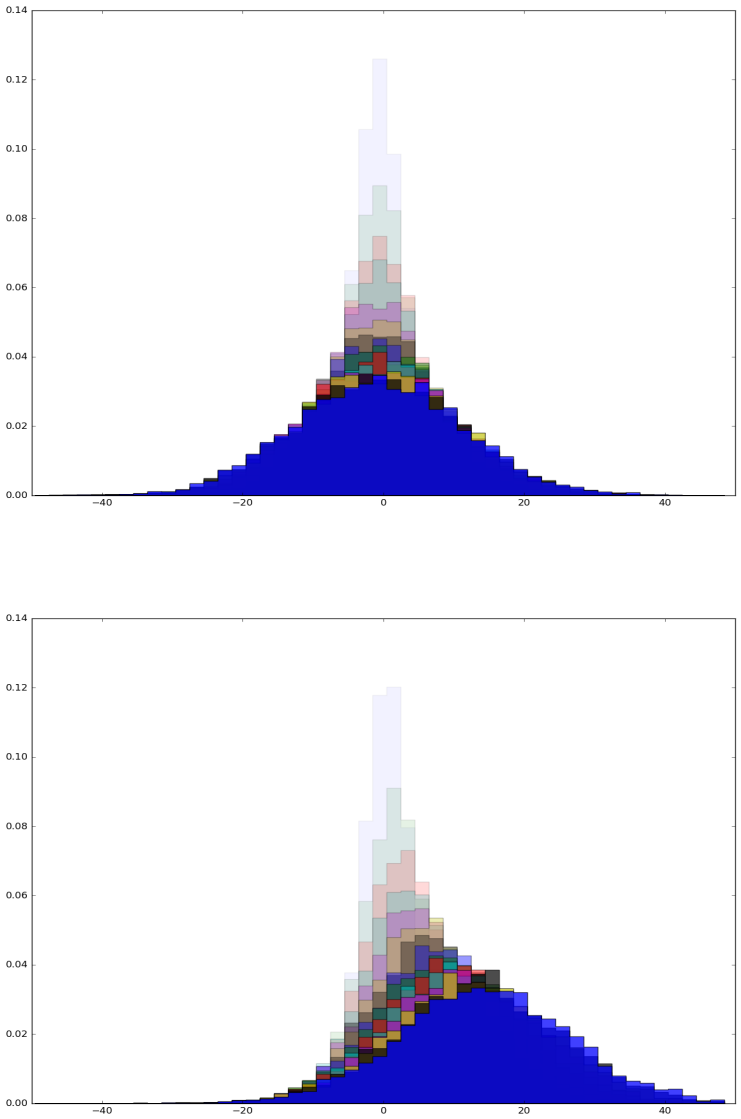


Figura 4.2: Difusão II: Histogramas obtidos dos caminhos simuladas da figura para valores $N = 10, 20....160$ com (Acima) $p = 1/2$, (Abaixo) $p = .55$.

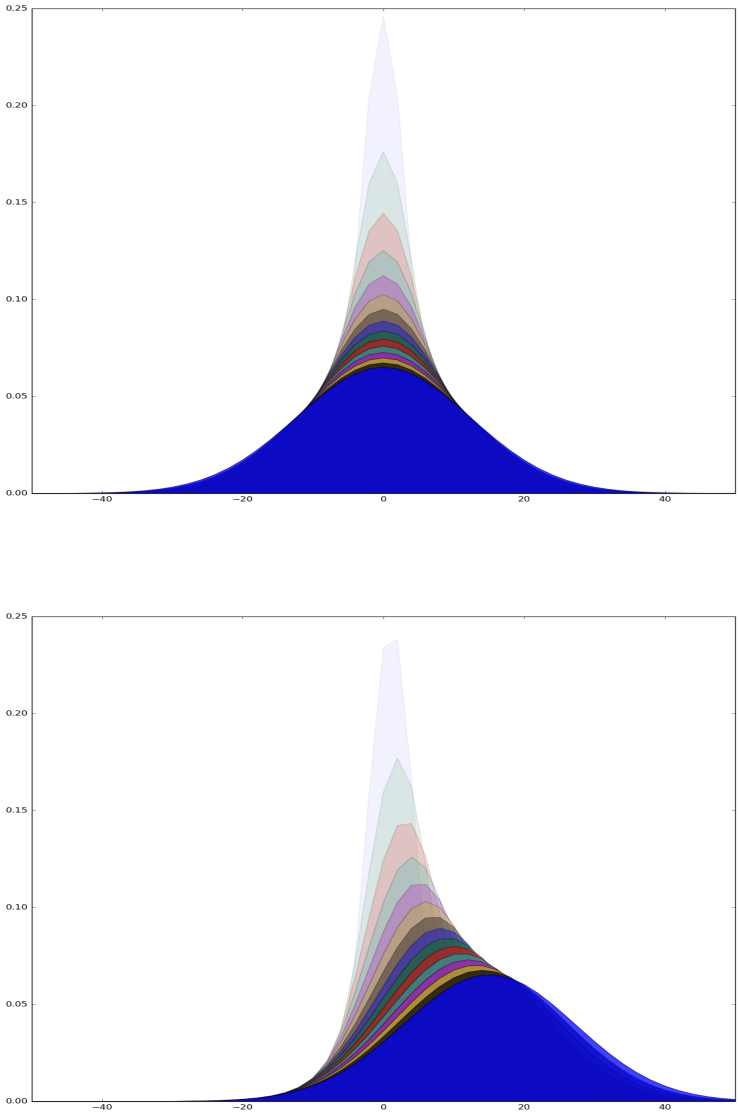


Figura 4.3: Difusão III: A distribuição binomial para valores $N = 10, 20....160$ com (Acima) $p = 1/2$, (Abaixo) $p = .55$.

$$\begin{aligned}
\langle m\rangle &= \sum_{m=0} mP(m|p,R,I_2) = \sum_{m=0}^{R}\binom{R}{m} mp^m(1-p)^{R-m} \\
&= \left(\sum_{m=0}^{R}\binom{R}{m} mp^m q^{R-m}\right)_{q=1-p} \\
&= \left(\sum_{m=0}^{R}\binom{R}{m} (p\frac{\partial}{\partial p}p^m) q^{R-m}\right)_{q=1-p} \\
&= \left(p\frac{\partial}{\partial p}\sum_{m=0}^{R}\binom{R}{m} p^m q^{R-m}\right)_{q=1-p} \\
&= \left(p\frac{\partial}{\partial p}(p+q)^R\right)_{q=1-p} \\
&= = pR(p+1-p)^{R-1} = pR \qquad (4.14)
\end{aligned}$$

O truque vale somente se colocarmos $q = 1-p$ no final[5]. Para calcular $\langle m^2\rangle$ vemos que dentro da soma aparece m^2p^m que podemos escrever como

$$p\frac{\partial}{\partial p}\left(p\frac{\partial}{\partial p}p^m\right) = m^2p^m$$

que permite escrever

$$\langle m^2\rangle = \left[\left(p\frac{\partial}{\partial p}\left(p\frac{\partial}{\partial p}(p+q)^R\right)\right]_{q=1-p}$$

que leva a $\langle m^2\rangle = R^2p^2 + Rp(1-p)$

A variância é comumente denotada $\mathrm{var}(m)$ ou σ^2 ou ainda σ_m^2 e definida por

$$\sigma_m^2 = \langle m^2\rangle - \langle m\rangle^2$$

e portanto para a distribuição binomial de m sucessos em R tentativas a raiz variância ou o desvio padrão é

$$\sigma_m = \sqrt{Rp(1-p)}. \qquad (4.15)$$

Olhe as figuras 7.8 e 4.3. Na primeira são mostradas trajetórias individuais e na segunda as distribuições binomiais para $p = 0.5$ e para $p = 0.55$, para valores de R cada vez maiores. Para $p \neq 0.5$ há deriva. O deslocamento, após R passos dos quais m são para a direita e $R-m$ são para a esquerda é

$$X = m - (R-m) = 2m - R$$

e o valor médio do deslocamento é

$$\mathbb{E}(X) = 2\langle m\rangle - R = (2p-1)R \qquad (4.16)$$

que é positivo para $p > 1/2$.

A comparação entre estas figuras das trajetórias e da distribuição permitirá começar a entender o processo de simulação conhecido como Monte Carlo, onde um processo individual, gerado muitas vezes permite estimar valores esperados de funções de uma variáveis estocásticas cuja distribuição pode ser muito difícil de tratar analiticamente. A raiz quadrada que aparece na equação 4.15 é extremamente importante. Não ocorre por acaso e de forma específica para a binomial. Somamos um número grande de passos gerados por Bernoulli. Toda vez que ocorrer uma soma de variáveis estocásticas, se a variância individual de cada termo for finita e sob condições de independência dos passos (suficiente mas não necessária) a variância crescerá com N e a largura da distribuição com $\sqrt{N}$. Voltaremos a isto no capítulo sobre o Teorema do Limite Central.

[5] É importante notar que a derivada parcial $(\partial f(p,q)/\partial p)_q$ é definida reduzindo a função de duas variáveis a uma função de uma só variável, que é feito ao declarar que q é mantido constante. Se pensarmos na superfície $z = f(p,q)$, notamos que em um dado ponto (p_1,q_1) podemos tomar a derivada em qualquer direção, em particular mantendo $q = q_1$ fixo, ou mantendo $p = p_1$ fixo que dá $(\partial f(p,q)/\partial q)_p$ ou ainda ao longo de qualquer direção, e.g $p = 1-q$, mas os resultados não são os mesmos.

Frequência não é probabilidade

Porque parece razoável confundir frequência e probabilidade? O que segue é importante. A probabilidade de bola vermelha ou de sucesso é p. O valor esperado do número de sucessos é $\langle m \rangle = Rp$, portanto

$$p = \frac{\langle m \rangle}{R} \tag{4.17}$$

ou seja

$$p = \langle \frac{m}{R} \rangle = \langle f \rangle \tag{4.18}$$

onde $f = m/R$ é a frequência de sucessos. Em palavras, o valor esperado da frequência é o parâmetro da binomial que por sua vez é a probabilidade de sucesso. A frequência não é a probabilidade. A frequência é um número que depende do experimento realizado. Isto caracteriza a frequência como um número aleatório. A variância da frequência é

$$\begin{aligned} \sigma_f^2 &= \langle f^2 \rangle - \langle f \rangle^2 \\ &= \langle (\frac{m}{R})^2 \rangle - \langle \frac{m}{R} \rangle^2 \\ &= \frac{1}{R^2}\sigma_m^2 \\ &= \frac{R}{R^2}p(1-p) = \frac{1}{R}p(1-p) \\ \sigma_f &= \frac{1}{\sqrt{R}}\sigma_m \end{aligned} \tag{4.19}$$

Isto significa que embora a frequência seja um número que depende do experimento particular e só o seu valor esperado seja a probabilidade de sucesso, à medida que o número de tentativas R aumenta, seu desvio padrão vai a zero com $1/\sqrt{R}$. Portanto qualquer experimento que meça a frequência encontrará valores perto da probabilidade para R grande o que pode levar alguns de vocês à possibilidade de confundir frequência com probabilidade. Isto porém não é perdoável.

O que significa perto e grande no parágrafo acima será discutido com mais cuidado no capítulo 7, onde faremos estas idéias mais precisas olhando para a desigualdade de Chebyshev e definindo convergência em probabilidade. Seremos, então, capazes de dizer o que significa que f converge para p quando R aumenta. Também olharemos o problema relacionado de inferência de p dada a frequência no capítulo 6

A distribuição Multinomial

Suponha que o processo seja descrito por I_{Multi} ="na urna há N bolas de no máximo C cores, M_c da cor c, $\sum_{c=1...C} M_c = N$. As bolas extraidas são repostas na urna".

Temos, analogamente ao caso de duas cores, que a probabilidade de extrair uma bola de uma cor c é $p_c = M_c/N$. Obviamente $\sum_{c=1...C} p_c = 1$, porque afinal uma bola extraida é de alguma cor. Para uma sequência de N extrações com reposição usamos o fato que as sequências são mutuamente exclusivas e a regra da soma para obter

$$P(m_1, ...m_C | I_{Multi}) = C_N^{m_1, m_2 ... m_C} p_1^{m_1} p_2^{m_2} .... p_C^{m_C}$$

Normalização leva a

$$\sum_{\sum_{m_C=N}} P(m_1,...m_C|I_{Multi}) = 1$$

Supomos novamente que já resolvemos o caso de de $R-1$ extrações e consideramos a extração de mais uma bola. O número total de casos deve satsifazer

$$C_N^{m_1,m_2...m_C} = C_{N-1}^{m_1-1,m_2,...m_C} + C_{N-1}^{m_1,m_2-1,...m_C} + ... + C_{N-1}^{m_1,m_2,...m_C-1} \quad (4.20)$$

onde o termo do lado direito em que aparece $m_c - 1$ é o número de sequências em que faltava uma bola da cor c para chegar ao caso denotado no lado esquerdo: $\{m_1, m_2...m_C\}$ em R extrações. As C condições iniciais $C_1^{0,...0,1,0...0} = 1$ são suficientes para girar a manivela da relação de recorrência 4.20. O resultado é que

$$C_N^{m_1,m_2...m_C} = \frac{N!}{m_1!m_2!....m_C!} \quad (4.21)$$

pois substituindo na relação de recorrência

$$\begin{aligned} C_N^{m_1,m_2...m_C} &\stackrel{?}{=} \sum_c \frac{(N-1)!}{m_1!m_2!..(m_c-1)!..m_C!} \\ &= \frac{\sum_c m_c (N-1)!}{m_1!m_2!...m_C!} \\ &= \frac{N(N-1)!}{m_1!m_2!...m_C!} \\ &= \frac{N!}{m_1!m_2!...m_C!} \end{aligned} \quad (4.22)$$

vemos que 4.20 é de fato satisfeita pelas expressões 4.21. Verifique que as condições iniciais são satisfeitas. Falta provar unicidade. Mas isso é simples e é deixado para os leitores interessados.

Urnas sem reposição: a distribuição hípergeometrica.

A diferença fundamental com relação aos casos anteriores é que vale $I_4 =$ "a extração de cada bola é feita sem reposição das anteriores, (inicialmente N bolas, M vermelhas)"e portanto em condições diferentes das anteriores. A primeira extração é igual ao caso anterior

$$P(x_1 = V|N, M, I_4) = \frac{M}{N}$$

Agora a diferença, no segundo passo o estado da urna e portanto as probabilidades dependem do resultado da primeira extração

$$P(x_2, x_1|N, M, I_4) = P(x_2|x_1, M, N, I_4)P(x_1|N, M, I_4)$$

Se as duas forem vermelhas, teremos

$$\begin{aligned} P(x_2 = V, x_1 = V|N, M, I_4) &= P(x_2 = V|x_1 = V, M, N, I_4)P(x_1 = V|N, M, I_4) \\ &= P(x_2|N-1, M-1, I_4)P(x_1 = V|N, M, I_4) \\ &= \frac{M-1}{N-1}\frac{M}{N}, \end{aligned} \quad (4.23)$$

pois na segunda extração há somente $N-1$ bolas, das quais $M-1$ são vermelhas. A probabilidade que as primeira r bolas extraidas

sejam vermelhas é

$$\begin{aligned} P(x_r = V, ...x_2 = V, x_1 = V|N, M, I_4) &= \frac{(M-r-1)...(M-1)M}{(N-r-1)...(N-1)N} \\ &= \frac{M!(N-r)!}{(M-r)!N!} \end{aligned} \tag{4.24}$$

que faz sentido mesmo que $r > M$ se for convencionado que o fatorial de números negativos é infinito. Continuamos, mas agora calculamos as probabilidades que as bolas seguintes sejam azuis. O estado da urna é de $N - r$ bolas, das quais $M - r$ são vermelhas, e a probabilidade de extrair uma bola azul é:

$$P(x_{r+1} = A|N - r, M - r, I_4) = \frac{N - r - (M - r)}{N - r} = \frac{N - M}{N - r}.$$

Repetindo

$$P(x_{r+b} = A, ..., x_{r+1} = A|N - r, M - r, I_4) = \frac{(N - M)!(N - r - b)!}{(N - M - b)!(N - r)!}.$$

Assim chegamos a que uma sequência de r vermelhas seguidas por b azuis tem probabilidade, pela regra do produto

$$\begin{aligned} P(x_{r+b} = A, ...x_{r+1} = A, x_r = V, ...x_1 = V|N, M, I_4) &= P(x_r = V, ...x_1 = V|N, M, I_4) \\ &\times \quad P(x_{r+b} = A, ...x_{r+1} = A|x_r = V, ...x_1 = V, N, M, I_4) \end{aligned}$$

que pode ser escrito como

$$\begin{aligned} &= \frac{M!(N-r)!}{(M-r)!N!} \frac{(N-M)!(N-r-b)!}{(N-M-b)!(N-r)!} \\ &= \frac{M!}{(M-r)!N!} \frac{(N-M)!(N-r-b)!}{(N-M-b)!}. \end{aligned}$$

Note que os fatoriais são de

- N e $(N - r - b)$ os números inicial e final de bolas na urna
- M e $N - M$, os números iniciais de bolas vermelhas e de azuis.
- $M - r$ e $(N - M - b)$ os números finais de bolas vermelhas e de azuis.

Ou seja não aparece nada que diga a ordem em que foram extraidas, primeiro as vermelhas depois as azuis. Isto deve ser verdade para qualquer ordem de extração,desde que os resultado finais de extração r e b sejam os mesmos. Vejamos se é assim. Suponha que numa sequência S_1 de $r + b$ a extração da k-ésima bola vermelha ocorreu na posição l e da k'-ésima bola azul na $l + 1$, e na sequência S_2 a k-esima bola vermelha foi extraida após $l + 1$ extrações e a k'-ésima bola azul após l. Aparte dessa troca, as sequências são iguais. Os fatores que contribuem à probabilidade são para a sequência S_1

$$\cdots \frac{M-k-1}{N-l} \frac{N-M-k'-1}{N-l-1} \cdots$$

e para a sequência S_2

$$\cdots \frac{N-M-k'-1}{N-l} \frac{M-k-1}{N-l-1} \cdots$$

que são iguais. Seque que a probabilidade de extrair r bolas vermelhas e b azuis, independentemente da ordem é dada pelo

produto do número de sequências possíveis, $\binom{r+b}{b}$ e da probabilidade de uma sequência:

$$P(\{r,b\}|N,M,I_4) = \frac{(r+b)!}{r!b!}\frac{M!}{(M-r)!N!}\frac{(N-M)!(N-r-b)!}{(N-M-b)!}$$

e simplificando, a probabilidade ao "extrair sem reposição $r+b$ bolas de uma urna com N bolas das quais M são vermelhas e $N-M$ azuis,exatamente r sejam vermelhas"é

$$P(\{r,b\}|N,M,I_4) = \binom{r+b}{r}\frac{\binom{N-r-b}{M-r}}{\binom{N}{M}} \tag{4.25}$$

É interessante que isto pode ser escrito como

$$P(\{r,b\}|N,M,I_4) = \frac{\binom{M}{r}\binom{N-M}{b}}{\binom{N}{r+b}} \tag{4.26}$$

onde o numerador é obtido pelo produto de todas as diferentes combinações de escolhas possíveis de r bolas do total de M vermelhas vezes o número de combinações de b do total de $N-M$ azuis, dividido pelo total de possibilidades das combinações de $r+b$ do total de N bolas. Podemos ainda escrever a mesma expressão de uma forma que fica simétrica e permite generalização para mais cores. Mudando a notação chamamos de M_1 (em lugar de M) o número de bolas vermelhas, M_2 o de azuis; de r_1 o número de bolas da primeira cor, de r_2 o da segunda cor:

$$P(\{r_1,r_2\}|M_1,M_2,I_4) := P(\{r,b\}|N,M,I_4) = \frac{\binom{M_1}{r_1}\binom{M_2}{r_2}}{\binom{M_1+M_2}{r_1+r_2}}. \tag{4.27}$$

É razoável supor, e facilmente demonstrável para o caso de C cores:

$$P(\{r_1,r_2...r_C\}|M_1,M_2,...M_C,I_4) = \frac{\prod_{c=1...C}\binom{M_c}{r_c}}{\binom{\sum_{c=1...C}M_c}{\sum_{c=1...C}r_c}} \tag{4.28}$$

Os caminhos hipergeométricos para urnas estão mostrados nas figuras 4.4, 4.5 e 4.6. Devido a que a urna não volta ao mesmo estado após a extração as figuras são diferentes dos caminhos binomiais. Há uma difusão inicial, mas as trajetórias convergem para o mesmo lugar. Não importa a história de extração, a urna vazia será a mesma em todos os casos.

Bolas escondidas

Voltemos ao caso sem reposição. N bolas, M vermelhas, $N-M$ azuis. Extraimos uma bola mas (agora a diferença) não somos informados da sua cor. A bola é escondida fora da urna. Qual é a probabilidade $P(x_2 = V|N,M,I_4)$ que a segunda bola seja vermelha? O interesse nestes casos está no método, não no jogo em si. As regras da probabilidade são suficientes para responder isto. Tivemos duas extrações portanto o nosso interesse deve começar por analisar a distribuição conjunta $P(x_2,x_1|N,M,I_4)$. A única coisa que sabemos sobre x_1 é que foi vermelho ou azul, possibilidades excusivas e exaustivas. Portanto

$$\begin{aligned} P(x_2|N,M,I_4) &= \sum_{x_1=V,A} P(x_2,x_1|N,M,I_4) \\ &= \sum_{x_1=V,A} P(x_2|x_1,N,M,I_4)P(x_1|N,M,I_4) \end{aligned} \tag{4.29}$$

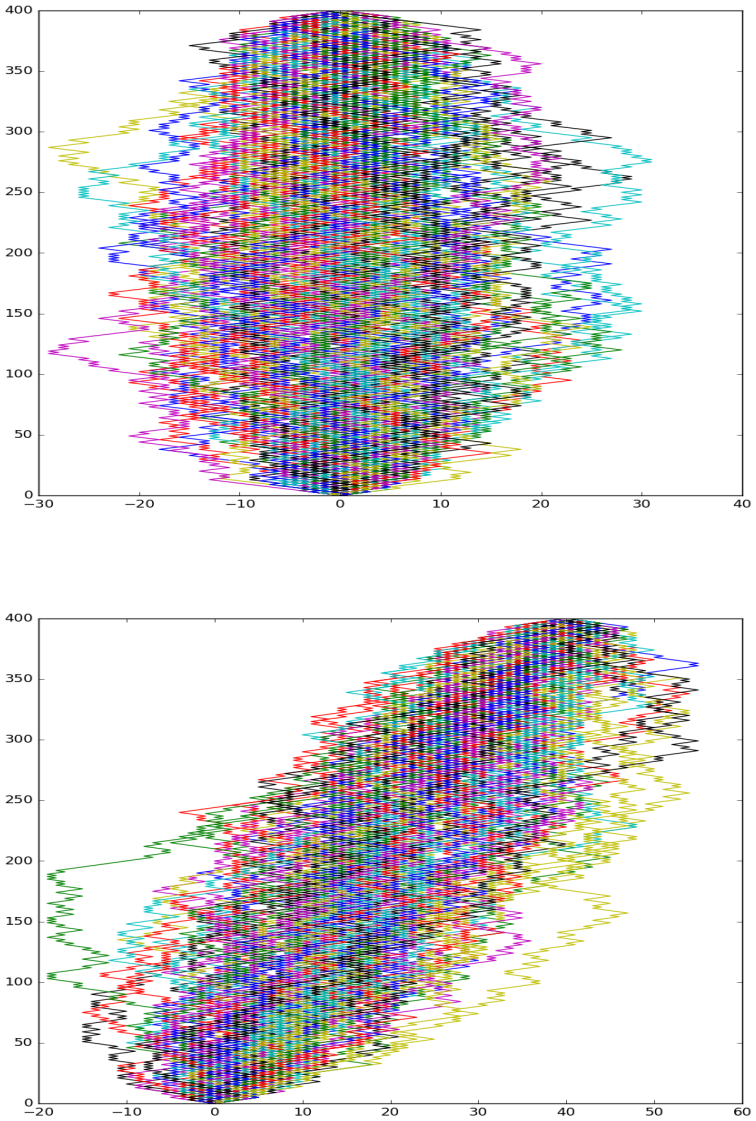


Figura 4.4: Caminhos hipergeométricos. Acima: Urna com $N = 400$ bolas das quais $M_1 = 200$ são vermelhas e $M_2 = 200$ azuis. Abaixo: Urna com $N = 400$, $M_1 = 220$ vermelhas e $M_2 = 180$ azuis. A cada bola vermelha extraida o caminhante anda para a direita, a cada bola azul, para a esquerda.

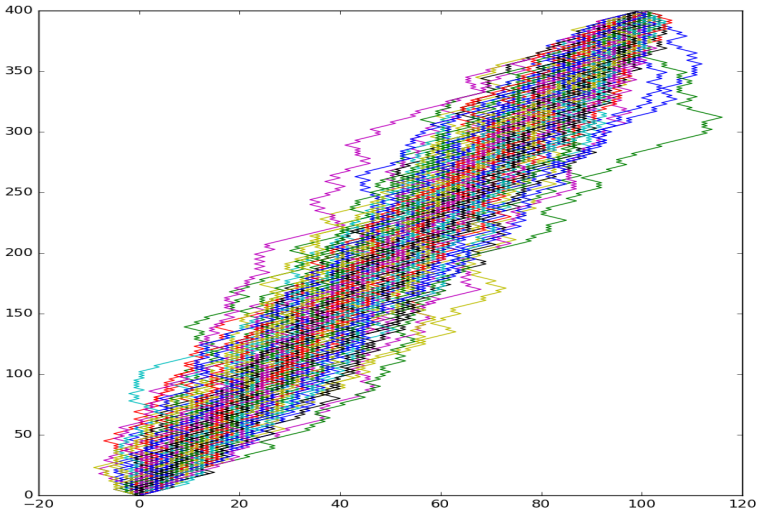


Figura 4.5: Caminhos hipergeométricos. Trajetórias para a urna com $N = 400$, $M_1 = 250$ vermelhas e $M_2 = 150$ azuis.

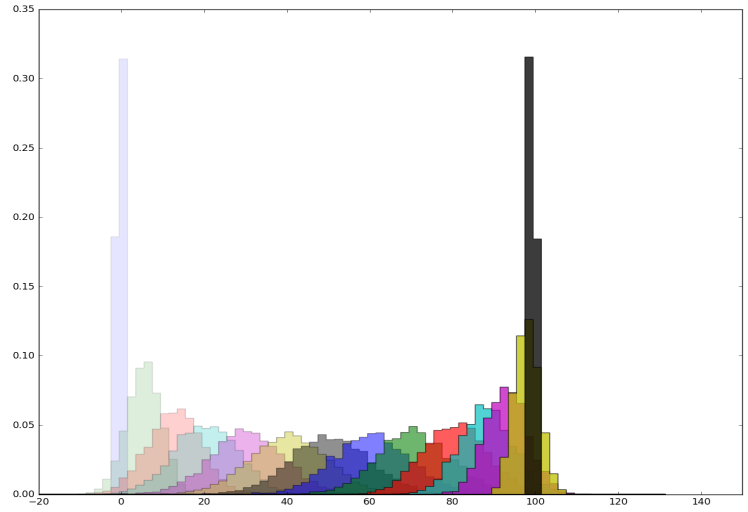


Figura 4.6: Histogramas dos caminhos hipergeométricos: simulaçao. Urna com $M = 400$, $M_1 = 250$ vermelhas e $M_2 = 150$ azuis. A situação é a mesma da figura anterior. Abaixo: Os histogramas foram gerados após extrair(1,20,50,80,120,160,200,240,280,320,350,370,390,399) bolas e olhar os resultados para 5000 urnas.

que fica escrita em termos de probabilidades que conhecemos. Isto é um exemplo ao contrário do uso de marginalização. Portanto

$$\begin{aligned} P(x_2 = V|N, M, I_4) &= P(x_2 = V|x_1 = V, N, M, I_4)P(x_1 = V|N, M, I_4) \\ &+ P(x_2 = V|x_1 = A, N, M, I_4)P(x_1 = A|N, M, I_4) \\ &= \frac{M-1}{N-1}\frac{M}{N} + \frac{M}{N-1}\frac{N-M}{N} \\ &= \frac{M}{N}. \end{aligned} \tag{4.30}$$

Ou seja, como a primeira extração não nos deu nenhuma informação, a probabilidade de extração no segundo passo continuou sendo M/N.

Podemos pensar sobre o que acontece se as duas primeiras bolas forem escondidas. O mesmo. Se não há informação não há alteração de probabilidades. Mas suponha que extraimos e escondemos uma bola. Extraimos uma segunda e é vermelha. O que isto nos diz sobre a bola escondida? Queremos saber sobre $P(x_1|x_2, N, M, I_4)$. Voltamos a pensar sobre a distribuição conjunta e usamos novamente a regra do produto

$$P(x_1|x_2, N, M, I_4) = \frac{P(x_2, x_1|N, M, I_4)}{P(x_2|N, M, I_4)}.$$

Especificamente, suponha que a segunda bola é vermelha, qual é a probabilidade que a primeira seja vermelha:

$$\begin{aligned} P(x_1 = V|x_2 = V, N, M, I_4) &= \frac{P(x_2 = V, x_1 = V|N, M, I_4)}{P(x_2 = V|N, M, I_4)} \\ &= \frac{\frac{M-1}{N-1}\frac{M}{N}}{\frac{M}{N}} = \frac{M-1}{N-1}, \end{aligned}$$

confirmando o que talvez podia ser desconfiado a o recuperar a probabilidade de extração de uma segunda bola conhecendo o resultado da primeira. Note então que o que é primeiro e o que é segundo não interessa. O que interessa é que informação está disponível. Se não há informação nenhuma é equivalente a uma primeira extração de bola, se há informação é equivalente a uma segunda extração sabendo a primeira.

Inversão: Urna com conteúdo desconhecido

Um problema em ciência pode ser descrito como "conhecido o sistema, que previsões podemos fazer sobre o resultado de experiências?"Outro tipo de problema é o inverso, "sabendo o resultado das experiências, o que podemos dizer sobre um sistema desconhecido?"

Considere que o conteúdo da urna é desconhecido e retiramos bolas com reposição. Como resposição significa que em cada extração o estado da urna é o mesmo, mesmo que o nosso estado de informação tenha mudado. Em um dado ponto temos um conjunto de dados, $D_R = \{V, V, V, A, A..\}$. O que podemos dizer sobre a fração de cores? Este tipo de problema constitui o tópico central do problema de análise de dados experimentais e inclui a idéia fundamental de modelo. Voltaremos com mais detalhes, uma e outra vez, ao longo destas notas. Precisamos definir a informação subjacente I_5. Consideramos que há somente C cores, cada cor com número M_c de bolas, $N = \sum_c M_c$ o número total de bolas. Portanto a probabilidade

de extração de uma bola de cor c seria $p_c = M_c/N$. Acabamos de ver que saberiamos calcular a probabilidade de qualquer sequência dados os p_c. Agora usamos a regra do produto, o truque que não parará de dar resultados. Por facilidade olhemos o caso de duas cores, teremos $p = M/N$ como parâmetro desconhecido. Obtivemos a distribuição binomial 4.12, para m bolas vermelhas em R extrações, quando a fração de bolas vermelhas é p:

$$P(m|p, R, I_2) = \binom{R}{m} p^m q^{R-m}. \tag{4.31}$$

A regra do produto nos dá para a distribuição conjunta de m e p

$$P(p, m|R, I_2) = P(p|R, I_2)P(m|p, R, I_2) = P(m|R, I_2)P(p|m, R, I_2), \tag{4.32}$$

de onde temos o resultado conhecido como a regra de Bayes

$$P(p|m, R, I_2) = \frac{P(p|R, I_2)P(m|p, R, I_2)}{P(m|R, I_2)}. \tag{4.33}$$

Aqui aparece algo novo. Vale a pena respirar e tomar o tempo necessário para assimilar algo que será fundamental no que segue. Temos a probabilidade do parâmetro da binomial, que por sua vez é uma probabilidade. Além disso temos duas probabilidades de p,

- A distribuição *a priori* $P(p|R, I_2)$
- A distribuição posterior $P(p|m, R, I_2)$. Posterior à inclusão da m nos condicionantes.

Ainda temos $P(m|p, R, I_2)$ que codifica a informação que temos sobre quão provável é um valor de m caso p tenha um valor dado. Esta probabilidade recebe o nome de verossimilhança. O mundo se divide em pessoas que ficam nervosas ao falar da probabilidade de p e aqueles que acham natural falar da probabilidade deste parâmetro. Claro que p é uma probabilidade, mas se lembrarmos que é a razão entre bolas vermelhas e o total, não há motivo para nervosismo. Ainda ficam mais nervosos ao falar da probabilidade de *a priori* - antes de levar em consideração os dados. As questões levantadas aqui serão atacadas no capítulo 6.

A regra de sucessão de Laplace

O que vem a seguir é interessante por pelo menos dois motivos. Primeiro porque mostra a aplicação dos métodos desenvolvidos a um problema de urna interessante, onde as hipóteses ficam claras, pois senão não é possível fazer as contas. O segundo é histórico. A previsão feita é usada agora em problemas que não tem nada a ver com as hipóteses e se chega a algo que viola as expectativas do bom senso. Para alguns autores isto é indicação que as regras da probabilidade usadas por Laplace não fazem sentido. Isso tem acontecido e a discussão sobre porque isto ocorre e como evitar este tipo de procedimento é instrutivo para o aluno. Faço hipóteses, calculo um resultado, aplico em outro problema onde a informação é diferente e portanto espero resultados diferentes, e como os dois não batem critico a teoria. Parece mais política que ciência.

Consideremos uma urna de composição desconhecida, exceto por ter bolas de somente duas possibilidades de cores e procedemos a extrações com reposição. A reposição significa que cada extração é

independente e em condições idênticas às anteriores. Outra forma de colocar o problema é considerando um processo de Bernoulli, dois estados $s = 1$ ou $s = -1$, ou sucesso e fracasso. Não sabemos o valor parâmetro p. Os dois casos são idênticos se na urna o número de bolas for infinito.

As asserções relevantes para o problema são as seguintes:

- N= "foram feitas N tentativas consecutivas de Bernoulli"
- n= "dado N, foram obtidos n sucessos"
- M= "foram feitas M tentativas consecutivas de Bernoulli"
- m= "dado M, foram obtidos m sucessos"
- I= descrição do processo

O fato de usar o mesmo símbolo para um número e uma asserção deve ser perdoado por simplificar a notação.

O objetivo do exercício é determinar com base em um primeiro experimento descrito por N e n, qual é a probabilidade $P(m|nMN)$ de obter m após M.

Começamos por identificar o que não sabemos, m e p, a probabilidade de sucesso, que é o parâmetro da binomial. Como dissemos na secção anterior alguns autores tentam evitar falar de probabilidade de uma probabilidade, enfatizarmos que p é um parâmetro de uma distrbuição, logo não deve haver resistência à sua estimativa e representação através de distribuições que codifiquem o que sabemos. Portanto estamos interessados na distibuição de probabilidades conjunta de m e p dado o que sabemos:$P(m, p|nMNI)$. Mas não estamos interessados em p, e portanto marginalizamos

$$P(m|nMNI) = \int_0^1 P(m, p|nMNI)dp.$$

A regra do produto leva a

$$P(m|nMNI) = \int_0^1 P(m|pnMNI)P(p|nMNI)dp. \tag{4.34}$$

Jogar M vezes o jogo sem saber o resultado não dá informação sobre p, portanto $P(p|nMNI) = P(p|nNI)$. Como sabemos que a probabilidade de obtenção de n sucessos em N tentativas é uma binomial $P(n|pNI)$ e podemos usar Bayes para inverter:

$$P(p|nNI) = \frac{P(p|NI)P(n|pNI)}{P(n|NI)} \tag{4.35}$$

O denominador $P(n|NI)$ pode ser obtido por normalização, portanto não nos preocupa. Novamente, é irrelevante saber N e não saber n, portanto temos $P(p|NI) = P(p|I)$ para o *a priori*. Fazemos a suposição que não temos, antes de ver os dados, nenhuma preferência por qualquer valor de p, portanto $P(p|I) = 1$, é a distribuição uniforme.

$$\begin{aligned} P(p|nNI) &\propto P(n|pNI) \propto p^n(1-p)^{N-n} \\ &= \frac{p^n(1-p)^{N-n}}{\int_0^1 p'^n(1-p')^{N-n}dp'} \\ &= \frac{(N+1)!}{n!(N-n)!}p^n(1-p)^{N-n}, \end{aligned} \tag{4.36}$$

que reconhecemos como a distribuição Beta(n, N) de p após n sucessos em N tentativas. Para a normalização usamos o resultado devido a Euler, ver equação 3.18

$$E_k^r = \int_0^1 p^r (1-p)^k dp \quad = \quad \frac{r!k!}{(r+k+1)!} \tag{4.37}$$

Voltamos ao cálculo de 4.34,notando que $P(m|pnMNI) = P(m|pMI)$, pois saber p torna desncessária a informação de n, N,

$$\begin{aligned} P(m|nMNI) &= \int_0^1 P(m|pMI)P(p|nNI)dp \\ &= \binom{M}{m} \frac{(N+1)!}{n!(N-n)!} \int p^m (1-p)^{M-m} p^n (1-p)^{N-n} dp \\ &= \binom{M}{m} \frac{(N+1)!}{n!(N-n)!} E_{N+M-n-m}^{m+n} \\ &= \frac{M!}{m!(M-m)!} \frac{(N+1)!}{n!(N-n)!} \frac{(m+n)!(N+M-n-m)!}{(N+M+1)!} \\ &= \binom{n+m}{n} \binom{N+M-n-m}{N-n} \frac{1}{\binom{N+M+1}{N+1}} \end{aligned}$$

Esta expressão horrível pode ser simplificada em casos particulares. Por exemplo, Lapalce considerou o caso em que após N eventos com n sucessos, queremos a probabilidade de $m = 1$ sucesso em $M = 1$ tentativas.

$$\begin{aligned} P(m=1|n, M=1, N, I) &= \binom{n+1}{n} \binom{N-n}{N-n} \frac{1}{\binom{N+2}{N+1}} \\ &= \frac{(n+1)!}{n!} \frac{(N+1)!}{(N+2)!} \\ &= \frac{n+1}{N+2} \end{aligned} \tag{4.38}$$

No caso particular, mas que concentrou a atenção de estudiosos por séculos, onde temos $n = N$ sucessos em N tentativas, a probabilidade de que a próxima seja um sucesso é

$$\begin{aligned} P(m=1|n=N, M=1, N, I) &= \binom{n+1}{n} \binom{N-n}{N-n} \frac{1}{\binom{N+2}{N+1}} \\ &= \frac{N+1}{N+2}. \end{aligned}$$

Este resultado recebe o nome de *regra da sucessão*. Aqui Laplace cometeu o seu maior erro, não no uso das regras da probabilidade nem de contas. Simplemente fez uma piada que foi mal entendida por muitos estudiosos que o seguiram. A estimativa bíblica da idade do universo era da ordem de 5000 anos $\approx 1.82613 \times 10^6$ dias. Em todos esses dias nasceu o sol. Qual seria a probabilidade de que o sol saisse amanhã? Pela regra da sucessão 4.38, seria $1 - 5. \times 10^{-7} = 0.9999995$. A chance de sair seria 182614 vezes maior que a de não sair. Na frase seguinte à piada, retomando um aspecto mais sério, disse que [6]

> Mas este número é incomparavelmente maior para ele que, reconhecendo na totalidade dos fenômenos o principal regulador dos dias e estações, visto que nada no momento presente pode deter a sua marcha"
>
> Laplace.

[6] "Mais ce nombre est incomparablement plus fort pour celui qui connaissant par l'ensemble des phénomènes, le principle régulateur des jours et des saisons, voit que rien dans le moment actuel, ne peut en arrêter le cours". Essai philosophique sur les probabilités Laplace

Isto significa que deve ficar claro ao usuário, que se tiver mais informação, no caso do sol todo o conhecimento de Dinâmica e Astronômia, deve por todos os meios usá-la. O cálculo acima então não se deveria aplicar a não ser a situações onde se deve aplicar: àquelas em que as hipóteses são justificáveis. Os críticos à regra da sucessão por dizer que dá resultados ridículos para a saida do sol amanhã, devem responder se acham natural dizer que tudo o que sabemos sobre o sol, e o que significa que ele sairá, pode ser descrito como uma urna com dois tipos de bolas, pretos e brancos. Mas se você usar frequência como definição de probabilidade pode estar tentado a dizer que o sol sempre sairá, pois sempre saiu. Mas isto é igualmente ridículo, pois temos informação, na forma de teorias de evolução estelar que isso mudará.

Outra crítica é sobre o uso da distribuição *a priori* uniforme. Retomaremos o efeito da distribuição *a priori* no capítulo 6. As mudanças para distribuições razoáveis mudam pouco. A queixa em partícular é que poderiamos fazer uma mudança não linear de variáveis e o que é uniforme agora deixaria de ser. Mas ao falar de urnas, parece natural falar do parâmetro p e se não há preferências *apriori* para acreditar num estado da urna, a uniforme parecebem justificada. Obviamente aquele que tiver informação diferente terá que fazer outras escolhas. Outras distribuições *a priori* podem e devem ser usadas, em outras condições de informação.

Poisson: um limite da binomial

Suponha que em um experimento temos N partículas que podem decair em um dado intervalo de tempo Δt e uma probabilidade p de detectar o resultado do decaimento da particula. O tempo morto do detector é nulo. O número m de sucessos, ou detecções em Δt é dado pela binomial

$$P(m|p,N,I_2) = \binom{N}{m} p^m (1-p)^{N-m} \tag{4.39}$$

Queremos tomar o limite de N muito grande, p muito pequeno. Há várias formas de fazé-lo, um resultado extremamente útil é quando

$$N \to \infty, p \to 0, Np \to \lambda = \text{constante}$$

pois, considerando que

$$\begin{aligned} p^m \frac{N!}{(N-m)!} &= p^m N(N-1)(N-2)\cdots(N-m+1) \\ &= pN(1-\frac{1}{N})pN(1-\frac{2}{N})\cdots pN(1-\frac{m-1}{N})pN \\ &\to \lambda^m \end{aligned} \tag{4.40}$$

e

$$\begin{aligned} (1-p)^{N-m} &= (1-\frac{\lambda}{N})^{N-m} \approx (1-\frac{\lambda}{N})^N \\ &\to e^{-\lambda}. \end{aligned}$$

Temos então a distribuição de Poisson (que talvez deveria também ter o nome de de Moivre)

$$
\begin{aligned}
P(m|p,N,I_2) &= \frac{N!}{m!(N-m)!}p^m(1-p)^{N-m} \\
&= \left(\frac{1}{m!}\right)\left(\frac{N!}{(N-m)!}p^m\right)\left((1-p)^{N-m}\right) \\
&\rightarrow \quad P(m|\lambda) = \frac{\lambda^m}{m!}e^{-\lambda}. \qquad (4.41)
\end{aligned}
$$

Lembramos que o valor médio

$$\langle m\rangle = \lambda, \qquad (4.42)$$

e o segundo momento

$$\langle m^2\rangle = \lambda + \lambda^2, \qquad (4.43)$$

que leva à variância

$$\sigma^2_{Poisson} = \lambda. \qquad (4.44)$$

Para calcular momentos superiores podemos usar

$$\lambda\frac{\partial}{\partial\lambda}P(m|\lambda) = -\lambda P(m|\lambda) + mP(m|\lambda) \qquad (4.45)$$

pois teremos, multiplicando por m^k e somando sobre m:

$$
\begin{aligned}
\langle m^{k+1}\rangle &= \lambda\langle m^k\rangle + \sum_m \lambda\frac{\partial}{\partial\lambda}m^k P(m|\lambda) \\
&= \lambda\langle m^k\rangle + \lambda\frac{\partial}{\partial\lambda}\langle m^k\rangle \qquad (4.46)
\end{aligned}
$$

Volteremos a falar desta distribuição ao analisar dados experimentais.

Sequências Imaginadas, mãos quentes e falácia do jogador.

É fácil imaginar o experimento de lançar uma moeda. Jogo a moeda bem para o alto, bate no teto e cai no chão. Observo e anoto o resultado. Agora surge a pergunta: é fácil imaginar um segundo lançamento? Parece fácil. Se o primeiro lançamento foi, porque não seria o segundo? E cem lançamentos? Este problema foi proposto aos estudantes do primeiro curso de Probabilidades no IFUSP em 2016. OS dados brutos são apresentados na figura

Notamos imediatamente que os dados gerados por pessoas não seguem o modelo pedido. Uma pequena deriva à direita nos dados é compatível com $p = .52$ poderia ser vista nos dados, na figura 4.7, mas ainda não temos instrumentos para estimar isto. Os histogramas dos dados e da binomial são bem diferentes, figura 4.8. O histograma dos dados é muito mais estreito que o da binomial. A figura 4.9 mostra um característica interessante do processo. Escolhemos um ponto qualquer na sequencia e perguntamos se os próximos $k-1$ tem o mesmo símbolo. Isto é, perguntamos se um dado sítio numa trajetoria é seguido por símbolos semelhantes de forma a que há uma sequência de k repetições. Fazemos isso para todos os sítios, e fazemos isso para k de 1 até 11. O logaritmo da razão f_I entre todas as vezes que ocorre e o total de símbolos KN aparece nas ordenadas da figura 4.9. Para a binomial, é simples de calcular, a razão vai como $f_B = p^{k-1}$, pois as jogadas são independentes (círculos). Mas a linha de baixo (quadrados), para os dados, mostra que há uma repressão sistemática por parte de uma pessoa em compensar uma sequência e

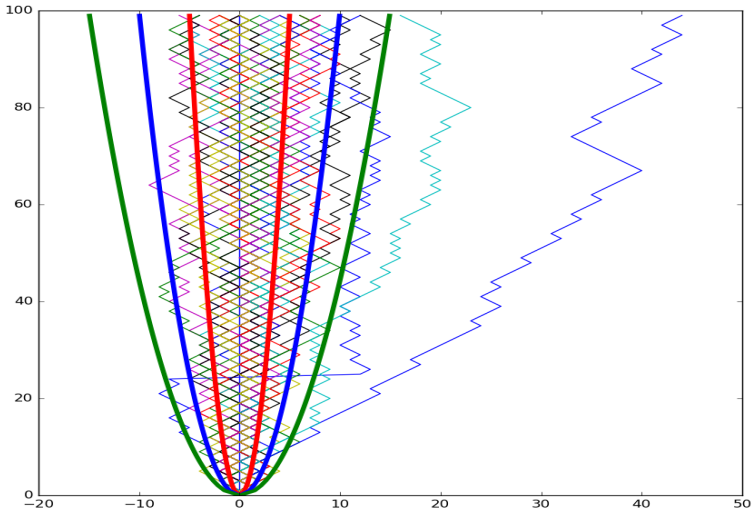


Figura 4.7: As trajetórias imaginadas por cada estudante. $K = 40$ estudantes responderam . As curvas sólidas são os desvios padrão $(\sigma, 2\sigma, 3\sigma)$ que uma binomial com $p = 0.5$ teria após $N = 100$ passos.

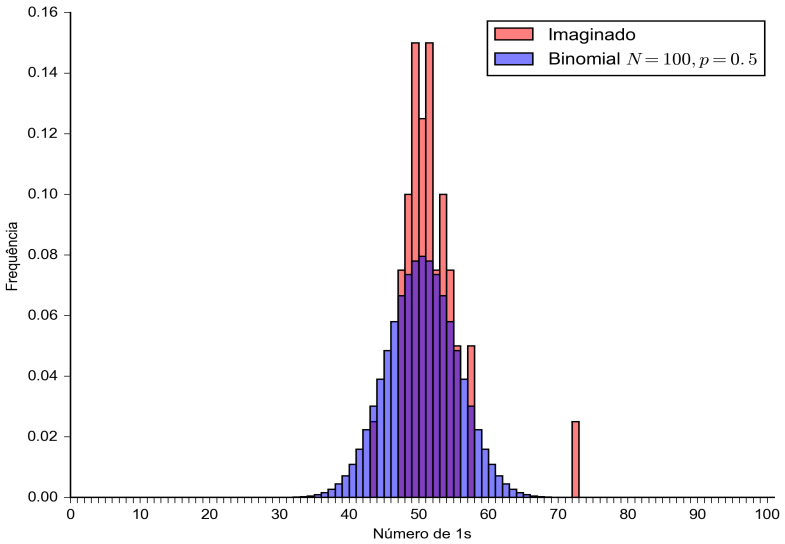


Figura 4.8: Histogramas do número de caras nos dados imaginados e o histograma do binomial simulado. $K = 40$ trajetorias de $N = 100$ jogadas.

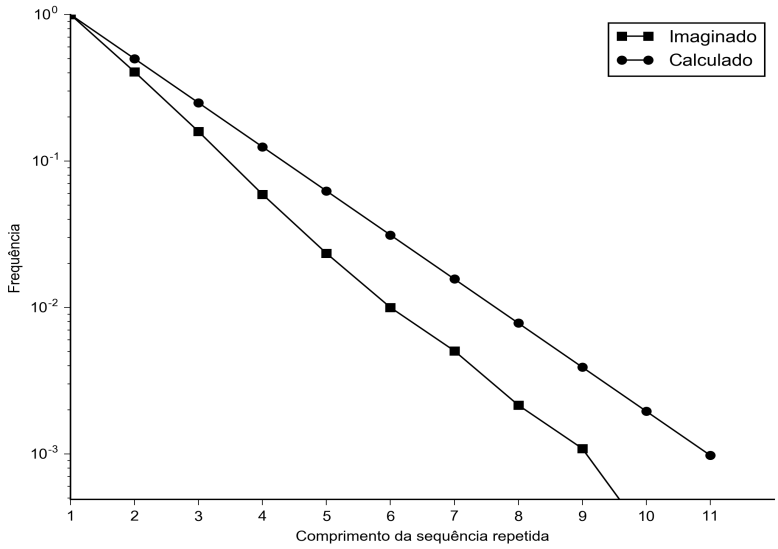


Figura 4.9: Frequências de sequências de k símbolos.

inverter o símbolo imaginado. A representação dos dados em termos do logarítmo da frequência é interessante para leis de formas exponenciais. Supomos um modelo $f = A2^{-\alpha(k-1)}$ e vemos da figura que $\alpha_B = 1$ para a binomial e $\alpha_I = 1.25$ para o processo imaginado. Explicar o valor não trivial pode ser interessante para quem estiver interessado nos aspectos psicológicos do problema. Um modelo muito simples é o de memória de um passo. Isto será retomado ao olhar para processos Markovianos. Neste caso podemos considerar como modelo apropriado um com $P(x_{t+1}|x_t)$ dado por

$$\begin{aligned} P(x_{t+1} = 1|x_t = 1) &= 1 - P(x_{t+1} = -11|x_t = 1) = \frac{1}{2^\alpha} \approx 0.42 \\ P(x_{t+1} = -1|x_t = -1) &= 1 - P(x_{t+1} = 1|x_t = -1) = \frac{1}{2^\alpha} \approx 0.42 \end{aligned}$$

Se você não lembrasse do passado isso não poderia ocorrer. Ver vários símbolos iguais sugere que o próximo deve ser diferente. Imaginem o contrário, um observador olha para um processo binomial e se surpreende que houve vários símbolos iguais. O que faz? Aposta que o próximo também deve ser iguais pois "se tudo fosse normal"deveria haver uma compensação. Acredita que o processo está quente e se dispõe a apostar mais alto, porque a máquina que gera as jogadas "está quente". Perceber que cometemos éstas falácias de análise pode ser útil para pessoas dispostas a perder dinheiro em apostas.

5
A distribuição Normal

A familia de distribuições Normal ou gaussiana é sem dúvida a mais importante na teoria e nas aplicações de probabilidade. Ela aparece tão frequentemente e comparativamente tem propriedades analíticas que permitem tantos resultados que muitas vezes parece natural que seja a única. Isso leva a erros também e portanto é necessário conhecer suas propriedades para usá-la ou não de forma adequada. É possível também encontrar referências à curva do sino (bell curve). Por favor refira-se à gaussiana como a curva do sino quando quiser deixar claro que você considera matemática e astrologia no mesmo patamar epistemológico.

Há dois motivos teóricos muito fortes que levam a considerar a modelagem da maior parte dos fenômenos aleatórios como gaussianos

- Teorema do Limite Central
- Entropia.

Há vários outros motivos que aparecerão nos desenvolvimentos futuros. Insistimos que haverá outras condições onde não deve ser usada. Veremos no capítulo 7 uma exposição sobre o teorema do limite central. O problema lida com somas de variáveis aleatórias. Por exemplo, $Y = \sum_{i=1}^{n} X_i$ sob condições bastante gerais, mas que aqui podemos reduzir a: I_X : as variáveis X_i tem segundo momento finito e são independentes (condição não necessária). Segue que a distribuição $P(Y|nI_X)$ de Y se aproxima da distribuição normal, numa região perto do máximo (região central) e que segundo critérios que podem ser definidos com cuidado, a aproximação melhora quando n cresce.

O segundo motivo tem a ver com entropia que é o tema central de qualquer curso de teoria de informação, mecânica estatística e termodinâmica e será o tema do capítulo 10. Para uma variável que toma valores no eixo real, com base na informação de que o seu primeiro momento é μ e a variância é σ^2, qual é a distribuição que deveriamos atribuir-lhe? Há infinitas distribuições compatíveis com essa informação. Poderiamos escolher qualquer uma delas. Por exemplo poderiamos escoher uma uniforme centrada em μ com a mesma variância, portanto largura $L = 2\sqrt{3}\sigma$. Ou seja fora do intervalo de tamanho L centrado em μ, X seria zero. Porque? O que levaria alguém a apostar que fora desse intervalo a probabilidade é nula? Pequena talvez, mas nula? O quê é que eles sabem que eu não sei? perguntaria o desconfiado. Iremos mostrar que a gaussiana é

aquela distribuição que satisfaz os vínculos informacionais e faz o menor número de hipóteses adicionais, que enfatizamos, não deveriamos fazer.

Por agora apresentaremos alguns resultados que permitirão manipular situações onde aparecem distribuições normais. Este capítulo portanto deve ser considerado como auxiliar não a teoria mas às ferramentas necessárias para o seu desenvolvimento.

Integrais Gaussiana

A distribuição gaussiana ou normal com parâmetros μ e σ é

$$P(x|\mu\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{5.1}$$

Dizemos que uma variável é gaussiana ou normal se a densidade de probabilidade é a gaussiana acima. Também escrevemos

$$P(x|\mu\sigma) = \mathcal{N}(\mu, \sigma^2), \tag{5.2}$$

ou ainda pode ser encontrada a notação $x \sim \mathcal{N}(\mu, \sigma^2)$.

É facil mostrar que a normalização é adequada:

$$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{5.3}$$

Veja a prova no Apêndice A no final deste capítulo.

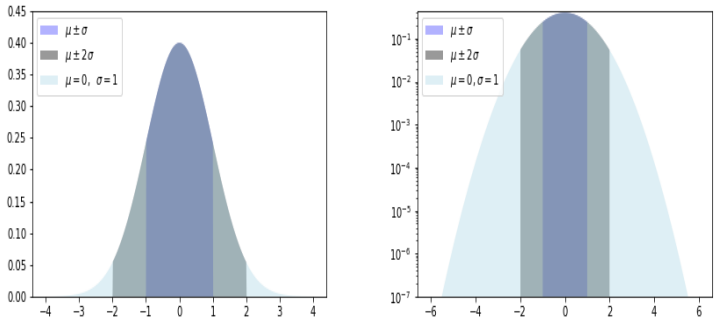


Figura 5.1: A distribuição gaussiana. As regiões marcadas mostram os valores que se afastam menos que 1σ e 2σ do valor médio. Do lado direito o eixo das ordenadas é logarítmico. As regiões centrais tem área ≈ 0.68 (1σ) e 0.95 (2σ). A região $3-\sigma$ tem área ≈ 0.99.

Limite da distribuição binomial

A história não cabe aqui, mas a distribuição Normal, assim como a de Poisson foram devidas a Abraham de Moivre.

Começamos com a distribuição binomial, que vimos no seção **??**. A uma variável de Bernoulli com dois estados, atribuimos probabilidade p de ser um *sucesso*. Vimos que a probabilidade de ter m sucessos em R tentativas é dada por

$$P(m|p, R, I_2) = \binom{R}{m} p^m q^{R-m} \tag{5.4}$$

com momentos

$$\langle m \rangle = pR, \tag{5.5}$$

$$\langle m^2 \rangle = R^2p^2 + Rp(1-p) \tag{5.6}$$

$$\text{var}(m) = \langle m^2 \rangle - \langle m \rangle^2 = Rp(1-p). \tag{5.7}$$

E conveniente introduzir os parâmetros μ e σ

$$\mu = \langle m \rangle \tag{5.8}$$
$$\sigma = \sqrt{Rp(1-p)}. \tag{5.9}$$

Estamos interessados em analizar o comportamento da probabilidade binomial para valores grandes do número de tentativas R. A idéia é perceber que ao analisar escalas onde $R >> 1$ é relevante, podemos tratar m/R como uma variável que toma valores no contínuo. Isto é devido a que diferenças entre $(m+1)/R$ e m/R nessa escala, da ordem de $1/R$ possam ser julgadas irrelevantes para alguns fins. De Moivre deu os primeros passos e Stirling (ver apêndice no final deste capítulo) mostrou que para valores grandes, os fatoriais podem ser bem aproximados por

$$\log n! = n \log n - n, \tag{5.10}$$

portanto o coeficiente binomial pode ser aproximado por

$$\begin{aligned} \log\left(\binom{R}{m}\right) &= (R\log R - R) - (m\log m - m) - ((R-m)\log(R-m) - R + m) \\ &= -m\log\frac{m}{R} - (R-m)\log(1-\frac{m}{R}). \end{aligned} \tag{5.11}$$

que leva a

$$\log P(m|p,R,I_2) = -m\log\frac{m}{R} - (R-m)\log(1-\frac{m}{R}) + m\log p + (R-m)\log(1-p). \tag{5.12}$$

Queremos aproximar esta expressão por uma expansão de Taylor em torno da moda, ou seja do seu máximo:

$$\begin{aligned} \frac{d}{dm}\log P(m|p,R,I_2) &= -\log\frac{m}{R} - 1 + \log(1-\frac{m}{R}) + 1 + \log p - \log(1-p) \\ &= -\log\frac{m/R}{1-m/R} + \log\frac{p}{1-p} \end{aligned} \tag{5.13}$$

que é zero para o valor óbvio $\mu = \langle m \rangle = Rp$. A segunda derivada em μ é

$$\begin{aligned} \frac{d^2}{dm^2}\log P(m|p,R,I_2)|_{m=\mu} &= -\frac{R}{\mu(R-\mu)} \\ &= -\frac{1}{\sigma^2} \end{aligned} \tag{5.14}$$

e até segunda ordem

$$\log P(m|p,R,I_2) \approx \log P(\mu|p,R,I_2) - \frac{1}{2\sigma^2}(m-\mu)^2 + ... \tag{5.15}$$

Vamos introduzir uma nova variável X que toma valores x nos reais $x \sim m$, com densidade de probabilidade

$$P(x|\mu\sigma)dx = P(m|p,R,I_2)$$

portanto a densidade de probabilidade será

$$P(x|\mu\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{5.16}$$

Note que da expressão anterior teriamos só que $P(x|\mu\sigma) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, mas sabemos qual deve ser a normalização correta. É claro que ter parado a expansão na segunda ordem não prova nada. Devemos

analisar as derivadas superiores. Continuamos derivando a partir da equação 5.14:

$$\begin{aligned}
\frac{d^2}{dm^2}\log P(m|p,R,I_2) &= -\frac{1}{m}-\frac{1}{R-m}\\
\frac{d^3}{dm^3}\log P(m|p,R,I_2) &= \frac{1}{m^2}-\frac{1}{(R-m)^2}\\
\frac{d^4}{dm^4}\log P(m|p,R,I_2) &= -2\frac{1}{m^3}+2\frac{1}{(R-m)^3}\\
\frac{d^k}{dm^k}\log P(m|p,R,I_2)| &= \mathcal{O}(\frac{1}{m^{k-1}}) \qquad (5.17)
\end{aligned}$$

o que significa que cada derivada, calculada em $m=\mu$ ganha um fator Rp no denominador. E sugere que para valores de R grandes a binomial é bem aproximada por uma gaussiana com a mesma média e variancia. Isto não substitui uma prova, mas serve como sugestão para o caminho de uma prova rigorosa. Isto é um caso particular do teorema do limite central. Duas características deste tipo de resultado devem ser notadas, pois são típicamente associadas a este tipo de aplicação do teorema. Primeiro, a aproximação pela série de Taylor é melhor na parte central, perto da moda. Em segundo o valor de m é a soma da variável que toma valores 1 para o sucesso e 0 para o fracasso. No capítulo 7 são apresentadas condições para este teorema de forma mais abrangente.

Momento Centrais da Gaussiana

O valor máximo da gaussiana ocorre para $x=\mu$. Portanto se a região onde há probabilidade razoável de encontrar o valor de X mudar, isto deve se refletir em mudança de μ. O parâmetro μ reflete portanto conhecimento sobre a *localização* de X. A medida que consideramos valores de x mais longe de μ a probabilidade cai. Mas o que quer dizer mais longe? O quanto $|x-\mu|$ é relevante depende do valor de σ. A distribuição depende de $y=\frac{|x-\mu|}{\sigma}$ na forma $\exp(-y^2/2)$. Portanto σ é um parâmetro de *escala.*

Os parâmetros da Gaussiana tem uma interpretação simples em termo dos momentos. Para calcular o valor esperado usamos a linearidade da integral e o fato que a distribuição está normalizada

$$\begin{aligned}
\mathbb{E}(x) &= \int_{-\infty}^{\infty} xP(x|\mu\sigma)dx\\
\mu &= \int_{-\infty}^{\infty} \mu P(x|\mu\sigma)dx\\
\mathbb{E}(x-\mu) &= \int_{-\infty}^{\infty} (x-\mu)P(x|\mu\sigma)dx\\
\text{mudando variável}\quad & x\to x'+\mu\\
\mathbb{E}(x-\mu) &= \int_{-\infty}^{\infty} x'P(x'|0\sigma)dx'\\
\mathbb{E}(x-\mu) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x' e^{-\frac{x'^2}{2\sigma^2}}dx'.\\
\mathbb{E}(x-\mu) &= 0\to \mathbb{E}(x)=\mu, \qquad (5.18)
\end{aligned}$$

pois o integrando é o produto de uma função impar por uma par [1]. Para o segundo momento, mostraremos que

$$\begin{aligned}
\mathbb{E}(x^2) &= \sigma^2+\mu^2\\
\mathbb{E}((x-\mu)^2) &= \mathbb{E}(x^2)-\mu^2=\sigma^2. \qquad (5.19)
\end{aligned}$$

[1] Uma função com a propriedade (i) $f_S(x)=f_S(-x)$ é chamada *par* ou simétrica, e uma função com a propriedade (ii) $f_A(x)=-f_A(-x)$ é chamada *impar* ou antissimétrica. Uma mudança da variável de integração $x\to x'=-x$ permite mostrar que integrais em intervalos $(-a,a)$ do produto $f_A(x)f_S(x)$ são nulas e em particular

$$\int_{-\infty}^{\infty} f_A(x)f_S(x)dx=0$$

Claro que supomos que as funções satisfazem propriedades que permitem as manipu-

Note a semelhança com as relações equivalentes para a binomial, mas lembre que agora o símbolo $\mathbb{E}$ significa uma integral e não a soma. Obviamente os momentos dependem da posição e da escala. Mas se olharmos para os momentos de y todas as distribuições tem os mesmos momentos. Os valores esperados

$$\begin{aligned} \mathbb{E}(x-\mu) &= 0 \\ \mathbb{E}((x-\mu)^2) &= \mathbb{E}(x^2) - \mu^2 = \sigma^2 \\ \mathbb{E}((x-\mu)^n) &= M_n(\sigma) \end{aligned} \tag{5.20}$$

de $x - \mu$ são chamados de momentos centrais:
$M_n(\sigma) = \mathbb{E}((x - \mathbb{E}(x))^n)$. Os momentos centrais impares M_{2n+1} são nulos. Em geral

$$\begin{aligned} M_n(\sigma) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2\sigma^2}} \frac{dx}{\sigma} \\ &= \sigma^n \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^n e^{-\frac{y^2}{2}} dy \\ &= \sigma^n M_n(\sigma = 1) \end{aligned} \tag{5.21}$$

mostrando que os momentos centrais dependem do desvio padrão σ de uma forma simples, σ^n vezes o momento central de y que denotaremos simplesmente $M_n(\sigma = 1) = M_n$. Estes são fáceis de calcular, integrando por partes [2]:

$$\begin{aligned} M_n &= \left(e^{-y^2/2} \frac{y^{n+1}}{n+1} \right)_{-\infty}^{\infty} - \frac{-1}{\sqrt{2\pi}} \int_{\infty}^{\infty} \frac{y^{n+2}}{n+1} e^{-\frac{y^2}{2}} dy \\ &= \frac{1}{n+1} M_{n+2}. \end{aligned} \tag{5.22}$$

[2] Lembrando: $\int_a^b u dv = uv|_a^b - \int_a^b v du$. Usaremos $u = e^{-y^2/2}$ e $dv = y^n dy$, portanto $v = \frac{y^{n+1}}{n+1}$ e $du = -ye^{-y^2/2}$.

Segue a relação de recorrência $M_{2n} = (2n-1)M_{2n-2}$. Começando de $M_2 = 1$ e iterando $M_{2n} = (2n-1)!!$ onde
$(2n-1)!! = 1 \times 3 \times 5 \times \cdots (2n-1) = (2n)!/(2^n n!)$. Logo

$$M_{2n}(\sigma) = \sigma^{2n}(2n-1)!! \tag{5.23}$$

Exercício: Função geratriz. Sem fazer o cálculo explícito, uma mudança de variáveis permite mostrar que se $x \sim \mathcal{N}(0, \sigma)$ então os valores esperados $\mathbb{E}(x^{2n}) \propto \sigma^{2n}$. A constante de proporcionalidade é M_{2n}. Calcule $\mathbb{E}(e^x) = f(\sigma)$, para isso complete os quadrados na exponencial (ver o resultado 5.67 no apêndice no fim deste capítulo.) Expanda em série de potências de x para mostrar que $\mathbb{E}(e^x) = \sum_{n=0}^{\infty} A_n(\sigma)\mathbb{E}(x^n)$. Expanda $f(\sigma)$ em potências de σ. Comparando termo a termo as duas séries reobtenha a expressão 5.23. O valor esperado $\mathbb{E}(e^x)$ é chamado de função geratriz dos momentos pois contém em si informação que permite gerar qualquer momento da distribuição. A nomenclatura não é uniforme e aqui chamaremos a função relacionada $\mathbb{E}(e^{ikx})$ de função característica que também pode ser chamada de geratriz dos momentos.

Herschel e Maxwell: um pouco de física

Mas ainda resta saber porque é razoável considerar erros gaussianos. Voltaremos a isto várias vezes. Agora veremos uma dedução a partir de algumas hipóteses razoáveis. O astrônomo John Herschel (filho de William) e o Maxwell deram argumentos que levam a gaussianas de forma muito elegante [3]. Herschel estava preocupado com caracterizar

[3] Ver Probability, E. T. Jaynes.

os erros de medida da posição de uma estrela e fez duas hipóteses. A primeira é que (i) os erros das coordenadas e_x e e_y, respectivamente os erros de medida da longitude (leste-oeste) e declinação (norte-sul) são supostos independentes e igualmente distribuidos. Portanto a distribuição conjunta deve ser fatorizável:

$$P(e_x, e_y|I) = P(e_x|I)P(e_y|I) = f(e_x)f(e_y),$$

para alguma função f ainda desconhecida. Se em lugar de coordenadas cartesianas ele usasse coordenadas polares:

$$\begin{aligned} e_x &= e_r \cos e_\theta \\ e_y &= e_r \sin e_\theta \\ e_r^2 &= e_x^2 + e_y^2 \\ P(e_x, e_y)de_x de_y &= P(e_r, e_\theta)e_r de_r de_\theta \end{aligned} \tag{5.24}$$

A segunda hipótese é que (ii) $P(e_r, e_\theta|I) = g(e_r)$ não depende do ãngulo, onde g é uma nova função igualmente desconhecida.

Temos, portanto , que para qualquer e_x e e_y temos uma equação funcional relacionando as duas funções desconhecidas

$$f(e_x)f(e_y) = g(\sqrt{e_x^2 + e_y^2})$$

e em particular, ao longo de um dos eixos cartesianos

$$f(e_x)f(0) = g(e_x),$$

que dermina g se f for conhecido. Eliminamos uma das funções desconhecidas e voltamos ao caso geral

$$f(e_x)f(e_y) = g(\sqrt{e_x^2 + e_y^2}) = f(\sqrt{e_x^2 + e_y^2})f(0)$$

$$\frac{f(e_x)f(e_y)}{f(0)^2} = \frac{f(\sqrt{e_x^2 + e_y^2})}{f(0)}$$

que é uma equação funcional com uma incógnita só. Definindo $\exp h(e_x) = f(e_x)/f(0)$, temos

$$h(e_x) + h(e_y) = h(\sqrt{e_x^2 + e_y^2}).$$

Obviamente $h(0) = 0$. Também vemos que $h(e_x) = h(-e_x)$. Supondo que h é duas vezes diferenciável, derivamos primeiro com respeito a e_x

$$h'(e_x) = h'(\sqrt{e_x^2 + e_y^2})\frac{e_x}{\sqrt{e_x^2 + e_y^2}}$$

e a seguir com respeito a e_y

$$0 = \frac{d}{de_y}\left(\frac{h'(\sqrt{e_x^2 + e_y^2})}{\sqrt{e_x^2 + e_y^2}}\right)$$

portanto $h'(x) \propto x$ e $h(x) \propto x^2$. A solução geral, que depende de um parãmero a, é

$$f(e_x) \propto e^{ae_x^2}.$$

como isto tem que ser uma distribuição normalizável, só podemos considerar $a < 0$, que escrevemos por motivos óbvios $a = -1/2\sigma^2$. Impondo normalização, chegamos à distribuição normal $e_x \sim \mathcal{N}(0, \sigma^2)$.

Maxwell fez um raciocínio similar sobre as velocidades de um átomo ou molécula num gás. Dentro do modelo que considerou as partículas que compoem o gás não são interagentes. Esse modelo recebe o nome de gás ideal. É ideal para o teórico que pode calcular tudo o que quiser, mas não deve ser para o experimental pois a maioria dos gases costuma ter propriedades muito mais complexas, a não ser que esteja em limites de baixas densidades. Ser ideal significa que é razoável supor que se as partículas não interagem, e são portanto independentes, devem ser igualmente distribuidas e podemos olhar para a distribuição de velocidades de uma partícula. Ainda mais, o mesmo se aplica ao olhar para uma partícula para as distribuições das componentes cartesianas. Segue que a distribuição de velocidade $\boldsymbol{V} = (v_x, v_y, v_z)$ fatoriza nas três dimensões:

$$P(\boldsymbol{V}|I) = P(v_x, v_y, v_z|I) = P(v_x|I)P(v_y|I)P(v_z|I).$$

A hipótese que isto é dado por uma função que só depende da magnitude $v = |\boldsymbol{V}|$ leva à generalização da equação funcional de Herschel e novamente à gaussiana. Assim

$$P(\boldsymbol{V}|I) = \prod_{i=x,y,z} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{v_i^2}{2\sigma^2}},$$

onde, e aqui a parte mais interessante que não provaremos, $\sigma = \sqrt{\frac{k_B T}{m}}$ onde T e m são a temperatura e a massa das partículas medidas em unidades apropriadas e k_B é a constante de Boltzmann que permite converter unidades de energia em graus de temperatura absoluta. [4] A densidade é gaussiana e para obter a probabilidade devemos incluir o elemento de volume $d^3\boldsymbol{V} = dv_x dv_y dv_z$. Passando para coordenadas esféricas

$$P(\boldsymbol{V}|I)d^3\boldsymbol{V} = P(v, \theta, \phi)v^2 dv d\Omega(\theta, \phi),$$

e dado que, por hipótese, $P(v, \theta, \phi)$ é função de $v^2 = v_x^2 + v_y^2 + v_z^2$ somente, não depende das variáveis angulares. Integrando sobre as variáveis angulares, temos que

$$\begin{aligned} P(v|I)dv &= \frac{4\pi}{(2\pi\sigma^2)^{\frac{3}{2}}} e^{-\frac{v^2}{2\sigma^2}} v^2 dv \\ &= 4\pi \left(\frac{m}{2\pi k_B T}\right)^{\frac{3}{2}} e^{-\frac{mv^2}{2k_b T}} v^2 dv \end{aligned}$$

[5] Este trabalho, no que se chama teoria cinética dos gases [6] , levou à Mecânica Estatística de Boltzmann e Gibbs e de muitos outros. Foi demostrado por Jaynes na década de 1950 que poderia ser entendido como um exemplo de teoria de informação.

[4] Suponha que voce coloque em contato térmico dois recipientes com o mesmo gás, n_i moles e temperaturas T_i, $i = 1, 2$ respectivamente. Suponha que é uma verificação empírica que nesse regime de temperaturas, entre T_1 e T_2 o calor específico é constante. Mostre que a temperatura final de equilíbrio é $T_f = \alpha T_1 + (1 - \alpha)T_2$, onde $\alpha = n_1/(n_1 + n_2)$. Mostre que a energia cinética média por partícula também satisfaz $e_f^c = \alpha e_1^c + (1 - \alpha)e_2^c$. Isto sugere que a energia cinética média por partícula é proporcional à temperatura. Por isso a variância da gaussiana é proporcional à temperatura.

[5] $d\Omega(\theta, \phi) = \sin\theta d\theta d\phi$, com limites $\theta \in (0, \pi)$ e $\phi \in (0, 2\pi)$. A área da esfera de raio 1 em 3d é:

$$\int d\Omega = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta = 4\pi$$

[6] Isto não passa de uma dedução rápida de uma gaussiana e da forma funcional da densidade de velocidades. Porque aparece a massa da partícula ou a temperatura será o tema de capítulos posteriores que não serão vistos num curso introdutório.

A Distribuição normal cumulativa

Uma variável $X \sim \mathcal{N}(0, 1)$ tem distribuição cumulativa

$$\Phi(x) = \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} \frac{dt}{\sqrt{2\pi}}. \tag{5.25}$$

É claro que $\Phi(-\infty) = 0$ e $\Phi(\infty) = 1$. A figura 5.2 o gráfico de $\Phi(x)$ no intervalo $-4 < x < 4$. Uma curva que vai de um valor assintótico constante a outro como ésta é chamada de sigmoide. Há várias outras sigmoides e um exemplo é a tangente hiperbólica. **Exercício:** Qual é a

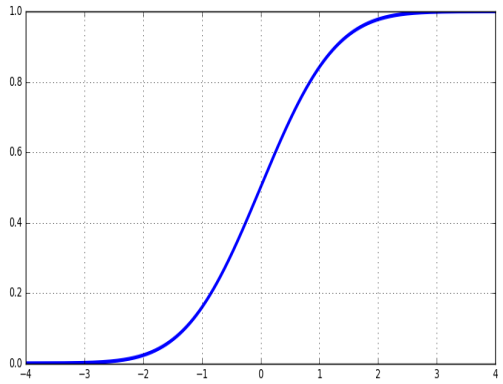


Figura 5.2: A distribuição cumulativa $\Phi(x)$

distribuição de densidade de probabilidades de uma variável cuja cumulativa é $\tanh(x)$?

Há várias funções relacionadas que foram introduzidas de forma independente e que são usadas na literatura:

A função erro:

$$\text{erf}(x) = \int_{-x}^{x} e^{-t^2} \frac{dt}{\sqrt{\pi}}. \tag{5.26}$$

A função erro complementar $\text{erfc}(x) = 1 - \text{erf}(x)$

Exercício: Mostre que

$$\Phi(x) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{x}{\sqrt{2}}) \tag{5.27}$$

Gauss e a gaussiana

Novamente a história pode ter sido outra, mas suponhamos que Gauss se fez a seguinte pergunta. Suponha que queiramos medir numa experiência o valor de uma quantidade z. O valor obtido na primeira medida é x_1. Podemos parar por ai, mas parece razoável medir novamente e obtemos x_2, diferente do primeiro resultado. Qual é o valor de z que devemos reportar? Porquê parar ai? Fazemos mais medidas e obtemos o conjunto de dados $D = \{x_1, x_2, ..., x_n\}$. Há uma grande tendência entre os praticantes de experiências para dizer que a média empírica

$$\hat{z} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

deva ser o valor estimado de z a ser reportado. Gauss achou isso razoável também, mas foi além. Cada medida pode ser descrita por

$$z = x_i + \xi_i$$

onde ξ_i é o erro. A pergunta que Gauss fez é: qual é a lei da distribuição de densidade de probabilidade de ξ que leva a que $\hat{z}$ seja uma boa resposta?

Vamos responder ésta pergunta usando as regras da probabilidade e algumas hipóteses. O teorema de Bayes dá

$$\begin{aligned} P(z|x_1,....x_n,I) &\propto P(z|I)P(x_1,....x_n|z,I) \quad \text{Regra do produto} \\ P(z|x_1,....x_n,I) &\propto P(x_1,....x_n|z,I) \quad \textit{a priori} \text{ uniforme} \\ &\propto \prod_i P(x_i|zI) \quad \text{Independência} \end{aligned} \tag{5.28}$$

A distribuição de ξ é um membro de uma família desconhecida $g(\xi, \theta)$ que supomos diferenciável com respeito a todos os argumento e definimos $\partial \log g(u,\theta)/\partial u = f(u,\theta)$. Tomando logaritmos podemos escrever

$$\begin{aligned} \log P(z|x_1,....x_n,I) &\propto \sum_i \log g(z-x_i) \\ \frac{\partial \log P(z|x_1,....x_n,I)}{\partial z} &\propto \sum_i f(z-x_i). \end{aligned}$$

Agora impomos que a densidade seja máxima em $\hat{z}$

$$\begin{aligned} 0 &= \frac{\partial \log P(z|x_1,....x_n,I)}{\partial z}|_{z=\hat{z}} \\ &= \sum_i f(\hat{z}-x_i) \\ 0 &= \sum_i f(\frac{1}{n}\sum_j x_j - x_i). \end{aligned} \tag{5.29}$$

A equação acima nos dá uma condição que a função desconhecida f deve satisfazer. Podemos ver que a se escolhermos $f(u,\theta) = au$, temos a identidade

$$0 = \sum_j x_j - \sum_i x_i. \tag{5.30}$$

Isto leva a $\log g(u|\theta) = au^2 + b$ e $g(u,\theta) \propto \exp(au^2+b)$ e portanto, como deve estar normalizada a deve ser negativo e b fica determinado por normalização. Revertendo à notação usual:

$$\begin{aligned} P(\xi|\theta) &\propto \exp(-\frac{1}{2\sigma^2}\xi^2) \\ P(z|x_i\sigma) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(z-x_i)^2) \end{aligned} \tag{5.31}$$

Surge naturalmente a gaussiana e segue o fato que $\mathbb{E}(\xi) = 0$ que significa que se a média empírica é um estimador adequado não deve haver erro sistemático. Deixamos para o próximo capítulo a estimativa de σ a partir dos dados.

Um pouco de análise de erros

O que trataremos nesta seção é de interesse por si só, mas pode ser visto como o começo das idéias que levam ao teorema do limite central que será tema do próximo capítulo.

Os alunos de Física são expostos no começo dos seus estudos à questões importantes sobre o significado dos números obtidos em um experimento. Dentro do contexto destas notas, uma pergunta sobre esse significado pode ser colocada como

> O que posso dizer sobre o valor numérico de X quando o resultado de uma medida deu x_1?

Note que se estivermos falando de um sistema cognitivo precisariamos modelar o que sabemos sobre o mundo externo (X) se os sistemas sensoriais se encontram num dado estado (x_1). Embora possa não parecer, são problemas da mesma natureza.

Se não soubermos nada sobre o aparelho de medida e como fizemos a experiência, não podemos dizer nada. Talvez a experiência tenha

sido usar uma régua milimetrada para medir a distância até o sol, ou entre dois átomos vizínhos num sólido. Precisamos mais informação.

Suponha que tenhamos motivos para achar que os erros de medida ϵ tem probabilidade gaussiana. Podemos escrever isso como

$$x = x_1 + \epsilon_x$$

onde $P(\epsilon_x|\mu_x\sigma_x I_e) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(\epsilon-\mu_x)^2}{2\sigma_x^2}}$. O valor esperado de ϵ é μ_x, isto costuma ser chamado de erro sistemático. Vamos supor que há motivos para achar que seja zero, o que poderia ser alcançado calibrando corretamente os aparelhos de medida. A raiz quadrada da variância σ_x descreve a dispersão dos erros dos valores medidos em torno da média. Segue que

$$P(x|x_1 I_e) = P(\epsilon_x|\mu_x\sigma_x I_e) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-x_1)^2}{2\sigma_x^2}} \tag{5.32}$$

Suponha que outra variável y seja medida com erros também gaussianos, caracterizados analogamente por μ_y e ϵ_y, mas de fato estamos interessados em z que por motivos teóricos achamos razoável descrever pelo modelo [7]

$$\mathcal{M} : z = x + y. \tag{5.33}$$

[7] Não estamos, neste momento questionando se o modelo é correto ou não. Isso é outro problema, veja o capítulo 8

Dado o que sabemos sobre medidas de x e y e sobre a caracterização dos seus erros de medida, o quê podemos dizer sobre z? Podemos escrever

$$z = z_1 + \epsilon_x + \epsilon_y. \tag{5.34}$$

Vamos mostrar a seguir (i) que z é uma variável com distribuição gaussiana, que não é óbvio neste momento, (ii) que o valor médio de z é $z_1 = x_1 + y_1$, que deve parecer óbvio e (iii) que a variância de z também é dada pela soma das variâncias de x e y, que também não deve parecer óbvio para o leitor.

A informação é codificada numa distribuição de probabilidades $P(z|x_1, y_1\mathcal{M}I)$ que queremos construir a partir das regras da teoria de probabilidade. Começamos pela marginalização:

$$P(z|x_1, y_1\mathcal{M}I) = \int dx \int dy P(x, y, z|x_1, y_1\mathcal{M}I). \tag{5.35}$$

Os limites das integrais são $-\infty$ e ∞, mas ficarão subentendidos.

A regra do produto aplicada ao integrando

$$\begin{aligned} P(z|x_1, y_1\mathcal{M}I) &= \int dx \int dy P(x, y|x_1, y_1\mathcal{M}I) P(z|x, y, x_1, y_1\mathcal{M}I) \\ &= \int dx \int dy P(x|x_1, y_1\mathcal{M}I) P(y|x, x_1, y_1\mathcal{M}I) P(z|x, y, x_1, y_1\mathcal{M}I) \end{aligned}$$

Caso acreditemos na indepêndencia entre as medidas de x e y, teremos $P(y|x, x_1, y_1\mathcal{M}I) = P(y|y_1 I_e)$. Também usamos que neste ponto o modelo $\mathcal{M}$ não tem influência. Usando a equação 5.32 e o equivalente para y, teremos

$$\begin{aligned} P(z|x_1, y_1\mathcal{M}I) &= \int dx \int dy P(x|x_1, I_e) P(y|y_1 I_e) P(z|x, y, \mathcal{M}I) \\ &= \int dx \int dy \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-x_1)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-y_1)^2}{2\sigma_y^2}} P(z|x, y, \mathcal{M}I). \end{aligned}$$

Dois comentários importantes sobre a expressão acima. Primeiro, a probabilidade de z que aparece do lado esquerdo é condicionada nos

dados x_1 e y_1. Não aparecem os valores *reais* de x nem y. Não temos acesso à realidade a não ser pelos dados. Para cada escolha de x e y temos uma probabilidade de z, mas integramos sobre todas as escolhas porque em maior ou menor grau, todas tem algum mérito, dentro do contexto da informação condicionante. Segundo, ainda não acabamos, temos que encontrar uma forma para $P(z|x,y,\mathcal{M}I)$. Ainda não temos ferramentas matemáticas para tornar isto imediato, e precisaremos um pouco de trabalho.

Um modelo como $\mathcal{M}$ denota conhecimento completo: dado x e y conhecemos z totalmente. Paradoxalmente, nesta altura do que se espera que o leitor saiba, o estado de conhecimento total torna as coisas mais difíceis. Então suponhamos que z não é determinado exatamente por x e y, mas por uma densidade de probabilidade, que tomaremos gaussiana de variância s^2 e média nula $P(z|x,y,\mathcal{M}_sI) = \mathcal{N}(0,s^2)$. Vamos usar a notação $\delta_s(z-x-y)$ para esta função. Agora, como veremos a seguir, podemos fazer as integrais necessárias, mas o resultado final dependerá de s. Qual é o valor de s? Tomaremos o limite de $s \to 0$, pois a dispersão de valores de z dados x e y é zero no caso de um modelo determinista como $\mathcal{M}$ [8].

[8] Este truque é conhecido desde o século XIX e associado a nomes como Frejet, Neuman, Landau. Mais recentemente, a Dirac

Juntando tudo

$$\begin{aligned} P(z|x_1,y_1\mathcal{M}I) &= \int dx \int dy \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-x_1)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-y_1)^2}{2\sigma_y^2}} \frac{1}{\sqrt{2\pi}s} e^{-\frac{(z-x-y)^2}{2s^2}} \\ &= \frac{1}{(2\pi)^{3/2}s\sigma_x\sigma_y} \int dx \int dy \exp\left(-\frac{(x-x_1)^2}{2\sigma_x^2} - \frac{(y-y_1)^2}{2\sigma_y^2} - \frac{(z-x-y)^2}{2s^2}\right). \end{aligned}$$

Esta integral pode ser um pouco assustadora, mas como mostramos no apêndice a seguir, estas integrais são fáceis. Elas aparecem de forma tão frequente, que convêm acostumar-se a fazé-las de forma automática. Começamos olhando para a integração em y. Temos a exponencial de um polinômio de segundo grau, e isso é fácil usando o resultado 5.67 [9]

[9] O resultado demostrado em 5.67 é $e^{\frac{h^2\sigma^2}{2}} = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}+xh} \frac{dx}{\sqrt{2\pi}\sigma}$.

$$P(z|x_1,y_1\mathcal{M}I) = \frac{1}{2\pi\sigma_x\sigma_y} \int dx e^{-\frac{(x-x_1)^2}{2\sigma_x^2}} \left(\int \frac{dy}{\sqrt{2\pi}s} e^{-\frac{(y-y_1)^2}{2\sigma_y^2}} e^{-\frac{(z-x-y)^2}{2s^2}}\right). \tag{5.36}$$

O expoente do integrando da integral em y é

$$\begin{aligned} \frac{(y-y_1)^2}{2\sigma_y^2} + \frac{((z-x)-y)^2}{2s^2} &= \frac{y^2}{2u^2} - yh + \frac{(z-x)^2}{2s^2} + \frac{y_1^2}{2\sigma_y^2} \\ \text{onde } h = \frac{y_1}{\sigma_y^2} + \frac{(z-x)}{s^2} &\quad \text{e} \quad \frac{1}{u^2} = \frac{1}{\sigma_y^2} + \frac{1}{s^2} \end{aligned} \tag{5.37}$$

A integral em y (termo entre parêntesis em 5.36), usando 5.67

$$\begin{aligned} &= e^{-\frac{y_1^2}{2\sigma_y^2}} e^{-\frac{(z-x)^2}{2s^2}} \int \frac{dy}{\sqrt{2\pi}s} e^{-\frac{y^2}{2u^2}-yh} \\ &= e^{-\frac{y_1^2}{2\sigma_y^2}} e^{-\frac{(z-x)^2}{2s^2}} \frac{u}{s} \int \frac{dy}{\sqrt{2\pi}u} e^{-\frac{y^2}{2u^2}-yh} \\ &\overset{5.67}{=} e^{-\frac{y_1^2}{2\sigma_y^2}} e^{-\frac{(z-x)^2}{2s^2}} \frac{u}{s} e^{\frac{h^2u^2}{2}} \end{aligned} \tag{5.38}$$

substituindo h e u

$$\begin{aligned} &= e^{-\frac{y_1^2}{2\sigma_y^2}(1-\frac{s^2}{s^2+\sigma_y^2})} e^{y_1(z-x)\frac{1}{s^2+\sigma_y^2}} e^{-(z-x)^2(\frac{1}{s^2}-\frac{\sigma_y^2}{s^2(s^2+\sigma_y^2)})} \frac{\sigma_y}{\sqrt{s^2+\sigma_y^2}} \\ &\overset{s\to 0}{\to} e^{-\frac{y_1^2}{2\sigma_y^2}} e^{2y_1(z-x)\frac{1}{2\sigma_y^2}} e^{-\frac{(z-x)^2}{\sigma_y^2}} \\ &= \exp(-\frac{1}{2\sigma_y^2}(z-x-y_1)^2) \end{aligned}$$

Pode parecer difícil, mas o que simplesmente acabamos de mostrar é

$$\int dy \exp\left(-\frac{(y-y_1)^2}{2\sigma_y^2}\right)\delta_s(z-x-y) \quad \overset{s\to 0}{\to} \quad \exp(-\frac{1}{2\sigma_y^2}(z-x-y_1)^2) \tag{5.39}$$

que mostra algo simples, no limite do modelo determinista quando $s \to 0$, eliminamos a variável y e a substituimos por $z - x$ que é o que o modelo indica [10].

Quase chegamos ao final do exercício. Voltamos à expressão 5.36, que agora toma a forma[11].

$$P(z|x_1, y_1\mathcal{M}I) = \frac{1}{2\pi\sigma_x\sigma_y}\int dx e^{-\frac{(x-x_1)^2}{2\sigma_x^2}} e^{-\frac{(z-x-y_1)^2}{2\sigma_y^2}}.$$

Novamente temos uma integral gaussiana, que podemos fazer recorrendo a 5.67. O expoente pode ser escrito

$$\begin{aligned} \frac{(x-x_1)^2}{2\sigma_x^2} + \frac{((z-y_1)-x)^2}{2\sigma_y^2} &= \frac{x^2}{2u'^2} - h'x + \left(\frac{x_1^2}{2\sigma_x^2} + \frac{(z-y_1)^2}{2\sigma_y^2}\right) \\ \text{onde } h' = \frac{x_1}{\sigma_x^2} + \frac{(z-y_1)}{\sigma_y^2} \quad &\text{e} \quad \frac{1}{u'^2} = \frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2} \end{aligned} \tag{5.40}$$

onde vemos que as expressões em 5.40 são análogas às 5.37. Completando quadrados novamente realizamos a integral em x usando 5.67

$$\begin{aligned} P(z|x_1, y_1\mathcal{M}I) &= \frac{\sqrt{2\pi}u'}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{x_1^2}{2\sigma_x^2} - \frac{(z-y_1)^2}{2\sigma_y^2}\right)\int \frac{dx}{\sqrt{2\pi}u'} e^{-\frac{x^2}{u'^2}+hx} \\ &= \frac{\sqrt{2\pi}u'}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{x_1^2}{2\sigma_x^2} - \frac{(z-y_1)^2}{2\sigma_y^2}\right) e^{\frac{h'^2u'^2}{2}} \end{aligned}$$

finalmente, substituindo h' e u' temos uma grande simplificação

$$P(z|x_1, y_1\mathcal{M}I) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{\frac{(z-x_1-y_1)^2}{2\sigma_z^2}} \tag{5.41}$$

onde introduzimos

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2. \tag{5.42}$$

Após as medidas x_1 e y_1 que tinham erros com variâncias σ_x^2 e σ_y^2 respectivamente, atribuiremos aos diferentes valores de z uma probabilidade gaussiana de vari ância σ_z^2 que é a soma das duas. O valor mais provável, a moda de z é

$$\mu_z = z_{\text{moda}} = x_1 + y_1, \tag{5.43}$$

o valor que teriamos atribuido a z, a partir do modelo $\mathcal{M}$ se as medidas não tivessem erros. Repetindo as contas para o caso em que o modelo é $w = x - y$, obteremos que w novamente é gaussiano com

[10] O leitor interessado deve procurar ler sobre a *"função"* ou melhor, distribuição δ de Dirac em particular e teoria de distribuições em geral. O objeto $\delta(z-x-y) = \lim_{s\to 0}\delta_s(z-x-y)$ tem propriedades interessantes. É preciso aumentar o conceito de função para chamar a δ de função. Para funções $f(x)$ (de forma bastante geral), entre as propriedades que δ tem, está

$$f(z) = \int_I f(x)\delta(x-z)dx$$

com $z \in I$. Esta propriedade que deduzimos para o caso particular permite chegar diretamente na equação 5.39.

[11] Expressões do tipo

$$h(x) = \int_\infty^\infty f(x-y)g(y)dy$$

são chamadas convoluções. Veremos mais disto no capítulo sobre o teorema do limite central 7. O resultado que obtemos na equação abaixo 5.41, é que a convolução de duas gaussianas é uma gaussiana, cujo parâmetro de localização é a soma e a variância também é soma dos das gaussianas originais.

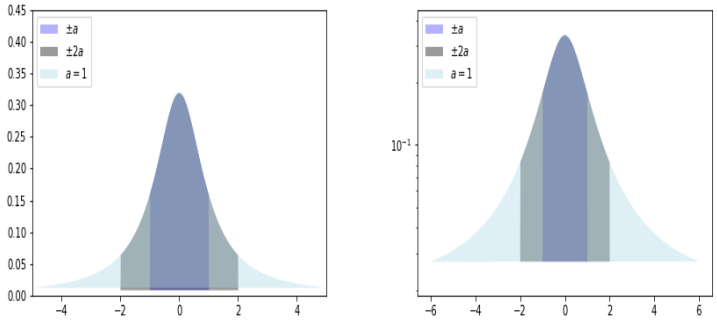


Figura 5.3: A distribuição de Cauchy. As regiões marcadas mostram os valores que se afastam menos que a e $2a$ da moda, que no caso é zero. Do lado direito o eixo das ordenadas é logarítmico.

a mesma variância de z e com moda $w_{\text{moda}} = x_1 - y_1$. Provavelmente o leitor reconheça 5.42 de cursos de Física Experimental.

Ainda falta saber como poderiamos ter escolhido o valor dos σs, que ficará para quando usarmos a regra de Bayes para estimar parâmetros de distribuições a partir de informação na forma de dados.

Propagação de erro

Há formas muito mais simples de fazer isto. Novamente $z = x + y$ e escrevemos $x = x_1 + \delta x$, $y = y_1 + \delta y$ e $z = z_1 + \delta z$ para fazer um ponto de contato com a notação usada na análise de erros em laboratório. Supomos os erros gaussianos como antes. Se não há erro sistemático vale que $\mathbb{E}(\delta x) = \mathbb{E}(\delta y) = \mathbb{E}(\delta z) = 0$. Temos que $\mathbb{E}(\delta x^2) = \sigma_x^2$, $\mathbb{E}(\delta y^2) = \sigma_y^2$ e $\mathbb{E}(\delta z^2) = \sigma_z^2$, e como

$$\begin{aligned} \sigma_z^2 &= \mathbb{E}(\delta z^2) = \mathbb{E}((\delta x + \delta y)^2) && (5.44) \\ &= \mathbb{E}(\delta x^2) + \mathbb{E}(\delta y^2) + \mathbb{E}(\delta x \delta y) \\ &= \sigma_x^2 + \sigma_y^2 && (5.45) \end{aligned}$$

onde usamos $\mathbb{E}(\delta x \delta y) = \mathbb{E}(\delta x)\mathbb{E}(\delta y) = 0$, devido à independência entre x e y. Reobtivemos a relação entre as variâncias da equação 5.42. Mas nada desta análise garante que z seja uma variável gaussiana.

A distribuição produto e a distribuição quociente ou razão de duas variáveis gaussianas

O estudo em geral das distribuiçoes de produtos ou quocientes faz parte do que se chama de álgebra de variáveis aleatórias. Obviamente, a distribuição do produto não é o produto de distribuições. Como vimos na soma, a distribuição de z, dado por $z = x + y$ não tem nada a ver com o soma das distribuições e nem faz referência às variáveis x nem y. Em geral para o produto ou quociente as contas ficam bem mais complicadas e faremos algumas simplificações. Em alguns casos de quocientes de variáveis gaussianas é possível fazer as contas, como veremos depois.

Suponha que x e y sejam como definidos acima, mas agora os modelos possíveis são

$$\begin{aligned} \mathcal{M}_{\text{prod}} &: s = xy && (5.46) \\ \mathcal{M}_{\text{razão}} &: r = \frac{x}{y} && (5.47) \end{aligned}$$

Agora esperamos que s e r sejam descritos por $s_1 = x_1 y_1$ e $r_1 = x_1/y_1$ mais algum erro:

$$\begin{aligned} s &= (x_1 + \epsilon_x)(y_1 + \epsilon_y) \\ r &= \frac{x_1 + \epsilon_x}{y_1 + \epsilon_y} \end{aligned}$$

Supondo que as amplitudes dos erros sejam *pequenos*, ou seja $x_1 >> \sigma_x$ e $y_1 >> \sigma_y$, então [12]

$$\begin{aligned} s &= (x_1 + \epsilon_x)(y_1 + \epsilon_y) = s_1 + x_1\epsilon_y + y_1\epsilon_x + \epsilon_x\epsilon_y \\ &\approx s_1 + x_1\epsilon_y + y_1\epsilon_x = s_1 + \eta_1 + \eta_2. \end{aligned}$$

[12] Não vamos fazer a conta exata pois entram funções que não são familiares aos leitores, eg. de Bessel

Jogar fora termos quadráticos permite manter as contas simples, e é justificado porque o ruído é pequeno. A nova forma para s inclui $\eta_1 = x_1\epsilon_y$ e $\eta_2 = y_1\epsilon_x$. Se uma variável tem distribuição gaussiana e é multiplicada por uma constante, ainda terá distribuição gaussiana mas o seus parâmetros de localização e escala serão multiplicados pelo mesmo fator. Assim $\sigma^2_{\eta_1} = x_1^2\sigma_y^2$ e $\sigma^2_{\eta_1} = y_1^2\sigma_x^2$. Teremos que s é aproximadamente gaussiano

$$P(s|x_1, x_2\mathcal{M}_{\text{prod}}) = \mathcal{N}(s_1, \sigma_s)$$

onde, pela equação 5.42, temos $\sigma_s^2 = \sigma^2_{\eta_1} + \sigma^2_{\eta_2}$. Logo

$$\sigma_s^2 = x_1^2\sigma_y^2 + y_1^2\sigma_x^2 \tag{5.48}$$

e fica mais bonito ao dividir por s_1^2, dando uma medida do desvio padrão relativo ao valor do do produto:

$$\frac{\sigma_s^2}{s_1^2} = \frac{\sigma_y^2}{x_1^2} + \frac{\sigma_y^2}{y_1^2} \tag{5.49}$$

Voltamos ao quociente:

$$\begin{aligned} r &= \frac{x_1}{y_1}\frac{1 + \frac{\epsilon_x}{x_1}}{1 + \frac{\epsilon_y}{y_1}} \approx r_1(1 + \frac{\epsilon_x}{x_1})(1 - \frac{\epsilon_y}{y_1}) \\ &\approx r_1(1 + \frac{\epsilon_x}{x_1} - \frac{\epsilon_y}{y_1}) \\ &\approx r_1 + \frac{\epsilon_x}{y_1} - x_1\frac{\epsilon_y}{y_1^2} \\ &= r_1 + \eta_1' + \eta_2' \end{aligned} \tag{5.50}$$

onde $\eta_1' = \frac{\epsilon_x}{y_1}$ e $\eta_2' = x_1\frac{\epsilon_y}{y_1^2}$, [13] portanto $\sigma_r^2 = \sigma^2_{\eta_1'} + \sigma^2_{\eta_2'}$ de onde segue que

[13] Porque não há erro de sinal?

$$\sigma_r^2 = \frac{\sigma_x^2}{y_1^2} + x_1^2\frac{\sigma_y^2}{y_1^4} \tag{5.51}$$

que ao dividir por r_1^2 fica igual à relação análoga para o produto

$$\frac{\sigma_r^2}{r_1^2} = \frac{\sigma_y^2}{x_1^2} + \frac{\sigma_y^2}{y_1^2}. \tag{5.52}$$

É talvez surpreendente que as variâncias para a razão e o produto, dadas por 5.48 e 5.52, tenham a mesma forma. Afinal nenhuma das duas é exata, mas a aproximação que o produto e a razão de duas variáveis gaussianas é por sua vez também gaussiana joga fora as diferenças. O estudante já deve ter visto as expressões para o *erro relativo* $\frac{\sigma_s}{s_1}$ e $\frac{\sigma_r}{r_1}$ em aulas de laboratório.

De volta ao quociente e caudas gordas

Vamos fazer as contas para encontrar a distribuição quociente $P(r|x_1, y_1, \sigma_x, \sigma_y, \mathcal{M}_{\text{razão}})$ no caso simples onde $x_1 = y_1 = 0$ e que denotaremos $P(r)$. Usaremos o atalho via a função δ, pois $P(r|xy) = \delta(r - x/y)$ [14] :

$$\begin{aligned} P(r) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(r,x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(r|x,y)P(x)P(y)dxdy \\ &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta(r-\frac{x}{y})e^{-\frac{x^2}{2\sigma_x^2}}e^{-\frac{(y^2}{2\sigma_y^2}}dxdy \\ &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\delta(r-\frac{x}{y})e^{-\frac{x^2}{2\sigma_x^2}}dx\right)e^{-\frac{y^2}{2\sigma_y^2}}dy \end{aligned} \tag{5.53}$$

[14] O aluno semiatento perguntará porque aparece xy como condicionante se estamos falando de quociente. Lembre se que como condicionante xy é a asserção "o valor de X está no intervalo $x, x+dx$ E o valor de Y está no intervalo $y, y+dy$", portanto um produto lógico das asserções que trazem a informação sobre x e y.

Mudamos variáveis de integração na integral interna $u = x/|y|$ onde y é mantida constante, portanto $|y|du = dx$ e

$$\frac{x^2}{2\sigma_x^2} = \frac{y^2u^2}{2\sigma_x^2}$$

$$\begin{aligned} P(r) &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\delta(r-u)e^{-\frac{y^2u^2}{2\sigma_x^2}}|y|du\right)e^{-\frac{y^2}{2\sigma_y^2}}dy \\ &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty}\left(e^{-\frac{y^2r^2}{2\sigma_x^2}}\right)e^{-\frac{y^2}{2\sigma_y^2}}|y|dy = \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty}e^{-\frac{y^2}{2A^2}}|y|dy, \\ &= \frac{1}{\pi\sigma_x\sigma_y}\int_0^{\infty}e^{-\frac{y^2}{2A^2}}|y|dy = \frac{1}{\pi\sigma_x\sigma_y}\int_0^{\infty}e^{-\frac{y^2}{2A^2}}ydy \end{aligned}$$

onde chamamos $A^{-2} = \frac{r^2}{\sigma_x^2} + \frac{1}{\sigma_y^2} = \frac{r^2\sigma_y^2+\sigma_x^2}{\sigma_x^2\sigma_y^2} = \frac{r^2+\sigma_x^2/\sigma_y^2}{\sigma_x^2}$. A integral em y é simples, pois fazendo a mudança de variáveis $\frac{y^2}{2A^2} = v$ obtemos

$$\begin{aligned} P(r) &= \frac{A^2}{\pi\sigma_x\sigma_y}\int e^{-v}dv \\ &= \frac{A^2}{\pi\sigma_x\sigma_y} \\ &= \frac{1}{\pi}\frac{\sigma_x}{\sigma_y}\frac{1}{r^2+\frac{\sigma_x^2}{\sigma_y^2}}. \end{aligned} \tag{5.54}$$

Obtivemos uma distribuição que não é gaussiana. Como vimos antes esta é a distribuição de Cauchy

$$P(r|a) = \frac{a}{\pi}\frac{1}{r^2+a^2}. \tag{5.55}$$

e é interessante notar que para valores de r grandes ela decai lentamente, que é chamado de cauda gorda. Isso faz com que as integrais fiquem mais complicadas. Por exemplo seu valor esperado $\mathbb{E}(r)$ tem que ser calculado redefinindo o que significa a integração de $-\infty$ a ∞ [15]

$$\mathbb{E}(r) = \lim_{R\to\infty}\int_{-R}^{R} rP(r|a)dr = 0. \tag{5.56}$$

[15] Este é o valor Principal de Cauchy.

Isso terá consequências importantes quando olharmos para aplicações. Mais importante ainda é que a variância é infinita, mas o parâmetro a mede a largura da distribuição no seguinte sentido. Para $r = a$, $P(r = a|a) = \frac{1}{2}P(r = 0|a)$ então a é o valor de r a *meia altura* e a largura da distribuição a meia altura é $2a$.

Apêndice A: Normalização Gaussiana

Começamos em 1 dimensão. Chamamos

$$I_c = c \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{5.57}$$

e queremos encontrar c tal que $I_c = 1$. Primeiro mudamos a variável x por um deslocamento μ: $x_{\text{novo}} = x_{\text{velho}} - \mu$. Nem a medida de integração nem os limites mudam, portanto

$$I_c = c \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx. \tag{5.58}$$

Não sabemos calcular analiticamente

$$\int_{-\infty}^{y} e^{-\frac{x^2}{2\sigma^2}} \tag{5.59}$$

em termos de funções simples conhecidas, o que força à introdução de uma nova função. A expressão acima está relacionada ao que é conhecida como função erro [16]. Gauss fez um truque que parece um retrocesso. Tentou calcular I_c^2:

[16] Para referência futura, a função erro é definida por $\text{erf}(y) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{y} \exp(-t^2) dt$.
Mudando variáveis, pode ser escrita como:
$\text{erf}(\frac{x}{\sqrt{2}}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{1}{2}t^2) dt$

$$I_c^2 = c^2 \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}} dy \tag{5.60}$$

Escrevemos a variável de integração na segunda integral como y, pois agora podemos escrever

$$I_c^2 = c^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{(x^2+y^2)}{2\sigma^2}} dxdy. \tag{5.61}$$

O truque vem de perceber que podemos interpretar a integral acima como a integral da função de duas variáveis $e^{-\frac{(x^2+y^2)}{2\sigma^2}}$ sobre todo o plano, (x, y). Podemos dividir o plano em pequenos elementos numa grade quadrada $\Delta x \Delta y$ e tomar os limites necessários, ou podemos dividí-lo em setores circulares onde $x^2 + y^2$ toma valor constante r^2. Isto é, usamos coordenadas polares. O estudante deve olhar o texto de cálculo necessário. As relações que permitem mudar as variáveis de integração são

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \\ r &= \sqrt{x^2+y^2} \\ \theta &= \arctan\frac{y}{x} \\ dxdy &\rightarrow rdrd\theta. \end{aligned} \tag{5.62}$$

Os limites de integração para as novas variáveis são $0 \leq r < \infty$ e $0 \leq \theta < 2\pi$ Assim

$$\begin{aligned} I_c^2 &= c^2 \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2\sigma^2}} rdrd\theta \\ &= 2\pi c^2 \int_0^{\infty} e^{-\frac{r^2}{2\sigma^2}} rdr \end{aligned}$$

pois a integral em θ é 2π. Agora o preço de fazer duas integrais não parece tão caro, pois a integral angular foi trivial. A vantagem de tudo isto é o aparecimento do fator r no elemento de área. Podemos

mudar novamente variáveis: $u = \frac{r^2}{2\sigma^2}$, e para o diferencial temos $du = \frac{rdr}{\sigma^2}$, que leva a

$$\begin{aligned} I_c^2 &= 2\pi\sigma^2 c^2 \int_0^\infty e^{-u} du \\ &= 2\pi\sigma^2 c^2 \left(-e^{-u}\right)_0^\infty \\ &= 2\pi\sigma^2 c^2 \end{aligned} \tag{5.63}$$

portanto, para que $I_c = 1$ devemos ter

$$c = \frac{1}{\sqrt{2\pi}\sigma} \tag{5.64}$$

Completando quadrados

É frequente encontrar integrais do tipo

$$A(h) = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2} + hx} \frac{dx}{\sqrt{2\pi}\sigma}.$$

É muito fácil de calcular pois podemos usar o resultado para a normalização da gaussiana

$$1 = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \frac{dy}{\sqrt{2\pi}\sigma} \tag{5.65}$$

Mudamos a variável de integração $y = x - a$ para obter

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} e^{-\frac{(x-a)^2}{2\sigma^2}} \frac{dx}{\sqrt{2\pi}\sigma} \\ &= \int_{-\infty}^{\infty} e^{-\frac{(x^2 - 2ax + a^2)}{2\sigma^2}} \frac{dx}{\sqrt{2\pi}\sigma} \\ &= e^{-\frac{a^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2} + x\frac{a}{\sigma^2}} \frac{dx}{\sqrt{2\pi}\sigma}. \end{aligned} \tag{5.66}$$

Agora podemos escolher a para fazer $h = \frac{a}{\sigma^2}$, de onde segue que

$$e^{\frac{h^2\sigma^2}{2}} = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2} + xh} \frac{dx}{\sqrt{2\pi}\sigma}. \tag{5.67}$$

O resultado também pode ser obtido de uma forma similar, mas que sugere o nome "completar quadrados."Podemos reescrever o expoente do integrando de $A(h)$

$$\begin{aligned} -\frac{x^2}{2\sigma^2} + hx &= -\frac{1}{2\sigma^2}\left(x^2 - 2(h\sigma^2)x\right) \\ &= -\frac{1}{2\sigma^2}\left(x^2 - 2(h\sigma^2)x + h^2\sigma^4 - h^2\sigma^4\right) \\ &= -\frac{1}{2\sigma^2}\left(x - (h\sigma^2)\right)^2 + \frac{h^2\sigma^2}{2}, \end{aligned} \tag{5.68}$$

ou seja, somamos e subtraimos $h^2\sigma^4$ para completar um quadrado perfeito. Muda variaveis, fazendo uma translação e usamos a integral da normalização.

Um comentário talvez fora de lugar, mas que torna 5.67 muito interessante, é notar que a variância, no termo exp $\frac{h^2\sigma^2}{2}$ aparece no numerador em lugar do denominador. Acreditando que os símbolos das integrais mantenham o significado mesmo para integração de variáveis complexas e que a expressão mantém-se válida mesmo que h seja complexo, em particular se substituirmos $h \to ik$ obtemos

$$e^{-\frac{k^2\sigma^2}{2}} = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2} + ixk} \frac{dx}{\sqrt{2\pi}\sigma}. \tag{5.69}$$

É fácil mostrar, mudando variáveis, que

$$e^{-\frac{x^2}{2\sigma^2}} = \int_{-\infty}^{\infty} e^{-\frac{k^2\sigma^2}{2} - ixk} \frac{\sigma dk}{\sqrt{2\pi}}. \tag{5.70}$$

Este é um exemplo, talvez o mais simples, de uma par de funções que estão relacionadas por uma operação que se chama transformada de Fourier. A importância desta área não pode ser exagerada, tanto pelas suas aplicações em ciência quanto pela beleza e riqueza em matemática. Nestas notas voltaremos a falar de transformada de Fourier ao tratar das distribuições de somas de variáveis estocásticas e o teorema do limite central pois a exponencial pode ser escrita em série de potências como

$$e^{ixk} = \sum_{n=0}^{\infty} \frac{(ixk)^n}{n!}$$

e portanto a transformada de Fourier de uma densidade de probabilidade é:

$$\begin{aligned} \Phi(k) := \langle e^{ixk} \rangle &= \int_{-\infty}^{\infty} P(x|I) e^{ixk} dx \\ &= \int_{-\infty}^{\infty} P(x|I) \sum_{n=0}^{\infty} \frac{(ixk)^n}{n!} dx \\ &= \sum_{n=0}^{\infty} \frac{\mathbb{E}(x^n)(ik)^n}{n!} \end{aligned} \tag{5.71}$$

A função $\Phi(k)$ é chamada função característica da variável aleatória X e sua expansão em série de potências de k tem coeficientes $\frac{E(x^n)i^n}{n!}$. Portanto se a função característica de uma variável for conhecida, uma simples expansão de Taylor nos dará os valores esperados. Este tipo de técnica é muito útil e está associado à ideia de função geratriz, usadas inicialmente por Euler em teoria de números e posteriormente por Laplace [17]. Esta parte do apêndice está um pouco acima do que se precisa neste curso, mas não do que se precisa na vida real dos físicos. O aluno não deve desanimar, mas ao contrário ficar animado com o fato que há um mundo de coisas interessantes para aprender que fornecerão ferramentas para quando for estudar os verdadeiros problemas de pesquisa.

[17] "A generating function is a device somewhat similar to a bag. Instead of carrying many little objects detachedly, which could be embarrassing, we put them all in a bag, and then we have only one object to carry, the bag. Quite similarly, instead of handling each term of the sequence a_0, a_1, $a_2, \cdots$ individually, we put them all in a power series $\sum a_n x^n$, and then we have only one mathematical object to handle, the power series."George Pólya, Induction and Analogy in Mathematics.

Mais de uma dimensão: Normais multivariadas

Em mais de uma dimensão o problema passa por algum conhecimento de propriedades de matrizes. Começamos por considerar N variáveis normais independentes, de média nula e variância 1. Formamos um arranjo $\boldsymbol{u} = (x_1, \cdots x_N)$. Alguns estarão tentados a usar o nome vetor para este arranjo, mas devemos resistir. Devemos guardar esse nome para situações em que há um significado especial para as variáveis, como por exemplo coordenadas em um espaço Cartesiano. Podemos simplesmente listar características de um sistema genérico e isso não faz do arranjo um vetor [18]. A distribuição de probabilidades é obtida pela regra do produto lógico

$$\begin{aligned} P(x_1 \cdots x_N, I) d^N x &= P(x_1 | x_2 \cdots x_N, I) P(x_2 \cdots x_N | I) d^N x \\ &= P(x_1|I) P(x_2|I) \cdots P(x_N|I) d^N x \\ &= \prod_{i=1}^{N} P(x_i|I) d^N x = \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}\sum_i x_i^2} d^N x \\ &= \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}\boldsymbol{u}^T \boldsymbol{u}} d^N x \end{aligned}$$

[18] A representação de uma mesa pode ser feita por um arranjo $\{$ altura , número de lados, número de pernas, número do varniz no catálogo da Acme,...$\}$. Isto não é um vetor.

Agora consideramos uma matriz $N \times N$ não singular, $\boldsymbol{A}$ e sua transposta $\boldsymbol{A}^T$ [19] e um arranjo constante $\boldsymbol{y}_0$ e fazemos a transformação de variáveis

$$\boldsymbol{u} = \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{y}_0), \quad \boldsymbol{u}^T = (\boldsymbol{y}^T - \boldsymbol{y}_0^T)\boldsymbol{A}^T.$$

[19] Os elementos estão relacionados por $(\boldsymbol{A})_{ij} = (\boldsymbol{A}^T)_{ji}$. Uma consequência é $\det \boldsymbol{A} = \det \boldsymbol{A}^T$.

O Jacobiano da transformação é [20]

$$|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}}| = |\det \boldsymbol{A}| = \frac{1}{\sqrt{|\det \boldsymbol{C}|}},$$

[20] Usamos $\det AB = \det A \det B$.

onde $\boldsymbol{C}^{-1} = \boldsymbol{A}^T\boldsymbol{A}$. A mudança de variáveis $\boldsymbol{u} \to \boldsymbol{y}$ leva a

$$\begin{aligned} P(\boldsymbol{y})d^N y &= P(\boldsymbol{u})|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}}|d^N y \\ &= \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{1}{\sqrt{\det|\boldsymbol{C}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{y}_0)^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{y}_0)\right). \end{aligned} \tag{5.72}$$

A matriz $\boldsymbol{C}$ tem um papel semelhante à variância σ^2 no caso de uma dimensão. Mas aqui inclui valores esperados de variáveis distintas

$$C_{ij} = \mathbb{E}((y_i - y_{0i})(y_j - y_{0j})) \tag{5.73}$$

que por descrever como *co*variam duas componentes é chamada de matriz de **covâriancia.** Para mostrar isso definimos

$$Z(\boldsymbol{J}) = \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{1}{\sqrt{\det|\boldsymbol{C}|}} \int \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{y}_0)^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{y}_0) + (\boldsymbol{y}-\boldsymbol{y}_0).\boldsymbol{J}\right). \tag{5.74}$$

que é útil pois ao derivar com respeito a J_i cai um fator $y_i - y_{0i}$, que ajuda a calcular os valores esperados:

$$\frac{\partial^2 \ln Z}{\partial J_i \partial J}|_{J=0} = \mathbb{E}((y_i - y_{0i})(y_j - y_{0j}))$$

Por simplicidade, podemos tomar $\boldsymbol{y}_0 = 0$. Completando quadrados, ao somar e subtrair $\frac{1}{2}\boldsymbol{J}^T\boldsymbol{C}\boldsymbol{J}$ no expoente, obtemos:

$$\begin{aligned} Z(\boldsymbol{J}) &= \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{1}{\sqrt{\det|\boldsymbol{C}|}} \int \exp\left(-\frac{1}{2}(\boldsymbol{y})^T\boldsymbol{C}^{-1}(\boldsymbol{y}) + (\boldsymbol{y}).\boldsymbol{J}\right) + \frac{1}{2}\boldsymbol{J}^T\boldsymbol{C}\boldsymbol{J} - \frac{1}{2}\boldsymbol{J}^T\boldsymbol{C}\boldsymbol{J} \\ &= \exp(\frac{1}{2}\boldsymbol{J}^T\boldsymbol{C}\boldsymbol{J})\left(\frac{1}{(2\pi)^{\frac{N}{2}}}\frac{1}{\sqrt{\det|\boldsymbol{C}|}} \int \exp\left(-\frac{1}{2}(\boldsymbol{y}-A^{-1}\boldsymbol{J})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-A^{-1}\boldsymbol{J})\right)\right) \\ &= \exp(\frac{1}{2}\boldsymbol{J}^T\boldsymbol{C}\boldsymbol{J}) \end{aligned} \tag{5.75}$$

onde usamos a integral da normalização. É fácil tomar as derivadas e mostrar 5.73. A função $Z(\boldsymbol{J})$ é novamente uma exemplo de uma função geratriz e $\boldsymbol{J}$ é chamado em Física de fonte. O estudante que já tenha estudado transformada de Laplace a reconhecerá no desenvolvimento acima.

É comum que em aplicações tenhamos informações sobre as variáveis y_i e sobre as covariâncias C_{ij}. Encontrar a transformação $\boldsymbol{A}^{-1}$ é interessante, pois leva de variáveis correlacionadas (y) a variáveis independentes (x) que são combinações lineares das variáveis que tipicamente são as variáveis originais na descrição de um problema. Estas transformações são de extrema importância em qualquer área da Física assim como em muitas outras áreas da ciência.

Exercício Para dois vetores $\mathbf{u}, \mathbf{v}$ num espaço vetorial onde o produto interno $\mathbf{u}.\mathbf{v}$ é definido, podemos provar a desigualdade de

Cauchy-Schwarz, que remonta a meados do século 19. Seja a norma definida por $|\mathbf{u}| = (\mathbf{u}.\mathbf{u})^{1/2}$, então

$$|\mathbf{u}.\mathbf{v}|^2 \leq |\mathbf{u}|^2|\mathbf{v}|^2 \tag{5.76}$$

Estude a prova deste teorema e use para provar que para cada elemento da matriz de covariância vale

$$|C_{ij}|^2 \leq I\!E((y_i - y_{0i})^2)I\!E((y_j - y_{0j})^2) = \text{Var}(y_i - y_{0i})\text{Var}(y_j - y_{0j}). \tag{5.77}$$

e como a variânia é invariante por translações

$$|C_{ij}|^2 \leq \text{Var}(y_i)\text{Var}(y_j), \tag{5.78}$$

ou

$$-1 \leq \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} \leq 1. \tag{5.79}$$

Pense sobre o que significa esta desigualdade em geral e o que significa caso seja satisfeita como igualdade.

Apêndice B: Distribuição χ^2

Defina $\chi^2 = \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}$, que por sua vez é uma variável aleatória. Qual é a sua distribuição? Devemos olhar para a densidade de probabilidade de que χ^2 caia num determinado intervalo $d\chi^2$ que é dada por

$$P(\chi^2) = \int ... \int \delta\left(\chi^2 - \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}\right) \frac{1}{\left(\sqrt{2\pi}\right)^n \sigma_1\sigma_2...\sigma_n} exp(-\sum \frac{x_i^2}{2\sigma_i^2}) dx_1...dx_n.$$

Mude variáveis $y = x/\sigma$. Mostre que o resultado é $P(\chi^2) = K_n \chi^{n-1} e^{-\frac{\chi^2}{2}}$. Não é necessário, mas sim um desafio, encontrar o valor de $K_n = (2^{\frac{(n-2)}{2}}(n/2-1)!)^{-1}$. {Você pode dar uma interpretação para esta distribuição? Considere n dados e uma hipótese sobre esses dados que dá uma estimativa -junto com uma estimativa das variâncias . A partir dos erros x_i podemos formar então χ^2.....Pense um pouco sobre o que ocorreria se há erros sistematicos. Qual é o valor de χ^2mais provável e qual a largura da distribuição }

Apêndice C: Stirling

A função Gama é definida para nossos propósitos imediatos sobre os números reais positivos:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt, \tag{5.80}$$

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

Podemos escrever

$$\Gamma(x+1) = \int_0^\infty e^{-t} t^x dt = \int_0^\infty e^{f(t;x)} dt \tag{5.81}$$

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 pela série de Taylor

$$f(t;x) = x\log x - x - \frac{1}{2x}(x-t)^2. \tag{5.82}$$

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

$$\begin{aligned}\Gamma(x+1) &\approx e^{x\log x - x}\int_{-x}^{\infty} e^{-\frac{1}{2x}(x-t)^2}dt \\ &\approx e^{x\log x - x}\int_{-\infty}^{\infty} e^{-\frac{1}{2x}(x-t)^2}dt \\ &= \sqrt{2\pi x}e^{x\log x - x}\end{aligned} \tag{5.83}$$

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

$$\log n! = \log\Gamma(n+1) = (n+\frac{1}{2})\log n - n + \frac{1}{2}\log 2\pi, \tag{5.84}$$

mostrando de forma intuitiva a equação 5.4. 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 5.84, mas o leitor deverá procurar outros textos (e.g. [21])

[21]

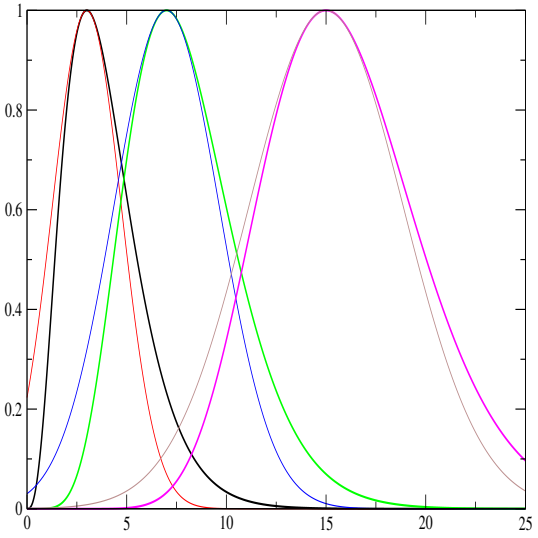


Figura 5.4: A figura mostra três pares de gráficos, para $x = 3,7$ e 15 respectivamente. Cada par é formado pelas funções $e^{-t}t^x/(x^xe^{-x})$ e $\exp(-(x-t)^2/(2x))$

Uma expansão sistemática, que não precisamos considerar aqui leva a

$$\log n! = (n+\frac{1}{2})\log n - n + \frac{1}{2}\log 2\pi + \log\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \mathcal{O}(\frac{1}{n^3})\right) \tag{5.85}$$

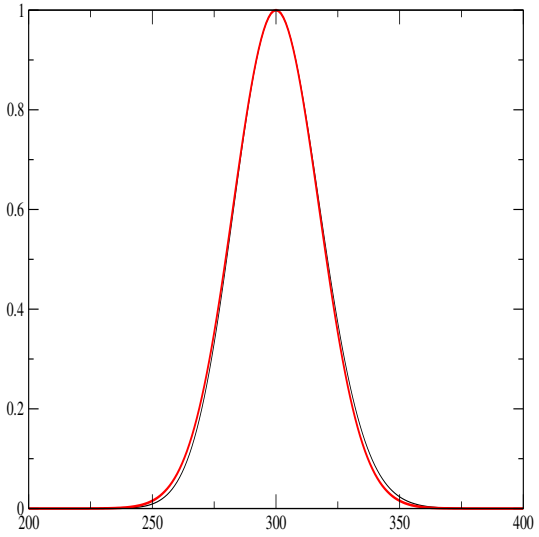


Figura 5.5: O mesmo que a figura anterior para x=300. Note aqui, assim como na figura anterior, que ao usar a gaussiana que é simétrica em relação ao máximo, o erro na integral cometido à esquerda do máximo tem sinal oposto ao cometido à direita.

6

Aplicações da regra de Bayes

Houve uma mudança muito grande nas últimas décadas quanto à difusão e popularidade de métodos de inferência Bayesianos. Enquanto ninguém nunca discutiu a validade do teorema de Bayes como uma relação entre diferentes probabilidades condicionais, houve e ainda há quem não o aceite como base de inferência. Aqui o ponto central é sobre a interpretação de probabilidade como uma frequência versus como representação de crenças em asserções. A discussão será feita em vários capítulos e no fim o estudante deverá escolher suas definições preferidas, fazer suas alianças ou talvez melhor, fornecer as suas próprias definições. Sobre o uso de um teorema para fazer inferência acredito que sempre devemos ter preocupação com a aplicação de teoremas no mundo real e tocaremos novamente neste tema ao falar de entropia. Do ponto de vista informacional, o formalismo de entropia é construido para fazer inferência, é um mecanismo mais geral e engloba inferência Bayesiana nos casos em que a informação, na forma de dados obtidos por medidas são usados. Por agora inferência deve ser entendida como um processo de mudar crenças, codificadas em distribuições de probabilidades. Então tentarei de forma sistemática me referir ao teorema de Bayes quando falar de uma relação entre probabilidades condicionais. Quando estiver interessado em fazer inferência usarei a regra de Bayes. A expressão matemática, a fórmula, do teorema e da regra são os mesmos. Mas é bom não confundí-los. A regra de Bayes é usar o teorema de Bayes para fazer inferência

É interessante olhar para vários casos em que a regra de Bayes nos fornece resultados de inferência em acordo com o bom senso, para poder se acostumar com esta forma de pensar. Primeiro olharemos se faz sentido para um lógico, ao sair de casa levar um guarda-chuva simplesmente porque há nuvens.

A regra de Bayes e Informação Incompleta

Exemplo 1: Chuva e Sol

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”

- $C =$ "$A \to B$"

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 a regra 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

$$P(A|BCI) = \frac{P(A|CI)P(B|ACI)}{P(B|CI)}, \tag{6.1}$$

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

$$P(B|ACI) = 1. \tag{6.2}$$

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

$$P(B|CI) \leq 1. \tag{6.3}$$

Substituindo estes resultados obtemos

$$P(A|BCI) \geq P(A|CI), \tag{6.4}$$

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: não chove. Continuaremos a insistir que não podemos dizer nada sobre B do ponto de vista da lógica? A regra de Bayes, nos diz

$$P(B|\overline{A}CI) = \frac{P(B|CI)P(\overline{A}|BCI)}{P(\overline{A}|CI)}, \tag{6.5}$$

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

$$\begin{aligned} 1 - P(\overline{A}|BCI) &\geq 1 - P(\overline{A}|CI) \\ P(\overline{A}|BCI) &\leq P(\overline{A}|CI) \\ \frac{P(\overline{A}|BCI)}{P(\overline{A}|CI)} &\leq 1 \end{aligned} \tag{6.6}$$

e

$$P(B|\overline{A}CI) \leq P(B|CI) \tag{6.7}$$

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 2: Teste Médico

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 que temos como dadas ou que queremos investigar

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

A validade destas asserções devera ser estudada na situação informacional descrita pelos dados sobre

- especificidade: $P(A|D) = .90$, a probabilidade de dar positivo no teste na condição de estar doente
- sensibilidade: $P(\overline{A}|\overline{D}) = 1 - P(A|\overline{D}) = 1. - .2 = .8$, a probabilidade de teste NÃO 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 a regra de Bayes. Assim poderemos calcular $P(D|AI)$ que é o que realmente interessa, a probabilidade de ter a doença quando o teste deu positivo,

$$P(D|AI) = \frac{P(D|I)P(A|DI)}{P(A|I)}, \tag{6.8}$$

e também

$$P(\overline{D}|AI) = \frac{P(\overline{D}|I)P(A|\overline{D}I)}{P(A|I)}, \tag{6.9}$$

os denominadores são inconvenientes e os eliminamos olhando para a razão

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)}. \tag{6.10}$$

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 informação 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 $\frac{P(D|I)}{P(\overline{D}|I)} = 0.01/.99$, só 1% da população tem a doença. Segue que

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)} = \frac{.01 \times .90}{.99 \times .20} = 0.045. \tag{6.11}$$

ou seja a probabilidade de não ter a doença é aproximadamente .95. Não devemos considerar que isto seja uma boa notícia, afinal a probabilidade que era de 0.01 de ter a doença passou para 0.045% : aumentou quase cinco vezes. Mas não devemos ainda entrar em pânico nem jogar fora a informação que ganhamos com o teste. O que fazer? A análise desta pergunta nos leva à questão de decisão, que não faz parte do objetivo destas notas. Certamente devemos passar a colher mais informação.

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, por isso acima a referência é à plausibilidade, mas podiamos ter simplesmente dito probabilidade.

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 $B|C = B|C'$, pois que influência pode ter saber sobre a água em Marte, sobre o que eu possa acreditar sobre o dia da semana. Também é razoável que a plausibilidade de que "haja vida em Marte e hoje seja segunda feira"dado que "há água em Marte"seja maior ou igual à plausibilidade que "haja vida em Marte e hoje seja segunda" dado que "não há agua em em Marte. "

Mas agora temos a regra de produto para as probabilidades ou plausibilidades regraduadas, e portanto podemos provar isto

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

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

Exemplo da regra de Bayes, ajuste de funções e estimativa de parâmetros

Uma das primeiras lições que os estudantes de física tem ao entrarem num laboratório é sobre ajuste de curvas e estimativa de parâmetros 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 complicados mais adiante nestas notas, olharemos para o caso em que obtemos um conjunto de dados

$$D = \{v_1, v_2, ...v_N\} \tag{6.12}$$

para as velocidades medidas em

$$T = \{t_1, t_2, ...t_N\}. \tag{6.13}$$

O modelo que temos em mente é

$$\mathcal{M} : v = v_0 + gt \tag{6.14}$$

Vamos supor que esse modelo está além da necessidade de discussões. Se não estivesse poderiamos querer avaliar, refutar ou aceitar, pelo menos até ter mais dados. Se houvesse outro modelo da mecânica poderiamos querer julgar o mérito entre os dois candidatos. Isto ajudaria a selecionar um modelo. Faremos isso mais tarde. Parece que a pergunta que queremos responder diz respeito a asserções do tipo

- $H(g)$:"O valor da aceleração da gravidade é g".

mas isto não está bem definido. O que queremos é analizar

- $H(g)$:"O valor da aceleração da gravidade esta entre o valor g e $g + \Delta g$".

Para cada valor de g que for inserido nessas 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 exercício é um exemplo do dia a dia dos físicos.

A regra de Bayes nos permite escrever

$$P(H|DI) = \frac{P(H|I)P(D|HI)}{P(D|I)}. \tag{6.15}$$

O que será discutido a seguir é fundamental para este curso. Será discutido em contextos mais compĺicados 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? Neste contexto temos o modelo $\mathcal{M}$ da equação 6.14.
- Quais são as características do aparelho de medida?
- Em que instantes de tempo t_i fizemos as medidas.
- Quais as incertezas que estas medidas têm?
- Em que planeta estamos?
- ...

e muito mais que ficará tácitamente escondido, mas ainda relevante.

D **é o conjunto de dados.** Representa a asserção sobre quais foram os dados medidos.

H **é a hipótese que queremos seja testada a respeito do parâmetro** g**.**

É importante notar que é fácil esquecer que é isto o que queremos avaliar.

Agora o significado das probabilidades que aparecem na equação 6.15.

Distribuição *a priori*

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 experiência 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 10 m/s, e pode ser constante entre 7 e 15. Muito mais que isso ou muito menos, deve ser erro, e é melhor jogar fora 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.

Verossimilhança[1]

A probabilidade $P(D|HI)$ descreve quão verossímel seria encontrar esse conjunto de dados se além de I, o valor particular de g

[1] *Likelihood* em inglés. Em linguagem corrente tem vários significados dependendo do contexto em que é usada, coloquialmente, em Direito, dentro de um texto literário, etc.

representado por H fosse o correto. Esta é a famosa contribuição do reverendo Thomas Bayes[2]: 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 porque se for pequeno podemos dizer que há poucas chances de que o valor do parâmetro dessa hipótese em particular tenha dado origem aos dados. Quando os nomes são dados em uma época e usados em outra pode ficar um pouco estranho. É obvio que esse termo é uma probabilidade e talvez o termo probabilidade inversa fosse mais útil para descrever seu significado. Na estatística da escola frequentista, a verossimilhança não é uma probabilidade porque os parâmetros de uma teoria tem uma existência ontologica e não se admite possam ser discutidos em termos de probabilidade.

[2] referencia de bayes

Evidência

O denominador $P(D|I)$ será interessante em outros contextos, em particular na comparação e seleção de modelos. Em geral é chamado de evidência. Mostraremos que é a evidência trazida pelos dados em favor do modelo considerado. 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 $g + dg$ e somas são substituidas por integrais.

Distribuição posterior

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. Em problemas de verdade as integrações são sobre espaços de dimensão muito grande. Os problemas práticos e teóricos associados a esta integração serão discutidos mais adiante, especificamente no capítulo de integração Monte Carlo.

Novamente, 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 [3]. 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 probabilidades codificam não é o valor de g, mas a crença que esse seja o valor correto.

[3] 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. Mas sobra ainda a utilidade de usar o modelo de Newton e nesse sentido queremos *determinar* g.

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,

$$v_i = v_0 + gt_i + \eta_i. \tag{6.16}$$

O resultado esperado puramente pelo modelo teórico (eq. 6.14) 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

$$\begin{aligned} P(\eta_1\eta_2....\eta_N|I_{exp}) &= P(\eta_1|I_{exp})P(\eta_2....\eta_N|\eta_1 I_{exp}) \\ &= P(\eta_1|I_{exp})P(\eta_2\eta_3....\eta_N|I_{exp}) \\ &= P(\eta_1|I_{exp})P(\eta_2|I_{exp})P(\eta_3....\eta_N|\eta_2 I_{exp}) \\ &... \\ &= \prod_i^N P(\eta_i|I_{exp}), \end{aligned} \tag{6.17}$$

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

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

Mas pelo modelo da equação 6.16, $\eta_i = v_i - v_0 - gt_i$. Isto pode ser interpretado como a probabilidade de obter v_i dada a informação (suposta verdeira) que a medida foi feita em t_i e a aceleração da gravidade é g. Portanto é a distribuição dos dados condicionada à hipótese. Portanto

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

Juntando tudo obtemos a posterior

$$P(H|DI) = \frac{P(H|I)}{PD|I)} \frac{e^{-\sum_{i=1}^N \frac{(v_i - v_0 - gt_i)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}. \tag{6.18}$$

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 podemos nos contentar com o valor de g que é mais provável g_{MAP}, isto recebe o nome de *máximo a posteriori* ou a moda da distribuição posterior. Se a distribuição *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,

$$g_{MV} = \arg\min_g \sum_{i=1}^{N} \frac{(v_i - v_0 - g t_i)^2}{2\sigma^2} \tag{6.19}$$

que é o velho método de mínimos quadrados. Mas escolher um valor sobre os outros esconde que não temos certeza absoluta. A largura da posterior nos dá uma medida da incerteza. Por simplicidade olhamos para *a priori* uniforme. Neste caso ou mesmo para distribuições *a priori* $\propto \exp(-ag^2 + bg)$ a distribuição de g é gaussiana. e temos

$$P(g|DI) \propto e^{-g^2(a+\sum \frac{t_i^2}{2\sigma_i^2})+\cdots},$$

ou seja, a variância da distribuição de g já vem com uma estimativa do *erro* da medida *estimativa* de g.

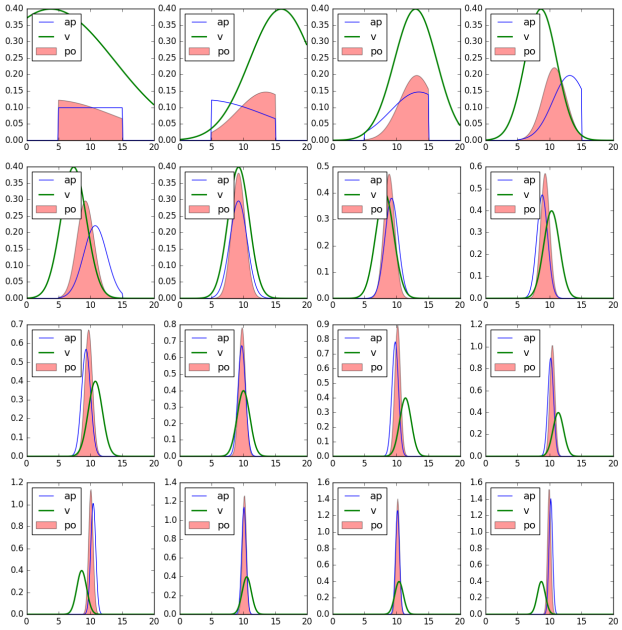


Figura 6.1: Distribuição *a priori* $P(g|D_n I)$ (ap, azul), verossimilhança $P(v_{n+1}|g, t_{n+1})$ (v, verde) e distribuição posterior $P(g|D_{n+1} I)$ (po, vermelho) como funções de g. Iniciando com uma distribuição *a priori* uniforme a posterior é obtida multiplicando pela verossimilhança e renormalizando. A posterior se transforma na *a priori* para a chegada de um novo dado. Isso significa que a curva azul de uma figura é a curva vermelha da figura anterior. A medida que os dados se acumulam a posterior se afina, dimimuindo a incerteza sobre g. As abscissas mostram a região de interesse de valores de g. Cada figura mostra as distribuições após a inclusão de um novo dado. A primeira é a distribuição *a priori* uniforme. Com o aumento do tempo a largura da verossimilhança também diminui. Note que como função de g a verossimilhança não está normalizada pois é uma distribuição de v_n e não de g.

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

$$g^* = \int g P(H|DI) dg, \tag{6.20}$$

que é o resultado da média das crenças de cada hipótese (valor de g) ponderado pelo peso da distribuição posterior, que representa quanto acreditamos em cada intevalo $(g, g + dg)$.

O painel 6.1 mostra o resultado de uma simulação do problema de estimativa de g no laboratório. É uma simulação Monte Carlo, técnica que será discutida mais adiante, do que esperamos ver no laboratório, caso as hipóteses descritas acima sejam razoáveis. A forma de inferência é sequencial ou *on-line*. Começamos de uma distribuição *a priori* uniforme de g entre 5 e 15 e fazemos uma medida

no instante de tempo t_1. O conjunto de dados é $D_1 = \{v_1\}$. Obtemos a posterior e o resultado é mostrado na figura na linha superior à esquerda de **??**. Colhemos um novo dado v_2 em t_2. Neste caso simples de dados independentes, tanto faz voltar ao começo e usar a *a priori* original e incluir a verossimilhança dos dois dados, ou usar a posterior obtida depois de um dado como a nova *a priori* e incluir este último dado. As figuras subsequentes mostram em verde a *a priori* depois de n dados e a posterior depois de $n+1$ dados.

Um pouco mais de mínimos quadrados

Ainda no problema de estudo da aceleração da gravidade podemos avançar um pouco mais. Vamos esquecer, pelo momento a distribuição a priori, de forma que a posterior é essencialmente a verosimilhança. Lembre que a posterior é distribuição probabilidade de g e a verossimilhança é dos dados, por isso uso a palavra essencialmente na frase anterior. O denominador (evidência) da regra de Bayes é importante. Assim

$$P(g|DI) \propto e^{-\sum_{i=1}^{N} \frac{(v_i - v_0 - g t_i)^2}{2\sigma^2}} \tag{6.21}$$

e a constante de proporcionalidade, que chamamos abaixo de C, é obtida por normalização. A relação acima vista como uma gaussiana da variável g leva a

$$\begin{aligned} P(g|DI) &= C \exp\left(-\frac{1}{2} g^2 \frac{\sum_{i=1}^{N} t_i^2}{\sigma^2} - g \frac{\sum_{i=1}^{N} t_i (v_i - v_0)}{\sigma^2} \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_G} e^{-\frac{(g - g_{MAP})}{2\sigma_G^2}} \end{aligned} \tag{6.22}$$

onde a média e variância da posterior são dadas por:

$$g_{MAP} = \frac{\sum_{i=1}^{N} t_i (v_i - v_0)}{\sum_{i=1}^{N} t_i^2} \tag{6.23}$$

$$\sigma_G^2 = \frac{\sigma^2}{\sum_{i=1}^{N} t_i^2} \tag{6.24}$$

Neste caso simples a média da posterior g^*, a estimativa de máxima verossimilhança g_{MV} e o máximo da posterior g_{MAP} coincidem

Exercício Mostre que igualando a derivada com respeito a g, como indicado na expressão 6.19, resulta em $g_{MV} = g_{MAP} = g^*$.

Chame $\bar{v^2} = \sum (v_i - v_0)^2 / N$, $\bar{vt} = \sum (v_i - v_0) t_i / N$ e $\bar{t^2} = \sum t_i^2 / N$, então

$$g_{MAP} = \frac{\bar{vt}}{\bar{t^2}} \tag{6.25}$$

$$\sigma_G^2 = \frac{\sigma^2}{N\bar{t^2}} \tag{6.26}$$

Podemos calcular o valor dos resíduos quadrados

$$\begin{aligned} \chi^2 &= \frac{N}{2\sigma^2} \left(\bar{v^2} - 2 g^* \bar{vt} + g^{*2} \bar{t^2} \right) \\ &= \frac{N}{2\sigma^2} \left(\bar{v^2} - \frac{(\bar{vt})^2}{\bar{t^2}} \right) \end{aligned} \tag{6.27}$$

Exercício É óbvio que χ^2 não pode ser negativo, pois é o valor mínimo do resíduos quadrados. Mostre isto sem usar essa informação. Procure saber sobre a desigualdade de Cauchy-Schwarz.

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.

Exercício Se a medida for repetida N vezes, sempre no mesmo tempo t_1, podemos ver que a variância cai com $1/N$, que é um resultado típico que encontraremos muitas vezes. Agora discuta se é razoável que os desvios de cada medida sejam iguais. Suponha que não sejam e encontre o valor de g_{MAP} sob essas condições.

Exercício O modelo que acabamos de estudar é linear e univariado. Podemos generalizar para casos não lineares ou multivariados ou ainda não lineares e multivariados. Tente generalizar o método acima.

Moeda recarregada

Retomamos o problema do final do capítulo anterior. Podemos falar de uma urna de composição incerta ou analogamente de jogadas similares de uma moeda. Usaremos a linguagem do lançamento de uma moeda [4]. Uma moeda que pertenceu a um executivo de uma grande empresa estatal é lançada e se a face que acaba ficando para cima é cara, $s = 1$. Se for coroa $s = -1$. Após N jogadas a informação é guardada em $D_N = (s_1, s_2, ..., s_n)$, que pode ser comprimido na informação (n, m), com n caras e $m = N - n$ coroas. Talvez haja motivo para achar que há algo estranho com a moeda e que as duas faces não sejam igualmente prováveis. Investiguemos. O problema com que nos defrontamos é encontrar o valor de um parâmetro p para o qual não temos informação completa e portanto procuramos uma distribuição $P(p|D_n)$ que codifique a informação disponível. Este problema é conjugado ao de, dado o valor de p e o número de jogadas N devemos atribuir probabilidades ao valor de n. Suponha que codificamos nossa crença sobre os diferentes valores de p com uma distribuição *a priori* $P_0(p)$. Uma vez colhido o primeiro valor s_1 passamos a

$$P(p|D_1) = \frac{P_0(p)P(s_1|p)}{P(D_1)} \tag{6.28}$$

A verossimilhança $P(s_1|p)$ é simples, especialmente se você expressar $P(s_1|p)$ em palavras. Se $s_1 = 1$ então queremos saber a probabilidade que saia cara quando a probabilidade de sair cara é p. Que é p.... E se $s_1 = -1$, queremos a probabilidade que saia coroa quando a probabilidade de sair cara é p. Portanto $1 - p$. Logo

$$P(s_1|p) = p^{\frac{s_1+1}{2}}(1-p)^{-\frac{s_1-1}{2}}$$

É talvez mais fácil introduzir a variável τ_i que conta as caras, é 1 se for cara e zero se não:

$$\tau_i = \frac{s_i + 1}{2}$$

[4] Sigo a apresentação em Sivia, após uma idéia de S. Gull . De fato o primeiro a pensar neste problema parece ter sido Laplace, que visitamos ao falar da regra da sucessão.

$$P(s_1|p) = p^{\tau_1}(1-p)^{1-\tau_1} \tag{6.29}$$

Cada vez que a moeda é jogada usamos a equação 6.28, mas usando como distribuição *a priori* a posterior obtida no passo anterior. Após N passos

$$P(p|D_N) \propto P(p|D_{N-1})P(s_N|p). \tag{6.30}$$

Como podemos escrever $P(p|D_{N-1})$ em termos de $P(p|D_{N-2})$ e podemos iterar, obtemos

$$\begin{aligned} P(p|D_1) &\propto P_0(p)P(s_1|p) \\ P(p|D_2) &\propto P_0(p)P(s_1|p)P(s_2|p) \\ P(p|D_3) &\propto P_0(p)P(s_1|p)P(s_2|p)P(s_3|p) \\ &\vdots \\ P(p|D_N) &\propto P_0(p)\prod_{i=1}^{N} P(s_i|p) \end{aligned} \tag{6.31}$$

Usando a equação 6.29, e notando que $\sum_{i=1}^{N} \tau_i = n$, o número de caras obtemos

$$P(p|D_N) = \frac{1}{\mathcal{N}} P_0(p) p^n (1-p)^{N-n}. \tag{6.32}$$

onde escrevemos explicitamente $\mathcal{N}$ que garante a normalização. O estudante pode neste momento achar que encontramos a função binomial que descrevia a probabilidade n caras e $m = N - n$ coroas quando a probabilidade de cara é p. Mas estaria enganado! A binomial dá $P(n|p, N)$ ou seja a probabilidade de n. A equação 6.32 é a probabilidade inversa. Neste momento não queremos entrar no tópico espinhoso de causalidade, que será deixado para mais tarde, mas Laplace teria dito que enquanto $P(n|p, N)$ descreve o efeito (n caras) devido à causa (p), a probabilidade $P(p|D_N)$ descreve a causa, dado o efeito. A variável sobre a qual não temos informação completa é o parâmetro contínuo p que toma valores no intervalo $[0, 1]$. Dito isso calculamos a normalização explicitamente:

$$P(p|D_N) = \frac{P_0(p)p^n(1-p)^{N-n}}{\int_0^1 P_0(p')p'^n(1-p')^{N-n}dp'}. \tag{6.33}$$

Para o caso em que a distribuição *a priori* é uniforme no intervalo

$$\begin{aligned} P(p|D_N) &= \frac{p^n(1-p)^{N-n}}{\int_0^1 p'^n(1-p')^{N-n}dp'} \\ &= \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)} p^n(1-p)^{N-n} \\ &= \frac{(N+1)!}{n!(N-n)!} p^n(1-p)^{N-n} \end{aligned} \tag{6.34}$$

portanto $p \sim Beta(n+1, m+1)$, pois reconhecemos a distribuição Beta com parâmetros $a = n+1$ e $b = m+1$.

A seguir simulamos alguns casos deste problema para investigar o efeito da escolha da distribuição *a priori* $P_0(p)$. O casos são

1. $P_0(p)$ é unforme refletindo total incerteza sobre o valor de p
2. $P_0(p)$ reflete a confiança na honestidade da moeda: é uma distribuição muito fina centrada em $1/2$.
3. $P_0(p)$ reflete a nossa certeza que o jogo não é honesto e portanto p está perto de 0, quando não sairão caras, ou perto de 1 quando não sairão coroas.

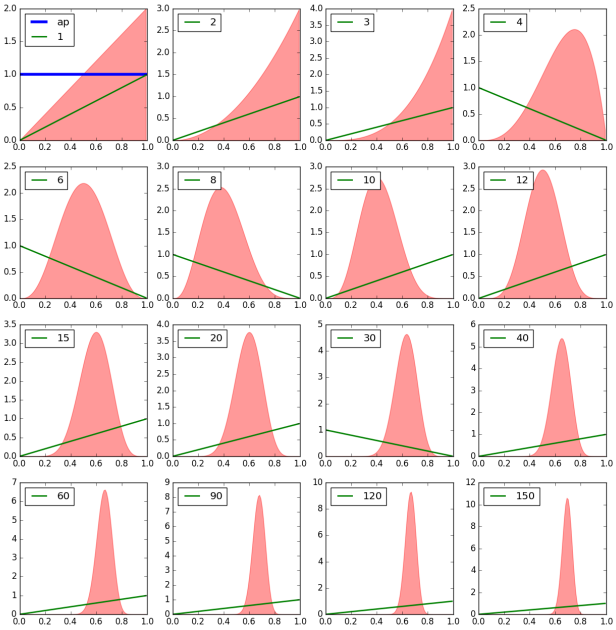


Figura 6.2: Distribuição posterior $P(p|D_N I)$. Iniciando com uma distribuição *a priori* uniforme a posterior é obtida multiplicando pela verossimilhança e renormalizando. A posterior se transforma na *a priori* para a chegada de um novo dado. A medida que os dados se acumulam a posterior se afina, dimimuindo a incerteza sobre p. O valor usado na simulação foi $p^* = 0.75$. A posterior se afina e fica concentrada na vizinhança de p^*. A legenda indica quanas jogadas da moeda foram levadas em conta. As curvas verdes (retas) são roporcionais à verossimilhança, p se o resultado foi cara e $1-p$ se foi coroa.

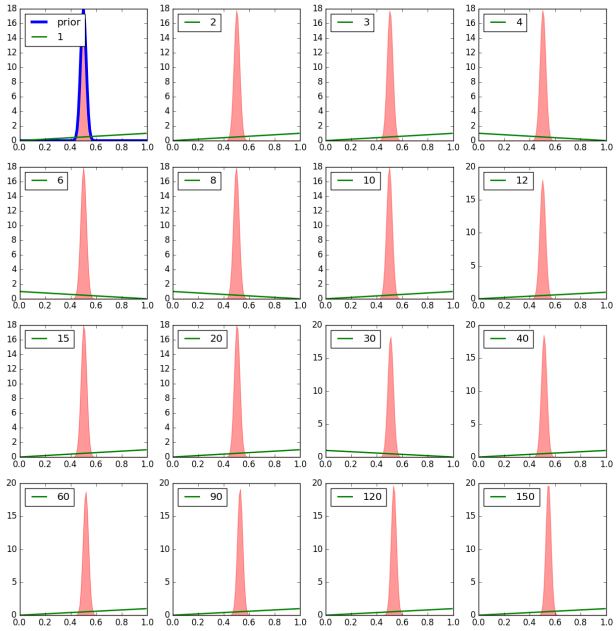


Figura 6.3: Igual que a figura anterior mas com uma distribuição *a priori* muito concentrada no centro. A convergência é muito mais lenta porque a crença inicial a região correta do parâmetro é muito pequena. Após 150 jogadas o valor MAP estimado é aproxiamdamente 0.55.

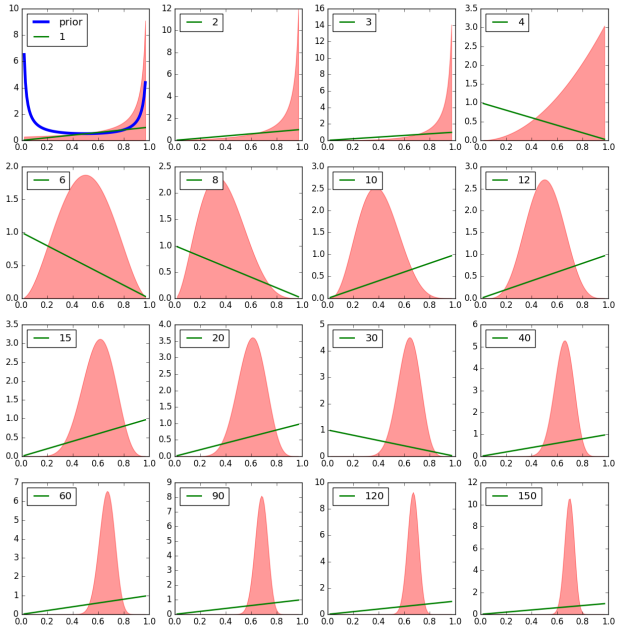


Figura 6.4: Igual que figura anterior mas com uma distribuição *a priori* muito desconfiada. A convergência é mais rápida que no caso anterior. Nos três casos a sequência de jogadas é a mesma.

Para o caso do prior uniforme, é fácil ver que o máximo da posterior, obtido a partir da derivada com respeito a p da equação 6.32

$$\begin{aligned} \frac{d}{dp}P(p|D_N)|_{p_{max}} &= 0 \\ np_{max}^{n-1}(1-p_{max})^{N-n} &= (N-n)p_{max}^{n}(1-p_{max})^{N-n-1} \\ p_{max} &= \frac{n}{N}. \end{aligned} \tag{6.35}$$

Isto está relacionado o que encontramos para a binomial, que o valor esperado $\langle n \rangle = pN$. Usando o resultado 3.18 de Euler

$$E_k^r = \int_0^1 p^r(1-p)^k dp = \frac{r!k!}{(r+k+1)!} \tag{6.36}$$

obtemos que

$$E_{N-n}^{n+1} = \int_0^1 p^{n+1}(1-p)^{N-n} dp = \frac{(n+1)!(N-n)!}{(N+2)!} \tag{6.37}$$

portanto o valor esperado de p é dado por

$$\begin{aligned} \mathbb{E}(p) &= \frac{E_{N-n}^{n+1}}{E_{N-n}^{n}} \\ &= \frac{(n+1)!(N-n)!}{(N+2)!}\frac{(N+1)!}{n!(N-n)!} \\ &= \frac{n+1}{N+2}. \end{aligned} \tag{6.38}$$

Os valores p_{max} e $\mathbb{E}(p)$ são bem próximos entre si para valores grandes de N e n. Obviamente a média empírica de τ_i é p_{max}:

$$\frac{1}{N}\sum_{i=1}^{N}\tau_i = \frac{n}{N} = p_{max}. \tag{6.39}$$

O interessante deste resultado não é que tenhamos obtido o óbvio, a média de sucessos é o valor mais provável do parâmetro p. O interessante é o método, pois na subseção 6.1.10 faremos um caso em que isso não é verdade. O cálculo da média não dá nenhuma informação sobre o parâmetro procurado. Mas antes uma nota de interesse histórico.

Bayes, a mesa de Bilhar e a distribuição a priori.

No trabalho de T. Bayes postumamente publicado por R. Price o problema atacado é essencialmente o da moeda, mas em outra linguagem. A informação a seguir será chamada como usualmente I. Considere uma mesa de bilhar quadrada de lado L, uniforme por construção. Uma bola rola em cima da superficie com atrito até parar. É jogada sem nenhum conhecimento das regras da dinâmica e pode parar em qualquer lugar. Chame $\theta = x/L$ onde x é a coordenada do ponto de parada. Trace uma linha reta paralela ao eixo y pelo ponto de parada. Jogue a bola novamente e defina, para $i = 1, 2...n$ os eventos $s_i = 1$ se o ponto de parada $x_i/L > \theta$ e $s_i = -1$ se não. Note que a probabilidade $P(s_i = 1|I) = \theta$ (equivalente ao parâmetro p anteriormente) e o problema de Bayes é determinar θ. Como não temos informação completa devemos, dar a probabilidade que θ pertença a cada itervalo $(\theta_a, \theta_b]$ condicionado ao dados contidos na sequência $\{s_i\}$. Por construção a distribuição *a priori* de θ é uniforme. Há grandes discussões sobre o uso da uniforme como afirmação de ignorância ou por indicar conhecimento explicito que por construção deve ser assim. Alguém pode ser ignorante enquanto outro não. Suponha que você veja a pessoa que joga a primeira bola e não saiba quem é. Outra pode saber que quem joga a bola é uma grande jogadora de bilhar, cuja filha foi raptada por um mafioso. Seu prior seria uniforme porque voce acreditou toda a história sobre a condução do experimento, mas outros priors podem não sé-lo.

Estimativas de parâmetros para distribuições gaussianas

As medidas x_i, $i = 1, .2....N$ de uma grandeza X tem erros com distribuição gaussiana. Queremos estimar os valores dos parâmetros de localização μ e/ou escala σ. Temos a seguinte informação:

- Conjunto de dados $D_N = \{x_1, x_2, x_3...., x_N\}$.
- Conhecemos a distribuição condicional $P(x|\mu\sigma I)$ dado qu $X \sim \mathcal{N}(\mu, \sigma)$. Todas as medidas são igualmente distribuidas.
- As medidas que levam aos dados são independentes.
- Temos as distribuições *a priori* $P(\mu|I)$ e $P(\sigma|I)$. Consideramos μ e σ *a priori* independentes.

Vamos considerar duas situações quando σ é (i) conhecido e (ii) desconhecido.

(i) Obviamente usamos a regra de Bayes e escrevemos

$$\begin{aligned} P(\mu|D_n\sigma I) &\propto P(\mu|\sigma I)P(D_n|\mu\sigma I) \\ &\propto P(\mu|\sigma I)\prod_i P(x_i|\mu\sigma I) \\ &\propto P(\mu|\sigma I)\exp(\sum_i \log P(x_i|\mu\sigma I)) \end{aligned} \tag{6.40}$$

Note que da estrutura da equação 6.40 vemos que a verossimilhança de μ é gaussiana, e isso continuará valendo se a distribuição *a priori* for constante ou gaussiana. Mas se não for podemos extrair informação útil. É útil trabalhar com o logaritmo dessa expressão. O valor μ_0 de μ que maximiza a probabilidade é determinado por

$$0 = \left(\frac{d \log P(\mu|I)}{d\mu} + \frac{d}{d\mu}(\sum_i \log P(x_i|\mu\sigma I)) \right)_{\mu=\mu_0} \tag{6.41}$$

Usaremos $P(\mu|\sigma I) = P(\mu|I)$ constante na região de interesse. Portanto podemos esquecer por enquanto a distribuição *a priori*, supondo-a constante no intervalo de interesse. Dado que a distribuição dos erros é gaussiana, o μ mais provável será determinado por

$$\begin{aligned} 0 &= \frac{d}{d\mu}(\sum_i (x_i - \mu)^2)|_{\mu_0} \\ 0 &= \sum_i (x_i - \mu_0) \\ \mu_0 &= \frac{1}{n}\sum_i x_i, \end{aligned} \tag{6.42}$$

ou seja o valor de μ mais provável a posteriori é a media dos dados ou a média empírica:

$$\mu_0 = \bar{x}.$$

Mas óbviamente esta dedução usou a informação que, além de independentes, os erros de medida eram gaussianos. Qual é a incerteza que temos sobre o valor de μ? Obviamente quando temos a posterior temos tudo o que podemos ter, mas queremos reduzir ao máximo o que deve ser comunicado. Talvez isso seja um residuo histórico dos tempos antes dos computadores. Um sumário da experiência é dizer a estimativa do parâmetro e sua incerteza. Dado que a distribuição de μ é gaussiana, a incerteza pode ser dada como o desvio padrão. Em princípio não devemos jogar informação fora. Mas dada a convicção de estar falando de uma variável com um valor real, queremos atribuir-lhe um valor. Para referência futura, o método para calcular a incerteza é olhar para a expansão de Taylor até segunda ordem do logaritmo da posterior, que é o inverso da variância da posterior ou uma medida da sua curvatura na região central:

$$\begin{aligned} -\frac{1}{\sigma_N^2} = \frac{1}{2}\frac{d^2}{d\mu^2}(\sum_i \log P(x_i|\mu\sigma I))|_{\mu=\mu_0} &= -\frac{1}{\sigma^2}\frac{d^2}{d\mu^2}\left(\sum_i (x_i - \mu)^2\right) \\ &= -\frac{N}{\sigma^2}, \end{aligned} \tag{6.43}$$

portanto é costume escrever o sumário da experiência como

$$\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}. \tag{6.44}$$

Notamos que a incerteza diminui com a raiz de N, justificando o custo da coleta de mais dados. A posterior se estreita, como vimos nas figuras 6.2-6.4.

Se as medidas tiverem, como no caso da estimativa de g, variâncias

diferentes, então devemos olhar para as derivadas de

$$\begin{aligned} 0 &= \frac{d}{d\mu}(\sum_i \frac{(x_i-\mu)^2}{\sigma_i^2})_{\mu_0} \\ \mu_0 &= \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}, \\ \frac{1}{\sigma_N^2} &= \sum \frac{1}{\sigma_i^2} \end{aligned} \tag{6.45}$$

No caso (ii) estamos interessados em $P(\mu|D_nI)$, que é obtida marginalizando a distribuição conjunta:

$$\begin{aligned} P(\mu|D_nI) &= \int P(\mu\sigma|D_nI)d\sigma \quad \text{e por Bayes:} \\ &\propto \int_0^\infty P(\mu\sigma|I)P(D_n|\mu\sigma I)d\sigma. \end{aligned} \tag{6.46}$$

Novamente supomos independência *a priori* de μ e σ, e podemos supor constância na região de interesse ou que a probabilidade *a priori* é $\propto \sigma^{-\alpha}$, $\alpha = 0$ ou 1. A motivação para esta última forma vem das idéias de Jeffreys e em resumo significa que esperamos que as probabilidades de σ estar entre x e $10x$ são independentes de x. Fazendo a mudança de variáveis $t = 1/\sigma$,

$$\begin{aligned} P(\mu|D_nI) &\propto \int d\sigma \frac{1}{\sigma^{N+\alpha}} e^{-\frac{a}{2\sigma^2}} \\ &\propto \int_0^\infty dt t^{-2} t^{N+\alpha} e^{\frac{-at^2}{2}} \\ &\propto (\frac{1}{a})^{\frac{N+\alpha-2+1}{2}} \int dt' t'^{N+\alpha-2} e^{\frac{-t'^2}{2}} \\ &\propto (\frac{1}{a})^{\frac{N+\alpha-1}{2}} \\ &\propto (\frac{1}{\sum_i (x_i-\mu)^2})^{\frac{N+\alpha-1}{2}}. \end{aligned} \tag{6.47}$$

e usando a notação $\bar{x} = \sum_i x_i/n$ e $\bar{x^2} = \sum_i x_i^2/n$ podemos escrever

$$\sum_i (x_i-\mu)^2) = n(\mu-\bar{x})^2 + n(\bar{x^2} - \bar{x}^2) \tag{6.48}$$

$$P(\mu|D_nI) \propto \frac{1}{\left((\mu-\bar{x})^2 + (\bar{x^2}-\bar{x}^2)\right)^{\frac{N+\alpha-1}{2}}} \tag{6.49}$$

Agora um comentário que atende à dúvida do leitor que se pergunta, e se em lugar de σ, fosse considerada a variância $u = \sigma^2$ como a desconhecida, o resultado seria o mesmo? Usamos um *a priori* de u *análogo* ao caso anterior du/u^α. Deve ser notado que uniforme em σ não é uniforme em σ^2,

$$\begin{aligned} P(\mu|D_nI) &\propto \int du \frac{1}{u^{\frac{N}{2}+\alpha}} e^{-\frac{a}{2u}} \\ &\propto \frac{1}{a} \int_0^\infty \int du' \frac{1}{u'^{\frac{N}{2}+\alpha}} e^{-\frac{1}{2u'}} \\ &\propto (\frac{1}{a})^{\frac{N}{2}+\alpha-1} \\ &\propto (\frac{1}{\sum_i (x_i-\mu)^2})^{\frac{N}{2}+\alpha-1}. \end{aligned} \tag{6.50}$$

Como pode ser visto, as expressões 6.47 e 6.50 são diferentes, a não ser que $\alpha = 1$, que é a prescrição de Jeffreys. Poderiamos ter

escolhido marginalizar sobre a variável $u_k = \sigma^k$ e o prior $1/u_k$ levaria ao mesmo resultado.

A distribuição de μ não é gaussiana, mas podemos calcular um valor e sua incerteza para sumarizar a informação da posterior, como se fosse:

$$\begin{aligned} \frac{d}{d\mu}\log P(\mu|D_n I) &= (N+\alpha-1)\frac{d}{d\mu}\log\sum_i (x_i-\mu)^2 \\ &= (N+\alpha-1)\frac{\sum_i (x_i-\mu)}{\sum_j (x_j-\mu)^2} \\ \ldots|_{\mu=\mu_0}=0 \Rightarrow \mu_0 &= \bar{x} = \frac{1}{N}\sum_i x_i \end{aligned} \tag{6.51}$$

e tomando a segunda derivada, notando que $\sum_j (x_j - \mu_0) = 0$

$$\begin{aligned} \frac{-1}{\sigma_\mu^2} &= \frac{d^2}{d\mu^2}\log P(\mu|D_n I)|_{\mu=\mu_0} \\ &= -\frac{N(N+\alpha-1)}{\sum_i (x_i-\mu_0)^2} = -\frac{N(N+\alpha-1)}{\sum_i (x_i-\bar{x})^2} \\ \sigma_\mu &= \frac{S}{\sqrt{N}} \end{aligned} \tag{6.52}$$

onde (usaremos a moda $\mu_0 = \bar{x}$ é a média empírica)

$$S^2 = \frac{1}{N+\alpha-1}\sum_i (x_i-\bar{x})^2$$

é a estimativa do σ desconhecido a partir dos dados e se $\alpha = 0$ é conhecido como variância amostral. Para N grandes não faz muita diferença tomar $\alpha = 0$ ou 1. A pdf depende da soma $\sum_i (x_i-\mu)^2$, que pode ser escrita de uma maneira mais reveladora

$$\begin{aligned} \sum_i (x_i-\mu)^2 &= \sum_i (x_i^2 - 2x_i\mu + \mu^2 + \bar{x}^2 - \bar{x}^2) \\ &= N(\bar{x}^2 - 2\bar{x}\mu + \mu^2) + \sum_i (x_i^2 - \bar{x}^2) \\ &= N(\bar{x}-\mu)^2 + \sum_i (x_i-\bar{x})^2 \end{aligned} \tag{6.53}$$

$$P(\mu|D_n I) \propto \frac{1}{[N(\bar{x}-\mu)^2 + \sum_i (x_i-\bar{x})^2]^{\frac{N+\alpha-1}{2}}}. \tag{6.54}$$

Note que para $N+\alpha = 3$ é reobtida a distribuição de Cauchy, de caudas gordas. Tem um máximo em $\mu = \bar{x}$ e uma largura a meia altura $l \sim S/\sqrt{N}$ pois $l^2 = 4(2^{\frac{2}{N+\alpha-1}} - 1)S^2(N+\alpha-1)/N$ onde para $\mu = \bar{x} \pm l/2$ a probabilidade cai à metade.

A introdução desta família de distribuições foi feita por W. S. Gosset, que assinava seus artigos como Student, e é conhecida como a distribuição T de Student. Definimos a variável t

$$\begin{aligned} t &= \frac{\mu-\bar{x}}{\sigma_\mu} \\ &= \frac{\mu-\bar{x}}{S/\sqrt{N}} \end{aligned} \tag{6.55}$$

e $\nu = N+\alpha-1$. Temos

$$\begin{aligned} P(\mu|D_n I) &\propto \frac{1}{[(\bar{x}-\mu)^2 + (N+\alpha-1)S^2]^{\frac{N+\alpha-1}{2}}} \\ &\propto \frac{1}{[\frac{N}{\nu}t^2+1]^{\frac{\nu}{2}}} \end{aligned} \tag{6.56}$$

que é a distribuição T com ν graus de liberdade.

Caudas gordas

A conclusões da secção anterior podem não se manter válidas para outras distribuições. Vamos mostrar um exemplo onde isso não ocorre.

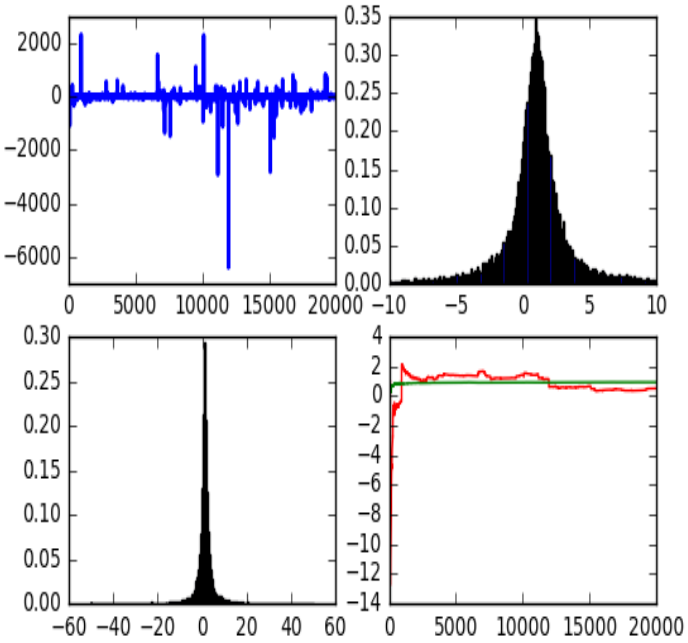


Figura 6.5: Um conjunto de amostras de uma distribuição de Cauchy. Esquerda-Superior: Série temporal, Dir-Sup: Histograma dos valores de $-10 < x < 10$ e Esq-Inf: de $-60 < x < 60$. Dir-Inf: A media amostral de x em vermelho, a média de x, para $-10 < x < 10$ em verde.

O exemplo descrito na figura 6.5 é para uma distribuição de Cauchy. Devido às caudas gordas é possível ter grandes erros se fizermos o que foi sugerido pela análise anterior; a 6.5 demostra que a localização da variável x pode ser estimada se fizermos a média com um corte de e.g. $|x| < 10$. A média amostral de x, dada por $\bar{x} = \sum_{i=1}^{n} x_i/n$ é realmente péssima pois a cada evento extremo a média dá pulos que fazem perder toda a informação conquistada até esse ponto. Os histogramas são bem comportados, mas ao aumentar a região analisada, novos eventos são encontrados, note que na série temporal um valor chega a -10^5. O corte nos valores extremos ao fazer a média é equivalente a declarar um a priori nulo fora desse intervalo.

Na figura 6.6 mostramos a evolução da posterior à medida que mais dados são incorporados. A linha vertical preta na abscissa $r = 1$ mostra o valor correto (usado na simulação). A linha azul mostra a média após n amostras. A linha verde está na moda da posterior. Após 60 amostras parece que a média convergiu para o valor correto, mas para $n = 90$ a média está bem longe e isso se mantém por longo tempo, mesmo em $n = 150$. Após 15 amostras, a posterior tem um um só pico e ele se estabiliza rapidamente perto de $r = 1$. O resultado assintótico para a posterior será independente da simulação, mas a parte inicial onde neste exemplo aparecem até três picos varia de caso a caso, assim com o comportamento da média amostral.

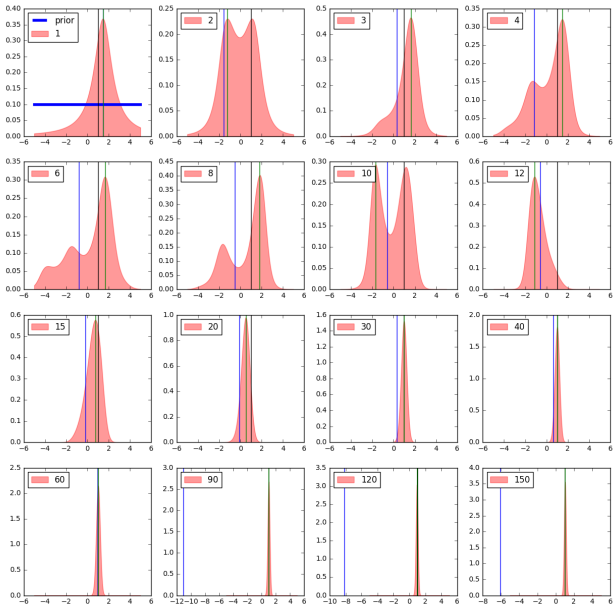


Figura 6.6: A posterior $P(r|D_k I)$ para o mesmo conjunto de amostras da distribuição de Cauchy da figura anterior. Apenas mostramos a posterior para k no conjunto $[1, 2, 3, 4, 6, 8, 10, 12, 15, 20, 30, 40, 60, 90, 120, 150]$

A posterior $P(r|D_k I)$ é bem comportada e não há pulos bruscos. Isso é devido que mesmo que a verossimilhança da última amostra seja grande numa região muito fora da posterior até o último dado ou a nova *a priori*, vai ser multiplicada por valores pequenos de todos os dados anteriores e não tem uma influência igual a todos os dados anteriores.

Vamos supor que a probabilidade de um valor de x seja dado por

$$P(x_i|arI) = \frac{a}{\pi}\frac{1}{1+\frac{(x-r)^2}{a^2}}$$

conforme a linguagem usada antes ésta é a verossimilhança do valor de x na $i-$ésima medida, dados o parâmetro de escala a, suposto conhecido e o parâmetro de localização r suposto desconhecido.

Pela regra de Bayes, após coletar o conjunto de dados $D_k = \{x_i\}_{i=1\ldots k}$, teremos

$$P(r|D_k, a, I) \propto P(r|a, I)P(D_k|r, a, I) \tag{6.57}$$

Supondo, por facilidade, primeiro que os dados são independentes e igualmente distribuidos e que a apriori pode ser tomada como constante, então

$$P(r|D_k, a, I) \propto \prod_i \frac{1}{1+\frac{(x_i-r)^2}{a^2}}. \tag{6.58}$$

A figura 6.6 mostra que a pesar da média ter pulos descontrolados a posterior de r se afina de forma bem comportada. O valor mais provável *a-posteriori* é dado pela solução de

$$r_{post} = \text{maxarg}\sum_i \log\left(\frac{1}{1+\frac{(x_i-r)^2}{a^2}}\right). \tag{6.59}$$

Portanto é dado pela solução da equação

$$\sum_i \frac{(x_i - r)}{1 + \frac{(x_i - r)^2}{a^2}} = 0, \tag{6.60}$$

que pode ser simplemente obtido numericamente mas não temos uma expressão tão fácil quanto no caso gaussiano. Podemos notar que a posterior dada pela equação 6.58 é bem comportada e sua moda muito mais estável do que a média amostral. O resultado a ser lembrado é que sempre podemos usar a média, mas nem sempre é uma boa idéia fazé-lo.

Exemplo simples de tomada de decisão e o problema das 3 portas

Este é um problema bastante conhecido e aparentemente contra intuitivo para muitos que o encontram pela primeira vez. Pode ser facilmente resolvido usando teoria de probabilidade e ajuda a ilustrar um problema de teste de hipóteses, que permite fazer uma decisão sobre que curso de ação deve-se tomar.

O nome do jogo está associado a um programa de TV e a seu anfitrião: Monty Hall. Considere o seguinte jogo assimétrico, entre um jogador J e a banca B. B monta o jogo da seguinte forma que é de conhecimento público. Há tres portas fechadas e atrás de uma delas há um prêmio e nada atrás das outras duas. B pede ao jogador J que aponte uma das portas. Então B, que sabe onde está prêmio, diz: "nao vou tocar a porta que voce escolheu e vou abrir uma das outras duas, que eu sei que estará vazia". Efetivamente, abre uma das outras duas e mostra que está vazia. Não existe a possibilidade de que abra a porta do prêmio. B agora dá uma nova chance a J e pergunta se quer mudar de opinião sobre a porta que escolheu. J tem à sua frente duas possibilidades (i) muda ou (ii) mantém a escolha inicial. O que deve ser decidido é se tanto faz mudar ou não ou seja se há uma melhor estratégia e caso haja, qual é.

A resposta é que sim há uma estratégia melhor e ela é que J deve MUDAR de porta. Muitas pessoas neste ponto reclamam e discordam. Pare para pensar se isto é óbvio ou não.

A solução do problema pode ser encontrada de várias maneiras, mas aqui estamos interessados em aprender sobre teste de hipóteses. A obtenção da resposta decorre de aplicar a seguinte estrategia: Escreva quais são as asserções possíveis e depois escreva as probabilidades e use quando e se possível, a regra de Bayes.

Considere as seguintes asserções, para cada $i = 1, 2$, ou 3:

Nome	Asserção
H_i:	"O prêmio está atrás da porta i"
D_i;	"a banca B abre a porta i"
I_i	"J aponta inicialmente a porta i"

Suponha que I_1 é verdade e deve ser entendido como dado do problema, pois determina simplesmente as condições em que se dará o jogo. Qualquer outra porta que tivesse escolhido seria chamada aqui de porta 1.

Qual é a probabilidade a priori de que o prêmio esteja na porta i? Raciocinado de acordo com um principio de razão insuficiente, J deve atribuir $P(H_i) = 1/3$, já que não há informação para diferenciar uma porta da outra. E após ter feito a escolha inicial da porta? Ainda deve ser $P(H_i|I_1) = 1/3$, pois o simples fato de apontar a porta não deveria mudar a probabilidade de esconder o prêmio.

A pergunta que J está se fazendo poderá ser respondida só após receber a informação de B, por exemplo que a porta 3 é aberta, ou seja D_3 é verdade. Isto é informação relevante: o prêmio não está atrás de 3. Esta informação forma o conjunto de dados que J usará para decidir.

O teste de hipótese deve decidir entre H_1 e H_2, sob a luz da informação recebida, o que significa que deve comparar $P(H_1|D_3I_1)$ e $P(H_2|D_3I_1)$. Definamos r por

$$r := \frac{P(H_1|D_3I_1)}{P(H_2|D_3I_1)}$$

Se $r > 1$, dada a informação disponível H_1 é mais provável e a porta 1 deverá se escolhida. Se $r < 1$, a porta 2, e se $r = 1$, tanto faz.

Para calcular essas probabilidades usaremos a regra de Bayes. A probabilidade conjunta de H_i e D_j dado I_k pode ser escrita de duas formas diferentes:

$$\begin{aligned} P(H_iD_j|I_k) &= P(H_i|I_k)P(D_j|H_iI_k) \\ &= P(D_j|I_k)P(H_i|D_jI_k), \end{aligned} \tag{6.61}$$

isto é, para analisar a probabilidade conjunta de H_ie D_j serem verdade dado I_k, primeiro podemos considerar H_i e depois, sendo H_i verdade, considerar a probabilidade de D_j ser verdade, dado H_i, sempre dado I_k, assim obtemos a primeira equação. Podemos inverter a ordem, começando por analisar D_j, obtendo a segunda equação. Os dois resultados devem ser iguais, pois se não fossem certamente seriamos levados a uma inconsistência, assim segue a regra de Bayes:

$$P(H_i|D_jI_k) = \frac{P(H_i|I_k)P(D_j|H_iI_k)}{P(D_j|I_k)}.$$

Novamente esta regra é a base para qualquer problema onde nova informação D_j nos leve a rever o que pensamos das diferentes hipóteses H_i. Isto é, descreve a forma como devemos mudar a atualização de probabilidades em face à nova informação D_j.

O teste de hipótese requer calcular

$$r = \frac{P(H_1|D_3I_1)}{P(H_2|D_3I_1)} = \frac{P(H_1|I_1)P(D_3|H_1I_1)}{P(H_2|I_1)P(D_3|H_2I_1)}.$$

Agora devemos calcular cada um dos quatro fatores que aparecem no lado direito da equação acima. Repetimos que *apriori* $P(H_1|I_1) = P(H_2|I_1) = P(H_3|I_1) = 1/3$ por simetria.

As verossimilhanças por outro lado nâo sao iguais e este é o ponto surpreendente. Consideremos primeiro $P(D_3|H_1I_1)$. Estas condições descrevem a situação em que B sabe que o prêmio está na porta 1 e que foi indicada por J. Então B poderá escolher entre as portas 2 e 3. Qual é o motivo para que B abra uma ou outra com diferente probabilidade? Não há. Então $P(D_2|H_1I_1) = P(D_3|H_1I_1)$, mas como B deve escolher entre uma e outra e nao há mais possibilidades temos que $P(D_2|H_1I_1) = P(D_3|H_1I_1) = 1/2$. Ou seja, sob as condições H_1I_1

a banca conclui que D_2 e D_3 são exaustivas, mutuamente exclusivas e simétricas.

O quarto fator no teste de hipóteses é $P(D_3|H_2I_1)$. Estamos na condição que J escolheu 1 (I_1), e o prêmio esta na 2 (H_2). Que escolha tem B? Somente uma: ele abrirá a porta 3, logo $P(D_3|H_2I_1) = 1$ e $P(D_2|H_2I_1) = 0$, pois se não fosse poderia ocorrer a revelação do prêmio, assim:

$$\frac{P(H_1|D_3I_1)}{P(H_2|D_3I_1)} = \frac{P(H_1|I_1)P(D_3|H_1I_1)}{P(H_2|I_1)P(D_3|H_2I_1)} = \frac{\frac{1}{3}\times\frac{1}{2}}{\frac{1}{3}\times 1} = \frac{1}{2},$$

logo concluimos que a probabilidade de estar atrás da porta 2 é o dobro que estar atrás da porta 1 e portanto a estrategia é que deverá mudar da porta 1 para a 2.

Considere a generalização do problema acima para N portas, k prêmios e a abertura de a portas. Considere um prêmio, um milhão de portas e a abertura de 999.998 portas. So ficam duas portas fechadas, uma indicada por J inicialmente e outra escolhida por B cuidadosamente. Chamemos $D_{/856.322}$ a asserção "*Abriu* 584.416 *portas e pulou a* 584.417 *que J tinha escolhido, depois abriu até a* 856.321 *e pulou a porta* 856.322 *e continuou abrindo até chegar à última*". Só temos duas hipóteses a confrontar: H_J que diz que o prêmio está atrás da porta escolhida pelo jogador, 584.417 e H_B que diz que o prêmio está atrás da porta 856.322, pulada pela banca. Fica mais intuitivo ? O teste de hipóteses dá

$$r = \frac{P(H_J|D_{/856.322}I_J)}{P(H_B|D_{/856.322}I_J)} = \frac{P(H_J|I_J)P(D_{/856.322}|H_JI_J)}{P(H_B|I_J)P(D_{/856.322}|H_BI_J)} =$$

$$= \frac{\frac{1}{N}\times\frac{1}{N-1}}{\frac{1}{N}\times 1} = \frac{1}{N-1} = \frac{1}{10^6-1} \simeq 10^{-6}.$$

É claro agora que o fato de J escolher aleatoriamente, sem informação nenhuma que quebre a simetria inicial entre as portas, leva a a que o prêmio tenha uma probabilidade muito baixa de estar na que escolheu. A banca sabe onde está o prêmio. Cuidadosamente evita uma porta em particular,abrindo todas as outras. O conjunto de portas não apontado inicialmente por J tem uma probabilidade $P(\overline{H}_J) = 1 - \frac{1}{N}$ de conter o prêmio. A informação que B fornece ao abrir as portas evita dizer qualquer informação sobre o domínio apontado por J, sem revelar exatamente a localização do prêmio. A cada porta aberta, a probabilidade que J tenha acertado não muda, nem a probabilidade $P(\overline{H}_J)$ mas esta se distribui igualmente por um número cada vez menor de portas. Quando só sobra uma porta fechada além da escolhida inicialmente, toda essa probabilidade se concentra.

Exercícios Propostos

- Laplace estudou o seguinte problema:

Considere três urnas idénticas. Com base unicamente na informação a seguir

1. Cada urma tem duas bolas. Uma tem duas bolas brancas (U_0), outra tem uma branca e uma preta (U_1) e a terceira duas pretas (U_2).

2. Independentemente de qualquer outra coisa, uma urna é escolhida sem que haja preferência por alguma delas.
3. Da urna escolhida uma bola é extraida. A bola é branca e colocada novamente na urna
4. Da mesma urna, uma bola é escolhida. Novamente a bola é branca.

calcule a probabilidade que a urna escolhida seja U_1 e a probabilidade que seja U_2. Valendo zero pontos, a probabilidade de ser U_3.

Tente estruturar a sua solução de forma a que seja útil para atacar outros problemas. Defina as asserções importantes no problema, por exemplo W_1= "a primeira bola extraida é branca". Defina qual é a asserção cuja probabilidade é pedida, incluindo (!!!) as condições. Use argumentos de simetria onde for necessário e não esqueça de identificar onde estes argumentos foram usados. Use as regras do produto e soma, use marginalização e independência.

1.b)Caso geral: Temos $M+1$ urnas indexadas por K que toma valores inteiros k de zero a M. Na k-ésima urna k bolas são pretas e as outras $M-k$ são brancas. Uma urna é escolhida sem que haja preferência por alguma em especial. Uma bola é retirada, sua cor anotada, sendo recolocada na urna. Isto é repetido N vezes com a mesma urna. O número de bolas pretas extraidas no total é J. Qual é a probabilidade da $k-$ésima urna ter sido escolhida dados M, N e J : $P(k|M, N, J)$? Dica: Suponha k conhecido e calcule a probabilidade de J. Faça a inversão.

7
Teorema do Limite Central

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 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 $A_1 A_2$ 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 alguns poucos gramas é da ordem de 10^{23}.

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 [1]. Estamos 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 7.1. 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

$$P(y|N = 2)dy = \int_{y \leq x_1 + x_2 \leq y + dy} dx_1 dx_2 P(x_1) P(x_2), \qquad (7.1)$$

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

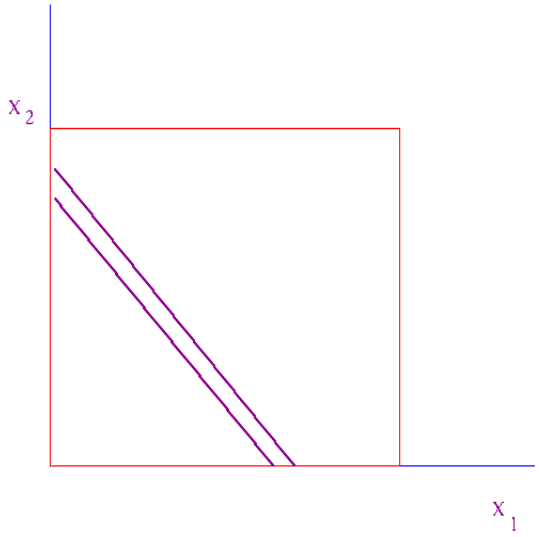


Figura 7.1: No plano X_1X_2 temos a região onde o valor de Y está entre y e $y+dy$. Todos os pares x_1 e x_2 nela contribuem para a probabilidade de Y

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 [2].

$$P(y|N=2)dy = \int \chi_{y\leq x_1+x_2\leq y+dy} dx_1 dx_2 P(x_1)P(x_2), \qquad (7.2)$$

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 $\delta_n(A)$:

$$\begin{aligned} \delta_n(y \leq x_1 + x_2 \leq y + \Delta y_n) &= \frac{1}{\Delta y_n}, \quad \text{se } y \leq x_1 + x_2 \leq y + \Delta y_n \\ &= 0, \quad \text{se não} \end{aligned} \qquad (7.3)$$

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

$$P(y|N=2) = \int dx_1 dx_2 P(x_1)P(x_2)\delta(y - x_1 + x_2), \qquad (7.4)$$

$$P(y|N=2) = \int dx P(x)P(y-x), \qquad (7.5)$$

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

[2] χ_A é chamada a função indicadora do intervalo ou conjunto A. Alguns autores chamam de função característica do intervalo mas não confunda com a função característica da distribuição de probabilidades definida abaixo.

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:

$$P(y) = \int dx_1 dx_2 P(y, x_1, x_2). \qquad (7.6)$$

Da regra do produto

$$P(y) = \int dx_1 dx_2 P(y|x_1, x_2)P(x_1, x_2). \qquad (7.7)$$

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 7.5. A vantagem disto é que podemos facilmente obter expressões para funções gerais. Se $y = f(x_1, x_2)$, então

$$P(y) = \int dx_1 dx_2 \delta(y - f(x_1, x_2)) P(x_1, x_2) \tag{7.8}$$

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
- $\langle x \rangle = \int_{-\infty}^{\infty} xP(x)dx < \infty$,
- $\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 P(x)dx < \infty$,

podemos introduzir a transformada de Fourier (TF) [3] e a inversa

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

$$\Phi(k) = \int_{-\infty}^{\infty} e^{-ikx} P(x)dx \tag{7.9}$$

$$P(x) = \int_{-\infty}^{\infty} e^{ikx} \Phi(k) \frac{dk}{2\pi} \tag{7.10}$$

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 7.9 e a troca da ordem de integração e somatória nos mostra que os coeficientes estão relacionados aos momentos:

$$\begin{aligned} \Phi(k) &= \int_{-\infty}^{\infty} \sum_{s=0}^{\infty} \frac{(-ikx)^s}{s!} P(x) \\ &= \sum_{s=0}^{\infty} \frac{(-ik)^s}{s!} \langle x^s \rangle \\ &= \sum_{s=0}^{\infty} \frac{(-ik)^s}{s!} M_s, \end{aligned} \tag{7.11}$$

em termos dos momentos $M_s = \langle x^s \rangle$. Tomemos a TF dos termos da equação 7.5, e usando :

$$\delta(k) = \int \frac{dx}{2\pi} e^{ikx} \tag{7.12}$$

obtemos

$$\begin{aligned} \Phi(k|N=2) &= \int dy dx e^{-iky} P(x) P(y-x), \\ &= \int \frac{dx dy dk_1 dk_2}{(2\pi)^2} \Phi(k_1|1)\Phi(k_2|1) e^{-iky + ik_1 x + ik_2 (y-x)}. \end{aligned}$$

Integrando sobre x e usando a representação da delta:

$$\begin{aligned} &= \int \frac{dy dk_1 dk_2}{2\pi} \Phi(k_1|1)\Phi(k_2|1) e^{-iky + ik_2 y} \delta(k_1 - k_2), \\ &= \Phi(k|1)\Phi(k|1) = \Phi^2(k|N=1) \end{aligned} \tag{7.13}$$

Para a soma de $N = n$ variáveis x_i

$$P(y|N=n) = \int \prod_{i=1...n} dx_i P(x_1)P(x_2)...P(y - \sum_{i=1}^{n-1} x_i), \qquad (7.14)$$

ou, introduzindo uma integral mais

$$P(y|N=n) = \int \prod_{i=1...n} dx P(x_1)P(x_2)...P(x_n)\delta(y - \sum_{i=1}^{n} x_i), \qquad (7.15)$$

obtemos

$$\Phi(k|N=n) = \Phi^n(k|N=1) \qquad (7.16)$$

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.

Podemos tomar o logarítmo de cada lado da equação 7.16 e dado que produtos, ao tomar logaritmos, viram somas, temos

$$\begin{aligned} \kappa_Y(k|N=n) &= \log\Phi(k|N=n) = \sum_i^n \log\Phi(k|N=1) \\ &= n\log\Phi(k|N=1) = n\kappa_X(k|N=1). \end{aligned} \qquad (7.17)$$

onde a segunda linha vale no caso que as variáveis x_i sejam igualmente distribuidas. Isto nos leva a discutir os cumulantes $\{C_s(n)\}$ de uma distribuição[4], definidos através da expansão em série de potências de ik da sua função característica:

$$\kappa(k|N=n) = \log\Phi(k|N=n) = \sum_{s=0}^{\infty} C_s(n)\frac{(-ik)^s}{s!}. \qquad (7.18)$$

[4] As funções κ são chamadas as funções geradoras dos cumulantes. A idéia de cumulantes teve várias origens independentes. Os nomes associados são T. N. Theile, H. Ursell, R. Fisher, J. Wishart.

$$\sum_{s=0}^{\infty} M_s(n)\frac{(-ik)^s}{s!} = e^{\sum_{s=0}^{\infty} C_s(n)\frac{(-ik)^s}{s!}}. \qquad (7.19)$$

A equação 7.17 nos indica o motivo do nome dos cumulantes: a aditividade (ou acúmulo) ante convoluções

$$\begin{aligned} C_s(Y = \sum_i X_i) &= \sum_i C_s(X_i), \\ C_s(n) &= nC_s(1), \end{aligned} \qquad (7.20)$$

onde a equação 7.20 segue porque as $\{x_i\}$ são identicamente distribuidas e usamos uma notação menos carregada. Concluimos que quando variáveis aleatórias independentes se somam, os cumulantes da distribuição da soma são a soma dos cumulantes das distribuições.

Podemos obter a relação entre os momentos e os cumulantes. Pela definição através da série de potências, vemos que em termos da função característica

$$C_s = \frac{1}{(-i)^s}\frac{d^s \log\Phi}{dk^s}|_{k=0} \qquad (7.21)$$

Podemos calcular alguns dos primeiros,

$$\begin{aligned}
\log \quad & \Phi(k|N=1) = \log \int e^{-ikx} P(x) dx \\
= \quad & \log \int \sum_{s=0}^{\infty} \frac{(-ikx)^s}{s!} P(x) dx \\
= \quad & \log(1 + \sum_{s=1}^{\infty} \frac{(-ik)^s}{s!} \langle x^s \rangle) \\
= \quad & \sum_{s_1=1}^{\infty} \frac{(-ik)^{s_1}}{s_1!} \langle x^{s_1} \rangle - \frac{1}{2} \sum_{s_1,s_2=1}^{\infty} \frac{(-ik)^{s_1+s_2}}{s_1! s_2!} \langle x^{s_1} \rangle \langle x^{s_2} \rangle \\
+ \quad & \frac{1}{3} \sum_{s_1,s_2,s_3=1}^{\infty} \frac{(-ik)^{s_1+s_2+s_3}}{s_1! s_2! s_3!} \langle x^{s_1} \rangle \langle x^{s_2} \rangle \langle x^{s_3} \rangle + ... \qquad (7.22)
\end{aligned}$$

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 $\langle x^s \rangle$:

$$\begin{aligned}
C_0 &= 0, \\
C_1 &= \langle x \rangle, \\
C_2 &= \langle x^2 \rangle - \langle x \rangle^2, \\
C_3 &= \langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3, \\
C_4 &= \langle x^4 \rangle - 4\langle x^3 \rangle \langle x \rangle - 3\langle x^2 \rangle^2 + 12\langle x^2 \rangle \langle x \rangle^2 \\
&- 6\langle x \rangle^4, \qquad (7.23)
\end{aligned}$$

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 homogeneidade:

$$nC_1^x = C_1^Y(n) = \sqrt{n} C_1^Z(n) = nC_1^W(n), \qquad (7.24)$$

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

$$nC_s^x = C_s^Y(n) = n^{s/2} C_s^Z(n) = n^s C_s^W(n), \qquad (7.25)$$

Portanto

$$\begin{aligned}
C_s^Y(n) &= nC_s^x \\
C_s^Z(n) &= \frac{1}{n^{\frac{s}{2}-1}} C_s^x \\
C_s^W(n) &= \frac{1}{n^{s-1}} C_s^x, \qquad (7.26)
\end{aligned}$$

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 W 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

$$u_s^Y(n) = \frac{C_s^Y(n)}{(C_2^Y(n))^{\frac{s}{2}}} = n^{1-\frac{s}{2}} u_s^x, \qquad (7.27)$$

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):

$$\begin{aligned} u_s^Z(n) &= \frac{C_s^Z(n)}{(C_2^Z(n))^{\frac{s}{2}}} = n^{1-\frac{s}{2}} u_s^x, \\ u_s^W(n) &= \frac{C_s^W(n)}{(C_2^W(n))^{\frac{s}{2}}} = n^{1-\frac{s}{2}} u_s^x, \end{aligned} \tag{7.28}$$

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.

O Teorema do Limite Central I

Começamos pela função característica de Z

$$\Phi_Z(k|n) = \exp\left(\sum_{s=1}^{\infty} C_s^Z(n) \frac{(-ik)^s}{s!}\right)$$

e a função geradora dos cumulantes

$$\begin{aligned} \kappa_Z(k|n) &= \sum_{s=1}^{\infty} C_s^Z(n) \frac{(-ik)^s}{s!} \\ &= \sum_{s=1}^{\infty} \frac{(-ik)^s}{s!} n^{1-\frac{s}{2}} C_s^Z(n) \\ &= \sum_{s=1}^{\infty} \frac{(-ik)^s}{s!} n^{1-\frac{s}{2}} C_s^X \\ &= n\kappa_X\left(\frac{k}{\sqrt{n}}\right). \end{aligned} \tag{7.29}$$

Note que a função κ é nula para argumentos nulos (segue da normalização da distribuição de probabilidades), portanto ao tomar o limite de n grande não está claro o que acontece com a expresão acima e portanto devemos investigar mais. Chamando $\mu_x = \langle X \rangle$ e $C_2^X = \sigma_x^2 = \langle X^2 \rangle - \langle X \rangle^2$

$$\begin{aligned} \kappa_Z(k|n) &= -ikC_1^Z(n) - k^2 C_2^Z(n) + \sum_{s=3}^{\infty} \frac{(-ik)^s}{s!} n^{1-\frac{s}{2}} (C_2^Z(n))^{\frac{s}{2}} u_2^X \\ &= -ik\sqrt{n}\mu_x - \frac{k^2}{2}\sigma_x^2 + \sum_{s=3}^{\infty} \frac{(-ik)^s}{s!} n^{1-\frac{s}{2}} C_s^X \end{aligned} \tag{7.30}$$

onde usamos os resultados 7.26.

$$\begin{aligned} |\kappa_Z(k|n) + ik\sqrt{n}\mu_x + \frac{k^2}{2}\sigma_x^2| &= |\sum_{s=3}^{\infty} \frac{(-ik)^s}{s!} n^{1-\frac{s}{2}} C_s^X| \\ |\kappa_Z(k|n) + ik\sqrt{n}\mu_x + \frac{k^2}{2}\sigma_x^2| &= \frac{1}{\sqrt{n}} |\sum_{s=3}^{\infty} \frac{(-ik)^s}{s!} n^{\frac{3-s}{2}} C_s^X| \end{aligned} \tag{7.31}$$

Os estudantes devem lembrar o critério de convergência de Dirichlet [5]. Note que se para cada k fixo a função característica for limitada, as condições de convergência são satisfeitas. Então para cada valor de k fixo, o termo do lado direito cai pelo menos como $n^{-1/2}$ portanto

$$\lim_{n\to\infty} |\kappa_Z(k|n) + ik\sqrt{n}\mu_x + \frac{k^2}{2}\sigma_x^2| = 0 \tag{7.32}$$

Isto forma a base para um teorema sobre a convergência da função geradora dos cumulantes de z na função geradora de uma disribuição normal. Também permite, agora pulando algumas etapas considerar razoável desprezar os cumulantes de ordem superior à segunda e obter pela

$$\begin{aligned} P(Z|N=n) &= \int_{-\infty}^{\infty} \exp(ikz + ikC_1^Z(n) - k^2C_2^Z(n)/2)\frac{dk}{\sqrt{2\pi}} \\ &= \int_{-\infty}^{\infty} \exp(ikz + ik\sqrt{n}\mu_x - k^2\sigma_x^2/2)\frac{dk}{\sqrt{2\pi}} \end{aligned} \tag{7.33}$$

para obter

$$P(Z|N=n) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(z-\sqrt{n}\mu)^2}{2\sigma_x^2}} \tag{7.34}$$

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

$$\begin{aligned} P(Y|N=n) &= \frac{1}{\sqrt{2\pi n\sigma_x^2}} e^{-\frac{(y-n\mu_x)^2}{2n\sigma_x^2}} \\ P(W|N=n) &= \frac{1}{\sqrt{2\pi\frac{\sigma_x^2}{n}}} e^{-\frac{(w-\mu_x)^2}{2\frac{\sigma_x^2}{n}}} \end{aligned} \tag{7.35}$$

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.

Isto é um esboço de uma prova. Vejamos agora um exemplo onde o cálculo é exato.

Exercício

Mostre que os resultados acima (eqs. 7.34 e 7.35) 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. No capítulo sobre a gaussiana encontramos explicitamente que a soma de duas variáveis normais é também normal. Este é um exemplo de uma distribuição dita **estável** sob adições. Somas de variaveis gaussianas são gaussianas.

Um teorema

Podemos fazer o argumento acima um pouco mais cuidadoso e obter algo mais parecido a um resultado rigoroso. O quê temos e o que falta?

[5] Considere uma sequência de números reais $\{a_s\}$, no caso que nos interessa aqui $a_s = n^{\frac{3-s}{2}}$ e uma sequência de números $\{b_s\}$, aqui tomaremos $b_s = \frac{(-ik)^s}{s!}C_s^X$, que satisfazem (i) $a_{s+1}/a_s \leq 1$,(ii) $\lim_{s\to\infty} a_s = 0$, (iii) para cada inteiro N vale $|\sum_{s=1}^{N} b_s| \leq M$, onde M é alguma constante. Então $\sum_s^{\infty} a_s b_s$ converge.

- Os cumulantes de ordem $s \geq 3$ de uma distribuição normal são nulos.
- Em algum sentido os cumulantes de ordem $s \geq 3$ para as somas tendem a diminuir com o aumento de n (eqs 7.28 e 7.27).

O que temos é que as funções características de Y, Z e W convergem, para cada k na função característica de uma gaussiana. Resta usar um teorema devido a Levy que diz que se há convergência pontual de uma sequência de funções características para uma função característica,

$$\Phi(k|n) \to \Phi(k)$$

então a sequência de distribuições cumulativas converge para a distribuição cumulativa

$$Prob(Y \in [a,b]|n) \to Prob(Y \in [a,b])$$

isso é chamado convergência em distribuição: $Y_n \to Y$. No caso em particular que estamos interessados a variável $Y_n = \sum_i^n X_i$ tende à Y que tem distribuição normal.

Exercício

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

A média e a concentração em torno da média

A distribuição de W dada pela eq. 7.34 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 7.26, temos que $C_2^W(n) = \sigma_x^2/n$, satisfaz

$$\begin{aligned} C_2^W(n) &= \int_{-\infty}^{\infty} dw(w^2 - \langle w\rangle^2)P(W=w|N=n) \\ &= \int_{-\infty}^{\infty} dw(w - \langle w\rangle)^2 P(W=w|N=n) \\ &\geq \int_{|w-\langle w\rangle|\geq\epsilon} dw(w - \langle w\rangle)^2 P(W=w|N=n) \\ &\geq \int_{|w-\langle w\rangle|\geq\epsilon} dw\epsilon^2 P(W=w|N=n) \\ &\geq \epsilon^2 Prob(|w - \langle w\rangle| \geq \epsilon). \end{aligned} \tag{7.36}$$

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

$$Prob(|w - \langle w\rangle| \geq \epsilon) \leq \frac{C_2^W(n)}{\epsilon^2} \tag{7.37}$$

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 $\langle w \rangle$ e que este valor se afaste do valor médio por mais que ϵ", isto é, , $Prob(|w - \langle w \rangle| \geq \epsilon)$ está limitada por:

$$Prob(|w - \langle w \rangle| \geq \epsilon) \leq \frac{C_2^x}{n\epsilon^2} \tag{7.38}$$

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'[6] onde p, q tem outro valores (entre zero e um). Consideremos as jogadas independentes, para duas jogadas i e j quaisquer $P(s_i|s_j I') = P(s_i|I')$. Chame m o número de caras para cima, quando a moeda é jogada n vezes. A frequência de caras é definida por $f = m/n$

[6] 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.

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

$$P(m|N = nI') = \frac{n!}{m!(n-m)!} p^m q^{n-m} \tag{7.39}$$

- **(B)** Calcule $\langle m \rangle$, $\langle m^2 \rangle$. [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: $\langle m \rangle = np$, $\langle m^2 \rangle = n^2p^2 + np(1-p)$]
- **(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* $\langle f \rangle = p$ 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.

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 $Z - \langle Z \rangle = \frac{Y - N\mu}{\sqrt{N}\sigma}$ 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
- $\int |dP(x)/dx|\, dx < A$
- $a < \langle x^2 \rangle = \sigma^2 < b$
- $\langle |x^3| \rangle < b$
- $\langle x^4 \rangle < b$
- $\langle |x^5| \rangle < b$
- $|\Phi(k)| > d$ 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

$$|\Phi(k)| < \rho.$$

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 **??** que foram *desprezados* anteriormente para chegar até a equação 7.33. Pense na diferença entre *desprezar* e *controlar*. Muitas vezes, em Física *desprezamos* sem *controlar*, pois tentar *controlar* pode ser tão díficil 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.

O Teorema do Limite Central III

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

A distribuição uniforme

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

$$\Phi(k|1) = \frac{1}{L}\int_{-L/2}^{L/2} e^{-ikx}dx = \frac{2}{kL}\sin(\frac{kL}{2}) \tag{7.40}$$

$$P(Y = y|N = n) = \int_{-\infty}^{\infty}[\Phi(k|1)]^n e^{iky}\frac{dk}{2\pi} = \int_{-\infty}^{\infty}[\frac{2}{kL}\sin(\frac{kL}{2})]^n e^{iky}\frac{dk}{2\pi} \tag{7.41}$$

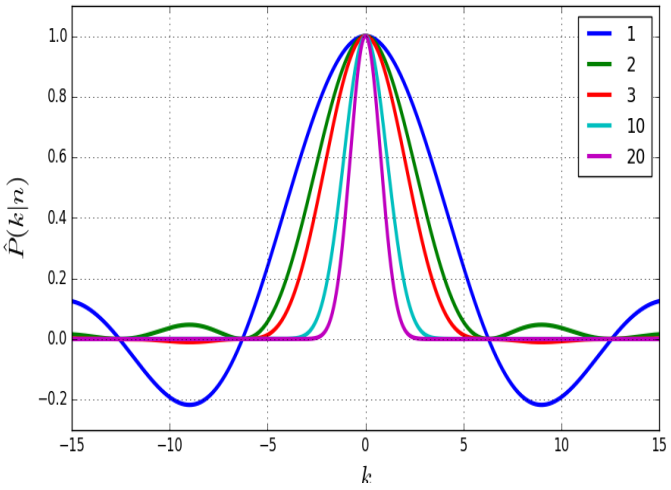


Figura 7.2: A função característica $\Phi(k|N = n)$ para a soma de n varíaveis uniformemente distribuidas.

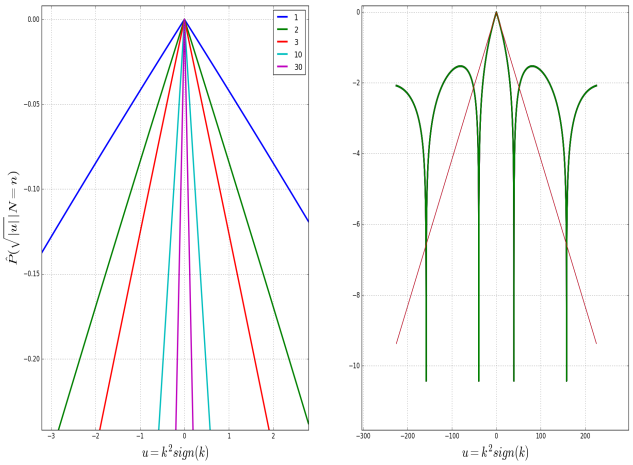


Figura 7.3: A função característica $\log(|\Phi(\sqrt{|u|}|N = n)|)$ como função de $u = k^2\text{sign}(k)$, para a soma de n varíaveis uniformemente distribuidas. Nesta representação gaussianas aparecem como retas $\propto -\text{abs}(u)$. Esquerda: na região central parecem gaussianas. Direita: fora da região central diferem de gaussianas. Nesta figura dividimos por $\sqrt{n}$ e todas as curvas colapsam. As retas são para a gaussiana mais proxima $\sigma^2/2n = 1/24$. Note a diferença da escala de u nas abscissas.

A figura 7.2 mostra que a função característica fica mais parecida com uma gaussiana e na figura 7.3 vemos que efetivamente o

$\log(|\Phi(\sqrt{|u|}|N=n)|)$ com $u=\pm k^2$ fica cada vez mais perto de $-\sigma^2|u|$ (gaussiana).

Exercício Mostre que $\sigma^2/2 = n/24$.(verificar??)

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

$$\Phi(k|1) = \int_0^\infty ae^{-xa}e^{-ikx}dx = \frac{a}{a+ik} \tag{7.42}$$

$$P(Y=y|N=n) = \int_{-\infty}^{\infty} (\frac{a}{a+ik})^n e^{iky}\frac{dk}{2\pi} \tag{7.43}$$

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

$$P(Y=y|N=n) = \Theta(y)a^n\frac{y^{n-1}e^{-ay}}{(n-1)!}. \tag{7.44}$$

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

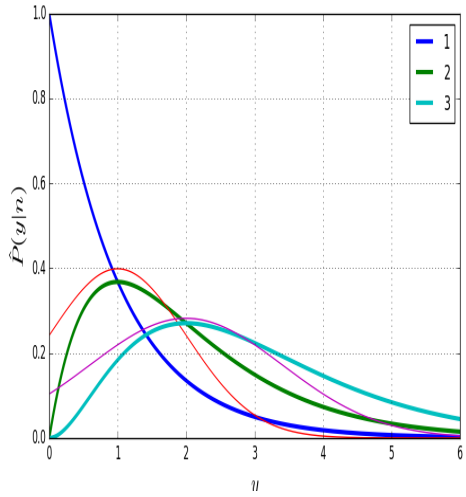


Figura 7.4: A densidade de probabilidade $P(y|N=n, a=1)$ para a soma de n varíaveis exponencialmente distribuidas, $n = 1, 2, 3$. Para as duas últimas mostramos as gaussianas com $\mu = \sigma^2 = n-1$

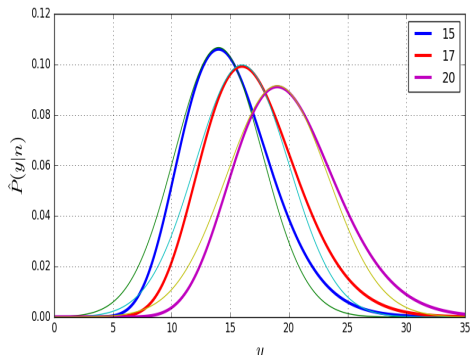


Figura 7.5: A densidade de probabilidade $P(y|N=n, a=1)$ para a soma de n varíaveis exponencialmente distribuidas, $n = 15, 17, 20$. Junto estão mostradas as gaussianas com $\mu = \sigma^2 = n-1$.

A figura 7.4 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 7.5. 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.

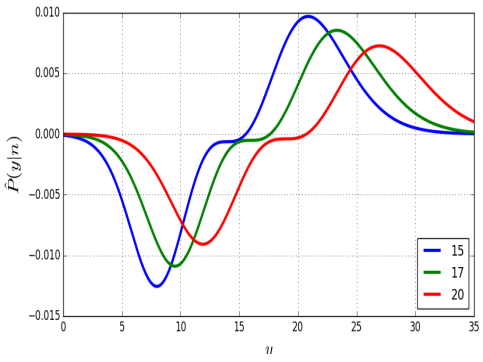


Figura 7.6: A diferença entre a densidade de probabilidade $P(y|N = n, a = 1)$ para a soma de n varíaveis exponencialmente distribuidas, $n = 15, 17, 20$ e as gaussianas com $\mu = \sigma^2 = n - 1$. Os mesmos parâmetros da figura anterior. Note que a região central é bem aproximada. Há uma região de transição, ao afastar-se para as caudas, e finalmente as caudas vão rapidamente para zero, assim como a sua diferença

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

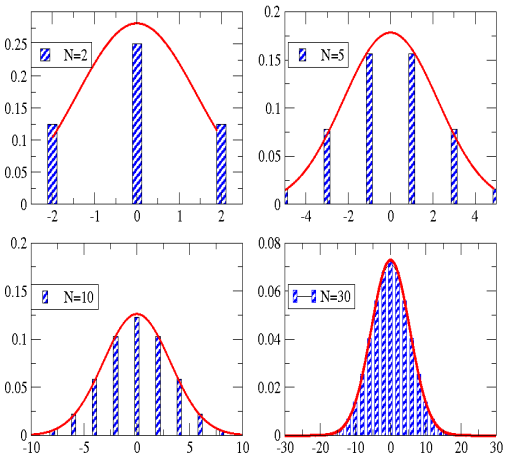


Figura 7.7: A binomial (dividida por $\Delta Y = 2$, barras) e a densidade gaussiana correspondente (linha contínua), para $N = 2, 5, 10$ e 30

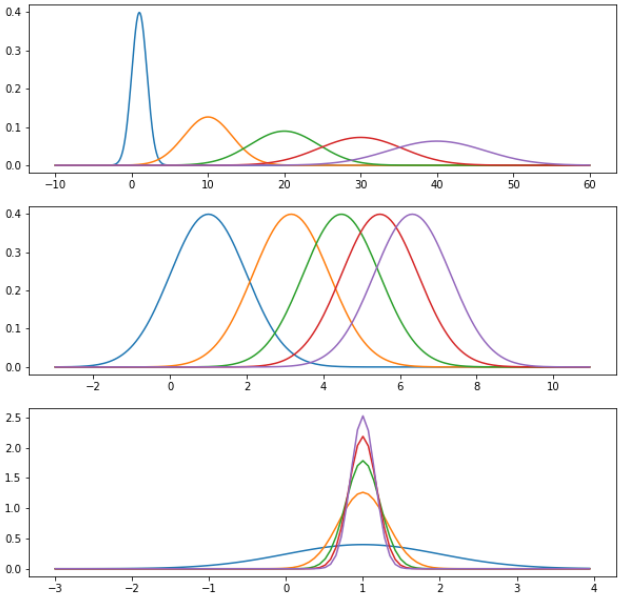
distribuição de probabilidades binomial deve ser comparada com a densidade de probabilidade gaussiana. Note que se N é par a probabilidade de que ocorra um valor de Y impar é zero, ou seja $\Delta Y = 2$.. Ao apresentar os gráficos da figura 7.7 a binomial foi dividida por ΔY. De outra forma: a probabilidade da binomial que Y tenha um dado valor num intervalo $(y, y+2)$ é aproximado pela integral da gaussiana entre y e $y+2$.

Caminho Aleatório

Novamente olhamos para a distribuição binomial. Olhe para a figura 7.8. Definimos o caminho aleatório através de

Difusão: K(= 10000 na figura 7.8) seqüências de N passos de um processo binomial, definidos por

Figura 7.8:



$$y_n = y_{n-1} + x_n \tag{7.45}$$

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,

$$\log N! = (N + \frac{1}{2}) \log N - N + \frac{1}{2} \log 2\pi + \log \left(1 + \frac{1}{12N} + \frac{1}{288N^2} + \mathcal{O}(\frac{1}{N^3})\right) \tag{7.46}$$

que é uma expansão assintótica, isto é melhora quando N aumenta [7].

[7] 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.

Não precisamos todos esses termos, basta $\log N \approx N \log N - N$ onde desprezamos $\mathcal{O}(\log N)$

$$\begin{aligned} \log P(m|N = nI') &= \log n! - \log m! - \log(n-m)! + m \log p + (n-m) \log q \\ &= n \log n - n - m \log m + m - (n-m) \log(n-m) \\ &\quad + n - m + m \log p + (n-m) \log q \end{aligned} \tag{7.47}$$

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 7.27, 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

$$P(m|N = nI) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(m - m^*)^2 \tag{7.48}$$

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

8

Seleção de Modelos

> "An ingenious Friend has communicated to me a Solution of the inverse Problem, in which he has shewn what the Expectation is, when an Event has happened p times, and failed q times, that the original Ratio of the Causes for the Happening or Failing of an Event should deviate in any given Degree from that of p to q. And it appears from this Solution, that where the Number of Trials is very great, the Deviation must be inconsiderable : Which shews that we may hope to determine the Proportions, and, by degrees, the whole Nature, of unknown Causes, by a sufficient Observation of the Effects."
> Observation of Man, His Frame, His Duty, and His Expectations
> David Hartley
> M.DCC.XLIX [1]

Um dos nossos objetivos finais é o de distinguir os méritos de modelos sobre a natureza. Não queremos atribuir certeza a um modelo e dizer que 'isto é a verdade sobre o universo.' Os modelos não devem ser julgados corretos. Eles são úteis e podem deixar de sé-los quando novas evidências permitirem a construção de novos modelos. Claro que o estudante *sabe* que a lei de Coulomb revela a verdade sobre como duas cargas interagem. Mas essa certeza será mudada quando aprender relatividade e teoria quântica de campos. A introdução da lei de gravitação universal de Newton e da sua dinâmica talvez tenha sido o evento de maior influência intelectual na sociedade ocidental. Pode ser que alguns achem que há outros candidatos, como Darwin. Para os gregos antigos as coisas caiam *porque* almejavam estar no seu lugar natural, ou seja o centro do universo que coincidia com o centro da Terra. Essa explicação deu lugar a outras e em algum ponto alguém, cujo nome talvez não deva ser mencionado, explicou a queda e o movimento em geral através de forças exercidas por anjinhos. Mas a partir de Galileu e certamente com Newton o enfoque mudou ao discutir *como* as coisas caem com a introdução do modelo de ação à distância e forças vetoriais de intensidades que decaem com o quadrado da distância. Mas com o trabalho de Einstein esse modelo ficou, se não obsoleto, pelo menos reduzido a uma boa aproximação da nova verdade, pois agora *sabemos* que a gravidade está relacionada à curvatura do espaço-tempo. Mas alguém quer apostar que essa *verdade* será a **verdade** no fim das suas carreiras científicas, quanto mais daqui a mil anos? Talvez nessa época não sejam as nossas respostas que não façam sentido, mas possivelmente as perguntas já tenham deixado de ser relevantes. Não precisamos esperar tanto. Nosso respeito pela

[1] Citado por S. Stiegler, *"Who discovered Bayes's Theorem ?"The American Statistician*, Vol 37, No. 4 (1983) 290-6.

Isto mostra que, além de um gosto exagerado pelas maiúsculas, o psicólogo Harley tinha escutado de um amigo uma primeira versão do que viria a ser conhecido através do trabalho de Bayes 15 anos depois, quando R. Price publicou os escritos de Bayes postumamente. Stiegler especula que o amigo ingenioso não tenha sido Bayes, mas talvez Nicolas Saunderson. Isto é de grande importância para os historiadores. Na minha opinião no entanto o culpado pelo teorema deve ser Laplace. Esta referência foi mencionada por S. Wechsler.

construção científica de Newton, Einstein e Coulomb não diminuirá mesmo percebendo que suas contribuições são temporariamente úteis e que futuras gerações terão outros modelos para temporariamente fazer as suas predições, que englobarão para certo regime de energia as dos modelos anteriores. Dito isso passamos ao problema de como escolher dentre os diferentes modelos aquele que deve ser preferido sobre os outros. Há muita coisa escrita sobre isto e muitos filósofos dedicaram suas vidas a este problema que está no centro do avanço científico e da epistemologia. A extensão da lógica a casos de informação incompleta nos permite atacar este problema como mais uma aplicação da regra de Bayes. [2]

[2] Dito de outra forma: "It is our responsibility as scientists, knowing the great progress which comes from a satisfactory philosophy of ignorance, the great progress which is the fruit of freedom of thought, to proclaim the value of this freedom; to teach how doubt is not to be feared but welcomed and discussed; and to demand this freedom as our duty to all coming generations."
R.P.Feynman
The Value of Science (1955)

Modelos

A construção de modelos em Física passa pela construção de teorias. Para a construção de teorias não há regras. Veja um belo exemplo da descrição de um processo de criação de Eletrodinâmica Quântica na Palestra de Prêmio Nobel de Feynman. A palestra poderia ser chamada de *The making of QED*. Uma teoria leva, ou deve levar, finalmente a que

$$y = f(x, \theta),$$

ou seja uma relação funcional entre duas variáveis x e y mediada por uma função f onde aparece ainda um conjunto de parâmetros, denotados coletivamente por θ. Chamaremos x variáveis de contrôle e a y de livres. Considere as seguintes perguntas, que podem ser de interesse em diferentes circunstâncias:

- (1) Previsão de y

Para f, θ e x conhecidos, y pode ser previsto. Ou seja podemos descrever o que sabemos de y através de $P(y|\,\theta, f)$.

- (2) Qual teria sido/deveria ser o valor da variável de controle x?

Conhecidos f, θ e y, queremos saber qual é o valor de x, ou em geral $P(x|y, \theta, f)$. Pense numa situação em que queremos escolher x para ter um y desejado. Impondo o valor de y determinamos quais ações (escolha de x) levam a um comportamento desejado do sistema.

- (3) Estimativa de parâmetros θ

f é dado. Para vários valores de x escolhidos, os de y são medidos. O que podemos dizer sobre o parâmetro θ (p. ex. a aceleração da gravidade g no problema analisado em 6.15)? Estamos interessados em $P(\theta|\{x_i, y_i\}, f)$.

Este tipo de problema é interessante pois também está ligado à descrição matemática de modelos de aprendizagem, por exemplo, em redes neurais.

- (4) Seleção de Modelos: f desconhecido

Temos, por exemplo duas formas funcionais diferentes $f_1(x, \theta_1)$ e $f_2(x, \theta_2)$ e devemos a partir de tudo o que sabemos e um conjunto de pares (x_i, y_i) escolher entre f_1 e f_2 e quem sabe determinar o θ correspondente. Este é o tema deste capítulo.

Escolha Bayesiana entre modelos

Com repeito à relação entre duas variáveis x e y contemplamos duas possibilidades. A relação funcional é descrita pela função $f_1(x, \Theta_1)$ ou pela função $f_2(x, \Theta_2)$. Os parâmetros por sua vez podem ser multidimensionais: $\Theta_1 = (\theta_{11}, \theta_{12}, ...\theta_{1d_1})$ e $\Theta_2 = (\theta_{21}, \theta_{22}, ...\theta_{2d_2})$ Além disso as medidas são corrompidas pela adição de *ruído*. Assim, as possibilidades de escolha são entre

- I) Modelo $\mathcal{M}_1$: $y = f_1(x, \Theta_1) + \xi$ descreve os dados.
- II) Modelo $\mathcal{M}_2$: $y = f_2(x, \Theta_2) + \eta$ descreve os dados.

com um conjunto de dados $D_n = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$.

A informação que temos é $I = f_1, f_2, d_1, d_2, P(\xi), P(\eta)$.

A maneira de proceder é um tipo de teste de hipóteses chamado de teste de hipóteses Bayesiano que consiste em comparar as probabilidades de cada modelo e a posterior escolha do mais provável. Queremos pois calcular $P(\mathcal{M}_i|D_n I)$ e calcular o que se costuma chamar de chances (*odds*)

$$\mathcal{O}_{12} = \frac{P(\mathcal{M}_1|D_n I)}{P(\mathcal{M}_2|D_n I)}. \tag{8.1}$$

Não tendo informação completa recorremos à regra de Bayes

$$P(\mathcal{M}_i|D_n I) \quad = \quad \frac{P(\mathcal{M}_i|I)P(D_n|\mathcal{M}_i I)}{P(D_n|I)}. \tag{8.2}$$

Como sempre, temos a distribuição de probabilidades *a priori* da qual não poderemos escapar; a verossimilhança de observar os dados caso $\mathcal{M}_i$ seja o modelo; e a probabilidade dos dados, que afortunadamente se cancela na expressão 8.1:

$$\mathcal{O}_{12}^{D_n} = \frac{P(\mathcal{M}_1|I)P(D_n|\mathcal{M}_1 I)}{P(\mathcal{M}_2|I)P(D_n|\mathcal{M}_2 I)} \tag{8.3}$$

Se definirmos de forma óbvia as chances *a priori* $\mathcal{O}_{12}^0$ dos modelos podemos escrever

$$\mathcal{O}_{12}^{D_n} = \mathcal{O}_{12}^0 \frac{P(D_n|\mathcal{M}_1 I)}{P(D_n|\mathcal{M}_2 I)} = \mathcal{O}_{12}^0 B_{12} \tag{8.4}$$

onde a razão $B_{12} = \frac{P(D_n|\mathcal{M}_1 I)}{P(D_n|\mathcal{M}_2 I)}$ é amiúde chamada de fator de Bayes.

Para a verossimilhança dos dados condicionado ao modelo, lembramos já ter visto algo parecido na estimativa de parâmetros:

$$P(\Theta_i|D_n \mathcal{M}_i I) \quad = \quad \frac{P(\Theta_i|\mathcal{M}_i I)P(D_n|\Theta_i \mathcal{M}_i I)}{P(D_n|\mathcal{M}_i I)}. \tag{8.5}$$

O que é verossimilhança em 8.2 é denominador em 8.5 que recebe o nome de evidência e estamos a um passo de ver por quê. Se as probabilidades *a priori* dos modelos forem iguais $\mathcal{O}_{12}^{D_n}$ é a razão entre os denominadores de 8.5, que são toda a evidência necessária para decidir entre os modelos. Portanto são chamados de evidência que os dados fornecem, não dos parâmetros mas dos modelos. Assim se pode escrever

$$\begin{aligned} \log \mathcal{O}_{12}^{D_n} &= \log \mathcal{O}_{12}^0 + \log B_{12} \\ &= \log \mathcal{O}_{12}^0 + E(D_n|\mathcal{M}_1 I) - E(D_n|\mathcal{M}_2 I) \end{aligned} \tag{8.6}$$

onde $E(D_n|\mathcal{M}_i I)$ é a evidência logaritmica do modelo $\mathcal{M}_i$, que fornece uma forma aditiva de considerar a informação na escolha de um modelo. Se as chances *a priori* forem muito maiores favorecendo um modelo do que outro, para inverter as preferências, a evidência dos dados tem que ser muito grande [3].

[3] De ce qui precede, nous devons generalement conclure que plus un fait est extraordinaire, plus il a besoin d'etre appuye de fortes preuves. Laplace

Agora veremos com as regras da soma e produto permitem calcular as evidências. O lado esquerdo de 8.5 deve ser normalizado, portanto a integral sobre todas as possibilidades dos parâmetros Θ_i, deve dar 1 e segue que

$$P(D_n|\mathcal{M}_i I) = \int P(\Theta_i|\mathcal{M}_i I) P(D_n|\Theta_i \mathcal{M}_i I) d\Theta_i. \tag{8.7}$$

Pela regra do produto, reescrevemos a distribuição conjunta

$$P(D_n|\mathcal{M}_i I) = \int P(D_n, \Theta_i|\mathcal{M}_i I) d\Theta_i, \tag{8.8}$$

e vemos que este resultado não é nada mais que a regra da marginalização. A expressão 8.7 é o que vamos usar para decidir entre dois modelos.

A integração sobre o espaço dos possíveis valores de Θ pode ser complicado, especialmente se a dimensão d_i do espaço é grande (possivelmente >3 e certamente se> 5) e merece um tratamento aparte que será retomado no capítulo 9 sobre métodos Monte Carlo. Este tipo de técnica só pode ser implementada usando computadores. Hoje em dia isto é corriqueiro, mas explica a necessidade de fugir destes métodos no século XIX e a tendência, por motivos históricos no século XX.

A navalha de Ocam

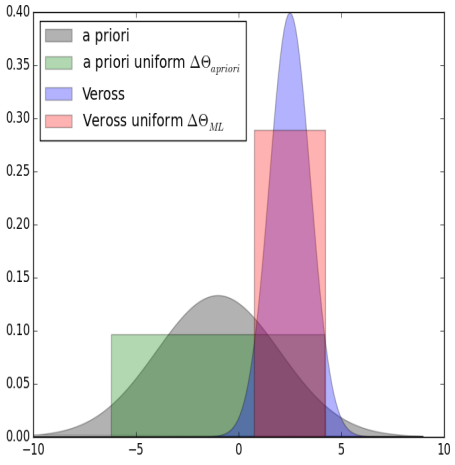


Figura 8.1: A figura mostra em uma dimensão como gaussianas são aproximadas por uniformes, tanto para a distribuição *a priori* quanto para a verossimilhança.

A ideia que simplicidade na explicação é algo desejável vem desde os gregos, mas recebe o nome de navalha de Ocam [4]. Deve este desejo ser algo novo a ser adicionado a nossa desiderta? Veremos que não, que de certa forma modelos mais complexos são automaticamente prejudicados pela sua complexidade ao ser julgados pelos métodos da teoria de probabilidades como foi mostrado por Jaynes.

[4] Procedimento epônimo de William of Ockham ou de Occam. Estas seriam as ortografias certas em inglés e latim respectivamente, mas decidimos agir de acordo a seus preceitos e escrever Ocam. O motivo de esta idéia ser associada a seu nome é por ter escrito *Numquam ponenda est pluralitas sine necessitate*

Se os dados têm informação a mais do que a distribuição *a priori*, esta deve ser moderadamente suave na vizinhança de

$$\Theta_{ML} = \underset{\Theta}{\text{argmax}} P(\Theta_i|\mathcal{M}_i I),$$

onde o subíndice ML significa máxima verossimilhança (maximum likelihood). Como a *a priori* está normalizada, o volume *a priori* plaussível deve ser tal que

$$P(\Theta_{iML}|\mathcal{M}_i I) \approx \frac{1}{|\Delta\Theta_i|^{d_i}_{apriori}}.$$

A largura da verossimilhança determina a região de integração que contribui para a evidência. Aproximadamene seu volume, lembrando que estamos em um espaço de d_i dimensões, é da ordem de $|\Delta\Theta_i|^{d_i}_{ML}$. Segue que é razoável aproximar a evidência

$$\begin{aligned} P(D_n|\mathcal{M}_i I) &= \int P(\Theta_i|\mathcal{M}_i I) P(D_n|\Theta_i \mathcal{M}_i I) d\Theta_i \\ &\approx P(\Theta_{iML}|\mathcal{M}_i I) \int P(D_n|\Theta_i \mathcal{M}_i I) d\Theta_i \\ &\approx P(\Theta_{iML}|\mathcal{M}_i I) P(D_n|\Theta_{iML} \mathcal{M}_i I) |\Delta\Theta_i|^{d_i}_{ML} \\ &\approx P(D_n|\Theta_{iML} \mathcal{M}_i I) \left(\frac{|\Delta\Theta_i|_{ML}}{|\Delta\Theta_i|_{apriori}} \right)^{d_i} \\ &\approx P(D_n|\Theta_{iML} \mathcal{M}_i I) e^{-\lambda_i d_i} \end{aligned} \tag{8.9}$$

onde $\lambda_i = -\log \frac{|\Delta\Theta_i|_{ML}}{|\Delta\Theta_i|_{apriori}}$ e deve ser positivo se os dados são informativos. Vemos que a evidência paga um preço exponencial no número de parâmetros do modelo. O significado de λ não positivo é que a região plausível *a priori* é menor que a região de parâmetros permitida pela verossimilhança. Os dados e o modelo aumentam a incerteza sobre os parâmetros. Há várias possíveis explicações mas todas elas devem ascender uma luz de alerta.

Na comparação de dois modelos, se tudo o resto for similar, a navalha de Ocam, brandida por Bayes decide em favor do mais simples.

Armazém de Critérios de Informação

Suponha que voce não esteja interessado nas minúcias de teoria de informação mas simplesmente queira saber se um ou outro modelo é apoiado pelos seus dados [5] . Há critérios de informação prontos que podem ser muito úteis. Podem também levar a conclusões que não encontram apoio nos dados. O problema em geral é que receitas prontas para usar são muito boas quando um certas hipóteses são satisfeitas. Se o método for usado em outras condições certamente não há garantias e pode ou não justificar as escolhas de modelos. Um dos nomes mais conhecidos nesta área é o de Akaike e seu critério de informação. No original era "An Information Criterion"mas agora é "Akaike IC"ou AIC. Não é justo atribuir-lhe as falhas no uso deste critério, seus trabalhos indicam as condições adequadas de uso. Usar o teorema de Pitagoras numa geometria errada não condena Pitagoras. Há outros critérios como TIC (Takeuchi's IC), BIC (Bayesian IC ou Schwarz Bayesian IC), DIC (deviance IC) que resultam de outras hipóteses simplificadoras. Talvez o leitor ache que neste ponto devamos ter um critério de seleção de critérios. O critério 8.3 para comparação é o que teoria de informação diz que deve ser

[5] Parece paradoxal, pois esta situação significaria que você quer saber o que os dados dizem, mas não quer ter o trabalho de extrair essa informação deles

feito. Pode ser que seja difícil, mas isso não altera o fato que é o que deveria ser usado. Os outros critérios devem ser usados com grãos de sal, pedindo desculpas por tomar atalhos.

Tomando o logaritmo da equação 8.9, mudando o sinal e multiplicando por 2, podemos definir

$$\begin{aligned} IC &= -2\log P(D_n|\mathcal{M}_i I) \\ &\approx -2\log P(D_n|\Theta_{iML}\mathcal{M}_i I) + 2\lambda_i d_i \end{aligned} \tag{8.10}$$

onde sabemos que há várias aproximações que nos levam de IC à expressão da direita. Isto ainda pode ser simplificado, substituindo λ_i por 1, chegamos ao AIC

$$AIC = -2\log P(D_n|\Theta_{iML}\mathcal{M}_i I) + 2d_i \tag{8.11}$$

[6] A aproximação de λ feita no AIC é suspeita e Schwarz considerou que talvez a largura da verossimilhança pudesse cair com o número n de dados como $1/\sqrt{n}$ com base na lei dos grandes números:

$$|\Delta\Theta_i|_{ML} = |\Delta\Theta_i|_{apriori}/\sqrt{n},$$

que leva ao BIC do modelo $\mathcal{M}_i$

$$BIC = -2\log P(D_n|\Theta_{iML}\mathcal{M}_i I) + d_i \log n. \tag{8.12}$$

O folclore é, então que quanto menor o AIC ou BIC de um modelo maior é seu mérito.

E devemos lembrar, olhando de volta à equação 8.3 que tudo isso foi feito sob a hipótese de que *a priori* os modelos são igualmente prováveis. Além disso é importante notar que foi usada a aproximação de que todos os parâmetros são igualmente importantes e que a escolha da escala para cada parâmetro faz com que a região em que a verossimilhança é relevante, tenha o mesmo tamanho $|\Delta\Theta_i|_{ML}$. O mesmo comentário vale para a distribuição *a priori*

Devemos ter cuidado ao usar estas receitas em casos que as aproximações que levaram à equação 8.9 não forem justificáveis. A próxima vez que alguém fale como você sobre telepatia e diga que os dados provam isso ou aquilo e você fique um frio na espinha, lembrará que o simples uso de uma receita pode levar a bobagens.

[6] Outros autores o escrevem o AIC como

$$AIC' = -2\log P(\Theta_{iML}|D_n\mathcal{M}_i I) + 2d_i$$

que difere por uma constante se a distribuição *a priori* de Θ for uniforme. Mas fica difícil entender como isso permitiria comparar modelos diferentes, com parâmetros diferentes se as constantes dependem do modelo.

Quantos picos no espectro?

INCOMPLETO Um dos problemas clássicos em Física Experimental é decidir se os dados são compatíveis com um pico, o que poderia indicar uma passagem para Estocolmo ou se simplesmente o pico é uma ficção. Estamos acostumados a ver coelhos nas nuvens e picos que não existem nos espectros. É melhor tomar cuidado [7].

Vamos começar com um problema mais simples e depois passaremos a outro conceitualmente igual mas um pouco mais difícil técnicamente.

[7] Uma referência muito útil é *Bayesian spectrum analyis* por G. L. Bretthorst. Springer (1988)

Picos num fundo de Ruido Gaussiano

Vimos já vários exemplos em que a distribuição normal aparece. Mais uma vez consideramos um caso em que temos que decidir entre dois modelos

- I) Modelo $\mathcal{M}_1$: $y = B_1(x,\Theta_1) + \xi$ descreve os dados.
- II) Modelo $\mathcal{M}_2$: $y = B_1(x,\Theta_1) + f(x,\rho) + \xi$ descreve os dados.

Exemplo: uma nova partícula ou ruído?

Uma experiência foi feita e temos um conjunto de dados $\{x_i, y_i\}$. Por motivos teoricos, isto é, usando a teoria T_1, suponha que um grupo de cientistas acredita que $y = B_1(x, \Theta_1)$, que é chamado de *fundo* ou *background*. Mas pode ser que conforme a teoria T_2 haja um pico que garantirá gloria e fama a seus descobridores. O pico por motivos teóricos teria uma forma

$$f(x, \rho) = f(x, A, x_0, \gamma) = \frac{A}{1 + (\frac{x - x_0}{\gamma})^2}$$

Esta forma recebe os nomes de vários pesquisadores. Como distribuição de probabilidade está associada a Cauchy e como perfil de linha a Lorentz e Breit e Wigner. Temos os seguintes candidatos como modelos:

- I) Modelo $\mathcal{M}_1$: $y = B_1(x, \Theta_1) + \xi$ descreve os dados.
- II) Modelo $\mathcal{M}_2$: $y = B_1(x, \Theta_1) + f(x, \rho) + \xi$ descreve os dados.

Um segundo grupo sugere que na realidade T_1 deveria ser T_3 e T_2 deveria ser T_4 e portanto acha que devemos considerar

- III) Modelo $\mathcal{M}_3$: $y = B_2(x, \Theta_2) + \xi$ descreve os dados.
- IV) Modelo $\mathcal{M}_4$: $y = B_2(x, \Theta_2) + f(x, \rho) + \xi$ descreve os dados.

Note que todos os candidatos tem o mesmo tipo de ruído ξ corrompendo os dados, pois tem a mesma informação sobre o aparelho experimental. Poderiamos ter grupos diferentes usando equipamento diferente mas por agora isso não é necessário.

Os backgrounds escolhidos pelos dois grupos são

$$\begin{aligned} B_1(x|\Theta_1) &= (1 - x^{\frac{1}{3}})^b x^{a_0 + a_1 \log x} \\ \Theta_1 &= (a_0, a_1, b) \\ B_2(x|\Theta_2) &= (1 - x^{\frac{1}{3}})^b (x^{c_0} + x^{a_0 + a_1 \log x}) \\ \Theta_2 &= (a_0, a_1, b, c_0) \end{aligned} \tag{8.13}$$

Para facilitar a notação chamamos $C_i(x) = B_i(x, \Theta_i) + f(x, \rho)$. Os grupos pelo menos concordam com o processo de ruído que contamina os dados

$$\xi \sim \text{Poisson}(\lambda) \tag{8.14}$$

mas que o parâmetro da distribuição de Poisson varia com a x :

$$\lambda(x) = \tag{8.15}$$

Verossimilhança dos Modelos

Para escrever uma forma para a Verossimilhança dos modelos

$$\begin{aligned} P(D_n|\mathcal{M}_i I) &= \int d\Theta P(\Theta|\mathcal{M}_i I) P(D_n|\Theta_i \mathcal{M}_i I) \quad \text{(sem pico)} \\ P(D_n|\mathcal{M}_i I) &= \int d\Theta d\rho P(\Theta|\mathcal{M}_i I) P(\rho|\mathcal{M}_i I) P(D_n|\Theta_i \rho \mathcal{M}_i I) \quad \text{(com pico)} \end{aligned}$$

precisamos introduzir os priors e as verossimilhanças sobre os parâmetros. Para as distribuições *a priori* é razoável usar distribuições uniformes em pequenas regiões compatíveis parâmetros b sejam positivos. Os parâmetros não são todos simétricos com alguns vínculos físicos que possamos saber, como por exemplo unitariedade

garante que os na sua importância, o que sugere que AIC e BIC não devem ser razoáveis.

A verossimilhança dos parâmetros

$$P(D_n|\Theta_i \mathcal{M}_i I) = \prod_{a=1}^{n} e^{-\lambda(x_a)}$$

9
Monte Carlo

A descrição das propriedades dos sistemas até aqui estudados foi reduzida ao cálculo de integrais em alta dimensão, valores esperados de funções

$$\langle f \rangle = \int f P d\mu.$$

Obviamente a medida $d\mu$ pode ser discreta, como é o caso quando tratamos variáveis discretas. É essencial encontrar meios numéricos de realzar estes cálculos, dado que o conjunto de modelos exatamente integráveis é bem menor que o de modelos interessantes. A classe de métodos de Monte Carlo é sem dúvida a mais importante de todas as ferramentas numéricas à nossa disposição. A arte de fazer Monte Carlos não será abordada aqui, simplesmente uma iniciação às idéias sem entrar em detalhes que se tornam necessários para seu uso profissional.

Integração Numérica em espaços de alta dimensão

Considere o método de integração numérica mais simples, chamado método do trapézio (ver de Vries). Aproximamos a integral

$$I = \int_a^b f(x)\, dx$$

por

$$I_T = \frac{1}{N}(\frac{1}{2}f(x_1) + \sum_{i=2}^{N-1} f(x_i) + \frac{1}{2}f(x_N)), \tag{9.1}$$

podemos mostrar que o erro cometido é proporcional a h^2, onde $h = (b-a)/N$, escrevemos então que

$$I = I_T + \vartheta(h^2).$$

Esta estimativa do erro também vale para integrais multidimensionais. Métodos mais sofisticados, baseados neste (e.g. estilo Romberg-Richardson), levam a melhorias no expoente de h, mas como veremos a seguir, não suficientes.

O custo computacional no cálculo de uma integral é proporcional ao número de vezes que a rotina que calcula o integrando é chamada dentro do programa. Na fórmula do trapézio acima este número de chamadas é N. Suponhamos um problema típico de Mecânica Estatística, por exemplo um gás dentro de uma caixa. Temos da ordem de $k = 10^{23}$ moléculas mas digamos que para poder lidar com o problema temos somente $k = 20$. Uma aproximação drástica, mas

veremos não suficiente. Neste caso é necessário lidar com integrais do tipo

$$Z = \int g(\{r_{ix}, r_{iy}, r_{iz}\}) dr_1^3 dr_2^3 ... dr_k^3$$

uma integral em $d = 3k = 60$ dimensões. Suponhamos que o volume da caixa seja $V = L^3$, e dividimos cada uma dos d eixos em intervalos de tamanho h . Isto significa uma grade com

$$N = \left(\frac{L}{h}\right)^d$$

pontos. Suponhamos que escolhemos um h extremamente grande, tal que $L/h = 10$, ou seja cada eixo será dividido em somente 10 intervalos. Assim temos

$$N = 10^{60}$$

pontos na grade.O quê significa um número tão grande como 10^{60}? Suponhamos que a máquina que dispomos é muito veloz, ou que a função que queremos integrar é muito simples, tal que cada chamada à subrotina demore somente 10^{-10}segundos. O tempo que demorará para calcular I_T é $10^{50}s$. Para ver que isso é muito basta lembrar que a idade do universo é da ordem de $4\,10^{17}s$, portanto nosso algoritmo levará da ordem de 10^{31}idades do universo. Não precisamos muito mais para que nos convençamos a procurar outro método de integração. Variantes do método de trapézio não ajudam muito. Infelizmente o que temos disponível, o Monte Carlo não é muito preciso, mas é muito melhor que isso.

Monte Carlo

Teorema Central do Limite: revisitado

Considere uma variável X com valores x em um intervalo dado e distribuição $P(x)$. Assumimos que os valores médios $\langle x \rangle$ e $\langle x^2 \rangle$existem e são finitos.[1]A variância σ_x é definida por

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$$

que também é finita.

[1] Definimos os momentos $\langle x^n \rangle = \int x^n P(x) dx$

Considere ainda uma sequência de N amostragens independentes de X: $\{x_i\}_{i=1,...N}$, e outra variável Y com valores y dados por

$$y = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Assintoticamente, isto é para N grande,a distribuição de y se aproxima de uma distribuição gaussiana, podemos escrever que aproximadamente

$$P(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\langle y \rangle)^2}{2\sigma_y^2}}$$

A aproximação é boa na região central da gaussiana e melhora quando N cresce. O valor médio de y e sua variancia são

$$\langle y \rangle = \langle x \rangle \quad \text{e} \quad \sigma_y = \frac{\sigma_x}{\sqrt{N}}$$

Notem que se o objetivo for encontrar o valor esperado de x, que é $\langle x \rangle$, e não for possível realizar a integral, podemos estimar $\langle x \rangle$ a partir de y (isso pode ser generalizado para o cálculo de

$\langle f \rangle = \int f P(x) dx$.) Qual é vantagem sobre simplesmente fazer uma medida (amostragem) de x? É que neste último caso o erro seria da ordem de σ_x, enquanto que a estimativa baseada em y terá erro estimado em $\sigma_y = \sigma_x / \sqrt{N}$, portanto **o erro da estimativa é independente da dimensão de x.** Para grandes dimensões isso é uma grande vantagem. O problema é que para reduzir o erro por um fator 2 é necessário trabalhar 4 vezes mais duro. E isso para o caso em que as amostras são independentes. O erro pode ser diminuido não só aumentando N mas também se mudarmos σ_x. Esse é o objetivo da técnica de amostragem por importância.

Exercício : Considere uma variável aleatória X que toma valores $-\infty < x < \infty$, com probabilidade $P(x)$. é dado que $\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$ é finito. Dado $y = \frac{1}{N} \sum_{i=1}^{N} x_i$ mostre, a partir de

$$P(y) = \int \cdots \int dx_i \cdots dx_N \delta \left(y - \frac{1}{N} \sum_{i=1}^{N} x_i \right) \prod_{i=1}^{N} P(x_i)$$

que $P(y)$ é aproximada por uma gaussiana para N grande. Determine a variancia de y.

Exercício:Distribuição de Cauchy Considere o problema acima, exceto que $\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$ é infinito pois $P(x) = \frac{b}{\pi(b^2+x^2)}$. Encontre a distribuição $P(y)$ de y, Note que não é gaussiana para nenhum valor de N. As integrais necessárias são relativamente fáceis de calcular pelo método dos resíduos.

Monte Carlo

A idéia básica é aproximar uma integral I por I_{MC}

$$I = \int_a^b f(x)\, dx \simeq I_{MC} = \frac{1}{N} \sum_{i=1}^{N} f(x_i) \tag{9.2}$$

onde os $\{x_i\}$ são escolhidos aleatoriamente de forma independente da distribuição uniforme em $[a, b]$. Se a integral de f^2 existir e for finita, e se as amostras $f(x_i)$ forem estatisticamente independentes - e isto é um grande *se* - então o erro da estimativa MC acima será dado por

$$\sigma_{I_{MC}} = \frac{\sigma_f}{\sqrt{N}}$$

e podemos estimar σ_f a partir dos dados da amostragem

$$\sigma_f^2 \approx \frac{1}{N} \sum f^2(x_i) - \left[\frac{1}{N} \sum f(x_i) \right]^2.$$

Embora eq. (9.2) possa ser usada para o cálculo da integral, em geral é necessário reduzir a variancia da função f. Isso é possível através de uma mudança de variáveis, que nem sempre pode ser implementada analiticamente e será descrita a seguir[2].

O método que iremos descrever não é útil, em geral, para realizar estimativas de Monte Carlo, mas servirá para motivar e sugerir novos caminhos. Imagine uma integral da forma

$$I = \int f(x) w(x) dx,$$

em geral essa separação do integrando em duas funções é muito natural. Tipicamente x é um vetor em um espaço de muitas dimensões mas $f(x)$ só depende de algumas poucas componentes de

[2] Uma forma trivial de conseguir a redução de σ_f é considerar variações da identidade $\int_0^1 f(x)dx = \int_0^1 g(x)dx$, onde $g(x) = \frac{1}{2}(f(x) + f(1-x))$. Note que o cálculo de g é duas vezes mais caro que o de f, portanto devemos ter $\frac{\sigma_f^2}{2\sigma_g^2} > 1$ para ter ganho efetivo

x, enquanto que $w(x)$ depende de todas. Suponha que $w(x)$ esteja normalizado. i.e:

$$\int w(x)dx = 1$$

Ilustraremos a separação em uma dimensão, tomemos o intervalo de integração $(0,1)$ e façamos a seguinte mudança de variáveis

$$y(x) = \int_0^x w(z)dz \tag{9.3}$$

$$y(0) = 0,\ y(1) = 1$$

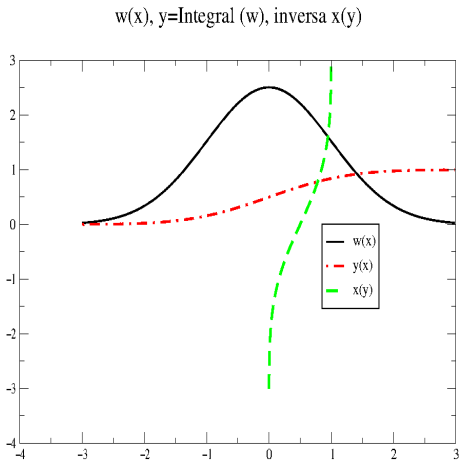
então $dy = w(x)dx$ e a integral toma a forma

$$I = \int f(x(y))dy$$

e a aproximação Monte Carlo é

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(x(y_i)) \tag{9.4}$$

onde os valores de y_i serão amostrados de uma distribuição uniforme no intervalo $(0,1)$. Depois basta calcular a função que relaciona y e x (eq. [9.3]). A função inversa permite calcular o valor de x onde deverá ser calculada a função $f(x)$. Este método assume que saibamos fazer a integral da equação 9.3, mas não é em geral possível fazé-lo de forma analítica.



Exemplos analíticos.

Ao realizar um cálculo MC teremos, tipicamente, acesso a um gerador de números aleatórios distribuidos uniformemente em $(0,1)$. O objetivo é, aqui de forma analítica e posteriormente, de forma numérica, mostrar como gerar números aleatórios distribuidos de acordo com uma distribuição dada a partir da distribuição disponível. Apresentaremos dois casos muito úteis que podem ser feitos de forma analítica.

Se duas variáveis (em e.g. R^N) tem uma relação funcional $y = \sigma(x)$, então suas densidades de probabilidade estão relacionadas assim

$$P_Y(y)dy = P_X(x)dx$$

$$P_Y(y)dy = P_X(x)\left|\frac{\partial x}{\partial y}\right| dy \tag{9.5}$$

onde $\left|\frac{\partial x}{\partial y}\right|$é o jacobiano da transformação e $dy = \prod_i dy_i$. No caso de interesse numérico temos aproximadamente

$$P_Y(y)dy = dy, \qquad 0 \leq y_i < 1, i = 1...N$$

e zero fora.

Distribuição Exponencial

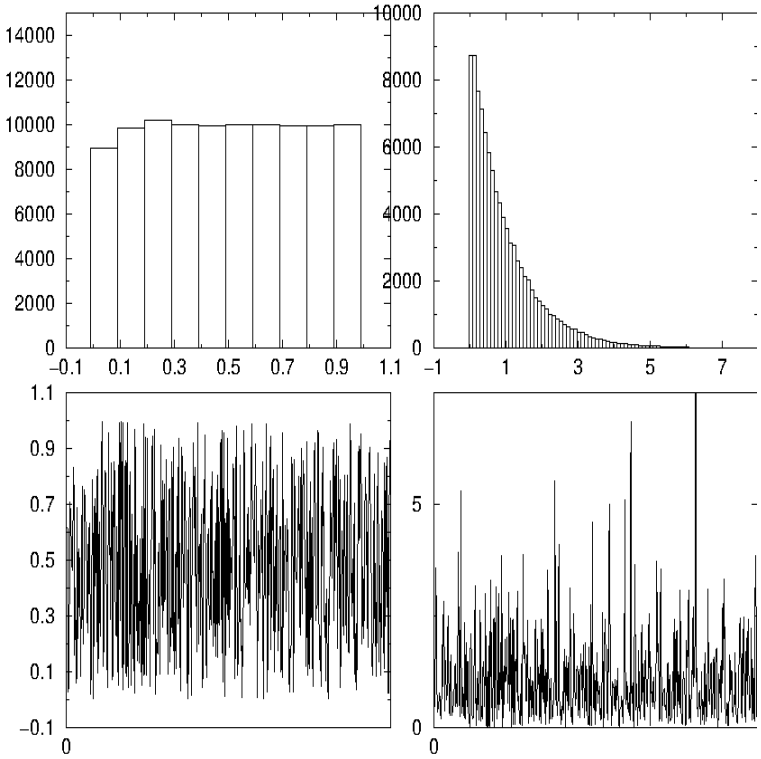
Suponha que queremos gerar amostras de uma distribuição exponencial. i.e $P_X(x) = exp(-x)$. Integrando a eq. (9.5) obtemos

$$y(x) = \int_0^{y(x)} P_X(x)\frac{dx}{dy}dy$$
$$y(x) = \int_0^x P_X(x)dx = \int_0^x e^{-z}dz$$
$$y(x) = 1 - exp(-x)$$

ou $x = -ln(y)$ terá a distribuição exponencial desejada, pois se y é uniforme em $(0,1)$ então $1-y$ também o é. Portanto é suficiente para gerar números distribuidos exponencialmente usar uma função que gera números aleatórios de distribuição uniforme `RAND(SEED)` e somente uma linha de (pseudo-) código

```
x=-log( RAND(SEED))
```

Compare na figura a distribuição uniforme (esquerda) e a a exponencial (direita) (abaixo : série temporal, acima : histogramas)



Distribuição Normal

Para gerar números distribuidos de acordo com a distribuição normal é tentador gerar um número grande de amostras de $P_Y(y)$ e definir $x = \frac{1}{\sqrt{N}} \sum y_i - \frac{\sqrt{N}}{2}$, que terá distribuição gaussiana (approximadamente). O problema é o custo computacional, pois requer N chamadas da função RAN para gerar uma só amostra de x. Portanto nunca gere números aleatórios gaussianos dessa maneira. Mais fácil, do ponto de vista computacional é partir da equação (9.5) . O método de Box-Muller, mostrado a seguir é muito mais eficiente, pois gera dois números gaussianos para duas chamadas da função geradora de uniformes. Dados y_1 e y_2 obtemos x_1e x_2 a partir da transformação:

$$\begin{aligned} x_1 &= \sqrt{-2\ln y_1}\cos 2\pi y_2 \\ x_2 &= \sqrt{-2\ln y_2}\sin 2\pi y_2 \end{aligned}$$

mostraremos que a sua distribuição conjunta será $P_X(x_1, x_2) = \frac{1}{2\pi} exp(-(x_1^2 + x_2^2)/2)$. Integrando a eq.(9.5) temos:

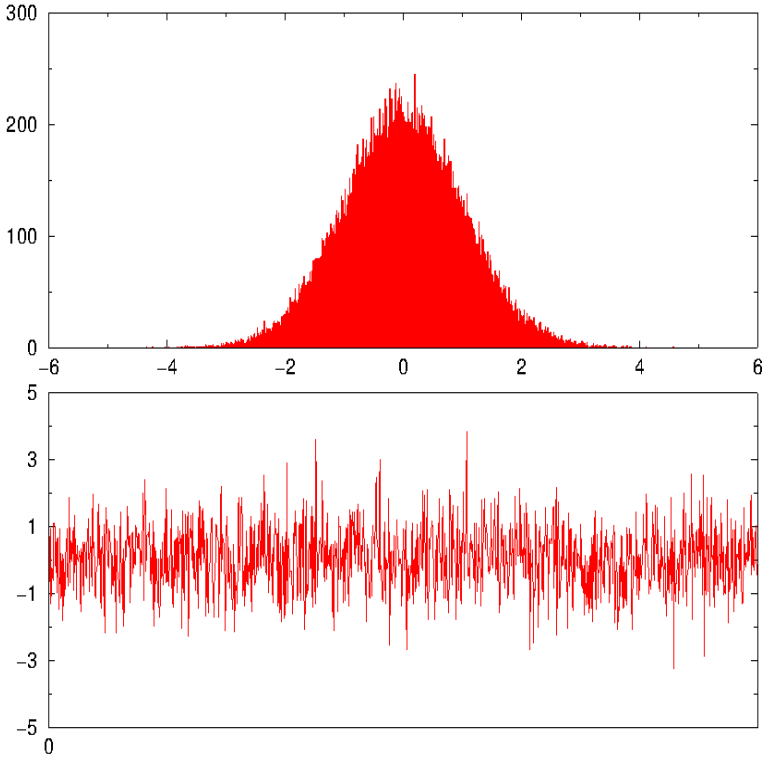
$$\int\int P_Y(y'(x_1, x_2)) \left|\frac{\partial y}{\partial x}\right| dy'_1 dy'_2 = \int\int P_X(x') dx'_1 dx'_2$$

segue o resultado pois o jacobiano é:

$$J = \left|\frac{\partial y}{\partial x}\right| = \frac{y_1}{2\pi} = \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}}$$

Usando este método obtemos a figura que segue, abaixo temos a série temporal e acima o histograma dos desvios normais:

Estes resultados de muita utilidade na simulação de distribuições gaussianas multivariadas, a ser discutidas posteriormente.



Métodos Estáticos: rejeição

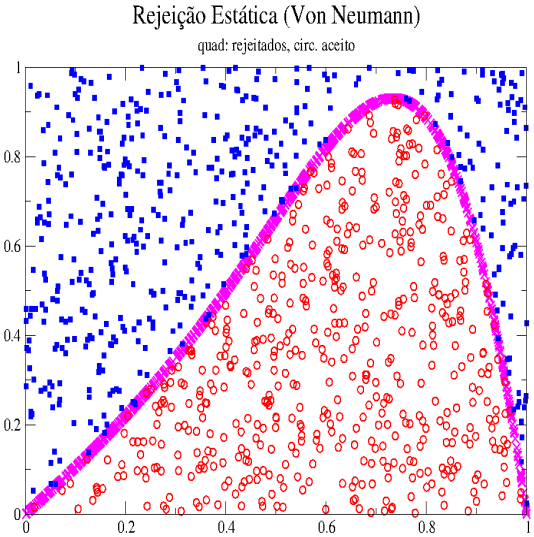
Raramente é possível realizar as integrais que permitem descobrir a transformação exata de variáveis e devemos então encontrar uma forma gerar diretamente os x com a distribuição $w(x)$. Os métodos que apresentaremos podem ser divididos em duas classes, estáticos e dinâmicos. Na primeira os números são gerados independentemente um dos outros[3] , enquanto que na segunda classe, construiremos um processo dinâmico que usara informação anterior para gerar o próximo número.

[3] Tão independentemente quanto o gerador de números pseudo-alealtórios o permitir.

Suponhamos que a região onde $w(x) \neq 0$ está contida em (a, b) e que ela é limitada, tal que $w(x) < c$. No método de rejeição estático geramos dois NAU ξ e η e definimos

$$\rho = a + (b - a)\xi, \quad \varphi = c\eta$$

o valor de ρ será aceito como o novo valor de x se $\varphi \leq w(\rho)$ e rejeitado se não.



A sequência de números aceitos x são as abicissas dos circulos na figura acima.

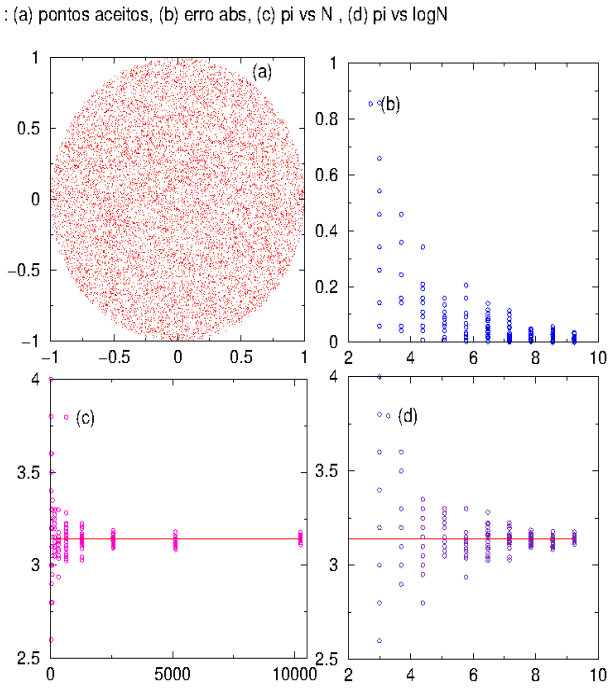
Círculo

Exemplo: calcule π

A figura mostra os resultados de algumas simulações para estimar π. Foram gerados N_{MC} pares de numeros aleatórios (x, y). Se $z = x^2 + y^2 \leq 1$ então o ponto é aceito, de outra forma é rejeitado.Os resultados foram obtidos para $N_{MC} = 10 * 2^m$ passos de Monte Carlo, com $m = 2, 4..., 20$. O resultado (figura (a) abaixo esq. acima) mostra os pares aceitos. Continuando no sentido horário, temos os resultados respectivamente :

- (b) do erro absoluto contra $\log(N_{MC})$
- (d) resultado de π_{MC}=(numero aceito/numero total) contra $\log(N_{MC})$
- (c) resultado de π_{MC}=(numero aceito/numero total) contra N_{MC}, os gráficos mostram os resultados de 20 corridas independentes. A

dispersão dos pontos nos dá uma idéia dos erros estatísticos. As barras horizontais mostram o valor 3.14159



Métodos Dinâmicos

A idéia por trás dos processos de Monte Carlo dinâmicos é a de um processo estocástico em tempo discreto. Um processo determinístico, em oposição, é tal que dado um certo conjunto de informações, é possível -em principio- determinar a evolução futura. Um processo estocástico serve para modelar o caso em que a informação é incompleta e às várias possibilidades de evolução são atribuidas probabilidades. O objetivo é construir um processo estocástico com distribuição de equilíbrio associada igual ao $w(x)$ dado. Note que o processo estocástico é uma caminhada aleatória. Consideremos um grande número de caminhadas independentes. O processo deve ser tal que a fração das caminhadas na vizinhança de x seja proporcional a $w(x)$, pelo menos se aproxime dela assintoticamente no tempo, e chamaremos de $P(x|t)$ à distribuição no instante t.

O conceito principal para entender o processo de MC dinâmico é a probabilidade de transição, $\Gamma(x|x_n, x_{n-1}, ...x_0, ...)$, que em princípio pode depender de toda a história da evolução. Um processo é chamado Markoviano (de 1 passo) se só depende da estado atual[4]

$$\Gamma(x_{n+1}|x_n, x_{n-1}, ...x_0) = \Gamma(x_{n+1}|x_n),$$

ou de forma vaga, para onde o processo vai (o futuro), depende somente de onde está agora (o presente) e não do passado. Chamaremos a sequência $\{x_0, x_1, ..., x_n, ...\}$ de cadeia de Markov[5]. . Para os nossos objetivos estas cadeias são ferramenta suficiente.

Um ingrediente necessário que o processo deverá satisfazer é convergência para o equilíbrio. A distribuição de equilíbrio ou

[4] outra notação comum é $\Gamma(x_n \to x_{n+1})$

[5] A cadeia de Markov é caracterizada pelas probabilidades de transição e pela distribuição inicial de probabilidades de x

invariante ou estacionária deve satisfazer a condição de estacionaridade

$$w(x) = \int w(z)\Gamma(x|z)dz, \tag{9.6}$$

mas se não for estacionária teremos a relação entre a probabilidade no instante t e no seguinte $t+1$ dada por

$$P(x|t+1) = \int P(z|t)\Gamma(x|z)dz,$$

Dado que as probabilidades de transição são normalizadas $1 = \int \Gamma(z|x)dz$ segue que

$$\Delta P(x|t) = P(x|t+1) - P(x|t) = \int P(z|t)\Gamma(x|z)dz - P(x|t)\int \Gamma(z|x)dz,$$

$$\Delta P(x|t) = P(x|t+1) - P(x|t) = \int \left[P(z|t)\Gamma(x|z) - P(x|t)\Gamma(z|x)\right] dz \tag{9.7}$$

A interpretação é imediata, a variação da probabilidade, de um instante para o outro, tem duas contribuições, de entrada e saída. O primeiro termo $[P(z|t)\Gamma(x|z)]\,dz$ representa o número de caminhadas em um volume dz em torno de z no instante t, que fizeram a sua transição para x no instante $t+1$. O segundo termo representa a saida, isto é os que estavam em x e escapam para z. A integral leva em conta todas as contribuições do espaço. É óbvio a partir das eqs. [9.6, 9.7]

$$\Delta w(x) = \int \left[w(z)\Gamma(x|z) - w(x)\Gamma(z|x)\right] dz = 0.$$

Há várias escolhas possíveis de Γ para satisfazer esta relação. A escolha mais simples sugere impor uma condição

$$w(z)\Gamma(x|z) = w(x)\Gamma(z|x) \tag{9.8}$$

que se a matriz de probabilidade de transições satisfizer então $w(x)$ será estacionaria. Esta condição, chamada de **balanceamento detalhado,** não é necessária, mas só suficiente. Além de haver motivações físicas para impó-la como condição deve ser ressaltado que é talvez a forma mais fácil de realizar o objetivo para construir a matriz de transição. Com qualquer escolha que satisfaça a condição eq. [9.8] $w(x)$ é um ponto fixo da dinâmica. Mas a pergunta que resta é sobre a estabilidade. É razoavel esperar a estabilidade dado que se em t , $P(x|t) > w(x)$, o número de caminhantes que sairão da região de x para z será maior que o que sairiam se a probabilidade fosse $w(x)$. Analogamente, se em t , $P(x|t) < w(x)$ então o número será menor.

Há várias maneiras de satisfazer a equação [9.8]. Embora todas levem a algoritmos corretos, no sentido que

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{9.9}$$

é uma aproximação que melhora para maiores valores de N, algumas serão eficientes enquanto outras não. Diferentes escolhas levam a diferentes sequências, e a pergunta relevante é: quanta informação nova é trazida por uma nova amostragem? A função de autocorrelação normalizada, que é fundamental para poder julgar a eficiência do MC, é definida por

$$C(k) \equiv \frac{\langle f_n f_{n+k}\rangle - \langle f\rangle^2}{\langle f_n f_n\rangle - \langle f\rangle^2}$$

onde

$$\langle f \rangle = \int f(x) w(x) dx$$

$$\langle f_n f_{n+k} \rangle = \int\int f(x_n) f(x_{n+k}) w(x_n) \Gamma^k(x_{n+k}|x_n) dx_{n+k} dx_n$$

e

$$\Gamma^k(x_{n+k}|x_n) = \int ... \int \Gamma(x_{n+k}|x_{n+k-1}) \Gamma(x_{n+k-1}|x_{n+k-2}) ... \Gamma(x_{n+1}|x_n) dx_{n+k-1} dx_{n+k-2} ... dx_{n-1}$$

é a probabilidade de transição em k passos. É óbvio que não é, em geral, possível calcular a autocorrelação, mas podemos estimá-la a partir das amostras colhidas:

$$C_{MC}(k) \equiv \frac{\langle f_n f_{n+k} \rangle_{MC} - \langle f \rangle^2_{MC}}{\langle f_n f_n \rangle_{MC} - \langle f \rangle^2_{MC}}$$

onde definimos a média (empírica) sobre a amostra de dados

$$\langle f_n f_{n+k} \rangle_{MC} = \frac{1}{N-k} \sum_{i=1}^{N-k} f(x_i) f(x_{i+k})$$

Tipicamente -mas não sempre - $C(k)$ tem um decaimento exponencial:

$$C(k) = e^{-k/\tau}$$

τ é tempo de correlação exponencial e mede a eficiência do processo em gerar números aleatórios independentes distribuidos de acordo com $w(x)$. Agora podemos escrever

$$I = \int_a^b f(x) w(x) dx \simeq I_{MC} = \frac{1}{N} \sum_{i=1}^{N} f(x_i) \pm \sigma_f \sqrt{\frac{2\tau}{N}}$$

onde assumimos que depois de um tempo (em unidades de 1 passo MC) aproximadamente 2τ as novas amostras serão estatisticamente independentes e o numero efetivo de amostras será reduzido por esse fator.

Outro tempo importante é τ_R, o tempo de relaxação para o equilíbrio. Este mede quanto tempo demora para que o processo estocástico perca memória das condições iniciais e os x sejam efetivamente reprsentativos de $w(x)$. Do ponto de vista de eficiência é razoavel não considerar e.g. os primeiros $10\tau_R$ passos gerados pelo processo. Se $C(k)$ efetivamente decair exponencialmente esses dois tempos são iguais, mas há casos em que não, e.g. perto de transições de fase críticas.

Algoritmo de Metropolis

O processo de geração dos números x_n será separado em duas partes. Em primeiro lugar definimos a probabilidade de *tentativa de mudança* $T(x_T|x_n)$, que determina a probabilidade de estando no tempo n em x_n, seja escolhido o ponto x_T como candidato ao próximo passo da sequência. Uma vez gerado x_T passamos à segunda parte, que é onde se decide se é feita a transição $x_n \to x_{n+1} = x_T$, ou seja x_T é aceito ou se não. Neste caso de rejeição fazemos a transição trivial $x_n \to x_{n+1} = x_n$, de forma que x_n é incluido novamente na sequência, Isto é feito introduzindo a *matriz de aceitação* $A(x_{n+1}|x_T)$. Ou seja

$$\Gamma(x|z) = A(x|z) T(x|z)$$

e a condição de balanceamento detalhado, para todo par de pontos $x \neq z$ toma a forma

$$A(x|z)T(x|z)w(z) = A(z|x)T(z|x)w(x)$$

que é satisfeita por uma familia de escolhas possíveis, em particular se definirmos

$$A(x|z) = F\left(\frac{w(x)T(z|x)}{w(z)T(x|z)}\right)$$

e F tal que

$$\frac{F(a)}{F(1/a)} = a, \text{ para todo } a \tag{9.10}$$

Para escolher $T(.|.)$ é útil definir uma distância entre configurações. Se estivermos falando de graus de liberdade no espaço euclideano, a soma das distâncias entre uma parícula nas configurações x e z é uma boa escolha. Para um sistema de Ising podemos medir a distância pelo número de spins diferentes entre duas configurações. Definimos uma região $\mathcal{B}(x|z)$, uma bola, e a função indicadora $\chi_{\mathcal{B}}$, que toma valores 1 dentro da bola e zero fora. A escolha mais comum, para a probabilidade de tentativa de mudança é tomar

$$T(z|x) = \text{constante dentro } \mathcal{B}(x|z) = \frac{1}{|\mathcal{B}(x|z)|}$$

e zero fora da de $\mathcal{B}$. Isso leva a uma taxa de tentativas simétricas ($T(z|x) = T(x|z)$,)e portanto basta tomar

$$\frac{A(x|z)}{A(z|x)} = \frac{w(x)}{w(z)}$$

A escolha associada ao nome de Metropolis ( ) é

$$F(a) = \min(1, a).$$

Para verificar que satisfaz 9.10, note que há dois casos:

- $a \geq 1$ segue que $\frac{\min(1,a)}{\min(1,a^{-1})} = \frac{1}{a^{-1}} = a$
- $a < 1$ segue que $\frac{\min(1,a)}{\min(1,a^{-1})} = a$

o que leva ao seguinte

Algoritmo de Metropolis:

1. escolha o valor inicial x_0
2. dado x_n determinaremos x_{n+1}: escolha um valor de tentativa x_T (uniformemente dentro de uma bola de raio d em torno de x_n)
3. verifique se $w(x_T)$ é maior ou menor que $w(x_n)$. Defina $r = \frac{w(x_T)}{w(x_n)}$ (para não ficar recalculando $w(x)$, que pode ser muito caro.)
   - Se $r \geq 1$ (i.e. $w(x_T) \geq w(x_n)$) então **aceita** : $x_{n+1} = x_T$
   - Se $r < 1$ (i.e. $w(x_T) < w(x_n)$) então escolhe um número aleatório uniforme $0 \leq \xi < 1$ e

     aceita : $x_{n+1} = x_T$ se $r \geq \xi$ (i.e. $w(x_T) \geq w(x_n)\xi$)
     rejeita : $x_{n+1} = x_n$ se $r < \xi$ (i.e. $w(x_T) \leq w(x_n)\xi$)
4. Guarda informação sobre x_n.
5. Se critério de parada não for satisfeito, volta a 2

Imagine o caso em que a função $w(x)$ pode ser parametrizada da forma

$$w(x) = \frac{e^{-\beta E(x)}}{Z}$$

esse é um dos casos mais interessantes (distribuição d Boltzmann-Gibbs) e a função $E(x)$ é interpretada como a energia de um sistema no estado x ou a função custo de um processo. Z é uma constante em relação a x mas depende do parâmetro β que em física é intepretado como o inverso da temperatura. Este tipo de função ocorre quando a probabilidade que devemos atribuir a uma dada configuração é baseada na informação que temos sobre o valor medio $< E(x) >$e é o resultado de encontrar a distribuição com a máxima entropia consistente com a informação dada.

O algoritmo de Metropolis pode ser redescrito da seguinte forma:

1. escolha o valor inicial x_0
2. dado x_n determinaremos x_{n+1}: escolha um valor de tentativa x_T (uniformemente dentro de uma bola de raio d em torno de x_n)
3. verifique se $E(x_T)$ é maior ou menor que $E(x_n)$.
   - Se $E(x_T) \leq E(x_n)$ então **aceita** : $x_{n+1} = x_T$
   - Se $E(x_T) \geq E(x_n)$ então escolhe um número aleatório uniforme $0 \leq \xi < 1$ e

     aceita : $x_{n+1} = x_T$ se $\exp(-\beta(E(x_T) - E(x_n)) \geq \xi$
     rejeita : $x_{n+1} = x_n$ se $\exp(-\beta(E(x_T) - E(x_n)) < \xi$.
4. Guarda informação sobre x_n.
5. Se um critério de parada não for satisfeito, volta a 2.

A processo realiza a caminhada aleatória de forma que uma diminuição na energia é sempre aceito, mas se há uma tentativa de escolha de um lugar de energia mais alta, a tentativa não é automaticamente rejeitada. Se o aumento de energia for muito grande é grande a probabilidade que seja rejeitada, mas se não for, é grande a de ser aceita. A escala de grande ou pequeno é determinada pela razão dos fatores de Boltzmann de cada configuração.

Atrator da dinâmica

Daremos argumentos em defesa da posição que a distribuição $w(x)$ é um ponto fixo atrativo da equação 9.7 para o algoritmo de Metropolis.

Defina os conjuntos

$$B_x(z) = \{z|T(z|x) > 0\} C_+(x) = \{x|P(x) > w(x)\}, \quad (9.11)$$
$$C_0(x) = \{x|P(x) = w(x)\}, \quad (9.12)$$
$$C_-(x) = \{x|P(x) < w(x)\}. \quad (9.13)$$

Defina $\delta P(x) = P(x|t) - w(x)$, a diferença entre a distribuiçao em um dado instante t e a de equilíbrio. É claro que dado que as distribuições estão normalizadas, teremos

$$\begin{aligned} 0 &= \int \delta P(x) dx \\ &= \int_{C_+(x)} \delta P(x) dx + \int_{C_-(x)} \delta P(x) dx \end{aligned}$$

onde o primeiro termo contém as contribuições postivas e o segundo as negativas. A equação da dinâmica

$$\begin{aligned}\Delta P(x|t) &= \int_{B_x(z)} [P(z|t)\Gamma(x|z) - P(x|t)\Gamma(z|x)]\, dz \\ &= \int_{B_x(z)} [(w(z)+\delta P(z))\Gamma(x|z) - (w(x)+\delta P(x))\Gamma(z|x)]\, dz \\ &= \int_{B_x(z)} [\delta P(z)\Gamma(x|z) - \delta P(x)\Gamma(z|x)]\, dz. \qquad (9.14)\end{aligned}$$

Agora integramos sobre o conjunto de configurações $C_+(x)$

$$\Delta P_+ = \int_{C_+(x)} \Delta P(x|t)dx \quad = \quad \int_{B_x(z),C_+(x)} [\delta P(z)\Gamma(x|z) - \delta P(x)\Gamma(z|x)]\, dzdx,$$

e separamos as configurações z em $C_\pm(z)$

$$\begin{aligned}\Delta P_+ &= \int_{C_+(x),C_+(z)} [\delta P(z)\Gamma(x|z) - \delta P(x)\Gamma(z|x)]\, dzdx \\ &+ \int_{C_+(x),C_-(z)} [\delta P(z)\Gamma(x|z) - \delta P(x)\Gamma(z|x)]\, dzdx.\end{aligned}$$

A integral sobre $C_+(x), C_+(z)$ é nula devido a que é simétrica e antisimétrica ante trocas $z \leftrightarrow x$. Portanto

$$\begin{aligned}\Delta P_+ &= \int_{C_+(x),C_-(z)} [\delta P(z)\Gamma(x|z) - \delta P(x)\Gamma(z|x)]\, dzdx. \\ &\leq 0, \qquad (9.15)\end{aligned}$$

pois o termo com sinal positivo tem $\delta P(z)$ que é sempre negativo em $C_-(z)$, e o termo com sinal negativo tem $\delta P(x)$ que é sempre positivo em $C_+(x)$.

Analogamente , integrando a equação 9.14 sobre o conjunto de configurações $C_-(x)$, vemos que

$$\begin{aligned}\Delta P_- &= \int_{C_-(x)} \Delta P(x|t)dx \qquad (9.16) \\ &= \int_{C_-(x),C_+(z)} [\delta P(z)\Gamma(x|z) - \delta P(x)\Gamma(z|x)]\, dzdx. \\ &\geq 0. \qquad (9.17)\end{aligned}$$

Estes resultados sugerem que a diferença entre $P(x|t)$ e $w(x)$ diminui ao iterar a dinâmica. A dinâmica é tal que nas regiões em que $P(x|t)$ é maior do que deveria ser, diminui. Onde é menor aumenta.

Modelo de Ising

Novamente visitaremos o laboratório de Ising. Há vários livros que tratam deste problema de forma mais completa. Aqui apresentaremos somente alguns detalhes que permitirão que o leitor comece seu simulador de Monte Carlo em um problema que tem propriedades críticas interessantes.

Que um algorimo funcione bem para o modelo de Ising em duas dimensões não é garantia que servirá para outros modelos. Do ponto de vista de um programa legível, convém escrevé-lo em módulos. Não buscamos velocidade mas sim facilidade de leitura. Os módulos necessários são descritos a seguir:

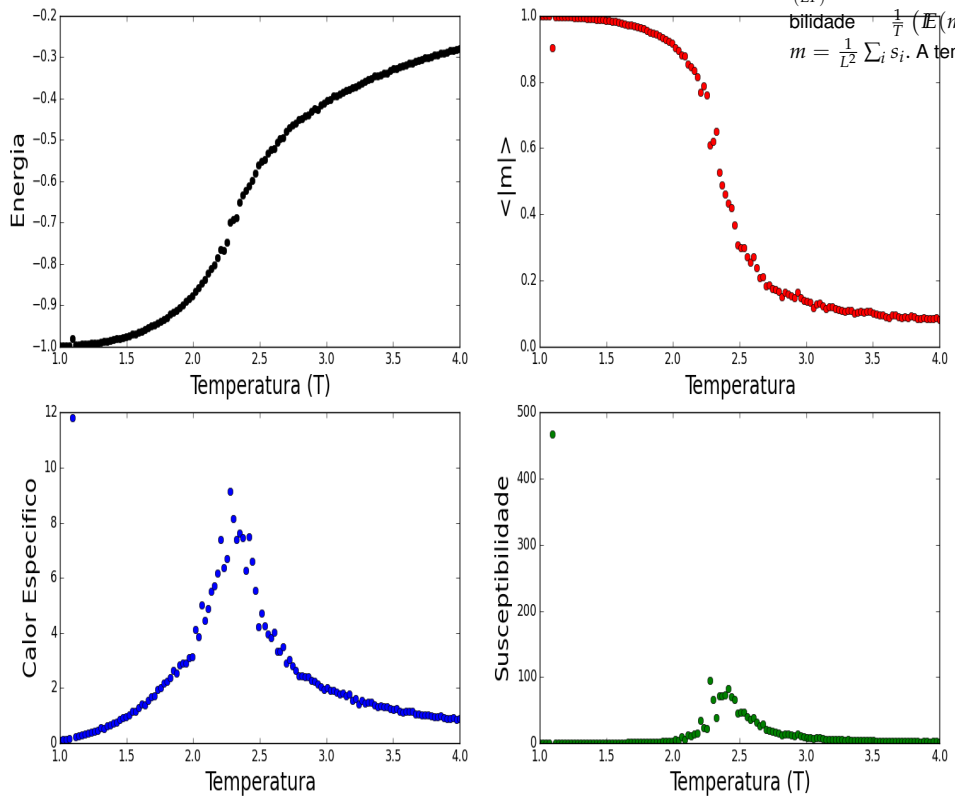


Figura 9.1: Ising 2d em rede $L \times L, L = 20$. Energia: $\frac{1}{L^2}\mathbb{E}(\mathcal{H})$, magnetização E(|m|), Calor Específico: $\frac{1}{(LT)^2}\left(\mathbb{E}(\mathcal{H}^2) - \mathbb{E}(\mathcal{H})^2\right)$, Susceptibilidade $\frac{1}{T}\left(\mathbb{E}(m^2) - \mathbb{E}(m)^2\right)$, com $m = \frac{1}{L^2}\sum_i s_i$. A temperatura é $T = 0.5$

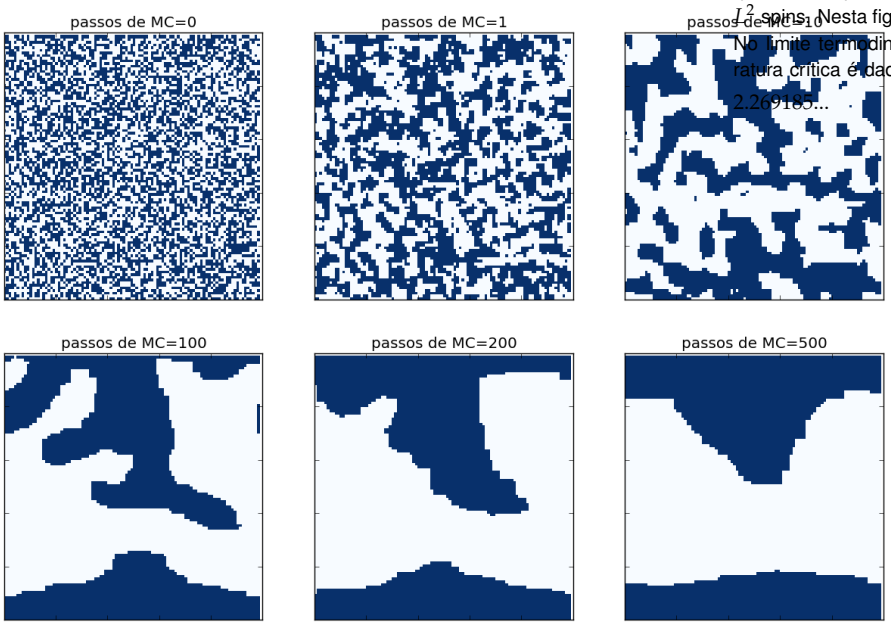


Figura 9.2: A matriz `sigma` em diferentes tempos. Unidade de tempo é 1 passo MC, que é dado pela tentativa de mudança de L^2 spins. Nesta figura $L = 100$ e $T = 0.50$. No limite termodinâmico, $L \to \infty$, temperatura crítica é dada por $T = \frac{2}{\log(1+\sqrt{2})} \approx 2.269185...$

Figura 9.3: O mesmo que na figura anteriorà temperatura é $T = 2.26 \approx T_c$

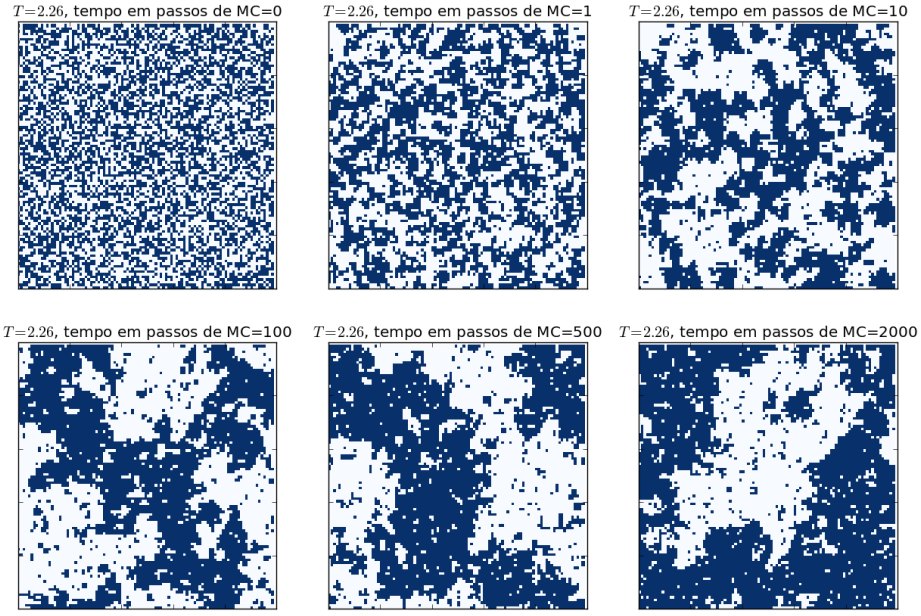
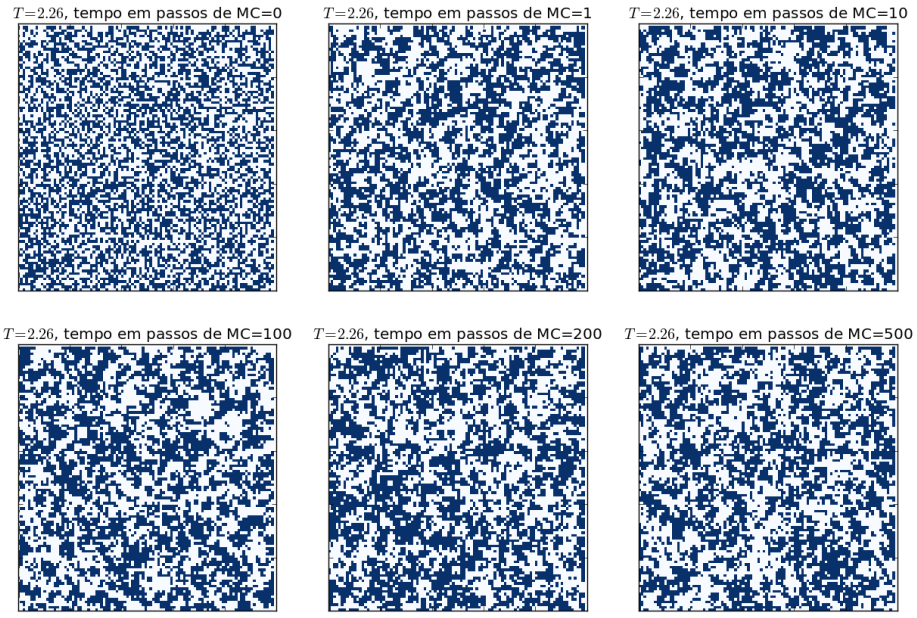


Figura 9.4: O mesmo que na figura anterior à temperatura é $T = 3.5$



Pseudo código

[6] A configuração do sistema é guardada numa matriz `sigma` de dimensões `L`$\times$ `L`. Isto é `sigma(Ix,Iy)=` $\pm$ `1`.

[6] Pode pegar o programa em http://rajeshrinet.github.io/blog/2014/ising-model/

A função `Controle(param)` controla que função é executada e com que paramêtros: `param=(L,Nterm,NMC,deltaNMC, interT)`

A rede tem tamanho L^2, `Nterm, NMC*deltaNMC` são respectivamente o número de configurações geradas para permitir termalização e o número de configurações que serão geradas na fase de medidas. As medidas são feitas a cada `deltaNMC` configurações. Portanto o número de configurações que são medidas é `NMC`.

`interT` determina o conjunto de temperaturas em que o sistema será simulado.

A função `Novaconf(sigma, param)` recebe uma configuração e devolve outra. Este é o coração do algoritmo. Abaixo mostramos como exemplo uma implementação do algoritmo de Metropolis.

A função `Acumula(sigma,f1,f2,...fl)` extrai informação de `config` e a acumula nos diversos `fs`. O objetivo da simulação é estimar quantidades de interesse termodinâmico através dos diferentes `fs`, por exemplo a energia, a magnetização, o calor específico, a susceptibilidade magnética e correlações, espaciais e temporais.

A função `GravaResultados` faz pequenas operações, como dividir pelo número de configurações, gravar os resultados. A função `Grafico` produz os gráficos para acabar de forma satisfatória. Fazer os gráficos é opcional, podem ser feitos em outro ambiente.

- `def Controle(param):`
  - para todo `1` $\leq$ `i,j` $\leq$`L : sigma(i,j) =`$\pm$ `1` com probabilidade meio
  - chama `Novaconf(sigma, param)` `Nterm` vezes para termalizar;
  - faz `NMC` vezes # coração do programa
    - chama `Novaconf(sigma, param)` `deltaNMC` vezes
    - chama `Acumula(sigma,f1,f2,...fl)`
  - chama `GravaResultados`.

A rotina que realiza a mudança de configuração

- `def Novaconf(sigma, param):`
  - para todo `1` $\leq$ `i,j` $\leq$`L` :
    - escolhe duas coordenadas `k,l` com probabilidade uniforme, e soma seus quatro vizinho `hkl=`$\sum_{\text{vizinhos}}$ `sigma (v(k),v(l))` . A definição de o que é um vizinho na borda da rede impõe a escolha de condições de contorno, por exemplo `v(k)` toma valores `(k`$\pm$`1)mod L`, para condições periódicas de contorno.
    - A mudança de energia ao tentar inverter o spin `sigma(k,l)` é `deltaE =-2hkl*sigma(k,l)`
    - Se `deltaE` $\leq$`0` ,
      - a nova configuração tera `sigma(k,l)=-sigma(k,l)`
    - `else` escolhe `aleat` aleatório uniforme entre `0` e `1`.
      - se `exp(-deltaE/T)` $\geq$ `aleat` então `sigma(k,l)=-sigma(k,l)`

O programa é de forma esquemática

- `param = L,Nterm,NMC,deltaNMC, interT`; inicializa parâmetros.
- para `t` no conjunto de temperaturas `interT`
- `Controle(param)`
- `Grafico`

10

Entropia

> I wish to propose for the reader's favourable consideration a doctrine which may, I fear, appear wildly paradoxical and subversive. The doctrine in question is this: that it is undesirable to believe in a proposition when there is no ground whatever for supposing it true.
>
> Bertrand Russell, in *Sceptical Essays*

É inquestionável a capacidade de Russell de criar frases de efeito para começar um livro. Como discordar da frase acima? "É indesejável acreditar numa proposição quando não há nenhuma base para supó-la verdadeira". Parece óbvia, mas ele usa algumas centenas de páginas para mostrar que a doutrina é efetivamente alheia à forma de pensar e agir das pessoas em geral e por isso ele está justificado em chamá-la paradoxal e subversiva. Ainda assim, parece óbvia para quem quer uma descrição científica de alguma área de conhecimento. É interessante notar as conseqüências que ela tem. Russell as analisa do ponto de vista das relações humanas. Shannon, Jaynes e toda a escola de teoria de informação o fazem do ponto de vista das conseqüências matemáticas que essa doutrina tem quando devemos raciocinar sob a luz de informação incompleta. Esse é o tema deste capítulo.

Especificamente queremos encontrar um método para atribuir números às probabilidades satisfazendo vínculos impostos por informação dada. Este problema também está relacionado à atualização de probabilidades para novos números devido à inclusão de nova informação. Novamente vamos pensar em casos particulares, que sendo suficientemente simples permitam expressar desejos de como deveria ser a teoria. Há várias formas de expressar informação e isso terá conseqüências sobre os métodos para atualizar ou atribuir probabilidades. Começamos descrevendo algo aparentemente simples, quando a informação não faz diferença entre as diferentes possibilidades. A partir dessa primeira atribuição continuamos em direção a um método geral. O resultado será que a cada distribuição de probabilidades será atribuido um número que indica quanta informação a mais seria necessário coletar para ter informação completa. Dentre aquelas distribuições que satisfazem os vínculos da informação conhecida como verdadeira, escolhemos a que é mais ignorante e portanto faz menos suposições não necessárias. A determinação desta distribuição requer o uso de algumas técnicas matemáticas. Nesse ponto o estudante deverá consultar algum livro de cálculo ou o Apêndice sobre o uso dos multiplicadores de Lagrange no cálculo de extremos sujeitos a vínculos. Nas últimas

seções do capítulo começaremos a estudar sistemas físicos e fazer a conexão com a Termodinâmica.

Incerteza, ignorância e Informação

Leia novamente a citação inicial deste capítulo. Suponha que sob dada informação I_A tenhamos que escolher entre diferentes distribuições de probabilidade que são igualmente compatíveis com os vínculos impostos pela informação I_A. Qual escolher? Para começar consideremos que temos só duas candidatas, $P_1(x_i)$ e $P_2(x_i)$. Suponha que $P_2(x_i)$, além de satisfazer os vínculos impostos por I_A também satisfaz àqueles impostos por informação adicional I_B, mas esta informação não é dada. Repetimos: só sabemos I_A. Qual das duas escolhemos?

Um critério poderia ser: "escolha P_2, pois pode ser que I_B seja verdadeiro." E se não for? Se você souber alguma coisa sobre a validade da asserção I_B claro que é bem vindo a usá-la. Mas se não souber, porque supor que é verdade? A escolha de P_2 é equivalente a assumir que I_B é verdadeira, mas isso não queremos fazer. Portanto parece natural que a escolha caia sobre P_1. Esta é, das distribuições compatíveis com os vínculos, aquela que faz menos suposições não garantidas pela informação disponível. Ao fazer esta escolha fazemos a escolha da distribuição que representa a maior ignorância possível. A outra escolha assume como certa informação que pode não ser verdadeira. É melhor não saber do que achar que se sabe algo que está errado [1].

[1] He who knows best knows how little he knows. Thomas Jefferson

Se você ainda insiste em escolher a outra distribuição, mesmo sem poder oferecer a veracidade de I_B como argumento, será por motivos não racionais e esse método, por mais que você o ache interessante, estará fora do alcance de nosso análise. Talvez a escolha seja por que você gostaria que I_B fosse verdade. Isso é ideologia, capricho ou dogma e não nos interessa discutir agora.

O método que procuramos - seguindo Shannon - procura atribuir a cada distribuição de probabilidade $P(x_i)$ uma *medida da ignorância* que essa distribuição representa sobre a variável x.

Se soubermos, por exemplo, que $x = 1$ não temos nenhuma ignorância. Se não soubermos nada sobre x, a não ser que está é uma variável que toma valores num conjunto discreto $\{x_i\}$ de n membros, então por simetria, $P(x_i) = 1/n$. Note que não é por símetria do sistema físico, mas porque a informação que temos não distingue as diferentes possibilidades i, que são simétricas quanto às preferências. Este é o princípio da razão insuficiente de Laplace, que Boltzmann seguiu, e que Jaynes discute de forma detalhada.

Máxima Entropia: Uma vez atribuida essa medida, que é chamada de *entropia,* tomaremos a distribuição com o maior valor possível dentre aquelas que satisfazem as restrições da informação conhecida. Esta é a distribuição mais incerta ou a que faz menos hipóteses e portanto está menos arriscada a fazer hipóteses incorretas.

Suponha que temos informação, na forma de um vínculo, no seguinte problema. I_A="Uma moeda indestrutível de três lados ($s = -1, 0$ ou 1) foi jogada muitas vezes, batendo no ventilador que gira no teto, e o valor médio $\overline{s}$ desse experimento é compatível com o valor 0."

É razoável atribuir valores $P(-1) = P(1) = 0$ e $P(0) = 1$? Se eu insistir que isso é razoável, a pergunta que você fará é: "porque foram

eliminadas as possibilidades ± 1?" e ainda "o quê é que ele sabe que eu não sei?". Simplesmente dizer que eu não gosto de ± 1 não é argumento razoável. A mesma coisa pode ser dito para a escolha de outra distribuição. Porque esta e não aquela?

Como proceder? Procuramos um critério que dê a cada distribuição $P(x)$, uma medida da falta de informação para determinar o valor de x. Queremos um método geral. Em princípio não sabemos se é possível tal método. Novamente usaremos a idéia que se um método geral existe, então deve ser aplicável a casos particulares. Se tivermos um número suficientemente grande de casos pode ser que o método geral se mostre incompatível com esse grande número de casos particulares. Se não fizermos a tentativa não saberemos.

Quais são os casos particulares que devemos adotar para este programa? A medida $H[P]$, a entropia, deve satisfazer, em primeiro lugar a transitividade, pois queremos escolher uma distribuição como sendo mais preferível que outra e se P_1 é preferida ante P_2 que por sua vez é preferida ante P_3, então P_1 deverá ser preferida ante P_3. Satisfazemos isto impondo que

- (1) A cada distribuição de probabilidades p associamos um número real $H[p]$

É tão razoável que é até difícil imaginar ante que críticas deveriamos defender este 'caso particular'. $H[p]$ é uma função de todo o conjunto de valores $p = \{p_i\}$. Para variáveis X que tomam valores reais, a entropia será um funcional da densidade $p(x)$. Isto é, a cada função $p(x)$ será atribuido um número.

Pequenas mudanças na distribuição $\{p_i\}$ não devem levar a grandes mudanças em $H[P]$.

- (2) $H[p]$ deve ser contínua nos p_i.

No caso particular em que a informação é simétrica ante troca dos rótulos i, teremos $P(x_i) = 1/n$. Neste caso, por simplicidade notacional denotamos $H[1/n, 1/n, ..., 1/n] = F(n)$. Suponha que temos dois problemas. No primeiro $n = 2$ e no segundo $n = 10000$. Quanta informação falta em cada problema para determinar o valor de x? Não sabemos, mas é razoável supor que no segundo caso falta mais. Logo, a medida que buscamos deve satisfazer

- (3) $F(n)$ é uma função crescente de n.

Outro caso particular onde sabemos algo pois achamos que se não fosse assim, algo estranho estaria acontecendo , é nosso velho amigo: Não queremos ser manifestamente inconsistentes, portanto o mais consistente possível é impor:

- (4) Se há mais de uma forma de analizar uma situação, todas devem dar o mesmo resultado.

Isto ainda não é suficiente e é preciso colocar uma condição que é menos óbvia e que diz respeito às possíveis maneiras de analisar a incerteza frente a diferentes agrupamentos dos estados. Suponha um dado cúbico. Os estados possíveis do sistema, para simplificar, são rotulados pelo número de pontos na face que aponta para cima. Podemos agrupar em dois grupos e.g. de { pares ou impares }, ou de { maiores ou menores que 3.5 }, ou de { primos ou não-primos }. Outra forma poderia ser menos uniforme, e.g. de { maiores ou

menores que 2.5 } ou de muitas outras maneiras, e.g. em mais de dois grupos.

Suponha que X possa ter n valores possíveis $\{x_i\}_{k=1\ldots n}$. Estes *micro*estados são mutuamente exclusivos. Queremos atribuir as probabilidades p_k. Seja $\{m_g\}$ um conjunto de números inteiros positivos tal que $\sum_{g=1}^{N} m_g = n$ e que denotam o tamanho de agrupamentos de estados de X. Escolha o conjunto de valores $\{m_g\}$ e mantenha-o fixo durante a análise. Depois esolheremos outro conjunto e o manteremos fixo novamente. Dentro de um grupo g os estados são indexados por i. Podemos escrever

$$k(i,g) = \sum_{g'}^{g-1} m_{g'} + i$$

e fixos os $\{m_g\}$, dado g e i obtemos k. A probabilidade de que X tenha um valor x_k é dada pela regra do produto em termos das probabilidades de estar em um grupo e de ter um índice i dado que está num grupo:

$$p_k = P(k) = P(i,g|\{m_g\}) = P(i|g\{m_g\})P(g|\{m_g\})$$

e pela regra da soma, a probabilidade que esteja no grupo g, por $P(g|\{m_g\}) = \sum_{i\in g} P(i|g\{m_g\}) = P_g$ assim

$$P(i|g\{m_g\}) = \frac{P(k)}{P(g|\{m_g\})}.$$

Se for dado que está no grupo g, a probabilidade que esteja no estado i é $p(i|g) = \frac{p_{k(i,g)}}{P_g}$. A incerteza associada à variável X pode ser medida em dois passos, o primeiro passo mede a incerteza de estar em um dos agrupamentos g, chamemos H_G esta entropia e o segundo, uma vez que está em g, sobre o estado i em particular esta entropia será chamada H_g.

A última imposição, que chamaremos de

- (5) aditividade sob agrupamento,

diz que a entropia dos dois passos é aditiva e portanto será $H_G + \sum_{g=1}^{N} P_g H_g$ que é a entropia do primeiro passo mais a média das entropias da incerteza associada a cada agrupamento g.

Isto, junto com a consistência do item (3) nos dá

$$H[p_k] = H_G[P_g] + \sum_{g=1}^{N} P_g H_g[p(i|g)] \tag{10.1}$$

Que isto é suficiente para determinar a forma funcional da entropia,

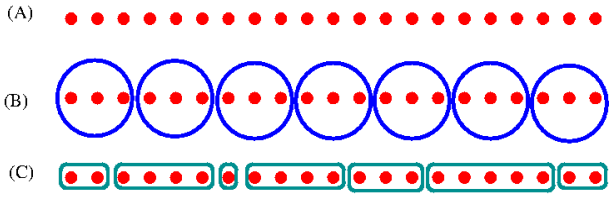


Figura 10.1: (A) Cada ponto (vermelho) representa um possível estado (mutuamente exclusivos) de uma variável. (B) Agrupamos em grupos com o mesmo tamanho. (C) Os agrupamentos são arbitrários.

será provado a seguir e será feito em dois passos, para cada um fazemos um dos dois tipos de agrupamento mencionados acima.

(I) Começamos analisando o caso particular (figura 10.1.B) em que a informação é simétrica para todos os estados i e os agrupamentos são

do mesmo tamanho $m_g = m = n/N$ para todos os N agrupamentos g. Então: $p_i = 1/n, P_g = 1/N$ e $p(i|g) = N/n = 1/m$

$$H[1/n...1/n] = H_G[1/N, ..., 1/N] + \sum_{g=1}^{N} \frac{1}{N} H_g[1/m, ...1/m] \quad (10.2)$$

como todos os termos lidam com entropias de distribuições uniformes, podemos introduzir a função monotônica desconhecida F:

$$F(n) = F(N) + \sum_{g=1}^{N} \frac{1}{N} F(m) \quad (10.3)$$

e, dado que $n = mN$, obtemos a equação funcional:

$$F(mN) = F(N) + F(m) \quad (10.4)$$

A solução óbvia é dada por

$$F(n) = k \log n. \quad (10.5)$$

Mas esta solução não é única a não ser que usemos o fato que $F(n)$ é monotônica pela imposição (2). A constante k não é muito importante neste estágio pois mede a escala das unidades, mudanças de k equivalem a mudanças na base do logarítmo que não alteram em nada a ordem de preferências de uma distribuição sobre outra. Mais para a frente veremos dentro da Mecânica Estatística, que tais mudanças equivalem a mudanças na escolha da escala de temperatura e nesse contexto poderá ser interpretada como k_B a constante de Boltzmann.

(II) Ainda não obtivemos a forma de H no caso geral. Passamos a analisar o caso em que os tamanhos dos agrupamentos m_g são arbitrários (figura 10.1.C), salvo o fato $\sum_{g=1}^{N} m_g = n$, mas ainda temos que $p_k = 1/n$ é uniforme. Temos então que a probabilidade $P_g = m_g/n$ é arbitrária, mas a probabilidade condicional dentro de cada grupo é uniforme: $p(i|g) = 1/m_g$

$$H[1/n, ..., 1/n] = H_G[P_g] + \sum_{g=1}^{N} P_g H_g[1/m_g, ..., 1/m_g] \quad (10.6)$$

e substituimos a entropia da distribuição uniforme:

$$F(n) = H_G[P_g] + \sum_{g=1}^{N} P_g F(p(i|g)), \quad (10.7)$$

então a entropia da distribuição de probabilidades arbitrária P_g é dada por

$$H_G[P_g] = F(n) - \sum_{g=1}^{N} P_g F(m_g). \quad (10.8)$$

Introduzimos um 1 através de $1 = \sum_{g=1}^{N} P_g$, substituimos $F(n) = k \log n$

$$H_G[P_g] = (\sum_{g=1}^{N} P_g) k \log n - k \sum_{g=1}^{N} P_g \log(m_g) \quad (10.9)$$

$$H_G[P_g] = -k \sum_{g=1}^{N} P_g \log(m_g/n) \quad (10.10)$$

e finalmente

$$H_G[P_g] = -k \sum_{g=1}^{N} P_g \log(P_g) \tag{10.11}$$

que é a forma da entropia de Shannon, mas pode ser chamada de entropia de Y, onde Y é um subconjunto de nomes extraídos de { Shannon, Wienner, Boltzmann, Gibbs, Jaynes }, o que torna a vida dos historiadores da ciência mais interessante.

Deve ficar claro que não há nenhuma controvérsia quanto à utilidade deste formalismo. No entanto há muita controvérsia sobre a interpretação desta fórmula em geral, da necessidade dos *axiomas*, da suficiência, de se o nome de Boltzmann e Gibbs, associado à entropia termodinâmica deveria ser associado neste ponto a esta forma. Há também discussões atuais sobre o efeito acarretado pela mudança de um ou outro dos casos particulares que usamos. Sobre se estes deveriam ser chamados de axiomas. Sobre como generalizar isto para o caso em que as variáveis x tomam valores em intervalos reais. Nesse caso a idéia de uniformidade fica relativa à escolha do sistema de coordenadas. A teoria fica muito mais interessante e a invariância antes transformações gerais de coordenadas leva as discussões para o contexto da geometria diferencial.

A contribuição de Boltzmann e Gibbs será discutida no próximo capítulo e não entraria aqui em um par de parágrafos. De Shannon e Wienner [2] vem idéia de discutir informação do ponto de vista de volume, quantidade com o objetivo de entender limitações para a transmissão em canais de comunicação. Isto inclui como codificar uma mensagem, comprimindo-a e como, no outro lado do canal de comunicação, reobté-la. Não foi discutida a qualidade da informação ou a utilidade da informação: se fosse, toda a teoria não teria utilidade para descrever por exemplo meios como a televisão.... Isto não é totalmente piada. Se olharmos uma máquina de processamento de informação, como por exemplo um sistema nervoso de um animal, o conceito do valor da informação toma uma posição muito mais central. Há vantagens evolutivas em realizar inferência de uma forma frente a outra. Mais sobre tudo isso em aulas posteriores [3].

[2] Aparentemente Alan Turing também teve estas idéias durante a 2da Guerra Mundial enquanto tentava decifrar códigos, mas esse trabalho teve menos influência porque era confidencial. Ver livro de Good.

[3] Não cabe no curso introdutório de Mecânica Estatística

Um método variacional

O caminho para as aplicações está aberto. O método de escolha das distribuições de probabilidade é de maximizar a ignorância. De todas as distribuições compatíveis com os vínculos, escolhemos a que introduz a menor quantidade de informação adicional aos vínculos. Reduzimos a inferência a um método variacional. Isto não deve vir como uma surpresa, pois toda a Física é essencialmente redutível a métodos variacionais. A pergunta profunda é se esses métodos estão relacionados ou se são independentes. A reposta que está emergindo é que todos esses métodos estão relacionados. E isto transcende a Física e se estende por outras áreas. Uma contribuição importante de Jaynes foi a percepção que a maximização da entropia sujeita aos vínculos da informação leva a um método fundamental de inferência, chamado por ele de MaxEnt. Outros nomes estão associados a extensão de idéias de entropia como método de inferência fundamental [4]. Em particular tem se conseguido nos últimos anos uma formulação muito mais satisfatória e menos ad hoc dos princípios por trás deste método.

[4] ver A.C

A generalização do método de máxima entropia para inferência em

problemas com variáveis em variedades contínuas pode levar a problemas conceituais que serão atacados a seguir.

Entropia Relativa

As deduções aqui deveriam ser lidas após treinamento sobre métodos variacionais, que será abordado nos próximos capítulos. Portanto talvez seja recomendável pular esta seção numa primeira leitura. A entropia de Shannon vai ser usada para inferências em quase todo o resto do livro. Mas esta não é a última palavra em mecanismos de inferência. A utilidade de entropia de Shannon foi a de *atribuição* de probabilidades. Isso deve dar a impressão que do nada, as probabilidades são construidas através da imposição de certos vínculos sobre funções dos graus de liberdade. Dificilmente a inferência é feita a partir de nenhum conhecimento prévio. Para qualquer situação é mais razoável que um estado de crenças seja modificado pela imposição de novos vínculos. Isso nos leva à definir Informação como tudo aquilo que leva a modificar as crenças codificadas em distribuições de probabilidades. Procuramos um mecanismo que sirva para *atualizar* as crenças a partir de uma crença *a priori*. A diferença importante pode ser vista no confronto de *atualizar* a partir de um estado de conhecimento *apriori* versus *atribuir* a partir do vácuo informacional.

Qual deve ser o mecanismo de mudar crenças? Há em geral muitas distribuições de probabilidade que satisfazem um certo conjunto de vínculos. A idéia é fazer um ranking de todas elas. Isto significa atribuir um número a cada distribuição de probabilidades e escolher aquela que maximiza o ranking de preferências. Ou seja devemos construir um funcional que atribui um número a cada distribuição. Claro que esse número deve depender no estado de crença antes da imposição dos novos vínculos. Como escolher o funcional? Novamente procuramos casos simples onde sabemos algo. Impomos como desejo que o funcional dê a resposta esperada neles. O método de obtenção do funcional é o de eliminação de todos os funcionais que falham em satisfazer esses casos simples. Resta algum funcional após este processo? Poderia ser que não, mas não é o caso. Chamaremos este funcional de Entropia Relativa $S[p||q]$, entre a distribuição p candidata a ser escolhida pelo processode inferêcnia e a distribuição *a priori* q. Também pode ser chamado de (menos) a divergência de Kullback-Leibler (KL). Por economia, será simplesmente chamada entropia. A menos de pequenos detalhes, a dedução segue A. Caticha de forma simplificada.

Desiderata Entrópica

Os desejos expostos a seguir são considerados em adição aos do Capítulo 1. Portanto a teoria de probabilidades é a estrutura matemática adequada. Mas agora faremos algumas demandas adicionais. Pedimos ao leitor que encontre argumentos contra os seguintes desejos para uma teoria de infêrencia:

- DE_1: **Generalidade**: O método deve ser Universal.

Queremos lidar com informação incompleta e o algoritmo de procedimento deve ser o mesmo independente de qual o tipo de problema de inferência que estamos tratando. A descrição do

problema, através da escolha dos graus de liberdade apropriados e dos vínculos impostos levará em conta a natureza explícita do problema.

- DE_2: **Parcimônia** A atualização deve ser minimalista.

De todas as distribuições $p(x)$ que satisfazem o vínculo devemos escolher a que menos mude a distribuição *a priori* $q(x)$. Claro que o que quer dizer mínimo deve ser definido. Mas há um caso simples em que mínimo é fácil de decidir. No caso em que não há nova informação, então a menor mudança é não fazer nada. Nosso método deve ser tal que $p(x) = q(x)$ após a incorporação de nenhuma informação. Parece muito trivial mas será útil.

- DE_3: **Localidade** no espaço de configurações.

Quando o espaço de configurações pode ser dividido em duas partes disjuntas, informação que diz respeito somente a uma das partes, ao ser incorporada, não deve alterar a atribuição de probabilidades às configurações da outra.

- DE_4: **Invariância**: A escolha dos rótulos dos graus de liberdade é uma convenção: não deve alterar o resultado da Inferência.

Para sistemas de variáveis que tomam valores num sub espaço de $\mathbb{R}^N$, a escolha do sistema de coordenadas não deve interferir no resultado do problema. Assim o funcional deve ser invariante ante mudanças (contínuas diferenciáveis) do sistema de coordenadas.

- DE_5: **Independência**: Sistemas independentes devem ser descritos da mesma forma quando estudados separadamente ou em conjunto.

Há situações em que sabemos que há sistemas independentes. Consideremos uma garrafa de café e a galáxia *X*9. Ao fazer predições sobre a termodinâmica do café podemos (i) tratar o sistema de café sozinho, ou (ii) incluir os graus de liberdade da galáxia *X*9 além dos do café, e marginalizar a distribuição, integrando sobre as configurações da galáxia. Os dois métodos devem dar o mesmo resultado sob pena de ao realizar medidas experimentais, poder distinguir entre os casos em que a galáxia foi ou não incluida. Note que poderiamos, caso dessem diferentes, fazer astronomia olhando para uma garrafa de café.

DE_3 : *Localidade*

Começamos com as consequências de DE_3. As configurações de um sistema formam o conjunto $\chi = \{x_1, x_2, ...x_n\}$. Podemos separar em dois conjuntos $i \in A$ e $j \in B$, tal que $A \wedge B = \emptyset$ e $A \vee B = \chi$, ou seja mutuamente exclusivos e exaustivos. Um caso simples é quando a informação é dada através dos seguintes vínculos

$$P_A = \sum_{x_i \in A} p_i, \tag{10.12}$$

$$P_B = \sum_{x_j \in B} p_j, \tag{10.13}$$

$$\text{com} \quad P_A + P_B = 1, \tag{10.14}$$

$$G = \sum_{x_j \in B} p_j g(x_j), \tag{10.15}$$

onde $p_k = p(x_k)$. Maximizar a entropia sujeita a estes vínculos leva a

$$0 = \delta \left[S - \lambda_1 \left(\sum_A p(x_i) - P_A \right) - \lambda_2 \left(\sum_B p(x_j) - P_B \right) - \lambda_3 \left(\sum_B p(x_j) g(x_j) - G \right) \right]$$

para variações independentes de quaisquer dos p. Temos n equações:

$$\frac{\partial S}{\partial p_i} = \lambda_1 \tag{10.16}$$

$$\frac{\partial S}{\partial p_j} = \lambda_2 + \lambda_3 g(x_j). \tag{10.17}$$

$$\tag{10.18}$$

Em geral teriamos, para $k \in A$ ou $k \in B$, chamando $\frac{\partial S}{\partial p_k} = f_k$, a depêndencia

$$\frac{\partial S}{\partial p_k} = f_k(p_1 ..... p_n, q_1 .... q_n). \tag{10.19}$$

Comparando as equações 10.16 e 10.19 vemos que

$$\lambda_1 = f_i(p_1 ..... p_n, q_1 .... q_n) \tag{10.20}$$

$$= f_i(p_i, q_i). \tag{10.21}$$

A segunda equação decorre de que mudando a informação que altera p_j do conjunto B, poderiamos mudar o lado direito sem alterar o multiplicador de Lagrange que é constante. Ainda mais, podemos considerar todas as possíveis divisões de χ em subconjuntos exaustivos e mutuamente exclusivos. Eliminariamos a dependência em todos os indices diferentes de i. Assim

$$\frac{\partial S}{\partial p_k} = f_k(p_k, q_k) \tag{10.22}$$

Integrando, vemos que a forma geral do funcional é

$$S = \sum F_k(p_k, q_k)$$

onde F_k com $f_k = \frac{\partial F_k}{\partial p_k}$ são funções ainda desconhecidas. A generalização para variáveis que tomam valores reais:

$$S = \int F(p(x), q(x), x) dx. \tag{10.23}$$

DE_4: Invariância

Vamos começar com um problema em uma dimensão. Uma densidade de probabilidades de uma variável X, $m(x)$ satisfaz

$$\int m(x) dx = 1$$

Podemos mudar variáveis $x \to x'$ e a densidade mudará $m(x) \to m'(x')$. A probabilidade de uma região dx em x deve ser preservada, portanto

$$m(x)dx = m'(x)dx'$$

e isso deve valer para a transformação de qualquer densidade. Em mais de uma dimensão teremos

$$m(x) = m'(x')J(x')$$

onde $J(x') = \det |\frac{\partial x'}{\partial x}|$ é o Jacobiano da transformação $x \to x'$.

A estratégia a seguir é a seguinte: fazemos a inferência no sistema de coordenadas x, obtendo um resultado. A seguir, a fazemos no referencial x'. Transformamos os x' de volta a x e comparamos. Os dois resultados deveriam dar igual. Mas não ocorre a não ser que imponhamos condições sobre a função F que aparece na equação 10.23. Isso restringe ainda mais os funcionais sobreviventes. Antes de prosseguir, facilita introduzir uma densidade $m(x)$ que por agora está à nossa disposição escolher. A motivação é que a razão de duas densidades tem propriedades simples de transformação

$$\frac{p(x)}{m(x)} = \frac{p'(x')}{m'(x')}, \tag{10.24}$$

pois o Jacobiano se cancela [5]. Assim consideramos que a equação 10.23 pode ser escrita

[5] Cabe ressaltar que deveriamos ter cuidado de afirmar que as regiões em que as diferentes distribuições se anulam são as mesmas e são excluidas destas considerações.

$$\begin{aligned} S &= \int F(p(x), q(x), x)dx = \int \frac{1}{m(x)} F(\frac{p(x)}{m(x)}m(x), \frac{q(x)}{m(x)}m(x), x)m(x)dx, \\ &= \int \Phi(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x)m(x)dx, \end{aligned} \tag{10.25}$$

onde introduzimos a ainda desconhecida função

$$\Phi(u, v, m, x) = \frac{1}{m} F(um, vm, x).$$

Queremos mudar as coordenadas de forma geral, mas é conveniente se restringir a um caso fácil. O vínculo imposto será

$$\int p(x)a(x) = A,$$

onde a função $a(x)$ é um escalar, o que significa que $a(x) \to a'(x') = a(x)$, não muda ante transformações de coordenadas. A normalização é um exemplo de vínculo escalar, portanto não é preciso considerá-lo separadamente. O cálculo variacional δ (S +vínculos) $= 0$, leva a

$$\begin{aligned} 0 &= \frac{\delta}{\delta p(x)} \left(S - \lambda \int p(x)a(x)dx \right), \\ &= \dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x) - \lambda a(x), \end{aligned} \tag{10.26}$$

onde

$$\dot{\Phi}(u, v, m, x) = \frac{d}{du}\Phi(u, v, m, x).$$

Analogamente, começando no sistema de coordenadas x', chegamos a

$$0 = \dot{\Phi}(\frac{p'(x')}{m'(x')}, \frac{q'(x')}{m'(x')}, m(x'), x') - \lambda' a'(x'). \tag{10.27}$$

Usando a equação 10.24, esta última equação pode ser reescrita como

$$0 = \dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m'(x'), x') - \lambda' a'(x'). \tag{10.28}$$

Para vínculos escalares, segue que, a razão

$$\frac{\dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x)}{\dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m'(x'), x')} = \frac{\lambda}{\lambda'} \tag{10.29}$$

é uma constante, pois os multiplicadores de Lagrange não dependem de x nem x'. Como isto vale para qualquer transformação contínua e diferenciável de $x' = \phi(x)$, podemos olhar para diferentes escolhas de

ϕ e eliminar candidatos. Primeiro, olhamos para transformações com Jacobiano igual a 1 em todo o espaço. Segue que $m(x) = m'(x')$, portanto a única diferença entre o numerador e o denominador é no quarto argumento. Como a razão é constante, $\dot{\Phi}$ não pode depender do quarto argumento x ou x'. Agora consideramos, novamente com Jacobiano igual a 1, a função $\phi(x) = x, x \in \bar{D}$ e $\phi(x) \neq x, x \in D$. Se $x \in \bar{D}$, a razão é 1:

$$\begin{aligned} 1 &= \frac{\dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x))}{\dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m'(x'))} \\ &= \frac{\lambda}{\lambda'} \end{aligned}$$

Mas para $x \in D$, não muda a razão. Portanto $\dot{\Phi}$ não pode depender agora do seu terceiro argumento e

$$1 = \frac{\dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)})}{\dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)})} \tag{10.30}$$

é trivialmente satisfeita. Segue que o funcional só pode ter esta estrutura:

$$S = \int \Phi(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}) m(x) dx. \tag{10.31}$$

DE_2 : *Parcimônia*

Temos à nossa disposição este intruso $m(x)$. Podemos escolher qualquer densidade? Impondo o desejo de parcimônia veremos que não é possível essa liberdade. Estamos interessados em não fazer nada. O aluno não deve se entusiasmar desnecessariamente. Se não chega informação, não deveria mudar nossa escolha de $p(x)$. Portanto a solução do problema variacional

$$\begin{aligned} \frac{\delta}{\delta p(x)} \left(S - \lambda \int p(x) dx \right) &= 0, \\ \dot{\Phi}(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}) &= \lambda, \end{aligned} \tag{10.32}$$

deve levar à escolha $p(x) = q(x)$. Mas nem todo funcional do tipo 10.31 leva a esse resultado. Isto restringe mais a familia de funcionais. O problema 10.32 é uma equação diferencial ordinária

$$\begin{aligned} \dot{\Phi}(u, v) &= \lambda \\ \Phi(u, v) &= \lambda u + c(v) \end{aligned}$$

portanto, se seguissemos este caminho teriamos

$$\begin{aligned} S &= \int \left(\lambda \frac{p(x)}{m(x)} + c(\frac{q(x)}{m(x)}) \right) m(x) dx \\ &= \lambda \int p(x) dx + \text{const}, \end{aligned}$$

o que é uma trivialidade, pois a normalização diz que isto é uma constante e toda distribuição normalizada teria a mesma preferência. Portanto escolher a função Φ leva a um beco sem saida. Mas há outra saida para satisfazer a equação 10.32, basta escolher $m(x) = q(x)$, neste caso

$$\dot{\Phi}(\frac{p(x)}{q(x)}, \frac{q(x)}{q(x)}) = \lambda, \tag{10.33}$$

que é constante de forma trivial quando $p(x) = q(x)$. Novamente temos um avanço, restringindo ainda mais as estruturas possíveis:

$$S = \int \Phi(\frac{p(x)}{q(x)})q(x)dx. \tag{10.34}$$

DE_5 : *Independência*

Este último desejo será usado para determinar a forma explícita de Φ na equação 10.34. Além da normalização, para o sistema composto pela garrafa de café, com graus de liberdade x_1 e a galáxia $X9$ com graus de liberdade x_2, temos os seguintes vínculos

$$\begin{aligned} \int p_1(x_1)f_1(x_1)dx_1 &= F_1 \\ \int p_2(x_2)f_2(x_2)dx_2 &= F_2. \end{aligned}$$

Tratando os dois sistemas de forma independente temos que $p_1(x_1)$ e $p_2(x_2)$ são dados pelas soluções de

$$\dot{\Phi}(\frac{p_1(x_1)}{q_1(x_1)}) = \lambda_{01} + \lambda_1 f_1(x_1) \text{ e } \dot{\Phi}(\frac{p(x_2)}{q_2(x_2)}) = \lambda_{02} + \lambda_2 f(x_2) \tag{10.35}$$

Tratando os sistemas juntos, o problema variacional é

$$\begin{aligned} 0 &= \frac{\delta}{\delta p(x_1,x_2)} \left\{ S - \mu_0 \left(\int p(x_1,x_2)dx - 1 \right) \right. \\ &- \mu_1 \left(\int f_1(x_1)p(x_1,x_2)dx_1dx_2 - F_1 \right) \\ &- \left. \mu_2 \left(\int f_2(x_2)p(x_1,x_2)dx_1dx_2 - F_2 \right) \right\}, \end{aligned}$$

$$\dot{\Phi}(\frac{p(x_1,x_2)}{q_1(x_1)q_2(x_2)}) = \mu_0 + \mu_1 f_1(x_1) + \mu_2 f_2(x_2) \tag{10.36}$$

Impomos que a solução deve satisfazer $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ e definimos $y = \frac{p_1(x_1)p_2(x_2)}{q_1(x_1)q_2(x_2)}$. Assim

$$\dot{\Phi}(y) = \mu_0 + \mu_1 f_1(x_1) + \mu_2 f_2(x_2). \tag{10.37}$$

Derivamos duas vezes, primeiro com respeito a x_2

$$\ddot{\Phi}(y)\frac{\partial y}{\partial x_2} = (\mu_2 f_2(x_2))', \tag{10.38}$$

e depois com respeito a x_1

$$\dddot{\Phi}(y)\frac{\partial y}{\partial x_1}\frac{\partial y}{\partial x_2} + \ddot{\Phi}(y)\frac{\partial^2 y}{\partial x_1 x_2} = 0. \tag{10.39}$$

Isto parece complicado mas é bem simples. Note que, com a linha indicando a derivada com respeito a x_1 ou x_2

$$\begin{aligned} \frac{\partial y}{\partial x_1} &= (\frac{p_1(x_1)}{q_1(x_1)})'\frac{p_2(x_2)}{q_2(x_2)} \\ \frac{\partial y}{\partial x_2} &= \frac{p_1(x_1)}{q_1(x_1)}(\frac{p_2(x_2)}{q_2(x_2)})' \\ \frac{\partial^2 y}{\partial x_1 x_2} &= (\frac{p_1(x_1)}{q_1(x_1)})'(\frac{p_2(x_2)}{q_2(x_2)})'. \end{aligned} \tag{10.40}$$

Multiplique as duas primeiras e divida pela terceira equação, segue que $\Phi(y)$ satisfaz a equação diferencial ordinária:

$$y\dddot{\Phi}(y) + \ddot{\Phi}(y) = 0. \tag{10.41}$$

A solução geral contém três constantes arbitrárias

$$\Phi(y) = Ay\log y + By + C. \tag{10.42}$$

A familia de funcionais entropia é reduzida a

$$S[p||q] = A\int p(x)\log\frac{p(x)}{q(x)}dx + B\int p(x)dx + C\int q(x)dx.$$

A constante C pode ser tomada igual a zero, pois simplesmente adiciona uma constante sem mudar preferências. A constante B também pode ser tomada zero, pois a normalização de $p(x)$ será sempre um vínculo. Finalmente e demonstrando otimismo os fisicos tomam $A < 0$ negativo para que o processo variacional seja o de maximização. Podemos tomar $A > 0$ e minimizar o que é chamado de divergência de Kullback-Leibler. A escolha de $A = -1$ é simplesmente uma escolha de unidades da entropia. Como veremos isto levará a que a temperatura e a energia sejam medidos nas mesmas unidades. O resultado final é dado por

$$S[p||q] = -\int p(x)\log\frac{p(x)}{q(x)}dx \tag{10.43}$$

É facil mostrar que os resultados $p_1(x_1)p_2(x_2)$ usando a equação 10.35 e $p(x_1, x_2)$ da equação 10.36, coincidem.

Exercício Para variáveis discretas mostre que $H_{SJ} = -\sum_i p_i \log p_i$ difere por uma constante de $S[p||q] = -\sum_i p_i \log\frac{p_i}{q_i}$, se $q_i = 1/N$ é constante e portanto o valor das distribuições que as extremizam coincidem.

Isto permite concluir que a entropia de Shannon faz referência a uma ditribuição *a priori*, a distribuição uniforme. O fato de não aparentar de forma explícita não quer dizer que não esteja lá. É claro que o ponto onde a distribuição uniforme faz a sua entrada no raciocínio é quando a função $F(n)$ crescente é introduzida.

Exercício Suponha $\mathcal{F}$ uma família paramétrica de densidades, e dois membros $p_1 = p(x|\theta_1)$ e $p_2 = p(x|\theta_2)$ de $\mathcal{F}$. Suponha que os parâmetros θ estejam em algum subconjunto de $\mathbb{R}^K$. Mostre que em geral $S[p_1||p_2] \neq S[p_2||p_1]$, mas para valores pequenos da distância Euclideana, $|\theta_1 - \theta_2|$ há simetria de S ante $\theta_1 \leftrightarrow \theta_2$. Isso permite definir uma distância entre duas densidades e uma geometria Riemanniana no espaço dos θ, que agora passa a ser mais que um simples subconjunto de $\mathbb{R}^K$, mas o que é conhecido como uma variedade estatística. Considere $\theta_2 = \theta_1 + d\theta$ e expanda a entropia até segunda ordem em $d\theta$. Dessa forma é possível definir o tensor métrico $g_{\mu\nu}$. Encontre uma expressão para o tensor métrico que é conhecido pelo nome de Fisher-Rao. Encontre a distância entre duas gaussianas multivariada com a mesma covariância e médias diferentes.

Apêndice: Multiplicadores de Lagrange

Uma montanha é descrita pela altura $z = f(x, y)$, com x e y as coordenadas de um ponto de altura z e f um função suficientemente bem comportada, tem um máximo que pode ser obtido igulando as derivadas parciais de f a zero. Suponha que uma estrada cruza a superfície da montanha e as coordenadas da estrada são descritas por $\phi(x, y) = c$. Qual é a altura máxima de um carro que viaja pela estrada?

O problema de encontrar pontos extremos ou só estacionários de funções sujeitos a vínculos é muito vasto. Damos algumas idéias básicas sem preocupação com rigor, para lembrar o estudante de técnicas que deveriam ser vistas em Cálculo 2 ou curso equivalente.

Seja o problema

- P_{livre}: Queremos encontrar um ponto (x^*, y^*) dentro de uma certa região C no plano real onde uma função $f(x, y)$ tem localmente um valor estacionário.

Fácil, tome as derivadas parciais e resolva o sistema $\partial_x f = 0$, $\partial_y f = 0$.

Queremos, a seguir resolver um caso mais difícil.

- P_{vinc}: Suponha que não procuramos o resultado em qualquer lugar de C, mas especificamente queremos um ponto estacionário entre aqueles pontos que satisfazem $\phi(x, y) = c$, que supomos descreva uma curva no plano que está parcialmente contida em C e chamaremos γ.

A solução do parágrafo anterior dificilmente nos dá a resposta pois seria uma coincidência se (x^*, y^*) caisse encima dessa curva.

A solução a esta classe de problema foi proposta por Lagrange. Considere a classe de funções $F_\lambda(x, y)$ que dependem do chamado multiplicador de Lagrange λ:

$$F_\lambda(x, y) = f(x, y) + \lambda(\phi(x, y) - c) \tag{10.44}$$

Note que se o ponto de coordenadas x e y estiver na curva γ então F_λ e f tem o mesmo valor. Repetindo: F_λ e f tem o mesmo valor se o vínculo que " x e y estão sobre γ" for respeitado.

Consideremos o P_{livre} mas para a função $F_\lambda(x, y)$. O problema é novamente simples [6]. Resolvemos o sistema

[6] a não ser que não seja....

$$\frac{\partial F_\lambda}{\partial x} = 0 \tag{10.45}$$

$$\frac{\partial F_\lambda}{\partial y} = 0, \tag{10.46}$$

onde λ ainda não foi especificado. A resposta depende do valor escolhido para λ, isto é define uma curva ρ, parametrizada por λ : $(x^*(\lambda), y^*(\lambda))$ onde F_λ é extremo . Agora voltamos ao problema P_{vinc}. Da resposta à dupla pergunta "onde f é máximo?" e "onde o vínculo é satisfeito?" , quando as duas são respondidas simultaneamente, decorre a solução. Substituimos a primeira por "onde F é máximo?" (resposta: em ρ) junto com a afirmação "$f = F_\lambda$ sob a condição de estar em γ". Segue que queremos encontrar o cruzamento de γ e ρ. Basta escolhermos $\lambda = \lambda_*$ tal que $\phi\left(x^*(\lambda_*), y^*(\lambda_*)\right) = c$, o resultado é um extremo para f qunado se satisfaz o vínculo.

Agora procure um livro de cálculo e prencha os detalhes. Discuta também como lidar com casos em que o extremo está na borda de C.

Há vínculos que são representados por desigualdades, os nomes de Kuhn e Tucker estão associados a esta extensão. Em muitos casos isto pode útil mas não no curso introdutório.