NESTOR CATICHA

PROBABILIDADES (NOTAS INCOMPLETAS)

Prólogo

Estas notas são complementares ao curso introdutório Probabilidades 04300xxx.

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 o objetivo do curso, fazer o aluno pensar e talvez modificar suas idéias sobre o que significa probabilidade. Rigor matemático não substitui rigor intelectual.

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 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 e Ciência Cognitiva 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. O objetivo destas notas é apresentar 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 a parte formal de Probabilidades como uma parte da Matemática, um ramo da análise funcional e teoria de medida não será explorada.

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 outras 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.

Sumário

1 Teoremas de Regraduação de Cox **7**

2 Outras definições de probabilidade **31**

3 Uso elementar de Probabilidades **37**

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 deviaser 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,

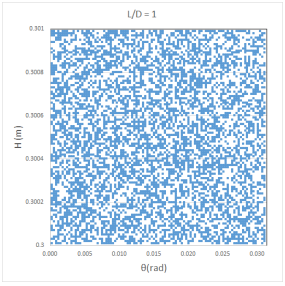


Figura 1.1: Integração numérica das equações de movimento de um modelo Newtoniano de uma "moeda"feita de massas (m) e molas (k). A figura mostra um espaço restrito de condições iniciais. H é a altura da moeda ao ser lançada e θ o ângulo com a horizontal, a moeda é solta do repouso. Nesta figura a altura 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.

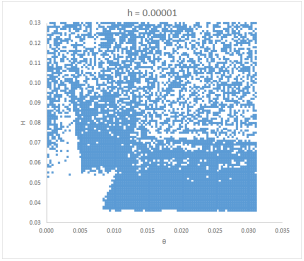


Figura 1.2: Igual à anterior, mas a moeda é solta de uma altura menor, para diferentes ângulos.

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

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

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. 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 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.

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 creer verdadeira. Mas quando se diz " a massa de Saturno está entre m_1 e m_2" ou "... entre m_3 e m_4" estamos usando asserções diferentes e a tarefa é determinar quanto acreditamos que uma ou a outra sejam verdade e aqui o estudante reconhece a linguagem científica.

[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 ferra-

menta 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 $C =$ "$A \to B$", isto é, A implica B. Traduzindo, isto significa "se souber que A é certamente verdade, segue que a proposição B também o é." Dado isso, o que podemos dizer sobre B? Nada com certeza, mas se também recebemos a informação adicional A, isto é, que "A é Verdade", então segue B, ou seja "B é Verdade".

Outro caso de informação completa, 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 $C =$ "$A \to B$" e "A é Falso", o quê pode ser concluido? Do ponto de vista lógico clássico nada podemos concluir sobre B. Da mesma forma se for dada a informação "B é Verdade", nada podemos concluir sobre A. Estamos frente a casos de informação incompleta e a lógica clássica não serve para chegar a uma conclusão. Não é possível deduzir nada. A indução, 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. 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*[7] com a informação e que corremos riscos de ser inconsistentes caso a regras de manipulação de probabilidades não sejam seguidas. Segue que **não há probabilidades que não sejam condicionais**, embora às vezes simplesmente a linguagem esqueça de deixar explícitas as relações de condicionalidade [8]. A amplidão da aplicabilidade da teoria que emerge é impressionante e por exemplo, quando o tipo de asserção for limitado àqueles entendidos em teoria de conjuntos as regras de manipulação serão não mais nem menos que aquelas ditadas pelos axiomas de Kolmogorov. 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 [9]. Eu seguirei a apresentação de A. Caticha, que é mais completa e clara, mas farei algumas pequenas mudanças[10].

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

[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*.

[7] 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.

[8] A maior fonte de erros será devido 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.

[9] Tribus, A. C

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

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

Em primeiro lugar queremos falar sobre uma asserção A no caso de informação incompleta. Nos referimos então à 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.

[11] 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.

DP_1 *Representação de crenças e transitividade*

Queremos analisar o primeiro caso simples que lida com o conceito de *mais plausível*. Se A dado B é mais plausível do que A dado C escrevemos $A|B \succ A|C$. Suponha ainda que $A|C \succ A|D$. 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 $C = \overline{A \wedge \overline{B}}$.

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 $\overline{a}$) [12], 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

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

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.

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 [13] 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 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. [14] 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:

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

[13] 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,

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

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

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

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:

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

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:

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

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

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

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

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

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

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

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

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

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:

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

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

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

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 $\phi(a(b \vee c)|d)$:

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

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 $G(x,w) = A(x)w = G(x,1)w$ [15]. 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

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

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

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

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 $c = e \wedge \overline{h}$ [16]. Portanto não ha restrições para o resultado que obtivemos.
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

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

o que permite regraduar mais uma vez os números associados as crenças sem mudar sua ordem. Crenças regraduadas, de forma

[15] 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 $h(x) = xh(1)$.

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

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? [17]

[17] 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. [18] Introduza a asserção

[18] 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.htm

> "... 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 uma coleção de elementos $A, B, C....$ que são eventos elementares [2] 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 campo se a soma, produto, interseção de dois elementos quaisquer pertencem ao sistema. Os axiomas de Kolmogorov para a teoria de Proabilidades são

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

- AK1) $\mathcal{F}$ é um campo de conjuntos 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.

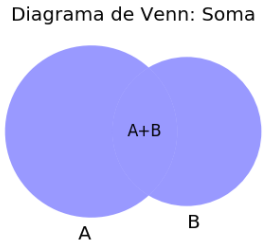
[3] Talvez a queixa é que as provas do capítulo 1 são para número finito de conjunções e disjunções. Isto porém não deve ser motivo de preocupação agora pois não é um empecilho irremovível.

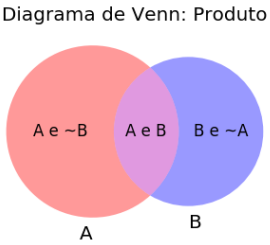
A probabilidade da *certeza* é 1 por AK4; a probabilidade está entre zero e um e a probabilidade da disjunção de asserções que não tem elementos em comum é a soma das probabilidades. Notamos, na introdução aos axiomas no livro de Kolmogorov, porém a falta de uma regra para o produto. Kolmogorov não a 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 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,

[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.

[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 $(M_0, M_0 + \Delta m)$

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].

[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 idênticas 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 $P(\boldsymbol{R}_1 = \boldsymbol{r}_1$ 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]
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*:

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

Obviamente $\text{Var}(X) \geq 0$. É fácil mostrar que $\text{Var}(X) = \langle x^2 \rangle - \langle x \rangle^2$. 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:

[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

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

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 $\{b_i\}_{i=1,\ldots N}$ 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 3.7 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

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

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

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

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

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

e o segundo momento

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

Portanto a variância

$$\sigma^2_{Poisson} = \lambda \tag{3.15}$$

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 **??**. 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:

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

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

Gaussiana ou Normal

Dedicaremos o capítulo **??** 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

$$I\!E(x) = a, \quad I\!E(x^2) - I\!E(x)^2 = a^2$$

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

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

por simetria, logo

$$I\!E(x) = x_0,$$

que coincide com a moda e a mediana. O interessante é que $I\!E((x-x_0)^2)) = \infty$ 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

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

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

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

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

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

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 **??** e **??** 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

[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.

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.

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(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

[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.

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.

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:

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

Para uma dada sequência usamos a regra do produto

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

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

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

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 **??**, 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 **??** é satisfeita. Estes coeficientes são chamados os coeficientes binomiais. O motivo disto é que

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

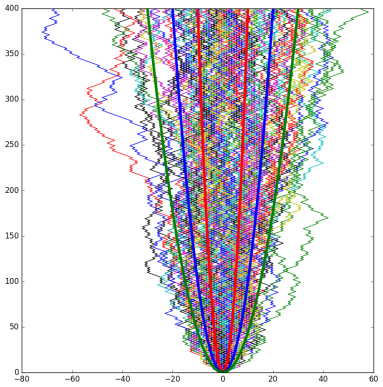
que é amplamente conhecida desde Newton. Mas é instrutivo provar este resultado, supondo-o válido para $R-1$, calculando

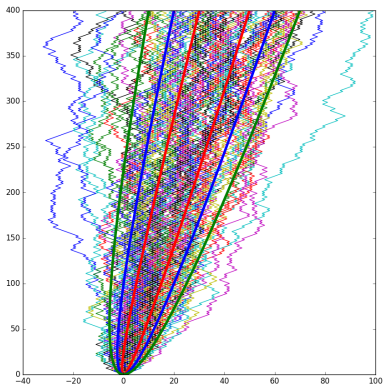
$(a+b)^{R-1}(a+b)$ e usando a relação de recorrência. A notação $C_R^m = \binom{R}{m}$ também é muito popular e é dito que representa o número de maneiras de escolher m elementos de um total de R ou o número de combinações de R, m a m.
Temos o resultado desejado,

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

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

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





difusão I: Ca-
iais, $N = 100$
0 passos cada
tante de uma
n caminhante
com $p = 1/2$
$q = 1/2$, in-
qualquer ou-
olas mostram
2σ respectiva-
tempo, onde
xo: $p = 0.55$.

Momentos da Binomial

É interessante calcular os valores esperados da distribuição binomial. A expressão da expansão binomial **??** 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 $p\frac{\partial}{\partial p} p^m = mp^m$:

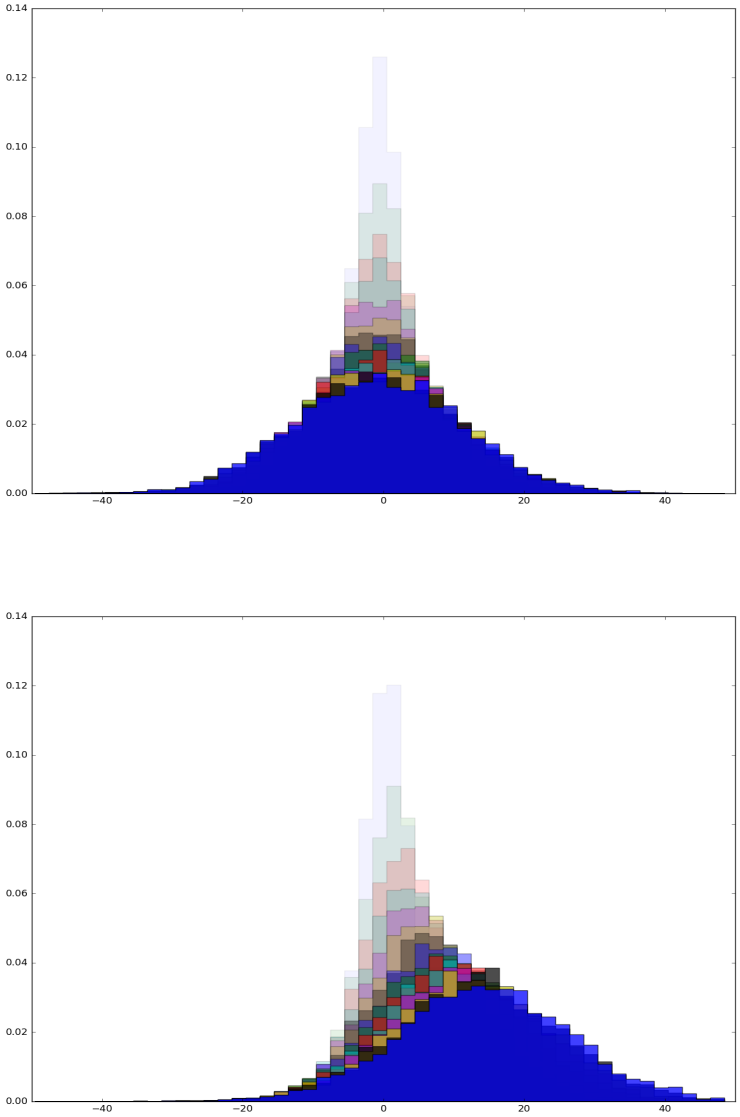
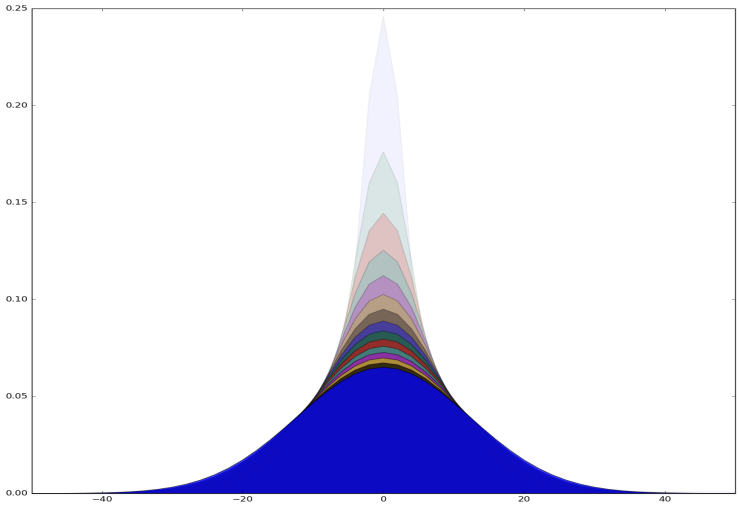
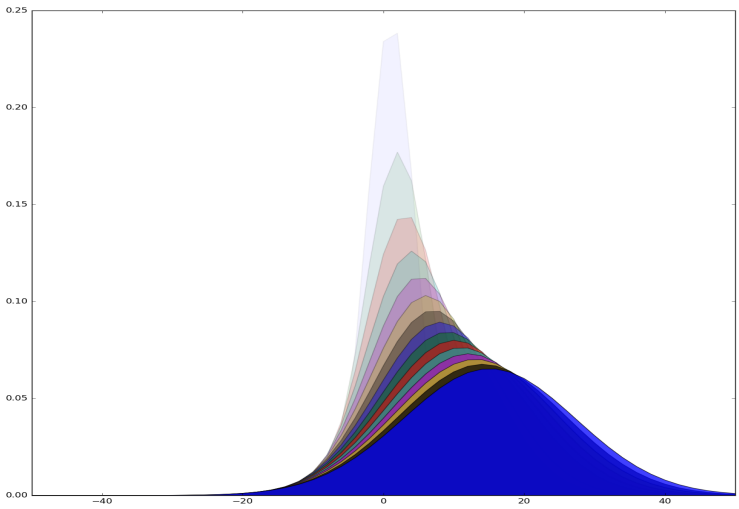


Figura 4.2: Difusão II: Histogramas obtidos dos caminhos simuladas da figura para valores $N = 10, 20....160$ com (Acima) $p = 1/2$, (Abaixo) $p = .55$.

são III: A dis-
l para valores
com (Acima)
xo) $p = .55$.





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

[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.

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 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

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

Olhe as figuras **??** e **??**. 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

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

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 **??** 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.

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 **??**, 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 **??**

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

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

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 **??**. O resultado é que

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

pois substituindo na relação de recorrência

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

vemos que **??** é de fato satisfeita pelas expressões **??**. 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

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

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

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

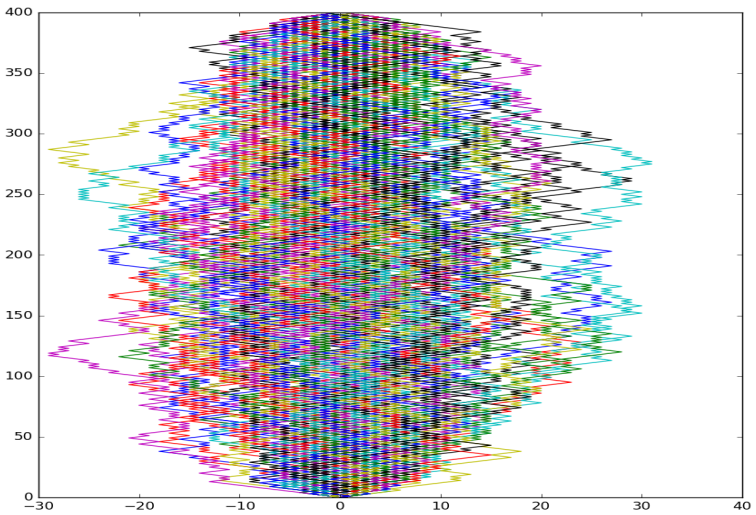
onde o numerador é obtido pelo produto de todas as diferentes combinações de escolhas possíveis de r bolas do total de M vermelhas vezes o número de combinações de b do total de $N-M$ azuis, dividido pelo total de possibilidades das combinações de $r+b$ do total de N bolas. Podemos ainda escrever a mesma 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:

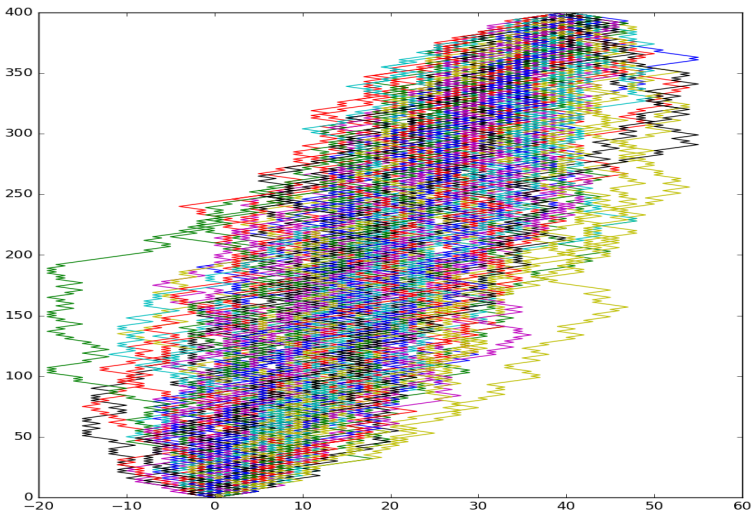
$$P(\{r_1,r_2\}|M_1,M_2,I_4) := P(\{r,b\}|N,M,I_4) = \frac{\binom{M_1}{r_1}\binom{M_2}{r_2}}{\binom{M_1+M_2}{r_1+r_2}}. \tag{4.27}$$

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

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

hipergeomé-
om $N = 400$
$= 200$ são ver-
zuis. Abaixo:
0, $M_1 = 220$
180 azuis. A
extraida o ca-
ra a direita, a
a a esquerda.





Os caminhos hipergeométricos para urnas estão mostrados nas figuras **??**, **??** e **??**. 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.

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

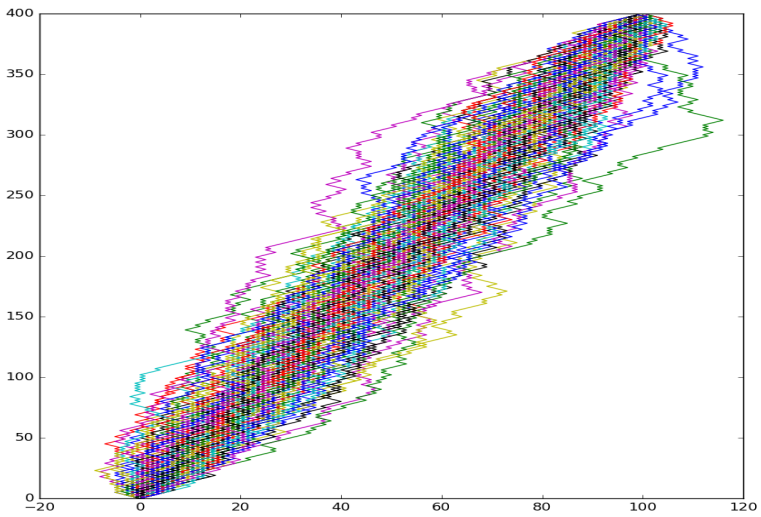
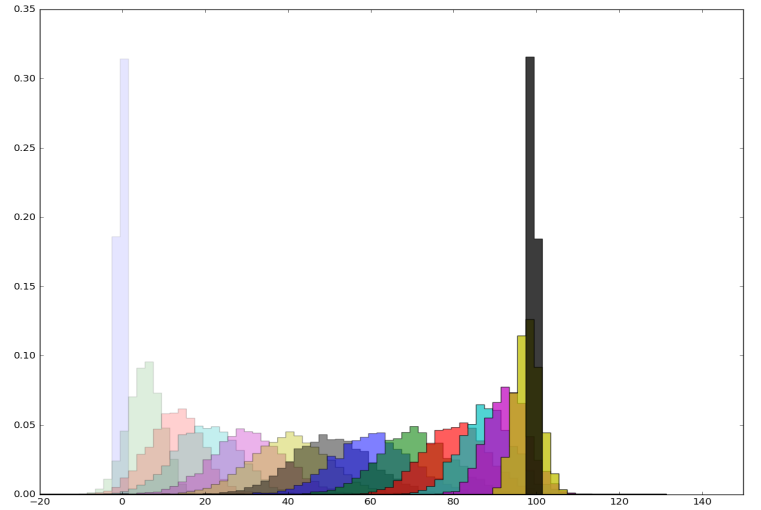


Figura 4.6: Histogramas dos caminhos hipergeométricos: simulaçao. Urna com $M = 400$, $M_1 = 250$ vermelhas e $M_2 = 150$ azuis. A situação é a mesma da figura anterior. Abaixo: Os histogramas foram gerados após extrair(1,20,50,80,120,160,200,240,280,320,350,370 bolas e olhar os resultados para 5000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



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

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

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 **??**, 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 **??**.

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

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

Voltamos ao cálculo de **??**,notando que $P(m|pnMNI) = P(m|pMI)$,

pois saber p torna desncessária a informação de n, N,

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

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.

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

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 **??**, 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.

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

[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

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 **??**. 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)

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

Lembramos que o valor médio

$$\langle m\rangle = \lambda, \tag{4.42}$$

e o segundo momento

$$\langle m^2\rangle = \lambda + \lambda^2, \tag{4.43}$$

que leva à variância

$$\sigma^2_{Poisson} = \lambda. \tag{4.44}$$

Para calcular momentos superiores podemos usar

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

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

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

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 **??**, mas ainda não temos instrumentos para estimar isto. Os histogramas dos dados e da binomial são bem diferentes, figura **??**. O histograma dos dados é muito mais estreito que o da binomial. A figura **??** 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

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.

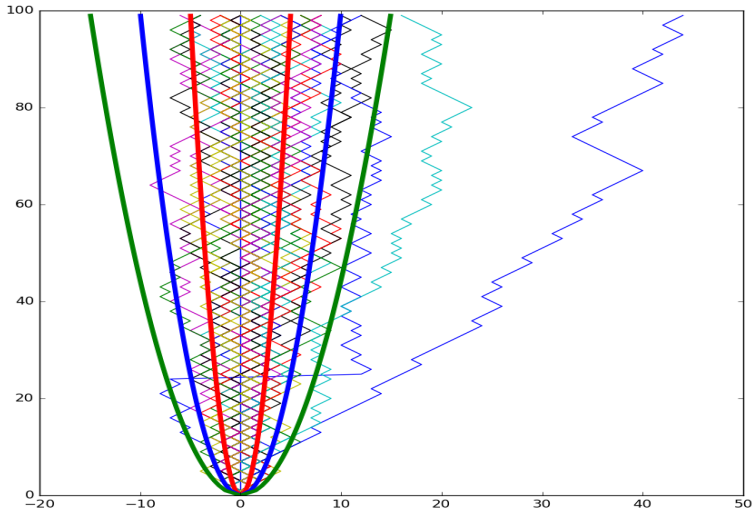
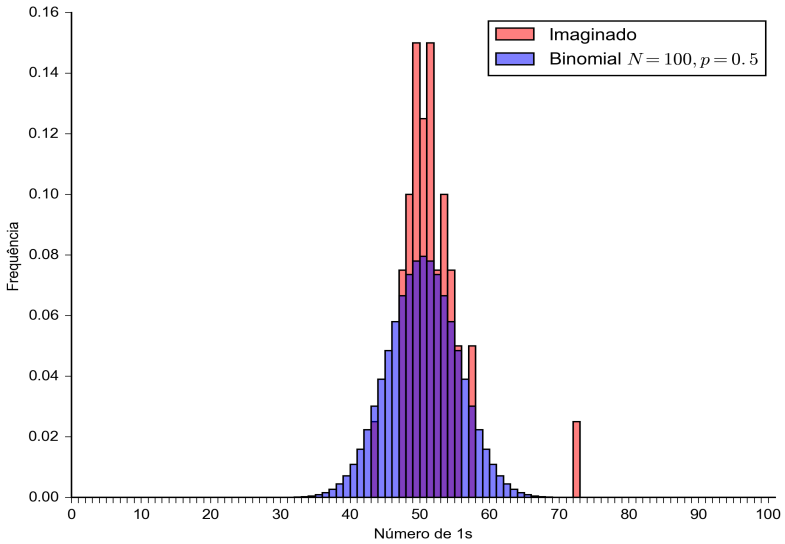
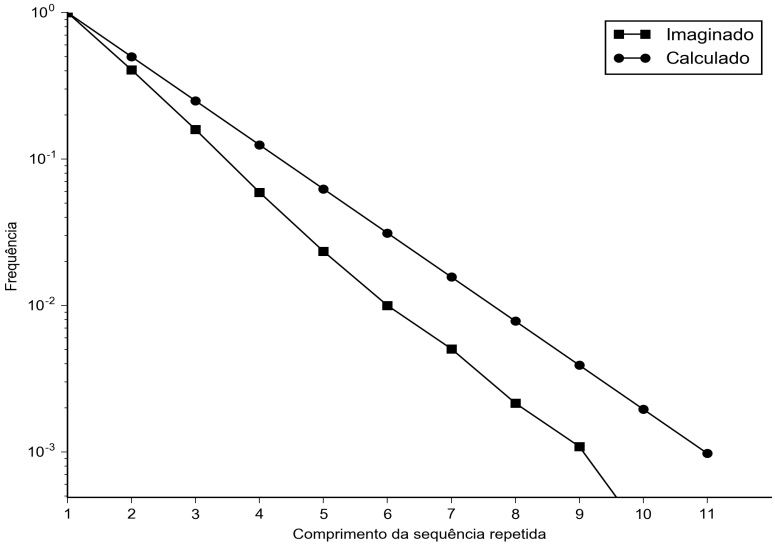


Figura 4.8: Histogramas do número de caras nos dados imaginados e o histograma do binomial simulado. $K = 40$ trajetorias de $N = 100$ jogadas.



requências de
le k símbolos.



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 **??**. 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 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.

Apêndice: (encontrar melhor lugar)Algumas desigualdades

Raramente ao lidar com modelos em Física que levam a problemas em probabilidades podemos fazer as contas de forma exata. Para progredir há algumas estratégias possíveis. Uma forma de aproximação é fazer a aproximação no modelo de interesse e depois tratar exatamente o modelo simplificado. Talvez seja estranho dizer que isto é um método de aproximar, pois as contas são exatas. Sim, exatas mas no modelo menos realista. Outra, que é encontrada em muitos dos métodos da Física Teórica, é fazer aproximações nas expressões associadas ao modelo original, que permitam avançar analiticamente e chegar a alguma conclusão. As aproximações tem a vantagem que muitas vezes, se bem sucedidas, capturam a essência do que um modelo tem que leva a algum fenômeno de interesse. A desvantagem é que ao fazer aproximações não sabemos se uma quantidade de interesse está sendo super ou subestimada. Outra estratégia é usar computadores e isso é a base da Física Computacional. Veremos um pouco disto em outros capítulos, em particular ao lidar com o método de Monte Carlo. Uma terceira estrategia usa a ideia de desigualdades, que obviamente tem sua origem em Análise. De forma simplificada são a essência de métodos rigorosos de Física Matemática.
Há uma literatura enorme sobre desigualdades e não temos nenhuma possibilidade de tratar o problema de forma exaustiva. Simplesmente queremos mostrar aos alunos alguns resultados simples e rigorosos.

Desigualdade de Jensen

Damos a seguir uma versão muito limitada de uma desigualdade que serve em muitos contextos. A função e^x tem curvatura sempre com o mesmo sinal ($d^2e^x/d^2x > 0$ para qualquer x). Funções com esta característica são ditas convexas. Verifique que a função $-\log(x)$ também é convexa para qualquer $x > 0$. O conceito pode ser estendido a mais dimensões. Aqui nos restringimos à função exponencial. Suponha uma variável aleatória x com densidade $p(x)$. Então vale

$$\mathbb{E}(e^x) \geq e^{E(x)} \tag{4.47}$$

Começamos com

$$\mathbb{E}(e^x) = \mathbb{E}(e^{x+E(x)-E(x)}) = e^{E(x)}\mathbb{E}(e^{x-E(x)}) \tag{4.48}$$

e no segundo fator usamos a desigualdade óbvia (verifique)

$$e^z \geq 1 + z, \tag{4.49}$$

com $z = x - \mathbb{E}(x)$, portanto

$$\mathbb{E}(e^x) = e^{E(x)}\mathbb{E}(e^{x-E(x)}) \geq e^{E(x)}\mathbb{E}(1 + x - \mathbb{E}(x)) = e^{E(x)}(1 + \mathbb{E}(x) - \mathbb{E}(x)), \tag{4.50}$$

que leva à desigualdade **??**. O valor da média da função exponencial (convexa) é maior que o valor da função exponencial calculada no valor médio da variável aleatória. Isto é devido ao simples fato que a curvatura é positiva. Portanto não é devido a nada além da convexidade da exponencial e por isso vale para outras funções convexas $g(x)$

$$\mathbb{E}(g(x)) \geq g(\mathbb{E}(x)). \tag{4.51}$$

Aqui só usaremos o caso da exponencial que foi provado e não o caso mais geral que não o foi.

Markov

A próxima desigualdade é devida a Markov (talvez a Chebyshev). Seja X uma variável aleatória não negativa que toma valores $[0, B)$ onde B pode ser infinito. A informação sobre X nos leva à densidade $p(x)$. Para qualquer $y \geq 0$ temos

$$P(X \geq y) = \int_y^B p(x)dx \tag{4.52}$$

$$= \int_0^B I(x \geq y)p(x)dx \tag{4.53}$$

onde $I(x \geq y)$ é a função indicadora do evento $x \geq y$. Isto é: $I(x \geq y) = 1$ se $x \geq y$ e $I(x \geq y) = 0$ se $x < y$. Note que se $x \geq y$, então $x/y \geq 1 = I(x \geq y)$, e se $x < y$ também vale $x/y \geq 0 = I(x \geq y)$. Portanto sempre vale que $x/y \geq I(x \geq y)$. Então podemos substituir a igualdade **??**, que talvez não saibamos calcular (se soubessemos para que fazer isto?), por uma desigualdade

$$P(X \geq y) = \int_0^B I(x \geq y)p(x)dx \tag{4.54}$$

$$\leq \int_0^B \frac{x}{y}p(x)dx \tag{4.55}$$

$$= \frac{1}{y}\int_0^B xp(x)dx \tag{4.56}$$

$$P(X \geq y) \leq \frac{\mathbb{E}(X)}{y}. \tag{4.57}$$

Esta é a desigualdade de Markov. Substituimos uma igualdade, a eq. **??**, talvez intratável analiticamente por uma desigualdade, que talvez nos dê algo concreto e rigoroso, embora possa não ser preciso. Talvez o estudante não esteja impressionado nesta altura. Mas, fazendo variações sobre o tema, este é o ponto de partida para várias outras desigualdades que muitas vezes representam a única forma de atacar rigorosamente alguns problemas.
Exercício Aplique ao caso da distribuição exponencial. Mostre que $e^{-y} \leq \frac{1}{y}$, para $y \geq 0$.

Mudança de variáveis

Seja X como acima. Suponha que Z é uma função de X, i.e. toma valores z dados por $z = f(x)$ e tem densidade de probabilidade $\tilde{p}(z)$. Lembre que a relação entre as densidades é

$$p(x)dx = \tilde{p}(z)dz.$$

Partimos da desigualdade de Markov para a variável aleatória Z e mudamos variáveis

$$\begin{aligned} P(Z \geq y) &\leq \frac{1}{y}\int_{f(0)}^{f(B)} z\tilde{p}(z)dz && (4.58)\\ &= \frac{1}{y}\int_0^B f(x)p(x)dx && (4.59)\end{aligned}$$

Portanto temos

$$P(f(x) \geq y) \leq \frac{\mathbb{E}(f(x))}{y} \quad (4.60)$$

onde $\mathbb{E}(f(x)) = \int f(x)p(x)dx$.

Chebyshev

Consideremos o caso particular $f(x) = (x - \mathbb{E}(X))^2$. Isto é interessante porque se X se afasta do valor esperado por mais que uma quantidade $t > 0$ temos duas possibilidades, ou

$$x \geq \mathbb{E}(x) + t$$

ou

$$x \leq \mathbb{E}(x) - t.$$

As duas condições podem ser escritas como $f(x) \geq t^2$. Portanto a desigualdade de Markov nos leva a

$$\begin{aligned} P((x - \mathbb{E}(X))^2 \geq t^2) &\leq \frac{\mathbb{E}((x - \mathbb{E}(X))^2)}{t^2}\\ P((x - \mathbb{E}(X))^2 \geq t^2) &\leq \frac{\mathrm{Var}(X)}{t^2} && (4.61)\end{aligned}$$

Esta última é conhecida como desigualdade de Chebyshev. A probabilidade de que X tenha um valor que se afasta da média por mais de um valor $t > 0$ cai com t^2 e a escala é dada pela variância de X. É claro que se a variância for infinita a desigualdade não diz nada.

Cotas exponenciais: Chernoff

Lembremos a função geradora dos momentos. Tomemos como definição

$$\Phi(\xi) = \mathbb{E}(e^{\xi x})$$

onde é melhor usar a versão real, embora seja às vezes conveniente introduzir a versão com exponenciais imaginárias.

Suponha que novamente olhemos para a probabilidade de $z \geq t$. Exponenciando, para $\xi > 0$, temos $e^{\xi x} \geq e^{\xi t}$. Portanto

$$P(x \geq t) = P(e^{\xi x} \geq e^{\xi t}) \leq \frac{\mathbb{E}(e^{\xi x})}{e^{\xi t}}. \tag{4.62}$$

mas o valor de ξ está à nossa disposição e

$$P(x \geq t) \leq \min_{\xi}\{\mathbb{E}(e^{\xi x})e^{-\xi t}\}. \tag{4.63}$$

Para avançar mais é preciso fazer hipóteses sobre as propriedades de $P(x)$. Faremos alguns exemplos simples para ver quão boa ou ruim pode ser esta desigualdade.
Caso exponencial: $x \sim \exp(-x)$.
Os valores médios $\mathbb{E}(x) = 1$ e $\mathbb{E}(x^2) = 2$ são fáceis de calcular. Perguntamos qual a probabilidade que x seja maior que o valor médio por uma quantidade t. Note que é fácil fazer as contas:

$$P(x \geq 1 + t) = \int_{1+t}^{\infty} e^{-x}dx = e^{-(1+t)},$$

mas suponha que não sabemos. Para $0 \leq \xi < 1$ temos

$$\mathbb{E}(e^{\xi x}) = \int_0^{\infty} e^{\xi x}e^{-x}dx = \frac{1}{1-\xi}$$

e

$$P(x \geq 1+t) = P(e^{\xi x} \geq e^{\xi(1+t)}) \leq \min_{\xi} \mathbb{E}(e^{\xi x})e^{-\xi(1+t)} = \min_{\xi} \frac{e^{-\xi(1+t)}}{1-\xi}.$$

O mínimo ocorre para $\xi = t/(1+t)$ e leva à cota de Chernoff

$$P(x \geq 1+t) \leq (1+t)e^{-t}$$

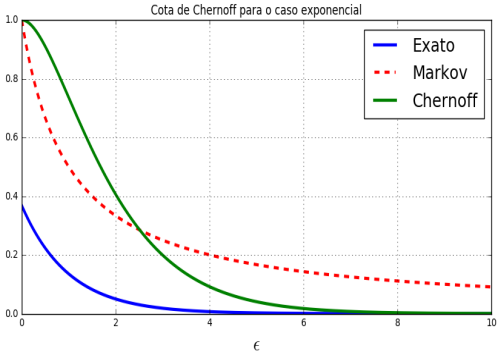
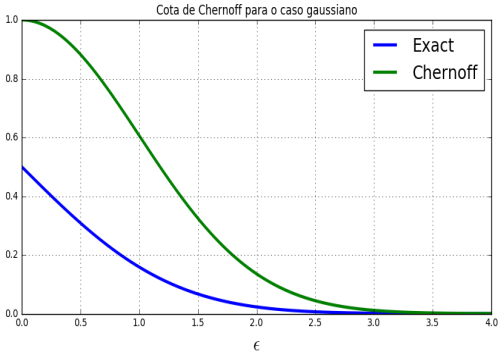


Figura 4.10: $P(x \geq \epsilon)$. Resultado exato e cotas de Markov de Chernoff para o caso exponencial

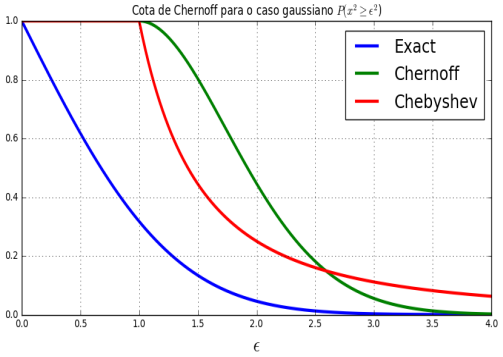
Caso Gaussiano: $x \sim \mathcal{N}(0, \sigma)$
A função geradora

$$\Phi(\xi) = \mathbb{E}(e^{\xi x}) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}+\xi x}dx$$

$x \geq \epsilon$). Re-
cota de Cher-
so gaussiano



$x \leq -\epsilon$). Re-
de Chebyshev
so gaussiano



Completando quadrados

$$-\frac{x^2}{2\sigma^2} + \xi x + \frac{A^2 - A^2}{2\sigma^2} = -\frac{(x-A)^2}{2\sigma^2} + \frac{A^2}{2\sigma^2}$$

onde foi necessário escolher A para cancelar o termo linear ξx:
$A = \xi\sigma^2$.

$$\Phi(\xi) = e^{\frac{A^2}{2\sigma^2}} = e^{\frac{\xi^2\sigma^2}{2}}$$

Note que na dedução da desigualdade de Markov usamos na desigualdade **??** que a variável aletória é não negativa para majorar a função indicadora. Aqui podemos fazer isto pois a exponencial é positiva. Voltando à equação **??**

$$P(x \geq t) \leq \min_{\xi}\{e^{\frac{\xi^2\sigma^2}{2}} e^{-\xi t}\}. \tag{4.64}$$

O mínimo é obtido para $\xi = t/\sigma^2$ e portanto, colocando $t = \epsilon\sigma$

$$P(z \geq \epsilon\sigma) \leq e^{-\frac{\epsilon^2}{2}}. \tag{4.65}$$

Neste caso sabemos que

$$P(z \geq \epsilon\sigma) = \frac{1}{\sqrt{2\pi}} \int_{\epsilon}^{\infty} e^{-\frac{x^2}{2}} dx := \phi(\epsilon), \tag{4.66}$$

a função complementar da cumulativa.

Podemos perguntar qual é a cota para o caso $f(x) = (x - \mathbb{E}(X))^2$ que nos levou à desigualdade de Chebyshev. O valor exato

$$P(x \geq \epsilon \text{ ou } x \leq -\epsilon) = \frac{2}{\sqrt{2\pi}} \int_{\epsilon}^{\infty} e^{-\frac{x^2}{2}} dx := 2\phi(\epsilon). \tag{4.67}$$

Isto é igual a $P(e^{\xi^2 x^2} \geq e^{\xi^2 \epsilon^2})$ que leva a uma desigualdade no estilo de Chernoff. Para $\xi < 1$:

$$\begin{aligned} P(e^{\frac{1}{2}\xi^2 x^2} \geq e^{\frac{1}{2}\xi^2 \epsilon^2}) &= \int_{e^{\frac{1}{2}\xi^2 x^2} \geq e^{\frac{1}{2}\xi^2 \epsilon^2}}^{\infty} e^{-\frac{1}{2}x^2} \frac{dx}{\sqrt{2\pi}} && (4.68) \\ &\leq \min_{\xi} \int_{-\infty}^{\infty} e^{\frac{1}{2}\xi^2 x^2} e^{-\frac{1}{2}\xi^2 \epsilon^2} e^{-\frac{1}{2}x^2} \frac{dx}{\sqrt{2\pi}} && (4.69) \\ &= \min_{\xi} \frac{e^{-\frac{1}{2}\xi^2 \epsilon^2}}{\sqrt{1-\xi^2}} && (4.70) \end{aligned}$$

que é alcançado para $\xi^2 = 1 - 1/\epsilon^2$ quando $\epsilon \geq 1$ e $\xi = 1$ para $0 < \epsilon \leq 1$. Desta forma temos uma cota trivial para $\epsilon < 1$ e $P(x^2 \geq \epsilon^2) \leq \epsilon \exp((1-\epsilon^2)^2)$, mostrado na figura **??**.
Os resultados obtidos acima não servem para outra coisa que desenvolver a intuição, pois não há sentido em usar cotas desta natureza quando as distribuições são tão simples que os cálculos podem ser feitos exatamente. O estilo de pensar sobre cotas requer um certo treinamento e será útil na prova de teoremas limite como o teorema central e em outros contextos.