# Notas das aulas de Probabilidades e Estatística

Salvatore Cosentino
Departamento de Matemática, Universidade do Minho,
Campus de Gualtar, 4710 Braga PORTUGAL
gab B.4023, tel +351 253 604086
E-mail scosentino@math.uminho.pt
URL http://w3.math.uminho.pt/~scosentino

25 de Abril de 2006

## Contents

# Advertências

A *teoria das probabilidades* é uma teoria física que trata das regularidades de fenómenos aleatórios (do latim ALEA, jogo de dados). Como toda teoria física, ela consiste num modelo matemático e regras operacionais que estabelecem a correspondência entre alguns objectos do modelo e os observáveis das experiências.

Praticamente todos os pensadores da idade moderna foram fascinados pelos problemas e paradoxos dos jogos, entre eles Galileo e Newton. Com Pascal, Fermat e Huygens nasceram os conceitos de probabilidades e esperança. A historia moderna começa com os trabalhos de James Bernoulli, De Moivre, Daniel Bernoulli, Euler, Gauss, Laplace e Poisson, quando nasceram os conceitos de probabilidade condicionada e independência, e a procura de teoremas limites ganhou o interesse dos matemáticos. A formulação axiomática moderna (mas não é a única!) é devida ao russo Kolmogorov que, pouco antes da segunda guerra mundial, elaborou a observação de Borel de que a linguagem natural da teoria das probabilidades era a teoria da medida recentemente fundada. A teoria dos processos estocásticos (desenvolvida entre outros por Chebyshev, Markov, Lyapunov, Khinchin, Wiener, Doob, Feller, Paul Pierre Lévy) e a teoria da informação (inventada por Shannon logo depois da guerra) fazem agora parte da bagagem de todas as ciências experimentais.

Uma das ocupações favoritas dos físicos (além de escalar montanhas) é decidir sobre a verdade de afirmações do tipo "numa experiência com certas condições de laboratório observo que acontece uma certa coisa", e chamar a isto uma lei da natureza. Se a experiência consiste em lançar um dado, não é claro como fazer previsões sobre a ocorrência de eventos como "sair uma pinta" ou "sair um número par de pintas". O que se observa, por outro lado, é que se repetimos a experiência muitas vezes as frequências dos eventos parecem ser números reproduzíveis. O tema central da teoria das probabilidades é esta transição da aleatoriedade à quase certeza, em presença de um número grande de provas. Regularidades deste tipo são bem conhecidas em física. Não há forma de saber o instante em que decai um núcleo de rádio, mas a fração de núcleos que decaem em $t$ anos dentro duma amostra grande segue uma lei empírica do tipo $1 - e^{-0.000436 \cdot t}$. É praticamente impossível fazer previsões sobre as trajectórias de um número pequeno de moléculas de gás fechadas numa caixa (o espaço das fases tem dimensão 6 vezes o número de moléculas, e ninguém sabe resolver uma equação diferencial não trivial em tantas variáveis), mas o estado "macroscópico" de um número enorme de moléculas (um litro de gás contém algo como 60000000000000000000000 moléculas!) é muito bem descrito por apenas dois parâmetros: o volume e a temperatura.

Um modelo destas situações consiste em associar a cada evento observável $A$ um número $\text{prob}(A)$ entre 0 e 1, e chamar este número "a probabilidade da ocorrência do evento ser observada numa dada experiência". Se os eventos $A$ e $B$ são incompatíveis, i.e. não podem acontecer seja $A$ seja $B$, é natural pôr

$$\text{prob}(A \text{ ou } B) = \text{prob}(A) + \text{prob}(B).$$

O modelo será portanto uma função aditiva "prob" definida sobre os eventos, uma álgebra de subconjuntos do espaço dos estados do sistema físico. A ideia central, que faz a teoria interessante, é o conceito de independência. Se julgamos que a informação acerca da ocorrência de $B$ não ajuda na previsão sobre a ocorrência do evento $A$, uma pequena reflexão sobre as frequências esperadas leva a conjecturar que

$$\text{prob}(A \text{ e } B) = \text{prob}(A) \times \text{prob}(B).$$

No modelo, portanto, a hipótese da independência de duas experiências será traduzida na multiplicatividade das relativas probabilidades (justificar esta hipótese a partir das leis da física é o trabalho de muitos matemáticos e físicos desde o tempo de Boltzmann!). O modelo permite calcular probabilidades de eventos. Resultados interessantes e, sobretudo, mais simples são obtidos idealizando a possibilidade de repetir infinitas vezes uma experiência, ou a possibilidade de ter um espaço dos estados infinito. De facto, a teoria trata modelos de evolução temporal, chamados processos estocásticos, cujas regularidades são fenómenos assimptóticos, ou modelos de sistemas com um número grande de componentes "microscópicas", cujas regularidades são devidas a fenómenos colectivos que são visíveis quando o número de componentes tende para infinito (e, por alguma razão misteriosa, a derivação de algumas das leis realmente observadas em termodinâmica parece precisar mesmo deste processo, dito limite termodinâmico). Isso convida/obriga a restringir o domínio da função "prob" a uma $\sigma$-álgebra, e a postular uma forma mais forte de aditividade, a aditividade enumerável.

Objectivo da teoria é fazer modelos de fenómenos físicos, e estimar probabilidades de eventos complicados, na esperança de provar que certos eventos particularmente significativos tenham

probabilidade muito grande ou muito pequena. Como sempre acontece em física, são os próprios teoremas do modelo matemático que sugerem a interpretação física dos objectos do modelo, i.e. sugerem quais e em que sentido alguns dos parâmetros do modelo são "observáveis". Entre outras coisas, o modelo faz previsões sobre as frequências em que ocorrem os eventos repetindo muitas vezes a experiência, a previsão é do tipo "a frequência está neste intervalo com uma dada probabilidade", e produz teoremas elegantes sobre eventos que dependem da observação de um número infinito de experiências. O interesse da teoria das probabilidades consiste na possibilidade das suas previsões, por vezes longe de serem intuitivas, serem realmente observadas... no sentido de que, se o modelo diz que um evento tem probabilidade 0.999, o evento é mesmo observado (se não for observado numa experiência, podemos pensar que tivemos muito azar, se não for observado em duas, três, quatro experiências seguidas, podemos tranquilamente jogar no lixo o nosso modelo).

A *estatística*, para quem faz ciência, é o pão nosso de cada dia, a "interface" entre um fenómeno da natureza (ou seja os resultados de uma experiência, afectados por erros que julgamos "casuais" ou que queremos detectar) e o seu modelo teórico (a mecânica newtoniana, a mecânica quântica, ...a teoria das probabilidades): um conjunto de ideias, interpretações, estratégias e receitas que têm o objectivo de testar o modelo (ou seja aceitar ou rejeitar relações entre observáveis), ajustar os seus parâmetros, fazer previsões sobre os resultados das experiências. Enfim, a estatística joga um papel decisivo dentro das regras operacionais que fazem a ligação entre o modelo teórico e o pedaço de natureza que ele pretende descrever. A validade das técnicas da estatística é fundamentada na observação empírica de que por vezes elas dão respostas credíveis, e, se utilizadas com honestidade, dão as únicas respostas credíveis. Outra observação empírica é que os erros casuais, supostamente devidos a pequenas variações nas condições do laboratório, "parecem" seguir leis simples. Por outro lado, a lei dos grandes números e o teorema do limite central, resultados formais da teoria das probabilidades, "parecem" oferecer argumentos em favor desta hipótese (uma piada atribuida a Henri Poincaré diz que "hà algo misterioso na distribuição gaussiana, pois todos estão convencidos que a distribuição gaussiana descreve o comportamento dos erros aleatórios: os matemáticos porque acham que os físicos a verificaram experimentalmente, os físicos porque acham que os matemáticos o demonstaram"). Melhor seria dizer que as leis simples são um bom compromisso entre a falta de informação acerca de como funciona realmente a natureza e a necessidade de ter modelos tratáveis. A verdade é que a decisão final acerca de uma experiência da física deve ser tomada com base no bom senso do cientista, na sua experiência, na sua honestidade (as revistas científicas nas prateleiras das universidades estão cheias de falsos anúncios revolucionários) e na sua intuição (histórias são conhecidas de observações "inexplicáveis" que deram origem a verdadeiras revoluções científicas).

Existe uma grande quantidade de óptimos livros de probabilidades. Uma excelente introdução, cheia de exemplos, discussões e problemas interessantes, é o clássico de Feller [Fe68]. Outra é o manual de Gnedenko [Gn73]. Uma introdução muito bem escrita em português é o livro de Pestana e Velosa [PV02]. Para um público matematicamente adulto, o manual de Shiryaev [Sh96]. Outros clássicos são os de Billingsley [Bi79], Breiman [Br68], Doob [Do53], Lamperti [La66], Loève [Lo55], Rényi [Ré74]. Para quem tiver pressa, o recente e "essencial" de Jacod e Protter [JP00]. Uma proposta herética para uma axiomática da teoria das probabilidades está no livro de De Finetti [DF91].

Introduções muito elementares mas honestas aos métodos da estatística estão no livrinho de Young [Yo62] e nas notas de McBane [McB01]. Um manual clássico sobre o tratamento dos dados experimentais é o de Bevington e Robinson [BR92]. Uma excelente introdução à estatística matemática é o já citado manual de Pestana e Velosa [PV02]. Um livro técnico é o de Mood, Graybill e Boes [MGB74]. Centenas de outros manuais podem ser encontrados nas prateleiras das bibliotecas. Cuidado: há alguns destes que parecem só fornecer certezas e receitas. Se isso acontecer, o meu conselho é fechar o livro e abrir o que está ao lado...

Isto não é um manual. Estas páginas contêm as minhas notas esboçadas preparando as aulas de "Métodos estatísticos" para alunos de engenharia, física e química, e depois de "Probabilidades e estatística" para alunos de matemática. Foram pensadas como uma introdução às ideias mais elementares da teoria das probabilidades e às principais técnicas da estatística utilizadas pelos físicos. Algumas matérias, que aliás não fazem parte do programa oficial, não foram leccionadas: umas por falta de tempo e outras (those written in english, so that if you are not curious you may avoid to print them and waste paper, ink and time) para poupar aos alunos uma discussão demasiado técnica. As notas foram escritas de maneira sintética, esquemática e muito informal, com a esperança que o leitor veja, nos livros sérios, como pôr correctamente os problemas e como demonstrar os resultados interessantes. Outra esperança é que o leitor fique com mais curiosidade,

e dúvidas, do que dogmas. Infelizmente, muitos dos resultados bonitos e surpreendentes da teoria das probabilidades são "difíceis", precisando da linguagem e das técnicas da teoria da medida e da integração (matéria que não faz parte do curriculum dos nossos cursos universitários), e muitos dos resultados quantitativos interessantes consistem em estimações laboriosas, que não cabem nas duas horas semanais de um semestre lectivo de três meses. É por isso que tive, com muita vergonha, que enunciar teoremas sem demonstrações, ou tentar definir objectos numa linguagem aproximativa. A minha única preocupação foi com o conteúdo "físico" da matéria.

Uma última advertência: a teoria das probabilidades e a estatística tratam e ajudam a compreender problemas bem mais interessantes e difíceis do que aqueles que aparecem nestas páginas, e que surgem naturalmente em física, biologia, engenharia, economia, linguística... O meu conselho é folhear pelo menos o primeiro volume do livro de Feller. Uma discussão informal e inteligente do papel que jogam "o acaso e o caos" na ciência moderna e na nossa visão do mundo está no livrinho de Ruelle [Ru91], um dos pais da mecânica estatística contemporânea.

Braga, 16 de Dezembro de 2004.
sal.

# 1 Introdução

**Observações e estimação.** Um físico tem uma teoria física, que contém um observável chamado $x$ (a constante de gravitação de Newton, a massa do electrão, o tempo característico do carbono $C_{14}$, ...a probabilidade de sair cara no lançamento de uma moeda), e quer estimar o seu valor. Repete várias vezes uma experiência em condições que ele julga idênticas (no sentido em que controla tudo o que é controlável) e obtém os resultados experimentais $x_1, x_2, ..., x_n$. A coisa mais honesta que ele pode dizer é que o observável está entre $x_{\min}$ e $x_{\max}$, mais ou menos. Os físicos costumam acreditar na existência do universo, e nas próprias teorias, portanto na existência do valor "verdadeiro" de $x$. Uma estimação natural é a *média aritmética* dos resultados

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right)$$

Os físicos também sabem que não faz sentido nenhum acreditar que o valor de $x$ seja exactamente $\overline{x}$ (as leis da física implicam que a posição de Vénus influencie a queda de uma pedra da torre de Pisa, embora não seja possível dizer qual é a sua influência!), só acreditam em afirmações como

$$\text{o observável } x \text{ é igual a } \overline{x} \pm \Delta x$$

que lêem: "o verdadeiro valor do observável $x$ está, com grande probabilidade, entre $\overline{x} - \Delta x$ e $\overline{x} + \Delta x$". Um dos problemas da estatística é

- estimar um valor razoável do "erro" $\Delta x$.

**Média aritmética e desvio padrão.** A média aritmética $\overline{x}$ é a média mais democrática entre os valores observados. É também o valor de $a$ que minimiza a soma

$$(x_1 - a)^2 + (x_2 - a)^2 + ... + (x_n - a)^2$$

dos quadrados dos "erros" nas distintas observações. Se acreditamos que $\overline{x}$ seja uma boa estimação do valor de $x$, então $x_k - \overline{x}$ pode ser interpretado como sendo o "erro cometido na $k$-ésima observação". A média aritmética dos "erros quadráticos" é

$$S^2 = \frac{1}{n}\left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... (x_n - \overline{x})^2\right)$$

e a sua raiz $S = \sqrt{S^2}$, dita *desvio padrão (standard deviation),* é uma medida de quanto cada valor $x_k$ difere de $\overline{x}$.

Uma apresentação honesta dos resultados das $n$ experiências é

$$x = \overline{x} \pm S$$

que pode ser lida como: "foram observadas flutuações da ordem de $S$ à volta de um valor médio $\overline{x}$". O valor de $S$ é uma medida da "sensibilidade" dos instrumentos do laboratório, ou melhor da reproduzibilidade das experiências.

**Desvio padrão da média .** O senso comun sugere que quanto maior for o número $n$ das observações quanto mais próxima a média $\overline{x}$ está do verdadeiro valor de $x$. Conjecturas razoáveis acerca da distribuição dos erros $x_k - x$ (sugeridas pelos histogramas dos dados experimentais) e considerações probabilísticas (o teorema do limite central) permitem quantificar esta expectativa. Por exemplo, se $n$ é grande e os histogramas dos dados experimentais fazem suspeitar que a distribuição dos erros é "gaussiana", o resultado é que as flutuações de $\overline{x}$ à volta de $x$ são da ordem de $S_m = S/\sqrt{n}$, dito *desvio padrão da média (standard deviation of the mean)*, e portanto podemos acreditar que

$$x = \overline{x} \pm S/\sqrt{n}$$

Justificar o factor $1/\sqrt{n}$ é um dos objectivos da teoria das probabilidades.

**Apresentação do resultado.** Dizer que um observável é igual a

$$x = 3.14159265359 \pm 0.062$$

não contém mais informação do que dizer que é igual a

$$x = 3.14 \pm 0.06$$

O "erro relativo" $\Delta x/\overline{x}$ indica a quantidade dos dígitos significativos, ou seja confiáveis, na estimação de $x$.

Por exemplo, uma tabela das constantes da física tem este valor da constante de gravitação de Newton:

$$G = 6.673(10) \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2} \text{ with relative standard uncertainty } 1.5 \times 10^{-3}$$

Isto quer dizer que, embora a média observada seja $6.67310 \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$, só podemos confiar nos primeiros três dígitos decimais deste valor.

**Modelização.** Uma lei física é uma relação entre um certo número de observáveis. Um exemplo é

$$y = f(x, a)$$

onde $y, x, a$ são certos observáveis (por exemplo, a lei de Hubble diz que a velocidade $v$ de afastamento de uma galáxia é igual a $H \cdot r$, onde $r$ é a distância entre a galáxia e a Via Lactea, e $H$ é a constante de Hubble). Uma experiência típica consiste em observar os valores $y_1, y_2, ..., y_n$ correspondentes a um certo número de valores $x_1, x_2, ..., x_n$ de $x$, considerada como variável independente sobre a qual temos um bom controlo, e portanto nenhum erro significativo. Se possível, cada $y_k$ é observada mais vezes, e portanto estimada com a sua média $\overline{y}_k$ e o seu desvio padrão $S_k$. O objectivo da experiência é

- estimar os valores dos "parâmetros livres" $a$ que mais concordam com as observações,
- decidir se a lei, i.e. a forma da função $f$, descreve bem os resultados da experiência.

**Mínimos quadrados.** A primeira coisa que um físico faz é desenhar no plano $x$-$y$, em correspondência de cada $x_k$, o intervalo $\overline{y}_k \pm S_k$. Depois, procura um valor $\alpha$ do parâmetro $a$ tal que a curva $y = f(x, \alpha)$ passe quanto mais próxima possível de todos os pontos $(x_k, \overline{y}_k)$, esperando que não se afaste mais do que $\pm S_k$ destes pontos. Uma receita razoável, dita método dos *mínimos quadrados (least-square fitting)*, é escolher o estimador $\alpha$ para o parâmetro $a$ de maneira tal que a soma

$$\sum_{k=1}^{n} \frac{\left(\overline{y}_k - f(x_k, a)\right)^2}{S_k^2}$$

seja a menor possível. Observe que acima cada "erro quadrático" $\left(\overline{y}_k - f(x_k, a)\right)^2$ é pesado com um factor inversamente proporcional ao quadrado da incerteza $S_k$ no valor $\overline{y}_k$.

Em teoria, desde que a função $f$ seja diferenciável, o valor de $\alpha$ é obtido calculando derivadas parciais e resolvendo um sistema de equações. Na prática, se a forma de $f$ não é simples, este é um problema difícil. O melhor é procurar soluções aproximadas, por exemplo utilizando técnicas de análise numérica.

**Qui-quadrado.** O valor de

$$\chi^2 = \sum_{k=1}^{n} \frac{\left(\overline{y}_k - f(x_k, \alpha)\right)^2}{S_k^2}$$

é uma medida da bondade do ajustamento. Quanto maior for $\chi^2$ quanto menos a curva $y = f(x, \alpha)$ está próxima dos dados $(x_k, \overline{y}_k)$. Conjecturas acerca da distribuição dos erros e considerações probabilísticas permitem quantificar quais valores de $\chi^2$ podem ser considerados aceitáveis, e quais nos fazem suspeitar que a lei não descreve bem os resultados da experiência.

**Probabilidade do homen na rua.** A teoria das probabilidades nasceu como a arte de utilizar a matemática para fazer previsões quantitativas acerca de fenómenos que são, por quanto podemos ver, aleatórios, como apostar na bolsa de valores, jogar cartas, lançar dados... A situação arquétipa é lançar uma moeda. Dois resultados são possíveis, "cara" ou "coroa", e ninguém sabe honestamente prever o resultado de um lançamento. Por outro lado, toda gente acha que se lançar uma moeda "honesta" $n$ vezes, e se $n$ for suficientemente grande, o número $S_n$ de vezes que sai cara será mais ou menos $n/2$. De facto, esperamos obter uma frequência $S_n/n$ da ordem de $1/2$, e isto é o que o homen na rua entende ao dizer que "a probabilidade de sair cara no lançamento de uma moeda honesta é igual a um-meio".

**Modelos probabilísticos.** O problema é que $n/2$ é apenas a nossa "melhor aposta" para $S_n$, e de facto ninguém espera obter "exactamente" o mesmo número de caras e de coroas em $n$ lançamentos. Isto seria ter muita sorte! Portanto ficamos na mesma: ainda não sabemos como fazer previsões. O que é preciso é inventar um modelo, e fazer contas. Com sorte, o modelo dirá que tipo de previsões temos o direito de fazer.

A ideia é quantificar a nossa expectativa acerca de um evento como "observar $k$ caras em $n$ moedas". Associamos um número entre zero e um a cada um destes eventos, que chamamos $\text{prob}\,(k \text{ caras em } n \text{ moedas})$ e lemos "probabilidade de observar $k$ caras em $n$ moedas". Uma maneira natural de o fazer é contar a cardinalidade dos casos favoráveis, todos os que levam ao resultado $S_n = k$, e dividir este número pela cardinalidade dos casos possíveis. Isto quer dizer definir

$$\text{prob}\,(k \text{ caras em } n \text{ moedas}) = \frac{|\text{casos favoráveis}|}{|\text{casos possíveis}|}$$

Naturalmente, somos livres de definir o que queremos, e até agora esta é apenas uma definição que não faz mal. Agora, deixando aos filósofos a tarefa de dizer o que a probabilidade "é", estabelecemos a seguinte "interpretação" do nosso modelo: "se o modelo diz que um certo evento tem probabilidade muito grande, como 0.99 ou 0.999 ou mais, então o evento é observado praticamente em todas as vezes que repetimos a experiência (se não for observado numa experiência, podemos pensar que tivemos muito azar, se não for observado em duas, três, quatro experiências seguidas, podemos tranquilamente jogar no lixo o nosso modelo)". Se conseguimos encontrar un tal evento, o que estamos a fazer é a previsão de que este evento vai acontecer.

**Regularidades probabilísticas.** Vamos calcular a nossa probabilidade $\text{prob}\,(k \text{ caras em } n \text{ moedas})$. O número dos casos possíveis é $2^n$, pois cada uma das $n$ moedas pode mostrar duas faces. O número dos casos favoráveis, e isto obriga a uma pequena reflexão, é

$$\frac{n!}{k! \cdot (n-k)!}$$

De facto, esta é a cardinalidade de todas as palavras de comprimento $n$ nas letras "cara" ou "coroa" que contêm $k$ vezes a letra "cara". O resultado é que o número que associamos ao evento "observar $k$ caras em $n$ moedas" é

$$\text{prob}\,(k \text{ caras em } n \text{ moedas}) = \frac{\frac{n!}{k! \cdot (n-k)!}}{2^n}$$

Quando $n$ é pequeno, este número não diz grande coisa. Por exemplo, a fórmula acima diz que $\text{prob}\,(1 \text{ cara em } 1 \text{ moeda}) = 1/2$ ou que $\text{prob}\,(1 \text{ cara em } 2 \text{ moedas}) = 1/2$, e o significado destas afirmações é o que encarregamos o nosso amigo filósofo de explicar-nos.

E' ao observar um histograma da função $k \mapsto \text{prob}\,(k \text{ caras em } n \text{ moedas})$ quando $n$ é grande que descobrimos um fenómeno interessante. O histograma tem a forma de um "sino" centrado no ponto $n/2$, e rapidamente decresce para valores praticamente nulos quando $|k - n/2|$ cresce. O máximo é no ponto que corresponde à nossa melhor aposta, mas é da ordem

$$\text{prob}\,(n/2 \text{ caras em } n \text{ moedas}) \sim 1/\sqrt{n}$$

um número muito pequeno se $n$ é grande. Por outro lado, ao somar todos os valores de $\text{prob}\,(k \text{ caras em } n \text{ moedas})$ num intervalo de comprimento $\sqrt{n}$ à volta de $n/2$ (os valores de $k$ para os quais a função é significativamente superior a zero) obtemos algo da ordem de $\sqrt{n} \cdot 1/\sqrt{n} \sim 1$,

$$\text{prob}\left(n/2 \pm \sqrt{n} \text{ caras em } n \text{ moedas}\right) \sim 1$$

Ou seja, encontramos um evento quase certo! Juntamente com a intrpretação acima esta é uma previsão: ao lançar um número grande $n$ de moedas, esperamos observar um número de caras no intervalo

$$S_n \simeq n/2 \pm \sqrt{n}$$

**Teorema do limite central.** Esta estimação pode ser melhorada utilizando um pouco de análise. O resultado, chamado "teorema do limite central", é que, oportunamente normalizada, a lei de $S_n/n - 1/2$ se estabiliza perto de uma lei universal dita "gaussiana" ao crescer $n$, no sentido em que

$$\sum_{k \text{ t.q. } a < \frac{k-n/2}{\sqrt{n/4}} \leq b} \text{prob}\,(k \text{ caras em } n \text{ moedas}) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

quando $n \to \infty$.

**Flutuações da marcha aleatória simétrica.** Uma interpretação interessante da experiência das moedas é a "marcha aleatória". O jogo consiste em passear pelos números inteiros dependendo dos resultados de lançamentos sucessivos de uma moeda honesta. A posição inicial no tempo 0 é $T_0 = 0$. Se estamos na posição $T_n$ no tempo $n$, a nossa posição $T_{n+1}$ no tempo $n+1$ será $T_n + 1$ ou $T_n - 1$ dependendo se a $(n+1)$-ésima moeda lançada mostra cara ou coroa, respectivamente. Se pensar um bocado, isto equivale a dizer que a posição $T_n$ no tempo $n$ é igual à diferença entre o número de caras e o número de coroas obtidas nos primeiros $n$ lançamentos, e portanto $T_n = S_n - (n - S_n)$. Também podemos pensar em $T_n$ como sendo o dinheiro que está a ganhar ou perder um jogador que aposta repetidamente um euro num jogo honesto.

O que é possível dizer acerca das trajectórias $n \mapsto T_n$ da marcha aleatória? A nossa melhor aposta para $S_n$ é $n/2$, logo a nossa melhor aposta para $T_n$ é zero. Também gostamos de afirmar isto dizendo que "a média" de $T_n$ é zero, a notação dos físicos sendo

$$\langle T_n \rangle = 0$$

Isto só diz que $T_n$ assume cada par de valores $\pm k$ com igual probabilidade. Também, a nossa melhor aposta para $S_n/n$ é $1/2$, e portanto a nossa melhor aposta para $T_n/n$ é zero. Logo, esperamos que o módulo de $T_n$ seja muito menor que $n$. Mas, quanto menor? Para o descobrir, uma boa estratégia é calcular a média do quadrado de $T_n$. Sabemos que $T_{n+1} = T_n \pm 1$, onde escolhemos $+$ ou $-$ dependendo do resultado da última moeda lançada. Ao fazer o quadrado, temos que

$$T_{n+1}^2 = T_n^2 \pm 2T_n + 1$$

Sendo as duas possibilidades acima equiprováveis, seja qual for que a nossa definição de "média" é natural esperar que

$$\langle T_{n+1}^2 \rangle = \langle T_n^2 \rangle + 1$$

Portanto, a média do quadrado da posição da marcha aleatória cresce de uma unidade em cada passo. Sendo obviamente $\langle T_1^2 \rangle = 1$, o resultado é que

$$\langle T_n^2 \rangle = n$$

e podemos dizer que, "em média", o módulo de $T_n$ é

$$|T_n| \sim \sqrt{n}$$

Ou seja, as trajectóriarias da marcha aleatória oscilam à volta de 0, e as oscilações são da ordem de $\sqrt{n}$. Em termos da frequência de caras redescobrimos a conjectura de que

$$S_n/n \sim 1/2 \pm 1/2\sqrt{n}$$

**Lei dos grandes números.** Se $n$ é grande, e os nossos instrumentos não são tão precisos para detectar um erro da ordem de $1/\sqrt{n}$, temos o direito de acreditar que

$$S_n/n \sim 1/2$$

quase certamente. Esta afirmação pode ser formalizada e é chamada "lei dos grandes números". E' o que um probabílista entende ao dizer que $1/2$ é a probabilidade de obter cara lançando uma moeda honesta.

**Intervalos de confiança.** Outra maneira de ler a nossa estimação é

$$1/2 \sim S_n/n \pm 1/2\sqrt{n}$$

Ou seja, podemos "estimar" a "probabilidade de obter cara" com a frequência observada $S_n/n$, uma vez que nos lembramos que a precisão da nossa estimação não pode ser melhor do que algo da ordem $1/\sqrt{n}$. Os estatísticos chamam isto "intervalos de confiança".

# 2 Espaços de probabilidades

**Eventos.** Seja $\Omega$ um conjunto não vazio. Uma família $\mathcal{E}$ de subconjuntos de $\Omega$ é uma *$\sigma$-álgebra* (ou *tribo*) se

*i)* $\emptyset \in \mathcal{E}$ e $\Omega \in \mathcal{E}$

*ii)* é *estável para passagem ao complementar*, ou seja se $A \in \mathcal{E}$ então $\Omega \backslash A \in \mathcal{E}$

*iii)* é *estável para reuniões e interseções enumeráveis*, ou seja se $(A_n)$ é uma família enumerável de elementos de $\mathcal{E}$ então

$$\cup_n A_n \in \mathcal{E} \qquad \text{e} \qquad \cap_n A_n \in \mathcal{E}$$

Observe que os três axiomas acima não são independentes. Por exemplo, $\emptyset \in \mathcal{E}$ e ii) implicam que $\Omega \in \mathcal{E}$. Também, dados i) e ii), a estabilidade para reuniões enumeráveis é equivalente à estabilidade para interseções enumeráveis.

Um par $(\Omega, \mathcal{E})$, formado por um conjunto não vazio $\Omega$ e uma $\sigma$-álgebra $\mathcal{E}$ de partes de $\Omega$, é chamado *espaço mensurável* (i.e. espaço onde é possível definir uma medida). Os elementos de $\mathcal{E}$ são ditos *conjuntos mensuráveis* (i.e. conjuntos que é possível medir), ou *eventos* no calão dos probabilistas.

**Exemplos.** $\{\emptyset, \Omega\}$ é a $\sigma$-álgebra trivial.

$\mathcal{P}(\Omega) = 2^\Omega = \{\text{subconjuntos de } \Omega\}$ é a maior das $\sigma$-álgebras de partes de $\Omega$.

**Experiências e álgebras.** No dialecto dos probabilistas, $\Omega$ representa o espaço dos estados de um sistema físico. Ao fazer uma experiência, o que fazemos é medir "observáveis", funções $\xi : \Omega \to \mathbf{R}$. A experiência mais simples é decidir se o estado do sistema satisfaz ou não uma certa propriedade, definida por meio de un certo número de observáveis. A esta propriedade está associado um subconjunto $A \subset \Omega$, e portanto a experiência consiste em decidir se $\omega \in A$ ou se $\omega \in \Omega \backslash A$, se "o evento $A$ aconteceu ou não". Ao fazer mais experiências deste tipo, por exemplo observando os eventos $A, B, C...$ , os conectores lógicos "e" e "ou" permitem obter informações acerca dos eventos $A \cap B$, $A \cup B$, $A \backslash B = A \cap (\Omega \backslash B)$, $A \cup B \cup C$ ... etc.

Uma família $\mathcal{A}$ de subconjuntos de $\Omega$, fechada com respeito às operações binárias $\cap$, $\cup$ e $\backslash$, e que contém os elementos neutros $\emptyset$ e $\Omega$, é dita *álgebra* (ou *álgebra de Boole*). É imediato verificar que uma álgebra é uma família $\mathcal{A}$ que satisfaz os axiomas

*i)* $\emptyset \in \mathcal{A}$ e $\Omega \in \mathcal{A}$

*ii)* se $A \in \mathcal{A}$ então $\Omega \backslash A \in \mathcal{A}$

*iii')* é *estável para reuniões e interseções finitas*, ou seja se $A$ e $B$ são elementos de $\mathcal{A}$ então também $A \cup B$ e $A \cap B$ são elementos de $\mathcal{A}$

O axioma iii') é, em geral, mais fraco do que o iii), a não ser que a família seja finita.

**Álgebras e partições.** Um exemplo simples de álgebra não trivial é $\{\emptyset, A, \Omega \backslash A, \Omega\}$, dita álgebra gerada por $A \subset \Omega$. Outros exemplos, de facto todos os exemplos finitos, são obtidos observando que as álgebras finitas estão em correspondência biunívoca com as partições finitas de $\Omega$.

Uma *partição* (ou *decomposição*) finita de $\Omega$ é uma família finita $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ de subconjuntos não vazios, ditos átomos, dois a dois disjuntos (i.e. tais que $D_i \cap D_j = \emptyset$ se $i \neq j$) e tais que

$$\Omega = D_1 \cup D_2 \cup ... \cup D_n$$

Dada uma partição finita $\mathcal{D}$, a família $\alpha(\mathcal{D})$, formada pelas reuniões de elementos de $\mathcal{D}$ (e pela reunião vazia), é uma álgebra, dita a *álgebra gerada por* $\mathcal{D}$. Por outro lado, se $\mathcal{A} \subset \mathcal{P}(\Omega)$ é uma álgebra finita, então existe uma partição finita $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ de $\Omega$ tal que $\mathcal{A} = \alpha(\mathcal{D})$.

**Álgebras e $\sigma$-álgebras.** Toda $\sigma$-álgebra é uma álgebra, e, por razões triviais, toda álgebra finita é uma $\sigma$-álgebra. É importante observar que uma $\sigma$-álgebra pode ser pensada como uma álgebra que é "completa" com respeito a uma operação natural de limite em teoria dos conjuntos.

Uma sucessão $(A_n)$ de subconjuntos de $\Omega$ é dita *crescente* se $... \subset A_n \subset A_{n+1} \subset ...$ , e *decrescente* se $... \supset A_{n+1} \supset A_n \supset ....$ É uma boa ideia utilizar a notação $A_n \uparrow A$ para dizer que o conjunto $A$ é igual à reunião $\cup_n A_n$ dos elementos da sucessão crescente $(A_n)$, e a notação $A_n \downarrow A$

para dizer que o conjunto $A$ é igual à interseção $\cap_n A_n$ dos elementos da sucessão decrescente $(A_n)$. Nos dois casos, o conjunto $A$ é dito *limite* da sucessão monótona (i.e. crescente ou decrescente) $(A_n)$.

Uma álgebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ é uma $\sigma$-álgebra sse, dada uma sucessão monótona $(A_n)$ de seus elementos tal que $A_n \uparrow A$ ou $A_n \downarrow A$, então também $A \in \mathcal{A}$. Uma implicação é trivial. Para provar a outra basta observar que, se $(B_n)$ é uma família enumerável de elementos de $\mathcal{A}$, então $\cup_n B_n$ é o limite da sucessão crescente de elementos de $\mathcal{A}$ definidos por $A_n = \cup_{k=1}^n B_n$, e que $\cap_n B_n$ é o limite da sucessão decrescente de elementos de $\mathcal{A}$ definidos por $A_n = \cap_{k=1}^n B_n$.

**Construção de $\sigma$-álgebras.** A possibilidade de "construir" $\sigma$-álgebras não triviais, e depois de "verificar" se determinados subconjuntos pertencem à determinadas $\sigma$-álgebras, depende das seguintes definições e observações.

Seja $\mathcal{C}$ uma família de partes de $\Omega$. A $\sigma$-álgebra *gerada por* $\mathcal{C}$, denotada por $\sigma(\mathcal{C})$, é a "menor" $\sigma$-álgebra que contém $\mathcal{C}$, ou seja a interseção de todas as $\sigma$-álgebras que contêm $\mathcal{C}$. Ela é bem definida, porque $\mathcal{P}(\Omega) \supset \mathcal{C}$, logo existe pelo menos uma $\sigma$-álgebra que contém $\mathcal{C}$, e porque a interseção de uma família de $\sigma$-álgebras é uma $\sigma$-álgebra.

Seja $\mathcal{E}$ uma $\sigma$-álgebra de partes de $\Omega$. Dada uma função $f : \Omega' \to \Omega$, a família

$$f^{-1}\mathcal{E} = \left\{ f^{-1}(A) \text{ com } A \in \mathcal{E} \right\}$$

é uma $\sigma$-álgebra de partes de $\Omega'$, dita *imagem inversa* ("pull-back") de $\mathcal{E}$ pela aplicação $f$. Isto acontece porque a função $f^{-1}$ entre as partes de $\Omega$ e as partes de $\Omega'$ comuta com as operações "interseção", "reunião" e "complementar".

Em particular, se $\Omega' \subset \Omega$ e $\mathcal{E}$ é uma $\sigma$-álgebra de partes de $\Omega$, então a família $\mathcal{E}' = \{A \cap \Omega' \text{ com } A \in \mathcal{E}\}$ é uma $\sigma$-álgebra de partes de $\Omega'$, dita $\sigma$-álgebra *traço* (observe que $\mathcal{E}' = i^{-1}\mathcal{E}$, onde $i$ denota a injeção $\Omega' \hookrightarrow \Omega$).

Outro caso interessante é quando a função $f$ é uma projeção $\pi : X \times Y \to X$, definida por $(x, y) \mapsto x$. Neste caso, se $\mathcal{E}$ uma $\sigma$-álgebra de partes de $X$, a imagem inversa $\pi^{-1}\mathcal{E}$ é uma $\sigma$-álgebra de partes do produto cartesiano $X \times Y$ composta por subconjuntos que "só dependem da primeira componente", pois são da forma $A \times Y$ com $A \subset X$.

É também possível puxar $\sigma$-álgebras para frente. Sejam $\mathcal{E}$ uma $\sigma$-álgebra de partes de $\Omega$, e $f : \Omega \to \Omega'$. A família

$$f\mathcal{E} = \left\{ A' \subset \Omega' \text{ t.q. existe } A \in \mathcal{E} \text{ t.q. } f^{-1}(A') = A \right\}$$

é uma $\sigma$-álgebra de partes de $\Omega'$, dita *imagem direta* ("push-forward") de $\mathcal{E}$ pela aplicação $f$.

A seguinte observação é conhecida como

**Lema do transporte** . *Se $f : \Omega' \to \Omega$ é uma função e $\mathcal{C}$ é uma família de partes de $\Omega$, então*

$$\sigma\left(f^{-1}\mathcal{C}\right) = f^{-1}\sigma(\mathcal{C})$$

**dem.** A inclusão $\sigma\left(f^{-1}\mathcal{C}\right) \subset f^{-1}\sigma(\mathcal{C})$ é obvia, sendo $f^{-1}\sigma(\mathcal{C})$ uma $\sigma$-álgebra que contém $f^{-1}\mathcal{C}$. Por outro lado, seja $\mathcal{E} = \left\{A \in \sigma(\mathcal{C}) \text{ t.q. } f^{-1}(A) \in \sigma\left(f^{-1}\mathcal{C}\right)\right\}$. É imediato verificar que $\mathcal{E}$ é uma $\sigma$-álgebra e que contém $\mathcal{C}$. As inclusões $\mathcal{C} \subset \mathcal{E} \subset \sigma(\mathcal{C})$ implicam que $\mathcal{E} = \sigma(\mathcal{C})$, donde $f^{-1}\sigma(\mathcal{C}) \subset \sigma\left(f^{-1}\mathcal{C}\right)$. □

**Exercícios.**

**a.** Prove que a interseção de uma família de $\sigma$-álgebras é uma $\sigma$-álgebra.

**b.** Dê exemplos de

$\sigma$-álgebras de partes de $\Omega$, onde $\Omega = \{a, b, c, d\}$ ou $\mathbf{N}$ ou $\mathbf{R}$.

**c.** Determine a cardinalidade da $\sigma$-álgebra $\mathcal{P}(\Omega)$ quando $\Omega$ é um conjunto finito ou um conjunto enumerável. Determine a $\sigma$-álgebra $\sigma(\mathcal{D})$ e a sua cardinalidade, quando $\mathcal{D} = \{D_1, D_2, ..., D_n, ...\}$ é uma partição enumerável de um conjunto $\Omega$. Existe uma $\sigma$-álgebra cuja cardinalidade é igual a cardinalidade de $\mathbf{N}$?

**Boreleanos.** Os físicos gostam de utilizar a recta real como modelo dos possíveis valores de observáveis físicos, pela simples razão de que é na recta real que podem fazer "análise" e portanto "calcular" as previsões dos próprios modelos. A métrica euclidiana na recta real (ou melhor a topologia induzida) convida à escolha de uma $\sigma$-álgebra particularmente significativa que contém os intervalos, os nossos subconjuntos preferidos.

Seja $\Omega$ um intervalo da recta real $\mathbf{R}$, munido da topologia standard. A $\sigma$-álgebra dos *boreleanos* de $\Omega$ é definida como a menor $\sigma$-álgebra que contém todos os subconjuntos abertos (e portanto os fechados, e reuniões e interseções enumeráveis de abertos e fechados, etc...) de $\Omega$, e denotada por $\mathcal{B}(\Omega)$.

Todo aberto da recta real é uma reunião enumerável de intervalos abertos dois a dois disjuntos, portanto $\mathcal{B}(\mathbf{R})$ é também a $\sigma$-álgebra gerada pela família dos intervalos abertos. De facto, e esta observação será útil a seguir, $\mathcal{B}(\mathbf{R})$ é a $\sigma$-álgebra gerada pela família de intervalos

$$\{]-\infty, t] \text{ com } t \in \mathbf{Q}\}$$

A prova consiste em mostrar que todo intervalo aberto pode ser obtido a partir de elementos desta família utilizando as operações de complementar, reunião e interseção enumerável. Por exemplo,

$$]a, b[ = \cup_{n=1}^{\infty} (]-\infty, a_n[^c \cap ]-\infty, b_n[)$$

onde $(a_n)$ e $(b_n)$ são sucessões estritamente monótonas de racionais tais que $a_n \downarrow a$ e $b_n \uparrow b$.

Em geral, seja $(\Omega, \tau)$ um espaço topológico (i.e. $\Omega$ é um conjunto não vazio e $\tau$ uma topologia definida em $\Omega$, uma coleção de subconjuntos de $\Omega$, ditos abertos, que contém $\emptyset$ e $\Omega$ e é estável para interseções finitas e reuniões enumeráveis). A $\sigma$-álgebra dos *boreleanos* de $\Omega$ é $\mathcal{B}(\Omega) = \sigma(\tau)$, definida como a menor $\sigma$-álgebra que contém todos os subconjuntos abertos. Se a família $\mathcal{C} \subset \mathcal{P}(\Omega)$ é uma base enumerável da topologia $\tau$ (i.e. se todo aberto $A \in \tau$ é uma reunião de elementos de $\mathcal{C}$), então $\sigma(\tau) = \sigma(\mathcal{C})$.

**Boreleanos em espaços produto.** Sejam $(X_\alpha, \tau_\alpha)$ espaços topológicos, com $\alpha \in \mathcal{I}$, e seja

$$\Omega = \prod_{\alpha \in \mathcal{I}} X_\alpha = \{x : \mathcal{I} \to \cup_{\alpha \in \mathcal{I}} X_\alpha \text{ t.q. } x_\alpha \in X_\alpha \text{ para todo } \alpha \in \mathcal{I}\}$$

o produto cartesiano dos $X_\alpha$ (onde utilizamos a notação $x_\alpha = x(\alpha)$ para a "coordenada" $\alpha$-ésima do ponto $x$). Um *cilindro aberto* de $\Omega$ é um conjunto $C$ formado da seguinte maneira: existem um conjunto finito de índices $\alpha_1, \alpha_2, ..., \alpha_n \in \mathcal{I}$ e uns abertos $C_{\alpha_i} \subset X_{\alpha_i}$ com $i = 1, 2, ..., n$ tais que

$$C = \{x = (x_\alpha)_{\alpha \in \mathcal{I}} \in X \text{ tais que } x_{\alpha_i} \in C_{\alpha_i} \text{ para todo } i = 1, 2, ..., n\}$$

A família $\mathcal{C}$, formada pelos cilindros abertos de $\Omega$, é uma base de uma topologia $\tau$ em $\Omega$, dita *topologia produto*. A $\sigma$-álgebra dos *boreleanos* de $\Omega$ é $\mathcal{B}(\Omega) = \sigma(\tau)$.

Por exemplo, a topologia produto em $\mathbf{R}^n$, onde cada cópia de $\mathbf{R}$ é munida da topologia standard, é equivalente à topologia euclidiana (pois a norma do supremo, que gera a topologia produto, é equivalente à norma euclidiana), e portanto a $\sigma$-álgebra dos boreleanos de $\mathbf{R}^n$ é também a $\sigma$-álgebra gerada pelos cilindros abertos.

Particularmente interessantes em probabilidades são espaços produto do tipo $\Omega = X^{\mathbf{N}}$ ou $X^{\mathbf{R}_{\geq 0}}$, que representam "provas repetidas" duma mesma esperiência descrita pelo espaço dos estados $X$, onde o parâmetro $\alpha \in \mathbf{N}$ ou $\mathbf{R}_{\geq 0}$ tem a interpretação de um "tempo". Tipicamente, $X$ é um conjunto finito, um conjunto enumerável, ou a recta real.

**Medidas de probabilidades.** Sejam $\Omega$ um conjunto não vazio e $\mathcal{E}$ uma $\sigma$-álgebra de partes de $\Omega$. Uma *probabilidade* (ou *medida de probabilidades*) no espaço mensurável $(\Omega, \mathcal{E})$ é uma função $\mathbf{P} : \mathcal{E} \to [0, 1]$ tal que

*i)* $\mathbf{P}(\Omega) = 1$ e $\mathbf{P}(\emptyset) = 0$

*ii)* é *σ-aditiva*, ou seja se $(A_n)$ é uma família enumerável de elementos de $\mathcal{E}$ dois a dois disjuntos então

$$\mathbf{P}\left(\cup_n A_n\right) = \sum_n \mathbf{P}\left(A_n\right)$$

Observe que a σ-aditividade implica a *aditividade* (finita): se $A_1$, $A_2$, ..., $A_n$ são elementos de $\mathcal{E}$ dois a dois disjuntos, então

$$\mathbf{P}\left(A_1 \cup A_2 \cup ... \cup A_n\right) = \mathbf{P}\left(A_1\right) + \mathbf{P}\left(A_2\right) + ... + \mathbf{P}\left(A_n\right)$$

(basta pôr $A_k = \emptyset$ para todo $k > n$ no axioma que define a σ-aditividade).

**Probabilidades em espaços enumeráveis.** Sejam $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ uma partição finita de $\Omega$, e $\mathcal{A} = \alpha\left(\mathcal{D}\right)$ a álgebra gerada por $\mathcal{D}$. Definir uma probabilidade em $\mathcal{A}$ é equivalente a escolher uma coleção $p_1, p_2, ...p_n$ de números $\geq 0$ tais que $p_1 + p_2 + ... + p_n = 1$ e declarar que $\mathbf{P}\left(D_k\right) = p_k$. Pois, cada $A \in \mathcal{A}$ é uma reunião disjunta de elementos de $\mathcal{D}$, e a aditividade determina o valor de $\mathbf{P}$ em $A$.

Se $\mathcal{D} = \{D_1, D_2, ..., D_n, ...\}$ é uma partição enumerável de $\Omega$, é também fácil definir uma probabilidade sobre a σ-álgebra $\sigma\left(\mathcal{D}\right)$: qualquer série convergente com termos não negativos e soma um. Pois, se $(p_n)$ é uma sucessão de números $\geq 0$ tais que $\sum_n p_n = 1$, então a função $\mathbf{P} : \sigma\left(\mathcal{D}\right) \to [0,1]$ definida por

$$\mathbf{P}\left(A\right) = \sum_{D_n \subset A} p_n$$

é uma probabilidade.

Um caso particular é quando $\Omega$ é um conjunto finito ou enumerável. Se $\Omega = \{\omega_1, \omega_2, \omega_3, ...\}$ e $p_1, p_2, p_3, ...$ são números $\geq 0$ tais que $\sum_n p_n = 1$, então a função $\mathbf{P} : \mathcal{P}\left(\Omega\right) \to [0,1]$ definida por

$$\mathbf{P}\left(A\right) = \sum_{\omega_k \in A} p_k$$

é uma probabilidade sobre as partes de $\Omega$. Por outras palavras, uma probabilidades nas partes de um espaço enumerável é definida fixando "a probabilidade" $p_k = \mathbf{P}\left(\{\omega_k\}\right)$ de cada um dos seus pontos.

**Delta de Dirac.** Se $\Omega$ é um conjunto e $x \in \Omega$, então a função $\delta_x : \mathcal{P}\left(\Omega\right) \to [0,1]$, definida por

$$\delta_x\left(A\right) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases}$$

é uma probabilidade sobre as partes de $\Omega$, dita *delta de Dirac* em $x$.

**Combinações convexas de medidas de probabilidades.** O espaço das medidas de probabilidades definidas sobre uma σ-álgebra $\mathcal{E}$ é um convexo: se $\mathbf{P}_0$ e $\mathbf{P}_1$ são medidas de probabilidades e $t \in [0,1]$, então também $\mathbf{P}_t = (1-t) \cdot \mathbf{P}_0 + t \cdot \mathbf{P}_1$, definida por

$$\mathbf{P}_t\left(A\right) = (1-t) \cdot \mathbf{P}_0\left(A\right) + t \cdot \mathbf{P}_1\left(A\right)$$

é uma medida de probabilidades. Também, dada uma família enumerável $(\mathbf{P}_n)$ de probabilidades definidas sobre $\mathcal{E}$ e uma sucessão $(p_n)$ de números $\geq 0$ tais que $\sum p_n = 1$, então $\mathbf{P} = \sum_n p_n \cdot \mathbf{P}_n$ é uma probabilidade sobre $\mathcal{E}$.

Por exemplo, seja $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_n\}$ um conjunto finito. O espaço das medidas de probabilidades definidas em $\mathcal{P}\left(\Omega\right)$ é naturalmente isomorfo ao simplexo $\Delta^n \subset \mathbf{R}^n$, definido por

$$\Delta^n = \left\{p = (p_1, p_2, ..., p_n) \in \mathbf{R}^n_{\geq 0} \text{ t.q. } p_1 + p_2 + ... + p_n = 1\right\}$$

Ou seja, toda medida de probabilidades é uma combinação convexa $\mathbf{P} = \sum_k p_n \delta_{\omega_k}$ de delta de Dirac nos pontos de $\Omega$, definida por

$$\mathbf{P}\left(A\right) = \sum_k p_k \delta_{\omega_k}\left(A\right) = \sum_{\omega_k \in A} p_k$$

**Medida de Lebesgue.** Probabilidades mais "interessantes" do que combinações convexas de delta de Dirac na recta real só podem ser definidas em $\sigma$-álgebras estritamente contidas em $\mathcal{P}(\mathbf{R})$, como a $\sigma$-álgebra dos boreleanos.

Existe uma única medida $\ell : \mathcal{B}(\mathbf{R}) \to [0,\infty]$ (i.e. uma função $\sigma$-aditiva tal que $\ell(\emptyset) = 0$), dita *medida de Lebesgue*, definida sobre os boreleanos da recta real tal que $\ell([a,b]) = |b-a|$ para todo intervalo $[a,b]$. Em particular, se $\Omega = [a,b] \subset \mathbf{R}$, existe uma única probabilidade $\mathbf{P}$ definida sobre os boreleanos $\mathcal{B}(\Omega)$ tal que $\mathbf{P}([s,t]) = |t-s| / |b-a|$ para todos os intervalos $[s,t] \subset [a,b]$.

**Funções mensuráveis e medida imágem.** Sendo funções, as medidas podem ser compostas. Isto fornece um método para construir medidas a partir de outras medidas.

Sejam $(\Omega, \mathcal{E})$ e $(\Omega', \mathcal{F})$ dois espaços mensuráveis. Uma aplicação $f : \Omega \to \Omega'$ é dita *mensurável* se $f^{-1}(A) \in \mathcal{E}$ para todo $A \in \mathcal{F}$, ou seja se $f^{-1}\mathcal{F} \subset \mathcal{E}$. Assim como as funções contínuas entre espaços topológicos são as aplicações que preservam os abertos, as aplicações mensuráveis são as aplicações que preservam os conjuntos mensuráveis.

Um critério para decidir se uma aplicação é mensurável é o seguinte: se a $\sigma$-álgebra $\mathcal{F}$ é gerada pela família $\mathcal{C}$, o lema do transporte implica que $f$ é mensurável sse $f^{-1}(C) \in \mathcal{E}$ para todo $C \in \mathcal{C}$.

Um caso particularmente importante é quando o contradomínio da aplicação é a recta real munida da $\sigma$-álgebra dos boreleanos. Uma aplicação $f : \Omega \to \mathbf{R}$, definida no espaços mensurável $(\Omega, \mathcal{E})$, é dita *Borel-mensurável* se $f^{-1}(B) \in \mathcal{E}$ para todo boreleano $B \in \mathcal{B}(\mathbf{R})$. O lema do transporte implica que é suficiente verificar que $f^{-1}(C) \in \mathcal{E}$ para todo elemento $C$ de uma família $\mathcal{C}$ que gera a $\sigma$-álgebra dos boreleanos (como a família dos abertos, a família dos intervalos, a família dos intervalos do tipo $]-\infty, t]$ com $t \in \mathbf{Q}$, ...).

Sejam $f : \Omega \to \Omega'$ uma aplicação mensurável entre os espaços mensuráveis $(\Omega, \mathcal{E})$ e $(\Omega', \mathcal{F})$, e $\mathbf{P} : \mathcal{E} \to [0,1]$ uma medida de probabilidades. A *medida imagem* de $\mathbf{P}$ pela aplicação $f$ é a função $\mathbf{P}' : \mathcal{F} \to [0,1]$ definida por $\mathbf{P}'(A) = \mathbf{P}\left(f^{-1}A\right)$ se $A \in \mathcal{F}$, i.e. é a função composta $\mathbf{P} \circ f^{-1}|_{\mathcal{F}}$. É imediato verificar que a medida imagem é uma medida de probabilidades.

**Espaços de probabilidades.** Um *espaço de probabilidades*, i.e. um modelo matemático de um fenómeno aleatório, é um terno $(\Omega, \mathcal{E}, \mathbf{P})$: um espaço dos *estados* (ou acontecimentos elementares) $\Omega$, uma $\sigma$-álgebra $\mathcal{E}$ de partes de $\Omega$, cujos elementos são ditos *eventos*, e uma *probabilidade* $\mathbf{P} : \mathcal{E} \to [0,1]$ definida sobre os eventos.

Se $A \in \mathcal{E}$, o número $\mathbf{P}(A)$ é chamado *probabilidade do evento* $A$.

As operações $\cap$, $\cup$, $\cdot^c$ e $\cdot\backslash\cdot$, assim como a relação binária $\subset$, têm interpretações naturais em termos de acontecimentos. $\Omega$ é o "evento certo", cuja probabilidade é 1, e $\emptyset$ é o "evento impossível", cuja probabilidade é 0. A interseção $A \cap B$ é o evento "aconteceram seja $A$ seja $B$". A reunião $A \cup B$ é o evento "aconteceu $A$ ou $B$". O complementar $A^c = \Omega\backslash A$ é o evento "não aconteceu $A$". A diferença $A\backslash B = A \cap B^c$ é o evento "aconteceu $A$ e não aconteceu $B$". A diferença simétrica $A \Delta B = (A\backslash B) \cup (B\backslash A)$ é o evento "aconteceu um e só um dos eventos $A$ e $B$". A inclusão $A \subset B$ quer dizer que a ocorrência do evento $A$ implica a ocorrência do evento $B$.

Particularmente significativas são afirmações do género "bla bla acontece com probabilidade um", o que quer dizer que o evento $A$, associado à descrição "bla bla", tem probabilidade $\mathbf{P}(A) = 1$. Um evento pode ter probabilidade 1 sem ser o evento certo, ou ter probabilidade 0 sem ser "impossível": em espaços de probabilidades não enumeráveis é natural acontecer que todos os pontos $\omega \in \Omega$ tenham probabilidade $\mathbf{P}(\{\omega\}) = 0$.

**obs.** Naturalmente, é possível fazer uma teoria elementar considerando só conjuntos finitos, e portanto medidas de probabilidades definidas sobre álgebras finitas (um exemplo muito bem escrito é o capítulo 1 do manual de Shiryaev). Infelizmente, isso não permite tratar com elegância fenómenos simples como o lançamento de uma moeda até sair cara pela primeira vez, ou o problema da ruína do jogador... e, sobretudo, isso torna mais complicado o formalismo e mais obscura a interpretação dos resultados. O preço a pagar em considerar medidas de probabilidades que satisfazem os axiomas de Kolmogorov é, por outro lado, alto. A existência e a construção de tais medidas são problemas técnicos nada triviais, dos quais trata a "teoria da medida". Um esboço da teoria da medida necessária para fazer probabilidades, sem muitas demonstrações, está mais à frente. Umas referências são os clássicos de Billingsley [Bi79], Doob [Do94], Halmos [Ha74] ou Rudin [Rud66].

**Propriedades elementares.** Propriedades elementares das medidas de probabilidades são as seguintes. Sejam $A$, $B$, $A_n$ com $n$ inteiro, eventos no espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$. Então

$$\begin{aligned}
\mathbf{P}(A^c) &= 1 - \mathbf{P}(A) \\
\mathbf{P}(\cup_n A_n) &= 1 - \mathbf{P}(\cap_n A_n^c) \\
\mathbf{P}(A) &= \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c) \text{ (\textit{fórmula da probabilidade total})} \\
\mathbf{P}(A) &\leq \mathbf{P}(B) \text{ se } A \subset B \text{ (\textit{monotonia})} \\
\mathbf{P}\left(\cup_n A_n\right) &\leq \sum_n \mathbf{P}\left(A_n\right) \text{ (\textit{$\sigma$-subaditividade})}
\end{aligned}$$

De facto, a primeira vem da normalização $\mathbf{P}(\Omega) = 1$ e da aditividade, observando que $\Omega = A \cup A^c$ com $A$ e $A^c$ disjuntos. A segunda vem da primeira e da fórmula de De Morgan $(\cup_n A_n)^c = \cap_n A_n^c$. A fórmula da probabilidade total vem da observação que $B \cup B^c = \Omega$ e $B$ e $B^c$ são disjuntos, e portanto $A$ é a reunião disjunta de $A \cap B$ e $A \cap B^c$. A monotonia vem de $\mathbf{P}(B) = \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) \geq \mathbf{P}(A)$, porque as probabilidades são não-negativas. A $\sigma$-subaditividade vem da seguinte observação: definidos os eventos $B_n = A_n \setminus \left(\cup_{k=1}^{n-1} A_k\right)$, ve-se que os $B_n$ são dois a dois disjuntos, que $\cup_n A_n = \cup_n B_n$ e que $\mathbf{P}\left(B_n\right) \leq \mathbf{P}\left(A_n\right)$ porque $B_n \subset A_n$, portanto a $\sigma$-aditividade e a monotonia implicam que $\mathbf{P}\left(\cup_n A_n\right) = \mathbf{P}\left(\cup_n B_n\right) = \sum_n \mathbf{P}\left(B_n\right) \leq \sum_n \mathbf{P}\left(A_n\right)$.

**Exercício.** Sejam $A$ e $B$ eventos do espaço de probabilidades $(\Omega, \mathcal{A}, \mathbf{P})$. Prove que:

$$\begin{aligned}
\mathbf{P}\left(A \cup B\right) &\geq \max\left\{\mathbf{P}\left(A\right), \mathbf{P}\left(B\right)\right\} \\
\mathbf{P}\left(A \cap B\right) &\leq \min\left\{\mathbf{P}\left(A\right), \mathbf{P}\left(B\right)\right\} \\
\mathbf{P}(A \cup B) &= \mathbf{P}(A) + \mathbf{P}(B \cap A^c) \\
\mathbf{P}(A \cup B) &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \\
\mathbf{P}\left(A \Delta B\right) &= \mathbf{P}\left(A\right) + \mathbf{P}\left(B\right) - 2 \cdot \mathbf{P}\left(A \cap B\right)
\end{aligned}$$

**Continuidade.** A consequência importante da $\sigma$-aditividade é a continuidade da medida de probabilidades, a possibilidade de calcular a probabilidade de um limite de certas sucessões de eventos calculando o limite das probabilidades dos elementos da sucessão.

A medida de probabilidade é *contínua*, ou seja

$$\text{se } A_n \uparrow A \text{ ou } A_n \downarrow A \text{ então } \mathbf{P}\left(A\right) = \lim_{n \to \infty} \mathbf{P}\left(A_n\right)$$

A segunda afirmação vem da primeira considerando os eventos complementares, portanto só temos que provar a primeira, i.e. o caso em que $A_n \uparrow A$. Sejam $B_n$ os eventos definidos por $B_1 = A_1$ e $B_n = A_n \setminus A_{n-1}$ se $n > 1$. Eles são dois a dois disjuntos, e é imediato verificar que $A_n = \cup_{k=1}^n B_k$ e $\cup_n A_n = \cup_k B_k$. Usando a $\sigma$-aditividade temos enfím

$$\mathbf{P}\left(\cup_n A_n\right) = \mathbf{P}\left(\cup_n B_n\right) = \sum_{k=1}^{\infty} \mathbf{P}\left(B_k\right) = \lim_{n \to \infty} \sum_{k=1}^{n} \mathbf{P}\left(B_k\right) = \lim_{n \to \infty} \mathbf{P}\left(A_n\right)$$

**Liminf e limsup.** Se $(A_n)$ é uma sucessão de eventos, também são eventos

$$\underline{\lim} A_n = \{\omega \in_{\text{f.o.}} A_n^c\} = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \quad \text{e} \quad \overline{\lim} A_n = \{\omega \in_{\text{i.o.}} A_n\} = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$$

O evento $\overline{\lim} A_n$ (o limsup dos $A_n$) é o conjunto dos $\omega \in \Omega$ que pertecem a uma infinidade dos $A_n$. O evento $\underline{\lim} A_n$ (o liminf dos $A_n$) é o conjunto dos $\omega \in \Omega$ que pertecem a uma quantidade finita dos $A_n^c$. Em particular $\underline{\lim} A_n \subset \overline{\lim} A_n$ .

Se acontece que $\underline{\lim} A_n = \overline{\lim} A_n$, então este evento é dito *limite* da sucessão $(A_n)$, e denotado por $\lim A_n$. Observe que $A = \lim A_n$ sse

$$\lim_{n \to \infty} 1_{A_n}(\omega) \to 1_A(\omega)$$

para todo $\omega \in \Omega$, onde $1_{A_n}$ e $1_A$ acima denotam as funções características dos eventos $A_n$ e $A$, respectivamente.

A continuidade da medida de probabilidades pode ser enunciada da seguinte forma:

$$\text{se } A = \lim A_n \text{ então } \mathbf{P}\left(A\right) = \lim_{n \to \infty} \mathbf{P}\left(A_n\right).$$

**Continuidade e $\sigma$-aditividade.** A continuidade da medida de probabilidades é, de facto, equivalente à $\sigma$-aditividade da medida no caso do axioma *ii)* ser substituido pela aditividade (finita):

*ii')* se $A$ e $B$ são disjuntos então $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

A continuidade da medida de probabilidades é, dada a aditividade, equivalente à "continuidade em $\emptyset$": se $A_n \downarrow \emptyset$ então $\lim_{n\to\infty} \mathbf{P}(A_n) = 0$.

**Aproximação.** Exceto em casos triviais, é muito diffícil "exibir" medidas de probabilidades. A estratégia é definir "explicitamente" $\mathbf{P}$ numa família pequena de eventos particularmente significativos ou simples, e depois utilizar teoremas que permitem "extender" a medida à $\sigma$-álgebra dos eventos todos. Isto explica a importância, quer prática quer teórica, de ter resultados que permitam "aproximar" a probabilidade de eventos arbitrários por meio de probabilidades que sabemos "calcular".

Sejam $(\Omega, \mathcal{E}, \mathbf{P})$ um espaço de probabilidades, e $\mathcal{A}$ uma álgebra de subconjuntos de $\Omega$ que gera a $\sigma$-álgebra $\mathcal{E}$, i.e. tal que $\mathcal{E} = \sigma(\mathcal{A})$. Então todo evento pode ser aproximado em probabilidade por um elemento de $\mathcal{A}$, i.e. para todo $E \in \mathcal{E}$ e todo $\varepsilon > 0$ existe $A \in \mathcal{A}$ tal que

$$\mathbf{P}(E \Delta A) < \varepsilon$$

De facto, seja

$$\mathcal{C} = \{C \in \mathcal{E} \text{ t.q. } \forall \varepsilon > 0 \ \exists\ A \in \mathcal{A} \text{ t.q. } \mathbf{P}(C \Delta A) < \varepsilon\}.$$

É imediato verificar que $\mathcal{C}$ é uma $\sigma$-álgebra, e o facto obvio que $\mathcal{A} \subset \mathcal{C}$ implica que $\sigma(\mathcal{A}) \subset \mathcal{C} \subset \mathcal{E}$, donde $\mathcal{C} = \mathcal{E}$.

Seja $(\Omega, \tau)$ um espaço topológico metrizável, munido da $\sigma$-álgebra dos boreleanos $\mathcal{B}$. Toda medida de probabilidades $\mathbf{P}$ definida em $\mathcal{B}$ é *regular*, i.e. para todo $B \in \mathcal{B}$ e todo $\varepsilon > 0$ existem um fechado $F$ e um aberto $A$ tais que $F \subset E \subset A$ e

$$\mathbf{P}(A \backslash F) < \varepsilon$$

De facto, seja

$$\mathcal{C} = \{C \in \mathcal{B} \text{ t.q. } \forall \varepsilon > 0 \ \exists\ F \text{ fechado e } \exists\ A \text{ aberto t.q. } F \subset E \subset A \text{ e } \mathbf{P}(A \backslash F) < \varepsilon\}.$$

A primeira observação é que $\mathcal{C}$ contém a família dos fechados de $\Omega$. Pois, se $d$ é uma métrica que gera a topologia $\tau$, e se $F$ é fechado, então as $\delta$-vizinhanças $A_\delta = \{\omega \in \Omega \text{ t.q. } d(\omega, F)\}$ são abertos tais que $F \subset A_\delta$ e $\mathbf{P}(A_\delta) \downarrow \mathbf{P}(F)$ pela continuidade da medida de probabilidade. Como a família dos fechados gera a $\sigma$-álgebra $\mathcal{B}$, basta agora verificar que $\mathcal{C}$ é uma $\sigma$-álgebra. A validade dos axiomas i) e ii) é obvia. Seja $(C_n)$ uma família enumerável de elementos de $\mathcal{C}$, e seja $\varepsilon > 0$. Existem fechados $F_n$ e abertos $A_n$ tais que $F_n \subset C_n \subset A_n$ e $\mathbf{P}(A_n \backslash F_n) < \varepsilon / 2^{n+1}$. Se $\overline{n}$ é suficientemente grande, $\mathbf{P}((\cup_n F_n) \setminus (\cup_{n \le \overline{n}} F_n)) < \varepsilon/2$. Então o fechado $\cup_{n \le \overline{n}} F_n$ e o aberto $\cup_n A_n$ satisfazem $\cup_{n \le \overline{n}} F_n \subset \cup_n C_n \subset \cup_n A_n$ e $\mathbf{P}((\cup_n A_n) \setminus \cup_{n \le \overline{n}} F_n) < \varepsilon$, o que mostra que $\mathcal{C}$ é também estável para reuniões enumeráveis.

**Exemplo: prova de Bernoulli.** Um modelo dum jogo com probabilidade $p$ de ganhar é:

$$\Omega = \{0 = \text{"perder"}, 1 = \text{"ganhar"}\}\ , \qquad \mathcal{E} = \mathcal{P}(\Omega)\ , \qquad \mathbf{P}(\{0\}) = 1 - p \quad \text{e} \quad \mathbf{P}(\{1\}) = p.$$

O caso em que $p = 1/2$ pode ser pensado como um modelo da experiência "lançar uma moeda honesta".

**Exemplo: dado.** Um modelo da experiência "lançar um dado" é:

$$\Omega = \{1, 2, ..., 6\}\ , \qquad \mathcal{E} = \mathcal{P}(\Omega), \qquad \mathbf{P}(\{1\}) = \mathbf{P}(\{2\}) = ... = \mathbf{P}(\{6\}) = 1/6.$$

**Exemplo: tempo de decaimento.** Um modelo do tempo em que decai um núcleo de uma substância radiactiva é

$$\Omega = \mathbf{R}_{\geq 0} \ , \qquad \mathcal{E} = \mathcal{B}\left(\mathbf{R}_{\geq 0}\right) \ , \qquad \mathbf{P}\left([a,b]\right) = e^{-a/\tau} - e^{-b/\tau}$$

onde $\tau > 0$ é o tempo característico do decaimento.

**Espaços de probabilidades uniformes.** Se $\Omega$ é um conjunto finito, uma probabilidade natural sobre as suas partes é

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}$$

(a notação é: $|A|$ ="cardinalidade do conjunto $A$"), dita *probabilidade uniforme*. Numa linguagem familiar, a probabilidade do evento $A$ é igual a "cardinalidade dos casos favoráveis a dividir pela cardinalidade dos casos possíveis". Em particular, todo ponto $\omega \in \Omega$ tem probabilidade $\mathbf{P}\left(\{\omega\}\right) = 1/\left|\Omega\right|$ . Os conjuntos formados por um só ponto de um espaço de probabilidades são por vezes ditos "átomos". A probabilidade uniforme num espaço finito é, portanto, definida pela condição: todos os átomos têm a mesma probabilidade.

**Fazer modelos.** É tradição pôr problemas de probabilidades em palavras da linguagem do dia a dia, como "numa aldeia vivem $n+1$ velhinhas. Uma velhinha inventa uma fofoca e conta-a a outra, escolhendo ao acaso entre as $n$ restantes, que por sua vez repete-a a uma terceira, também escolhendo ao acaso entre as $n$ restantes, etc... Calcule a probabilidade de a fofoca ser contada $k$ vezes sem voltar a ser contada à velhinha que a inventou, e sem ninguém a ouvir duas vezes."

A resposta consiste em fazer um modelo da experiência e calcular a probabilidade do evento dentro do modelo. O modelo preferido, nos casos em que o espaço dos acontecimentos é finito, é um espaço de probabilidades uniforme (simplesmente porque é a probabilidade mais "democrática"). Se a situação é pouco clara, ou o espaço dos acontecimentos não é finito, fazer um modelo precisa de mais cuidado e de considerações "físicas". Não faz muito sentido querer refutar um modelo com base em considerações teóricas. Decidir se um modelo descreve adequatamente uma experiência real é um problema ao qual só e possível dar respostas empíricas, e isto é um dos objectivos da estatística (ou, em geral, da física). Nas palavras de Doob: "Finally, it is important to keep mathematics and real life apart. It is an interesting facet of human behaviour that, even when actual coin tossing is analyzed, the analysis has almost always been philosophical, ignoring the laws of mechanics, which quite unphilisophically govern the motion of real-world coins, under initial conditions imposed by real-world humans, and thereafter subject to the laws of motion of a real body falling under the influence of real gravity. The point is that the impossible-to-make-precise description of the actual result of coin tossing has a precise mathematical counterpart, in which mathematical theorems can be proved, some of which suggest real-world observational results."

**Exemplo: fofocas.** Numa aldeia vivem $n+1$ velhinhas. Uma velhinha inventa uma fofoca e conta-a a outra, escolhendo ao acaso entre as $n$ restantes, que por sua vez repete-a a uma terceira, também escolhendo ao acaso entre as $n$ restantes, etc... Calcule a probabilidade de a fofoca ser contada $k$ vezes sem voltar a ser contada à velhinha que a inventou, e sem ninguém a ouvir duas vezes.

Seja $X = \{0, 1, 2, ..., n\}$ o conjunto das velhinhas, e seja 0 a velhinha que inventou a fofoca. O espaço dos possíveis acontecimentos é o espaço $\Omega$ dos caminhos $\omega : \{0, 1, 2, ..., k\} \to X$ tais que $\omega\left(0\right) = 0$ e $\omega\left(i\right) \neq \omega\left(i-1\right)$ se $i = 1, 2, ..., k$. A cardinalidade de $\Omega$ é $n^k$. O evento $A =$"a fofoca é contada $k$ vezes sem voltar a ser contada à velhinha que a inventou, e sem ninguém a ouvir duas vezes" é o subconjunto de $\Omega$ formados pelos caminhos $\omega$ tais que $\omega\left(i\right) \neq \omega\left(j\right)$ se $i \neq j$. A cardinalidade de $A$ é $n \cdot (n-1) \cdot (n-2) \cdot ... \cdot (n-k+1)$, desde que $k \leq n$. Portanto, dentro do modelo "probabilidade uniforme nas partes de $\Omega$", a resposta é

$$\mathbf{P}\left(A\right) = \frac{n \cdot (n-1) \cdot (n-2) \cdot ... \cdot (n-k+1)}{n^k}$$

se $k \leq n$ e $\mathbf{P}\left(A\right) = 0$ se $k > n$.

Se sabemos que a velhinha 3 brigou com a velhinha 7 e já não fala com ela, o modelo tem que ser mudado...

**Exemplo: as duas moedas.** Um modelo do lançamento de duas moedas é:

$$\Omega = \{H,T\} \times \{H,T\} = \{(H,H),(H,T),(T.H),(T,T)\}$$

onde $H$ está por "cara" e $T$ por "coroa", com probabilidade uniforme $\mathbf{P}$ sobre as suas partes. Em particular, o evento "obter uma cara e uma coroa", ou seja $A = \{(H,T),(T,H)\}$, tem probabilidade $\mathbf{P}(A) = 1/2$. Ninguém duvida que este modelo descreve bem a experiência.

"E se as moedas são iguais e são lançadas simultaneamente?" O que acontece é o seguinte. Se "eu que observo" não sei distinguir entre as duas moedas, nem dizer qual caiu primeiro, então a álgebra de eventos que "eu observo" é a álgebra $\mathcal{A}$ gerada pela partição

$$\Omega = \{(H,H)\} \cup \{(H,T),(T,H)\} \cup \{(T,T)\}$$

ou seja uma sub-álgebra estricta das partes de $\Omega$. Isto não obriga a mudar a medida de probabilidade, basta dizer que agora a probabilidade é a restrição de $\mathbf{P}$ à álgebra $\mathcal{A}$. Em particular, o evento $A$ continua a ter probabilidade $1/2$. Aliás, as moedas, coitadas, não sabem que eu não sei distinguir-las, nem têm relojos para decidir se cairam no mesmo instante!

Quem não acreditar nesta resposta, é convidado a lançar muitas vezes duas moedas que acha iguais, o mais simultaneamente que pode, e observar a frequência com que acontece o evento $A$. Esta é a única maneira de decidir se o modelo é credível.

Existem na natureza objectos que são "intrinsecamente" indistinguíveis, são as partículas da física subatómica de acordo com a mecânca quântica, e têm efectivamente um comportamento estatístico muito pouco intuitivo para nós que vivemos num mundo macroscópico...

**Exemplo: provas repetidas.** Um bêbado tem 7 chaves, das quais só uma abre a porta da sua casa, e começa a experimentá-las uma a uma. Qual é a probabilidade $p_n$ de ele conseguir abrir a porta à $n$-ésima tentativa?

A resposta depende da estratégia que o bêbado utiliza.

Se decide não voltar a pôr no bolso as chaves já experimentadas, uma resposta é $p_n = 1/7$ se $n = 1,2,...,7$, e portanto ele tem a certeza de abrir a porta dentro de 7 tentativas. A probabilidade de ele abrir a porta dentro de $n$ tentativas, com $n \leq 7$, é $n/7$.

Se bebeu muito, e volta a pôr no bolso as chaves já experimentadas, não pode ter a certeza de conseguir abrir a porta dentro de um número fixado de tentativas. Abrir a porta (pela primeira vez) à $n$-ésima tentativa quer dizer falhar nas primeiras $n-1$ e acertar a $n$-ésima, e portanto uma resposta é $p_n = (6/7)^{n-1} \cdot 1/7$. A probabilidade de ele abrir a porta dentro de $n$ tentativas, e desta vez $n$ pode ser arbitrariamente grande, é $1-(6/7)^n$.

As duas estratégias acima são designadas como "escolher objectos sem reposição" e "escolher objectos com reposição", respectivamente. As respostas acima são "intuitivas" e "razoáveis", mas é importante reconhecer as hipóteses escondidas por trás. A primeira resposta assume que, em cada prova, cada uma das chaves ainda não experimentadas tem a mesma probabilidade de ser escolhida, i.e. cada prova é descrita por um espaço de probabilidade uniforme (embora a maneira menos ambígua de ver o problema é esquecer o "tempo", e reparar que se trata da probabilidade uniforme no espaço das permutações das sete chaves). A segunda resposta assume, além da uniformidade em cada prova, que as diferentes provas são "independentes", i.e. que a $n$-ésima tentativa não tem memória das $n-1$ tentativas falhadas anteriores.

**! Paradoxo de Bertrand.** Escolho ao acaso uma corda numa circunferência. Qual é a probabilidade de o comprimento dela ser maior do que o raio da circunferência?

Resposta 1. Fixo um extremo da corda e escolho o outro com probabilidade uniforme com respeito ao comprimento do arco $d\theta/2\pi$. A probabilidade é $2/3$.

Resposta 2. Escolho ao acaso a linha afim que suporta a corda. Pela simetria rotacional, considero só as linhas horizontais que cortam a circunferência, uniformemente com respeito à medida de Lebesgue $dy/2$ no intervalo $[-1,1]$. A probabilidade é $1/2$.

Resposta 3. Escolho ao acaso o ponto central da corda, uniformemente com respeito à àrea, a medida de Lebesgue $dxdy/\pi$ na bola. A probabilidade é $1/4$.

Moral: a palavra "acaso" é ambígua. As resposta 1, 2 e 3 são respostas a três distintas perguntas que a nossa linguagem do dia a dia confunde.

**Exercícios.**

**a.** Defina modelos probabilísticos (ou seja espaços de probabilidades) das seguintes experiências:
- lançamento de um dado,
- lançamento de dois dados,
- lançamento de 3 moedas,
- lançamento de um dado e uma moeda,
- extracção de uma bola de uma caixa que contém $b$ bolas brancas e $p$ bolas pretas,
- lançamentos de uma moeda até sair cara pela primeira vez.

**d.** Defina um modelo probabílisticos da experiência "lançar duas vezes um dado" ou "lançar dois dados", e determine a probabilidade dos seguintes eventos:
- observar faces distintas,
- obter 6 pelo menos uma vez,
- a soma dos valores observados ser $> 10$,
- o maior dos valores obtidos ser $\leq 4$.

**c.** Sejam $A$, $B$ e $C$ eventos de um espaço de probabilidades tais que $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = 1/4$, $\mathbf{P}(A \cap B) = \mathbf{P}(B \cap C) = 0$ e $\mathbf{P}(A \cap C) = 1/8$. Determine a probabilidade de ocorrer pelo menos um deles.

**d.** Quantos filhos deve ter um casal de modo a ter, com probabilidade $\geq 0.99$, pelo menos um rapaz e uma rapariga?

**e.** Uma enciclopedia em 24 volumes é posta ao acaso numa estante. Com que probabilidade a obra é ordenada correctamente, de esquerda para direita ou de direita para esquerda?

**f.** Escrevo $n$ cartas para $n$ pessoas distintas, meto-as em $n$ envelopes, e escrevo ao acaso as $n$ direcções dos destinatários. Com que probabilidade pelo menos uma das cartas chega ao destinatário? E todas?

# 3 Probabilidade condicionada e independência

**Probabilidade condicionada.** Sejam $(\Omega, \mathcal{E}, \mathbf{P})$ um espaço de probabilidades e $B$ um evento com $\mathbf{P}(B) > 0$. A *probabilidade condicionada* do evento $A$ com respeito ao evento $B$ é definida por

$$\mathbf{P}(A\,|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

(que os probabilistas lêem "a probabilidade de $A$ sabendo (que) $B$ (aconteceu)", ou "a probabilidade de $A$ dado $B$").

A ideia da probabilidade condicionada é a de definir uma nova medida de probabilidades tal que $B$ joga o papel do evento certo. Pois, fixado o evento $B$, a função $\mathbf{P}_B : A \mapsto \mathbf{P}(A\,|B)$ é uma probabilidade sobre $\mathcal{E}$, e $\mathbf{P}_B(B) = 1$.

Saber que um evento aconteceu é uma informação que, em geral, muda a nossa expectativa acerca dos outros. Esta é a ideia formalizada na definição de probablidade condicionada. De facto, se $B$ aconteceu, a $\sigma$-álgebra dos eventos possíveis é $\mathcal{E}_B = \{A \cap B \text{ com } A \in \mathcal{E}\}$, e a função $\mathbf{P}_B$ pode ser pensada como uma probabilidade definida sobre $\mathcal{E}_B$.

**Árvores de probabilidades.** A definição de probabilidade condicionada, na forma

$$\mathbf{P}(A \cap B) = \mathbf{P}(A\,|B) \cdot \mathbf{P}(B)$$

lê-se, da direita para a esquerda, "a probabilidade de acontecer seja $A$ seja $B$ é igual ao produto da probabilidade de acontecer $B$ vezes a probabilidade de acontecer $A$ sabendo que aconteceu $B$".

Em geral, se $A_1, A_2, ..., A_n$ são eventos e as seguintes probabilidades condicionadas fazem sentido,

$$\mathbf{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbf{P}(A_n\,|A_1 \cap ... \cap A_{n-1}) \cdot ... \cdot \mathbf{P}(A_3\,|A_1 \cap A_2) \cdot \mathbf{P}(A_2\,|A_1) \cdot \mathbf{P}(A_1)$$

Esta observação justifica o uso das "árvores de probabilidades".

**Partições e fórmula da probabilidade total.** Uma *partição* de um espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$ é uma família enumerável $B_1, B_2, B_3, ...$ de eventos dois a dois disjuntos, com $\mathbf{P}(B_n) > 0$ para todo $n$, e tais que $\Omega = \cup_n B_n$.

Se $B_1, B_2, B_3, ...$ é uma partição de $\Omega$, então todo evento $A$ é igual a reunião disjunta $\cup_n (A \cap B_n)$. Utilizando a aditividade e a definição de probabilidade condicionada temos que

$$\begin{aligned} \mathbf{P}(A) &= \sum_n \mathbf{P}(A \cap B_n) \\ &= \sum_n \mathbf{P}(A\,|B_n) \cdot \mathbf{P}(B_n) \end{aligned}$$

Esta identidade é dita *fórmula da probabilidade total*. Embora elementar, é muito útil para calcular a probabilidade de um evento que parece complicado: divide-se o evento em "casos" mutualmente exclusivos...

**Fórmula da Bayes.** Em problemas de estatística é também interessante a identidade, válida na situação tratada acima se também $\mathbf{P}(A) > 0$,

$$\mathbf{P}(B_n\,|A) = \frac{\mathbf{P}(A\,|B_n) \cdot \mathbf{P}(B_n)}{\mathbf{P}(A)}$$

e conhecida como *fórmula de Bayes*. A fórmula da probabilidade total então implica o *teorema de Bayes*

$$\mathbf{P}(B_n\,|A) = \frac{\mathbf{P}(A\,|B_n) \cdot \mathbf{P}(B_n)}{\sum_n \mathbf{P}(A\,|B_n) \cdot \mathbf{P}(B_n)}$$

Os eventos $B_n$ têm a interpretação de hipóteses, e a probabilidade condicionada $\mathbf{P}(B_n\,|A)$ a de probabilidade "a posteriori" de $B_n$, depois de ter observado o evento $A$.

**Exercícios.**

**a.** Duas bolinhas são retiradas da uma caixa que contém $a$ bolinhas brancas e $b$ bolinhas pretas. Sejam $A$ e $B$ os eventos "a primeira bolinha retirada é branca" e "a segunda bolinha retirada é branca". Calcule $\mathbf{P}(B\,|A)$, $\mathbf{P}(B\,|A^c)$, $\mathbf{P}(A)$ e $\mathbf{P}(B)$.

**b.** Tenho duas moedas honestas e uma moeda falsa que tem "cara" em cada face. Escolho ao acaso uma das três moedas, lanço-a $n$ vezes, e observo $n$ vezes cara. Qual a probabilidade de eu ter escolhido a moeda falsa? Observe o que acontece quando $n \to \infty$.

**Independência.** Seja $(\Omega, \mathcal{E}, \mathbf{P})$ um espaço de probabilidades. Os eventos $A$ e $B$ são ditos *independentes* quando

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$$

"Ser independentes" é uma relação simétrica, mas não é reflexiva nem transitiva.

A interpretação é a seguinte: se $\mathbf{P}(B) > 0$, os eventos $A$ e $B$ são independentes sse $\mathbf{P}(A\,|B) = \mathbf{P}(A)$ (ou seja, "toda informação acerca do evento $B$ não muda as expectativas acerca do evento $A$").

A família de eventos $(A_k)$ é uma *família independente* se para todo natural $i$ e toda escolha de $k_1, k_2, ..., k_i$ distintos

$$\mathbf{P}(A_{k_1} \cap A_{k_2} \cap ... \cap A_{k_i}) = \mathbf{P}(A_{k_1}) \cdot \mathbf{P}(A_{k_2}) \cdot ... \cdot \mathbf{P}(A_{k_i})$$

As $\sigma$-álgebras $\mathcal{F}$ e $\mathcal{G}$, contidas em $\mathcal{E}$, são *independentes* se todo $A \in \mathcal{F}$ é independente de todo $B \in \mathcal{G}$. De maneira analoga define-se a independência de uma família de $\sigma$-álgebras. A família de $\sigma$-álgebras $(\mathcal{E}_k)$, contidas em $\mathcal{E}$, é uma *família independente* se para todo $n$ e toda escolha de $A_1 \in \mathcal{E}_1$, $A_2 \in \mathcal{E}_2$, ..., $A_n \in \mathcal{E}_n$ acontece que

$$\mathbf{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbf{P}(A_1) \cdot \mathbf{P}(A_2) \cdot ... \cdot \mathbf{P}(A_n)$$

**Exercícios.**

**a.** O evento $A$ é independente de $A$ sse $\mathbf{P}(A) = 0$ ou $1$ (ou seja, um evento é independente de si mesmo sse a sua probabilidade é trivial).

**b.** Os eventos $A$ e $B$ são independentes sse $A$ e $B^c$ são independentes.

**c.** Sejam $A$ e $B$ dois eventos tais que $\mathbf{P}(A \cap B) > 0$. Então

$$\mathbf{P}(C|A \cap B) = \mathbf{P}(C|A)$$

implica que

$$\mathbf{P}(C \cap B|A) = \mathbf{P}(C|A) \cdot \mathbf{P}(B|A)$$

Interprete este resultado.

**d.** Considere o espaço de probabilidades uniforme que descreve a experiência "lançar $n$ moedas honestas". Verifique que os eventos "cara na $i$-ésima moeda" e "cara na $j$-ésima moeda" são independentes se $i \neq j$.

**e.** Considere o espaço de probabilidades uniforme que descreve a experiência "lançar 2 moedas honestas". Determine a probabilidade condicionada de

- obter duas caras sabendo que a primeira moeda mostra cara,
- obter duas caras sabendo que pelo menos uma das moedas mostra cara.

**f.** *(urna de Polya)* Uma caixa contém $a$ bolinhas brancas e $b$ bolinhas pretas. Uma bolinha é escolhida ao acaso, e é posta novamente na caixa junto com $d$ bolinhas da mesma cor. Mais uma bolinha é escolhida ao acaso, e é posta novamente na caixa junto com $d$ bolinhas da mesma cor. E assim a seguir...

- Determine a probabilidade da segunda bolinha retirada ser preta.
- Mostre que a probabilidade da $n$-ésima bolinha retirada ser preta é igual a probabilidade da primeira bolinha retirada ser preta.
- Determine a probabilidade da primeira bolinha retirada ser preta sabendo que a a segunda bolinha retirada é preta.
- Determine a probabilidade da primeira bolinha retirada ser preta sabendo que as sucessivas $n$ bolinhas retiradas são preta, e calcule o limite desta probabilidade quando $n \to \infty$.

**g.** Retiro uma carta de um baralho francês de 52 cartas. Os eventos "a carta é um 7" e "a carta é um ♣" são independentes, no modelo de probabilidade uniforme. As coisas mudam se o 7 de ♡ não está no baralho.

**h.** Família com $n$ filhos, que podem ser meninas ou meninos. Um modelo é o espaço das palavras de comprimento $n$ nas letras "menina" e "menino" munido da probabilidade uniforme. Os eventos "a família não tem mais do que uma menina" e "a família tem pelo menos uma menina e um menino" são independentes se $n = 3$, mas isso não acontece se $n = 2$. Este exemplo mostra que a independência de dois eventos não é uma questão "semántica", mas uma propriedade que pode ser verificada, ou não, dentro de um modelo.

**i.** No lançamento de dois dados, sejam $A$ o evento "ímpar no primeiro dado", $B$ o evento "ímpar no segundo dado" e $C$ o evento "a soma é ímpar". É facil verificar que os eventos $A$, $B$ e $C$ são dois a dois independentes e têm probabilidade positiva, mas

$$\mathbf{P}(A \cap B \cap C) \neq \mathbf{P}(A) \cdot \mathbf{P}(B) \cdot \mathbf{P}(C)$$

sendo $A \cap B \cap C$ o evento impossível. Este exemplo mostra que a independência de uma família de eventos não é uma consequência da independência entre pares de eventos, mas uma condição mais forte.

**j.** Seja $(A_k)_{k=1,\dots,n}$ uma família de eventos independentes. Prove que

$$\mathbf{P}(\cup_{k=1}^{n} A_k) = 1 - \prod_{k=1}^{n} \mathbf{P}(A_k^c)$$

Deduza que a probabilidade de nenhum dos $A_k$ acontecer é $\prod_{k=1}^{n} \mathbf{P}(A_k^c)$.

**k.** Um sistema é composto por $n$ componentes em série, e funciona só se cada componente funciona. As componentes avariam independentemente uma das outras, com probabilidade $p \in ]0,1[$. Calcule a probabilidade de

- o sistema não funcionar,
- apenas a primeira componente ter avariado sabendo que o sistema não funciona,
- todas as componentes terem avariado sabendo que o sistema não funciona,
- o sistema não funcionar sabendo que as primeiras $k$ componentes funcionam.

**l.** Um sistema é composto por $n$ componentes em paralelo, e funciona desde que pelo menos uma das componentes funciona. As componentes avariam independentemente uma das outras, com probabilidade $p \in ]0,1[$. Calcule a probabilidade de

- o sistema não funcionar,
- apenas a primeira componente ter avariado, sabendo que o sistema não funciona,
- todas as componentes terem avariado, sabendo que o sistema não funciona.
- o sistema não funcionar sabendo que as primeiras $k$ componentes funcionam.

**obs.** A importância das noções de independência e probabilidade condicionada está no facto delas fazerem de guia na construção de modelos de fenómenos físicos. Por um lado, permitem construir modelos de "experiências independentes", como mostra o exemplo a seguir e, mais à frente, o das provas de Bernoulli. Por outro lado, permitem codificar de que maneira o futuro "depende" do presente e eventualmente do passado em modelos de sistemas dinâmicos aleatórios, como mostra o exemplo a seguir das moedas com memória e, mais à frente, o das cadeias de Markov.

**Exemplo: um dado e uma moeda independentes.** Como fazer um modelo do lançamento de um dado e uma moeda que diga que "o dado e a moeda são independentes"? Sejam $(\Omega_d, \mathcal{P}(\Omega_d), \mathbf{P}_d)$ o modelo do lançamento de um dado, e $(\Omega_m, \mathcal{P}(\Omega_m), \mathbf{P}_m)$ o modelo do lançamento da moeda. Todo subconjunto de $\Omega = \Omega_d \times \Omega_m$ é uma reunião disjunta de conjuntos do tipo $A_d \times A_m$ com $A_d \subset \Omega_d$ e $A_m \subset \Omega_m$. Por outro lado $A_d \times A_m = (A_d \times \Omega_m) \cap (\Omega_d \times A_m)$, e $A_d \times \Omega_m$ pode ser interpretado como "um evento que só depende do dado", assim como $\Omega_d \times A_m$ "um evento que só depende da moeda". Portanto, postulando a aditividade, a receita

$$\mathbf{P}(A_d \times A_m) = \mathbf{P}_d(A_d) \cdot \mathbf{P}_m(A_m)$$

define uma probabilidade $\mathbf{P} : \mathcal{P}(\Omega) \to [0,1]$, dita *probabilidade produto*, sobre as partes de $\Omega$ tal que "todos os eventos que só dependem do dado são independentes de todos os eventos que só dependem da moeda" (é um exercicio de álgebra provar que $\mathbf{P}$ é uma probabilidade). Esta receita corresponde a multiplicar as probabilidades dos átomos dos dois espaços: se os átomos de $\Omega_d$ e de $\Omega_m$ têm probabilidades $p_i^d = \mathbf{P}_d(\{\omega_i^d\})$ e $p_j^m = \mathbf{P}_m(\{\omega_j^m\})$, então os átomos do produto cartesiano têm probabilidades $\mathbf{P}(\{(\omega_i^d, \omega_j^m)\}) = p_i^d \cdot p_j^m$.

Se $\mathbf{P}_d$ e $\mathbf{P}_m$ são as probabilidades uniformes em $\Omega_d$ e $\Omega_m$ respectivamente, i.e. modelos de um dado e uma moeda honesta, então a probabilidade produto em $\Omega = \Omega_d \times \Omega_m$ é a probabilidade uniforme: cada resultado possível tem probabilidade $1/|\Omega|$.

**Espaços de probabilidades produto.** A mesma construção acima pode ser feita para um número finito de espaços de probabilidades finitos: o resultado é um modelo de "experiências independentes". Seja $(\Omega_k, \mathcal{A}_k, \mathbf{P}_k)$, com $k = 1, 2, ..., n$, uma coleção de espaços de probabilidades finitos (i.e. $\mathcal{A}_k$ são algebras finitas), pensados como modelos de $n$ experiências distintas. Seja $\Omega$ o produto cartesiano $\prod_{k=1}^n \Omega_k = \Omega_1 \times \Omega_2 \times ... \times \Omega_n$. A cada $\mathcal{A}_k$ pode ser associada, de maneira natural, uma sub-álgebra $\mathcal{A}'_k$ de partes de $\Omega$: basta associar a cada elemento $A_k \in \mathcal{A}_k$ o evento

$$A'_k = \{\omega = (\omega_1, \omega_2, ..., \omega_n) \in \Omega \text{ t.q. } \omega_k \in A_k\} \subset \Omega$$

que só depende da $k$-ésima experiência, a $k$-ésima coordenada de $\omega$. A *álgebra produto* em $\Omega$ é a álgebra $\mathcal{A} = \otimes_{k=1}^n \mathcal{A}_k$, definida como sendo a menor álgebra que contém $\mathcal{A}'_1 \cup \mathcal{A}'_2 \cup ... \cup \mathcal{A}'_n$. Todo elemento $A \in \mathcal{A}$ é uma reunião disjunta de eventos do tipo $A_1 \times A_2 \times ... \times A_n$ com $A_k \in \mathcal{A}_k$, portanto, postulando a aditividade, a receita

$$\mathbf{P}(A_1 \times A_2 \times ... \times A_n) = \mathbf{P}_1(A_1) \cdot \mathbf{P}_2(A_2) \cdot ... \cdot \mathbf{P}_n(A_n)$$

define uma probabilidade $\mathbf{P} : \mathcal{A} \to [0,1]$, dita *probabilidade produto*. O espaço de probabilidade $(\Omega, \mathcal{A}, \mathbf{P})$ é dito *espaço de probabilidades produto*. A família $(\mathcal{A}'_k)_{k=1,...,n}$, formada pelas álgebras de eventos que dependem das diferentes experiências, é uma família de sub-álgebras independentes do espaço de probabilidades produto.

Vale a pena observar que, se os $\Omega_k$ são conjuntos finitos e as $\mathbf{P}_k$ são as probabilidades uniformes em $\mathcal{A}_k = \mathcal{P}(\Omega_k)$, então a probabilidade produto $\mathbf{P}$ construida acima é a probabilidade uniforme em $\mathcal{A} = \mathcal{P}(\Omega)$. Neste caso, todo átomo $\omega \in \Omega$ tem probabilidade $\mathbf{P}(\{\omega\}) = 1/|\Omega|$.

**Exemplo: moedas com memória.** Sejam $p_1$ a probabilidade de sair cara no primeiro lançamento de uma moeda, e $p$ a probabilidade de obter num lançamento o mesmo resultado do lançamento precedente. Esta informação é suficiente para calcular a probabilidade $p_n$ de sair cara no $n$-ésimo lançamento, para todo natural $n$, esquecendo, por enquanto, o problema não trivial de definir rigorosamente o espaço de probabilidades. A fórmula da probabilidade total diz que "a probabilidade de obter cara no $(n+1)$-ésimo lançamento é igual à soma de $p$ vezes a probabilidade

de obter cara no $n$-ésimo lançamento mais $1-p$ vezes a probabilidade de obter coroa no $n$-ésimo lançamento", ou seja

$$p_{n+1} = p \cdot p_n + (1-p) \cdot (1-p_n)$$

e esta equação recursiva, junto com a condição inicial $p_1$, determina as probabilidades $p_n$ para todo $n \in \mathbf{N}$. A solução é

$$p_n = p_1 \cdot \delta^{n-1} + (1-p) \cdot \left(1 + \delta + \delta^2 + ... + \delta^{n-2}\right) = (p_1 - 1/2) \cdot \delta^{n-1} + 1/2$$

onde $\delta = 2p - 1$.

É interessante observar que, se $p \neq 0$ ou 1, o limite $\lim_{n\to\infty} p_n$ existe, e é independente de $p_1$. Este é um caso simples do teorema ergódico para cadeias de Markov transitivas, que descreve a "perda de memória" e a "convergência para um estado estacionário" de um sistema dinâmico suficientemente caótico. É a procura deste tipo de regularidades um dos objectivos da teoria das probabilidades.

**Lema de Borel-Cantelli.** O resultado seguinte é o protótipo de uma "lei zero-um", um teorema que diz que um evento (neste caso o limsup de uma sucessão de eventos independentes) só pode ter probabilidade 0 ou 1 (ou seja, é independente de si mesmo).

**Lema de Borel-Cantelli** . *Seja $(A_n)$ uma sucessão de eventos.*
*i) se a série $\sum \mathbf{P}(A_n)$ é convergente então $\mathbf{P}\left(\overline{\lim} A_n\right) = 0$*
*ii) se $(A_n)$ é uma família independente e a série $\sum \mathbf{P}(A_n)$ é divergente então $\mathbf{P}\left(\overline{\lim} A_n\right) = 1$.*

**dem.** i) Os eventos $\cup_{k=n}^{\infty} A_k$ formam uma sucessão decrescente, logo

$$\mathbf{P}\left(\overline{\lim} A_n\right) = \lim_{n\to\infty} \mathbf{P}\left(\cup_{k=n}^{\infty} A_k\right) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbf{P}(A_k)$$

e, se a série é convergente, o seu resto converge para 0.

ii) Se $\mathbf{P}\left(\overline{\lim} A_n\right) < 1$ então existe $\varepsilon > 0$ tal que

$$\begin{aligned} \varepsilon &\leq \mathbf{P}\left(\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k^c\right) \leq \lim_{n\to\infty} \lim_{m\to\infty} \mathbf{P}\left(\cap_{k=n}^{n+m} A_k^c\right) \\ &\leq \lim_{n\to\infty} \lim_{m\to\infty} \prod_{k=n}^{m} \left(1 - \mathbf{P}(A_k)\right) \leq \lim_{n\to\infty} \lim_{m\to\infty} \exp\left(-\sum_{k=n}^{n+m} \mathbf{P}(A_k)\right) \end{aligned}$$

(porque $1 - x \leq e^{-x}$). Portanto

$$\lim_{n\to\infty} \lim_{m\to\infty} \sum_{k=n}^{n+m} \mathbf{P}(A_k) \leq -\log \varepsilon$$

e isso implica que a série $\sum \mathbf{P}(A_n)$ é convergente. □

# 4 Modelos finitos e provas de Bernoulli

**Cálculo combinatório.** Para calcular probabilidades em espaços de probabilidades uniformes é preciso calcular cardinalidades de conjuntos finitos. Sejam $K$ e $N$ conjuntos finitos de cardinalidade respetivamente $k$ e $n$.

A cardinalidade do produto cartesiano $K \times N$ é $k \cdot n$.

A cardinalidade de $N^K = \{\text{funções } K \to N\}$, isomorfo ao produto cartesiano $N \times N \times ... \times N$ de $k$ cópias de $N$, é

$$\left|N^K\right| = n^k$$

A cardinalidade de $D_k^n = \{\text{funções injetivas } K \to N\}$ é

$$|D_k^n| = n \cdot (n-1) \cdot ... \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

desde que $k \leq n$, tendo decidido que $0! = 1$.

Em particular, a cardinalidade de $D_n^n$, o espaço das *permutações* de $N$, é

$$|D_n^n| = n!$$

A cardinalidade de $C_k^n = \{\text{subconjuntos } K \subset N \text{ com } |K| = k\}$ é

$$|C_k^n| = \frac{n!}{k!(n-k)!}$$

desde que $k \leq n$, pois $C_k^n \simeq D_k^n$ módulo $D_k^k$ (i.e. duas funções injetivas $K \to N$ definem o mesmo subconjunto de $N$, a imagem, sse diferem por uma permutação de $K$).

**Coeficiente binomial.** O número $|C_k^n|$, usualmente denotado por $\binom{n}{k}$, é dito *coeficiente binomial*, por via da fórmula do binómio de Newton

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Em particular, se $a + b = 1$, vale a identidade $\sum_{k=0}^{n} \binom{n}{k} a^k (1-a)^{n-k} = 1$ .

**Fórmula de Stirling.** É útil saber a fórmula de Stirling, que diz que

$$n! = \sqrt{2\pi n} \cdot n^n \cdot e^{-n} \cdot e^{x_n/12n}$$

onde $x_n \in ]0, 1[$. Em particular, se $n$ é grande,

$$n! = \sqrt{2\pi n} \cdot n^n \cdot e^{-n} \cdot (1 + \mathcal{O}(1/n)) \simeq \sqrt{2\pi n} \cdot n^n \cdot e^{-n}$$

**Exercícios.**

**a.** Calcule a cardinalidade de $\{\text{funções sobrejetivas } N \to K\}$.

**b.** Calcule a cardinalidade de $\{\text{funções } N \to N \text{ sem pontos fixos}\}$.

**Exemplo: o problema dos aniversários.** $k$ bolinhas caem em $n$ caixas. Cada bolinha escolhe uma das caixas, independentemente do que fazem as outras. O problema é calcular a probabilidade do evento $A =$"alguma caixa contém mais do que uma bolinha" (para que $A$ não seja o evento certo temos que pôr $k \leq n$).

Um modelo desta experiência é $\Omega = N^K$ com probabilidade uniforme, onde $N$ é um conjunto de $n$ elementos (o conjunto das caixas) e $K$ é um conjunto de $k$ elementos (o conjunto das bolinhas). Um ponto de $\Omega$ é uma função $\omega : K \to N$, e o valor $\omega(i)$ é a caixa escolhida pela $i$-ésima bolinha. Portanto, $A_{i,j} = \{\omega \in \Omega \text{ t.q. } \omega(i) = j\}$ representa o evento "a $i$-ésima bolinha cae na $j$-ésima caixa". Observe que a probabilidade uniforme em $\Omega$ verifica $\mathbf{P}(A_{i,j}) = 1/n$, o que quer dizer que cada bolinha tem probabilidade $1/n$ de cair em cada uma das caixas, e que a família de eventos $(A_{i,j_i})_{i\in K,\ j_i \in N}$ é uma família independente para cada escolha de $i \mapsto j_i$, o que traduz a "independência das diferentes bolinhas".

O evento $A^c =$"nenhuma caixa contém mais do que uma bolinha" tem cardinalidade igual à cardinalidade de $D_k^n$, portanto a resposta é

$$\mathbf{P}(A) = 1 - \mathbf{P}(A^c) = 1 - \frac{n!}{n^k(n-k)!}$$

Uma boa aproximação de $\mathbf{P}(A^c)$, se $k \ll n$, é

$$1 - \frac{(1 + 2 + ... + (k-1))}{n} = 1 - \frac{k \cdot (k-1)}{2n} \simeq \exp\left(-\frac{k \cdot (k-1)}{2n}\right)$$

Uma curiosidade: se $n = 365$ e $k \geq 23$ então $\mathbf{P}(A) > 0.50$, se $n = 365$ e $k \geq 64$ então $\mathbf{P}(A) > 0.99$.

**Exemplo: estatísticas de Fermi-Dirac e de Bose-Einstein.** As partículas da física subatómica são "indistinguíveis", e isso dá lugar a estatísticas menos intuitivas do que a estatística das bolinhas. Elas dividem-se em bosões (com spin inteiro, como os fotões), que podem estar em grupos num mesmo estado físico, e fermiões (com spin semi-inteiro, como os electrões) que, de acordo com o "princípio de exclusão de Pauli", não podem estar num mesmo estado com outros fermiões.

Temos $k$ partículas que podem ocupar $n$ estados, com $k \leq n$. Queremos calcular a probabilidade $p$ de elas ocuparem os primeiros $k$ estados, e a probabilidade $q$ de elas ocuparem $k$ estados diferentes.

Na estatística de Maxwell-Boltzmann, a estatística dos objectos macroscópicos e portanto a mesma das bolinhas, as respostas são

$$p = \frac{k!}{n^k} \quad \text{e} \quad q = \frac{n!}{n^k(n-k)!}$$

Na estatística de Bose-Einstein, em que as partículas são indistinguíves, as respostas são

$$p = \frac{k!(n-1)!}{(n+k-1)!} \quad \text{e} \quad q = \frac{n!(n-1)!}{(n-k)!(n+k-1)!}$$

Na estatística de Fermi-Dirac, em que as partículas são indistinguíves e em que duas partículas não podem ocupar o mesmo estado, as respostas são

$$p = \frac{k!(n-k)!}{n!} \quad \text{e} \quad q = 1$$

**Provas de Bernoulli.** É um modelo de $n$ experiências repetidas e independentes de um jogo com probabilidade de sucesso $p$ (para evitar trivialidades $0 < p < 1$). É tradição chamar $q = 1 - p$ a probabilidade de insucesso. O espaço dos acontecimentos é

$$\Omega^n = \{0,1\}^n = \{\omega = (\omega_1, \omega_2, ..., \omega_n) \text{ com } \omega_i = 0 \text{ ou } 1\}$$

o espaço das palavras de comprimento $n$ nas letras 0 ("insucesso") e 1 ("sucesso"). A família dos eventos é $\mathcal{P}(\Omega^n)$. Seja $A_i = \{\omega \in \Omega^n \text{ t.q. } \omega_i = 1\}$ o evento "sucesso na $i$-ésima prova". A receita "a família $(A_i)_{i=1,...,n}$ é uma família independente e $\mathbf{P}(A_i) = p$ para todo $i = 1, 2, ..., n$ " define

uma probabilidade $\mathbf{P}$ sobre $\mathcal{P}(\Omega^n)$. De facto, cada palavra, por exemplo $\omega = (1,0,1,...,0) \in \Omega$, é da forma

$$\{\omega\} = A_1 \cap A_2^c \cap A_3 \cap ... \cap A_n^c$$

i.e. é uma interseção de $A_i$ ou $A_i^c$ com $i = 1,2,...,n$. Pela hipótese de independência, a sua probabilidade tem que ser um produto do género $p \cdot q \cdot p \cdot ... \cdot q$, com um número de fatores $p$ igual ao número de vezes que a letra 1 aparece na palavra. O resultado é que

$$\mathbf{P}(\{\omega\}) = p^{\sum_{i=1}^n \omega_i} q^{n-\sum_{i=1}^n \omega_i}$$

Verificar os axiomas é um exercício de álgebra, aliás, $\mathbf{P}$ é a probabilidade produto em $\{0,1\}^n$, onde cada factor $\{0,1\}$ é munido da probabilidade "$\mathbf{P}_i(\{1\}) = p$ e $\mathbf{P}_i(\{0\}) = q$". Este espaço de probabilidades é dito *esquema de Bernoulli*.

Se $\omega \in \Omega^n$ é uma palavra que contém $k$ vezes a letra 1 (logo $n-k$ vezes a letra 0), a sua probabilidade é $p^k q^{n-k}$ e não depende das posições das letras, mas só da quantidade de letras 1. Por outro lado, o número de palavras de $\Omega^n$ com $k$ letras 1 é igual à cardinalidade dos subconjuntos de tamanho $k$ de um conjunto de tamanho $n$. Portanto a probabilidade do evento "$k$ sucessos em $n$ provas" é

$$\mathbf{P}\{\omega \in \Omega^n \text{ t.q. } \omega_1 + \omega_2 + ... + \omega_n = k\} = \binom{n}{k} p^k q^{n-k}$$

A lei associada às provas de Bernoulli é dita *lei binomial*, e joga um papel central na teoria das probabilidades.

Uma observação importante é que o esquema de Bernoulli com $p = 1/2$ (pensado como um modelo de $n$ lançamentos de uma moeda "honesta") é equivalente à probabilidade uniforme nas partes de $\Omega^n$, pois cada palavra $\omega \in \Omega^n$ tem probabilidade $\mathbf{P}(\{\omega\}) = 2^{-n}$.

**Provas independentes com mais resultados possíveis, lei multinomial.** Obviamente, as "letras" 0 e 1 do esquema de Bernoulli podem ser substituidas por outras... Seja $X = \{x_1, x_2, ..., x_z\}$ um "alfabeto" finito, seja

$$\Omega = X^n = \{\omega = (\omega_1, \omega_2, ..., \omega_n) \text{ com } \omega_i \in X\}$$

o espaço das palavras de comprimento $n$ nas letras de $X$, e seja $p = (p_1, p_2, ..., p_z)$ uma probabilidade nas partes de $X$, i.e. uma coleção de números não negativos tais que $\sum_{i=1}^z p_i = 1$. A probabilidade produto $\mathbf{P}$ nas partes de $\Omega^n$, onde cada $X$ é munido da probabilidade $p$, é um modelo de $n$ esperiências repetidas e independentes com $z$ resultados possíveis, também dito esquema de Bernoulli. A probabilidade produto é determinada por

$$\mathbf{P}(\{\omega\}) = p_0^{k_0(\omega)} \cdot p_1^{k_1(\omega)} \cdot ... \cdot p_z^{k_z(\omega)}$$

onde $k_i(\omega)$, com $i \in X$, denota o número de vezes que a letra $x_i$ está contida na palavra $\omega$. A probabilidade do evento formado pelas palavras que contêm $k_1$ vezes a letra $x_1$, $k_2$ vezes a letra $x_2$, ... e $k_z$ vezes a letra $x_z$ é

$$\mathbf{P}\{\omega \in \Omega^n \text{ t.q. } k_1(\omega) = k_1 \text{ , } k_2(\omega) = k_2 \text{ ,..., } k_z(\omega) = k_z\} = \frac{n!}{k_1! \cdot k_2! \cdot ... \cdot k_z!} \cdot p_0^{k_0} \cdot p^{k_2} \cdot ... \cdot p_z^{k_z}$$

**Exercícios.**

**a.** $k$ bolinhas caem em $n$ caixas numeradas, e cada bolinha tem probabilidade $1/n$ de cair em cada caixa (independentemente do que fazem as outras bolinhas). Calcule as probabilidades dos eventos:

- a primeira caixa está vazia,
- as primeiras $k$ caixas estão ocupadas,
- pelo menos uma das caixas está vazia,
- pelo menos uma das caixas contém mais do que uma bolinha.

Responda às mesmas perguntas sabendo que cada caixa não pode conter mais do que uma bolinha (ou seja, as bolinha caem, uma após a outra, e cada uma tem a mesma probabilidade de cair em cada caixa vazia).

**b.** Defina um modelo probabílistico de $n$ lançamentos independentes de uma moeda, e calcule as probabilidades dos seguintes eventos:
- sair a sequência "cara, coroa, cara, coroa,...",
- sair $k$ vezes cara,
- sair pelo menos uma vez cara e uma vez coroa,
- sair pelo menos uma vez cara sabendo que saiu $k$ vezes coroa,
- nunca sair cara.

**c.** É mais provável obter pelo menos um ás em 6 lançamentos de um dado ou obter pelo menos dois ases em 12 lançamentos de um dado?

**d.** *(paradoxo de De Méré)* É mais provável obter exactamente um ás em 4 lançamentos de um dado ou obter exactamente dois ases em 24 lançamentos de dois dados?

**Marcha aleatória.** Um homenzinho passeia dentro dos inteiros $\mathbf{Z}$ com a seguinte estratégia. Começa na posição 0. A cada instante $i = 1, 2, 3, ..., n$ lança uma moeda, com probabilidade $p$ de sair cara, e depois de cada lançamento faz um passo para a frente se saiu cara ou um passo para trás se saiu coroa.

Um modelo desta marcha é assim. Seja $\Omega^n = \{-1, 1\}^n$, $\mathcal{E} = \mathcal{P}(\Omega^n)$ e $\mathbf{P} : \mathcal{E} \to [0, 1]$ o esquema de Bernoulli determinado por $\mathbf{P}(\omega_i = 1) = p$ e $\mathbf{P}(\omega_i = -1) = 1 - p$. A cada palavra $\omega = (\omega_1, \omega_2, ..., \omega_n) \in \Omega$ está associada a "trajectória" $T = (T_0, T_1, T_2, ..., T_n)$, definida por

$$T_0 = 0 \ , \ T_1 = \omega_1 \ , \ T_2 = \omega_1 + \omega_2 \ , \ ... \ , \ T_n = \omega_1 + \omega_2 + ... + \omega_n$$

onde $T_k$ é a "posição" do homenzinho no "tempo" $k$. A medida de probabilidade $\mathbf{P}$ pode ser pensada como uma probabilidade no espaço das trajectórias da marcha aleatória, definido por

$$\Omega' = \{T : \{0, 1, 2, ..., n\} \to \mathbf{Z} \text{ t.q. } T_0 = 0 \text{ e } T_k = T_{k-1} \pm 1 \text{ se } 0 < k \leq n\}$$

Particularmente interessante é a *marcha simétrica*, quando $p = 1/2$ e portanto $\mathbf{P}$ é a probabilidade uniforme nas partes de $\Omega'$: cada trajectória possível tem probabilidade $2^{-n}$.

A marcha aleatória, modelada no esquema de Bernoulli, é um modelo paradigmático em teoria das probabilidades. Representa o modelo mais simples de um "sistema dinâmico aleatório", e as suas "regularidades" são protótipos de fenómenos observados em situações mais complexas. Não é muito longe da realidade dizer que o objectivo da teoria das probabilidades é uma descrição qualitativa das "trajectórias típicas", da "maioria das trajectórias", da marcha aletória e das suas generalizações.

É também possível fazer modelos "contínuos" (i.e. o tempo vive em $\mathbf{R}_+$ e a posição em $\mathbf{R}$ ou $\mathbf{R}^n$ ou num espaço ainda mais geral) de uma marcha deste tipo, conhecidos pelos físicos como "movimento Browniano" e pelos matemáticos como "processo de Wiener".

**Lei hipergeométrica.** De uma caixa, que contém $a$ bolinhas brancas e $b$ bolinhas pretas, são retiradas $n$ bolinhas (sem reposição, e portanto $n \leq a + b$). Um modelo desta experiência é $\Omega = C_n^{a+b}$ com probabilidade uniforme. O evento $A =$"$k$ das $n$ bolinhas são brancas" tem cardinalidade $\left|C_k^a \times C_{n-k}^b\right|$ , logo a sua probabilidade é

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}$$

É confortável observar que, no limite quando $a + b \to \infty$, com $a/(a + b) \to p$ e $k$ e $n$ finitos,

$$\mathbf{P}(A) \to \binom{n}{k} p^k (1 - p)^{n-k}$$

que é a probabilidade do evento "$k$ sucessos em $n$ provas de Bernoulli com probabilidade de sucesso $p$". Ou seja, como a intuição sugere, escolher objectos sem reposição não difere muito de escolher objectos com reposição quando a população é muito grande.

**Exercício.** De uma caixa, que contém $a$ bolinhas brancas e $b$ bolinhas pretas, são retiradas $n$ bolinhas. Sejam $A$ e $B$ os eventos "$k$ das $n$ bolinhas são brancas" e "a $i$-ésima bolinha retirada é branca", respetivamente. Prove que

$$\mathbf{P}(B|A) = k/n$$

quer as bolinhas sejam retirada sem reposição quer sejam retiradas com reposição.

**Infinitas provas de Bernoulli.** Em problemas como a ruína do jogador ou o lançamento de uma moeda até sair coroa pela primeira vez convém fazer um modelo de infinitas provas de Bernoulli, porque não é honesto dizer que o evento "nunca sai coroa em 1000 lançamentos de uma moeda" é impossível (seria como dizer "se nos primeiros 999 lançamentos de uma moeda saiu sempre cara, no 1000-ésimo lançamento tem que sair coroa"!).

O espaço dos acontecimentos é

$$\Omega^\infty = \{0,1\}^{\mathbf{N}} = \{\omega = (\omega_1, \omega_2, ..., \omega_n, ...) \text{ com } \omega_n = 0 \text{ ou } 1\}$$

o espaço das palavras infinitas nas letras 0 e 1. Seja $A_i = \{\omega \in \Omega^\infty \text{ t.q. } \omega_i = 1\}$ o evento "sucesso na $i$-ésima prova". Um *cilindro* em $\Omega^\infty$ é uma interseção finita de conjuntos $A_i$ e $A_j^c$. Seja $\mathcal{B}(\Omega^\infty)$ a menor $\sigma$-álgebra de partes de $\Omega^\infty$ que contém todos os cilindros, dita $\sigma$-álgebra dos boreleanos de $\Omega^\infty$. É possível provar (mas não é óbvio!, é um caso particular do teorema de extensão de Kolmogorov, esboçado mais a frente) que a receita "a família $(A_n)_{n\in\mathbf{N}}$ é uma família independente e $\mathbf{P}(A_i) = p$ para todo $i \in \mathbf{N}$" define uma probabilidade $\mathbf{P}$ sobre $\mathcal{B}(\Omega^\infty)$, dita *probabilidade produto*.

Neste modelo, toda palavra $\omega \in \Omega^\infty$ tem probabilidade $\mathbf{P}(\{\omega\}) = 0$, desde que $p$ seja diferente de 0 ou 1 (casos que não têm muito interesse). Em particular, como era de se esperar, o evento "nunca sai coroa" tem probabilidade 0, mesmo não sendo vazio. Por outro lado, todo cilindro não vazio tem probabilidade positiva, que pode ser facilmente calculada como no caso das provas de Bernoulli finitas.

**Exemplo: a ruína do jogador.** Os jogadores $A$ e $B$, que têm capital inicial de respectivamente $a$ e $b$ rublos, apostam cada vez 1 rublo num jogo onde $A$ tem probabilidade $p$ de ganhar, e $B$ probabilidade $q = 1 - p$ de ganhar. Um jogador perde o jogo quando acabar o seu dinheiro. Podemos calcular as probabilidades de $A$ ou $B$ perder o jogo?

A priori, o jogo pode não acabar nunca, embora tenhamos a ideia de que isso é um evento pouco provável. O modelo natural é o das infinitas provas de Bernoulli, onde "sucesso na $i$-ésima prova" quer dizer, por exemplo, o jogador $A$ ganhou um rublo na $i$-ésima aposta. Sejam

$$p_k(n) = \mathbf{P}(\text{"o jogador } A\text{, com capital inicial de } k \text{ rublos, perde dentro das primeiras } n \text{ apostas"})$$
$$q_k(n) = \mathbf{P}(\text{"o jogador } B\text{, com capital inicial de } a+b-k \text{ rublos, perde dentro das primeiras } n \text{ apostas"})$$
$$r_k(n) = \mathbf{P}(\text{"nenhum dos jogadores, tendo } A \text{ capital inicial de } k \text{ rublos, perde dentro das primeiras } n \text{ apostas"})$$

Estes eventos só dependem das primeiras $n$ provas de Bernoulli, i.e. são cilindros, e as respectivas probabilidades podem ser facilmente calculadas. A reunião dos eventos "o jogador $A$, com capital inicial de $k$ rublos, perde dentro das primeiras $n$ apostas" para $n = 1, 2, 3, ...$, que pertence à $\sigma$-álgebra $\mathcal{B}(\Omega^\infty)$, pode ser interpretado como sendo o evento "o jogador $A$, com capital inicial de $k$ rublos, perde o jogo". Então, pela continuidade da medida de probabilidades,

$$p_k = \lim_{n\to\infty} p_k(n)\ , \quad q_k = \lim_{n\to\infty} q_k(n)\ , \quad r_k = \lim_{n\to\infty} r_k(n)\ ,$$

representam as probabilidades dos eventos "o jogador $A$, com capital inicial de $k$ rublos, perde o jogo", "o jogador $B$, com capital inicial de $a+b-k$ rublos, perde o jogo" e "o jogo não acaba nunca", , respetivamente,.

É evidente que $p_k(n) + q_k(n) + r_k(n) = 1$ para todo $n$, e portanto $p_k + q_k + r_k = 1$. Também, é uma trivialidade observar que

$$p_{a+b} = 0\ , \quad q_{a+b} = 1\ , \quad r_{a+b} = 0$$

$$p_0 = 1 \ , \quad q_0 = 0 \ , \quad r_0 = 0$$

Por outro lado, se depois de $n$ apostas nenhum dos jogadores perdeu, a ruína de $A$ pode vir de dois eventos complementares: $A$ ganha a aposta seguinte e perde depois, $A$ perde a aposta seguinte e perde depois. Pela fórmula da probabilidade total temos

$$p_k = p \cdot p_{k+1} + q \cdot p_{k-1}$$

se $k$ é o capital de $A$ depois das $n$ apostas. Portanto, a sucessão $(p_k)$ satisfaz a equação às diferenças finitas

$$p(p_{k+1} - p_k) = q(p_k - p_{k-1})$$

com as condições na fronteira acima. As soluções são

$$p_a = \frac{b}{a+b}$$

se $p = 1/2$ e

$$p_a = \frac{1 - (p/q)^b}{1 - (p/q)^{a+b}}$$

se $p \neq 1/2$. Nos dois casos temos $q_a = 1 - p_a$ e portanto $r_a = 0$.

Em particular, se o jogo é honesto ($p = 1/2$) e $b \gg a$, a ruína de $A$ é quase certa. Se $A$ joga melhor do que $B$ (se $p > 1/2$), ele tem mais probabilidade de ganhar, mesmo tendo um capital inicial muito menor, pois

$$p_a \sim (q/p)^a$$

se $b \to \infty$. Nos dois casos, aumentar as apostas favorece o jogador mais fraco.

**Exercício.** Traduza o problema da ruína do jogador na linguagem da marcha aleatória.

**Exemplo: la biblioteca total.** Na biblioteca de Babel estão todos os livros possíveis dum certo tamanho. Os homens passam a vida tentando decifrar os livros, ou à procura do livro que conta o próprio passado, na esperança de que também conte o futuro. "...Una secta blafema sugerió que cesaran todas las buscas y que todos los hombres barajaran letras y símbolos, hasta construir, mediante un notable don del azar, esos libros canónicos. Las autoridades se vieron obligadas a promulgar órdenes severas. La secta desapareció, pero en mi niñez he visto hombres viejos que largamente se ocultaban en las letrinas, con unos discos de metal en un cubilete prohibido, y debilmente remedaban el divino desorden..." (Jorge Luis Borges, *La Biblioteca de Babel,* em *Ficciones,* 1941). Seja $b = (b_1, b_2, ..., b_n)$ uma palavra de $n$ letras 0 ou 1 (por exemplo "Muchos anos después, frente al peloton de fusilamiento, el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevó a conocer el hielo..." em binário), e $B_k$ os cilindros

$$B_k = \{\omega \in \Omega^\infty \text{ tais que } \omega_{kn+1} = b_1 \ , \ \omega_{kn+2} = b_2 \ , \ ..., \ \omega_{kn+n} = b_n\}$$

(i.e. as palavras infinitas que contêm a palavra $b$ a partir da $(kn+1)$-ésima letra) para $k = 0, 1, 2, ...$ No modelo das infinitas provas de Bernoulli, por exemplo com probabilidade uniforme em cada prova, os eventos $B_k$ são independentes e têm todos a mesma probabilidade $\mathbf{P}(B_k) = 2^{-n} > 0$. Pelo lema de Borel-Cantelli, $\mathbf{P}\left(\overline{\lim} B_k\right) = 1$. Ou seja, com probabilidade 1 um livro infinito contém uma infinidade de vezes a palavra $b$. O profeta da seita blasfema não estava tão louco?

A afirmação acima parece paradoxal, mas na verdade é muito intuitiva. Um livro finito, por mais comprido que seja, tem probabilidade positiva $\varepsilon$, por mais pequena que seja, de aparecer escolhendo ao acaso $n$ letras do alfabeto. A interpretação física da probabilidade sugere que o tal livro aparece, em média, uma vez por cada $\varepsilon^{-1}$ tentativas. Logo, ao escrever um livro de $k \cdot n \cdot \varepsilon^{-1}$ letras escolhidas ao acaso, nós esperamos ver o livro aparecer $k$ vezes...

**Exercício..** Uma moeda, tal que a probabilidade de se obter cara num lançamento é $p \in ]0, 1[$, é lançada um número infinito de vezes. Calcule a probabilidade de:

- se obter cara um número infinito de vezes,
- se obter cara pelo menos uma vez.
- se obter uma sucessão de 1000 caras seguidas, pelo menos uma vez,
- não se obter nunca cara a partir do $n$-ésimo lançamento, para algum $n$.

# 5 Construction of (probability) measures

Unlike what happens in other branches of mathematics, one of the most delicate technical issue in probability is to prove the mere "existence" of the objects we want to deal with. Indeed, although probability measures on countable spaces are simply countable collections of nonnegative numbers summing up one, it is far from clear how to "exhibit", i.e to show existence and to construct, probability measures on large $\sigma$-algebras. Here we sketch the relevant results dating back to the beginning of the XX century. Almost all straightforward verifications are left to the reader. Standard references are the books by Billingsley [Bi79], Doob [Do94], Halmos [Ha74], Parthasarathy [Pa67] and Rudin [Rud66].

**Measures.** Let $(\Omega, \mathcal{E})$ be a measurable space. A (positive) *measure* on $\mathcal{E}$ is a function $\mu : \mathcal{E} \to [0, \infty]$ such that

i) $\mu(\emptyset) = 0$

ii) it is *$\sigma$-additive,* i.e. if $(S_n)$ is a countable family of pairwise disjoint elements of $\mathcal{E}$ then

$$\mu(\cup_n S_n) = \sum_n \mu(S_n)$$

A measure $\mu$ is called *finite* if $\mu(\Omega) < \infty$ , and *probability* (measure) if $\mu(\Omega) = 1$. It is called *$\sigma$-finite* if $\Omega$ is a countable union of subsets $S_n \in \mathcal{E}$ with $\mu(S_n) < \infty$.

A *measure space* is a triple $(\Omega, \mathcal{E}, \mu)$, a nonempty set $\Omega$, a $\sigma$-algebra $\mathcal{E}$ of subsets of $\Omega$, a measure $\mu$ on $\mathcal{E}$ . If $S \in \mathcal{E}$, the nonnegative number $\mu(S)$ is called the *measure*, or *mass*, of the set $S$. When $\mu = \mathbf{P}$ is a probability measure, then the triple $(\Omega, \mathcal{E}, \mathbf{P})$ is said a *probability space*, measurable sets are called *events*, and $\mathbf{P}(S)$ is called *probability of the event $S$*.

Basic properties of measures are the following: measures are *monotone*, i.e. $\mu(S) \leq \mu(T)$ if $S \subset T$, and *$\sigma$-subadditive*, i.e. if $(S_n)$ is a countable family of elements of $\mathcal{E}$ then

$$\mu(\cup_n S_n) \leq \sum_n \mu(S_n)$$

Measures are continuous from below and from above, in the following sense: if $S_n \uparrow S$ then $\mu(S_n) \uparrow \mu(S)$, if $S_n \downarrow S$ and $\mu(S_n) < \infty$ for some $n$ then $\mu(S_n) \downarrow \mu(S)$. In the case of a probability measure, both continuity properties are equivalent, and indeed a simple argument shows that they are equivalent to continuity from above at $\emptyset$: if $S_n \downarrow \emptyset$ then $\mu(S_n) \downarrow 0$. Moreover, continuity is equivalent to $\sigma$-aditivity if the set function $\mu$ is only assumed (finitely) additive.

**Null sets and complete measures.** A subset $E \subset \Omega$ *has zero measure* if it is contained in a measurable set $S \in \mathcal{E}$ with $\mu(S) = 0$. If any set with zero measure belongs to $\mathcal{E}$, then the measure space $(X, \mathcal{E}, \mu)$ is said *complete*. Any measure space can be canonically completed, extending the measure to the $\sigma$-algebra $\overline{\mathcal{E}}$ made of unions of elements of $\mathcal{E}$ with subsets of zero measure.

A property (like continuity of a function, or convergence of a sequence of functions defined on $\Omega$) holds *$\mu$-a.e.* ("almost everywhere" with respect to the measure $\mu$) if the set of points of $\Omega$ where it does not hold has zero measure.

**Exterior measures.** Let $\Omega$ be a nonempty set. A set function $\mu : \mathcal{P}(\Omega) \to [0, \infty]$ is an *exterior measure* if

i) $\mu(\emptyset) = 0$,

ii) it is monotone, i.e. $\mu(S) \leq \mu(T)$ whenever $S \subset T$,

iii) it is $\sigma$-subadditive, i.e. for any countable family $(S_n)$ of subsets of $\Omega$

$$\mu(\cup_n S_n) \leq \sum_n \mu(S_n)$$

Given an exterior measure $\mu$, the family of *$\mu$-measurable sets* is defined as

$$\mathcal{E}_\mu = \{E \subset \Omega \text{ such that } \mu(S) = \mu(S \cap E) + \mu(S \cap E^c) \text{ for any } S \subset \Omega\}$$

**Theorem.** *Let $\mu$ be an exterior measure on $\Omega$. The family $\mathcal{E}_\mu$ of $\mu$-measurable sets is a $\sigma$-algebra, and the restriction of $\mu$ on $\mathcal{E}_\mu$ is a complete measure.*

**proof.** The proof is quite long and delicate, but the only idea it uses is the following: in order to check that $E \in \mathcal{E}_\mu$ it is sufficient to check that $\mu(S) \geq \mu(S \cap E) + \mu(S \cap E^c)$ for any $S \subset X$. Indeed, the reverse inequality comes for free from monotonicity and subadditivity of $\mu$.

First, observe that $\emptyset \in \mathcal{E}_\mu$, and that $\mathcal{E}_\mu$ is trivially stable under complementary.

Then, given $A, B \in \mathcal{E}_\mu$ and an arbitrary subset $S \subset \Omega$, use the $\mu$-measurability of $A$ to obtain

$$\mu\left((A \cup B) \cap S\right) = \mu\left((A \cup B) \cap S \cap A\right) + \mu\left((A \cup B) \cap S \cap A^c\right)$$

There follows that, using again the $\mu$-measurability of both $B$ and $A$,

$$\begin{aligned}
\mu\left((A \cup B) \cap S\right) + \mu\left((A \cup B)^c \cap S\right) &= \mu\left((A \cup B) \cap S \cap A\right) + \mu\left((A \cup B) \cap S \cap A^c\right) + \mu\left((A \cup B)^c \cap S\right) \\
&= \mu\left(S \cap A\right) + \mu\left(B \cap S \cap A^c\right) + \mu\left(A^c \cap B^c \cap S\right) \\
&= \mu\left(S \cap A\right) + \mu\left(S \cap A^c\right) \\
&= \mu\left(S\right)
\end{aligned}$$

This shows that $\mathcal{E}_\mu$ is stable under finite unions.

Also observe that, if $A$ and $B$ are disjoint elements of $\mathcal{E}_\mu$, then

$$\mu\left((A \cup B) \cap S\right) = \mu\left(A \cap S\right) + \mu\left(B \cap S\right)$$

(as follows from $\mu$-measurability of $A$, against the subset $(A \cup B) \cap S$). By induction,

$$\mu\left((\cup_n A_n) \cap S\right) = \sum_n \mu\left(A_n \cap S\right)$$

for any finite family $(A_n)$ of pairwise disjoint $\mu$-measurable sets and any $S \subset \Omega$.

Now, let $(A_n)$ be a countable family of pairwise disjoint $\mu$-measurable sets, and let $S \subset \Omega$. Since any finite union of the $A_n$'s is $\mu$-measurable, the above remark and monotonicity of $\mu$ imply that

$$\begin{aligned}
\mu\left(S\right) &= \mu\left(\left(\cup_{n=1}^N A_n\right) \cap S\right) + \mu\left(\left(\cup_{n=1}^N A_n\right)^c \cap S\right) \\
&\geq \sum_{n=1}^N \mu\left(A_n \cap S\right) + \mu\left(\left(\cup_n A_n\right)^c \cap S\right)
\end{aligned}$$

for any $N$. Letting $N \to \infty$ and using subadditivity of $\mu$ we get

$$\begin{aligned}
\mu\left(S\right) &\geq \sum_n \mu\left(A_n \cap S\right) + \mu\left(\left(\cup_n A_n\right)^c \cap S\right) \\
&\geq \mu\left(\left(\cup_n A_n\right) \cap S\right) + \mu\left(\left(\cup_n A_n\right)^c \cap S\right)
\end{aligned}$$

Our first observation implies that this last inequality is indeed an equality. This proves that the countable union $\cup_n A_n$ belongs to $\mathcal{E}_\mu$, hence, by an obvious argument, that $\mathcal{E}_\mu$ is also stable under countable unions. There follows that $\mathcal{E}_\mu$ is a $\sigma$-algebra.

The fact that the restriction $\mu\big|_{\mathcal{E}_\mu}$ is a measure (we only have to check $\sigma$-additivity) follows from the last (in)equality, setting $S = \cup_n A_n$. Also, one easily checks that any set of $\mu$-measure zero is $\mu$-measurable, hence that $\mu\big|_{\mathcal{E}_\mu}$ is complete. □

**Carathéodory's extension theorem.** A strategy to construct interesting measures on uncountable spaces is: start with an exterior measure (it is very easy to produce exterior measures, for example by means of variational principles) and then check that the $\sigma$-algebra of measurable sets is sufficiently big for our purpose. The idea of Carathéodory is the following.

Let $\mathcal{A}$ be an algebra of subsets of $\Omega$. A function $m : \mathcal{A} \to [0, 1]$ is a *probability measure* on $\mathcal{A}$ if

i) $m\left(\emptyset\right) = 0$ and $m\left(\Omega\right) = 1$,

ii) it is additive, i.e. $m\left(A \cup B\right) = m\left(A\right) + m\left(B\right)$ whenever $A$ and $B$ are disjoint,

iii) it is continuous at $\emptyset$, i.e. $A_n \downarrow \emptyset$ implies $m\left(A_n\right) \downarrow 0$.

Given a probability measure $m$ on a algebra $\mathcal{A}$ , the recipe

$$\mu(S) = \inf\left\{\sum_n m(A_n) \text{ where } S \subset \cup_n A_n \text{ with } A_n \in \mathcal{A}\right\}$$

defines an exterior measure on $\mathcal{P}(\Omega)$, hence the above construction produces a measure $\mu$ on the $\sigma$-algebra of $\mu$-measurable sets, which contains $\mathcal{A}$ and so contains $\sigma(\mathcal{A})$. One then checks that $\mu(A) = m(A)$ for any $A \in \mathcal{A}$, so that $\mu$ is an "extension" of the measure $m$. Uniqueness of the extension is an obvious consequence of the approximation theorem.

If the measure of $\Omega$ is not finite, the construction must be slightly different. We must substitute the continuity at $\emptyset$ of $m$ with the condition that if $(A_n)$ is a sequence of elements of $\mathcal{A}$ such that $S_n \downarrow \emptyset$ and $m(A_1) < \infty$ then $m(A_n) \downarrow 0$, ask that $\Omega$ is the union of an increasing sequence $(\Omega_n)$ of elements of $\mathcal{A}$ with finite measure such that $m(A) = \lim_{n\to\infty} m(A \cap \Omega_n)$ for any $A \in \mathcal{A}$, and call such a function $m$ a *$\sigma$-finite measure* on the algebra $\mathcal{A}$. Carathéodory's extension theorem is then stated in the following general form.

**Carathéodory's extension theorem.** *Given a $\sigma$-finite measure $m$ on a algebra $\mathcal{A}$ of subsets of $\Omega$, there exists a unique $\sigma$-finite measure $\mu$ on $\sigma(\mathcal{A})$ which extends $m$.*

**Lebesgue measure.** The collection $\mathcal{I}$ of intervals of the real line is a *semi-algebra*, i.e. the intersection of two elements of $\mathcal{I}$ is in $\mathcal{I}$ and the complement of an element of $\mathcal{I}$ is a union of elements of $\mathcal{I}$. The function $m : \mathcal{I} \to [0, \infty]$, defined as $m([a, b]) = |b - a|$ if $a$ e $b$ are finite, and $\infty$ if the interval is unbounded, is monotone and gives value zero to the empty set. Postulating additivity, the function $m$ extends to a measure on the algebra $\mathcal{A}$ made of disjoint unions of elements of $\mathcal{I}$ (this is not trivial!, the proof uses the Heine-Borel theorem about compact subsets of the real line). The function $\mu : \mathcal{P}(\mathbf{R}) \to [0, \infty]$, defined as

$$\mu(E) = \inf\left\{\sum m(C_n) \text{ with } E \subset \cup_n C_n \text{ e } C_n \in \mathcal{A}\right\}$$

is then an exterior measure on the real line. The $\sigma$-algebra $\mathcal{L}$ of $\mu$-measurable sets, called *Lebesgue $\sigma$-algebra*, contains the Borel sets, because it contains the intervals. The restriction $\ell = \mu|_{\mathcal{L}}$, as well as $\mu|_{\mathcal{B}(\mathbf{R})}$, is called *Lebesgue measure*.

Observe that Lebesgue measure on the real line is not a probability measure, having infinite mass. Nevertheless, one can easily define probability measures on bounded intervals taking normalized restrictions of Lebesgue measure. For example, take $\Omega = [0, 1]$, and $\mathcal{E} = \mathcal{B}(\Omega) = \{\Omega \cap B \text{ with } B \in \mathcal{B}(\mathbf{R})\}$, the Borel subsets of the interval. The restriction of $\ell$ to $\mathcal{E}$ is a probability measure, called Lebesgue measure on the unit interval.

The very same construction works in $\mathbf{R}^n$, starting with the semi-algebra of "rectangles" measured by the "euclidean volume", and produces a measure $\ell$ on $\mathcal{B}(\mathbf{R}^n)$, also called Lebesgue measure. Lebesgue measure is the unique measure over the Borel sets of the euclidean space which is invariant under traslations, i.e. $\ell(\lambda + B) = \ell(B)$ for any $\lambda \in \mathbf{R}^n$ and any Borel set $B$, and which is normalized to give measure one to the unit square, i.e. $\ell([0, 1]^n) = 1$.

The axiom of choice allows one to "give examples" of subsets wich are not Lebesgue-measurable (for example, the set made of one point for each orbit of an irrational rotation of the circle $\mathbf{R}/\mathbf{Z}$ ... ).

**Lebesgue-Stieltjes probability measures.** Let $\mathbf{P} : \mathcal{B}(\mathbf{R}) \to [0, 1]$ be a probability measure defined on the Borel subsets of the real line. The function $F : \mathbf{R} \to [0, 1]$, defined as

$$F(t) = \mathbf{P}(]-\infty, t])$$

satisfies the following properties: it is nondecreasing (as follows from monotonicity of measures), is right-continuous and admits limits of the left (as follows from monotonicity and continuity of measures), and is normalized so to have

$$\lim_{t\to-\infty} F(t) = 0 \qquad \text{and} \qquad \lim_{t\to\infty} F(t) = 1$$

(as follows from the normalization of probability measures). Any function with these properties is said a *distribution function*.

The following result says that probability measures on $\mathcal{B}(\mathbf{R})$ are indeed in one-to-one correspondence with distribution functions, and gives a method to construct nontrivial probability measures on the real line.

**Theorem.** *Given a distribution function $F : \mathbf{R} \to [0,1]$, there exists a unique probability measure $\mathbf{P} : \mathcal{B}(\mathbf{R}) \to [0,1]$ such that*

$$\mathbf{P}(]a,b]) = F(b) - F(a)$$

*for any $-\infty \leq a < b < \infty$.*

**proof.** Let $\mathcal{A}$ be the algebra made of finite unions of intervals of the real line. It is clear that $\sigma(\mathcal{A}) = \mathcal{B}(\mathbf{R})$, so that, in order to apply Carathéodory theorem, we just have to check that the formula above does define a probability measure on $\mathcal{A}$. The continuity properties of $F$ implies that it is possible to define $\mathbf{P} : \mathcal{A} \to [0,1]$, postulating finite additivity, in such a way that $\mathbf{P}(]a,b]) = F(b) - F(a)$ for any $a, b \in \mathbf{R}$. Normalization being trivial, we must verify $\sigma$-additivity, or, what is the same, continuity at the empty set. Let $(A_n)$ be a decreasing sequence of elements of $\mathcal{A}$ such that $A_n \downarrow \emptyset$. For any $\varepsilon > 0$ there exists a (sufficiently large) compact interval $K \subset \mathbf{R}$ such that $\mathbf{P}(K^c) < \varepsilon$. There follows that, setting $B_n = A_n \cap K$,

$$\mathbf{P}(A_n) \leq \mathbf{P}(B_n) + \varepsilon$$

Right continuity of $F$ implies that there exist compact subsets $K_n \subset B_n$ such that

$$\mathbf{P}(B_n \backslash K_n) \leq \varepsilon \cdot 2^{-n}$$

Since $A_n \downarrow \emptyset$ also $K_n \downarrow \emptyset$, and by the Cantor intersection theorem this implies that there exists $\overline{n}$ such that $K_n = \emptyset$ for any $n > \overline{n}$. Collecting the inequalities we get, for $n > \overline{n}$,

$$\begin{aligned}
\mathbf{P}(A_n) &= \mathbf{P}(A_n \backslash B_n) + \mathbf{P}(B_n) \\
&= \mathbf{P}(A_n \backslash B_n) + \mathbf{P}\left(B_n \backslash \cap_{k=1}^{\overline{n}} K_n\right) + \mathbf{P}\left(\cap_{k=1}^{\overline{n}+1} K_n\right) \\
&\leq \mathbf{P}(A_n \backslash B_n) + \mathbf{P}\left(\cup_{k=1}^{\overline{n}} (B_k \backslash K_k)\right) \\
&\leq \mathbf{P}(A_n \backslash B_n) + \sum_{k=1}^{\overline{n}} \mathbf{P}(B_n \backslash K_n) \\
&\leq \varepsilon + \sum_{k=1}^{\overline{n}} \varepsilon \cdot 2^{-k} \leq 2\varepsilon
\end{aligned}$$

Hence, since $\varepsilon$ was arbitrary, $\mathbf{P}(A_n) \downarrow 0$. □

**Lebesgue-Stieltjes probability measures on $\mathbf{R}^n$.** Given a probability measure $\mathbf{P}$ on the Borel $\sigma$-algebra of $\mathbf{R}^n$, one can still define its distribution function $F : \mathbf{R}^n \to [0,1]$ as

$$F(x_1, x_2, ..., x_n) = \mathbf{P}(]-\infty, x_1] \times ]-\infty, x_2] \times ... \times ]-\infty, x_n])$$

Such functions have definite continuity properties that are analogous (but slightly more complicated to write down!) to the ones of one-dimensional distribution functions. It turns out that, again, that probability measures on $\mathcal{B}(\mathbf{R}^n)$ are in one-to-one correspondence with distribution functions.

**Kolmogorov extension (finite outcomes case).** Let $X$ be a finite space, equipped with the discrete topology, and let $\Omega$ be the product

$$\Omega = X^{\mathbf{N}} = \{x : \mathbf{N} \to X\}$$

its point indentified with sequences $x = (x_1, x_2 \, ..., x_n, ...)$ with $x_n \in X$. Let $\mathcal{C}$ be the collection of *cylinders* of $X$, the subsets of the form

$$C_B = \{x \in \Omega \text{ s.t. } (x_1, x_2 \, ..., x_n) \in B\}$$

with $B \in \mathcal{B}(X^n)$. Cylinders form a basis of the product topology of $\Omega$, which makes $\Omega$ a compact metrizable space. In particular, the Borel $\sigma$-álgebra of $\Omega$ is $\mathcal{B}(\Omega) = \sigma(\mathcal{C})$

Let $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_n, ...$ be probability measures defined on the Borel sets of $X, X^2, ..., X^n, ...$ respectively. The sequence $(\mathbf{P}_n)$ is said *consistent* if

$$\mathbf{P}_{n+1}(B \times X) = \mathbf{P}_n(B)$$

for any $n$ and any Borel subset $B$ of $X^n$. The (most elementary version of) Kolmogorov extension theorem says that

**Kolmogorov extension theorem.** *Given a consistent family of probability measures as above, there exists a unique probability measure* $\mathbf{P}$*, defined on the Borel* $\sigma$*-algebra of* $\Omega$*, such that*

$$\mathbf{P}(C_B) = \mathbf{P}_n(B)$$

*for any cilinder* $C_B$*.*

**proof.** The proof consists in the following two steps. First, observe that cylinders form an algebra, and use consistency of the $\mathbf{P}_n$'s to verify that the formula above does define a function $\mathbf{P} : \mathcal{C} \to [0,1]$ on cylinders (i.e. it does not depend on the different ways the same cylinder may be presented) which is additive and properly normalized. Then, use compactness of $X$ to check that $\mathbf{P}$ is continuous at $\emptyset$, in order to apply Carathéodory theorem. Indeed, let $(A_n)$ be a sequence of cilinders such that $A_n \downarrow \emptyset$, and assume by contradiction that $\mathbf{P}(A_n) > \delta > 0$ for any $n$. This implies that $A_n \neq \emptyset$ for any $n$, but, since the $A_n$ are compact, then the Cantor intersection theorem says that $\cap_n A_n \neq \emptyset$, contrary to the hypothesis.□

Kolmogorov theorem is the key tool in probability theory, since it allows one to construct measures which describe an infinite sequence of trials starting with some rule which gives information about the $n$-th trial given the knowledge of the first $n-1$. It actually works with much more general spaces and in a more general setting.

Also, one can easily adapt the construction to the case where $\Omega = \prod_{n \in \mathbf{N}} X_n$, the topological product of a countable family of finite spaces. In some precise sense, this is a universal model of a dynamical system.

**Example: Bernoulli trials.** If $X = \{0,1\}$, then $\Omega = X^{\mathbf{N}}$ is the state space of infinite Bernoulli trials with two possible outcomes: sucess and failure. Let $\mathbf{P}_1 : \mathcal{P}(X) \to [0,1]$ be a any probability measure, defined by $\mathbf{P}_1(\{1\}) = p$. Kolmogorov construction can be applied postulating the independence of different trials, i.e. declaring that the family formed by the cilinders $\{x_n = 1\}$ is an independent family, and giving measure $p$ to each $\{x_n = 1\}$. The resulting probability space $(\Omega, \mathcal{B}(\Omega), \mathbf{P})$ describes the infinite independent Bernoulli trials. If $p = 1/2$, this is a model of infinite tosses of a fair coin.

Of course, the very same construction can be made when $X$ is a finite space with any finite numer $N$ of elements. The probability space constructed by means of the Kolmogorov theorem is a model of repeated independent experiences with $N$ possible outcomes, still called Bernoulli trials.

**A mathematical model for Bernoulli trials.** Although no physicist has ever tossed a coin infinitely often, mathematicians are able to "imagine" such an experiment! Indeed, consider the base 2 representation of a number in the unit interval $[0,1]$. It is a map $\{0,1\}^{\mathbf{N}} \to [0,1]$ given by

$$(x_k) \mapsto \sum_{k=1}^{\infty} \frac{x_k}{2^k}$$

(if you like, you can make it bijective allowing only one of the two possible representations for the ambiguous points). The push-forward of Bernoulli measure with $p = 1/2$ turns out to be Lebesgue measure on the interval. Hence, a point "randomly chosen" w.r.t. Lebesgue measure in the unit interval has a binary expansion that represents the result of an infinite sequence of coin tosses.

**Kolmogorov extension (countable real outcomes case).** Let $\Omega = \mathbf{R}^{\mathbf{N}}$

be the direct product

$$\mathbf{R}^{\mathbf{N}} = \{x = (x_1, x_2 \ ..., x_n, ...) \text{ with } x_n \in \mathbf{R}\}$$

Let $\mathcal{C}$ be the collection of *cylinders* of $\mathbf{R}^{\mathbf{N}}$, the subsets of the form

$$C_B = \{x \in X \text{ s.t. } (x_1, x_2 \ ..., x_n) \in B\}$$

with $B \in \mathcal{B}(\mathbf{R}^n)$. Open cylinders, those with $B$ open, form a basis of the product topology of $\mathbf{R}^{\mathbf{N}}$, which makes it a complete metrizable space. Let $\mathcal{B}\left(\mathbf{R}^{\mathbf{N}}\right) = \sigma(\mathcal{C})$ be the Borel $\sigma$-álgebra of $\mathbf{R}^{\mathbf{N}}$. Let $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_n, ...$ be probability measures defined on $\mathcal{B}\left(\mathbf{R}^1\right), \mathcal{B}\left(\mathbf{R}^2\right), ..., \mathcal{B}(\mathbf{R}^n)$ respectively. The sequence $(\mathbf{P}_n)$ is said *consistent* if

$$\mathbf{P}_{n+1}(B \times \mathbf{R}) = \mathbf{P}_n(B)$$

for any $n$ and any Borel subset $B \subset \mathbf{R}^n$. The Kolmogorov extension theorem now takes the following form.

**Kolmogorov extension theorem.** *Given a consistent family of probability measures as above, there exists a unique probability measure* $\mathbf{P}$*, defined on the Borel* $\sigma$*-algebra of* $\mathbf{R}^{\mathbf{N}}$*, such that*

$$\mathbf{P}(C_B) = \mathbf{P}_n(B)$$

*for any cilinder* $C_B$*.*

**proof.** As in the previous version of the theorem, we must prove that the function $\mathbf{P}$, as defined on the algebra of cylinders by the formula above, is $\sigma$-additive. Let $(A_n)$, with $A_n = C_{B_n}$ and $B_n \in \mathcal{B}(\mathbf{R}^n)$, be a decreasing sequence of cilinders such that $A_n \downarrow \emptyset$, and assume by contradiction that $\mathbf{P}(A_n) \geq \delta > 0$ for any $n$. Since each $\mathbf{P}_n$ is a probability measure on $\mathcal{B}(\mathbf{R}^n)$, and compact sets generate the Borel $\sigma$-algebra of $\mathbf{R}^n$, we can find compact sets $K_n \subset B_n$ such that

$$\mathbf{P}_n(B_n \backslash K_n) \leq \delta \cdot 2^{-n-1}$$

There follows that

$$\begin{aligned} \mathbf{P}\left(A_n \backslash \left(\cap_{k=1}^n C_{K_k}\right)\right) &\leq \mathbf{P}\left(\cup_{k=1}^n A_n \backslash C_{K_k}\right) \\ &\leq \sum_{k=1}^n \delta \cdot 2^{-n-1} \leq \delta/2 \end{aligned}$$

hence, from $\mathbf{P}(A_n) \geq \delta$, that

$$\mathbf{P}\left(\cap_{k=1}^n C_{K_k}\right) \geq \delta/2$$

for any $n$. Now, the sequence $\left(\cap_{k=1}^n C_{K_k}\right)_{n \in \mathbf{N}}$ is a decreasing sequence of nonempty compact subsets of $\mathbf{R}^{\mathbf{N}}$, so that by the Cantor intersection theorem its intersection is nonempty, contradicting the hypothesis $A_n \downarrow \emptyset$. □

# 6 Variáveis aleatórias, leis

**Variáveis aleatórias.** Seja $(\Omega, \mathcal{E}, \mathbf{P})$ um espaço de probabilidades. Uma *variável aleatória* (com valores na recta real) é uma função $\xi : \Omega \to \mathbf{R}$ tal que

$$\{\omega \in \Omega \text{ t.q. } \xi(\omega) \in A\} \in \mathcal{E}$$

para todo intervalo $A \subset \mathbf{R}$.

Por razões de economia, é uma boa ideia simplificar a notação e escrever $\{\xi \in A\}$ em vez de $\xi^{-1}(A) = \{\omega \in \Omega \text{ t.q. } \xi(\omega) \in A\}$. Outra liberdade será a de poupar os parênteses, e escrever $\mathbf{P}(\xi \in A)$ ou $\mathbf{P}\{\xi \in A\}$ em vez de $\mathbf{P}(\{\omega \in \Omega \text{ t.q. } \xi(\omega) \in A\})$.

**Observáveis e funções mensuráveis.** Se $\Omega$ é um modelo dos possíveis estados de um sistema físico, os observáveis da física são funções reais $\xi$ definidas em $\Omega$. Fazer observações quer dizer ler resultados experimentais do tipo $\xi = a$, ou $\xi \leq a$ ou $a < \xi < b$ nos instrumentos do laboratório. Se o modelo físico é um modelo probabilístico, o que queremos é saber calcular as probabilidades de obter certos resultados. Se $\xi : \Omega \to \mathbf{R}$ é uma função arbitrária e $A \subset \mathbf{R}$, o conjunto $\{\xi \in A\}$ pode não ser um evento. Afinal, a imagem inversa é só uma função $\xi^{-1} : \mathcal{P}(\mathbf{R}) \to \mathcal{P}(\Omega)$, e em geral não tem mais propriedades. A definição de variável aleatória diz que os conjuntos do tipo $\{\xi > x\}$, $\{y < \xi \leq x\}$, $\{\xi = x\}$, $\{y < \xi < x\}$, ... são eventos.

De facto, pelo lema do transporte, a definição implica (e portanto é equivalente à condição) que são eventos todos os conjuntos $\{\xi \in A\}$ em que $A$ é um boreleano da recta real, pois a $\sigma$-álgebra dos borelanos é gerada pela família dos intervalos. Por outras palavras, se $\mathcal{B}(\mathbf{R})$ denota a $\sigma$-álgebra dos boreleanos da recta real, uma variável aleatória é uma função mensurável de $(\Omega, \mathcal{E})$ em $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$, uma função $\xi : \Omega \to \mathbf{R}$ tal que $\xi^{-1}(A) \in \mathcal{E}$ para todo $A \in \mathcal{B}(\mathbf{R})$. Como a $\sigma$-álgebra dos boreleanos da recta real é gerada pela família de intervalos $\{]-\theta, t] \text{ com } t \in \mathbf{Q}\}$, o lema do trasporte implica que uma variável aleatória pode ser definida como uma função $\xi : \Omega \to \mathbf{R}$ tal que

$$\{\omega \in \Omega \text{ t.q. } \xi(\omega) \leq t\} \in \mathcal{E}$$

para todo $t \in \mathbf{Q}$.

Observe que, se $\mathcal{E} = \mathcal{P}(\Omega)$, então toda função $\xi : \Omega \to \mathbf{R}$ é uma variável aleatória.

Uma ideia intuitiva do significado da mensurabilidade é sugerida pela seguinte observação: se $\mathcal{E}$ é a $\sigma$-álgebra gerada pela partição $\mathcal{D} = \{S_1, S_2, ..., S_n, ...\}$, então $\xi$ é mensurável sse é constante nos átomos de $\mathcal{D}$.

Se $\xi : \Omega \to \mathbf{R}$ é uma variável aleatória, a família $\mathcal{E}_\xi = \xi^{-1}(\mathcal{B}(\mathbf{R}))$ é uma sub-$\sigma$-álgebra de $\mathcal{E}$, dita a $\sigma$-álgebra *gerada* por $\xi$. É a coleção de eventos acerca dos quais a observação da variável fornece informações. Observe que $\mathcal{E}_\xi$ pode ser caracterizada como sendo a menor das $\sigma$-álgebras $\mathcal{E}' \subset \mathcal{P}(\Omega)$ tais que $\xi : \Omega \to \mathbf{R}$ seja uma função mensurável de $(\Omega, \mathcal{E}')$ em $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$

Se $\xi : \Omega \to \mathbf{R}$ é uma variável aleatória e $\varphi : \mathbf{R} \to \mathbf{R}$ uma função arbitrária, a função composta $\varphi \circ \xi$ pode não ser uma variável aleatória. Uma condição suficiente para que $\varphi \circ \xi$ seja uma variável aleatória é que $\varphi$ seja Borel-mensurável, i.e. tal que $\varphi^{-1}(A) \in \mathcal{B}(\mathbf{R})$ para todo $A \in \mathcal{B}(\mathbf{R})$. Isto acontece se, por exemplo, $\varphi$ é contínua ou contínua por pedaços.

**Valores infinitos.** Existem situações físicas em que é natural admitir valores infinitos para uma variável (por exemplo, um tempo de espera). Uma *variável aleatória generalizada* é uma função $\xi : \Omega \to \overline{\mathbf{R}} = \mathbf{R} \cup \{\pm\infty\}$ tal que $\{\omega \in \Omega \text{ t.q. } \xi \leq x\} \in \mathcal{E}$ para todo $x \in \mathbf{R}$ (a ordem na recta real extendida é tal que $-\infty < x < \infty$ para todo $x \in \mathbf{R}$). Em particular, $\{\xi = \infty\}$ e $\{\xi = -\infty\}$ são eventos.

**Variáveis simples.** Uma função constante é uma variável aleatória. Se $S$ é um evento, a função característica de $S$, definida por

$$1_S(\omega) = \begin{cases} 1 & \text{se } \omega \in S \\ 0 & \text{se } \omega \notin S \end{cases}$$

é uma variável aleatória.

Uma variável aleatória $\xi : \Omega \to \mathbf{R}$ que assume uma quantidade finita de valores, i.e. tal que $|\xi(\Omega)| < \infty$, é dita *simples*. É imediato verificar que toda variável simples é da forma

$$\xi = \sum_{k=1}^{n} x_k \cdot 1_{S_k}$$

onde $\mathcal{D} = \{S_1, S_2, ..., S_n\}$ é uma partição de $\Omega$ com $S_k \in \mathcal{E}$, e $x_1, x_2, ..., x_n$ são números reais. A representação acima é única se decidimos que $x_i \neq x_j$ quando $i \neq j$, pois, neste caso, $\{x_1, x_2, ..., x_n\} = \xi(\Omega)$ e podemos escrever

$$\xi = \sum_{x_k \in \xi(\Omega)}^{n} x_k \cdot 1_{\{\xi = x_k\}}$$

A $\sigma$-álgebra $\mathcal{E}_\xi$ é a álgebra gerada pela partição $\mathcal{D}$.

É imediato verificar que combinações lineares, funções arbitrárias, assim como produtos e quocientes (desde que sejam definidos), de variáveis aleatórias simples são variáveis aleatórias simples.

**Variáveis discretas.** Uma variável que assume uma quantidade enumerável de valores é dita *discreta*. Toda variável discreta é da forma

$$\xi = \sum_{k} x_k \cdot 1_{S_k}$$

onde $\mathcal{D} = \{S_1, S_2, ..., S_k, ...\}$ é uma partição enumerável de $\Omega$, e $x_1, x_2, ..., x_k, ...$ são números reais, os seus valores. A $\sigma$-álgebra gerada por $\xi$ é $\mathcal{E}_\xi = \sigma(\mathcal{D})$.

Toda função $\xi : \Omega \to \{x_1, x_2, x_3, ...\} \subset \mathbf{R}$ cuja imagem é um subconjunto enumerável da recta real e tal que $\{\xi = x_k\} \in \mathcal{E}$ para todo $k = 1, 2, ...$ é uma variável aleatória discreta, pois $\{\xi \leq x\} = \cup_{x_i \leq x} \{\xi = x_i\}$ para todo $x \in \mathbf{R}$.

**Exercícios.**

**a.** Verifique as seguintes relações, que descrevem a correspondência entre a álgebra (de Boole) das partes de $\Omega$ e a álgebra das funções características (observe que a função constante e igual a um pode ser interpretada como sendo $1 = 1_\Omega$ ):

$$1_{\overline{A}} = 1 - 1_A \qquad 1_{A \cap B} = 1_A \cdot 1_B \qquad 1_{A \cup B} = 1 - 1_{A^c \cap B^c} = 1_A + 1_B - 1_{A \cap B}$$

Determine a função característica do evento $A \triangle B$ em função de $1_A$ e $1_B$. Mostre que a função característica do complementar de $A_1 \cup A_2 \cup ... \cup A_n$ é

$$\prod_{k=1}^{n} (1 - 1_{A_k})$$

**b.** Sejam

$$\xi = \sum_{k=1}^{n} x_k 1_{A_k} \qquad \text{e} \qquad \eta = \sum_{k=1}^{m} y_k 1_{B_k}$$

duas variáveis aleatórias simples, onde $\{A_1, A_2, ..., A_n\}$ e $\{B_1, B_2, ..., B_m\}$ são partições de $\Omega$ e os $x_k$ e $y_k$ números reais. Determine expressões análogas para as variáveis

$$a\xi + b \qquad \text{e} \qquad \xi + \eta.$$

onde $a, b \in \mathbf{R}$. Deduza que o espaço das variáveis aleatórias simples (com valores reais) é uma $\mathbf{R}$-álgebra.

Determine também expressões análogas para as variáveis $\xi^k$ com $k \in \mathbf{N}$, $\sin \xi$, $\exp \xi$, $\xi\eta$, $\max\{\xi, \eta\}$ e $\min\{\xi, \eta\}$.

**Aproximação.** Toda variável aleatória $\xi : \Omega \to \mathbf{R}_{\geq 0}$ com valores não negativos é o limite pontual de uma sucessão crescente de variáveis aleatórias simples e não negativas. Por exemplo, a sucessão $(\xi_n)$ de variáveis simples definidas por

$$\xi_n = \sum_{k=0}^{n2^n} \frac{k}{2^n} \cdot 1_{\{k/2^n \leq \xi < (k+1)/2^n\}}$$

é tal que[1]

$$0 \leq \xi_1 \leq \xi_2 \leq ... \leq \xi_n \leq ... \leq \xi \qquad \text{e} \qquad \xi_n \uparrow \xi$$

**álgebra das variáveis aleatórias e limites.** Se $\xi$ e $\eta$ são variáveis aleatórias (definidas no mesmo espaço de probabilidades), as funções $\max\{\xi, \eta\}$ e $\min\{\xi, \eta\}$ são variáveis aleatórias. Em particular, toda variável aleatória $\xi$ é igual a diferença $\xi^+ - \xi^-$ de duas variáveis aleatórias não negativas, definidas por

$$\xi^+ = \max\{\xi, 0\} \qquad \xi^- = -\min\{\xi, 0\}$$

Em particular, se $\xi$ é uma variável aleatória, o seu módulo $|\xi| = \xi^+ + \xi^-$ é também uma variável aleatória. Aproximando $\xi^+$ e $\xi^-$ com variáveis simples, vê-se que toda variável aleatória $\xi$ é o limite pontual de uma sucessão $(\xi_n)$ de variáveis simples tais que $|\xi_n| \uparrow |\xi|$.

Se $(\xi_n)$ é uma sucessão de variáveis aleatórias, então $\sup \xi_n$ , $\overline{\lim}\xi_n$ , $\inf \xi_n$ e $\underline{\lim}\xi_n$ são variáveis aleatórias (admitindo $\pm\infty$ como valores possíveis!), e portanto também o limite pontual $\lim \xi_n$, se existir. A tal fim, basta observar que

$$\{\sup \xi_n > x\} = \cup_n \{\xi_n > x\} \qquad \text{e} \qquad \{\inf \xi_n < x\} = \cup_n \{\xi_n < x\}$$

e que

$$\overline{\lim}\xi_n = \inf_n \sup_{k \geq n} \xi_k \qquad \text{e} \qquad \underline{\lim}\xi_n = \sup_n \inf_{k \geq n} \xi_k$$

Combinações lineares, assim como produtos e quocientes (desde que sejam definidos), de variáveis aleatórias são variáveis aleatórias. A prova pode ser dada aproximando as variáveis com variáveis simples, ou directamente brincando com a definição.

**Exercícios.**

**a.** Se $\xi$ e $\eta$ são variáveis aleatórias, então $\{\xi = \eta\}$ e $\{\xi > \eta\}$ são eventos.

**b.** Se $\xi_1, \xi_2, ..., \xi_n$ são variáveis aleatórias e $\varphi : \mathbf{R}^n \to \mathbf{R}$ é uma função Borel-mensurável, então $\varphi(\xi_1, \xi_2, ..., \xi_n)$ é uma variável aleatória.

**Função de repartição e lei.** Seja $\xi : \Omega \to \mathbf{R}$ uma variável aleatória definida no espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$. A *função de repartição* (ou de *distribuição*) da variável aleatória $\xi$ é a função $F_\xi : \mathbf{R} \to [0, 1]$ definida por

$$F_\xi(x) = \mathbf{P}\{\xi \leq x\}$$

(a "probabilidade da variável $\xi$ ser menor ou igual a $x$").

A *lei* da variável aleatória $\xi$ é a função $\mathbf{P}_\xi : \mathcal{B}(\mathbf{R}) \to [0, 1]$ que associa a um boreleano $A \subset \mathbf{R}$ a probabilidade

$$\mathbf{P}_\xi(A) = \mathbf{P}\{\xi \in A\}$$

(a "probabilidade da variável $\xi$ pertencer a $A$"). A lei uma medida de probabilidades sobre os boreleanos da recta real, a medida imagem de $\mathbf{P}$ pela aplicação $\xi$.

A relação entre a lei e a função de repartição de uma variável aleatória $\xi$ é a seguinte:

$$F_\xi(x) = \mathbf{P}_\xi(]-\infty, x]) \qquad \text{e} \qquad \mathbf{P}_\xi(]a, b]) = F_\xi(b) - F_\xi(a)$$

(ou seja, a função de repartição é a restrição da lei à família dos intervalos do género $]-\infty, x]$).

[1] "$\leq$" denota a ordem parcial natural no espaço das funções com valores reais, i.e. $\xi \leq \eta$ quer dizer que $\xi(\omega) \leq \eta(\omega)$ para todo $\omega \in \Omega$. A notação $\xi_n \uparrow \xi$ quer dizer que a sucessão de funções $(\xi_n)$ é crescente, i.e. $\xi_n(\omega) \leq \xi_{n+1}(\omega)$ para todo $\omega \in \Omega$ e todo $n \in \mathbf{N}$, e que $\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)$ para todo $\omega \in \Omega$.

**Funções de repartição.** A função de repartição $F_\xi$ de uma variável aleatória satisfaz as seguintes propriedades:

i) é uma função crescente, porque $\{\xi \leq x\} \subset \{\xi \leq x'\}$ se $x < x'$ e porque a medida de probabilidades é monótona,

ii) é contínua à direita, porque

$$F_\xi(x) = \mathbf{P}\left(\cap_{n\geq 1}\{\xi \leq x + 1/n\}\right) = \lim_{n\to\infty} F_\xi(x + 1/n) = \lim_{y\downarrow x} F_\xi(y)$$

(onde utilizamos a monotonia de $F_\xi$ e a continuidade da medida de probabilidades),

iii) admite o limite à esquerda, porque

$$\lim_{y\uparrow x} F_\xi(y) = \lim_{n\to\infty} F_\xi(x - 1/n) = \mathbf{P}\left(\cup_{n\geq 1}\{\xi \leq x - 1/n\}\right) = \mathbf{P}\{\xi < x\}$$

(onde utilizamos a monotonia de $F_\xi$ e a continuidade da medida de probabilidades),

iv) e satisfaz a normalização

$$\lim_{x\to-\infty} F_\xi(x) = 0 \quad \text{e} \quad \lim_{x\to\infty} F_\xi(x) = 1$$

porque $\cap_{n\geq 1}\{\xi \leq -n\} = \emptyset$ e $\cup_{n\geq 1}\{\xi \leq n\} = \Omega$.

Em geral, uma função $F : \mathbf{R} \to [0, 1]$ com estas propriedades é dita uma *função de repartição*.

Uma função de repartição pode não ser contínua. O que acontece é que

$$\mathbf{P}\{\xi = x\} = F_\xi(x) - \lim_{y\uparrow x} F_\xi(y)$$

e portanto, se $F_\xi$ é contínua em $x$, a probabilidade $\mathbf{P}\{\xi = x\}$ é igual a zero.

**Construção de variáveis, medidas de Lebesgue-Stieltjes.** A lei, assim como a função de repartição, de uma variável aleatória contém toda a informação relevante acerca da variável. Ou seja, é sempre possível "construir" uma variável aleatória, i.e. definir um espaço de probabilidades e uma função mensurável, que tem uma dada lei ou uma dada função de repartição. A construção esboçada em baixo é dita a "realização standard" da variável.

Por um lado, dada uma probabilidade $\mathbf{P}$ sobre os boreleanos da recta real, podemos definir um espaço de probabilidades como $(\mathbf{R}, \mathcal{B}(\mathbf{R}), \mathbf{P})$ e uma variável aleatória $\xi : \mathbf{R} \to \mathbf{R}$ como $x \mapsto \xi(x) = x$. É evidente que então a lei de $\xi$ é $\mathbf{P}$.

Por outro lado, dada uma função de repartição $F : \mathbf{R} \to [0, 1]$, então existe uma única medida de probabilidade $\mathbf{P}$ definida sobre os boreleanos da recta real tal que

$$\mathbf{P}(]a, b]) = F(b) - F(a)$$

para todos $a < b$. Esta medida é dita *medida de Lebesgue-Stieltjes* associada à função $F$, e é construída por meio do teorema de Carathéodory. A construção acima então produz uma variavel aleatória, definida no espaço de probabilidades $(\mathbf{R}, \mathcal{B}(\mathbf{R}), \mathbf{P})$, com função de repartição $F$.

**Densidade discreta.** Seja $\xi : \Omega \to \{x_1, x_2, x_3, ...\}$ uma variável aleatória discreta. A *densidade discreta* (ou *distribuição*) de $\xi$ é a função $p_\xi : \{x_1, x_2, x_3, ...\} \to [0, 1]$ definida por

$$p_\xi(x_k) = \mathbf{P}(\xi = x_k)$$

(a "probabilidade da variável $\xi$ ser igual a $x_k$").

A densidade discreta de $\xi$ determina (e é determinada por) a lei de $\xi$, pois

$$\mathbf{P}(\xi \in A) = \mathbf{P}\left(\cup_{x_k\in A}\{\xi = x_k\}\right) = \sum_{x_k\in A} \mathbf{P}(\xi = x_k)$$

para todo boreleano $A \subset \mathbf{R}$, sendo os eventos $\{\xi = x_k\}$ dois a dois disjuntos. Em particular, a densidade discreta determina (e é determinada por) a função de repartição, pois

$$F_\xi(x) = \mathbf{P}(\xi \leq x) = \sum_{x_k\leq x} \mathbf{P}(\xi = x_k)$$

Se os valores da variável são ordenados de tal maneira que $... < x_n < x_{n+1} < ...$, então $F_\xi$ é constante em cada intervalo $[x_n, x_{n+1}[$ e satisfaz

$$F_\xi(x_n) = F_\xi(x_{n-1}) + \mathbf{P}(\xi = x_n) \qquad \text{e} \qquad \mathbf{P}(\xi = x_n) = F_\xi(x_n) - F_\xi(x_{n-1})$$

**Construção de variáveis discretas.** Toda função $p : \{x_1, x_2, x_3, ...\} \to [0,1]$ tal que $\sum_{x_k} p(x_k) = 1$ é a densidade discreta de uma variável aleatória. Basta pôr $\Omega = \{x_1, x_2, x_3, ...\}$ , $\mathcal{E} = \mathcal{P}(\Omega)$ , $\mathbf{P}(A) = \sum_{x_k \in A} p(x_k)$ para todo $A \subset \Omega$, e $\xi : \Omega \to \mathbf{R}$ definida por $\xi(x_k) = x_k$. Portanto, a densidade discreta contém toda a informação sobre a variável aleatória discreta (podemos esquecer o espaço de probabilidades onde ela foi definida!). Por outras palavras, uma variável aleatória discreta é para todos os efeitos uma variável aleatória definida num espaço de probabilidades enumerável.

Isso explica por que nos manuais elementares de estatística uma variável aleatória discreta é um "objecto que pode assumir os valores $x_1, x_2, x_3, ...$ com probabilidades $p_1, p_2, p_3, ...$".

**Leis.** Os estatísticos utilizam a palavra "lei" também num sentido genérico. Duas variáveis definidas em espaços de probabilidades diferentes que têm a mesma lei são essencialmente indistinguíves. É por isso que, uma vez definido um conjunto de variáveis significativas (binomial, geométrica, de Poisson, gaussiana, exponencial, ...), utilizam expressões do tipo "seja $\xi$ uma variável com lei de Poisson", e poupam, justamente, o trabalho de especificar o espaço de probabilidades onde a variável está definida.

**Exercícios.**

**a.** Determine a densidade discreta da variável aleatória $\xi$ com valores em $\mathbf{N}$ e função de repartição $F_\xi(k) = 1 - p^k$.

**b.** Sejam $\xi$ uma variável aleatória, e $\eta = a\xi + b$ onde $a, b \in \mathbf{R}$ com $a > 0$. Mostre que

$$\mathbf{P}\{\eta = t\} = \mathbf{P}\left\{\xi = \frac{t-b}{a}\right\} \qquad \text{e} \qquad F_\eta(t) = F_\xi\left(\frac{t-b}{a}\right)$$

**c.** Seja $\xi$ uma variável aleatória com função de repartição $F_\xi$. Determine a função de repartição das variáveis

$$\xi^+ = \max\{\xi, 0\} \qquad \xi^- = -\min\{\xi, 0\} \qquad |\xi| \qquad \xi^k \qquad \sin\xi \qquad \exp\xi \qquad a\xi + b$$

onde $a, b \in \mathbf{R}$ e $k \in \mathbf{N}$.

**Famílias de variáveis, processos estocásticos.** Os teoremas interessantes da teoria das probabilidades são afirmações acerca de famílias de variáveis aleatórias. Dependendo do contexto, ou seja do fenómeno físico do qual é um modelo, uma coleção de variáveis é pensada como um vector aleatório, um processo, um sistema de partículas...

Um *vector aleatório* é uma função

$$\xi = (\xi_1, \xi_2, ..., \xi_n) : \Omega \to \mathbf{R}^n$$

tal que as suas $n$ coordenadas $\xi_1, \xi_2, ...$ e $\xi_n$ são variáveis aleatória com valores reais. A função de repartição do vector aleatório $\xi$ é a função $F_\xi : \mathbf{R}^n \to [0,1]$ definida por

$$F_\xi(x_1, x_2, ..., x_n) = \mathbf{P}(\{\xi_1 \le x_1\} \cap \{\xi_2 \le x_2\} \cap ... \cap \{\xi_n \le x_n\})$$

A lei de $\xi$ é a função $\mathbf{P}_\xi : \mathcal{B}(\mathbf{R}^n) \to [0,1]$, que associa $\mathbf{P}(\xi \in A)$ a cada boreleano $A$ de $\mathbf{R}^n$. Um vector aleatório é, portanto, uma família de $n$ variáveis aleatórias com valores reais definidas num mesmo espaço de probabilidades.

Um *processo estocástico* é uma família $\xi = (\xi_t)_{t \in T}$ de variáveis aleatórias com valores reais, definidas num espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$, onde $T$ é um subconjunto da recta real. O parâmetro $t \in T$ neste caso tem a interpretação de um "tempo", e tipicamente $T = \mathbf{R}$, ou $\mathbf{Z}$, ou $\mathbf{R}_{\ge 0}$, ou $\mathbf{N}$. A cada ponto $\omega \in \Omega$ está associada uma *trajectória* $\xi(\omega) : T \to \mathbf{R}$, definida por

$$t \mapsto \xi_t(\omega)$$

e a probabilidade $\mathbf{P}$ pode ser pensada como uma probabilidade definida numa $\sigma$-álgebra de partes do espaço das trajectórias.

Outra interpretação, por exemplo em mecânica estatística, é pensar $(\xi_t)_{t\in T}$ como uma coleção de variáveis que descrevem as "partículas", ou as componentes "microscópicas", de um sistema "macroscópico". Neste caso o parâmetro $t \in T$ é pensado como uma etiqueta que identifica as diferentes partículas, ou a posição delas, e vive em $\mathbf{N}$, $\mathbf{Z}$ ou em outros retículos como por exemplo $\mathbf{Z}^n$.

**Independência.** As variáveis aleatórias $\xi_1, \xi_2, ..., \xi_n$ são *independentes* (ou *formam uma família de variáveis independentes*) se para todos boreleanos (ou para todos intervalos) $A_1, A_2, ..., A_n \subset \mathbf{R}$

$$\mathbf{P}\left(\{\xi_1 \in A_1\} \cap \{\xi_2 \in A_2\} \cap ... \cap \{\xi_n \in A_n\}\right) = \mathbf{P}\left(\xi_1 \in A_1\right) \cdot \mathbf{P}\left(\xi_2 \in A_2\right) \cdot ... \cdot \mathbf{P}\left(\xi_n \in A_n\right)$$

O significado desta definição é o seguinte. Cada uma das variáveis define uma $\sigma$-álgebra $\mathcal{E}_{\xi_i} = \xi_i^{-1}\left(\mathcal{B}\left(\mathbf{R}\right)\right)$, contida em $\mathcal{E}$, que tem a interpretação da família de eventos observados ao observar $\xi_i$. A condição acima é equivalente à independência das $\sigma$-álgebras $\mathcal{E}_{\xi_1}$, $\mathcal{E}_{\xi_2}$, ..., $\mathcal{E}_{\xi_n}$.

A sucessão de variáveis aleatórias $(\xi_n)$ é uma *sucessão de variáveis independentes* se, para cada $n \in \mathbf{N}$, as variáveis $\xi_1, \xi_2, ..., \xi_n$ são independentes. Mais em geral, a família $(\xi_t)_{t\in T}$ de variáveis aleatórias é uma *família independente* se toda subfamília finita $\xi_{t_1}, \xi_{t_2}, ..., \xi_{t_n}$ é uma família de variáveis independentes.

**Exercício.** Se $\xi_1$, $\xi_2$, ...,$\xi_n$ são independentes e $\varphi_1, \varphi_2, ..., \varphi_n$ são funções reais Borel mensuráveis (por exemplo contínuas) então as variáveis $\varphi_1 \circ \xi_1, \varphi_2 \circ \xi_2, ..., \varphi_n \circ \xi_n$ também são independentes.

**V. a.'s i.i.d.** Tem interesse, sobretudo para formular teoremas de convergência, considerar sucessões $(\xi_n)_{n\in\mathbf{N}}$ de variáveis aleatórias tais que: são definidas num mesmo espaço de probabilidades, têm todas a mesma lei (a saber, a lei de uma variável $\xi$ fixada), e são independentes (ou seja, todo subconjunto finito delas é um conjunto de variáveis independentes). Os probabilistas chamam tais sucessões *variáveis aleatórias independentes e identicamente distribuídas*, e utilizam expressões como "sejam $\xi_1, \xi_2, \xi_3, ...$ v.a.'s i.i.d.".

**Construção de processos.** Dado um espaço de probabilidades é sempre possível "inventar" uma família $(\xi_t)_{t\in T}$ de variáveis aleatórias, portanto os processos estocásticos "existem". Menos óbvio é que existam processos "interessantes", ou seja, modelos úteis de determinados fenómenos físicos. Um fenómeno físico sugere que as $\xi_t$ satisfaçam certas propriedades, codificadas por exemplo por meio da coleção de *distribuções conjuntas* de dimensão finita

$$\mathbf{P}\left(\{\xi_{t_i} \leq x_1\} \cap \{\xi_{t_2} \leq x_2\} \cap ... \cap \{\xi_{t_n} \leq x_n\}\right)$$

onde $t_1, t_2, ..., t_n \in T$, ou por meio de condições mais sintéticas que determinam a "dependência" entre as variáveis. A "construção" de processos a partir desta informação é possível utilizando o teorema de extensão de Kolmogorov, cujo caso mais geral tem o nome de "teorema de Kolmogorov-Ionesco-Tulcea".

**Densidade conjunta e independência de variáveis discretas.** Seja $\{\xi, \eta, ..., \varsigma\}$ uma família finita de variáveis aleatórias discretas. A *densidade conjunta* das variáveis $\xi, \eta, ..., \varsigma$ é a função $p : \xi\left(\Omega\right) \times \eta\left(\Omega\right) \times ... \times \varsigma\left(\Omega\right) \to [0, 1]$ definida por

$$p(x_i, y_j, ..., z_k) = \mathbf{P}\left(\{\xi = x_i\} \cap \{\eta = y_j\} \cap ... \cap \{\varsigma = z_k\}\right)$$

A densidade conjunta das variáveis $\xi, \eta, ..., \varsigma$ determina as densidades de cada uma delas, pois, por exemplo,

$$\begin{aligned} \mathbf{P}\left(\xi = x_i\right) &= \sum_{y_j,...,z_k} \mathbf{P}\left(\{\xi = x_i\} \cap \{\eta = y_j\} \cap ... \cap \{\varsigma = z_k\}\right) \\ &= \sum_{y_j,...,z_k} p(x_i, y_j, ..., z_k) \end{aligned}$$

pela fórmula da probabilidade total. O contrário é, em geral, falso.

A função $(\xi, \eta, ..., \varsigma) : \Omega \to \mathbf{R}^n$ definida por

$$(\xi, \eta, ..., \varsigma)(\omega) = (\xi(\omega), \eta(\omega), ..., \varsigma(\omega))$$

é um vector aleatório, e portanto a densidade conjunta das variáveis $\xi, \eta, ..., \varsigma$ pode ser pensada como a densidade discreta de $(\xi, \eta, ..., \varsigma)$. As densidades discretas $p_\xi, p_\eta, ..., p_\varsigma$ das variáveis $\xi, \eta, ..., \varsigma$ são ditas *densidades marginais* do vector aleatório $(\xi, \eta, ..., \varsigma)$.

As variáveis aleatórias discretas $\xi, \eta, ..., \varsigma$ são independentes sse a densidade conjunta é da forma

$$p(x_i, y_j, ..., z_k) = p_\xi(x_i) \cdot p_\eta(y_j) \cdot ... \cdot p_\varsigma(z_k)$$

**Exercícios.**

**a.** Uma variável aleatória $\xi$ é independente de si mesma sse é constante com probabilidade um, i.e. se existe $a \in \mathbf{R}$ tal que $\mathbf{P}(\xi = a) = 1$.

**b.** Pode acontecer que, dada uma função Borel mensurável $\varphi$, as variáveis aleatória $\xi$ e $\varphi \circ \xi$ são independentes, sem que $\xi$ seja constante. Por exemplo, se $\xi$ tem valores $\pm 1$, então $\xi$ é independente de $|\xi|$.

**c.** Determine quando as variáveis $\xi$ e $\sin \xi$ são independentes.

**d.** *(min e max)* Sejam $\xi_1, \xi_2, ..., \xi_n$ variáveis aleatórias independentes, e sejam

$$\xi_{\max} = \max\{\xi_1, \xi_2, ..., \xi_n\} \qquad \text{e} \qquad \xi_{\min} = \min\{\xi_1, \xi_2, ..., \xi_n\}$$

Mostre que

$$\mathbf{P}\{\xi_{\min} > x\} = \prod_{k=1}^{n} \mathbf{P}\{\xi_k > x\} \qquad \text{e} \qquad \mathbf{P}\{\xi_{\max} < x\} = \prod_{k=1}^{n} \mathbf{P}\{\xi_k < x\}$$

**e.** *(um dado e uma moeda)* Um modelo do lançamento de um dado e uma moeda é: $\xi = 1, 2, ..., 6$ e $\eta = 0, 1$ (cara ou coroa) com densidade conjunta $\mathbf{P}(\xi = i \ , \ \eta = j) = p(i, j) = 1/12$ para todos $i = 1, 2, ..., 6$ e $j = 0, 1$. As variáveis $\xi$ e $\eta$ são independentes, e têm densidades (marginais) $p_\xi(i) = 1/6$ para todos $i = 1, 2, ..., 6$ e $p_\eta(j) = 1/2$ para todos $j = 0, 1$.

**f.** *(escolher bolinhas com e sem reposição)* Retiro duas vezes uma bolinha duma caixa com 6 bolinhas numeradas de 1 até 6. Sejam $\xi =$"número da primeira bolinha" e $\eta =$"número da segunda bolinha". As variáveis $\xi$ e $\eta$ são independentes, e têm densidade conjunta $p(i, j) = 1/36$ para todos pares $i, j$.

Retiro duas bolinhas duma caixa com 6 bolinhas numeradas de 1 até 6. Neste caso as variáveis $\xi$ e $\eta$, defnidas como acima, não são independentes, e a densidade conjunta é $p(i, j) = 1/30$ se $i \neq j$ e 0 se $i = j$.

As densidades (marginais) de $\xi$ e $\eta$ são iguais nas duas experiências!

**g.** *(densidade da soma de duas variáveis)* Sejam $\xi$ e $\eta$ variáveis aleatórias discretas independentes com valores inteiros. Então a variável $\xi + \eta$ tem densidade discreta

$$\mathbf{P}(\xi + \eta = k) = \sum_{i+j=k} \mathbf{P}(\xi = i) \cdot \mathbf{P}(\eta = j)$$

**h.** *(provas de Bernoulli)* No modelo das $n$ provas de Bernoulli, as variáveis $\xi_1$, $\xi_2$, ...,$\xi_n$, definidas por $\xi_k(\omega) = \omega_k$ formam uma família de variáveis independentes e identicamente distribuídas. A soma $S_n = \xi_1 + \xi_2 + ... + \xi_n$ é a variável que representa "o número de sucessos em $n$ provas".

**i.** *(marcha aleatória)* A marcha aleatória modelada no esquema de Bernoulli é o processo estocástico $(T_n)$ definido por $T_n = \xi_1 + \xi_2 + ... + \xi_n$, onde $(\xi_k)$ são v.a.'s i.i.d. com valores $\pm 1$ e lei $\mathbf{P}(\xi_k = 1) = p$.

**k.** Sejam $\xi_1, \xi_2, \xi_3, ...$ v.a.'s i.i.d. com valores $\pm 1$ e densidade discreta $\mathbf{P}(\xi_k = \pm 1) = 1/2$, e sejam $\eta_n = \prod_{k=1}^{n} \xi_k$. Mostre que $(\eta_n)$ é uma família de variáveis independentes.

# 7 Valor médio, variância e covariância

**Valor médio.** Seja $\xi = \sum_{k=1}^{n} x_k \cdot 1_{S_k}$ uma variável aleatória simples definida no espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$. O *valor médio* (ou *média,* ou *esperança*) de $\xi$ é

$$\mathbf{E}\xi = \sum_{k=1}^{n} x_k \cdot \mathbf{P}\left(S_k\right)$$

A variável aleatória discreta $\xi = \sum_k x_k \cdot 1_{S_k}$, é dita *integrável* se

$$\sum_k |x_k| \cdot \mathbf{P}\left(S_k\right) < \infty$$

O *valor médio* (ou *média,* ou *esperança*) da variável aleatória discreta integrável $\xi$ é

$$\mathbf{E}\xi = \sum_k x_k \cdot \mathbf{P}\left(S_k\right)$$

Observe que, se os valores $x_k$ são dois a dois distintos, então a média admite a seguinte expressão em termos da densidade discreta:

$$\begin{aligned} \mathbf{E}\xi &= \sum_k x_k \cdot \mathbf{P}\left(\xi = x_k\right) \\ &= \sum_k x_k \cdot p_\xi\left(x_k\right) \end{aligned}$$

Uma notação tradicional é $\mathbf{E}\xi = m$, ou $m_\xi$ se é importante lembrar que é o valor médio da variável $\xi$.

A definição da média para variáveis arbitrárias é mais delicada, e é esboçada mais à frente. Na linguagem da análise, $\mathbf{E}\xi$ é "o integral de Lebesgue da função $\xi$ com respeito à medida $\mathbf{P}$", e é denotado por $\int_\Omega \xi\left(\omega\right) d\mathbf{P}\left(\omega\right)$.

**Porque a esperança se chama esperança?** Se $\xi$ é um modelo dos possíveis resultados de uma experiência, e repetimos a experiência um número grande de vezes, a interpretação física da lei dos grandes números diz que com probabilidade muito grande a média aritmética dos resultados observados, ou seja a "média empírica" observada, está próxima de $\mathbf{E}\xi$.

**Integrabilidade.** Evidentemente, seria possível definir o valor médio de uma variável que não é integrável (a soma de uma série convergente que não é absolutamente convergente), mas isto não é particularmente interessante... A seguir, a afirmação $\mathbf{E}\left|\xi\right| < \infty$ será sinónimo de "a variável $\xi$ é integrável", e a afirmação $\mathbf{E}\xi = m$ será sinónimo de "a variável $\xi$ é integrável e o seu valor médio é igual a $m$".

**Propriedades da média.** A média deve ser pensada como um operador

$$\mathbf{E} : \{\text{variáveis aleatórias (discretas) integráveis}\} \to \mathbf{R}$$

uma função que associa um valor $\mathbf{E}\xi$ a cada variável integrável $\xi$. As seguintes propriedades do valor médio são triviais para variáveis simples, e facilmente generalizadas às variáveis discretas utilizando a álgebra das séries absolutamente convergentes.

Se $A$ é um evento e $1_A$ denota a função característica de $A$, então

$$\mathbf{E}1_A = \mathbf{P}\left(A\right)$$

A média é "definida positiva": se $\xi \geq 0$, ou se pelo menos $\mathbf{P}\left\{\xi \geq 0\right\} = 1$, então

$$\mathbf{E}\xi \geq 0$$

e a igualdade é possivel sse $\mathbf{P}\left\{\xi = 0\right\} = 1$.

A média é "linear": se $\xi$ é integrável e $a, b \in \mathbf{R}$, então

$$\mathbf{E}\left(a \cdot \xi + b\right) = a \cdot \mathbf{E}\xi + b$$

e se $\xi$ e $\eta$ são integráveis então

$$\mathbf{E}\left(\xi + \eta\right) = \mathbf{E}\xi + \mathbf{E}\eta$$

A média é "monótona": se $\xi \geq \eta$, ou se pelo menos $\mathbf{P}\left\{\xi \geq \eta\right\} = 1$, e se $\xi$ e $\eta$ são integráveis, então

$$\mathbf{E}\xi \geq \mathbf{E}\eta$$

e a igualdade é possível sse $\mathbf{P}\left\{\xi = \eta\right\} = 1$. Em particular,

$$\left|\mathbf{E}\xi\right| \leq \mathbf{E}\left|\xi\right|$$

**Exercícios.**

**a.** Prove as propriedades da média enunciadas acima.

**b.** *(fórmula da probabildade total)* Se $\xi$ é integrável e $A$ é um evento, então

$$\mathbf{E}\xi = \mathbf{E}1_A\xi + \mathbf{E}1_{A^c}\xi$$

Em geral, se $\mathcal{D} = \{D_1, D_2, ..., D_k, ...\}$ é uma partição enumerável de $\Omega$, então

$$\mathbf{E}\xi = \sum_k \mathbf{E}1_{D_k}\xi$$

**Funções de variáveis aleatórias.** Se $\xi : \Omega \to \{x_1, x_2, x_3, ...\}$ é uma variável aleatória discreta e $\varphi : \mathbf{R} \to \mathbf{R}$ é uma função arbitrária, então $\eta = \varphi \circ \xi$ é também uma variável aleatória discreta. A densidade discreta de $\eta$ é

$$\mathbf{P}\left(\eta = y_j\right) = \sum_{x_k \in \varphi^{-1}\{y_j\}} \mathbf{P}\left(\xi = x_k\right)$$

onde $\{y_1, y_2, y_3...\} = \varphi\left(\{x_1, x_2, x_3, ...\}\right)$ é o conjunto dos valores de $\eta$. Em geral é falso que se $\xi$ é integrável também $\eta$ é, assim como é falso que $\mathbf{E}\eta$ seja igual a $\varphi\left(\mathbf{E}\xi\right)$. Se $\eta$ é integrável, podemos calcular $\mathbf{E}\eta$ a partir da densidade discreta de $\xi$, pois, dado que a série é absolutamente convergente,

$$\begin{aligned}
\mathbf{E}\eta &= \sum_{y_j} y_j \cdot \mathbf{P}\left(\eta = y_j\right) \\
&= \sum_{y_j} y_j \cdot \sum_{x_k \in \varphi^{-1}\{y_j\}} \mathbf{P}\left(\xi = x_k\right) \\
&= \sum_{x_k} \varphi(x_k) \cdot \mathbf{P}\left(\xi = x_k\right)
\end{aligned}$$

**Exercícios.**

**a.** *(média aritmética)* Seja $\xi : \Omega \to \{x_1, x_2, ..., x_n\}$ uma variável aleatória simples com *lei uniforme*, i.e. com densidade discreta $\mathbf{P}\left(\xi = x_k\right) = 1/n$ para todo $k = 1, 2, ..., n$. Verifique que a média de $\xi$ é a média aritmética dos seus valores, ou seja

$$\mathbf{E}\xi = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right)$$

**b.** Escrevo $n$ cartas para $n$ pessoas distintas, meto-as em $n$ envelopes, e escrevo ao acaso as $n$ direções dos destinatários. Com que probabilidade pelo menos uma das cartas chega ao destinatário? E todas?

Defina $A_k$ como sendo o evento "a $k$-ésima carta chega ao seu destinatário" e $A = A_1 \cup A_2 \cup ... \cup A_n$, e utilize a fórmula $\mathbf{P}(A) = \mathbf{E}1_A$, assim como a expressão de $1_A$ em termos dos $1_{A_k}$, para calcular a probabilidade de $A$. O que acontece quando $n \to \infty$?

**c.** Seja $\xi$ uma variável aleatória discreta com valores em $\mathbf{N}$. Prove que

$$\mathbf{E}\xi = \sum_{n=1}^{\infty} \mathbf{P}(\xi \geq n)$$

**d.** (la biblioteca total) Um livro de $10^6$ letras no alfabeto castelhano tem probabilidade $25^{-10^6}$ no espaço de todos os livros possíveis deste comprimento (no modelo uniforme em que cada letra tem probabilidade $1/25$). O tempo que um fiel da seita blasfema de Borges tem que esperar para ver o seu livro aparecer pela primeira vez é da ordem de $25^{10^6}$ letras aleatórias... (na verdade é um pouco menor!)

**e.** (ruína do jogador) Já vimos que se $A$ joga contra um jogador com capital muito grande, a sua ruína é quase certa. Uma boa ideia para não perder muito dinheiro é usar estratégias. Por exemplo, o jogador $A$, com capital inicial de $a$ rublos, pode decidir acabar o jogo se estiver ganhando $b$ rublos. Nesse caso, a esperança da variável aleatória $\gamma =$"ganho do jogador $A$ depois do jogo ter acabado" é

$$\mathbf{E}\gamma = b \cdot (1 - p_a) - a \cdot p_a$$

A conclusão é que, mesmo usando esta estratégia, o jogo é honesto (ou seja $\mathbf{E}\gamma = 0$) sse já era honesto (ou seja se $p = 1/2$).

**Momentos e variância.** Seja $\xi$ uma variável aleatória discreta. Se $\xi^k$ é integrável, podemos definir o *momento* de grau $k$ da variável aleatória $\xi$ como sendo $\mathbf{E}\xi^k$. Se $\xi^k$ é integrável, então também $\xi^{k'}$ com $k' \leq k$ é integrável (basta utilizar a desigualdade elementar $|x|^{k'} \leq 1 + |x|^k$ válida para todo $x \in \mathbf{R}$). Mais interessantes são os momentos centrados, definidos por $\mathbf{E}(\xi - \mathbf{E}\xi)^k$, porque são invariantes por translações, e ainda mais interessantes os momentos centrados absolutos, definidos por $\mathbf{E}|\xi - \mathbf{E}\xi|^k$.

A *variância* da variável aleatória $\xi$ é o momento centrado de grau dois, ou seja

$$\begin{aligned} \mathbf{V}\xi &= \mathbf{E}(\xi - \mathbf{E}\xi)^2 \\ &= \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2 \end{aligned}$$

É evidente que $\mathbf{V}\xi \geq 0$, sendo a esperança de uma variável não-negativa. Uma notação tradicional para a variância é $\mathbf{V}\xi = \sigma^2$. A raiz positiva da variância, $\sigma = \sqrt{\mathbf{V}\xi}$, é dita *desvio padrão* de $\xi$. Outra notação será $\sigma_\xi$ se queremos lembrar que é o desvio padrão da variável $\xi$.

**O significado "matemático" da média e da variância.** A variância, ou o desvio padrão, é uma medida da "variabilidade" de $\xi$, de quanto os seus valores $x_k$ estão espalhados ao redor da "média" $\mathbf{E}\xi$.

Em particular, $\mathbf{V}\xi = 0$ sse $\mathbf{P}\{\xi = \mathbf{E}\xi\} = 1$ (ou seja "uma variável com variância nula é muito pouco aleatória!"). De facto, se $\xi = \sum_k x_k \cdot 1_{S_k}$,

$$0 = \mathbf{V}\xi = \sum_k (x_k - \mathbf{E}\xi)^2 \cdot \mathbf{P}(S_k)$$

implica que $\mathbf{P}(S_k) = 0$ para todo $x_k \neq \mathbf{E}\xi$.

Mais interessante é observar que, se $\xi$ é uma variável aleatória discreta com variância finita, então

$$\mathbf{V}\xi = \inf_{x \in \mathbf{R}} \mathbf{E}|\xi - x|^2$$

Ou seja, a variância é o valor mínimo da função $x \mapsto \mathbf{E}\,|\xi - x|^2$, e o valor de $x$ onde o mínimo é atingido é $\mathbf{E}\xi$. A interpretação "física" deste facto é que a média é o "baricentro" da lei da variável, pensada como uma distribuição de massas na recta real, e a variância é o seu "momento de inércia".

**Variáveis adimensionais.** Se $\xi$ é uma variável aleatória discreta e $a, b \in \mathbf{R}$, então

$$\mathbf{V}(a \cdot \xi + b) = a^2 \cdot \mathbf{V}\xi$$

Por vezes é interessante estudar, em vez da variável $\xi$ (que na interpretação física do modelo pode ter uma "dimensão"), a variável "adimensional" $\xi^*$ definida, desde que $\mathbf{V}\xi > 0$, por

$$\xi^* = \frac{\xi - \mathbf{E}\xi}{\sqrt{\mathbf{V}\xi}}$$

A variável $\xi^*$ tem média 0 e variância 1. A ideia é que $\xi^*$ tem todas as propridades "qualitativas" de $\xi$, as propriedades que não dependem nem da escolha da origem nem da escolha da unidade de medida.

**Exercícios.**

**a.** Seja $\xi$ é uma variável aleatória discreta com variância finita. Verifique que $\mathbf{V}\xi = \inf_{x \in \mathbf{R}} \mathbf{E}\,|\xi - x|^2$.

**b.** Seja $\xi$ é uma variável aleatória com valores $1, 2, ..., n$ e lei uniforme. Calcule $\mathbf{V}\xi$.

**Produtos de variáveis aleatórias.** Em geral, mesmo se $\xi$ e $\eta$ são integráveis, a variável $\xi\eta$ pode não ser integrável. O produto $\xi\eta$ é integrável se $\xi$ e $\eta$ têm variância finita, e neste caso vale a *desigualdade de Cauchy-Schwarz*

$$\mathbf{E}\,|\xi\eta| \leq \sqrt{\mathbf{E}\xi^2} \cdot \sqrt{\mathbf{E}\eta^2}$$

De facto, se $\mathbf{E}\xi^2 = 0$, então $\xi = 0$ com probabilidade um, e portanto também $\mathbf{E}\,|\xi\eta| = 0$. Se, por outro lado, $\mathbf{E}\xi^2$ e $\mathbf{E}\eta^2$ são positivas, e definimos $\xi' = \xi/\sqrt{\mathbf{E}\xi^2}$ e $\eta' = \eta/\sqrt{\mathbf{E}\eta^2}$, a desigualdade elementar $2\,|\xi'\eta'| \leq \xi'^2 + \eta'^2$ e a monotonia da média implicam que $2\mathbf{E}\,|\xi'\eta'| \leq \mathbf{E}\xi'^2 + \mathbf{E}\eta'^2 = 2$, que é equivalente à desigualdade acima.

Observe que a igualdade $\mathbf{E}\,|\xi\eta| = \sqrt{\mathbf{E}\xi^2} \cdot \sqrt{\mathbf{E}\eta^2}$ é possível sse $\mathbf{P}\,(\xi' \pm \eta' = 0) = 1$, ou seja quando as variáveis $\xi$ e $\eta$ são proporcionais, no sentido em que existe um real $\lambda$ tal que $\xi = \lambda\eta$ com probabilidade um.

Se $\xi$ e $\eta$ são integráveis e independentes, então $\xi\eta$ é integrável e

$$\mathbf{E}\xi\eta = \mathbf{E}\xi \cdot \mathbf{E}\eta$$

De facto, sejam $\xi = \sum_k x_k \cdot 1_{S_k}$ e $\eta = \sum_k y_k \cdot 1_{T_k}$. Então $\xi\eta = \sum_{k,j} x_k y_j \cdot 1_{S_k \cap T_j}$, e portanto

$$\begin{aligned}
\mathbf{E}\xi\eta &= \sum_{k,j} x_k y_j \cdot \mathbf{P}\,(S_k \cap T_j) \\
&= \sum_{k,j} x_k y_j \cdot \mathbf{P}\,(S_k) \cdot \mathbf{P}\,(T_j) \\
&= \left(\sum_k x_k \cdot \mathbf{P}\,(S_k)\right) \cdot \left(\sum_j y_j \cdot \mathbf{P}\,(T_j)\right) \\
&= \mathbf{E}\xi \cdot \mathbf{E}\eta
\end{aligned}$$

(este é um resultado standard sobre as séries absolutamente convergentes: se $\sum_i a_i$ e $\sum_j b_j$ são absolutamente convergentes, então a série $\sum_{i,j} a_i b_j$ é absolutamente convergente e tem soma igual ao produto das somas das duas séries). Por indução, segue que se $\xi_1$, $\xi_2$, ...,$\xi_n$ são integráveis e independentes então

$$\mathbf{E}\xi_1\xi_2...\xi_n = \mathbf{E}\xi_1 \cdot \mathbf{E}\xi_2 \cdot ... \cdot \mathbf{E}\xi_n$$

(basta observar que a independência da família $\{\xi_1, \xi_2, ..., \xi_n\}$ implica a independência das variáveis $\xi_1 \cdot \xi_2 \cdot ... \cdot \xi_{n-1}$ e $\xi_n$).

**Covariância.** Sejam $\xi$ e $\eta$ duas variáveis aleatórias com variância finita. A variância da soma é

$$\mathbf{V}(\xi+\eta) = \mathbf{V}\xi + \mathbf{V}\eta + 2\cdot \mathrm{Cov}(\xi,\eta)$$

onde a *covariância* de $\xi$ e $\eta$ (ou "entre" $\xi$ e $\eta$) é definida por

$$\begin{aligned}\mathrm{Cov}(\xi,\eta) &= \mathbf{E}\left((\xi-\mathbf{E}\xi)\cdot(\eta-\mathbf{E}\eta)\right)\\ &= \mathbf{E}\xi\eta - \mathbf{E}\xi\cdot\mathbf{E}\eta\end{aligned}$$

Se $\xi$ e $\eta$ são independentes, então $\mathbf{E}\xi\eta = \mathbf{E}\xi\cdot\mathbf{E}\eta$, e portanto $\mathrm{Cov}(\xi,\eta)=0$ e

$$\mathbf{V}(\xi+\eta) = \mathbf{V}\xi + \mathbf{V}\eta$$

As vaiáveis $\xi$ e $\eta$ são ditas *não correlacionadas* se $\mathrm{Cov}(\xi,\eta)=0$. Infelizmente, $\mathrm{Cov}(\xi,\eta)=0$ não implica que $\xi$ e $\eta$ sejam independentes!

**Exercício.** Se as variáveis aleatórias $\xi_1$, $\xi_2$, ...,$\xi_n$ são independentes então

$$\mathbf{V}(\xi_1+\xi_2+...+\xi_n) = \mathbf{V}\xi_1 + \mathbf{V}\xi_2 + ... + \mathbf{V}\xi_n$$

(observe que a independência da família $\{\xi_1,\xi_2,...,\xi_n\}$ implica a independência das variáveis $\xi_1+\xi_2+...+\xi_{n-1}$ e $\xi_n$).

**Correlação.** Sejam $\xi$ e $\eta$ duas variáveis aleatórias com variâncias positivas $\mathbf{V}\xi=\sigma_\xi^2$ e $\mathbf{V}\eta=\sigma_\eta^2$. O *coeficiente de correlação* é definido por

$$\varrho(\xi,\eta) = \frac{\mathrm{Cov}(\xi,\eta)}{\sigma_\xi\cdot\sigma_\eta}$$

É imediato verificar que o coeficiente de correlação é "adimensional" e "invariante por translações", ou seja

$$\varrho(a\xi+b, c\eta+d) = \varrho(\xi,\eta)$$

para todos $a,b,c,d\in\mathbf{R}$ com $a$ e $c\neq 0$. Da identidade

$$0\le \mathbf{V}(\xi/\sigma_\xi \pm \eta/\sigma_\eta) = 2(1\pm\varrho(\xi,\eta))$$

segue que $-1\le\varrho(\xi,\eta)\le 1$.

O interesse do coeficiente de correlação está na seguinte observação: $\varrho(\xi,\eta)=\pm1$ sse $\xi$ e $\eta$ são linearmente dependentes com probabilidade um, ou seja se existem $a,b\in\mathbf{R}$, $a\neq0$, tais que $\mathbf{P}(\eta=a\xi+b)=1$. Pois, $\varrho(\xi,\eta)=\pm1$ implica que

$$\mathbf{V}(\xi/\sigma_\xi \mp \eta/\sigma_\eta) = 0$$

mas uma variável aleatória tem variância zero sse é constante com probabilidade um.

A informação contida no coeficiente de correlação é a seguinte: se as variáveis são independentes, então $\varrho(\xi,\eta)=0$; se $|\varrho(\xi,\eta)|=1$, então as variáveis são linearmente dependentes. O que o coeficiente de correlação "detecta" é, portanto, a "correlação linear" entre duas variáveis.

**Exercícios.**

**a.** Sejam $\xi_n$ o número de caras e $\eta_n$ o número de coroas obtidas lançando $n$ vezes uma moeda. Assuma que, em cada lançamento, a probabilidade de sair cara é igual a $p$. Determine o coeficiente de correlação $\varrho(\xi_n,\eta_n)$. (Observe que $\mathbf{V}(\xi_n+\eta_n)=0$, pois $\xi_n+\eta_n=n$ com probabilidade um)

**b.** *(moedas correlacionadas)* As duas moedas de um mago funcionam assim: a primeira moeda é honesta, e mostra cara com probabilidade 1/2; a segunda moeda mostra a mesma face da primeira com probabilidade $p$. Sejam $\xi$ e $\eta$ as variáveis aleatória definidas por: $\xi=1$ se a primeira moeda mostra cara e 0 se a primeira moeda mostra coroa, $\eta=1$ se a segunda moeda mostra cara e 0 se a segunda moeda mostra coroa. Determine a densidade conjunta, as densidades marginais e a covariância de $\xi$ e $\eta$. Existe um valor de $p$ tal que $\xi$ e $\eta$ são independentes?

**Estimadores e método dos mínimos quadrados.** Sejam $\xi$ e $\eta$ duas variáveis aleatórias. Numa experiência física observamos a variável $\xi$. Se $\xi$ e $\eta$ não são independentes, podemos esperar "estimar" $\eta$ a partir de informações sobre $\xi$. Uma função $\varphi : \xi \mapsto \varphi(\xi)$ é dita *estimador de* $\eta$. Uma medida da bontade do estimador $\varphi$ é o "erro quadrático médio"

$$\mathbf{E}\left(\eta - \varphi(\xi)\right)^2$$

Procurar o estimador que torna mínimo o erro pode ser difícil. Mais fácil é minimizar o erro na classe dos estimadores lineares, i.e. das funções $\lambda : \xi \mapsto a + b\xi$. O *melhor estimador linear no sentido dos mínimos quadrados* é

$$\lambda^*(\xi) = a^* + b^*\xi$$

onde os parâmetros são (obtidos calculando uma derivada)

$$a^* = \mathbf{E}\eta \qquad b^* = \frac{\mathrm{Cov}\,(\xi, \eta)}{\mathbf{V}\xi}$$

O erro quadrático médio cometido ao estimar a variável $\eta$ com $\lambda^*(\xi)$ resulta ser

$$\mathrm{erro}^2 = \mathbf{E}\left(\eta - \lambda^*(\xi)\right)^2 = \mathbf{V}\xi \cdot \left(1 - \rho^2(\xi, \eta)\right)$$

# 8 Modelos discretos

**Lei de Bernoulli.** A *lei de Bernoulli* é a lei de uma variável $\xi$ que assume os valores 0 (insucesso) ou 1 (sucesso) com probabilidades $\mathbf{P}(\xi = 0) = 1 - p$ e $\mathbf{P}(\xi = 1) = p$. É tradição denotar $q = 1 - p$ a probabilidade de "insucesso".

A média e a variância de $\xi$ são

$$\begin{aligned} \mathbf{E}\xi &= 0 \cdot q + 1 \cdot p = p \\ \mathbf{V}\xi &= \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2 = \left(0^2 \cdot q + 1^2 \cdot p\right) - p^2 = pq \end{aligned}$$

**Provas de Bernoulli: lei binomial.** Na esperiência das $n$ provas de Bernoulli com probabilidade de sucesso $p$ em cada prova, o espaço de probabilidades é $\Omega^n = \{\omega = (\omega_1, \omega_2, ..., \omega_n) \text{ com } \omega_k = 0, 1\}$, $\mathcal{E} = \mathcal{P}(\Omega^n)$ e a probabilidade é definida por

$$\mathbf{P}(\omega) = p^{\sum_k \omega_k} q^{n - \sum_k \omega_k}$$

As variáveis $\xi_k : \Omega^n \to \{0, 1\}$, definidas por $\xi_k(\omega) = \omega_k$, têm lei de Bernoulli (o evento $\{\xi_k = 1\}$ é o evento "sucesso na $k$-ésima prova"), e são independentes (por construção!, mas é um bom exercício provar a independência).

A variável

$$S_n = \xi_1 + \xi_2 + ... + \xi_n$$

definida por $S_n(\omega_1, \omega_2, ..., \omega_n) = \omega_1 + \omega_2 + ... + \omega_n$, representa o "número de sucessos em $n$ provas". Tem valores $k = 0, 1, 2, ..., n$ com probabilidades

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k q^{n-k}$$

A lei de $S_n$ é dita *lei binomial*. Uma notação para uma variável com lei binomial é "$S_n \sim B(n, p)$", que os estatísticos lêem "a variável $S_n$ tem lei binomial com parâmetros $n$ e $p$". A lei de Bernoulli é uma lei $B(1, p)$.

A média e a variância de $S_n$ são

$$\begin{aligned} \mathbf{E}S_n &= \mathbf{E}(\xi_1 + \xi_2 + ... + \xi_n) = np \\ \mathbf{V}S_n &= \mathbf{V}(\xi_1 + \xi_2 + ... + \xi_n) = npq \end{aligned}$$

(onde utilizamos a independência das $\xi_k$).

As variáveis acima podem ser pensadas como definidas no espaço das infinitas provas de Bernoulli,

$$\Omega^\infty = \{\omega = (\omega_1, \omega_2, ..., \omega_k, ...) \text{ com } \omega_k = 0, 1\}$$

munido da probabilidade produto $\mathbf{P} : \mathcal{B}(\Omega^\infty) \to [0, 1]$. Neste caso, a variável $S_n$ tem o significado de "o número de sucessos nas primeiras $n$ provas". É interessante ver a família $(S_n)_{n \in \mathbf{N}}$ como um processo estocástico, pensando em $n$ como um tempo. A variável $S_n/n$ representa a "frequência de sucessos nas primeiras $n$ provas".

**Exercícios.**

**a.** Seja $S_n$ o número de sucessos obtidos em $n$ provas de Bernoulli com probabilidade de sucesso $p$. Determine a probabilidade dos eventos $\{S_n = np\}$, $\{S_n = 0\}$ e $\{|S_n| = k\}$. Determine a probabilidade de $S_n$ ser par.

**b.** Seja $S_n$ uma variável aleatória com lei binomial $B(n, p)$. Determine o máximo da densidade discreta $k \mapsto \mathbf{P}(S_n = k)$.

**c.** *(somas de binomiais independentes)* Se $\xi_1, \xi_2, ..., \xi_n$ são independentes com lei binomial $\xi_i \sim B(k_i, p)$, então a soma $S_n = \xi_1 + \xi_2 + ... + \xi_n$ tem lei binomial $B(k, p)$ com $k = k_1 + k_2 + ... + k_n$.

**Marcha aleatória.** A posição da marcha aleatória ao tempo $n$ é a variável aleatória

$$T_n = \xi_1 + \xi_2 + ... + \xi_n$$

onde as variáveis $\xi_1, \xi_2, ...$ são independentes e identicamente distribuídas, com valores $\pm 1$ e lei definida por $\mathbf{P}(\xi_i = 1) = p$ e $\mathbf{P}(\xi_i = -1) = q$. Em particular, a posição da marcha ao tempo $n$ é igual à posição ao tempo $n-1$ mais $\pm 1$, o incremento devido ao $n$-ésimo "passo", i.e.

$$T_n = T_{n-1} + \xi_n$$

Os valores de $T_n$ são $-n$, $-n+2$, $-n+4$, ..., $n-2$, $n$ (e portanto são pares ou ímpares dependendo da paridade de $n$). A lei de $T_n$ pode ser calculada observando que $T_n$ é igual à diferença entre o número de sucessos e o número de insucessos em $n$ provas de Bernoulli. Isto implca que $S_n = (T_n + n)/2$ tem lei binomial $B(n,p)$, logo a densidade discreta de $T_n$ é

$$\mathbf{P}(T_n = 2k - n) = \binom{n}{k} p^k q^{n-k}$$

onde $k = 0, 1, 2, ..., n$.

A média e a variância de $S_n$ são

$$\begin{aligned} \mathbf{E}T_n &= n(2p-1) \\ \mathbf{V}T_n &= 4npq \end{aligned}$$

**Exercício.** Seja $T_n$ a posição no tempo $n$ de uma marcha aleatória simétrica, i.e. $T_n = \xi_1 + \xi_2 + ... + \xi_n$ onde as variáveis $\xi_k$ são independentes e têm valores $\pm 1$ com probabilidade uniforme.

Determine a lei, a média e a variância de $T_n$.

Determine a probabilidade dos eventos $\{T_n = 0\}$ e $\{|T_n| = n\}$

Determine as probabilidades condicionadas

$$\mathbf{P}(T_{n+1} = k+1 \,|T_n = k) \qquad \mathbf{P}(T_{n+1} = k+1 \,|T_n = k,\, T_{n-1} = k-1)$$

**Tempo de espera: lei geométrica.** Seja $\tau$ o número de provas de Bernoulli necessárias para obter "sucesso" pela primeira vez. É possível definir a variável aleatória $\tau$ no modelo das infinitas provas de Bernoulli, i.e. como a função no espaço $\Omega^\infty = \{\omega = (\omega_1, \omega_2, ..., \omega_k, ...) \text{ com } \omega_k = 0, 1\}$ das palavras infinitas nas letras 0 e 1 definida por

$$\tau(\omega) = \min\{k \text{ tal que } \omega_k = 1\}$$

se o mímimo acima for finito, e $\tau(\omega) = \infty$ no ponto $\omega = (0, 0, 0, ...)$. Os valores possíveis são $\mathbf{N} \cup \{\infty\}$. O evento $\{\tau = k\}$ com $k \in \mathbf{N}$ é o cilindro

$$\{\omega \in \Omega^\infty \text{ t.q. } \omega_1 = \omega_2 = ... = \omega_{k-1} = 0 \text{ e } \omega_k = 1\}$$

e portanto a sua probabilidade é

$$\mathbf{P}(\tau = k) = (1-p)^{k-1} p$$

Esta é a densidade discreta da variável aleatória "tempo de espera" em infinitas provas de Bernoulli. A lei de $\tau$ é dita *lei geométrica*. Uma notação pode ser $\tau \sim$geométrica$(p)$.

Observe que a probabilidade do evento $\{\tau = \infty\}$ é igual a

$$\mathbf{P}(\tau = \infty) = 1 - \mathbf{P}(\tau < \infty) = 1 - \sum_{k=1}^{\infty} (1-p)^{k-1} p = 0$$

desde que $p$ seja diferente de 0, caso pouco interessante em que $\mathbf{P}(\tau = \infty) = 1$.

Em algum livro é chamada "geométrica" também a lei da variável aleatória $\xi = \tau - 1$, com valores $0, 1, 2, ...$ e densidade discreta $\mathbf{P}(\xi = k) = (1-p)^k p$.

A lei geométrica é caracterizada pela propriedade de "falta de memória". Por um lado, observando que

$$\mathbf{P}(\tau > k) = \sum_{n=k+1}^{\infty} (1-p)^{n-1} p = (1-p)^k$$

temos que para todos $k, n \in \mathbf{N}$

$$\begin{aligned} \mathbf{P}(\tau = k+n \,|\, \tau > k) &= \frac{\mathbf{P}(\{\tau = k+n\} \cap \{\tau > k\})}{\mathbf{P}(\tau > k)} \\ &= \frac{\mathbf{P}(\tau = k+n)}{\mathbf{P}(\tau > k)} \\ &= (1-p)^{n-1} p \end{aligned}$$

e portanto

$$\mathbf{P}(\tau = k+n \,|\, \tau > k) = \mathbf{P}(\tau = n)$$

Por outro lado, uma variável aleatória $\tau$ com valores em $\mathbf{N}$ que satisfaz a condição acima para todos $n$ e $k$ tem lei geométrica. De facto, chamando $p$ a probabilidade do evento $\{\tau = 1\}$, esta propriedade fornece a equação recursiva $\mathbf{P}(\tau = n+1) = (1-p) \cdot \mathbf{P}(\tau = n)$, que determina as probabilidades dos eventos $\{\tau = n\}$ para todo $n$.

Isso diz que esperar mais um tempo $n$ depois de ter esperado um tempo $k$ é a mesma coisa que esperar um tempo $n$ à partida, i.e. "o conhecimento do passado não influi nas previsões sobre o futuro uma vez que o presente é conhecido" (ao contrário do que acham muitos jogadores do jogo do bicho!). Se pensamos em $\tau$ como um "tempo de vida" de um sistema físico, então esta propriedade quer dizer algo como "o sistema não tem idade: se o sistema está vivo hoje, o seu futuro é igual ao futuro de um sistema recém nascido".

A média de $\tau$ é

$$\begin{aligned} \mathbf{E}\tau &= \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\, p = -p \cdot \frac{d}{dp} \left( \sum_{k=0}^{\infty} (1-p)^k \right) \\ &= -p \cdot \frac{d}{dp}\,(1/p) = 1/p \end{aligned}$$

desde que $p \neq 0$.

**Exercícios.**

**a.** Considere um modelo de "lançamentos sucessivos e independentes" de uma moeda tal que a probabilidade de obter cara em cada lançamento seja $p$.

Determine a probabilidade de obter cara nos primeiros $k-1$ lançamentos.

Determine a probabilidade de obter coroa pela primeira vez no $k$-ésimo lançamento.

Determine a probabilidade de obter coroa pela segunda vez no $k$-ésimo lançamento.

Determine o valor esperado do número de lançamentos necessários até obter coroa pela primeira vez.

**b.** No jogo da roleta russa, põe-se uma bala no carregador de uma pistola que contém seis entradas. O jogador faz rolar o tambor da pistola e dispara na sua têmpora.

Calcule a probabilidade do jogador morrer se repetir o jogo uma, duas ou tres vezes.

Calcule a probabilidade do jogador morrer à $n$-ésima vez que repete o jogo sabendo que ainda está vivo depois da $(n-1)$-ésima vez.

Quantas vezes é que o jogador tem que repetir o jogo para ter probabilidade de morrer maior de 0.999 ?

**c.** Sejam $\tau_1$ e $\tau_2$ duas variáveis aleatórias independentes com leis geométrica($p_1$) e geométrica($p_2$) respectivamente. Mostre que a variável $\min\{\tau_1, \tau_2\}$ tem lei geométrica e determine o seu parâmetro e a sua média. Determine a lei da variável $\max\{\tau_1, \tau_2\}$.

**d.** A NASA estima que uma em cada 68 missões do vaivém pode falhar (dando lugar a um desastre) devido a problemas técnicos. Calcule a probabilidade de assistir a um desastre dentro das primeiras dez missões do vaivém.

**Aproximação e lei de Poisson.** Calcular densidades de uma variável aleatória binomial $S_n \sim B(n,p)$ com $n$ muito grande pode ser pouco prático, mesmo com a ajuda de uma máquina. Uma boa aproximação, se $n \gg 1$ e $pn = \lambda \ll n$, é considerar uma variável $S_n$ com lei $B(n, \lambda/n)$ e calcular o limite da sua densidade quando $n \to \infty$.

$$\begin{aligned} \mathbf{P}(S_n = k) &= \binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{n-k} \\ &= \frac{\lambda^k}{k!} (1-\lambda/n)^n \frac{n(n-1)...(n-k+1)}{n^k} (1-\lambda/n)^{-k} \\ &\to \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

Este resultado é conhecido como *teorema de Poisson*. A observação de que $\sum_{k\geq 0} \frac{\lambda^k}{k!} e^{-\lambda} = 1$ justifica a seguinte definição.

A variável aleatória $\xi$ com valores $0, 1, 2, ...$ tem *lei de Poisson* com parâmetro $\lambda$ se a sua densidade discreta é

$$\mathbf{P}(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda} .$$

Uma notação é $\xi \sim$Poisson($\lambda$). A lei de Poisson é uma boa aproximação da lei binomial se a probabilidade $p$ é pequena e $n$ é grande (ou seja se $n \gg np$). Tem também uma interpretação física interessante, e é um modelo natural de muitos fenómenos.

A média e a variância de $\xi$ são

$$\begin{aligned} \mathbf{E}\xi &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \\ \mathbf{V}\xi &= \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2 = \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 \\ &= \lambda \cdot \sum_{k=0}^{\infty} (k+1) \cdot \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 = \lambda(\lambda+1) - \lambda^2 = \lambda \end{aligned}$$

**Exercícios.**

**a.** Seja $\xi$ uma variável aleatória com lei de Poisson Poisson($\lambda$).
Determine o máximo da densidade discreta $k \mapsto \mathbf{P}(\xi = k)$.
Fixado $k$, que valor de $\lambda$ maximiza $\mathbf{P}(\xi = k)$ ?

**b.** *(limite termodinâmico)* Uma caixa de volume $v$ contém $n$ moléculas de gas. A probabilidade de cada molécula estar numa certa região da caixa, cujo volume é $q$, é igual à $q/v$. Então a probabilidade de observar $k$ moléculas na região é dada por

$$\binom{n}{k} (q/v)^k (1-q/v)^{n-k}$$

O "limite termodinâmico" consiste em fazer $n \to \infty$ e $v \to \infty$ mantendo constante a densidade média $\rho = n/v$. Utilizando o teorema de Poisson, mostre que no limite termodinâmico a probabilidade de observar $k$ moléculas na região, suposta fixada, é

$$\frac{(\rho q)^k}{k!} e^{-\rho q}$$

**c.** Cada núcleo, dentro de uma amostra de $10^{20}$ núcleos, tem probabilidade $10^{-18}$ de decair no espaço de uma hora. Estime a probabilidade de decairem 10 núcleos no espaço de uma hora, e a probabilidade de decairem pelo menos 99% dos núcleos no espaço de uma hora.

**d.** Duas telefonistas recebem, cada hora, respectivamente $\xi$ e $\eta$ chamadas com lei de Poisson e parâmetros $\lambda$ e $m$. Determine a probabilidade de: as duas telefonistas receberem, no total, menos do que três chamadas numa hora; nenhuma das duas telefonistas receber mais do que uma chamada numa hora; cada uma das telefonistas receber pelo menos uma chamada numa hora.

**Distribuição de Gibbs.** A energia de um sistema termodinâmico em equilibrio térmico é uma variável aleatória $\xi$ com valores $\{e_1, e_2, e_3, ...\}$ e lei determinada por

$$\mathbf{P}(\xi = e_n) = \frac{e^{-\beta e_n}}{Z(\beta)}$$

onde $\beta > 0$ e a série

$$Z(\beta) = \sum_n e^{-\beta e_n}$$

é suposta convergente. A função $\beta \mapsto Z(\beta)$ é dita "função de partição" do sistema, e o parâmetro $\beta$, na interpretação física, é igual a $1/k_B T$, onde $T$ é a temperatura e $k_B$ a constante de Boltzmann. A função de partição "contém" a termodinâmica do sistema, pois os potenciais termodinâmicos são obtidos à partir da "energia livre"

$$F = -\beta^{-1} \log Z(\beta)$$

As ideias físicas que justificam este modelo podem ser encontradas no manual de Landau e Lifshitz. Tratados críticos e problemáticos sobre os fundamentos (matemáticos e físicos) da mecânica estatísticas estão no clássico de Khinchin [Kh57] ou no recente tratado de Gallavotti [Ga99].

**Exercício.** Mostre que, no modelo de Gibbs, a "energia média", definida por $E = \mathbf{E}\xi$, é

$$E = -\frac{\partial}{\partial \beta} \log Z(\beta)$$

A "entropia" do sistema é definida como

$$S = -\mathbf{E} \log \pi$$

onde $\pi$ é a variável aleatória $\pi = \sum_n \mathbf{P}(\xi = e_n) \cdot 1_{\{\xi = e_n\}}$. Verifique que a "energia livre", definida como sendo $F = E - \beta^{-1} S$, é igual a $-\beta^{-1} \log Z(\beta)$.

# 9 Função geradora de probabilidades

**Função geradora de probabilidades.** Seja $\xi$ uma variável aleatória discreta com valores inteiros não negativos. A *função geradora de probabilidades* de $\xi$ é a função $z \mapsto \psi_\xi(z)$ definida por

$$\begin{aligned}\psi_\xi(z) &= \mathbf{E}z^\xi \\ &= \sum_{n=0}^{\infty} z^n \cdot \mathbf{P}(\xi = n)\end{aligned}$$

na região da recta real, ou do plano complexo, onde a série é absolutamente convergente. Os valores $\mathbf{P}(\xi = n)$ da densidade discreta são os coeficientes de Taylor da série que define $\psi_\xi$, e portanto $\psi_\xi$ é uma função analítica no intervalo $]-1, 1[$, porque a série converge absolutamente se $|z| \leq 1$, sendo $\sum \mathbf{P}(\xi = n) = 1$. A função geradora determina (e é determinada por) a densidade discreta, porque os coeficientes de Taylor de uma função analítica são únicos:

$$\mathbf{P}(\xi = n) = \frac{1}{n!} \cdot \frac{d^n \psi_\xi}{dz^n}(z)\,|_{z=0}$$

Se o raio de convergência da função geradora $\psi_\xi(z)$ é $> 1$, as derivadas no ponto 1 fornecem os momentos da variável $\xi$. Por exemplo,

$$\psi'_\xi(1) = \sum_{n=0}^{\infty} n \cdot \mathbf{P}(\xi = n) = \mathbf{E}\xi$$

$$\psi''_\xi(1) = \sum_{n=0}^{\infty} (n^2 - n) \cdot \mathbf{P}(\xi = n) = \mathbf{E}\xi^2 - \mathbf{E}\xi$$

e portanto

$$\mathbf{V}\xi = \psi''_\xi(1) - \psi'_\xi(1) - \big(\psi'_\xi(1)\big)^2$$

**Exercícios.**

**a.** *(somas de variáveis independentes)* Se $\xi$ e $\eta$ são variáveis aleatórias independentes com valores inteiros não negativos então

$$\psi_{\xi+\eta}(z) = \psi_\xi(z) \cdot \psi_\eta(z)$$

**b.** (binomial) Se $\xi$ tem lei binomial $B(n, p)$ então $\psi_\xi(z) = (1 - p + zp)^n$.

**c.** (Poisson) Se $\xi$ tem lei Poisson($\lambda$) então $\psi_\xi(z) = e^{\lambda(z-1)}$.

**d.** (geométrica) Se $\xi$ tem lei geométrica($p$) então $\psi_\xi(z) = \frac{p}{1-z(1-p)}$.

**e.** (geométrica) Se $\xi \sim$ geométrica($p$) então $\mathbf{V}\xi = \frac{1-p}{p^2}$.

**f.** A soma de duas variáveis de Poisson independentes com parâmetros $\lambda$ e $m$ é uma variável de Poisson com parâmetro $\lambda + m$.

**Somas aleatórias.** Seja $(\xi_n)$ uma sucessão de variáveis aleatórias com valores inteiros não negativos, e $\tau$ um "tempo aleatório", i.e. uma variável aleatória com valores $0, 1, 2, ....$ Se $S_n = \xi_1 + \xi_2 + ... + \xi_n$, então a variável $S_\tau$, pensada como "a soma do processo até o tempo $\tau$", é definida como

$$S_\tau(\omega) = \xi_1(\omega) + \xi_2(\omega) + ... + \xi_{\tau(\omega)}(\omega)$$

se $\tau(\omega) > 0$ e $S_\tau(\omega) = 0$ se $\tau(\omega) = 0$. Usando a fórmula da probabilidade total temos

$$\mathbf{P}\left(S_\tau = k\right) = \sum_{i=0}^{\infty} \mathbf{P}\left(S_i = k \,|\tau = i\,\right) \cdot \mathbf{P}\left(\tau = i\right)$$

Se as variáveis $\xi_1, \xi_2, ...$ e $\tau$ são independentes, então a densidade de $S_\tau$ é

$$\mathbf{P}\left(S_\tau = k\right) = \sum_{i=0}^{\infty} \mathbf{P}\left(S_i = k\right) \cdot \mathbf{P}\left(\tau = i\right)$$

Em particular, se as variáveis $\xi_i$ são identicamente distribuídas

$$\begin{aligned} \mathbf{E}S_\tau &= \sum_{i=0}^{\infty} \mathbf{P}\left(\tau = i\right) \sum_k k \cdot \mathbf{P}\left(S_i = k\right) \\ &= \sum_{i=0}^{\infty} \mathbf{P}\left(\tau = i\right) \cdot \mathbf{E}S_i = \sum_{i=0}^{\infty} i \cdot \mathbf{P}\left(\tau = i\right) \cdot \mathbf{E}\xi_1 \\ &= \mathbf{E}\tau \cdot \mathbf{E}\xi_1 \end{aligned}$$

Nessas condições, a função geradora da soma aleatória é muito simples. De facto, sejam $\psi_{\xi_1}$ a função geradora de $\xi_1$ e $\psi_\tau$ a função geradora de $\tau$. Então $\psi_{S_i}(z) = \psi_{\xi_1}(z)^i$, e portanto

$$\begin{aligned} \psi_{S_\tau}(z) &= \sum_{k=0}^{\infty} z^k \sum_{i=0}^{\infty} \mathbf{P}\left(S_i = k\right) \cdot \mathbf{P}\left(\tau = i\right) \\ &= \sum_{i=0}^{\infty} \psi_{\xi_1}(z)^i \cdot \mathbf{P}\left(\tau = i\right) \\ &= \psi_\tau(\psi_{\xi_1}(z)) \end{aligned}$$

ou seja, a função geradora de $S_\tau$ é $\psi_{S_\tau} = \psi_\tau \circ \psi_{\xi_1}$.

**Exercício.** Repito umas provas de Bernoulli um número $\tau$ de vezes, onde $\tau$ tem lei de Poisson. As variáveis $S_\tau$ ="número de sucessos em $\tau$ provas" e $T_\tau$ ="número de insucessos em $\tau$ provas" são independentes!

# 10 Integration

The useful definition of mean value for arbitrary random variables goes under the technical name of "Lebesgue integration theory". A standard reference is the books by Rudin [Rud66], or almost any manual of probability.

**Mean of nonnegative random variables.** Let $(\Omega, \mathcal{E}, \mathbf{P})$ be a probability space, and let $\varphi : \Omega \to \mathbf{R}$ be a *simple* random variables, i.e. $\varphi = \sum_{k=1}^{n} x_k \cdot 1_{S_k}$ with $S_k \in \mathcal{E}$ e $x_k \in \mathbf{R}$. The mean of $\varphi$ is defined as

$$\mathbf{E}\varphi = \sum_{k=1}^{n} x_k \cdot \mathbf{P}(S_k)$$

The function $\varphi \mapsto \mathbf{E}\varphi$ is homogeneous and additive on the space of nonnegative simple random variables. Observe that if $\varphi$ is simple and $S$ is an event, then $1_S\varphi$ is again a simple random variable. If, moreover, $\varphi$ is nonnegative, then the set function $S \mapsto \mathbf{E}1_S\varphi$ is a measure over $\mathcal{E}$.

Let $\xi : \Omega \to [0, \infty]$ be a nonnegative random variable. The mean of $\xi$ is defined as the extended real number

$$\mathbf{E}\xi = \sup_{\varphi \text{ simple s.t. } 0 \le \varphi \le \xi} \mathbf{E}\varphi$$

(note that this number may be infinite!). It is immediate to check the following properties. If $0 \le \eta \le \xi$ on $S$, then $\mathbf{E}1_S\xi \le \mathbf{E}1_S\eta$. If $S \subset T$ then $\mathbf{E}1_S\xi \le \mathbf{E}1_T\xi$. The mean is additive and homogeneous, i.e. $\mathbf{E}(\xi + \eta) = \mathbf{E}\xi + \mathbf{E}\eta$ and $\mathbf{E}\lambda\xi = \lambda \cdot \mathbf{E}\xi$ for any $\lambda \ge 0$. Moreover, $\mathbf{E}1_S\xi = 0$ iff $\xi = 0$ on $S$.

**Lebesgue monotone convergence theorem (a.k.a. Beppo Levi theorem).** *Let $(\xi_n)$ be an increasing sequence of nonnegative random variables such that $\xi_n \uparrow \xi$. Then $\mathbf{E}\xi_n \uparrow \mathbf{E}\xi$.*

**proof.** The sequence $(\mathbf{E}\xi_n)$ is increasing. Let $a$ be its limit. Clearly $a \le \mathbf{E}\xi$, because $\xi_n \le \xi$ for any $n$, so that we are left with showing the reverse inequality. Let $\varphi$ be a simple random variable such that $0 \le \varphi \le \xi$, and let $0 < \varepsilon < 1$. Since $\xi_n \uparrow \xi$ pointwise, the space $\Omega$ is the union of the increasing sequence of events $\Omega_n = \{\xi_n \ge (1-\varepsilon)\varphi\}$. There follow that

$$\mathbf{E}\xi_n \ge \mathbf{E}1_{\Omega_n}\xi_n \ge (1-\varepsilon) \cdot \mathbf{E}1_{\Omega_n}\varphi$$

Taking the limit as $n \to \infty$ we get

$$a \ge (1-\varepsilon) \cdot \mathbf{E}\varphi$$

From the arbitrariness of $\varphi$ (smaller than $\xi$) and $\varepsilon$, this gives $a \ge \mathbf{E}\xi$. □

**Approximation by simple radom variables.** Any nonnegative random variable $\xi$ is the pointwise limit of an increasing sequence of nonnegative simple random variables. Indeed, the sequence of simple random variables defined by

$$\xi_n = \sum_{k=0}^{n2^n} \frac{k}{2^n} \cdot 1_{\{k/2^n \le \xi < (k+1)/2^n\}}$$

is increasing and has limit $\xi_n \uparrow \xi$. This observation, together with the monotone convergence theorem, shows that the mean of a nonnegative random variable is, and indeed could have been defined as, a limit of means of a sequence of simple random variable, since $\mathbf{E}\xi_n \uparrow \mathbf{E}\xi$.

**Exchanging mean and sum.** The monotone convergence theorem also implies the following useful remark: if $(\xi_n)$ is a sequence of nonnegative random variables, then $\mathbf{E}(\sum_n \xi_n) = \sum_n \mathbf{E}\xi_n$.

**Fatou lemma.** *If $(\xi_n)$ is a sequence on nonnegative random variables, then*

$$\mathbf{E}\underline{\lim}\xi_n \le \underline{\lim}\mathbf{E}\xi_n$$

**proof.** If $\psi_n = \inf_{k\leq n} \xi_k$ then $\mathbf{E}\psi_n \leq \mathbf{E}\xi_n$. The sequence $(\psi_n)$ is increasing and its pointwise limit is $\underline{\lim}\xi_n$, so that the monotone convergence theorem gives the result. □

**Mean.** A real valued random variable $\xi$ is *integrable* if $\mathbf{E}\,|\xi| < \infty$. The *mean* (or *expectation*) of an integrable real valued random variable $\xi$ is defined as

$$\mathbf{E}\xi = \mathbf{E}\xi^+ - \mathbf{E}\xi^-$$

where $\xi^+$ and $\xi^-$ are the positive and negative parts of $\xi$, respectively.

Most inequalities involving the mean remain valid admitting $\pm\infty$ as possible values and using the natural order of the extended real line $\overline{\mathbf{R}} = [-\infty, \infty]$. We say that the mean of the random variable $\xi$ *exists* if

$$\min\left\{\mathbf{E}\xi^+, \mathbf{E}\xi^-\right\} < \infty$$

and, if this is the case, we define the mean of $\xi$ as $\mathbf{E}\xi = \mathbf{E}\xi^+ - \mathbf{E}\xi^- \in \overline{\mathbf{R}}$.

Sometimes it is useful to consider complex valued random variables $\xi : \Omega \to \mathbf{C}$. They are called integrable if, again, $\mathbf{E}\,|\xi| < \infty$, and their mean is defined as $\mathbf{E}\xi = \mathbf{E}\,(\Re\xi) + i\mathbf{E}\,(\Im\xi)$.

In analysis, the mean $\mathbf{E}\xi$ is known as the "abstract Lebesgue integral of the measurable function $\xi$ with respect to the measure $\mathbf{P}$", and denoted by $\int_\Omega \xi\,(\omega)\,d\mathbf{P}\,(\omega)$ or simply $\int \xi d\mathbf{P}$. Also, the Lebesgue integral makes sense with respect to any positive measure, finite or not. If $\ell$ denotes the Lebesgue measure on the real line, the Lebesgue integral of a measurable function $\xi$ w.r.t. $\ell$ is denoted by $\int_{\mathbf{R}} \xi\,(x)\,dx$. If $\mu$ is the Lebesgue-Stieltjes measure defined by the distribution function $F$, then the Lebesgue integral of a measurable function $\xi$ w.r.t. $\mu$ is denoted by $\int_{\mathbf{R}} \xi\,(x)\,dF\,(x)$.

**Lebesgue dominated convergence theorem.** *Let $(\xi_n)$ be a sequence of random variables such that the pointwise limit $\xi_n \to \xi$ exists. If $|\xi_n| \leq |\psi|$ for all $n$ and some integrable random variable $\psi$, then $\xi$ is integrable, $\mathbf{E}\,|\xi_n - \xi| \to 0$ and $\mathbf{E}\xi_n \to \mathbf{E}\xi$.*

**proof.** The inequality $|\xi| \leq |\psi|$ implies that $\xi$ is integrable. Then, just apply Fatou lemma to the nonnegative variables $2\,|\psi| - |\xi_n - \xi|$ and take the limit as $n \to \infty$. □

**Properties of the mean.** Properties of the mean value of integrable variables follow directly from the elementary properties that holds for simple random variables. The mean value is linear, i.e.

$$\mathbf{E}\,(\xi + a\eta) = \mathbf{E}\xi + a \cdot \mathbf{E}\eta$$

for any constant $a$ and any integrable random variables $\xi$ and $\eta$, and is monotone, i.e.

$$\mathbf{E}\xi \leq \mathbf{E}\eta$$

if $\xi \leq \eta$ on a set of probability one. In particular, $|\mathbf{E}\xi| \leq \mathbf{E}\,|\xi|$. Observe also that $\mathbf{E}\,|\xi| = 0$ implies that $\xi = 0$ on a set of probability one.

If the random variables$\xi$ and $\eta$ are integrable and independent, then

$$\mathbf{E}\xi\eta = \mathbf{E}\xi \cdot \mathbf{E}\eta$$

as follows approximating with simple variables.

**Inequalities.** Particularly interesting, both in analysis and in probability, are inequalities involving means of random variable. The most elementary is

**Chebyshev inequality.** *Let $\xi$ be an integrable nonnegative random variable. Then, for any $\varepsilon > 0$,*

$$\mathbf{P}\,\{\xi \geq \varepsilon\} \leq \frac{1}{\varepsilon}\mathbf{E}\xi$$

**proof.** Just observe that

$$\mathbf{P}\{\xi \geq \varepsilon\} = \mathbf{E}1_{\{\xi \geq \varepsilon\}} \leq \mathbf{E}\left(1_{\{\xi \geq \varepsilon\}} \cdot \frac{\xi}{\varepsilon}\right) \leq \mathbf{E}\left(\frac{\xi}{\varepsilon}\right) = \frac{1}{\varepsilon}\mathbf{E}\xi$$

□

The mean w.r.t. a probability measure is a kind of convex combination of the values of a random variable. It is not surprising that we can deduce inequalities from convexity arguments. The prototype is

**Jensen's inequality** . *If $\xi$ is a real valued integrable random variable and $h : \mathbf{R} \to \mathbf{R}$ is a convex function, then*

$$h(\mathbf{E}\xi) \leq \mathbf{E}h(\xi)$$

**proof.** Indeed, a convex function satisfies $h(x') \geq h(x) + \lambda(x) \cdot (x' - x)$ for any $x, x' \in \mathbf{R}$ and some $\lambda(x) \in \mathbf{R}$. Taking $x = \mathbf{E}\xi$ and $x' = \xi$ we get

$$h(\xi) \geq h(\mathbf{E}\xi) + \lambda(\mathbf{E}\xi) \cdot (\xi - \mathbf{E}\xi)$$

and integrating we obtain the result. □

**Lyapunov's inequality.** *Let $\xi$ be a random variable and $p' > p > 0$. Then*

$$(\mathbf{E}|\xi|^p)^{1/p} \leq \left(\mathbf{E}|\xi|^{p'}\right)^{1/p'}$$

**proof.** Apply Jensen inequality to the convex function $h(\xi) = |\xi|^{p'/p}$. □

**Hölder (and Cauchy-Schwarz) inequality.** *Let $p > 1$ and $1/p + 1/q = 1$ (such pair of exponents are called conjugate). Let $\xi$ and $\eta$ be a random variable with $\mathbf{E}|\xi|^p < \infty$ and $\mathbf{E}|\eta|^q < \infty$. Then*

$$\mathbf{E}|\xi\eta| \leq (\mathbf{E}|\xi|^p)^{1/p} \cdot (\mathbf{E}|\eta|^q)^{1/q}$$

*The particular case $p = q = 2$ is called Cauchy-Schwarz inequality, and reads*

$$\mathbf{E}|\xi\eta| \leq \sqrt{\mathbf{E}\xi^2} \cdot \sqrt{\mathbf{E}\eta^2}$$

**proof.** If $\mathbf{E}|\xi|^p$ or $\mathbf{E}|\eta|^q$ are zero, there is nothing to prove, since then $|\xi\eta| = 0$ outside a set of probability zero. If both are positive, we may define $\xi' = \xi/(\mathbf{E}|\xi|^p)^{1/p}$ and $\eta' = \eta/(\mathbf{E}|\eta|^q)^{1/q}$. The elementary inequality $a^{1/p}b^{1/q} \leq \frac{1}{p}a + \frac{1}{q}b$ (holding for positive $a$ and $b$) applied to $a = |\xi'|^p$ and $b = |\eta'|^q$ gives $|\xi'\eta'| \leq \frac{1}{p}|\xi'|^p + \frac{1}{q}|\eta'|^q$. Integrating we get $\mathbf{E}|\xi'\eta'| \leq 1$, which is equivalent to Hölder inequality. □

**Minkowski inequality.** *Let $p \geq 1$, and let $\xi$ and $\eta$ be a random variable with $\mathbf{E}|\xi|^p < \infty$ and $\mathbf{E}|\eta|^p < \infty$. Then*

$$(\mathbf{E}|\xi + \eta|^p)^{1/p} \leq (\mathbf{E}|\xi|^p)^{1/p} + (\mathbf{E}|\eta|^p)^{1/p}$$

**proof.** The elementary inequality $\left(\frac{a+b}{2}\right)^p \leq \frac{1}{2}(a+b)^p$ gives

$$|\xi+\eta|^p \leq (|\xi|+|\eta|)^p \leq 2^{p-1}(|\xi|^p+|\eta|^p)$$

Taking the mean, we see that $\mathbf{E}|\xi+\eta|^p$ is finite. Now write

$$|\xi+\eta|^p = |\xi|\cdot|\xi+\eta|^{p-1} + |\eta|\cdot|\xi+\eta|^{p-1}$$

The Hölder inequality applied to the two terms on the right gives

$$\mathbf{E}|\xi+\eta|^p \leq (\mathbf{E}|\xi+\eta|^p)^{1/q}\cdot\left((\mathbf{E}|\xi|^p)^{1/p}+(\mathbf{E}|\eta|^p)^{1/p}\right)$$

where $q$ is such that $1/p+1/q=1$. □

**A.e. convergence.** Pointwise convergence is not particularly meaningful in probability. Interesting sequences of random variables, nevertheless, and this is the content of most theorems of the theory, do converge in some weaker but significative sense.

The sequence of random variables $(\xi_n)$ *converge almost everywhere* , or *with probability one*, to the random variable $\xi$, notation $\xi_n \to_{\text{a.e.}} \xi$, if the set of points where it does not converge has probability zero, i.e. if

$$\mathbf{P}\{\xi_n \to \xi\} = 1$$

This is the strongest nontrivial convergence. Pointwise convergence, of course, implies convergence a.e. Quite surprisingly, a.e. convergence implies uniform convergence outside sets of small probability, as the following theorem says.

**Egorov theorem.** *Let $(\xi_n)$ be a sequence of real (or complex) valued measurable functions defined in a probability space, such that $\xi_n \to_{\text{a.e.}} \xi$. For any $\varepsilon>0$ there exists a measurable set $E_\varepsilon$ such that $\xi_n \to \xi$ uniformly on $\Omega\backslash E_\varepsilon$ and $\mathbf{P}(E_\varepsilon)<\varepsilon$.*

**proof.** Fixed $k$, the sequence of events $E_n^k = \cap_{m\geq n}\{|\xi_m-\xi|>1/k\}$ is decreasing and has intersection of zero measure. From the continuity of the probability measure there follows that, given $\varepsilon>0$, there exists $n_k$ such that $\mathbf{P}\left(E_{n_k}^k\right)<\varepsilon 2^{-k}$. Then, the set $E_\varepsilon = \cup_{k\geq 1}E_{n_k}^k$ has measure $<\varepsilon$, and $\xi_n\to\xi$ uniformly on $\Omega\backslash E_\varepsilon$. □

$L^p$

**Convergence.** Minkowski inequality shows that the function

$$\xi \mapsto \|\xi\|_p = (\mathbf{E}|\xi|^p)^{1/p}$$

is a pseudonorm on the space of random variables with $\mathbf{E}|\xi|^p<\infty$, as long as $1\leq p<\infty$.

The essential supremum of the real valued random variable $\xi$ is defined as

$$\operatorname{ess\,sup}\xi = \inf\{\lambda \text{ s.t. } \mathbf{P}(\xi>\lambda)=0\}$$

The function $\xi\mapsto\|\xi\|_\infty = \operatorname{ess\,sup}|\xi|$ also is a pseudonorm on the vector space of random variables with $\operatorname{ess\,sup}|\xi|<\infty$.

Let $1\leq p<\infty$. The sequence of random variables $(\xi_n)$ *converge in $L^p$ norm* to the random variable $\xi$, notation $\xi_n\to_{L^p}\xi$, if

$$\mathbf{E}|\xi_n-\xi|^p\to 0$$

as $n\to\infty$. When $p=2$ this is called *mean square convergence.*

**Theorem.** *Let $1\leq p\leq\infty$, and let $(\xi_n)$ be a Cauchy sequence in the pseudonorm $\|\cdot\|_p$. Then there exists a subsequence $(\xi_{n_k})$ and a random variable $\xi$ such that $\xi_{n_k}\to_{\text{a.e.}}\xi$. Moreover, $\|\xi_n-\xi\|_p\to 0$ as $n\to\infty$.*

**proof.** Let $1 \leq p < \infty$, and let $(\xi_n)$ be a Cauchy sequence w.r.t. $\|\cdot\|_p$. There exists a subsequence $(\xi_{n_k})$such that $\left\|\xi_{n_{k+1}} - \xi_{n_k}\right\|_p < 2^{-k}$ for any $k$. There follows from the (easy half of the) Borel-Cantelli lemma that the series

$$|\xi_{n_1}| + \sum_{k=1}^{\infty} \left|\xi_{n_{k+1}} - \xi_{n_k}\right|$$

converges on a set of probabilty one, hence there exists a random variable $\xi$ such that

$$\xi_{n_k} = \xi_{n_1} + \sum_{k=1}^{k-1} \left(\xi_{n_{k+1}} - \xi_{n_k}\right) \to \xi$$

almost everywhere. One then uses the Fatou theorem and triangular inequality to show that indeed $\|\xi_n - \xi\|_p \to 0$.

Let $(\xi_n)$ be a Cauchy sequence w.r.t. $\|\cdot\|_\infty$. The sequence converges uniformly outside the union of the sets $\{|\xi_n| > \|\xi_n\|_\infty\}$ and $\left\{|\xi_k - \xi_j| > \|\xi_k - \xi_j\|_\infty\right\}$, for $n, k, j \in \mathbf{N}$. Since they are sets of probability zero, their union is, hence there exists a bounded random variable $\xi$ such that $\xi_n \to \xi$ on a set of probability one (indeed, one can define $\xi = 0$ where the sequence does not converge). There follows easily that $\|\xi_n - \xi\|_\infty \to 0$. □

**Spaces of random variables.** Consider the space of random variables $\xi$ such that $\mathbf{E}\,|\xi|^p < \infty$, equipped with the pseudonorm $\|\cdot\|_p$. One can define the quotient normed space $L^p(\mathbf{P})$ made of equivalence classes of random variables, where $\xi$ and $\eta$ are identified if $\|\xi - \eta\|_p = 0$, i.e. if the set where $\xi \neq \eta$ has probability zero. The above theorem shows that $L^p(\mathbf{P})$ is complete, hence is a Banach space. In particular, $L^p$-convergence is actually convergence in the space $L^p(\mathbf{P})$.

Hölder inequality takes the form

$$\mathbf{E}\,|\xi\eta| \leq \|\xi\|_p \cdot \|\eta\|_q$$

if $1 < p < \infty$ and $1/p + 1/q = 1$, and generalizes as

$$\mathbf{E}\,|\xi\eta| \leq \|\xi\|_1 \cdot \|\eta\|_\infty$$

if $p = 1$.

Observe that, as follows from the very definition of mean, simple random variables are dense in any of the spaces $L^p$.

The space $L^2(\mathbf{P})$, equipped with the inner product defined by $\langle \xi, \eta \rangle = \mathbf{E}\xi\overline{\eta}$, is a Hilbert space, as follows from the Cauchy-Schwarz inequality.

**Weak forms of convergence.** The sequence of random variables $(\xi_n)$ *converge in probability*, to the random variable $\xi$, notation $\xi_n \to_{\mathbf{P}} \xi$, if, given any $\varepsilon > 0$,

$$\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \to 0$$

as $n \to \infty$. This definition is motivated by the law of large number. Trivially, convergence almost everywhere implies convergence in distribution. Also, $L^p$ convergence for some $p \geq 1$ implies convergence in probability, as follows from Chebyshev inequality.

The sequence of random variables $(\xi_n)$ *converge in distribution* , or *in law*, to the random variable $\xi$, notation $\xi_n \to_{\mathrm{d}} \xi$ or $\xi_n \to_{\mathcal{L}} \xi$, if

$$\mathbf{E}f(\xi_n) \to \mathbf{E}f(\xi)$$

as $n \to \infty$ for any bounded continuous function $f$. This is equivalent to the condition that the distribution functions $F_{\xi_n}(t)$ converge to the distribution function $F_\xi(t)$ for any point of continuity $t$ of the latter. Convergence in probability implies convergence in distribution.

**Densities.** Let $(\Omega, \mathcal{E})$ be a measurable space. Given a probability measure $\mathbf{P}$ over $\mathcal{E}$ and a nonnegative random variable $\eta : \Omega \to [0, \infty]$, the set function $\mathbf{Q} : A \mapsto \mathbf{E}1_A\eta$ is a measure (not necessarily finite) over $\mathcal{E}$. The random variable $\eta$ is said the *density*, or *Radon-Nikodym derivative*, of $\mathbf{Q}$ with respect to $\mathbf{P}$, and denoted by $d\mathbf{Q}/d\mathbf{P}$. The measure $\mathbf{Q}$ as above is *absolutely continuos* with respect to the measure $\mathbf{P}$, namely satisfies $\mathbf{Q}(A) = 0$ for any $A \in \mathcal{E}$ with $\mathbf{P}(A) = 0$. The converse of this observation is the following

**Radon-Nikodym theorem.** *Let $(\Omega, \mathcal{E}, \mathbf{P})$ be a probability space, and let $\mathbf{Q}$ be a finite measure over $\mathcal{E}$ which is absolutely continuous with respect to $\mathbf{P}$. Then there exists a nonnegative integrable random variable $\eta$ such that $\mathbf{Q}(A) = \mathbf{E}1_A\eta$ for any $A \in \mathcal{E}$.*

**proof.** Let $\mathbf{Q}'$ be the positive measure $\mathbf{Q} + \mathbf{P}$. If $\xi \in L^2(\mathbf{Q}')$, the Cauchy-Schwarz inequality yelds

$$\left|\int \xi d\mathbf{Q}\right| \leq \int |\xi| \, d\mathbf{Q} \leq \int |\xi| \, d\mathbf{Q}' \leq \left(\int |\xi|^2 \, d\mathbf{Q}'\right)^{1/2} \cdot (\mathbf{Q}'(\Omega))^{1/2}$$

Since $\mathbf{Q}'(\Omega)$ is bounded, the map

$$\xi \mapsto \int \xi d\mathbf{Q}$$

define a bounded linear functional in $L^2(\mathbf{Q}')$. There follows from the standard theory of Hilbert spaces that there exists a random variable $\psi \in L^2(\mathbf{Q}')$ such that

$$\int \xi d\mathbf{Q} = \int \xi\psi d\mathbf{Q}'$$

for any $\xi \in L^2(\mathbf{Q}')$. Taking $\xi = 1$ in the above formula, we see that $0 \leq \int \psi d\mathbf{Q}' = \mathbf{Q}(\Omega) \leq \mathbf{Q}'(\Omega)$, so that we may take $0 \leq \psi \leq 1$ for $\mathbf{Q}'$-almost any point. Using the definition of $\mathbf{Q}'$, we rewrite the above identity as

$$\int \xi(1-\psi) \, d\mathbf{Q} = \int \xi\psi d\mathbf{P}$$

Since $\mathbf{Q}$ is absolutely continuous w.r.t. $\mathbf{P}$, $\mathbf{P}(\psi = 1) = 0$ (as follows taking $\xi = 1_{\{\psi=1\}}$ above) hence $\psi < 1$ for $\mathbf{P}$-almost any point. Now, let $A \in \mathcal{E}$ be an arbitrary event and take $\xi = \left(1 + \psi + \psi^2 + ... + \psi^n\right)$. The identity reads

$$\int 1_A\left(1 - \psi^{n+1}\right) d\mathbf{Q} = \int 1_A\left(\psi + \psi^2 + ... + \psi^{n+1}\right) \psi d\mathbf{P}$$

Since $0 \leq \psi < 1$ for $\mathbf{P}$-almost any point, the left side converges, by the monotone convergence theorem, to $\mathbf{Q}(A)$ as $n \to \infty$. Also, the monotone convergence theorem says that $\psi + \psi^2 + ... + \psi^{n+1}$ convergences to a nonnegative random variable $\eta$ as $n \to \infty$, so that we get the desired result

$$\mathbf{Q}(A) = \int 1_A \eta d\mathbf{P}$$

Taking $A = \Omega$ one finally sees that $\mathbf{E}\eta = \mathbf{Q}(\Omega) < \infty$, so that $\eta$ is integrable. □

**Conditional mean.** Let $(\Omega, \mathcal{E}, \mathbf{P})$ be a probability space, and let $\mathcal{F}$ be a sub-$\sigma$-algebra of $\mathcal{E}$. Given an integrable random variable $\xi$, there exists a unique random variable $\mathbf{E}(\xi|\mathcal{F})$, called the *conditional mean* of $\xi$ w.r.t. $\mathcal{F}$, which is $\mathcal{F}$-measurable (i.e. the inverse image of any Borel set belongs to $\mathcal{F}$) and such that

$$\mathbf{E}(1_A\mathbf{E}(\xi|\mathcal{F})) = \mathbf{E}1_A\xi$$

for any $A \in \mathcal{F}$. Indeed, if $\xi \geq 0$ then one can define $\mathbf{E}(\xi|\mathcal{F})$ as equal to the Radon-Nikodym derivative of the measure $A \mapsto \mathbf{E}1_A\xi$, defined on $\mathcal{F}$, with respect to the restriction of $\mathbf{P}$ to $\mathcal{F}$. The general case is defined by linearity, writing $\xi$ as a difference of two nonegative random variables. Uniqueness is intended a.e., i.e. modulo sets of zero probability.

The conditional mean is monotone, i.e. if $\xi \geq 0$ then $\mathbf{E}(\xi|\mathcal{F}) \geq 0$, and preserves the mean value, since

$$\mathbf{E}\mathbf{E}(\xi|\mathcal{F}) = \mathbf{E}\xi$$

It can be considered as a "projection" of $\xi$ onto the space of $\mathcal{F}$-measurable random variable, equipped with the $L^1$-norm. In particular, if $\mathcal{F}$ is the trivial $\sigma$-algebra $\{\emptyset, \Omega\}$, then $\mathbf{E}(\xi|\mathcal{F})$ is constant a.e. and equal to $\mathbf{E}\xi$.

To understand the meaning of the conditional mean, observe that if $\mathcal{F}$ is the $\sigma$-algebra generated by a partition $\{D_1, D_2, ..., D_k, ...\}$ with atoms of positive measure, then $\mathbf{E}(\xi|\mathcal{F})$ is constant on each atom $D_k$ and takes value

$$\mathbf{E}(\xi|\mathcal{F})(\omega) = \frac{\mathbf{E}1_{D_k}\xi}{\mathbf{P}(D_k)}$$

on a.e. $\omega \in D_k$. Hence, the above property reads

$$\mathbf{E}\xi = \mathbf{E}\mathbf{E}\left(\xi|\mathcal{F}\right) = \sum_k \mathbf{E}1_{D_k}\xi$$

and may be thought as a generalization of the formula of the total probability.

Given a random variable $\eta$, we may consider the $\sigma$-algebra generated by $\eta$, namely $\mathcal{E}_\eta = \eta^{-1}\mathcal{B}\left(\mathbf{R}\right)$. The conditional mean $\mathbf{E}\left(\xi|\mathcal{E}_\eta\right)$ is then denoted by $\mathbf{E}\left(\xi|\eta\right)$.

# 11 Leis dos grandes números

**Desigualdades de Chebyshev.** Seja $\xi$ uma variável aleatória discreta não negativa. Se a variável $\xi$ é integrável, então

$$\mathbf{P}\left\{\xi \geq \varepsilon\right\} \leq \frac{1}{\varepsilon}\mathbf{E}\xi$$

para todo $\varepsilon > 0$. A demonstração é simplesmente a sequência de estimações elementares

$$\mathbf{P}\left\{\xi \geq \varepsilon\right\} = \mathbf{E}1_{\{\xi\geq\varepsilon\}} \leq \mathbf{E}\left(1_{\{\xi\geq\varepsilon\}} \cdot \frac{\xi}{\varepsilon}\right) \leq \mathbf{E}\left(\frac{\xi}{\varepsilon}\right) = \frac{1}{\varepsilon}\mathbf{E}\xi$$

Este é o protótipo de uma família de desigualdades, obtidas escolhendo oportunamente a variável $\xi$ em função de outras.

Um caso particular é a *desigualdade de Markov*: se $\xi$ é uma variável aleatória discreta integrável e $\varepsilon > 0$ então

$$\mathbf{P}\left\{|\xi| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon}\mathbf{E}\left|\xi\right|$$

Outro caso particular, obtido considerando a variável $\left|\xi - \mathbf{E}\xi\right|^2$, é a *desigualdade de Chebyshev*: se $\xi$ é uma variável aleatória discreta com variância finita e $\varepsilon > 0$ então

$$\mathbf{P}\left\{|\xi - \mathbf{E}\xi| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2}\mathbf{V}\xi$$

A desigualdade de Chebyshev não é, em geral, uma boa estimação. Melhores costumam ser as desigualdades

$$\mathbf{P}\left\{|\xi - \mathbf{E}\xi| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^k}\mathbf{E}\left|\xi - \mathbf{E}\xi\right|^k$$

quando $k$ cresce. A sua importância é teórica: permite provar uma forma da lei dos grandes números com um esforço mínimo.

Ainda melhor costuma ser a seguinte *desigualdade de Chebyshev exponencial*: se a variável $\xi$ é tal que $e^{\beta\xi}$ tém esperança finita para todo $\beta > 0$, então

$$\mathbf{P}\left(\xi \geq \varepsilon\right) = \mathbf{P}\left(e^{\beta\xi} \geq e^{\beta\varepsilon}\right) \leq e^{-\beta\varepsilon}\mathbf{E}e^{\beta\xi}$$

para todo $\beta > 0$, e portanto

$$\mathbf{P}\left(\xi \geq \varepsilon\right) \leq e^{-H(\varepsilon)}$$

onde a função "entropia" $H$ é a transformada de Legendre da função $\beta \mapsto \log \mathbf{E}e^{\beta\xi}$, definida por

$$H\left(\lambda\right) = \sup_{\beta>0}\left(\beta\lambda - \log \mathbf{E}e^{\beta\xi}\right)$$

Esta desigualdade joga um papel central na teoria dos grandes desvios.

**Médias empíricas.** Seja $(\xi_k)$ uma sucessão de variáveis aleatórias definidas num espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$, e sejam $S_n$ "as somas parciais das $\xi_k$", definidas por $S_n = \xi_1 + \xi_2 + ... + \xi_n$. As leis dos grandes números são afirmações acerca da convergência das "médias empíricas" $S_n/n$ quando $n$ é grande.

Se as variáveis são identicamente distribuídas, i.e. são "réplicas" de uma variável fixada $\xi$, então a esperança de $S_n/n$ é igual a $\mathbf{E}\xi$, ou seja é constante, não depende de $n$. Se as variáveis são também independentes, a variância $\mathbf{V}\left(S_n/n\right)$ é igual a $\frac{1}{n}\mathbf{V}\xi$, e portanto decresce quando $n$ cresce. Isto leva a conjecturar que, quando $n$ é grande, a variável $S_n/n - \mathbf{E}\xi$ é "pequena" com grande probabilidade, i.e, numa linguagem muito informal,

$$\frac{S_n}{n} \sim \mathbf{E}\xi$$

Por exemplo, se $\xi_k$ são provas de Bernoulli com probabilidade de sucesso $p$, então $S_n$ é o número de sucessos em $n$ provas, e $S_n/n$ tem a interpretação da "frequência de sucessos em $n$ provas". A sua esperança é $\mathbf{E}(S_n/n) = p$ e a sua variância é

$$\mathbf{E}\left|\frac{S_n}{n} - p\right|^2 = \frac{pq}{n}$$

A conjectura agora é

$$\frac{S_n}{n} \sim p$$

A lei dos grandes números, o resultado que dá razão de existir à teoria das probabilidades e que explica o significado "físico" da esperança, formaliza esta expectativa.

**Lei dos grandes números.** *Sejam $\xi_1, \xi_2, ...$ variáveis aleatórias independentes e identicamente distribuídas, com média $\mathbf{E}\xi = m$ e variância finita, e seja $S_n = \xi_1 + \xi_2 + ... + \xi_n$. Então para todo $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{S_n}{n} - m\right| < \varepsilon\right\} = 1$$

**dem.** Um cálculo mostra que $\mathbf{E}(S_n/n) = \mathbf{E}\xi$ e $\mathbf{V}(S_n/n) = \frac{1}{n}\mathbf{V}\xi$. A desigualdade de Chebyshev diz que, dado $\varepsilon > 0$,

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right\} \leq \frac{\mathbf{V}\xi}{n\varepsilon^2}$$

e portanto

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - m\right| < \varepsilon\right\} \geq 1 - \frac{\mathbf{V}\xi}{n\varepsilon^2} \to 1$$

quando $n \to \infty$. □

**obs.** *(convergência em probabilidade)* A lei dos grandes números costuma ser enunciada como "sejam ... então $S_n/n \to_{\mathbf{P}} m$", que se lê: as médias empíricas $S_n/n$ convergem para o valor médio $m$ "em probabilidade".

**Lei dos grandes números de Chebyshev.** Se $\xi_1, \xi_2, ...$ são variáveis independentes mas não necessariamente identicamente distribuídas, ainda é possível provar uma "lei dos grandes números". De facto, uma condição uniforme sobra as variâncias, como $\mathbf{V}\xi_k \leq K < \infty$ para todo $k$, é suficiente para utilizar a desigualdade de Chebyshev e provar que

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{S_n}{n} - \mathbf{E}\left(\frac{S_n}{n}\right)\right| < \varepsilon\right\} = 1$$

para todo $\varepsilon > 0$.

**Lei dos grandes números de Markov.** A hipótese de independência também não é necessária. Se $\xi_1, \xi_2, ...$ são variáveis aleatórias tais que $\lim_{n\to\infty} \frac{1}{n^2}\mathbf{V}S_n = 0$, então

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{S_n}{n} - \mathbf{E}\left(\frac{S_n}{n}\right)\right| < \varepsilon\right\} = 1$$

para todo $\varepsilon > 0$.

**Lei dos grandes números de Bernoulli.** Se $\xi_1, \xi_2, ...$ são variáveis independentes e identicamente distribuídas com lei Bernoulli $B(1,p)$, então $S_n$ tem lei binomial $B(n,p)$ e representa o número de sucessos em $n$ provas de Bernoulli. A lei dos grandes números lê-se então

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| < \varepsilon\right\} = 1$$

para todo $\varepsilon > 0$, e mostra em que sentido a frequência dos sucessos em $n$ experiências repetidas e independentes "aproxima" a probabilidade de sucesso $p$.

E natural considerar "típicos" os eventos com $|S_n/n - p| < \varepsilon$, cuja probabilidade é assimptoticamente igual a um. Isto não quer dizer que os eventos com $|S_n - np| \neq 0$ sejam desprezáveis! Pelo contrário, a variância de $S_n$ é proporcional a $n$, e portanto é razoavel suspeitar que $|S_n - np| \sim \sqrt{n}$. Aliás, se pensamos em $S_n$ como a posição ao tempo $n$ de uma marcha aleatória (dar um passo

para a frente por cada sucesso, e ficar parado por cada insucesso), a lei dos grandes números só diz que é muito provável observar trajectórias "encaixadas" entre as retas $n(p \pm \varepsilon)$, i.e.

$$n(p-\varepsilon) < S_n < n(p+\varepsilon)$$

quando $n$ é suficientemente grande, onde $\varepsilon$ é um número positivo arbitrário.

Se $p = 1/2$, a variável $T_n = 2S_n - n$ representa a posição ao tempo $n$ de uma marcha aleatória simétrica. A lei dos guandes números diz que, se $n$ é suficientemente grande, as trajecórias satisfazem $|T_n| < n\varepsilon$ com probabilidade muito próxima de um.

**Lei dos grandes números e observações.** A lei dos grandes números é um resultado bonito. Um matemático lê o teorema, que diz "se $\xi_1, \xi_2, ...$ bla bla bla ... então $S_n/n \to_{\mathbf{P}} p$", e fica feliz. Um físico não, ele quer saber qual é a informação "física" do teorema. O teorema diz que, fixado um "erro" $\varepsilon$ e uma probabilidade $\alpha$, se $n$ é suficientemente grande a probabilidade de observar uma frequência de sucessos $S_n/n$ que difere de $p$ por mais que $\varepsilon$ é menor que $\alpha$. A previsão física é, se $\alpha$ é muito pequeno, "a frequência satisfaz $|S_n/n - p| < \varepsilon$ na esmagadora maioria das vezes que repetimos as $n$ experiências". Enfím, o enunciado "$S_n/n \to_{\mathbf{P}} p$" é simplesmente uma maneira elegante de enunciar a previsão "ao repetir um número muito grande de vezes a experiência, é muito provável observarvar uma frequência de sucessos muito perto de $p$". É neste sentido que $p$, a probabilidade do evento "sucesso", é um observável físico. Acontece que a informação quantitativa está contida na demonstração, e a desigualdade de Chebyshev fornece uma relação entre $\varepsilon$, $\alpha$ e $n$, embora não seja a melhor possível. Determinar o $n$ optimal em função de $\varepsilon$ e $\alpha$, ou seja a velocidade de convergência na lei dos grandes números é um problema físico relevante, pois se a convergência for muito lenta a lei pode não ser observavel! Este problema é tratado pela teoria dos grandes desvios.

**Lei dos grandes números de Poisson: provas com probabilidade de sucesso variável.** Sejam $\xi_1, \xi_2, ..., \xi_k, ...$ variáveis independentes com leis de Bernoulli com parâmentros variáveis, por exemplo $\xi_k \sim B(1, p_k)$ onde $p_k \in [0,1]$, e seja $S_n = \xi_1 + \xi_2 + ... + \xi_n$. Sejam $\overline{p}_n = \frac{1}{n}(p_1 + p_2 + ... + p_n)$ a "probabilidade média nas primeiras $n$ provas", e $\overline{\sigma}_n^2 = \overline{p}_n(1 - \overline{p}_n)$. As frequências empíricas $S_n/n$ satisfazem a lei dos grandes números, i.e.

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{S_n}{n} - \overline{p}_n\right| < \varepsilon\right\} = 1$$

porque $\mathbf{V}\xi_k = p_k(1-p_k) \leq \sup_{0\leq x\leq 1} x(1-x) = 1/4$ para todo $k$.

Mais interessante é observar que $\mathbf{V}(S_n/n) \leq \overline{\sigma}_n^2/n$, e a igualdade é satisfeita sse todos os $p_k$ com $k = 1, 2, ..., n$ são iguais à média $\overline{p}_n$. Portanto, embora isto pareça paradoxal, a variabilidade dos parâmetros diminui a incerteza sobre a frequência $S_n/n$ (mas na verdade isto é intuitivo, se eu lançar 50 moedas com $p = 1$ e 50 moedas com $p = 0$, com "certeza" observo 50 caras e 50 coroas, ou seja uma frequência exactamente igual a $1/2$).

**Lei dos grandes números de Khinchin.** *Sejam $\xi_1, \xi_2, ...$ variáveis aleatórias independentes e identicamente distribuídas, com média finita $\mathbf{E}\xi = m$, e seja $S_n = \xi_1 + \xi_2 + ... + \xi_n$. Então para todo $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{S_n}{n} - m\right| < \varepsilon\right\} = 1$$

**dem.** Fixados $\delta > 0$ e $n \in \mathbf{N}$, a ideia é escrever cada $\xi_k$ como uma soma $\xi_k^{\mathrm{b}} + \xi_k^{\mathrm{u}}$ onde

$$\xi_k^{\mathrm{b}} = \xi_k \cdot 1_{\{|\xi_k| < \delta n\}} \qquad \text{e} \qquad \xi_k^{\mathrm{u}} = \xi_k \cdot 1_{\{|\xi_k| \geq \delta n\}}$$

A variável $\xi_k^{\mathrm{b}}$ é limitada, e satisfaz

$$\mathbf{E}\xi_k^{\mathrm{b}} \to m \text{ quando } n \to \infty$$

$$\mathbf{V}\xi_k^{\mathrm{b}} \leq \delta n \cdot \mathbf{E}|\xi| < \infty$$

Utilizando a desigaldade de Chebyshev e a desigualdade do triângulo, vê-se que

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k^{\mathrm{b}}-m\right|\geq\varepsilon\right)\leq\frac{4\delta}{\varepsilon^2}\cdot\mathbf{E}\left|\xi\right|$$

se $n$ é suficientemente grande. Por outro lado,

$$\mathbf{P}\left(\xi_k^{\mathrm{u}}\neq 0\right)=\mathbf{P}\left(\left|\xi_k\right|\geq\delta n\right)\leq\frac{1}{\delta n}\cdot\mathbf{E}\left|1_{\{|\xi_k|\geq\delta n\}}\xi_k\right|\leq\frac{\delta}{n}$$

se $n$ é suficientemente grande, donde

$$\mathbf{P}\left(\sum_{k=1}^{n}\xi_k^{\mathrm{u}}\neq 0\right)\leq\sum_{k=1}^{n}\mathbf{P}\left(\xi_k^{\mathrm{u}}\neq 0\right)\leq\delta$$

Então

$$\begin{aligned}\mathbf{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k-m\right|\geq\varepsilon\right) &\leq \mathbf{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k^{\mathrm{b}}-m\right|\geq\varepsilon\right)+\mathbf{P}\left(\sum_{k=1}^{n}\xi_k^{\mathrm{u}}\neq 0\right)\\ &\leq \frac{4\delta}{\varepsilon^2}\cdot\mathbf{E}\left|\xi\right|+\delta\end{aligned}$$

e o resultado vem da arbitrariedade de $\delta$. □

**Um sermão.** Vale a pena insistir sobre o conteúdo "físico" da lei dos grandes números, o resultado que dá razão de existir à teoria das probabilidades, e que muita confusão gera até em pessoas instruídas. A proposição 5.154 do *Tractatus Logico-Philosophicus* de Wittgenstein começa assim: "Suppose that an urn contains black and white balls in equal numbers (and none of any other kind). I draw one ball after another, putting them back into the urn. By this experiment I can establish that the number of black balls drawn and the number of white balls drawn approximate to one another as the drawn continues..." (na tradução de D.F. Pears e B.F. McGuinnes, ed. Routledge 1974). Esta afirmação é correcta ou falsa dependendo do significado das palavras "number" e "approximate". Ou elas não têm o mesmo significado em que um matemático pensa, ou o senhor nunca na vida teve a preocupação de fazer a experiência (eu aposto na primeira das hipóteses, embora igualmente grave, visto o habitual cuidado do autor acerca da utilização da linguagem!). Um modelo razoável da experiência em causa é o das $n$ provas de Bernoulli, com $p=1/2$. A diferença entre o número de bolas brancas e o número de bolas pretas é $T_n$, a posição de uma marcha aleatória simétrica. O que a lei dos grandes números diz é que, dado $\varepsilon>0$, se $n$ é suficientemente grande $T_n/n$ esta a distância menor que $\varepsilon$ de 0 com probabilidade muito grande. O que a lei dos grandes números não diz é que $|T_n|$ é pequeno! De facto, a diferença $|T_n|$ entre o número de bolas pretas e o número de bolas brancas escolhidas é, com grande probabilidade, da ordem de $\sqrt{n}$, ou seja muito grande, embora as possibilidades $T_n>0$ e $T_n>0$ sejam igualmente prováveis... Para compreender este fenómeno, é suficiente observar que, fixado $K>0$ arbitrário, a probabilidade $\mathbf{P}\left(|T_n|<K\right)\to 0$ quando $n\to\infty$.

**Demonstração de Bernstein do teorema de Weierstrass.** Um teorema de Weierstrass diz que os polinómios são densos no espaço das funções contínuas definidas no intervalo $[0,1]$ munido da norma do supremo. Isto quer dizer que toda função contínua $f:[0,1]\to\mathbf{R}$ pode ser aproximada abitrariamente bem por polinómios na norma do supremo. Bernstein descobriu a seguinte demonstração "probabilística" deste clássico teorema de análise, que tem a vantagem de ser "constructiva".

Sejam $f:[0,1]\to\mathbf{R}$ uma função contínua, e seja $\varepsilon>0$. Dado $n\in\mathbf{N}$, seja $Q_n:[0,1]\to\mathbf{R}$ o polinómio definido por

$$\begin{aligned}Q_n\left(x\right) &= \mathbf{E}f\left(S_{n,x}/n\right)\\ &= \sum_{k=0}^{n}f\left(k/n\right)\binom{n}{k}x^k\left(1-x\right)^{n-k}\end{aligned}$$

onde a variável $S_{n,x}$ tem lei binomial $B(n,x)$. Pela continuidade uniforme (pois o domínio é compacto) de $f$, existe $\delta > 0$ tal que $|f(x) - f(x')| < \varepsilon$ se $|x - x'| < \delta$. A diferença $|f(x) - Q_n(x)|$ é igual a

$$\begin{aligned} |f(x) - Q_n(x)| &= |f(x) - \mathbf{E}f(S_{n,x}/n)| = |\mathbf{E}(f(x) - f(S_{n,x}/n))| \\ &= \left|\mathbf{E}1_{\{|S_{n,x}/n-x|<\delta\}}(f(x) - f(S_{n,x}/n)) + \mathbf{E}1_{\{|S_{n,x}/n-x|\geq\delta\}}(f(x) - f(S_{n,x}/n))\right| \\ &\leq \left|\mathbf{E}1_{\{|S_{n,x}/n-x|<\delta\}}(f(x) - f(S_{n,x}/n))\right| + \left|\mathbf{E}1_{\{|S_{n,x}/n-x|\geq\delta\}}(f(x) - f(S_{n,x}/n))\right| \end{aligned}$$

O primeiro termo é estimado utilizando a continuidade uniforme de $f$, donde

$$\begin{aligned} \left|\mathbf{E}1_{\{|S_{n,x}/n-x|<\delta\}}(f(x) - f(S_{n,x}/n))\right| &\leq \varepsilon \cdot \left|\mathbf{E}1_{\{|S_{n,x}/n-x|<\delta\}}\right| \\ &\leq \varepsilon \end{aligned}$$

O segundo termo é estimado utilizando a lei dos grande números, de facto a desigualdade de Chebyshev, e resulta

$$\begin{aligned} \left|\mathbf{E}1_{\{|S_{n,x}/n-x|\geq\delta\}}(f(x) - f(S_{n,x}/n))\right| &\leq 2\|f\|_\infty \cdot \mathbf{P}\left(\left|\frac{S_{n,x}}{n} - x\right| \geq \delta\right) \\ &\leq 2\|f\|_\infty \cdot \frac{1/4}{n\delta^2} \end{aligned}$$

porque a variância de $S_{n,x}$ é $x(1-x) \leq 1/4$. Se $n$ é suficientemente grande, basta pôr $n \geq \|f\|_\infty / 2\varepsilon\delta^2$, temos que $\|f - Q_n\|_\infty \leq 2\varepsilon$ .

# 12 Teorema limite de De Moivre e Laplace

**Estimação da probabilidade de obter $k$ sucessos em $n$ provas de Bernoulli quando $n$ é grande.** Seja $S_n$ o número de sucessos em $n$ provas de Bernoulli, i.e. a variável com lei $B(n,p)$ onde $0<p<1$. Quando $n$ é grande, a expressão de $\mathbf{P}(S_n=k)$ é um horror, problemática até para um computador muito potente. Por outro lado, é intuitivo conjecturar que existem certos valores de $k$ que são muito pouco prováveis em relação a outros...

A lei dos grandes números sugere que os valores mais prováveis de $S_n/n$ são da ordem de $p$. De facto, é facil ver que $\mathbf{P}(S_n=k)$ é crescente se $k<np-q$ e é decrescente se $k>np-q$, onde utilizamos a notação tradicional $q=1-p$. Utilizando a fórmula de Stirling[2], ve-se que, quando $k\simeq np$ e $n$ é grande, a probabilidade $\mathbf{P}(S_n=k)$ é da ordem de

$$\mathbf{P}(S_n=k)=\binom{n}{k}p^k(1-p)^{n-k}\simeq\frac{1}{\sqrt{2\pi npq}}$$

Isto sugere que, embora a variável $S_n$ pode assumir $n+1$ valores, a probabilidade dela estar num intervalo de amplitude $\mathcal{O}(\sqrt{n})$ à volta de $np$ é da ordem de um, pelo menos se $n$ é grande[3]. De facto, a variância de $S_n$ é igual a $npq$, e, no mesmo espírito da lei dos grandes números, é natural conjecturar que

$$|S_n-np|\sim\sqrt{npq}$$

O teorema do limite central formaliza esta expectativa, e forneçe um "modelo assimptótico" para a lei da variável "normalizada" $S_n^*$, definida por

$$S_n^*=\frac{S_n-\mathbf{E}S_n}{\sqrt{\mathbf{V}S_n}}=\frac{S_n-np}{\sqrt{npq}}$$

A cada valor possível $k=0,1,2,...,n$ de $S_n$ corresponde um (e só um) valor

$$x=\frac{k-np}{\sqrt{npq}}$$

de $S_n^*$, entre $-\sqrt{np/q}$ e $\sqrt{nq/p}$. Quando $k$ varia num intervalo de amplitude $\mathcal{O}(\sqrt{n})$ à volta de $np$, o parâmetro $x$ varia num intervalo de amplitude $\mathcal{O}(1)$ (i.e. limitado) à volta de 0. De facto, a aproximação a seguir, será boa também para valores maiores, desde que o módulo de $x$ cresça sensivelmente menos do que $\sqrt{n}$.

Pela fórmula de Stirling, chamando $p'=k/n$ e $q'=1-p'$,

$$\begin{aligned}\mathbf{P}(S_n=k) &= \frac{1}{\sqrt{2\pi np'q'}}\left(p/p'\right)^k\left(q/q'\right)^{n-k}\cdot\alpha\\ &= \frac{1}{\sqrt{2\pi npq}}e^{-n\left(p'\cdot\log\left(p'/p\right)+q'\cdot\log\left(q'/q\right)\right)}\cdot\alpha\cdot\beta\\ &= \frac{1}{\sqrt{2\pi npq}}e^{-nH\left(p'\right)}\cdot\alpha\cdot\beta\end{aligned}$$

onde

$$\alpha=e^{\theta_1/n+\theta_2/k+\theta_3/(n-k)}\text{ com }0<\theta_i<1/12\ ,\qquad\beta=\sqrt{pq/p'q'}$$

e a função "entropia" $H$ é definida por

$$H(p')=p'\cdot\log(p'/p)+q'\cdot\log(q'/q)$$

É imediato ver que

$$\alpha\cdot\beta=1+\mathcal{O}\left(x/\sqrt{n}\right)$$

[2] A fórmula de Stirling diz que

$$n!=\sqrt{2\pi n}\cdot n^n\cdot e^{-n}\cdot e^{x_n/12n}$$

onde $x_n\in]0,1[$.

[3] A notação de Landau dos "O-grandes" e "o-pequenos" é a seguinte. Sejam $f$ e $g$ duas funções definidas numa vizinhança de $\infty$.

"$f(x)=\mathcal{O}(g(x))$ quando $x\to\infty$" quer dizer que o quociente $f/g$ é limitado numa vizinhança de $\infty$, ou seja que existem $K>0$ e $R\in\mathbf{R}$ tais que $|f(x)|\leq K\cdot|g(x)|$ para todo $x>R$.

"$f(x)=o(g(x))$ quando $x\to\infty$" quer dizer que o quociente $f/g$ converge para 0 em $\infty$, ou seja que $\lim_{x\to\infty}|f(x)|/|g(x)|=0$.

Por outro lado, dado que $p' - p = x\sqrt{pq}/\sqrt{n}$ e nos estamos interessados em valores pequenos de $x/\sqrt{n}$, o desenvolvimento de Taylor da função $H$ diz que

$$\begin{aligned} H(p') &= \frac{1}{2pq}(p'-p)^2 + \mathcal{O}\left((p'-p)^3\right) \\ &= \frac{1}{2n}x^2 + \mathcal{O}\left(\left(x/\sqrt{n}\right)^3\right) \end{aligned}$$

e portanto

$$e^{-nH(p')} = e^{-x^2/2}\cdot\gamma$$

onde

$$\gamma = 1 + \mathcal{O}\left(x^3/\sqrt{n}\right)$$

Juntando estas informações temos enfim que

$$\begin{aligned} \mathbf{P}(S_n = k) &= \frac{1}{\sqrt{2\pi npq}}e^{-x^2/2}\cdot\alpha\cdot\beta\cdot\gamma \\ &= \frac{1}{\sqrt{2\pi npq}}e^{-x^2/2}\cdot\left(1+\mathcal{O}\left(x/\sqrt{n}\right)\right)\cdot\left(1+\mathcal{O}\left(x^3/\sqrt{n}\right)\right) \end{aligned}$$

Em particular,

$$\sup_{|x|\le f(n)}\left|\frac{\mathbf{P}\left(\frac{S_n-np}{\sqrt{npq}}=x\right)}{\frac{1}{\sqrt{2\pi npq}}e^{-x^2/2}} - 1\right| \to 0$$

quando $n\to\infty$, se $f(n) = o\left(n^{1/6}\right)$. Este resultado é o "teorema limite local de De Moivre e Laplace", e costuma ser enunciado da seguinte maneira.

**Teorema limite local de De Moivre e Laplace.** *Seja $S_n$ o número de sucessos em $n$ provas de Bernoulli, onde $p\in]0,1[$ é a probabilidade de sucesso em cada prova, e $q = 1-p$. Então*[4]

$$\mathbf{P}\left(\frac{S_n - np}{\sqrt{npq}} = x\right) \sim \frac{1}{\sqrt{2\pi npq}}e^{-x^2/2}$$

*quando $n\to\infty$, uniformemente para valores admissíveis de $x$ (i.e. tais que $np + x\cdot\sqrt{npq}$ seja um inteiro entre 0 e $n$) tais que $|x| = o\left(n^{1/6}\right)$.*

**Teorema integral de De Moivre e Laplace.** O teorema limite local tem, em si, um interesse limitado. É importante porque permite provar o resultado seguinte, o "teorema integral de De Moivre e Laplace", que é um caso particular do moderno "teorema do limite central".

**Teorema integral de De Moivre e Laplace.** *Seja $S_n$ o número de sucessos em $n$ provas de Bernoulli, onde $p\in]0,1[$ é a probabilidade de sucesso em cada prova, e $q = 1-p$. Então*

$$\mathbf{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \le b\right) \to \int_a^b \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

*quando $n\to\infty$, uniformemente em $-\infty \le a < b \le \infty$.*

**dem.** Os valores admissíveis de $S_n$ são inteiros, e a cada valor $k$ corresponde um valor $x_k = (k-np)/\sqrt{npq}$ da variável $S_n^*$. O teorema limite local diz que

$$\mathbf{P}\left(\frac{S_n - np}{\sqrt{npq}} = x_k\right) = \frac{1}{\sqrt{2\pi}}e^{-x_k^2/2}\cdot(x_{k+1}-x_k)(1+\delta)$$

[4] A notação "$f(x)\sim g(x)$ quando $x\to\infty$", cujo significado intuitivo é "a função $f$ é assimptótica à função $g$ quando $x$ é grande", quer dizer que $\lim_{x\to\infty} f(x)/g(x) = 1$.

e fornece uma estimação para o erro $\delta = \alpha \cdot \beta \cdot \gamma - 1$. De facto $\delta$, que depende de $n$ e de $x_k$, é tal que $\delta \to 0$ uniformemente quando $n \to \infty$ e $x_k$ varia num intervalo limitado. Nos queremos é estimar

$$\mathbf{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \leq b\right) = \sum_{a<x_k\leq b} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \cdot (x_{k+1} - x_k) + \sum_{a<x_k\leq b} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \cdot (x_{k+1} - x_k) \cdot \delta$$

A primeira soma no termo à direita converge para o integral de Riemann

$$\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

quando $n \to \infty$, uniformemente para valores de $a$ e $b$ dentro dum intervalo limitado. Sabendo que

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

é fácil ver que a segunda soma converge para 0 quando $n \to \infty$. De facto, o mesmo acontece quando $x_k$ varia num intervalo do genero $]-\infty, b]$, pois o integral de $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ é arbitrariamente pequeno fora dum intervalo limitado suficientemente grande. □

**Aproximação normal.** A função $x \mapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ é chamada *gaussiana*. Os integrais definidos da função gaussiana não admitem expressões em termos de funções "simples". É por isso que os valores aproximados da sua primitiva, a função $\Phi : \mathbf{R} \to [0, 1]$ definida por

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

uma vez calculados numericamente, costumam ser reproduzidos em tabelas nos livros de probabilidade e estatística.

A função $x \mapsto \Phi(x)$ é uma função de repartição, pois tem valores em $[0, 1]$, é contínua, e satisfaz $\Phi(-\infty) = 0$ e $\Phi(\infty) = 1$. Uma variável aleatória cuja função de repartição é $\Phi$ é dita *gaussiana*, ou *normal*. O teorema do limite central então fornece a aproximação

$$\mathbf{P}(np + a\sqrt{npq} < S_n \leq np + b\sqrt{npq}) \simeq \Phi(b) - \Phi(a)$$

ou, de maneira equivalente,

$$\mathbf{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \leq b\right) \simeq \Phi(b) - \Phi(a)$$

válida quando $n$ é grande. Numa linguagem sugestiva: "a lei de $S_n^*$ é assimptótica à lei de uma variável normal".

Para ter uma ideia da previsão quantitativa que o teorema permite fazer, é bom saber alguns valores do integral definido da gaussiana, como

$$\int_{-1.64}^{1.64} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \simeq 0.90 \qquad \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \simeq 0.95 \qquad \int_{-2.58}^{2.58} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \simeq 0.99$$

$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \simeq 0.683 \qquad \int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \simeq 0.954 \qquad \int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \simeq 0.997$$

Por exemplo,

$$\mathbf{P}(|S_n - np| \leq 2\sqrt{npq}) \geq 0.95 \qquad \mathbf{P}(|S_n - np| \leq 3\sqrt{npq}) \geq 0.99$$

ou, em termo da frequência,

$$\mathbf{P}\left(\left|\frac{S_n}{n} - p\right| \leq 2 \cdot \sqrt{\frac{pq}{n}}\right) \geq 0.95 \qquad \mathbf{P}\left(\left|\frac{S_n}{n} - p\right| \leq 3 \cdot \sqrt{\frac{pq}{n}}\right) \geq 0.99$$

**Velocidade da convergência.** É importante ter uma ideia da velocidade da convergência no teorema integral de De Moivre e Laplace, que de facto é "lenta". Uma análise mais detalhada da demonstração mostra que o erro

$$\text{erro}_n = \sup_{-\infty<x<\infty} \left| \mathbf{P}\left\{S_n^* \leq x\right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \right|$$

na aproximação normal é da ordem de $1/\sqrt{npq}$. O resultado optimal é a *desigualdade de Berry e Esseen*, que neste caso particular das provas de Bernoulli assume a forma

$$\text{erro}_n \leq \frac{p^2+q^2}{\sqrt{npq}}$$

Se $p$ é pequeno, ou muito perto de um, este número pode ser grande, a não ser que $n \gg 1/pq$. De facto, neste caso a densidade de $S_n$ é fortemente assimétrica, e a aproximação de Poisson fornece uma estimação melhor da lei de $S_n$.

**Eventos típicos e eventos estáveis.** É interessante observar que a desigualdade de Chebyshev fornece uma estimação

$$\mathbf{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\sqrt{pq}\right) \leq \frac{1}{\varepsilon^2 n}$$

muito fraca, embora suficiente para provar a lei dos grandes números. Se $n$ é grande, o teorema integral de De Moivre e Laplace diz mais, pois[5]

$$\begin{aligned} \mathbf{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\sqrt{pq}\right) &= \mathbf{P}\left(\left|\frac{S_n - np}{\sqrt{npq}}\right| \geq \varepsilon\sqrt{n}\right) \\ &\simeq 2 \cdot \int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ &\leq \frac{1}{\varepsilon\sqrt{n}\sqrt{2\pi}} e^{-n\varepsilon^2/2} \end{aligned}$$

e portanto, a probabilidade dos eventos que a lei dos grandes números considera "não típicos", ou seja desprezáveis, decresce para zero exponencialmente em $n$. Isto implica que os eventos "típicos", aqueles tais que $|S_n/n - p| < \varepsilon$, têm probabilidade que converge muito rapidamente para um quando $n \to \infty$. Em particular, o teorema integral de De Moivre e Laplace implica a lei dos grandes números.

O conteúdo qualitativo do teorema integral de De Moivre e Laplace é que, se $n$ é grande,

$$\mathbf{P}\left(\left|\frac{S_n}{n} - p\right| \leq \varepsilon\sqrt{\frac{pq}{n}}\right) \simeq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

e este número é $\mathcal{O}(1)$, não depende de $n$. Ou seja, eventos com probabilidade "assimptoticamente estável" (e que portanto pode ser arbitrariamente grande ou pequena, dependendo do valor de $\varepsilon$) são tais que a desvio da frequência é da ordem de $1/\sqrt{n}$. Claro que também os eventos complementares têm probabilidade assimptoticamente estável...

**Eventos típicos e eventos estáveis da marcha aleatória.** Seja $(T_n)_{n\in\mathbf{N}}$ a trajectória de uma marcha aleatória simétrica, i.e. $T_n = \xi_1 + \xi_2 + ... + \xi_n$ onde as variáveis $\xi_k$ são independentes e identicamente distribuídas com valores $\pm 1$ e lei determinada por $\mathbf{P}(\xi_k = 1) = 1/2$.

Os eventos que a lei dos grandes números considera "típicos" são os eventos com $|T_n| < \varepsilon n$, onde $\varepsilon > 0$ é arbitrário, cuja probablidade é assimptoticamente igual a um. Os eventos complementares têm probabilidade exponencialmente pequena, pois

$$\begin{aligned} \mathbf{P}(|T_n| \geq \varepsilon n) &\simeq 2 \cdot \int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ &\leq \frac{1}{\varepsilon\sqrt{n}\sqrt{2\pi}} e^{-n\varepsilon^2/2} \end{aligned}$$

[5] Observe que, se $x > 0$,

$$\int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt < \frac{1}{x} \cdot \int_x^{\infty} \frac{t}{\sqrt{2\pi}} e^{-t^2/2} dt = \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}$$

(basta observar que $T_n$ é igual a $2 \cdot S_n - n$, onde $S_n$ é o número de sucessos em $n$ provas de Bernoulli com $p = 1/2$).

Os eventos "estáveis" são os eventos tais que $|T_n| \leq \varepsilon\sqrt{n}$, onde $\varepsilon > 0$ é arbitrário, pois

$$\mathbf{P}\left(|T_n| \leq \varepsilon\sqrt{n}\right) \simeq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

quando $n$ é grande.

**Exercício.** Seja $S_n$ o número de "caras" obtidas lançando $n$ vezes uma moeda honesta.

Estime a probabilidade de obter um número de caras ìgual ao número de coroas, quando $n$ é grande e par, utilizando a fórmula de Stirling.

Estime, utilizando o teorema integral de De Moivre e Laplace, a probabilidade de obter um número de caras que difere do número de coroas por menos de $K$, quando $K$ é um número positivo arbitrário e $n$ é grande. Que acontece quando $n \to \infty$?

Estime, utilizando o teorema integral de De Moivre e Laplace, a probabilidade de obter um número de caras que difere do número de coroas por menos de $K\sqrt{n}$, quando $K$ é um número positivo arbitrário e $n$ é grande. Que acontece quando $n \to \infty$?

Utilizando o teorema integral de De Moivre e Laplace, determine intervalos $[a, b]$ tais que

$$\mathbf{P}\left(a \leq S_n \leq b\right) \geq 90\% \text{ ou } 95\% \text{ ou } 99\%$$

quando $n$ é grande. Determine os intervalos correspondentes para a frequência $f_n = S_n/n$.

Quantos lançamentos de uma moeda honesta é preciso fazer para observar uma frequência $f_n = S_n/n$ tal que

$$|f_n - 1/2| \leq \varepsilon$$

com probabilidade $\geq 90\%$? E $\geq 99\%$? Deduza valores numéricos quando $\varepsilon = 0.1$ ou $0.01$ ou $0.001$.

Responda às mesmas perguntas (oportunamente modificadas, se necessário) no caso em que a probabilidade de sair cara é $p$.

**Mais um sermão: oscilações da marcha aleatória.** Seja $(T_n)_{n \in \mathbf{N}}$ a trajectória de uma marcha aleatória simétrica. A lei dos grandes números diz que as trajectórias mais prováveis são aquelas que estão entre as retas $\pm n\varepsilon$, com $\varepsilon > 0$ arbitrário, desde que $n$ seja suficientemente grande.

A variância de $T_n$ é igual a $n$, portanto é razoável esperar valores de $T_n$ da ordem de $\sqrt{n}$ com probabilidade grande. Uma ideia das oscilações típicas é dada pelos seguintes resultados (cuja demonstração não é elementar, e utiliza o lema de Borel-Cantelli assim como o teorema do limite central):

$$\mathbf{P}\left(\overline{\lim} \frac{T_n}{\sqrt{n}} = \infty\right) = \mathbf{P}\left(\underline{\lim} \frac{T_n}{\sqrt{n}} = -\infty\right) = 1$$

e

$$\mathbf{P}\left(\lim \frac{T_n}{\sqrt{n}\log n} = 0\right) = 1$$

O significado é: com probabilidade um, as trajectórias da marcha aleatória "intersectam" uma infinidade de vezes as curvas $\pm\alpha\sqrt{n}$ e deixam só uma quantidade finita de vezes as regiões limitadas pelas curvas $\pm\alpha\sqrt{n}\log n$, onde $\alpha$ é um número positivo arbitrário. Em particular, com probabilidade um, as trajectorias da marcha aleatória simétrica passam uma infinidade de vezes pelo valor $T_n = 0$. Ou seja, a esmagadora maioria das trajectórias oscilam à volta de 0 e a amplitude das oscilações cresce pelo menos como a raiz de $n$.

Uma ideia mais precisa acerca das oscilações é fornecida pela *lei do logaritmo iterado* (Khinchin, 1924), que diz que

$$\mathbf{P}\left(\overline{\lim} \frac{T_n}{\sqrt{2n\log\log n}} = 1\right) = 1$$

Uma pergunta natural é estimar a proporção de tempo que as trajectórias passam na região $T_n > 0$, i.e. estimar a lei da variável $\eta_n = \tau_n/n$ onde $\tau_n = |\{i = 1, 2, ..., n \text{ t.q. } T_i > 0\}|$. O resultado surpreendente é a *lei do* arcsin (Paul Lévy, 1939), que diz que

$$\mathbf{P}\left(\eta_n \leq x\right) \to \frac{2}{\pi} \arcsin\sqrt{x}$$

quando $n \to \infty$. O gráfico da função $\arcsin\sqrt{x}$ mostra que é mais provável que $\eta_n$ esteja perto de 0 ou de 1, do que perto de 1/2! Claro que, a "surpresa" só mostra que o nosso senso comum não foi treinado para lidar com sequências aleatórias muito compridas...

**Estimação da probabilidade de sucesso nas provas de Bernoulli.** O primeiro problema da estatística das provas de Bernoulli é: observados $k$ sucessos em $n$ provas, i.e. uma sequência $\omega = (\omega_1, \omega_2, ..., \omega_n)$ de 0's e 1's tal que $S_n(\omega) = \omega_1 + \omega_2 + ... + \omega_n = k$, estimar a probabilidade de sucesso $p$ e fazer uma afirmação quantitativa acerca da "confiança" da estimação. Isto é um típico problema de física: temos um modelo, o espaço de probabilidades das provas de Bernoulli, e queremos estimar um dos seu parâmetro, neste caso a probabilidade $p$, fazendo umas experiências.

A lei dos grandes números sugere que uma primeira estimação de $p$ seja a frequência observada $f_n(\omega) = S_n(\omega)/n$, e portanto $k/n$. Fixado um "nível de confiança" $\alpha$, por exemplo 0.95 ou 0.99, procuramos um valor de $\varepsilon > 0$ tal que

$$\mathbf{P}\left(|f_n - p| \leq \varepsilon \cdot \sqrt{\frac{pq}{n}}\right) \geq \alpha$$

De facto, se $n$ é grande, o teorema integral de De Moivre e Laplace diz que

$$\mathbf{P}\left(|f_n - p| \leq \varepsilon \cdot \sqrt{\frac{pq}{n}}\right) \simeq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

e portanto determina o $\varepsilon$ correspondente ao nível de confiança: qualquer $\varepsilon$ superior à raiz da equação $\Phi(t) - \Phi(-t) = \alpha$. Esta aproximação é razoável se sabemos "a priori" que $p$ não é nem muito grande nem muito pequena, i.e. se $\min\{p, q\} \geq \delta > 0$, pois neste caso o erro cometido na aproximação normal é $\leq 1/\delta\sqrt{n}$. Desprezando este erro, podemos afirmar que, com probabilidade $\geq \alpha$, o parâmetro $p$ é tal que

$$p - \varepsilon \cdot \sqrt{\frac{p(1-p)}{n}} \leq f_n \leq p + \varepsilon \cdot \sqrt{\frac{p(1-p)}{n}}$$

e portanto $p_- \leq p \leq p_+$ onde $p_\pm$ são as duas raizes da equação

$$f_n^2 - 2pf_n + p^2 - \varepsilon^2 \frac{p(1-p)}{n} = 0$$

Iterando as desigualdades, e desprezando os termos $\mathcal{O}\left(1/n^{3/4}\right)$, uma resposta é o "intervalo de confiança"

$$f_n - \varepsilon \cdot \sqrt{\frac{f_n(1-f_n)}{n}} \leq p \leq f_n + \varepsilon \cdot \sqrt{\frac{f_n(1-f_n)}{n}}$$

Para ter uma ideia quantitativa da estimação, as tabelas dizem que $\varepsilon \simeq 2$ para um intervalo de confiança com nível $\alpha \geq 95\%$, e $\varepsilon \simeq 3$ para um intervalo de confiança com nível $\alpha \geq 99\%$. Portanto, a afirmação física será, por exemplo,

$$p = f_n \pm 2 \cdot \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}} \qquad \text{com probabilidade } \geq 95\%$$

Duas observações importantes. A primeira é que "o nível de confiança não é confiável", só é determinado com um erro da órdem de $1/\sqrt{n}$ (é por isto que, desde que $n$ não seja muito grande, algo como $n \gg 10^4$, não faz muito sentido querer utilizar o valor verdadeiro $\varepsilon = 1.96$ em vez de $\varepsilon = 2$ num intervalo de nível 95%). A segunda é que também "a amplitude do intervalo não é confiável", sendo uma aproximação do verdadeiro valor $|p_+ - p_-|$ com um erro da ordem de $1/n^{3/4}$. Um físico só pode acreditar numa afirmação do género "o verdadeiro valor de $p$ é igual a $f_n$ mais ou menos um multiplo pequeno de $\sqrt{f_n(1-f_n)}/\sqrt{n}$ com probabilidade muito grande".

# 13 Grandes desvios e entropia

**Grandes desvios.** Sejam $\xi_1, \xi_2, ...$ variáveis independentes e identicamente distribuídas, com média $\mathbf{E}\xi = m$, e seja $S_n = \xi_1 + \xi_2 + ... + \xi_n$. Se a lei dos grandes números é aplicável, dado $\varepsilon > 0$, o evento com $|S_n/n - m| < \varepsilon$ tem probabilidade próxima de 1 quando $n$ é grande, i.e. é o "evento típico". O evento complementar $|S_n/n - m| \geq \varepsilon$ é "desprezável", i.e. tem probabilidade assimptótica a 0 quando $n$ cresce. A teoria dos grandes desvios (*large deviations* em inglês) trata do problema de estimar a velocidade optimal com que a probabilidade deste evento decresce para zero quando $n \to \infty$. A relevância física do problema é evidente: da velocidade de convergência depende a possibilidade prática de "observar" a lei dos grandes números.

A desigualdade de Chebyshev fornece a estimação

$$\mathbf{P}\left(\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right) \leq \text{const.}/n$$

se a variância de $\xi$ é finita. Se a variável tem momentos finitos de todos os graus, então a probabilidade acima é $\leq \text{const.}/n^k$ para todos $k > 1$, ou seja, decresce mais rapidamente do que o inverso de qualquer polinómio.

É possivel dizer mais se, por exemplo, a variável $\xi$ é limitada, ou mais em geral se $\mathbf{E}e^{\beta|\xi|} < \infty$ para algum $\beta > 0$, e portanto $\beta \mapsto \mathbf{E}e^{\beta\xi}$ é uma função analítica de $\beta$ num aberto $B \subset \mathbf{R}$ que contém a origem. A *energia livre*[6] do sistema é a função $\beta \mapsto F(\beta)$, definida por

$$F(\beta) = \log \mathbf{E}e^{\beta\xi}$$

se $\beta \in B$ e por $F(\beta) = \infty$ fora desta região. Se as $\xi_k$ têm variância positiva, i.e. não são triviais, então um cálculo mostra que $F$ é estritamente convexa nesta região. De facto,

$$\begin{aligned} F'(\beta) &= e^{-F(\beta)}\mathbf{E}\left(\xi e^{\beta\xi}\right) \\ F''(\beta) &= e^{-F(\beta)}\left(e^{-F(\beta)}\mathbf{E}\left(\xi^2 e^{\beta\xi}\right) - \left(\mathbf{E}\left(\xi e^{\beta\xi}\right)\right)^2\right) \end{aligned}$$

A desigualdade de Cauchy-Schwarz, aplicada as variáveis $\xi e^{\beta\xi/2}$ e $e^{\beta\xi/2}$, implica que $F''(\beta) > 0$, a não ser que $\xi e^{\beta\xi/2}$ seja proporcional a $e^{\beta\xi/2}$ com probabilidade um, o que é equivalente a dizer que $\xi$ é constante com probabilidade um. Também temos que $F(0) = 0$ e $F'(0) = m$.

Utilizando a desigualdade de Chebyshev exponencial vê-se que

$$\begin{aligned} \mathbf{P}\left(\frac{S_n}{n} \geq \lambda\right) &= \mathbf{P}\left(e^{\beta S_n} \geq e^{n\beta\lambda}\right) \\ &\leq e^{-n\beta\lambda} \cdot \mathbf{E}e^{\beta S_n} \\ &\leq e^{-n\beta\lambda} \cdot \left(\mathbf{E}e^{\beta\xi}\right)^n \\ &\leq e^{-n\beta\lambda} \cdot e^{n \log \mathbf{E}e^{\beta\xi}} \end{aligned}$$

para todo $\beta > 0$, e portanto

$$\begin{aligned} \mathbf{P}\left(\frac{S_n}{n} \geq \lambda\right) &\leq \inf_{\beta>0} e^{-n(\beta\lambda - F(\beta))} \\ &\leq e^{-n \sup_{\beta>0}(\beta\lambda - F(\beta))} \end{aligned}$$

A *entropia*[7] é a transformada de Legendre da energia livre, a função $\lambda \mapsto H(\lambda)$ definida por

$$H(\lambda) = \sup_{\beta}(\beta\lambda - F(\beta))$$

Como $F$ é estritamente convexa e satisfaz $F(0) = 0$ e $F'(0) = m$, a entropia é também estritamente convexa e tem um mínimo em $\lambda = m$, onde $H(m) = 0$. Em particular, se $\lambda > m$ então

$$H(\lambda) = \sup_{\beta>0}(\beta\lambda - F(\beta)) > 0$$

[6] A linguagem vem da mecânica estatística, a menos de um factor dimensional $\beta^{-1}$ proporcional à temperatura, se consideramos as $\xi_k$ como es energias das partículas de um sistema termodinâmico em equilibrio.

[7] Os probabilistas também chamam $H$ a *transformada de Cramér* da função de repartição de $\xi$.

Logo, se $\varepsilon > 0$,

$$\mathbf{P}\left(\frac{S_n}{n} - m \geq \varepsilon\right) \leq e^{-nH(m+\varepsilon)}$$

onde $H(m+\varepsilon) > 0$. Repetindo um argumento análogo com $\leq m-\varepsilon$ em vez de $\geq m+\varepsilon$ e juntando as desigualdades, o resultado é que

$$\mathbf{P}\left(\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right) \leq 2e^{-n\min\{H(m+\varepsilon),H(m-\varepsilon)\}}$$

onde também $H(m-\varepsilon) > 0$. Ou seja, a probabilidade dos "grandes desvios" decresce para 0 exponencialmente em $n$, e a velocidade da convergência é determinada pela entropia.

**Grandes desvios nas provas de Bernoulli.** Sejam $\xi_1, \xi_2, ...$ variáveis independentes e identicamente distribuídas com lei de Bernoulli $B(1,p)$, e seja $S_n = \xi_1 + \xi_2 + ... + \xi_n$ o número de sucesso em $n$ provas. A energia livre é

$$F(\beta) = \log\left(e^{\beta}p + q\right)$$

Um cálculo mostra que, se $p < p' < 1$ e $q' = 1 - p'$, a entropia é dada por

$$\begin{aligned} H(p') &= \sup_{\beta>0}\left(\beta p' - F(\beta)\right) \\ &= p'\log\left(p'/p\right) + q'\log\left(q'/q\right) \end{aligned}$$

a mesma função encontrada na prova do teorema local de De Moivre e Laplace. A desigualdade elementar $H(p') \geq 2(p'-p)^2$, e um argumento análogo com $0 < p' < p$, fornece enfim a estimação

$$\mathbf{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

para a probabilidade dos grandes desvios nas provas de Bernoulli. Observe que este resultado é assimptoticamente equivalente à estimação obtida utilizando o teorema limite integral de De Moivre e Laplace.

**Entropia.** Uma *experiência* é uma partição finita $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ do espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$. A *entropia* de $\mathcal{D}$ é definida por

$$h(\mathcal{D}) = -\sum_i p_i \log p_i$$

onde $p_i = \mathbf{P}(D_i)$ e decidimos que $0 \cdot \log 0 = 0$.

A concavidade da função $x \mapsto -x\log x$ implica que a entropia satisfaz

$$h(\mathcal{D}) \leq \log n$$

onde $n$ é a cardinalidade da partição, e que a igualdade acontece sse $p_i = 1/n$ para todo $i$. Ou seja, a entropia é máxima sse a medida de probabilidade em $\mathcal{D}$ é uniforme.

Dadas duas expriências $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ e $\mathcal{C} = \{C_1, C_2, ..., C_m\}$, a experiência produto $\mathcal{D} \vee \mathcal{C}$ é a partição $\{E_{1,1}, E_{1,2}, ..., E_{n,m}\}$ onde $E_{i,j} = D_i \cap C_j$. A entropia da experiência produto satisfaz

$$h(\mathcal{D} \vee \mathcal{C}) = h(\mathcal{D}) + h(\mathcal{C}|\mathcal{D})$$

se a entropia de $\mathcal{C}$ dado $\mathcal{D}$ é definida por

$$h(\mathcal{C}|\mathcal{D}) = -\sum_{i,j} p_i q_{ij} \log q_{ij}$$

onde $q_{ij} = \mathbf{P}(C_j\,|D_i\,)$. É fácil ver que

$$h(\mathcal{D} \vee \mathcal{C}) \leq h(\mathcal{D}) + h(\mathcal{C})$$

Se as experiências são independentes, i.e. se $\mathbf{P}(D_i \cap C_j) = \mathbf{P}(D_i) \cdot \mathbf{P}(C_j)$ para todos $i$ e $j$, então

$$h(\mathcal{D} \vee \mathcal{C}) = h(\mathcal{D}) + h(\mathcal{C})$$

A entropia é uma medida da incerteza da experiência, ou seja da informação que contêm os resultados da experiência. Assume o máximo precisamente quando a experiência é a mais incerta, ou seja quando todos os átomos da partição têm a mesma probabilidade. Assume o mínimo, o valor zero, quando um dos átomos a partição tem probabilidade um, ou seja quando a experiência não é nada "aleatória". A definição é tal que a informação que vem de duas experiências independentes é a soma das duas informações. Saber o resultado de uma experiência, diminui a incerteza sobre uma outra experiência, a não ser que elas sejam independentes.

**Entropia e palavras típicas.** Seja $\Omega = \{x_1, x_2, ..., x_n\}$, munido da probabilidade $\mathbf{P}'$ definida por $\mathbf{P}'(\{x_k\}) = p_k$, onde os $p_k$ são números não negativos tais que $\sum_k p_k = 1$. Seja

$$\Omega^N = \{\omega = (\omega_1, \omega_2, ..., \omega_n) \text{ com } \omega_k \in \Omega\}$$

o espaço das palavras de comprimento $N$ nas letras $x_1, x_2, ..., x_n$, munido da probabilidade produto $\mathbf{P}$, i.e. o espaço das $N$ provas independentes de uma experiência com $n$ resultados possíveis. Seja

$$h = h(\mathcal{D}) = -\sum_i p_i \log p_i$$

onde $\mathcal{D}$ é a partição $\{\{x_1\}, \{x_2\}, ..., \{x_n\}\}$ de $\Omega$. A lei dos grandes números diz que, fixado $\delta' > 0$, o conjunto das palavras que contêm cada letra $x_i$ um número de vezes $S_{N,i}$ tal que

$$\left| \frac{S_{N,i}}{N} - p_i \right| \leq \delta'$$

tem probabilidade perto de um quando $n$ é grande. É natural chamar "típicas" tais palavras. A entropia $h$ fornece uma medida do "tamanho" das palavras típicas.

**Teorema de McMillan.** *Para todos $\varepsilon > 0$ e $\delta > 0$ existe $\overline{N}$ tal que para todo $N > \overline{N}$ existe um conjunto de palavras típicas $\Omega^N_{\text{tip}} \subset \Omega^N$ que satisfaz:*

*i) se $\omega \in \Omega^N_{\text{tip}}$ então a sua probabilidade é $e^{-N(h+\delta)} < \mathbf{P}\{\omega\} < e^{-N(h-\delta)}$*

*ii) a probabilidade das palavras típicas é $\mathbf{P}(\Omega^N_{\text{tip}}) > 1 - \varepsilon$.*

**dem.** Dada uma palavra $\omega \in \Omega^N$, seja $S_{N,i}$ o número de letras $x_i$ contidas em $\omega$. Pela lei dos grandes números (ou seja, pela desigualdade de Chebyshev), para todos $\varepsilon' > 0$ e $\delta' > 0$ existe $\overline{N}$ tal que para todo $N > \overline{N}$

$$\mathbf{P}\left\{ \left| \frac{S_{N,i}}{N} - p_i \right| > \delta' \right\} < \varepsilon'$$

Seja

$$\Omega^N_{\text{tip}} = \left\{ \omega \in \Omega^N \text{ t.q. } \left| \frac{S_{N,i}}{N} - p_i \right| \leq \delta' \right\}$$

Então, se $\omega \in \Omega^N_{\text{tip}}$ a sua probabilidade $\mathbf{P}\{\omega\} = p_1^{S_{N,1}} \cdot p_2^{S_{N,2}} \cdot ... \cdot p_n^{S_{N,n}}$ satisfaz

$$\log \mathbf{P}\{\omega\} = \sum_i f_i \log p_i = -Nh + \sum_i (S_{N,i} - Np_i) \log p_i$$

e portanto

$$\left| \frac{\log(1/\mathbf{P}\{\omega\})}{N} - h \right| < \delta$$

se $\delta' < \delta N / |\sum_i \log p_i|$. Por outro lado

$$\mathbf{P}(\Omega^N \setminus \Omega^N_{\text{tip}}) \leq \sum_i \mathbf{P}\left\{ \left| \frac{S_{N,i}}{N} - p_i \right| > \delta' \right\} < n\varepsilon'$$

e portanto $\mathbf{P}(\Omega^N_{\text{tip}}) > 1 - \varepsilon$ se $\varepsilon' < \varepsilon/n$. □

Corolário deste teorema é uma caracterização "dinâmica" da entropia. Seja $\rho$ um número arbitrário no intervalo $(0,1)$ e $E_\rho(N)$ a menor cardinalidade de um conjunto de palavras de $\Omega^N$ cuja probabilidade é $> \rho$. Então

$$h = \lim_{N\to\infty} \frac{\log E_\rho(N)}{N}$$

Pois, se $N$ é suficientemente grande, para chegar a ter probabilidade $\rho$ temos que usar pouco menos do que um número $K$ de palavras típicas tal que

$$Ke^{-Nh} \sim \rho$$

logo $E_\rho(N) \sim \rho e^{Nh}$.

**Codificação.** O teorema de McMillan diz que $h$ é o logaritmo do número $m$ de letras suficientes para codificar as palavras típicas (ou seja um conjunto de palavras com probabilidade muito perto de um) usando palavras do mesmo comprimento $N$, pois a cardinalidade das palavras típicas é $\sim e^{Nh}$. Também podemos mudar o ponto de vista, e codificar a mensagem com palavras de comprimento menor no mesmo alfabeto. Então o comprimento pode ser da ordem de $M$ se

$$n^M \sim e^{Nh}$$

Ou seja, podemos reduzir o comprimento da mensagem de um *factor de compressão*

$$\frac{M}{N} \sim \frac{h}{\log n}$$

(observe que $\log n$ é o máximo valor possível da entropia), dito *entropia relativa*.

Modelos menos ingénuos de mensagens devem incluir uma "gramática", um conjunto de regras que dizem que letras $\omega_k$ são permitidas depois das letras $\omega_1\omega_2...\omega_{k-1}$. Modelos razoáveis são as cadeias de Markov.

# 14 Modelos contínuos

**Variáveis aleatórias absolutamente contínuas.** Seja $(\Omega, \mathcal{E}, \mathbf{P})$ um espaço de probabilidades. A variável aleatória $\xi : \Omega \to \mathbf{R}$ é *absolutamente contínua* se existe uma função "integrável" $f_\xi : \mathbf{R} \to \mathbf{R}_{\geq 0}$, dita *densidade* de $\xi$, tal que

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(t)\, dt$$

para todo $x \in \mathbf{R}$.

Se $\xi$ é absolutamente contínua então $F_\xi$ é contínua, e portanto $\mathbf{P}(\xi = x) = 0$ para todo $x$. Em particular, se $-\infty \leq a < b \leq \infty$,

$$\mathbf{P}(a \leq \xi \leq b) = F_\xi(b) - F_\xi(a) = \int_a^b f_\xi(x)dx$$

A variável aleatória $\xi : \Omega \to \mathbf{R}^n$ é *absolutamente contínua* se existe uma função integrável $f_\xi : \mathbf{R}^n \to \mathbf{R}_{\geq 0}$, dita *densidade de probabilidade*, tal que para todo boreleano $A \subset \mathbf{R}^n$

$$\mathbf{P}(\xi \in A) = \int_A f_\xi(x)\, dx$$

onde $dx = dx_1 dx_2 ... dx_n$ denota a medida de Lebesgue em $\mathbf{R}^n$.

**Integrabilidade.** Acima, a palavra "integrável" tem um significado técnico preciso: "integrável com respeito à medida de Lebesgue". A condição $\mathbf{P}(\xi \in A) = \int_A f_\xi(x)\, dx$ diz que "a lei de $\xi$ é uma medida absolutamente contínua com respeito à medida de Lebesgue". Para as aplicações elementares que temos em mente, basta pensar no caso particular de uma função $f_\xi$ que é contínua em todos os pontos da recta (ou do espaço euclidiano) exceto, possivelmente, num conjunto enumerável de pontos, ou seja numa função integrável no sentido do integral de Riemann. A seguir, o símbolo $\int_a^b f(x)dx$ será sinónimo de "o integral de Riemann, eventualmente impróprio, da função $f$ no intervalo $]a, b[$".

A densidade de uma variável absolutamente contínua não é única, duas densidades podem diferir em um conjunto de medida de Lebesgue nula (por exemplo num subconjunto enumerável da recta).

Se $F_\xi$ é uma função diferenciável, ou pelo menos diferenciável por pedaços, uma densidade de $\xi$ é a derivada de $F_\xi$ (teorema fundamental do cálculo).

**Exemplos esquisitos.** Exemplos de variáveis aleatórias que não são contínuas são as discretas. Por exemplo, uma variável constante não é absolutamente contínua (a sua lei é um delta de Dirac). Também existem variáveis que não são nem discretas nem absolutamente contínuas... Um exemplo famoso é uma variável que é igual a um no conjunto de Cantor $K = \left\{\sum_{k=1}^{\infty} \frac{x_k}{3^k} \text{ com } x_k \in \{0, 2\}\right\} \subset [0, 1]$ com probabilidade um. A sua função de repartição é a "escada do diabo", uma função crescente que tem derivada nula fora dum conjunto de medida de Lebesgue nula!

**Construção de variáveis absolutamente contínuas.** Uma densidade de uma variável aleatória é uma função $f$ integrável, não negativa e tal que

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Uma função com estas propriedades define uma variável aleatória absolutamente contínua. Pois podemos pôr $\Omega = \mathbf{R}$, $\mathcal{E} = \mathcal{B}(\mathbf{R})$, $\mathbf{P}(A) = \int_A f(x)dx$ para todo $A \in \mathcal{B}(\mathbf{R})$, e definir $\xi : \Omega \to \mathbf{R}$ como $\xi(x) = x$. Mais uma vez, toda a informação relevante acerca de uma variável aleatória absolutamente contínua está contida na sua densidade.

**Valor médio e variância.** A variável aleatória absolutamente contínua $\xi : \Omega \to \mathbf{R}$ com densidade $f_\xi$ é dita *integrável* se

$$\int_{-\infty}^{\infty} |x| \cdot f_\xi(x)dx < \infty$$

O *valor médio* (ou *esperança*, ou *média*) da variável aleatória absolutamente contínua e integrável $\xi$ é

$$\mathbf{E}\xi = \int_{-\infty}^{\infty} x \cdot f_\xi(x)dx$$

As propriedades do valor médio de variáveis aleatórias absolutamente contínuas são análogas às propriedades das variáveis discretas (aliás, o integral de Lebesgue é definido aproximando as variáveis com variáveis simples!).

Se $\xi$ e $\eta$ são variáveis aleatória absolutamente contínuas e integráveis, e $a, b \in \mathbf{R}$, então

$$\mathbf{E}\left(a\xi + b\eta\right) = a\mathbf{E}\xi + b\mathbf{E}\eta$$

Se $\xi$ e $\eta$ são independentes e integráveis, então $\xi\eta$ é integrável e

$$\mathbf{E}\xi\eta = \mathbf{E}\xi \cdot \mathbf{E}\eta$$

Se $\xi^2$ é integrável, a *variância* da variável aleatória $\xi$ é definida por $\mathbf{V}\xi = \mathbf{E}\left(\xi - \mathbf{E}\xi\right)^2 = \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2$, i.e.

$$\mathbf{V}\xi = \int_{-\infty}^{\infty} (x - m)^2 \cdot f_\xi(x)dx$$

onde $m$ é o valor médio de $\xi$.

A desigualdade de Chebyshev, a desigualdade de Cauchy-Schwarz, assim como a definição e as propriedades da covariância, continuam válidas para as variáveis contínuas.

**Leis uniformes.** A variável aleatória $\xi$ tem *lei uniforme* no intervalo $[a, b]$ da recta real se a sua função de repartição é

$$F_\xi(x) = \begin{cases} 0 & \text{se } x < a \\ \frac{x-a}{b-a} & \text{se } a \leq x \leq b \\ 1 & \text{se } x > b \end{cases}$$

Uma sua densidade é $f_\xi(x) = 1/(b-a)$ se $x \in [a, b]$ e 0 se $x \notin [a, b]$.

Seja $dx$ a medida de Lebesgue em $\mathbf{R}^n$, e $C \subset \mathbf{R}^n$ um domínio suficientemente regular tal que $\mathrm{vol}(C) = \int_C dx < \infty$. A variável aleatória $\xi : \Omega \to \mathbf{R}^n$ tem *lei uniforme* em $C$ se para todo boreleano $A \subset \mathbf{R}^n$

$$\mathbf{P}\left(\xi \in A\right) = \frac{\mathrm{vol}(A \cap C)}{\mathrm{vol}(C)}$$

Ou seja, uma densidade de $\xi$ é $f_\xi(x) = 1/\mathrm{vol}(C)$ se $x \in C$ e 0 se $x \notin C$.

**Mudança de variável.** Se $\xi : \Omega \to \mathbf{R}$ é uma variável aleatória, e $\varphi : \mathbf{R} \to \mathbf{R}$ uma função, a função composta $\eta = \varphi \circ \xi$ pode não ser uma variável aleatória. Acontece que $\eta$ é uma variável aleatória se $\varphi$ é suficientemente regular (se a imagem inversa de todo intervalo aberto é um boreleano), por exemplo se é contínua. Se $\xi$ tem densidade $f_\xi$, então a função de repartição de $\eta$ pode ser calculada por

$$\begin{aligned} F_\eta(x) &= \mathbf{P}\left(\eta \leq x\right) = \mathbf{P}\left(\varphi(\xi) \leq x\right) \\ &= \int_{\varphi^{-1}((-\infty, x])} f_\xi(y)dy \end{aligned}$$

Se $\xi$ é absolutamente contínua e $\varphi$ é um difeomorfismo, então $\eta = \varphi \circ \xi$ é também absolutamente contínua. A lei de $\eta$ é determinada por meio da mudança de variável no integral, pois

$$\mathbf{P}\left(\eta \in A\right) = \int_{\varphi^{-1}(A)} f_\xi(x)dx = \int_A \left|\det \mathrm{Jac}\varphi^{-1}(x)\right| \cdot f_\xi\left(\varphi^{-1}(x)\right) dx$$

e portanto uma densidade de $\eta$ é

$$f_\eta(x) = \left|\det \mathrm{Jac}\varphi^{-1}(x)\right| \cdot f_\xi\left(\varphi^{-1}(x)\right)$$

Também, se $\eta$ é integrável, a sua média pode ser calculada por meio do integral

$$\mathbf{E}(\eta) = \int_{-\infty}^{\infty} \varphi(x)\, f_\xi(x) dx$$

**Exercícios.**

**a.** Se $\xi$ tem densidade $f_\xi$, então $\eta = a\xi + b$, com $a \neq 0$, tem densidade

$$f_\eta(x) = \frac{1}{|a|} f_\xi\left(\frac{x-b}{a}\right)$$

**b.** Se $\xi$ tem densidade $f_\xi$ e função de repartição $F_\xi$, então $\eta = \xi^2$ tem função de repartição

$$F_\eta(t) = F_\xi\left(\sqrt{t}\right) - F_\xi\left(-\sqrt{t}\right)$$

e portanto densidade

$$f_\eta(x) = \frac{1}{2\sqrt{x}}\left(f_\xi\left(\sqrt{x}\right) + f_\xi\left(-\sqrt{x}\right)\right)$$

se $x > 0$ e $f_\eta(x) = 0$ se $x \leq 0$.

**c.** Seja $\xi$ uma variável aleatória com densidade uniforme na bola unitária $B^2(1) = \left\{z \in \mathbf{R}^2 \text{ t.q. } |z| \leq 1\right\}$. Então a variável $\eta : \Omega \to \mathbf{R}^2$ definida por

$$\eta = \xi \cdot \sqrt{\frac{-2\log|\xi|^2}{|\xi|^2}}$$

(se $\xi \neq 0$) tem densidade

$$f_\eta(z) = \frac{1}{2\pi} e^{-|z|^2/2}$$

Em particular, isto prova que $\iint_{\mathbf{R}^2} e^{-|z|^2/2} dz = \left(\int_{-\infty}^{\infty} e^{-x^2/2} dx\right)^2 = 2\pi$.

**d.** *(lei de Cauchy)* Seja $\xi$ uma variável aleatória com lei uniforme no intervalo $\left]\frac{-\pi}{2}, \frac{\pi}{2}\right[$. Mostre que a variável $\eta = \tan\xi$ tem "lei de Cauchy", ou seja tem densidade

$$f_\eta(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$$

Prove que $\eta$ não é integrável. Determine a lei da variável $1/\eta$.

**Lei exponencial.** A variável $\xi$ tem *lei exponencial* se a sua função de repartição é

$$F_\xi(x) = 1 - e^{-x/\tau}$$

se $x \geq 0$ e 0 se $x < 0$, onde $\tau$ é um parâmetro positivo. A lei exponencial é o análogo contínuo da lei da variável tempo de espera (em física é um modelo de um tempo de decaimento de uma substância radioactiva, ou em geral de um tempo de vida). Uma densidade é

$$f_\xi(x) = \frac{1}{\tau} e^{-x/\tau}$$

se $x \geq 0$ e 0 se $x < 0$. A esperança é $\mathbf{E}\xi = \tau$. Uma notação é $\xi \sim \exp(\tau)$. A lei exponencial também tem (e é caracterizada por) a propriedade da falta de memória, ou seja

$$\mathbf{P}\{\xi > x + y \,|\, \xi > y\} = \mathbf{P}\{\xi > x\}$$

para todos $x, y \geq 0$. Um sistema físico cujo tempo de vida tem lei exponencial não tem idade: se viveu até hoje, o seu futuro é igual ao futuro de um sistema recém nascido!

**Uma interpretação da lei exponencial.** Em média, caem $N$ estrelas ao longo dum tempo $T$. Quanto tempo é preciso esperar para ver a primeira estrela cair? Qual a lei do tempo em que cai a primeira estrela?

Sejam $\xi_1, \xi_2, ...\xi_N$ variáveis independentes com lei uniforme no intervalo $[0, T]$, onde cada $\xi_k$ é pensada como o tempo em que cai a estrela $k$-ésima. O tempo em que cai a primeira estrela é a variável $\xi_{\min} = \min\{\xi_1, \xi_2, ...\xi_N\}$. Pela hipótese de independência, dado $0 \leq x \leq T$,

$$\mathbf{P}(\xi_{\min} \geq x) = \prod_{k=1}^{N} \mathbf{P}(\xi_k \geq x) = \left(1 - \frac{x}{T}\right)^N$$

No limite termodinâmico, quando $N \to \infty$ e $T \to \infty$ mantendo constante a "frequência" $1/\tau = N/T$, temos que

$$\mathbf{P}(\xi_{\min} \geq x) \to e^{-x/\tau}$$

i.e. a lei de $\xi_{\min}$ tende para uma lei exponencial com esperança $\tau$.

**Tempos de vida.** Uma máquina é composta por $n$ componentes em série. O tempo de vida de cada componente é suposto ser uma variável aleatória $\xi_k$, com lei exponencial e esperança $\tau_k$. O tempo de vida da máquina é a variável $\xi_{\min} = \min\{\xi_1, \xi_2, ...\xi_n\}$. Se as $\xi_k$ são independentes, dado $t > 0$,

$$\mathbf{P}(\xi_{\min} \geq t) = \prod_{k=1}^{n} \mathbf{P}(\xi_k \geq t) = \prod_{k=1}^{n} e^{-t/\tau_k} = e^{-t/\tau}$$

onde $\tau$ é $n^{-1}$ vezes a média harmónica dos $\tau_k$, i.e.

$$\tau = \left(\sum_{k=1}^{n} \frac{1}{\tau_k}\right)^{-1}$$

Portanto, $\xi_{\min}$ tem lei exponencial e esperança $\tau$. Observem que $\tau < \min\{\tau_1, \tau_2, ...\tau_n\}$, e que, se os $\tau_k$ são todos iguais, então $\tau = \tau_1/n$ e portanto diminui quando $n$ cresce.

Diferente é o caso de uma máquina composta por $n$ componentes em paralelo. O seu tempo de vida é $\xi_{\max} = \max\{\xi_1, \xi_2, ...\xi_n\}$, cuja função de repartição é

$$\mathbf{P}(\xi_{\max} \leq t) = \prod_{k=1}^{n} \mathbf{P}(\xi_k \leq t) = \prod_{k=1}^{n} \left(1 - e^{-t/\tau_k}\right)$$

A média de $\xi_{\max}$ pode ser calculada integrando $\int_0^\infty t \cdot F'_{\xi_{\max}}(t)\, dt$.

**Exercício.** $N$ estrelas estão distribuidas dentro de uma bola de raio $R$ centrada no sol. Assuma que a posição de cada estrela tem lei uniforme na bola e que as posições das diferentes estrelas sejam independentes.

Determine a probabilidade da estrela mais próxima do sol estar à distância $\geq x$, onde $0 \leq x \leq R$.

Determine a mesma probabilidade no limite termodinâmico, quando $N \to \infty$ e $R \to \infty$ mantendo constante a "densidade média"

$$\rho = \frac{N}{\mathrm{vol}\left(B^3(R)\right)}$$

onde $\mathrm{vol}\left(B^3(R)\right)$ é o volume da bola de raio $R$ centrada no sol, e portanto determine a lei da variável que representa a distância entre o sol e a estrela mais próxima.

Estime o valor esperado da distância entre o sol e a estrela mais próxima, sabendo que uma estimação da densidade média da nossa galáxia numa vizinhança do sol é $\rho \simeq 0.0063$ parsec$^{-3}$.

**Lei normal.** A variável aleatória $\xi : \Omega \to \mathbf{R}$ tem *lei normal* $N(0,1)$ (ou *gaussiana*) se uma sua densidade é

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Pela sua importância teórica, a função de repartição de uma variável gaussiana merece um nome, e costuma ser indicada por

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

Se $\xi$ tem lei normal $N(0,1)$, então a variável aleatória $\eta = \sigma\xi + m$, onde $m, \sigma \in \mathbf{R}$ e $\sigma > 0$, tem lei $N(m, \sigma^2)$, ou seja tem densidade

$$f_\eta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2}$$

Uma variável $\eta$ com lei $N(m, \sigma^2)$ tem $\mathbf{E}\eta = m$ e $\mathbf{V}\eta = \sigma^2$.

**Exercícios.**

**a.** Prove que

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1 \qquad \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 0 \qquad \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$$

e deduza que uma variável $\eta$ com lei $N(m, \sigma^2)$ tem $\mathbf{E}\eta = m$ e $\mathbf{V}\eta = \sigma^2$.

**b.** Se $\xi_1, \xi_2, ..., \xi_n$ são independentes e têm leis $N\left(m_1, \sigma_1^2\right)$, $N\left(m_2, \sigma_2^2\right)$, ...,$N\left(m_n, \sigma_n^2\right)$ respectivamente, então a variável $\xi_1 + \xi_2 + ... + \xi_n$ tem lei $N\left(m, \sigma^2\right)$ com $m = m_1 + m_2 + ... + m_n$ e $\sigma^2 = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2$.

**Uma interpretação geométrica da lei normal.** Seja $B_{\sqrt{n}}^{n+1}$ a bola de raio $\sqrt{n}$ centrada na origem de $\mathbf{R}^{n+1}$, seja $\mathbf{P}_n$ a probabilidade uniforme em $B_{\sqrt{n}}^{n+1}$, e seja $\pi : \mathbf{R}^{n+1} \to \mathbf{R}$ a projeção $x = (x_0, x_1, ..., x_n) \mapsto x_0$. A medida imagem $\pi\mathbf{P}_n$ tem suporte $[-\sqrt{n}, \sqrt{n}]$ e densidade

$$\gamma_n \cdot \left(\sqrt{1 - \frac{x_0^2}{n}}\right)^n$$

onde $\gamma_n$ é um fator de normalização. No limite quando $n \to \infty$ a densidade da medida $\pi\mathbf{P}_n$ converge para a gaussiana

$$\frac{1}{\sqrt{2\pi}} e^{-x_0^2/2}$$

Observe que o mesmo fenómeno acontece a partir da medida uniforme (i.e. invariante por rotações) na esfera $S_{\sqrt{n}}^{n+1} = \partial B_{\sqrt{n}}^{n+1}$.

**Distribuição de Maxwell-Boltzmann.** O conjunto de nível $E$ da energia (cinética) de um sistema de $n$ partículas clássicas não interagentes é (no espaço dos momentos) a esfera

$$S_{\sqrt{E}}^{3n-1} = \left\{ x \in \mathbf{R}^{3n} \text{ t.q. } |x|^2 = E \right\}$$

Fixar uma "energia média por partícula", por exemplo 1, e fazer crescer o número de partículas equivale a estudar as esferas $S_{\sqrt{E}}^{3n-1}$ no limite termodinâmico quando $n \to \infty$. Seja $\mathbf{P}_n$ a probabilidade uniforme em $S_{\sqrt{E}}^{3n-1}$, e seja $\pi : \mathbf{R}^{3n} \to \mathbf{R}^3$ a projeção $x = (x_1, ..., x_n) \mapsto x_1$, onde $x_1$ é o momento da primeira partícula. A medida imagem $\pi\mathbf{P}_n$ tem densidade

$$\gamma_n \cdot \left(\sqrt{1 - \frac{|x_1|^2}{n}}\right)^{3n-4}$$

onde agora $|x_1|$ denota a norma euclidiana de $x_1$ em $\mathbf{R}^3$. No limite quando $n \to \infty$, esta densidade converge para

$$\frac{1}{\sqrt{2\pi/3}} e^{-3|x_1|^2/2}$$

que é chamada *distribuição de Maxwell-Boltzmann.*

**Leis gamma.** A variável aleatória $\xi : \Omega \to \mathbf{R}$ tem *lei gamma* $\Gamma(\alpha, \lambda)$ com parâmetros $\alpha > 0$ e $\lambda > 0$ se a sua densidade é

$$f(x) = cx^{\alpha-1} e^{-\lambda x}$$

se $x > 0$ e 0 se $x \leq 0$. O valor da constante $c$ é determinado pela normalização: $c = \lambda^\alpha / \Gamma(\alpha)$ onde a função "Gamma" é definida por

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

(é uma extensão da função "factorial", pois $\Gamma(n+1) = n!$ se $n$ é um inteiro não negativo). Não existem fórmulas explícitas para a função de repartição da lei gamma, a não ser que $\alpha$ seja um inteiro positivo, e nesse caso

$$F_\xi(x) = 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x}$$

Valor médio e variância são $\mathbf{E}\xi = \alpha/\lambda$ e $\mathbf{V}\xi = \alpha/\lambda^2$.

A lei $\Gamma(1, \lambda)$ é a lei exponencial $\exp(1/\lambda)$.

Se $\xi$ tem lei $N(0, \sigma)$ então $\xi^2$ tem lei $\Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$.

Se $\xi_1, \xi_2, ..., \xi_n$ são independentes e têm leis gamma $\Gamma(\alpha_1, \lambda)$, $\Gamma(\alpha_2, \lambda)$, ...,$\Gamma(\alpha_n, \lambda)$ respectivamente, então a variável $\xi_1 + \xi_2 + ... + \xi_n$ tem lei $\Gamma(\alpha, \lambda)$ com $\alpha = \alpha_1 + \alpha_2 + ... + \alpha_n$.

**Uma interpretação da lei de Poisson.** Sejam $\xi_1, \xi_2, ...$ variáveis independentes com lei $\exp(\tau)$, e seja $\eta$ a variável com valores $0, 1, 2, ...$ definida por

$$\eta = \sup\{n \text{ tais que } \xi_1 + \xi_2 + ... + \xi_n \leq t\}$$

onde $t > 0$. Então a função de repartição de $\eta$ é

$$F_\eta(k) = \mathbf{P}\{\eta \leq k\} = \sum_{i=0}^{k} \frac{(t/\tau)^i}{i!} e^{-t/\tau}$$

e portanto $\eta$ tem lei Poisson $(t/\tau)$.

**Qui-quadrado.** Se $\xi_1, \xi_2, ..., \xi_n$ são independentes e têm lei $N(0,1)$, então a variável

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + ... + \xi_n^2$$

tem lei $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, dita lei do *qui-quadrado.* Em particular $\mathbf{E}\chi_n^2 = n$ e $\mathbf{V}\chi_n^2 = 2n$.

Se $n$ é grande, a lei de $\sqrt{2\chi_n^2}$ é muito bem aproximada pela lei normal $N\left(\sqrt{2n-1}, 1\right)$.

*T*

**de Student.** A *lei de Student* $\mathrm{t}_n$ é a lei da variável

$$\varsigma = \frac{\xi}{\sqrt{\eta/n}}$$

onde $\xi$ e $\eta$ são independentes, $\xi$ tem lei $N(0,1)$ e $\eta$ tem lei $\chi_n^2$.

Se $n$ é grande, a lei de Student é muito bem aproximada pela lei normal $N(0,1)$.

**Quantis e tabelas.** Em geral, não é possível obter fórmulas explícitas para os valores das funções de repartição das leis interessantes (normal, gamma, qui-quadrado, ...). Seja $\alpha$ um número entre 0 e 1, uma probabilidade. O *quantil de ordem* $\alpha$ da variável aleatória $\xi$ é o maior dos valores $x$ tais que $\mathbf{P}\{\xi \leq x\} \leq \alpha$, ou seja

$$q_\alpha = \sup\{x \text{ t.q. } \mathbf{P}\{\xi \leq x\} \leq \alpha\}$$

Observe que, se a densidade de $\xi$ é estrictamente positiva, então $q_\alpha$ é o único valor tal que $\mathbf{P}\{\xi \leq q_\alpha\} = \alpha$. Os livros de estatística costumam ter tabelas dos quantis das leis normal, qui-quadrado, $T$ de Student e outras.

**Quantis da lei normal.** Seja $\phi_\alpha$ o quantil de ordem $\alpha$ da lei normal $N(0,1)$, i.e.

$$\Phi(\phi_\alpha) = \alpha$$

Como a densidade da normal é uma função par, temos que $\phi_{1-\alpha} = -\phi_\alpha$ e portanto, se $\xi \sim N(0,1)$,

$$\mathbf{P}\left\{|\xi| \leq \phi_{1-\alpha/2}\right\} = 1 - \alpha$$

Não faz mal lembrar pelo menos a ordem de alguns quantis da normal: $\phi_{0.95} \simeq 1.64$ , $\phi_{0.975} \simeq 1.96$ e $\phi_{0.995} \simeq 2.58$. Portanto,

$$\mathbf{P}\{|\xi| \leq 1.64\} \simeq 0.90 \quad \mathbf{P}\{|\xi| \leq 1.96\} \simeq 0.95 \quad \mathbf{P}\{|\xi| \leq 2.58\} \simeq 0.99$$

Outra observação útil é que, se $\eta$ tem lei $N(m, \sigma^2)$, então os seus quantis $q_\alpha$ são simplesmente $q_\alpha = \sigma\phi_\alpha + m$. Também observamos que

$$\mathbf{P}\{|\eta - m| \leq \sigma\} \simeq 0.683 \quad \mathbf{P}\{|\eta - m| \leq 2\sigma\} \simeq 0.954 \quad \mathbf{P}\{|\eta - m| \leq 3\sigma\} \simeq 0.997$$

# 15 Convergência e aproximação

**Convergências.** Seja $(\xi_n)$ uma sucessão de variáveis aleatórias com valores reais definidas no espaço de probabilidades $(\Omega, \mathcal{E}, \mathbf{P})$. A noção mais ingénua de convergência (a convergência pontual), $\xi_n \to \xi$ se $\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)$ para todo $\omega \in \Omega$, não é muito interessante em probabilidades. Por exemplo, se $\xi_n$ é o número de sucessos em $n$ provas de Bernoulli, a sucessão $(\xi_n)$ não converge nunca.

A sucessão $(\xi_n)$ converge para $\xi$ *quase certamente* (ou em *quase todo ponto*), notação $\xi_n \to_{\text{qtp}} \xi$ , se

$$\mathbf{P}\left(\omega \in \Omega \text{ t.q. } \lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)\right) = 1$$

A sucessão $(\xi_n)$ converge para $\xi$ *em probabilidades*, notação $\xi_n \to_{\mathbf{P}} \xi$, se para todo $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\left(\omega \in \Omega \text{ t.q. } |\xi_n(\omega) - \xi(\omega)| < \varepsilon\right) = 1$$

É importante ter uma noção de convergência para uma sucessão de variáveis aleatórias definidas em espaços de probabilidades distintos. A sucessão $(\xi_n)$ converge para $\xi$ *em lei* (ou *em distribuição*), notação $\xi_n \to_{\text{d}} \xi$, se para toda função real contínua e limitada $\varphi$ acontece que

$$\mathbf{E}\varphi(\xi_n) \to \mathbf{E}\varphi(\xi)$$

Uma definição equivalente é: $\xi_n \to_{\text{d}} \xi$ sse para todo ponto de continuidade $x$ da função de repartição $F_\xi$ de $\xi$, acontece que

$$\lim_{n\to\infty} F_{\xi_n}(x) = F_\xi(x)$$

onde $F_{\xi_n}$ são as funções de repartição das $\xi_n$.

A hierarquia entre estas noções é a seguinte: a convergência q.t.p. implica a convergência em probabilidade, e a convergência em probabilidade implica a convergência em lei.

**Leis dos grandes números.** Dada uma sucessão de variáveis aleatórias $\xi_1, \xi_2, \ldots$ , e definidas as somas parciais $S_n = \xi_1 + \xi_2 + \ldots + \xi_n$, as leis dos grandes números são afirmações sobre a convergência das variáveis $S_n/n$. Existem muitas versões da lei dos grandes números, provadas ao longo da história com condições cada vez mais fracas sobre as variáveis $\xi_n$. Se as $\xi_n$ são independentes e identicamente distribuídas e têm variância finita, então a desigualdade de Chebyshev implica que

$$\frac{S_n}{n} \to_{\mathbf{P}} m$$

onde $m$ é o valor médio das $\xi_i$. Em particular, as variáveis $S_n/n$ convergem em lei para a constante $m$.

De facto, com hipóteses mais fracas é possível provar uma convergência mais forte. A *lei dos grandes números de Kolmogorov* (também dita *lei forte dos grandes números*) diz que, se $\xi_1, \xi_2, \ldots$ são variáveis aleatórias independentes e identicamente distribuídas com valor médio $\mathbf{E}\xi_1 = m$, então

$$\frac{S_n}{n} \to_{\text{qtp}} m$$

**Um sermão de J.L. Doob.** "...it is true that, in the mathematical context, the number of heads tossed in $n$ tosses of a balanced coin, divided by $n$, has almost sure limit $1/2$. Whether this is true or not in real life must await an examination of an experiment, a nonmathematical concept (although that fact is sometimes not made clear in elementary probability texts), in which a coin is tossed infinitely often. Up to the present time, no one has been able to toss a coin that often, and this is sufficient reason for mathematicians to hand the problem to philosophers and ingenious physicists."

**Teorema do limite central.** Dada uma sucessão de variáveis aleatórias $\xi_1, \xi_2, \ldots$ , e definidas as somas parciais $S_n = \xi_1 + \xi_2 + \ldots + \xi_n$, uma pergunta natural é se é possível dizer alguma coisa acerca da lei de $S_n$ quando $n$ é grande. Para poder comparar as informações, uma boa ideia é considerar as somas "adimensionais"

$$S_n^* = \frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{V}S_n}}$$

que têm valor médio 0 e variância 1. O primeiro resultado nesse sentido foi obtido por De Moivre e Laplace, no caso de variáveis $\xi_n$ independentes e identicamente distribuídas com lei de Bernoulli, e diz que a lei de $S_n^*$ é bem aproximada por uma lei normal, quando $n$ é grande. A versão moderna deste teorema tem uma demonstração elegante, baseada no teorema de Lévy, um resultado que diz essencialmente que a convergência pontual das funções características é equivalente à convergência em lei.

**Teorema do limite central.** *Sejam $\xi_1, \xi_2, \ldots$ varáveis aletórias independentes e identicamente distribuídas, com $\mathbf{E}\xi_n = m$ e $\mathbf{V}\xi_n = \sigma^2 > 0$. Então as variáveis*

$$S_n^* = \frac{\xi_1 + \xi_2 + \ldots + \xi_n - nm}{\sigma\sqrt{n}}$$

*convergem em lei para uma variável normal $N(0,1)$ quando $n \to \infty$.*

**dem.** A demonstração, esboçada a seguir, "explica" por que a lei normal é uma lei muito especial. A função característica de uma variável aleatória $\xi$ é definida por $\phi_\xi(\theta) = \mathbf{E}e^{i\theta\xi}$. Somar $n$ variáveis independentes e identicamente distribuídas com média 0 e variância 1 corresponde, no mundo das funções características, a passar de $\phi(\theta)$ para $\phi(\theta/\sqrt{n})^n$. Esta sequência de funções caracerísticas, sob condições oportunas, converge para um ponto fixo da transformada de Fourier: a gaussiana!

A ideia é aplicar o *teorema de Lévy*, que diz: sejam $\phi, \phi_1, \phi_2, ..$ as funções características das variáveis aleatórias $\xi, \xi_1, \xi_2, \ldots$ Então $\xi_n \to^{\mathcal{L}} \xi$ sse $\phi_n(\theta) \to \phi(\theta)$ para todo $\theta \in \mathbf{R}$.

Seja $\phi$ a função característica de $\eta_k = (\xi_k - m)/\sigma$. Então

$$\phi_{S_n^*}(\theta) = \phi\left(\theta/\sqrt{n}\right)^n$$

Sabemos que $\phi(\theta) = 1 - \frac{1}{2}\theta^2 + \mathrm{o}(\theta^2)$, porque $\mathbf{E}\eta_k = 0$ e $\mathbf{V}\eta_k = 1$. Calculando o limite temos

$$\lim_{n\to\infty} \phi\left(\theta/\sqrt{n}\right)^n = \lim_{n\to\infty} \exp\left(n\left(-\frac{\theta^2}{2n} + \mathrm{o}\left(\theta^2/n\right)\right)\right) = e^{-\theta^2/2}$$

que é a função característica da lei normal. Pelo teorema de Lévy, $S_n^*$ converge em lei para uma variável normal $N(0,1)$.

□

**Aproximação normal.** O teorema do limite central sugere que, se $n$ é grande, a probabilidade $\mathbf{P}\{a < S_n^* < b\}$ pode ser aproximada por $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$, no sentido em que

$$\mathbf{P}\{a < S_n^* < b\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

quando $n \to \infty$.

A velocidade de convergência depende, obviamente, das leis das variáveis $\xi_n$, e portanto não é possível, em geral, dizer a partir de quais valores de $n$ a aproximação começa a ser boa. É bom saber que a convergência costuma ser lenta. De facto, um *teorema de Berry e Esseen* diz que uma estimação do erro

$$\mathrm{erro}_n = \sup_{-\infty<x<\infty} \left| \mathbf{P}\{S_n^* < x\} - \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \right|$$

é $\mathrm{erro}_n \leq \mathrm{const}/\sqrt{n}$, onde a constante depende dos primeiros três momentos das variáveis, e em geral não pode ser melhor.

**Histogramas.** Sejam $\xi_1, \xi_2, ...$ variáveis independentes e identicamente distribuídas, com densidade $f_\xi$. Sejam $[a, b]$ um intervalo da recta e $\eta_n$ a variável

$$\eta_n = \frac{1}{n} \sum_{k=1}^{n} 1_{[a,b]} \circ \xi_k$$

onde $1_{[a,b]}(t) = 1$ se $t \in [a, b]$ e $1_{[a,b]}(t) = 0$ se $t \notin [a, b]$. Então a lei dos grande números de Kolmogorov implica que

$$\eta_n \to^{\text{qtp}} \int_a^b f_\xi(t)dt$$

**Montecarlo.** Sejam $\xi_1, \xi_2, ...$ variáveis independentes e identicamente distribuídas, com lei uniforme no intervalo $[0, 1]$, e seja $\varphi : [0, 1] \to \mathbf{R}$ uma função limitada e integrável. Então as variáveis $\varphi(\xi_k)$ têm valor médio

$$\mathbf{E}\varphi(\xi_k) = \int_0^1 \varphi(t)dt$$

Pela lei dos grandes números

$$\frac{1}{n} \sum_{k=1}^{n} \varphi(\xi_k) \to^{\text{qtp}} \int_0^1 \varphi(t)dt$$

Portanto, se temos um gerador de números aleatórios com lei uniforme no intervalo (todos os computadores têm sucessões de números "aleatórios" em memória!) podemos aproximar numericamente o integral de $\varphi$ calculando as somas à esquerda. Estes algoritmos são chamados *métodos de Montecarlo.* A velocidade de convergência destes algoritmos é inferior à velocidade dos métodos de integração usuais, mas a implementação é muito mais fácil, sobretudo em dimensão maior que um.

**Simulações, geradores de números aleatórios.** Problemas de física particularmente difíceis conduzem à necessidade de fazer simulações de experiências aleatórias. As linguagem como pascal, fortran, C, contêm "routines" que produzem sucessões de números "aleatórios" (isso mesmo: uma máquina pode ser programada para produzir sucessões de números que parecem aleatórios!) com distribuição uniforme em $[0, 1]$. A partir destas sucessões é possível simular sucessões de números aleatórios com outras leis (e também existe muito software com sucessões de números aleatórios com as leis mais importantes).

**Exercícios.**

**a.** Se $\xi$ tem lei uniforme em $[0, 1]$ então $a + (b - a)\xi$ tem lei uniforme em $[a, b]$.

**b.** Se $\xi$ tem lei uniforme em $[0, 1]$ e $\tau > 0$, então $-\tau \log(1 - \xi)$ tem lei exponencial $\exp(\tau)$.

**c.** Se $\xi_1, \xi_2, ...$ são independentes e identicamente distribuídas, com lei exponencial de parâmetro $\tau$, então $\eta$, definida por

$$\eta = \sup \{k \text{ t.q. } \xi_1 + \xi_2 + ... + \xi_k \leq 1\}$$

tem lei Poisson $(1/\tau)$.

**d.** Se $\xi_1, \xi_2, ...$ são independentes e identicamente distribuídas, com lei uniforme em $[0, 1]$, então $\eta_1, \eta_2, ...$ definidas por

$$\eta_k = \begin{cases} 1 & \text{se } \xi_k \leq p \\ 0 & \text{se } \xi_k > p \end{cases}$$

são independentes e têm lei de Bernoulli $B(1, p)$. Portanto $\eta_1 + \eta_2 + ... + \eta_n$ tem lei binomial $B(n, p)$. Pelo teorema do limite central a variável

$$\frac{\eta_1 + \eta_2 + ... + \eta_n - np}{\sqrt{npq}}$$

é uma boa aproximação de uma variável normal $N(0, 1)$ se $n$ é grande.

**e.** Seja $A \subset [0,1] \times [0,1]$. Para estimar a área de $A$, uma boa ideia é produzir uma sucessão de pontos aleatórios com lei uniforme no quadrado $[0,1] \times [0,1]$ e calcular a fração dos pontos que pertencem a $A$. Esta variável converge em quase todo ponto para $\text{area}\,(A)$.

# 16 Estimação

**Observações.** Um físico tem uma teoria física, que contém um observável chamado $x$ (a constante de gravitação, a massa do electrão, o tempo característico do carbono $C_{14}$, ...a probabilidade de sair cara no lançamento de uma moeda). Repete várias vezes uma experiência em condições que ele julga idênticas (no sentido em que controla tudo o que é controlável) e obtém os resultados experimentais $x_1, x_2, ..., x_n$. A coisa mais honesta que ele pode dizer é que o observável está entre $x_{\min}$ e $x_{\max}$, mais ou menos. Os físicos costumam acreditar na existência do universo, e nas próprias teorias, portanto na existência do valor "verdadeiro" de $x$. Uma estimação natural é a *média aritmética* dos resultados

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right)$$

Os físicos também sabem que não faz sentido nenhum acreditar que o valor de $x$ seja exactamente $\overline{x}$ (as leis da física implicam que a posição de Vénus influencie a queda de uma pedra da torre de Pisa, embora não seja possível dizer qual é a sua influência!), só acreditam em afirmações como

$$\text{o observável é igual a } \overline{x} \pm \Delta x$$

que lêem: o verdadeiro valor do observável $x$ está, "com grande probabilidade", entre $\overline{x} - \Delta x$ e $\overline{x} + \Delta x$. Um dos problemas da estatística é estimar um valor razoável do "erro" $\Delta x$.

**Média aritmética e erros quadráticos.** A média aritmética $\overline{x}$ é a média mais democrática entre os valores observados. É também o valor de $a$ que minimiza a soma

$$(x_1 - a)^2 + (x_2 - a)^2 + ... + (x_n - a)^2$$

dos quadrados dos "erros" nas distintas observações. O mínimo

$$(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... (x_n - \overline{x})^2$$

é uma medida de quanto os valores de $x$ diferem da média aritmética, e a sua média

$$\frac{1}{n}\left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... (x_n - \overline{x})^2\right)$$

é uma medida de quanto cada valor $x_k$ difere de $\overline{x}$.

**Exemplo.** Lanço $n$ vezes uma moeda e observo $k$ vezes coroa. Uma conjectura é que a probabilidade $p$ de sair coroa em cada lançamento é a frequência observada $f = k/n$. Se tivesse lançado a moeda mais uma vez, os resultados possíveis teriam sido $k/(n+1)$ ou $(k+1)/(n+1)$. Portanto, não faz sentido ficar com a resposta $k/n$, é mais honesto dizer que a probabilidade pode ser $f \pm \Delta f$, com $\Delta f \geq 1/n$. Acabo de lançar dez vezes uma moeda de 50 liras, e obtive 5 vezes coroa (juro!): tudo o que posso esperar é que $p = 0.5 \pm 0.1$. Mesmo assim, esta não é uma estimação honesta...

**Exemplo.** Estudando os dados dos astrofísicos do seu tempo, Hubble observou que as velocidades $v$ em que as galáxias fogem de nós parecem ser proporcionais às distâncias $r$ entre elas e nós. Ele conjecturou a lei $v = H \cdot r$. Nós temos observações das distâncias e das velocidades. Como estimar $H$ sabendo que a distância é $r \pm \Delta r$ e a velocidade é $v \pm \Delta v$? A resposta correcta é que $H$ está entre $H_{\min}$ e $H_{\max}$, o mínimo e o máximo da função $H = v/r$ no domínio $[r - \Delta r, r + \Delta r] \times [v - \Delta v, v + \Delta v]$. Se os erros relativos são pequenos, uma boa aproximação é

$$\text{a constante de Hubble } H \text{ é } \frac{v}{r} \pm \left(\frac{1}{r}\Delta v + \frac{v}{r^2}\Delta r\right)$$

porque os outros termos no desenvolvimento de Taylor da função $v/r$ são mais pequeninos. A receita acima corresponde a somar os erros relativos. Esta também, em geral, não é a melhor estimação possível...

**Apresentação do resultado.** Dizer que um observável é igual a

$$3.14159265359 \pm 0.062$$

não contém mais informação do que dizer que é igual a

$$3.14 \pm 0.06$$

A final, o erro $\Delta x$ é uma medida da "sensibilidade" dos instrumentos do laboratório, ou melhor da reproduzibilidade das experiências.

**Modelo das observações.** Porque é uma boa ideia usar a média aritmética das observações para estimar o valor de um observável? Ou temos fé, ou temos que fazer um "modelo das observações" para justificar a nossa escolha. Um modelo é assim.

Existe um valor verdadeiro do observável $x$, que chamamos $m$. Cada observação é uma experiências aleatória, descrita pela variável $\xi$ com esperança $\mathbf{E}\xi = m$ (ou seja acreditamos que os instrumentos observam mesmo o parâmetro $x$, não há erros sistemáticos). O nosso controlo das condições do laboratório não é, não pode ser, perfeito, portanto a variável $\xi$ é mesmo variável, e tem uma certa lei. Não podemos saber a lei de $\xi$, logo podemos supôr que tem uma certa variância $\mathbf{V}\xi = \sigma^2$. Também podemos supôr que as diferentes observações são independentes (fazer física é possível precisamente na medida em que físicos que vivem em laboratórios distintos, um em Braga e outro em Guimarães, podem reproduzir e verificar as experiências dos outros: a "independência" das experiências é uma das hipóteses necessárias para poder falar de física). Então a média aritmética

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right)$$

das $n$ observações, que os estatísticos chamam *média amostral*, tem boa probabilidade de estar perto de $m$ quando $n$ é grande (lei dos grandes números).

Os histogramas dos resultados das experiências reais são muito parecidos com o gráfico de uma distrubuição normal. Portanto uma hipótese de trabalho pode ser que $\xi$ tem lei normal $N(m, \sigma^2)$. Então a média aritmética $\overline{x}$ tem lei $N(m, \sigma^2/n)$. Por outro lado, mesmo se $\xi$ não fôr normal, o teorema do limite central diz que, quando $n$ é muito grande, a lei de $\overline{x}$ é bem aproximada pela lei normal. Se $n$ não é muito grande, também existe uma "justificação" para esta hipótese. É razoável pensar que a variável $\xi$ seja igual a $m$ mais uma soma de muitos "erros aleatórios" pequenos devidos a pequenas perturbações incontroláveis das condições de laboratório (uma borboleta que posou no aparelho, uma eclipse de lua, a vizinha que prepara um cafezinho, o eixo do bem que bombardeia uma aldeia do eixo do mal, ...). Mais uma vez, se os erros são muitos, e se "em média" são nulos, o teorema do limite central sugere que a lei de $\xi$ é bem aproximada por uma lei normal com média $m$. A variância $\sigma^2$ é uma medida da "sensibilidade" dos instrumentos de laboratório.

O modelo também diz que quanto maior for a amostra tanto maior é a probabilidade de $\overline{x}$ estar perto de $m$ (lei dos grandes números). O modelo faz previsões quantitativas: diz que

$$\frac{\overline{x} - m}{\sigma/\sqrt{n}}$$

tem lei $N(0, 1)$. O problema é que nós não temos nenhuma ideia do valor de $\sigma$. Como estimar $\sigma$? Os estatísticos chamam

$$S^2 = \frac{1}{n-1}\left(\left(x_1 - \overline{x}\right)^2 + \left(x_2 - \overline{x}\right)^2 + ... \left(x_n - \overline{x}\right)^2\right)$$

*variância amostral*. Dentro do nosso modelo da observação esta é uma variável aleatória, porque cada $x_k$ é uma variável. A esperança de $S^2$ é $\mathbf{E}S^2 = \sigma^2$, portanto $S^2$ é uma estimação razoável de $\sigma^2$. Também útil é saber que a variância de $S^2$ é $\mathbf{V}S^2 = \frac{2\sigma^2}{n-1}$, e portanto se $n$ é grande a variância amostral tem boa probabilidade de estar perto da variância de $\xi$ (lei dos grandes números). O modelo diz que a variável

$$\frac{n-1}{\sigma^2} \cdot S^2$$

tem lei $\chi^2_{n-1}$, mas esta variável ainda contém a variância desconhecida $\sigma^2$. Mais interessante, enfim, é saber que o modelo diz que a variável

$$\frac{\overline{x} - m}{S/\sqrt{n}}$$

onde só aparece o parâmetro $m$, o observável que queremos estimar, tem lei de Student $\mathrm{t}_{n-1}$ (teorema de Cochran).

Sumário: um modelo razoável das experiências repetidas diz que $\frac{\overline{x}-m}{S/\sqrt{n}}$ é uma variável aleatória com lei de Student. Um físico, depois de ter feito as $n$ experiências, pode dizer, por exemplo (se $n$ é grande, os quantis da lei de Student não diferem significativamente dos quantis da lei normal), que "o valor do observável $x$ está entre $\overline{x}-2\frac{S}{\sqrt{n}}$ e $\overline{x}+2\frac{S}{\sqrt{n}}$ com probabilidade $\geq 95\%$", e , julgando esta uma probabilidade mesmo grande, escrever

$$\text{o valor do observável } x \text{ é igual a } \overline{x} \pm 2\frac{S}{\sqrt{n}}$$

**Exemplo: a agulha de Buffon.** O chão tem linhas paralelas a distância $\ell$. Lanço $n$ vezes uma agulha de comprimento $\ell$ e registo a frequência $f$ das vezes que a agulha toca uma das linhas. Num modelo natural estas são provas de Bernoulli. Em cada prova, uma probabilidade natural é: lei uniforme pela posição do centro da agulha entre uma linha e a sucessiva, e lei uniforme pelo ângulo que a agulha forma com a direcção das linhas. A resposta é que a probabilidade de sucesso em cada prova é $p = 2/\pi$. Uma estimação da probabilidade $p$ é

$$\text{o observável } p \text{ é igual a } f \pm \Delta f$$

com probabilidade 95%, onde $\Delta f \simeq 1/\sqrt{n}$ (porque $1/4$ é a maior variância de uma variável de Bernoulli). O observável $\pi$ é igual a $2/p$, portanto um físico que quer estimar $\pi$ escreve

$$\pi = \frac{2}{f} \pm \frac{2}{f^2}\Delta f$$

**Exemplo: sondagens.** $N$ americanos podem escolher entre os candidatos $B$ ou $K$ nas eleições presidenciais. Uma amostra de $n$ eleitores é entrevistada: $b'$ eleitores da amostra afirmam estar intencionados em votar o senhor $B$ e os outros $k' = n - b'$ afirmam estar intencionados em votar o senhor $K$. O problema é estimar o numero $b$ de eleitores, dentro da população total, que estão intencionados em votar $B$, e portanto a percentagem $b/N$. A variável $b'$ tem lei hipergeométrica, que, se $n \ll N$, é bem aproximada pela lei binomial $B(n, b/N)$. Portanto, um intervalo de confiança 95% para $b/N$ é

$$b'/N \pm \frac{1}{\sqrt{n}}$$

O $\pm 1/\sqrt{n}$ é o que os técnicos chamam "margem de erro da sondagem". Naturalmente, o verdadeiro problema é arranjar uma amostra representativa da população, ou seja simular uma escolha aleatória dentro de uma população cujas intenções são determinadas por factores sociais...

**Estimadores.** A teoria do físico é mesmo um modelo probabilístico, ou seja uma variável aleatória $\xi$. Os resultados $x_1, x_2, ..., x_n$ das experiências são portanto valores possíveis de uma sucessão $\xi_1, \xi_2, ..., \xi_n$ de variáveis aleatórias independentes com a lei de $\xi$. A lei de $\xi$ contém parâmetros, por exemplo a média $m$, ou a variância $\sigma^2$. Os observáveis são os parâmetros da lei de $\xi$. Esta é a situação mais geral: o modelo físico contém o modelo das experiências.

Estimar o parâmetro $\theta$ quer dizer encontrar uma função $x_1, x_2, ..., x_n \mapsto t(x_1, x_2, ..., x_n)$ tal que o seu valor com boa probabilidade seja perto do valor de $\theta$. A coisa mais natural é procurar $t$ de modo que a sua esperança seja $\mathbf{E}_\theta t(\xi_1, \xi_2, ..., \xi_n) = \theta$, onde $\mathbf{E}_\theta$ denota a esperança com respeito a lei determinada pelo valor $\theta$ do parâmetro. Em geral esta é uma boa estratégia, mas não há razão para excluir outras soluções (a forma da função $t$ pode ser muito complicada, e nós só queremos é estimar... muitas vezes temos que nos contentar com aproximações, e estratégias nas experiências podem ajudar). Os estatísticos chamam *estimadores* a estas funções $t$, e chamam *centrados* aos estimadores tais que $\mathbf{E}_\theta t = \theta$. No espírito da lei dos grandes números, um estimador $t$ é dito *consistente* se para todo $\varepsilon > 0$

$$\mathbf{P}_\theta(|t - \theta| \geq \varepsilon) \to 0$$

quando $n \to \infty$.

**Exemplos.** Se $\xi$ tem lei normal $N\left(m,\sigma^2\right)$, a média amostral $\overline{x}$ é um bom estimador da esperança: tem esperança $m$, logo é centrado, e variância $\sigma^2/n$, logo é consistente (pela desigualdade de Chebyshev).

Também, a variância amostral $S^2$ é um bom estimador da variância: tem esperança $\sigma^2$, logo é centrado, e variância $\frac{2\sigma^2}{n-1}$, logo é consistente.

**Desigualdade de Rao-Cramér.** Uma medida da bontade do estimador centrado $t$ é a sua variância $\mathbf{V}_\theta t$. O estimador centrado $t^*$ do parâmetro $\theta$ é dito *eficiente* se

$$\mathbf{V}_\theta t^* = \inf_t \mathbf{V}_\theta t$$

onde o ínfimo é sobre todos os estimadores centrados de $\theta$.

Seja $p_\theta$ a densidade discreta das variáveis $\xi_1, \xi_2, ..., \xi_n$. A probabilidade de observar os valores $x_1, x_2, ..., x_n$ é

$$\mathbf{P}_\theta\left(\omega\right) = \mathbf{P}_\theta\left(\xi_1 = x_1, \xi_2 = x_2, ..., \xi_n = x_3\right) = \prod_{i=1}^{n} p_\theta\left(x_i\right)$$

e a probabilidade do evento certo é

$$1 = \sum_\omega \mathbf{P}_\theta\left(\omega\right)$$

Assumindo que seja possível derivar em ordem a $\theta$ a igualdade acima e que $\mathbf{P}_\theta\left(\omega\right) > 0$ para todo $\omega$, temos que

$$0 = \sum_\omega \mathbf{P}_\theta\left(\omega\right) \cdot \frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right)\right)$$

Se $t$ é um estimador centrado de $\theta$, então

$$\theta = \mathbf{E}_\theta t = \sum_\omega t\left(\omega\right) \mathbf{P}_\theta\left(\omega\right)$$

e derivando obtemos

$$1 = \sum_\omega t\left(\omega\right) \cdot \mathbf{P}_\theta\left(\omega\right) \cdot \frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right) = \mathbf{E}_\theta\left(t\left(\omega\right) \cdot \frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right)\right)$$

Juntando as duas expressões temos que

$$1 = \mathbf{E}_\theta\left((t-\theta) \cdot \frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right)\right)$$

e, pela desigualdade de Cauchy-Schwarz,

$$1 \leq \mathbf{E}_\theta\left(t-\theta\right)^2 \cdot \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right)\right)^2$$

A conclusão é a *desigualdade de Rao-Cramér*

$$\inf_{t \text{ centrado}} \mathbf{V}_\theta t \geq \frac{1}{I_\theta}$$

onde a *informação de Fisher* é definida como

$$I_\theta = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta} \log \mathbf{P}_\theta\left(\omega\right)\right)^2$$

**Exemplo.** A informação de Fisher no modelo das $n$ provas de Bernoulli com probabilidade de sucesso $p$ é $I_p = \frac{n}{p(1-p)}$, e isto mostra que a média amostral $\overline{x}$ é um estimador eficiente de $p$.

**Estimadores de máxima verosimilhança.** Uma receita que produz estimadores do parâmetro $\theta$ é a seguinte. Dado $\theta$, é possível calcular a densidade de probabilidade

$$p_\theta(x_1, x_2, ..., x_n)$$

de obter os valores $x_1, x_2, ..., x_n$ em $n$ provas independentes e identicamente distribuídas com a lei determinada por $\theta$. Um *estimador de máxima verosimilhança* é uma função $x_1, x_2, ..., x_n \mapsto t(x_1, x_2, ..., x_n)$ tal que

$$p_{t(x_1,x_2,...,x_n)}(x_1, x_2, ..., x_n) = \max_\theta p_\theta(x_1, x_2, ..., x_n)$$

para todos os possíveis valores de $x_1, x_2, ..., x_n$.

**Exemplo.** Se $x$ tem lei normal $N(m, \sigma^2)$, o estimador de máxima verosimilhança para $m$ é a média amostral $\overline{x}$. De facto, a densidade de probabilidade $p_\theta(x_1, x_2, ..., x_n)$ é proporcional à exponencial da soma

$$-\sum_{k=1}^{n}(x_k - m)^2$$

que é máxima quando $m = \overline{x}$. O estimador de máxima verosimilhança para $\sigma^2$ é

$$\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2$$

que difere pouco da variância amostral quando $n$ é grande.

**Exercícios.**

**a.** Se $\xi$ tem lei de Poisson Poisson$(\lambda)$, o estimador de máxima verosimilhança para $\lambda$ é a média amostral $\overline{x}$.

**b.** Se $\xi$ tem lei exponencial $\exp(\tau)$, o estimador de máxima verosimilhança para o tempo característico $\tau$ é a média amostral $\overline{x}$.

**c.** Determine o estimador de máxima verosimilhança para o valor médio da variável "tempo de espera" nas provas de Bernoulli.

**Exemplo.** Dois namorados querem saber quantas estrelas caem durante a noite de S. Lorenço. Ela contou $k_a$ estrelas, ele $k_e$ estrelas, e o número das estrelas que foram vistas pelos dois é $k_{ae}$. Um modelo simples é fazer a hipótese de que a observação de cada estrela, entre as $n$ que caíram, por parte de cada um deles seja uma prova de Bernoulli com probabilidade $p_a$ e $p_e$ respetivamente. A lei dos grandes números sugere que $k_a \sim np_a$ e $k_e \sim np_e$. Mas também $k_{ae} \sim np_ap_e$. Portanto uma estimação de $n$ pode ser

$$\frac{k_ak_e}{k_{ae}}$$

e este número não depende dos $p$, mas só da hipótese de independência!

**Exemplo.** Uma caixa contém $N$ bolinhas numeradas de 1 até $N$. Retiro $n$ bolinhas, e quero estimar $N$. A variável aleatória $\xi =$"o maior dos números que trazem as $n$ bolinhas retiradas" tem densidade

$$\mathbf{P}(\xi = k) = (k^n - (k-1)^n)N^{-n}$$

porque $\mathbf{P}(\xi \leq k) = (k/N)^n$. Aproximando a soma com um integral (o que não faz mal se $N \gg 1$), a esperança de $\xi$ é $\mathbf{E}\xi \simeq \frac{n}{n+1}N$. Por exemplo, se os talibãs capturam 10 tanques americanos, e o maior número de matrícula deles é 910, podem estimar que os americanos têm um arsenal de $\simeq 1000$ tanques.

**Intervalos de confiança.** Os resultados de uma experiência podem ser apresentados da seguinte forma: o valor $m$ do observável $x$ está no intervalo $a \leq m \leq b$, dito *intervalo de confiança*, com probabilidade $\geq 1-\alpha$, dita *nível* (do intervalo de confiança). Um intervalo de confiança simétrico, i.e. do tipo $a-\varepsilon \leq m \leq a+\varepsilon$, costuma ser apresentado pela expressão $m = a \pm \varepsilon$.

**Intervalos para a média.** Se no nosso modelo das observações a variável $\frac{\overline{x}-m}{\sigma/\sqrt{n}}$ tem lei normal $N(0,1)$, um intervalo de confiança de nível $1-\alpha$ é

$$m = \overline{x} \pm \phi_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

onde $\phi_{1-\alpha/2}$ é o quantil da lei normal. Valores típicos são $\phi_{0.975} \simeq 1.96$ se o nível é 95%, ou $\phi_{0.995} \simeq 2.6$ se o nível é 99%.

Num modelo em que temos que estimar a variância amostral (ou seja, sempre!), um intervalo de nível $1-\alpha$ é

$$m = \overline{x} \pm t_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

onde $t_{1-\alpha/2}$ é o quantil da lei de Student $\mathrm{t}_{n-1}$.

**Intervalos para a "probabilidade".** Um caso particular é uma experiência em que temos que estimar uma probabilidade $p$ (a probabilidade de sucesso no modelo das provas de Bernoulli), ou seja os resultados possíveis são $x_k = 0$ ou 1 e o resultado das experiências é a frequência $f = \overline{x} =$"número de sucessos em $n$ provas"$/n$. O teorema do limite central diz que, se $n$ é grande, a lei da variável $\frac{\overline{x}-p}{\sqrt{p(1-p)}/\sqrt{n}}$ é bem aproximada pela lei normal $N(0,1)$. Uma boa ideia é estimar a variância $p(1-p)$ com o seu máximo 1/4. Um intervalo "generoso" de nível $\geq 1-\alpha$ é portanto

$$p = \overline{x} \pm \phi_{1-\alpha/2} \cdot \frac{1}{2\sqrt{n}}$$

Se suspeitamos que a probabilidade $p$ seja pequena, ou grande, o intervalo acima é sobreestimado. Um intervalo melhor é dado calculando a variância amostral como no caso geral. Uma aproximação razoável é dada pela fórmula

$$p = \overline{x} \pm \phi_{1-\alpha/2} \cdot \frac{\sqrt{f(1-f)}}{\sqrt{n}}$$

porque a variância amostral é $n/(n-1)$ vezes $f\,(1-f)$.

Se suspeitamos que a probabilidade $p$ seja muito, mas muuuuuito, pequena, intervalos melhores devem ser estimados utilizando a aproximação de Poisson.

**Intervalos para a variância.** Às vezes é importante estimar a variância $\sigma^2$ dos resultados das experiências (é uma medida da reproducibilidade da experiência, ou da sensibilidade dos instrumentos do laboratório). No modelo, a variável $\frac{n-1}{\sigma^2} \cdot S^2$ tem lei $\chi^2_{n-1}$. Fixado um nível $1-\alpha$, dois intervalos de confiança são

$$0 \leq \sigma^2 \leq \frac{n-1}{q_\alpha} \cdot S^2$$

e

$$\frac{n-1}{q_{1-\alpha/2}} \cdot S^2 \leq \sigma^2 \leq \frac{n-1}{q_{\alpha/2}} \cdot S^2$$

onde $q_\alpha$, $q_{1-\alpha/2}$ e $q_{\alpha/2}$ são os quantis da lei $\chi^2_{n-1}$.

**Relative standard uncertainty.** Se não temos tabelas no laboratório, basta lembrar que intervalos de confiança "generosos" com nível de confiança $\geq 95\%$ e $\geq 99\%$ são da ordem de

$$m = \overline{x} \pm 2 \cdot \frac{S}{\sqrt{n}} \quad \text{e} \quad m = \overline{x} \pm 3 \cdot \frac{S}{\sqrt{n}}$$

(os quantis da lei de Student não dão erros relativos significativamente diferentes dos quantis da lei normal, se $n$ é grande). Aliás, em primeiro lugar, a arbitrariedade do nível dos intervalos de confiança não tem nenhum significado físico. Em segundo lugar, a velocidade de convergência no teorema do limite central, que supostamente justifica o modelo das observações, é lenta. Por exemplo, o "erro" no nível de confiança para a probabilidade em $n$ provas de Bernoulli com probabilidade de sucesso perto de $1/2$ é da ordem de $1/\sqrt{n}$. Se $n$ é 10000, o que nós acreditamos ser um intervalo de nível 95% pode muito bem ser em realidade um intervalo de nível 94% ou 96%. Moral: se $n$ não é muito, mas mesmo muito, grande, o nível de confiança não é confiável!

Os parâmetros físicos são a média $\overline{x}$ e o desvio padrão $S$ observados (que, além de serem uns estimadores centrados e consistentes, são os estimadores mais "democráticos", dando peso igual às diferentes observações). O *desvio padrão da média* $S_m = S/\sqrt{n}$ é uma medida da incerteza na estimação de $m$, o suposto valor verdadeiro de $x$. Mais significativo do que o desvio padrão é o *desvio padrão relativo* $S_m/\overline{x} = S/(\overline{x}\sqrt{n})$ da média, que diz a quantidade dos dígitos significativos na estimação de $m$.

Por exemplo, uma tabela das constantes da física tem este valor da constante de gravitação de Newton:

$$G = 6.673(10) \times 10^{-11} \mathrm{m}^3 \mathrm{kg}^{-1} \mathrm{s}^{-2} \text{ with relative standard uncertainty } 1.5 \times 10^{-3}$$

**Quantas observações fazer.** Vale a pena observar que o erro (e o erro relativo) na estimação de um observável é inversamente proporcional à raiz quadrada $\sqrt{n}$ do número $n$ de experiências. Portanto, para reduzir o erro de um factor 10, um físico tem que multiplicar por 100 as observações...

Por exemplo, para estimar $\pi$ com um erro da ordem de 0.01 tenho que lançar algo como 10000 agulhas de Buffon. Para chegar a ter informações sobre o quarto dígito decimal de $\pi$ tenho que lançar 100000000 agulhas!

**Propagação dos erros.** Se os observáveis $x, y, ..., z$ são estimados ser $\overline{x} \pm \Delta x$, $\overline{y} \pm \Delta y$, ..., $\overline{z} \pm \Delta z$, e se julgamos que os erros nos diferentes observáveis são independentes, uma boa ideia é estimar o observável $f = f(x, y, ..., z)$ com $\overline{f} \pm \Delta f$, onde $\overline{f} = f(\overline{x}, \overline{y}, ..., \overline{z})$ e o erro é a raiz quadrada de

$$(\Delta f)^2 = \left(\frac{\partial f}{\partial x}(\overline{x}, \overline{y}, ..., \overline{z})\right)^2 \cdot \Delta x^2 + \left(\frac{\partial f}{\partial y}(\overline{x}, \overline{y}, ..., \overline{z})\right)^2 \cdot \Delta y^2 + ... + \left(\frac{\partial f}{\partial z}(\overline{x}, \overline{y}, ..., \overline{z})\right)^2 \cdot \Delta z^2$$

Uma justificação desta receita vem das hipóteses: os erros $\Delta x$, $\Delta y$, ..., $\Delta z$ são gaussianos, independenes e pequenos.

# 17 Testes estatísticos

**Testes.** Às vezes, mais do que estimar uns observáveis $x$ e $y$, um físico está interessado em testar uma hipótese do tipo $x = y$, ou $x > a$ (por exemplo, não queremos estimar o valor da constante de Hubble $H$, mas só saber se o universo está em expansão, i.e. se $H > 0$).

Num teste, temos que tomar uma decisão, sim ou não, aceitar ou rejeitar a hipótese, dependendo dos valores obtidos nas observações. O *nível de significância* $\alpha$ do teste é a maior das probabilidades

$$\mathbf{P}\,(\text{rejeitar a hipótese} \mid \text{a hipótese é verdadeira})$$

(os estatístico chamam esta "a probabilidade de fazer um erro do primeiro tipo"). A hipótese determina a lei, ou uma família de leis, da variável observada. O resultado das observações é uma variável aleatória $z$, função dos resultados experimentais $x_1, x_2, ..., x_n$. Fazer um teste consiste em fixar uma região $R$, dita *região crítica* do teste, do valor observado $z$ que consideramos não aceitável se a hipótese for verdadeira. O complementar desta região é dita região de aceitação do teste. A receita do teste é: se $z \in R$ rejeitamos a hipótese, se $z \notin R$ aceitamos a hipótese. A escolha da região crítica determina o nível de significância $\alpha$ do teste.

Um físico honesto testa a hipótese mais conservadora (se quero anunciar ao mundo que a água tem memória, testo a hipótese de que a água não tem memória!), e portanto é importante ter valores pequenos de $\alpha$, tipicamente 10%, 5% ou 1%.

Pode acontecer que as duas hipóteses alternativas sejam igualmente razoáveis. Neste caso também é significativo o parâmetro $\beta$, definido como a maior das probabilidades

$$\mathbf{P}\,(\text{aceitar a hipótese} \mid \text{a hipótese é falsa})$$

(os estatísticos chamam esta "a probabilidade de fazer um erro do segundo tipo"). Uma boa estratégia é, então, construir uma região crítica tentando minimizar a soma $\alpha + \beta$.

**Exemplo.** Uma maneira ingénua de testar a hipótese $x = y$ consiste em calcular os intervalos de confiança de nível $1-\alpha$ para os dois observáveis, e aceitar a hipótese se estes intervalos têm pontos comuns. Da mesma forma, parece razoável aceitar a hipótese $x > a$ se o intervalo de confiança de nível $1-\alpha$ de $x$ contém valores $> a$. Esta ideia é a base do procedimento formal (só aparentemente mais complicado) que os estatísticos chamam testes de hipótese.

**Exemplo: as moedas com spin inteiro.** Há livros (e professores) de estatística que dizem que um modelo do lançamento "simultâneo" de duas moedas iguais é o seguinte: três acontecimentos possíveis, "duas caras", "duas coroas" e "uma cara e uma coroa", com probabilidade uniforme. Este modelo diz que a probabilidade do evento "uma cara e uma coroa" é 1/3. Como decidir se as moedas que temos no bolso são bosões? Eu suspeito que um modelo melhor é aquele que diz que a probabilidade do evento "uma cara e uma coroa" é 1/2. O senso comum sugere a seguinte estratégia. Se $n$ é muito grande, uns intervalos de confiança para a probabilidade do evento nos dois modelos são disjuntos (basta pôr $1/\sqrt{n} \ll |1/2 - 1/3|$, para um nível de confiança de 95%). Se eu lançar uma centena de vezes as duas moedas (o mais "simultaneamente" possível, claro!), posso razoavelmente esperar que a frequência observada "escolha" um dos dois intervalos alternativos, ou quem sabe nenhum, e portanto o modelo melhor.

**Testes sobre médias.** Uma estratégia para testar a hipótese $m > a$ é assim. Obtidos os resultados $x_1, x_2, ..., x_n$ das experiências, podemos calcular

$$z = \frac{\overline{x} - a}{S/\sqrt{n}}$$

O modelo nos diz que a variável $t = \frac{\overline{x}-m}{S/\sqrt{n}}$ tem lei de Student $\mathrm{t}_{n-1}$. Podemos ver, nas tabelas, o valor $t_\alpha$ tal que $\mathbf{P}\,(t \le t_\alpha) = \alpha$. Se a hipótese é verdadeira, i.e. se $m > a$, então

$$\begin{aligned} \mathbf{P}\,(z \le t_\alpha) &= \mathbf{P}\left(t + \frac{m-a}{S/\sqrt{n}} \le t_\alpha\right) \\ &\le \mathbf{P}\,(t \le t_\alpha) = \alpha \end{aligned}$$

Portanto,

$$R = \{z \in \mathbf{R} \text{ t.q. } z < t_\alpha\}$$

é uma região crítica de um teste com nível de significância $\alpha$. Enfim, aceitamos a hipótese se $z > t_\alpha$ e rejeitamos a hipótese se $z < t_\alpha$.

Se a hipótese é $m = a$, uma região crítica de um teste com nível de significância $\alpha$ é

$$R = \left\{z \in \mathbf{R} \text{ t.q. } |z| > t_{1-\alpha/2}\right\}$$

Aceitamos a hipótese se $-t_{1-\alpha/2} < z < t_{1-\alpha/2}$.

**Comparação de dados.** Outro problema é testar a hipótese $x = y$ a partir das observações $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_k$ de dois observáveis. Se as variâncias são $\mathbf{V}\xi = \sigma_\xi^2$ e $\mathbf{V}\eta = \sigma_\eta^2$, então a variável

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_\xi^2/n + \sigma_\eta^2/k}}$$

tem lei normal $N(0,1)$ na hipótese de que $m_\xi = m_\eta$. Fixado um nível de significância $\alpha$, as tabelas indicam-nos o valor $\phi_{1-\alpha/2}$ tal que $\mathbf{P}\left(|z| \geq \phi_{1-\alpha/2}\right) = \alpha$. Aceitamos a hipótese se $|z| < \phi_{1-\alpha/2}$.

Se as variâncias são desconhecidas (o que acontece praticamente sempre), temos que estimá-las com as variâncias amostrais $S_x^2$ e $S_y^2$. Se

$$S_{\text{tot}}^2 = \frac{1}{n+m-2}\left((n-1)S_x^2 + (k-1)S_y^2\right)$$

então a variável

$$z = \frac{\overline{x} - \overline{y}}{S_{\text{tot}}\sqrt{1/n + 1/k}}$$

tem lei de Student $\mathrm{t}_{n+k-2}$ na hipótese de que $m_\xi = m_\eta$. Fixado um nível de significância $\alpha$, as tabelas indicam-nos o valor $t_{1-\alpha/2}$ tal que $\mathbf{P}\left(|z| \geq t_{1-\alpha/2}\right) = \alpha$. Uma região crítica do teste é

$$R = \left\{z \in \mathbf{R} \text{ t.q. } |z| > t_{1-\alpha/2}\right\}$$

Aceitamos a hipótese se $|z| < t_{1-\alpha/2}$.

**Teste de Fisher-Snedecore.** Há situações em que é importante testar hipóteses sobre a variância de um observável $x$, suposto gaussiano (a variância é uma medida da precisão dos instrumentos do laboratório).

Por exemplo, num modelo em que $\xi$ tem lei $N(m, \sigma^2)$, queremos testar a hipótese $\sigma^2 \leq b^2$. Obtidos os resultados $x_1, x_2, ..., x_n$ das experiências, podemos calcular a variância amostral $S$. A variável $(n-1)\frac{S^2}{\sigma^2}$ tem lei $\chi_{n-1}^2$. Fixado um nível de significância $\alpha$, uma tabela fornece-nos o valor $q_{1-\alpha}$ tal que $\mathbf{P}\left(\chi_{n-1}^2 < q_{1-\alpha}\right) = 1 - \alpha$. Se

$$z = (n-1)\frac{S^2}{b^2}$$

na hipótese $\sigma^2 \leq b^2$ temos

$$\alpha = \mathbf{P}\left((n-1)\frac{S^2}{\sigma^2} > q_{1-\alpha}\right) \geq \mathbf{P}\left(z > q_{1-\alpha}\right)$$

Portanto, aceitamos a hipótese se $z < q_{1-\alpha}$.

Se a hipótese for $\sigma^2 = b^2$, um argumento análogo dá-nos uma região de aceitação $q_{\alpha/2} < z < q_{1-\alpha/2}$.

**Teste do qui-quadrado, ou de Pearson.** O problema é testar um modelo probabilístico (por exemplo, saber se um dado é honesto). As observações $x_1, x_2, ..., x_n$ são valores possíveis de uma sucessão de experiências independentes descritas pela variável aleatória $\xi$. A variável $\xi$ tem valores $1, 2, ..., M$ com densidade discreta $\mathbf{P}(\xi = k) = p_k$. Sejam $f_1, f_2, ..., f_M$ as frequências empíricas em $n$ observações, i.e.

$$f_k = \frac{1}{n}\text{card}\{i \text{ tais que } x_i = k\}$$

e $T_n$ a variável aleatória

$$T_n = n\sum_{k=1}^{M}\frac{(f_k - p_k)^2}{p_k}$$

Um *teorema de Pearson* diz que, quando $n \to \infty$, as variáveis $T_n$ convergem em lei para uma variável $\chi^2_{M-1}$.

Queremos testar a hipótese $\mathbf{P}(\xi = k) = p_k$. Fixado um nível de significância $\alpha$, uma tabela fornece-nos o valor $q_{1-\alpha}$ tal que $\mathbf{P}\left\{\chi^2_{M-1} < q_{1-\alpha}\right\} = 1 - \alpha$. A região de aceitação com nível de significância $\alpha$ é $T_n < q_{1-\alpha}$.

Os estatísticos concordam em dizer que a aproximação de Pearson começa a ser boa (e portanto o teste é significativo) desde que os $np_k$ sejam maiores de 5.

**Outros testes não paramétricos.** Quase todas as receitas elementares da estatística utilizam umas hipóteses acerca da distribuição dos dados observados (tipicamente a hipótese gaussiana).

Por exemplo, um intervalo de confiança $\overline{x} \pm t_{1-\varepsilon/2}(n-1) \cdot S_x/\sqrt{n}$ assume que os dados seguem uma lei gaussiana. Se isto não acontecer e se $n$ não for muito grande, esta estimação não é credível. Seria desejável ter instrumentos que permitam decidir se e quando tais hipóteses são credíveis.

Os testes de Student e de Fisher-Snedecore sobre a resposta a um certo tratamento também utilizam a hipótese gaussiana. Aceitam a hipótese de "falta de eficâcia" desde que a diferença entre as médias $\overline{x} - \overline{y}$ seja da ordem de $\sqrt{S_x^2 + S_y^2}$ e que as variâncias $S_x^2$ e $S_y^2$ sejam comparáveis, mas são completamente insensíveis às outras possíveis diferenças entre a distribuição das respostas $y_k$ e a distribuição dos $x_k$. Seria também desejável ter instrumentos mais sensíveis para decidir se dois conjuntos de dados podem ser considerados estatisticamente homogéneos ou não.

Uma série de testes particularmente potentes, robustos, e sobretudo simples de serem utilizados foram desenvolvidos a partir de resultados de Kolmogorov e Smirnov.

**Teste de Kolmogorov-(Smirnov).** Observados os resultados $x_1, x_2, ..., x_n$ de $n$ experiências, a *função de repartição empírica*, ou *distribuição de frequência acumulada*, ou *curva cumulativa* (em inglês, *empirical distribution function* ou *cumulative fraction function*) é a função $F_n : \mathbb{R} \to [0, 1]$ definida por

$$F_n(t) = \frac{1}{n}\text{card}\{k = 1, 2, ..., n \text{ t.q. } x_k \leq t\}$$

Observe que $F_n$ é uma função crescente que toma valores $0 \leq F_n(t) \leq 1$, e que $F_n(t) = 0$ se $t < \min_k x_k$ e $F_n(t) = 1$ se $t \geq \max_k x_k$.

O problema é testar uma hipótese acerca da distribuição dos dados: "os $x_1, x_2, ..., x_n$ são valores de uma sucessão de v.a.'s i.i.d. com a lei de uma certa variável $\xi$ com função de repartição $F : \mathbb{R} \to [0, 1]$".

Lembre que $F(t)$ é, por definição, a probabilidade de observar um valor $\xi \leq t$. O valor $F_n(t)$ é a proporção de observações com $x_k \leq t$. Portanto, fixado $t$, o produto $nF_n(t)$ é uma variável aleatória com lei binomial $B(n, F(t))$. A lei dos grandes números sugere que, se $n$ é grande, $F_n(t)$ aproxime a probabilidade $\mathbb{P}(\xi \leq t) = F(t)$. O teorema limite central sugere que as flutuações de $F_n(t)$ à volta de $F(t)$ sejam da ordem de

$$|F_n(t) - F(t)| \sim \frac{1}{\sqrt{n}}$$

pois o desvio padrão de $F_n(t) - F(t)$ é $\frac{1}{\sqrt{n}} \cdot \sqrt{F(t) \cdot (1 - F(t))} \leq 1/2\sqrt{n}$. Uma ideia ingénua é: a hipótese é credível se as flutuações $|F_n(t) - F(t)|$ não são superiores a 2 ou 3 vezes $1/2\sqrt{n}$.

A flutuação máxima observada, dita *discrepância*, é definida por

$$D_n = \sup_t |F_n(t) - F(t)|$$

Um facto interessante acerca de $D_n$ é que a sua lei é independente da lei $F$, desde que $F$ seja a função de repartição de uma variável $\xi$ absolutamente contínua. De facto, nesta hipótese $F$ é uma função invertível (no domínio, de probabilidade um, onde é estritamente crescente, ou seja onde a sua densidade $F'$ é estritamente positiva), e a variável aleatória $\eta = F(\xi)$ tem lei uniforme no intervalo $[0,1]$, pois

$$\mathbb{P}(\eta \leq t) = \mathbb{P}(F(\xi) \leq t) = \mathbb{P}\left(\xi \leq F^{-1}(t)\right) = F\left(F^{-1}(t)\right) = t$$

quando $t \in [0,1]$. Se os dados $x_k$ seguem a lei $F$ então os $y_k = F(x_k)$ têm lei uniforme no intervalo $[0,1]$. A função de repartição empírica das variáveis $y_k = F(x_k)$ é

$$\begin{aligned} G_n(t) &= \frac{1}{n}\text{card}\{k = 1,2,...,n \text{ t.q. } y_k \leq t\} \\ &= \frac{1}{n}\text{card}\left\{k = 1,2,...,n \text{ t.q. } x_k \leq F^{-1}(t)\right\} \\ &= F_n\left(F^{-1}(t)\right) \end{aligned}$$

e a discrepância é igual à discrepância das variáveis $x_k$, pois

$$\begin{aligned} \sup_t |G_n(t) - t| &= \sup_t \left|F_n\left(F^{-1}(t)\right) - t\right| \\ &= \sup_{F(t)} |F_n(t) - F(t)| \end{aligned}$$

Isto prova que a lei da discrepância é universal.

Mais interessante é que a lei de $\sqrt{n}D_n$ pode ser estimada, pois um teorema de Kolmogorov diz que quando $n \to \infty$ a lei da variável $\sqrt{n}D_n$ converge. A demonstração acima implica em particular que a lei de $D_n$ pode ser calculada no caso em que $\xi$ é suposta ter lei uniforme no intervalo $[0,1]$. Neste caso a lei da família de variáveis $t \mapsto \sqrt{n}(G_n(t) - t)$, com $t \in [0,1]$, é assimptótica à lei de um "laço Browniano" de comprimento um, uma família de variáveis $t \mapsto B(t)$ que pode ser pensada como limite contínuo de uma marcha aleatória que começa e termina na origem. Sem entrar em detalhes técnicos, resulta que a lei de $\sup_{0\leq t\leq 1}|B(t)|$ pode ser calculada (embora de uma maneira não elementar), e em particular

$$\mathbb{P}\left(\sup_{0\leq t\leq 1}|B(t)| > d\right) = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2d^2}$$

Na prática isto quer dizer que temos uma estimação dos valores $d_\varepsilon$ tais que $\mathbb{P}(\sqrt{n}D_n > d_\varepsilon) = \varepsilon$, válida quando $n$ é suficientemente grande.Isto sugere um método para testar a hipótese "os dados têm a lei de $\xi$" com nível de significância $\varepsilon$: uma região crítica é

$$\sqrt{n}D_n > d_\varepsilon$$

Os limites inferiores $d_\varepsilon$ da região crítica para os valores 10%, 5% e 1% do nível de significância (se $n$ é suficientemente grande, da ordem de algumas dezenas), são

$$d_{0.10} \simeq 1.22 \qquad d_{0.05} \simeq 1.36 \qquad d_{0.01} \simeq 1.63$$

**Teste de (Kolmogorov)-Smirnov.** Um problema muito parecido é decidir se "dois conjuntos de observações, $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_m$, são descritos pela mesma distribuição".

Sejam

$$F_n(t) = \frac{1}{n}\text{card}\{k = 1,2,...,n \text{ t.q. } x_k \leq t\} \qquad \text{e} \qquad G_m(t) = \frac{1}{m}\text{card}\{k = 1,2,...,m \text{ t.q. } y_k \leq t\}$$

as funções de repartição empíricas dos dados $x_k$ e $y_k$, respectivamente. A lei dos grandes números sugere que, se $n$ e $m$ são grandes, $F_n(t)$ e $G_m(t)$ sejam próximos. O teorema limite central sugere que as flutuações sejam da ordem de

$$|F_n(t) - G_m(t)| \sim \sqrt{\frac{1}{n} + \frac{1}{m}}$$

A *discrepância* neste caso é definida como

$$D_{n,m} = \sup_t |F_n(t) - G_m(t)|$$

Também a lei de $D_{n,m}$, na hipótese de que os dois conjuntos de dados têm a mesma lei, é independente da tal lei! Um teorema de Smirnov, que generaliza o teorema de Kolmogorov, diz que também a lei de $\sqrt{\frac{nm}{n+m}}D_{n,m}$ converge para a lei de $\sup_{0\leq t\leq 1}|B(t)|$. Portanto, a região crítica de um teste sobre a nossa hipótese com nível de significância $\varepsilon$ é

$$\sqrt{\frac{nm}{n+m}}D_{n,m} > d_\varepsilon$$

Observe que este teste não utiliza nenhuma hipótese acerca da lei.

# 18 Modelização dos dados

**Modelização.** Uma lei física é uma relação entre um certo número de observáveis. Um exemplo muito geral é

$$y = f(\mathbf{x}, \mathbf{a})$$

onde $y$, $\mathbf{x} = (x_1, x_2, ..., x_l)$ e $\mathbf{a} = (a_1, a_2, ..., a_M)$ são certos observáveis. Uma experiência típica consiste em observar os valores $y_1, y_2, ..., y_n$ correspondentes a um certo número de valores $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ de $\mathbf{x}$, considerada como variável independente sobre a qual temos um bom controlo, e portanto nenhum erro significativo. O objectivo da experiência é estimar os valores dos "parâmetros livres" $\mathbf{a}$ que mais concordam com as observações, e decidir se a lei, i.e. a forma da função $f$, descreve bem os resultados da experiência.

**Princípio da máxima verosimilhança.** Uma hipótese de trabalho razoável é assumir que cada $y_i$ tem lei normal com esperança $f(\mathbf{x}_i, \mathbf{a})$ e variância $\sigma_i^2$ (por exemplo estimada a partir da variância amostral, se para cada $\mathbf{x}_i$ temos muitas observações de $y_i$). Neste caso, a densidade de probabilidade de obter o resultado $y_i$ é

$$p(y_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y_i - f(\mathbf{x}_i, \mathbf{a}))^2}{\sigma_i^2}}$$

Na hipótese de que as diferentes observações são independentes, a densidade de probabilidade de obter os resultados $y_1, y_2, ..., y_n$ é proporcional a

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - f(\mathbf{x}_i, \mathbf{a}))^2}{\sigma_i^2}\right)$$

O *princípio da máxima verosimilhança* é uma receita que consiste em escolher os parâmetros livres de maneira tal que a densidade acima seja a maior possível, o que é equivalente a escolher os estimadores $\alpha$ para os parâmetros $\mathbf{a}$ de maneira tal que a soma dos "erros quadráticos"

$$\sum_{i=1}^{n}\frac{(y_i - f(\mathbf{x}_i, \mathbf{a}))^2}{\sigma_i^2}$$

seja a menor possível (daqui o nome de "least-square fitting"). Em teoria, desde que a função $f$ seja diferenciável, os valores de $\alpha$ são obtidos calculando derivadas parciais e resolvendo um sistema de $M$ equações. Na prática, se a forma de $f$ não é simples, este é um problema difícil. O melhor é procurar soluções aproximadas, por exemplo utilizando técnicas de análise numérica.

**Teste do qui-quadrado.** O valor de

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - f(\mathbf{x}_i, \alpha))^2}{\sigma_i^2} = \min_{\mathbf{a}}\sum_{i=1}^{n}\frac{(y_i - f(\mathbf{x}_i, \mathbf{a}))^2}{\sigma_i^2}$$

é uma medida da bontade do modelo. De facto, se $\alpha$ são os valores verdadeiros dos $\mathbf{a}$, então a hipótese gaussiana implica que $\chi^2$ tem lei qui-quadrado $\chi^2_{n-M}$. Uma tabela dos quantis da lei $\chi^2_{n-M}$ fornece a probabilidade

$$Q = \mathbf{P}\left(\chi^2_{n-M} > \chi^2\right)$$

Os físicos consideram aceitáveis valores $Q \geq 0.1$ (esta regra é equivalente a aceitar a hipótese "a lei $y = f(\mathbf{x}, \alpha)$ é verdadeira" com nível de significância da ordem de 10%, um valor típico de um teste acerca de uma hipótese conservadora). Por outro lado, valores grandes de $\chi^2$, por exemplo $Q \ll 0.01$, são fortes indícios de que a conjectura $f$ não é uma lei que descreve bem os dados observados.

Se as observações de $y_i$, para cada valor $\mathbf{x}_i$, são poucas, não temos uma estimação credível das variâncias $\sigma_i^2$. O que os físicos fazem nesse caso é pôr as variâncias iguais a 1 nas fórmulas acima, e depois estimar

$$\sigma_i^2 \simeq \frac{\chi^2}{n - M}$$

(o que significa fazer a hipótese de que as $\sigma_i^2$ são todas iguais). A partir destas variâncias é possível, usando a fórmula da propagação dos erros, estimar os erros nos parâmetros $\alpha$.

**Que funções testar.** É bom lembrar que a lei não é ditada pelos deuses: pode ser uma previsão de uma teoria física que queremos testar, ou simplesmente uma conjectura sugerida pelos resultados da experiência. É claro que, na segunda hipótese, uma função $f$ suficientemente irregular e um número muito grande de parâmetros livres **a** permite "ajustar" com óptima precisão qualquer dado (basta que $f$ seja um polinómio de grau muito grande!): é costume entre os físicos experimentar leis simples, possivelmente com poucos parâmetros livres...

**Modelos lineares.** Modelos que são tratáveis analiticamente são os modelos lineares, onde a lei é da forma

$$y = \sum_{j=1}^{M} a_j f_j(\mathbf{x})$$

Os valores $\alpha$ de **a** que minimizam o erro quadrático médio são obtidos calculando o zero das derivadas parciais em ordem aos $a_j$. O resultado é

$$\alpha = \mathbf{C} \cdot \mathbf{h}$$

onde $\mathbf{C} = \mathbf{A}^{-1}$, $\mathbf{A}$ é a matriz $M \times M$ com entradas

$$\mathbf{A}_{kj} = \sum_{i=1}^{n} \frac{f_k(\mathbf{x}_i) f_j(\mathbf{x}_i)}{\sigma_i^2}$$

(é muito azar observar uma matriz não invertivel!) e $\mathbf{h}$ é o vector de componentes

$$h_k = \sum_{i=1}^{n} \frac{y_i f_j(\mathbf{x}_i)}{\sigma_i^2}$$

Uma estimação do desvio padrão dos $\alpha_j$ e das covariâncias entre eles é fornecida pela fórmula da propagação dos erros, e é

$$\sigma^2_{\alpha_j} = \mathbf{C}_{kk} \qquad \text{e} \qquad \sigma^2_{\alpha_j \alpha_k} = \mathbf{C}_{jk}$$

**Regressão linear.** Um exemplo simples é uma lei linear $y = a + bx$ entre os observáveis $x$ e $y$. O objectivo de uma experiência pode ser: ajustar os parâmetros $a$ e $b$, e testar a validade da lei. Repetimos $n$ vezes a experiência e obtemos os resultados $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$. A primeira coisa que fazem os físicos é traçar os pontos $(x_k, y_k)$ no plano, e observar se estão todos perto de uma recta.

Um modelo da experiência é: em cada observação, $y_k$ é uma variável aleatória igual a $a + bx_k + \text{erro}_k$, e os "erros" são independentes e têm lei normal $N(0, \sigma^2)$ (o valor médio dos erros é nulo porque julgamos que a experiência é bem feita, i.e. não estamos à espera de "erros sistemáticos"). Uma receita razoável para estimar os parâmetros (que os estatísticos chamam *método dos mínimos quadrados*) é escolher $a$ e $b$ de maneira tal que a soma dos quadrados dos erros $\sum \text{erro}_k^2 = \sum (a + bx_k - y_k)^2$ seja a menor possível. Derivando, obtemos a resposta $y = \alpha + \beta x$, onde

$$\beta = \frac{\overline{\sigma}^2_{xy}}{\overline{\sigma}^2_x} \qquad \alpha = \overline{y} - \beta \overline{x}$$

e

$$\overline{\sigma}^2_{xy} = \sum (x_k - \overline{x})(y_k - \overline{y}) \quad \text{e} \quad \overline{\sigma}^2_x = \sum (x_k - \overline{x})(x_k - \overline{x})$$

No modelo (em que os erros são independentes e têm lei normal) $\alpha$ e $\beta$ são bons estimadores de $a$ e $b$ respectivamente, porque $\alpha$ tem lei $N\left(a, \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\overline{\sigma}^2_x}\right)\right)$ e $\beta$ tem lei $N\left(b, \sigma^2/\overline{\sigma}^2_x\right)$. Naturalmente não sabemos o valor de $\sigma^2$, mas um seu estimador é a *variância residual*

$$S^2 = \frac{1}{n-2} \sum (\alpha + \beta x_k - y_k)^2$$

A variável $\frac{n-2}{\sigma^2} \cdot S^2$ tem lei qui-quadrado $\chi^2_{n-2}$. O resultado final é que

$$\frac{\alpha - a}{S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\overline{\sigma}^2_x}}} \quad \text{e} \quad \frac{\beta - b}{S/\overline{\sigma}_x}$$

têm lei de Student $\mathrm{t}_{n-2}$. Intervalos de confiança de nível $1-\varepsilon$ pelos parâmetros da lei são

$$a = \alpha \pm t_{1-\varepsilon/2} \cdot S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\overline{\sigma}_x^2}} \qquad \text{e} \qquad b = \beta \pm t_{1-\varepsilon/2} \cdot \frac{S}{\overline{\sigma}_x}$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student $\mathrm{t}_{n-2}$.

Este modelo estima os valores "mais prováveis" de $a$ e $b$, e portanto a lei na forma da *recta de regressão*

$$y = \alpha + \beta x$$

Como decidir que $y = a + bx$ é mesmo uma lei, ou seja que o observável $y$ depende do observável $x$? Uma boa ideia é testar a hipótese $b = 0$, ou seja "o observável $y$ é independente do observável $x$". Fixado um nível de significância $\varepsilon$, a região de aceitação da hipótese é

$$\left| \beta \cdot \frac{\overline{\sigma}_x}{S} \right| < t_{1-\varepsilon/2}$$

se $t_{1-\varepsilon/2}$ é o quantil da lei de Student $\mathrm{t}_{n-2}$. Portanto, admitimos que a variável $y$ depende de $x$ se encontramos um valor de $\beta$ maior que $t_{1-\varepsilon/2} \cdot \frac{S}{\overline{\sigma}_x}$. Para valores típicos do nível de significância este limite é da ordem de duas ou três vezes a razão entre as incertezas nas variáveis $(y_k - \alpha + \beta x_k)$ e $x_k$, o que é muito razoável.

A falta de simetria das fórmulas acima reflecte o facto de considerar $x$ como variável independente da lei $y = a + bx$. A regressão tipicamente é utilizada quando temos um bom controlo do observável $x$, e por isto podemos pensar que os erros na sua determinação são desprezáveis. Caso contrário, ao escrever a lei na forma $x = a' + b'y$, o argumento acima produz a recta de regressão $x = \alpha' + \beta' y$, onde agora

$$\beta' = \frac{\overline{\sigma}_{xy}^2}{\overline{\sigma}_y^2} \qquad \text{e} \qquad \alpha' = \overline{x} - \beta' \overline{y}$$

A relação teórica entre $b$ e $b'$ é $bb' = 1$. É razoável suspeitar que observar um número $\beta\beta'$ pequeno, mais perto de 0 que de 1, é indício de que alguma coisa não está a correr bem. A raiz deste número é dita coeficiente de correlação empírico, e fornece uma outra maneira de testar a validade da lei.

**Linearização.** Uma lei não linear pode ficar linear depois de uma mudança de variável, por exemplo a lei $y = ae^{bx}$ é $\log y = \log a + bx$. Se os instrumentos medem os $y_k$ em escala logarítmica, a regressão linear fornece uma estimação correcta de $\log a$ e $b$. Caso contrário, é de se esperar que os erros, definidos por $y_k = ae^{bx_k} + \mathrm{erro}_k$, não sejam identicamente distribuídos, e modelos mais cuidadosos são necessários para estimar os parâmetros de regressão.

**Correlação.** Uma medida empírica da "correlação linear" entre $x$ e $y$ é o *coeficiente de correlação empírico*

$$\rho = \frac{\overline{\sigma}_{xy}^2}{\overline{\sigma}_x \overline{\sigma}_y} = \frac{\sum (x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum (x_k - \overline{x})^2}\sqrt{\sum (y_k - \overline{y})^2}}$$

Um valor de $\rho$ perto de $\pm 1$ é indício de correlação linear efectiva entre as variáveis. Um valor de $\rho$ perto de 0 é indício que as variáveis podem ser independentes (e portanto $y = a + bx$ não é uma lei da física!). Desenhando no plano os pontos de coordenadas

$$\left( \frac{(x_k - \overline{x})}{\overline{\sigma}_x}, \frac{(y_k - \overline{y})}{\overline{\sigma}_y} \right)$$

a primeira situação corresponde a ter pontos mais concentrados num dos quatro quadrantes (parecem mesmo seguir uma recta!), e a segunda a ter pontos uniformemente espalhados na bola de raio 1.

O coeficiente de correlação empírico fornece uma outra maneira de testar a hipótese "o observável $y$ é independente do observável $x$". Os livros de estatística contêm tabelas das probabilidades de duas amostras de tamanho $n$ de variáveis aleatórias independentes com lei normal ter coeficiente de correlação $\geq \delta$.

Por exemplo, se $n = 10$ as tabelas dizem que

$$\mathbf{P}(\rho \geq 0.55) \simeq 0.1 \quad \text{e} \quad \mathbf{P}(\rho \geq 0.76) \simeq 0.01$$

Portanto, a hipótese é rejeitada (logo a lei é aceite) com nível de significância 10% se é observado um coeficiente de correlação $\rho > 0.55$. A hipótese é rejeitada com nível de significância 1% se é observado um coeficiente de correlação $\rho > 0.76$.

# References


[BR92] P.R. Bevington and D.K. Robinson, *Data reduction and error analysis for the physical science,* McGraw-Hill, New York 1992.

[Bi68] P. Billingsley, *Convergence of probability measures,* J. Wiley & Sons, New York 1968.

[Bi79] P. Billingsley, *Probability and measure,* J. Wiley & Sons, New York 1979.

[Br68] L. Breiman, *Probability*, Addison-Wesley, Reading, MA 1968.

[DF91] B. De Finetti, *Theory of probability: a critical introduction treatment*, John Wiley & Sons, Chichester 1991.

[Do53] J.L. Doob, *Stochastic processes*, John Wiley, New York 1953.

[Do94] J.L. Doob, *Measure theory*, Springer, New York 1994.

[El85] R. Ellis, *Entropy, large deviations, and statistical mechanics*, Springer, New York 1985.

[Fe68] W. Feller, *An introduction to probability theory and its applications, vol. 1 & 2,* John Wiley & Sons, New York 1968.

[Fe63] R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman lectures on physics*, Addison-Wesley, Reading, 1963.

[Ga99] G. Gallavotti, *Statistical mechanics, a short treatise*, Springe-Verlag, 1999.

[Gn73] B.V. Gnedenko, *The theory of probability,* Mir, Moscow 1973.

[Ha74] P. Halmos, *Measure theory*, Springer-Verlag, New York 1974.

[JP00] J. Jacod and P. Protter, *Probability essentials,* Springer, Berlin 2000.

[Ka57] M. Kac, *Probability and related topics in physical sciences*, Lectures in Applied Mathematics, Interscience Publishers, New York 1957.

[Kh57] A.I. Khinchin, *Mathematical foundations of statistical mechanics*, Dover, New York 1957.

[Kh57'] A.I. Khinchin, *Mathematical foundations of information theory*, Dover, New York 1957.

[Ko33] A.N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitrechnung*, Ergebnisse Der Mathematik, Berlin 1933.

[KF82] A.N. Kolmogorov and S.V. Fomin, *Elementos da teoria das funções e de análise funcional,* MIR, Moscou 1982.

[La66] J. Lamperti, *Probability*, Benjamin, New York 1966.

[La77] J. Lamperti, *Stochastic processes*, Springer-Verlag, New York 1977.

[LL] L.D. Landau and E.M.Lifshitz, *Statistical physics*,

[Lo55] M. Loève, *Probability theory*, Springer, New York 1955.

[Mc01] G.C. McBane, *Treatment of experimental data in the physical chemistry laboratory*, lecture notes Chemistry 353/355/455 (the "green book"), http://faculty.gvsu.edu/mcbaneg/greenbook.pdf 2001.

[MGB74] A.M. Mood, F.A. Graybill and D.C. Boes, *Introduction to the theory of statistics,* McGraw-Hill, New York 1974.

[Pa67] K. Parthasarathy, *Probability measures on metric spaces*, Academic Press, New York 1967.

[PV02] D.D. Pestana e S.F. Velosa, *Introdução à probabilidade e à estatística*, Fundação Calouste Gulbenkian, Lisboa 2002.

[Re70] A. Rényi, *Probability theory*, North-Holland, Amsterdam 1970.

[Ru91] D. Ruelle, *Chance and chaos,* Princeton University Press, Princeton N.J. 1991.

[Rud66] W. Rudin, *Real and complex analysis*, McGraw-Hill, New York 1966.

[Sh96] A.N. Shiryaev, *Probability*, Springer-Verlag, New York 1996.

[Si76] Ya.G. Sinai, *Introduction to ergodic theory*, Princeton University Press, Princeton 1976.

[Yo62] H.D. Young, *Statistical treatment of experimental data*, McGraw-Hill, New York 1962.