Notas das aulas de Métodos Estatísticos

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

3 de Março de 2006

Contents

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" Δx.

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 "desvios" 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 "desvios 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*, ou *standard uncertainty),* 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 α 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 α 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 "desvio 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 α 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 χ^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 χ^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, 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.

É 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 da função 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 $\sqrt{n} \cdot 1/\sqrt{n} \sim 1$,

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

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 "normal" ou "gaussiana" ao crescer n, no sentido em que

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

quando $n \to \infty$.

Flutuações da marcha aleatória. 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

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

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

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

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 $\sqrt{n}$.

Lei dos grandes números. Em termos da frequência de caras redescobrimos a conjectura de que

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

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

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

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 $1/\sqrt{n}$.

2 Espaços de probabilidades

Ingredientes. Os ingredientes de um modelo probabilístico são:

• um *espaço dos estados*, ou acontecimentos, Ω

• uma família $\mathcal{E}$ de *eventos*, que são subconjuntos de Ω

• uma *probabilidade*, ou seja uma função $\mathbb{P} : \mathcal{E} \to [0,1]$ que associa um número $\mathbb{P}(A)$ a cada evento $A \in \mathcal{E}$

• uns observáveis, ou seja funções $\xi : \Omega \to \mathbb{R}$, ditas *variáveis aleatórias*.

Estes ingredientes satisfazem certas propriedades naturais, codificadas em "axiomas" da teoria das probabilidades.

Experiências e álgebras. No dialecto dos probabilistas, Ω representa o espaço dos estados de um sistema físico. Ao fazer uma experiência, o que fazemos é medir "observáveis", funções $\xi : \Omega \to \mathbb{R}$. 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 Ω, fechada com respeito às operações binárias $\cap$, $\cup$ e $\backslash$, e que contém os elementos neutros $\emptyset$ e Ω, 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 $\mathcal{A}$.

Eventos. Seja Ω um conjunto não vazio. Uma família $\mathcal{E}$ de subconjuntos de Ω é uma *σ-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

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

Um par $(\Omega, \mathcal{E})$, formado por um conjunto não vazio Ω e uma σ-álgebra $\mathcal{E}$ de partes de Ω, 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.

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

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

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

$$\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$$

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

$$\mathbb{P}(A_1 \cup A_2 \cup ... \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + ... + \mathbb{P}(A_n)$$

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

Probabilidades em espaços finitos ou enumeráveis. Sejam $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_k, ...\}$ um conjunto finito ou enumerável, e $\mathcal{E} = \mathcal{P}(\Omega)$ a família dos subconjuntos de Ω. Se $p_1, p_2, ...p_k...$ é uma coleção de números ≥ 0 tais que $\sum_k p_k = 1$, então a função $\mathbb{P} : \mathcal{E} \to [0,1]$ definida por

$$\mathbb{P}(A) = \sum_{\omega_k \in A} p_k$$

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

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

Se $A \in \mathcal{E}$, o número $\mathbb{P}(A)$ 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. Ω 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 $\mathbb{P}(A) = 1$. 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 $\mathbb{P}(\{\omega\}) = 0$.

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

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

De facto, a primeira vem da normalização $\mathbb{P}(\Omega) = 1$ 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 $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B) \geq \mathbb{P}(A)$, porque as probabilidades são não-negativas. A σ-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 $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$ porque $B_n \subset A_n$, portanto a σ-aditividade e a monotonia implicam que $\mathbb{P}(\cup_n A_n) = \mathbb{P}(\cup_n B_n) = \sum_n \mathbb{P}(B_n) \leq \sum_n \mathbb{P}(A_n)$.

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

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

Continuidade. A consequência importante da σ-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.

Uma sucessão (A_n) de subconjuntos de Ω 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).

A medida de probabilidade é *contínua*, ou seja

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

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 σ-aditividade temos enfím

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

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

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

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:

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

Espaços de probabilidades uniformes. Se Ω é um conjunto finito, e seja $|\Omega| = n$ a sua cardinalidade. Uma probabilidade natural sobre as suas partes é

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

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 $\mathbb{P}(\{\omega\}) = 1/n$. 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 theorem can be proved, some of which suggest real-world observational results" (J.L. Doob, *Measure Theory*, Springer-Verlag, New York 1994).

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 Ω dos caminhos $\omega : \{0, 1, 2, ..., k\} \to X$ tais que $\omega(0) = 0$ e $\omega(i) \neq \omega(i-1)$ se $i = 1, 2, ..., k$. A cardinalidade de Ω 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 Ω formados pelos caminhos ω tais que $\omega(i) \neq \omega(j)$ 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 Ω", a resposta é

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

se $k \leq n$ e $\mathbb{P}(A) = 0$ 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 $\mathbb{P}$ sobre as suas partes. Em particular, o evento "obter uma cara e uma coroa", ou seja $A = \{(H, T), (T, H)\}$, tem probabilidade $\mathbb{P}(A) = 1/2$. 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 Ω. Isto não obriga a mudar a medida de probabilidade, basta dizer que agora a probabilidade é a restrição de $\mathbb{P}$ 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 ≤ 4.

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

d. Quantos filhos deve ter um casal de modo a ter, com probabilidade ≥ 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 $(\Omega, \mathcal{E}, \mathbb{P})$ um espaço de probabilidades e B um evento com $\mathbb{P}(B) > 0$. A *probabilidade condicionada* do evento A com respeito ao evento B é definida por

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

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

Saber que um evento aconteceu é uma informação que, em geral, muda a nossa expectativa acerca dos outros. 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 $\mathbb{P}_B : A \mapsto \mathbb{P}(A\,|B)$ é uma probabilidade sobre $\mathcal{E}$, e $\mathbb{P}_B(B) = 1$.

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

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

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,

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

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

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

Se $B_1, B_2, B_3, ...$ é uma partição de Ω, 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

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

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 $\mathbb{P}(A) > 0$,

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

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

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

Os eventos B_n têm a interpretação de hipóteses, e a probabilidade condicionada $\mathbb{P}(B_n\,|A)$ 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 $\mathbb{P}(B\,|A)$, $\mathbb{P}(B\,|A^c)$, $\mathbb{P}(A)$ e $\mathbb{P}(B)$.

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

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

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

A interpretação é a seguinte: se $\mathbb{P}(B) > 0$, os eventos A e B são independentes sse $\mathbb{P}(A\,|B) = \mathbb{P}(A)$ (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

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

Exercícios.

a. O evento A é independente de A sse $\mathbb{P}(A) = 0$ 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 $\mathbb{P}(A \cap B) > 0$. Então

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

implica que

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

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

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

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

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

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

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.

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 $(\Omega_d, \mathcal{P}(\Omega_d), \mathbb{P}_d)$ o modelo do lançamento de um dado, e $(\Omega_m, \mathcal{P}(\Omega_m), \mathbb{P}_m)$ 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

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

define uma probabilidade $\mathbb{P} : \mathcal{P}(\Omega) \to [0,1]$, dita *probabilidade produto*, sobre as partes de Ω 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 $\mathbb{P}$ é uma probabilidade). Esta receita corresponde a multiplicar as probabilidades dos átomos dos dois espaços: se os átomos de Ω_d e de Ω_m têm probabilidades $p_i^d = \mathbb{P}_d\left(\{\omega_i^d\}\right)$ e $p_j^m = \mathbb{P}_m\left(\{\omega_j^m\}\right)$, então os átomos do produto cartesiano têm probabilidades $\mathbb{P}\left(\left\{\left(\omega_i^d, \omega_j^m\right)\right\}\right) = p_i^d \cdot p_j^m$.

Se $\mathbb{P}_d$ e $\mathbb{P}_m$ são as probabilidades uniformes em Ω_d e Ω_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 $1/|\Omega|$.

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 $n \in \mathbb{N}$. 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.

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,

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

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 Ω 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 Ω verifica $\mathbb{P}(A_{i,j}) = 1/n$, 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

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

Uma boa aproximação de $\mathbb{P}(A^c)$, 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 $\mathbb{P}(A) > 0.50$, se $n = 365$ e $k \geq 64$ então $\mathbb{P}(A) > 0.99$.

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 $\mathbb{P}(A_i) = p$ para todo $i = 1, 2, ..., n$ " define uma probabilidade $\mathbb{P}$ 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

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

Verificar os axiomas é um exercício de álgebra, aliás, $\mathbb{P}$ é a probabilidade produto em $\{0,1\}^n$, onde cada factor $\{0,1\}$ é munido da probabilidade "$\mathbb{P}_i(\{1\}) = p$ e $\mathbb{P}_i(\{0\}) = q$". 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 Ω^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

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

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 Ω^n, pois cada palavra $\omega \in \Omega^n$ tem probabilidade $\mathbb{P}(\{\omega\}) = 2^{-n}$.

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 $\mathbb{P}$ nas partes de Ω^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

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

onde $k_i(\omega)$, com $i \in X$, denota o número de vezes que a letra x_i está contida na palavra ω. 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

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

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?

Marcha aleatória. Um homenzinho passeia dentro dos inteiros $\mathbb{Z}$ 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 $\mathbb{P} : \mathcal{E} \to [0, 1]$ o esquema de Bernoulli determinado por $\mathbb{P}(\omega_i = 1) = p$ e $\mathbb{P}(\omega_i = -1) = 1 - p$. 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 $\mathbb{P}$ pode ser pensada como uma probabilidade no espaço das trajectórias da marcha aleatória, definido por

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

Particularmente interessante é a *marcha simétrica*, quando $p = 1/2$ e portanto $\mathbb{P}$ é a probabilidade uniforme nas partes de Ω': 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.

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

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

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

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

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

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

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

5 Variáveis aleatórias, leis

Variáveis aleatórias. Seja $(\Omega, \mathcal{E}, \mathbb{P})$ um espaço de probabilidades. Se Ω é um modelo dos possíveis estados de um sistema físico, os observáveis da física são funções reais ξ definidas em Ω. 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.

Uma *variável aleatória* (com valores na recta real) é uma função $\xi : \Omega \to \mathbb{R}$ tal que

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

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

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 $\mathbb{P}(\xi \in A)$ ou $\mathbb{P}\{\xi \in A\}$ em vez de $\mathbb{P}(\{\omega \in \Omega \text{ t.q. } \xi(\omega) \in A\})$.

Se Ω é um conjunto finito ou enumerável, e $\mathcal{E}$ é a família das suas parte, então toda função $\xi : \Omega \to \mathbb{R}$ é uma variável aleatória.

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 $\xi : \Omega \to \mathbb{R}$ que assume uma quantidade finita de valores 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 $S_1, S_2, ..., S_n$ é uma partição de Ω com $S_k \in \mathcal{E}$, e $x_1, x_2, ..., x_n$ são números reais.

Combinações lineares, funções arbitrárias, assim como produtos e quociêntes (desde que sejam definidos), de variáveis aleatórias simples são variáveis aleatórias simples.

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

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

onde $S_1, S_2, ..., S_k, ...$ é uma partição enumerável de Ω, e $x_1, x_2, ..., x_k, ...$ são números reais, os seus valores. 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

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

Toda função $\xi : \Omega \to \{x_1, x_2, x_3, ...\} \subset \mathbb{R}$ 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 $x \in \mathbb{R}$.

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

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

(a "probabilidade da variável ξ ser menor ou igual a x").

A *lei* da variável aleatória ξ é a função $\mathbb{P}_\xi$ que associa a um intervalo $A \subset \mathbb{R}$ a probabilidade

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

(a "probabilidade da variável ξ pertencer a A").

A relação entre a lei e a função de repartição de uma variável aleatória ξ é a seguinte:

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

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

Funções de repartição. A função de repartição F_ξ 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

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

(onde utilizamos a monotonia de F_ξ e a continuidade da medida de probabilidades),

iii) admite o limite à esquerda, porque

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

(onde utilizamos a monotonia de F_ξ 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 $F : \mathbb{R} \to [0,1]$ com estas propriedades é dita uma *função de repartição*.

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

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

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

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

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

(a "probabilidade da variável ξ ser igual a x_k").

A densidade discreta de ξ determina (e é determinada por) a lei de ξ, pois

$$\mathbb{P}(\xi \in A) = \mathbb{P}(\cup_{x_k\in A}\{\xi = x_k\}) = \sum_{x_k\in A} \mathbb{P}(\xi = x_k)$$

para todo intervalo $A \subset \mathbb{R}$, 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

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

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

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

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 $\Omega = \{x_1, x_2, x_3, ...\}$, $\mathcal{E} = \mathcal{P}(\Omega)$, $\mathbb{P}(A) = \sum_{x_k \in A} p(x_k)$ para todo $A \subset \Omega$, e $\xi : \Omega \to \mathbb{R}$ 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 ξ 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 ξ com valores em $\mathbb{N}$ e função de repartição $F_\xi(k) = 1 - p^k$ se $k \in \mathbb{N}$, onde $p \in]0,1[$..

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

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

c. Seja ξ uma variável aleatória com função de repartição F_ξ. 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 $a, b \in \mathbb{R}$ e $k \in \mathbb{N}$.

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

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

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

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

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 $(\Omega, \mathcal{E}, \mathbb{P})$, onde T é um subconjunto da recta real. O parâmetro $t \in T$ neste caso tem a interpretação de um "tempo", e tipicamente $T = \mathbb{R}$, ou $\mathbb{Z}$, ou $\mathbb{R}_{\ge 0}$, ou $\mathbb{N}$. A cada ponto $\omega \in \Omega$ está associada uma *trajectória* $\xi(\omega) : T \to \mathbb{R}$, definida por

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

e a probabilidade $\mathbb{P}$ pode ser pensada como uma probabilidade definida no 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 $\mathbb{N}$, $\mathbb{Z}$ ou em outros retículos como por exemplo $\mathbb{Z}^n$.

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

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

A sucessão de variáveis aleatórias (ξ_n) é uma *sucessão de variáveis independentes* se, para cada $n \in \mathbb{N}$, 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.

V.a.'s i.i.d. Tem interesse, sobretudo para formular teoremas de convergência, considerar sucessões $(\xi_n)_{n\in\mathbb{N}}$ de variáveis aleatórias tais que: são definidas num mesmo espaço de probabilidades, têm todas a mesma lei (a saber, a lei de uma variável ξ 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.".

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 $p : \xi(\Omega) \times \eta(\Omega) \times ... \times \varsigma(\Omega) \to [0, 1]$ definida por

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

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

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

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

A função $(\xi, \eta, ..., \varsigma) : \Omega \to \mathbb{R}^n$ 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 ξ é independente de si mesma sse é constante com probabilidade um, i.e. se existe $a \in \mathbb{R}$ tal que $\mathbb{P}(\xi = a) = 1$.

b. *(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

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

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 $\mathbb{P}(\xi = i\ ,\ \eta = j) = p(i, j) = 1/12$ para todos $i = 1, 2, ..., 6$ e $j = 0, 1$. Mostre que as variáveis ξ e η 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 ξ =“número da primeira bolinha” e η =“número da segunda bolinha”. As variáveis ξ e η 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 ξ e η, defnidas como acima, não são independentes, e a densidade conjunta é $p(i, j) = 1/30$ se $i \neq j$ e 0 se $i = j$.

Mostre que as densidades (marginais) de ξ e η são iguais nas duas experiências!

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

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

6 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 $(\Omega, \mathcal{E}, \mathbb{P})$. O *valor médio* (ou *média,* ou *esperança*) de ξ é

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

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

$$\sum_k |x_k| \cdot \mathbb{P}(S_k) < \infty$$

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

$$\mathbb{E}\xi = \sum_k x_k \cdot \mathbb{P}(S_k)$$

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

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

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

Porque a esperança se chama esperança? Se ξ 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 $\mathbb{E}\xi$.

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

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

uma função que associa um valor $\mathbb{E}\xi$ a cada variável integrável ξ. 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

$$\mathbb{E}1_A = \mathbb{P}(A)$$

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

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

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

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

$$\mathbb{E}(a \cdot \xi + b) = a \cdot \mathbb{E}\xi + b$$

e se ξ e η são integráveis então

$$\mathbb{E}(\xi + \eta) = \mathbb{E}\xi + \mathbb{E}\eta$$

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

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

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

$$|\mathbb{E}\xi| \leq \mathbb{E}|\xi|$$

Exercícios.

a. Prove as propriedades da média enunciadas acima.

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

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

onde $\{y_1, y_2, y_3...\} = \varphi(\{x_1, x_2, x_3, ...\})$ é o conjunto dos valores de η.

Em geral é falso que se ξ é integrável também η é, assim como é falso que $\mathbb{E}\eta$ seja igual a $\varphi(\mathbb{E}\xi)$. Se η é integrável, podemos calcular $\mathbb{E}\eta$ a partir da densidade discreta de ξ, pois, dado que a série é absolutamente convergente,

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

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 $\mathbb{P}(\xi = x_k) = 1/n$ para todo $k = 1, 2, ..., n$. Verifique que a média de ξ é a média aritmética dos seus valores, ou seja

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

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 $\mathbb{P}(A) = \mathbb{E}1_A$, 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 ξ uma variável aleatória discreta com valores em $\mathbb{N}$. Prove que

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

Momentos e variância. Seja ξ uma variável aleatória discreta. Se ξ^k é integrável, podemos definir o *momento* de grau k da variável aleatória ξ como sendo $\mathbb{E}\xi^k$. Se ξ^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 $x \in \mathbb{R}$). Mais interessantes são os momentos centrados, definidos por $\mathbb{E}(\xi - \mathbb{E}\xi)^k$, porque são invariantes por translações, e ainda mais interessantes os momentos centrados absolutos, definidos por $\mathbb{E}|\xi - \mathbb{E}\xi|^k$.

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

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

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

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

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

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

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

Mais interessante é observar que, se ξ é uma variável aleatória discreta com variância finita, então

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

Ou seja, a variância é o valor mínimo da função $x \mapsto \mathbb{E}\,|\xi - x|^2$, e o valor de x onde o mínimo é atingido é $\mathbb{E}\xi$. 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 ξ é uma variável aleatória discreta e $a, b \in \mathbb{R}$, então

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

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

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

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

Exercícios.

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

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

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

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

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

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

Produtos de variáveis independentes. Se ξ e η são integráveis e independentes, então $\xi\eta$ é integrável e

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

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

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

(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 ξ_1, ξ_2, ...,ξ_n são integráveis e independentes então

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

(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 ξ_n).

Covariância. Sejam ξ e η duas variáveis aleatórias com variância finita. A variância da soma é

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

onde a *covariância* de ξ e η (ou "entre" ξ e η) é definida por

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

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

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

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

Por indução, se as variáveis aleatórias ξ_1, ξ_2, ...,ξ_n são independentes então

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

Correlação. Sejam ξ e η duas variáveis aleatórias com variâncias positivas $\mathbb{V}\xi = \sigma_\xi^2$ e $\mathbb{V}\eta = \sigma_\eta^2$. 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 $a, b, c, d \in \mathbb{R}$ com a e $c \neq 0$.

Da identidade

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

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

O interesse do coeficiente de correlação está na seguinte observação: $\varrho(\xi,\eta) = \pm 1$ sse ξ e η são linearmente dependentes com probabilidade um, ou seja se existem $a, b \in \mathbb{R}$, $a \neq 0$, tais que $\mathbb{P}(\eta = a\xi + b) = 1$. Pois, $\varrho(\xi,\eta) = \pm 1$ implica que

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

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 ξ_n o número de caras e η_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 $\mathbb{V}(\xi_n + \eta_n) = 0$, pois $\xi_n + \eta_n = n$)

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 ξ e η 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 ξ e η. Existe um valor de p tal que ξ e η são independentes?

7 Modelos discretos

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

A média e a variância de ξ são

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

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

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

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

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

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

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

(onde utilizamos a independência das ξ_k).

Exercício.

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.

Determine o máximo da densidade discreta $k \mapsto \mathbb{P}(S_n = k)$.

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 ± 1 e lei definida por $\mathbb{P}(\xi_i = 1) = p$ e $\mathbb{P}(\xi_i = -1) = q$. Em particular, a posição da marcha ao tempo n é igual à posição ao tempo $n - 1$ mais ± 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

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

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

A média e a variância de S_n são

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

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 ξ_k são independentes e têm valores ± 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

$$\mathbb{P}(T_{n+1} = k+1 \,|T_n = k) \qquad \mathbb{P}(T_{n+1} = k+1 \,|T_n = k,\, T_{n-1} = k-1)$$

Tempo de espera: lei geométrica. Seja τ o número de provas de Bernoulli necessárias para obter "sucesso" pela primeira vez. É possível definir a variável aleatória τ 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 $\mathbb{N} \cup \{\infty\}$. O evento $\{\tau = k\}$ com $k \in \mathbb{N}$ 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

$$\mathbb{P}(\tau = k) = (1-p)^{k-1} p$$

Esta é a densidade discreta da variável aleatória "tempo de espera" em infinitas provas de Bernoulli. A lei de τ é dita *lei geométrica*. Uma notação pode ser $\tau \sim$geométrica(p).

Observe que a probabilidade do evento $\{\tau = \infty\}$ é igual a

$$\mathbb{P}(\tau = \infty) = 1 - \mathbb{P}(\tau < \infty) = 1 - \sum_{k=1}^{\infty} (1-p)^{k-1} p = 0$$

desde que p seja diferente de 0, caso pouco interessante em que $\mathbb{P}(\tau = \infty) = 1$.

Em algum livro é chamada "geométrica" também a lei da variável aleatória $\xi = \tau - 1$, com valores $0, 1, 2, ...$ e densidade discreta $\mathbb{P}(\xi = k) = (1-p)^k p$.

A lei geométrica é caracterizada pela propriedade de "falta de memória". Por um lado, observando que

$$\mathbb{P}(\tau > k) = \sum_{n=k+1}^{\infty} (1-p)^{n-1} p = (1-p)^k$$

temos que para todos $k, n \in \mathbb{N}$

$$\begin{aligned}\mathbb{P}(\tau = k+n \,|\, \tau > k) &= \frac{\mathbb{P}(\{\tau = k+n\} \cap \{\tau > k\})}{\mathbb{P}(\tau > k)} \\ &= \frac{\mathbb{P}(\tau = k+n)}{\mathbb{P}(\tau > k)} \\ &= (1-p)^{n-1}p\end{aligned}$$

e portanto

$$\mathbb{P}(\tau = k+n \,|\, \tau > k) = \mathbb{P}(\tau = n)$$

Por outro lado, uma variável aleatória τ com valores em $\mathbb{N}$ 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 $\mathbb{P}(\tau = n+1) = (1-p) \cdot \mathbb{P}(\tau = n)$, 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 τ 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 τ é

$$\begin{aligned}\mathbb{E}\tau &= \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\,p = -p \cdot \frac{d}{dp}\left(\sum_{k=0}^{\infty}(1-p)^k\right) \\ &= -p \cdot \frac{d}{dp}\,(1/p) = 1/p\end{aligned}$$

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 τ_1 e τ_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 a sua média. Determine a lei da variável $\max\{\tau_1, \tau_2\}$.

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 $n \to \infty$.

$$\begin{aligned} \mathbb{P}(S_n = k) &= \binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{n-k} \\ &= \frac{\lambda^k}{k!} (1-\lambda/n)^n \frac{n(n-1)...(n-k+1)}{n^k} (1-\lambda/n)^{-k} \\ &\to \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

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 ξ com valores $0, 1, 2, ...$ tem *lei de Poisson* com parâmetro λ se a sua densidade discreta é

$$\mathbb{P}(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ .$$

Uma notação é $\xi \sim$Poisson(λ). 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 ξ são

$$\begin{aligned} \mathbb{E}\xi &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \\ \mathbb{V}\xi &= \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 \\ &= \lambda \cdot \sum_{k=0}^{\infty} (k+1) \cdot \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 = \lambda(\lambda+1) - \lambda^2 = \lambda \end{aligned}$$

Exercícios.

a. Seja ξ uma variável aleatória com lei de Poisson Poisson(λ).
Determine o máximo da densidade discreta $k \mapsto \mathbb{P}(\xi = k)$.
Fixado k, que valor de λ maximiza $\mathbb{P}(\xi = k)$?

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 ξ e η chamadas com lei de Poisson e parâmetros λ 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.

e. Um sinal é uma sucessão de bytes, uma palavra nas letras 0 e 1. Na transmissão de um sinal, cada byte pode ser distorcido com probabilidade $p = 0.01$, independentemente uns dos outros.

Estime

- a probabilidade de que um sinal de 1000 bytes contenha bytes distorcidos (ou seja, pelo menos um),

- e a probabilidade de que um sinal de 1000 bytes contenha pelo menos 10 bytes distorcidos.

Calcule o valor esperado e a variância do número de bytes distorcidos na transmissão de um sinal de 1000 bytes.

Distribuição de Gibbs. A energia de um sistema termodinâmico em equilíbrio térmico é uma variável aleatória ξ com valores $e_1, e_2, e_3, ...$ e lei determinada por

$$p_n = \mathbb{P}(\xi = e_n) = \frac{e^{-\beta e_n}}{Z(\beta)}$$

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 β, na interpretação física, é igual a $1/k_B T$, onde T é a temperatura e k_B a constante de Boltzmann.

A "energia média", definida por $E = \mathbb{E}\xi$, é igual a

$$E = -\frac{\partial}{\partial \beta} \log Z(\beta)$$

A "entropia" do sistema é definida como

$$S = -\sum_n \log p_n \cdot p_n$$

A "energia livre", definida como sendo $F = E - \beta^{-1} S$, é igual a

$$F = -\beta^{-1} \log Z(\beta)$$

8 Leis dos grandes números

Desigualdades de Chebyshev. Seja ξ uma variável aleatória discreta não negativa. Se a variável ξ é integrável, então

$$\mathbb{P}\{\xi \geq \varepsilon\} \leq \frac{1}{\varepsilon}\mathbb{E}\xi$$

para todo $\varepsilon > 0$. A demonstração é simplesmente a sequência de estimações elementares

$$\begin{aligned}\mathbb{P}\{\xi \geq \varepsilon\} &= \sum_{x_k \geq \varepsilon} \mathbb{P}(\xi = x_k) \\ &\leq \sum_{x_k \geq \varepsilon} \frac{x_k}{\varepsilon} \cdot \mathbb{P}(\xi = x_k) \\ &\leq \sum_{k} \frac{x_k}{\varepsilon} \cdot \mathbb{P}(\xi = x_k) = \frac{1}{\varepsilon}\mathbb{E}\xi\end{aligned}$$

Este é o protótipo de uma família de desigualdades, obtidas escolhendo oportunamente a variável ξ em função de outras.

Um caso particular é a *desigualdade de Markov*: se ξ é uma variável aleatória discreta integrável e $\varepsilon > 0$ então

$$\mathbb{P}\{|\xi| \geq \varepsilon\} \leq \frac{1}{\varepsilon}\mathbb{E}|\xi|$$

Outro caso particular, obtido considerando a variável $|\xi - \mathbb{E}\xi|^2$, é a *desigualdade de Chebyshev*: se ξ é uma variável aleatória discreta com variância finita e $\varepsilon > 0$ então

$$\mathbb{P}\{|\xi - \mathbb{E}\xi| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2}\mathbb{V}\xi$$

A desigualdade de Chebyshev não é, em geral, uma boa estimação. Melhores costumam ser as desigualdades

$$\mathbb{P}\{|\xi - \mathbb{E}\xi| \geq \varepsilon\} \leq \frac{1}{\varepsilon^k}\mathbb{E}|\xi - \mathbb{E}\xi|^k$$

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 ξ é tal que $e^{\beta\xi}$ tém esperança finita para todo $\beta > 0$, então

$$\mathbb{P}(\xi \geq \varepsilon) = \mathbb{P}\left(e^{\beta\xi} \geq e^{\beta\varepsilon}\right) \leq e^{-\beta\varepsilon}\mathbb{E}e^{\beta\xi}$$

para todo $\beta > 0$, e portanto

$$\mathbb{P}(\xi \geq \varepsilon) \leq e^{-H(\varepsilon)}$$

onde a função "entropia" H é a transformada de Legendre da função $\beta \mapsto \log \mathbb{E}e^{\beta\xi}$, definida por

$$H(\lambda) = \sup_{\beta > 0}\left(\beta\lambda - \log \mathbb{E}e^{\beta\xi}\right)$$

Esta desigualdade joga um papel central na teoria dos grandes desvios.

Médias empíricas. Seja (ξ_k) uma sucessão de variáveis aleatórias definidas num espaço de probabilidades $(\Omega, \mathcal{E}, \mathbb{P})$, e sejam S_n "as somas parciais das ξ_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 ξ, então a esperança de S_n/n é igual a $\mathbb{E}\xi$, ou seja é constante, não depende de n. Se as variáveis são também independentes, a variância $\mathbb{V}(S_n/n)$ é igual a $\frac{1}{n}\mathbb{V}\xi$, e portanto decresce quando n cresce.

Isto leva a conjecturar que, quando n é grande, a variável $S_n/n - \mathbb{E}\xi$ é "pequena" com grande probabilidade, i.e, numa linguagem muito informal,

$$\frac{S_n}{n} \sim \mathbb{E}\xi$$

Por exemplo, se ξ_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 $\mathbb{E}(S_n/n) = p$ e a sua variância é

$$\mathbb{E}\left|\frac{S_n}{n} - p\right|^2 = \frac{pq}{n}$$

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 $\mathbb{E}\xi = m$ e variância finita, e seja $S_n = \xi_1 + \xi_2 + ... + \xi_n$. Então para todo $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{S_n}{n} - m\right| < \varepsilon\right\} = 1$$

dem. Um cálculo mostra que $\mathbb{E}(S_n/n) = \mathbb{E}\xi$ e $\mathbb{V}(S_n/n) = \frac{1}{n}\mathbb{V}\xi$. A desigualdade de Chebyshev diz que, dado $\varepsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right\} \leq \frac{\mathbb{V}\xi}{n\varepsilon^2}$$

e portanto

$$\mathbb{P}\left\{\left|\frac{S_n}{n} - m\right| < \varepsilon\right\} \geq 1 - \frac{\mathbb{V}\xi}{n\varepsilon^2} \to 1$$

quando $n \to \infty$. □

obs. *(convergência em probabilidade)* A lei dos grandes números costuma ser enunciada como "sejam ... então $S_n/n \to_{\mathbb{P}} m$", que se lê: as médias empíricas S_n/n convergem para o valor médio m "em probabilidade".

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

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{S_n}{n} - p\right| < \varepsilon\right\} = 1$$

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 ε 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 $S_n/n \to_{\mathbb{P}} p$", e fica feliz. Um físico não, ele quer saber qual é a informação "física" do teorema. O teorema diz que, fixado um "erro" ε e uma probabilidade α, se n é suficientemente grande a probabilidade de observar uma frequência de sucessos S_n/n que difere de p por mais que ε é menor que α. A previsão física é, se α é muito pequeno, "a frequência satisfaz $|S_n/n - p| < \varepsilon$ na esmagadora maioria das vezes que repetimos as n experiências". Enfím, o enunciado "$S_n/n \to_{\mathbb{P}} p$" 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 ε, α e n, embora não seja a melhor possível. Determinar o n optimal em função de ε e α, 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.

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{S_n}{n} - \overline{p}_n\right| < \varepsilon\right\} = 1$$

porque $\mathbb{V}\xi_k = p_k\,(1 - p_k) \leq \sup_{0\leq x\leq 1} x\,(1 - x) = 1/4$ para todo k.

Mais interessante é observar que $\mathbb{V}(S_n/n) \leq \overline{\sigma}_n^2/n$, 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 $1/2$).

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 ε 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 $\mathbb{P}(|T_n| < K) \to 0$ quando $n \to \infty$.

9 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 $\mathbb{P}(S_n=k)$ 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 $\mathbb{P}(S_n=k)$ 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[1], ve-se que, quando $k\simeq np$ e n é grande, a probabilidade $\mathbb{P}(S_n=k)$ é da ordem de

$$\mathbb{P}(S_n=k)=\binom{n}{k}p^k(1-p)^{n-k}\simeq\frac{1}{\sqrt{2\pi npq}}$$

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[2]. 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

$$S_n^*=\frac{S_n-\mathbb{E}S_n}{\sqrt{\mathbb{V}S_n}}=\frac{S_n-np}{\sqrt{npq}}$$

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 $q'=1-p'$,

$$\begin{aligned}\mathbb{P}(S_n=k) &= \frac{1}{\sqrt{2\pi np'q'}}\left(p/p'\right)^k\left(q/q'\right)^{n-k}\cdot\alpha\\ &= \frac{1}{\sqrt{2\pi npq}}e^{-n\left(p'\cdot\log\left(p'/p\right)+q'\cdot\log\left(q'/q\right)\right)}\cdot\alpha\cdot\beta\\ &= \frac{1}{\sqrt{2\pi npq}}e^{-nH\left(p'\right)}\cdot\alpha\cdot\beta\end{aligned}$$

[1] A fórmula de Stirling diz que

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

onde $x_n\in\,]0,1[$.

[2] A notação de Landau dos "O-grandes" e "o-pequenos" é a seguinte. Sejam f e g duas funções definidas numa vizinhança de ∞.

"$f(x)=\mathcal{O}(g(x))$ quando $x\to\infty$" quer dizer que o quociente f/g é limitado numa vizinhança de ∞, ou seja que existem $K>0$ e $R\in\mathbb{R}$ 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 ∞, ou seja que $\lim_{x\to\infty}|f(x)|/|g(x)|=0$.

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

$$\alpha \cdot \beta = 1 + \mathcal{O}\left(x/\sqrt{n}\right)$$

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

$$\begin{aligned} \mathbb{P}(S_n = k) &= \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2} \cdot \alpha \cdot \beta \cdot \gamma \\ &= \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2} \cdot \left(1 + \mathcal{O}\left(x/\sqrt{n}\right)\right) \cdot \left(1 + \mathcal{O}\left(x^3/\sqrt{n}\right)\right) \end{aligned}$$

Em particular,

$$\sup_{|x| \le f(n)} \left| \frac{\mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} = x\right)}{\frac{1}{\sqrt{2\pi npq}} e^{-x^2/2}} - 1 \right| \to 0$$

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*[3]

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} = x\right) \sim \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2}$$

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*

$$\mathbb{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \le b\right) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

quando $n \to \infty$, uniformemente em $-\infty \le a < b \le \infty$.

[3] 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 $\lim_{x\to\infty} f(x)/g(x) = 1$.

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

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} = x_k\right) = \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \cdot (x_{k+1} - x_k)\,(1+\delta)$$

e fornece uma estimação para o erro $\delta = \alpha \cdot \beta \cdot \gamma - 1$. De facto δ, 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

$$\mathbb{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \leq b\right) = \sum_{a<x_k\leq b} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \cdot (x_{k+1} - x_k) + \sum_{a<x_k\leq b} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \cdot (x_{k+1} - x_k) \cdot \delta$$

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 $\Phi : \mathbb{R} \to [0,1]$ 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 Φ é dita *gaussiana*, ou *normal*. O teorema do limite central então fornece a aproximação

$$\mathbb{P}\left(np + a\sqrt{npq} < S_n \leq np + b\sqrt{npq}\right) \simeq \Phi(b) - \Phi(a)$$

ou, de maneira equivalente,

$$\mathbb{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \leq b\right) \simeq \Phi(b) - \Phi(a)$$

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,

$$\mathbb{P}\left(|S_n - np| \leq 2\sqrt{npq}\right) \geq 0.95 \qquad \mathbb{P}\left(|S_n - np| \leq 3\sqrt{npq}\right) \geq 0.99$$

ou, em termo da frequência,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq 2 \cdot \sqrt{\frac{pq}{n}}\right) \geq 0.95 \qquad \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq 3 \cdot \sqrt{\frac{pq}{n}}\right) \geq 0.99$$

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

$$\text{erro}_n = \sup_{-\infty<x<\infty} \left|\mathbb{P}\left\{S_n^* \leq x\right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt\right|$$

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

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\sqrt{pq}\right) \leq \frac{1}{\varepsilon^2 n}$$

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[4]

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\sqrt{pq}\right) &= \mathbb{P}\left(\left|\frac{S_n - np}{\sqrt{npq}}\right| \geq \varepsilon\sqrt{n}\right) \\
&\simeq 2 \cdot \int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\
&\leq \frac{1}{\varepsilon\sqrt{n}\sqrt{2\pi}} e^{-n\varepsilon^2/2}
\end{aligned}$$

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,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \leq \varepsilon\sqrt{\frac{pq}{n}}\right) \simeq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

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 ε) são tais que a desvio da frequência é da ordem de $1/\sqrt{n}$. Claro que também os eventos complementares têm probabilidade assimptoticamente estável...

[4] Observe que, se $x > 0$,

$$\int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt < \frac{1}{x} \cdot \int_x^{\infty} \frac{t}{\sqrt{2\pi}} e^{-t^2/2} dt = \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}$$

Eventos típicos e eventos estáveis da marcha aleatória. Seja $(T_n)_{n\in\mathbb{N}}$ a trajectória de uma marcha aleatória simétrica, i.e. $T_n = \xi_1 + \xi_2 + ... + \xi_n$ onde as variáveis ξ_k são independentes e identicamente distribuídas com valores ± 1 e lei determinada por $\mathbb{P}(\xi_k = 1) = 1/2$.

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

$$\begin{aligned}\mathbb{P}(|T_n| \geq \varepsilon n) &\simeq 2 \cdot \int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ &\leq \frac{1}{\varepsilon\sqrt{n}\sqrt{2\pi}} e^{-n\varepsilon^2/2}\end{aligned}$$

(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

$$\mathbb{P}\left(|T_n| \leq \varepsilon\sqrt{n}\right) \simeq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

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

$$\mathbb{P}(a \leq S_n \leq b) \geq 90\% \text{ ou } 95\% \text{ ou } 99\%$$

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 $(T_n)_{n\in\mathbb{N}}$ 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):

$$\mathbb{P}\left(\overline{\lim}\frac{T_n}{\sqrt{n}} = \infty\right) = \mathbb{P}\left(\underline{\lim}\frac{T_n}{\sqrt{n}} = -\infty\right) = 1$$

e

$$\mathbb{P}\left(\lim \frac{T_n}{\sqrt{n}\log n} = 0\right) = 1$$

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 α 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

$$\mathbb{P}\left(\overline{\lim}\frac{T_n}{\sqrt{2n\log\log n}} = 1\right) = 1$$

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

$$\mathbb{P}\left(\eta_n \leq x\right) \to \frac{2}{\pi}\arcsin\sqrt{x}$$

quando $n \to \infty$. O gráfico da função $\arcsin\sqrt{x}$ mostra que é mais provável que η_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" α, por exemplo 0.95 ou 0.99, procuramos um valor de $\varepsilon > 0$ tal que

$$\mathbb{P}\left(|f_n - p| \leq \varepsilon \cdot \sqrt{\frac{pq}{n}}\right) \geq \alpha$$

De facto, se n é grande, o teorema integral de De Moivre e Laplace diz que

$$\mathbb{P}\left(|f_n - p| \leq \varepsilon \cdot \sqrt{\frac{pq}{n}}\right) \simeq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

e portanto determina o ε correspondente ao nível de confiança: qualquer ε 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".

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

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

Se ξ^2 é integrável, a *variância* da variável aleatória ξ é definida por $\mathbb{V}\xi = \mathbb{E}\left(\xi - \mathbb{E}\xi\right)^2 = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2$, e resulta ser igual ao integral

$$\mathbb{V}\xi = \int_{-\infty}^{\infty} (x-m)^2 \cdot f_\xi(x)dx$$

onde m é o valor médio de ξ.

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 ξ 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 $x \notin [a, b]$.

Mudança de variável. Se $\xi : \Omega \to \mathbb{R}$ é uma variável aleatória, e $\varphi : \mathbb{R} \to \mathbb{R}$ uma função, a função composta $\eta = \varphi \circ \xi$ pode não ser uma variável aleatória. Acontece que η é uma variável aleatória se φ é suficientemente regular (se a imagem inversa de todo intervalo aberto é um boreleano), por exemplo se é contínua. Se ξ tem densidade f_ξ, então a função de repartição de η pode ser calculada por

$$\begin{aligned} F_\eta(x) &= \mathbb{P}\left(\eta \leq x\right) = \mathbb{P}\left(\varphi(\xi) \leq x\right) \\ &= \int_{\varphi^{-1}((-\infty,x])} f_\xi(y)dy \end{aligned}$$

Se ξ é absolutamente contínua e φ é um difeomorfismo, então $\eta = \varphi \circ \xi$ é também absolutamente contínua. A lei de η é determinada por meio da mudança de variável no integral, pois

$$\mathbb{P}\left(\eta \in A\right) = \int_{\varphi^{-1}(A)} f_\xi(x)dx = \int_A \left|\det \mathrm{Jac}\varphi^{-1}(x)\right| \cdot f_\xi\left(\varphi^{-1}(x)\right) dx$$

e portanto uma densidade de η é

$$f_\eta(x) = \left|\det \mathrm{Jac}\varphi^{-1}(x)\right| \cdot f_\xi\left(\varphi^{-1}(x)\right)$$

Também, se η é integrável, a sua média pode ser calculada por meio do integral

$$\mathbb{E}\left(\eta\right) = \int_{-\infty}^{\infty} \varphi\left(x\right) f_\xi(x)dx$$

Exercícios.

a. Se ξ tem densidade f_ξ, 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 ξ tem densidade f_ξ e função de repartição F_ξ, 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. *(lei de Cauchy)* Seja ξ 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 η não é integrável.

Lei exponencial. A variável ξ 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 τ é um parâmetro positivo. A lei exponencial é o análogo contínuo (e de facto um limite, como mostra o exemplo a seguir) 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 $\mathbb{E}\xi = \tau$. Uma notação é $\xi \sim \exp(\tau)$. A lei exponencial também tem (e é caracterizada por) a propriedade da falta de memória, ou seja

$$\mathbb{P}\{\xi > x + y \,|\xi > y\} = \mathbb{P}\{\xi > x\}$$

para todos $x, y > 0$. 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. Repito provas de Bernoulli em intervalos de tempo de comprimento $\varepsilon > 0$ até obter sucesso pela primeira vez. Se ξ denota o tempo necessário, então

$$\mathbb{P}(\xi > n\varepsilon) = (1-p)^n$$

onde p é a probabilidade de sucesso em cada prova. O tempo médio de espera é $\mathbb{E}\xi = \varepsilon/p$. No limite quando $\varepsilon \to 0$ e $p \to 0$ mantendo constante a média $\tau = \varepsilon/p$

$$\begin{aligned}
\mathbb{P}(\xi > t) &= \left(1 - \frac{\varepsilon}{\tau}\right)^{t/\varepsilon} \\
&= \left(1 - \frac{t/\tau}{t/\varepsilon}\right)^{t/\varepsilon} \\
&\to e^{-t/\tau}
\end{aligned}$$

i.e. a lei de ξ tende para uma lei exponencial com esperança τ.

Outra 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 ξ_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 $0 \leq x \leq T$,

$$\mathbb{P}(\xi_{\min} \geq x) = \prod_{k=1}^{N} \mathbb{P}(\xi_k \geq x) = \left(1 - \frac{x}{T}\right)^N$$

No limite termodinâmico, quando $N \to \infty$ e $T \to \infty$ mantendo constante a "frequência" $1/\tau = N/T$, temos que

$$\mathbb{P}(\xi_{\min} \geq x) \to e^{-x/\tau}$$

i.e. a lei de $\xi_{\min}$ tende para uma lei exponencial com esperança τ.

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 ξ_k, com lei exponencial e esperança τ_k. O tempo de vida da máquina é a variável $\xi_{\min} = \min\{\xi_1, \xi_2, ...\xi_n\}$. Se as ξ_k são independentes, dado $t > 0$,

$$\mathbb{P}(\xi_{\min} \geq t) = \prod_{k=1}^{n} \mathbb{P}(\xi_k \geq t) = \prod_{k=1}^{n} e^{-t/\tau_k} = e^{-t/\tau}$$

onde τ é n^{-1} vezes a média harmónica dos τ_k, i.e.

$$\tau = \left(\sum_{k=1}^{n} \frac{1}{\tau_k}\right)^{-1}$$

Portanto, $\xi_{\min}$ tem lei exponencial e esperança τ. Observem que $\tau < \min\{\tau_1, \tau_2, ...\tau_n\}$, e que, se os τ_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

$$\mathbb{P}(\xi_{\max} \leq t) = \prod_{k=1}^{n} \mathbb{P}(\xi_k \leq t) = \prod_{k=1}^{n} \left(1 - e^{-t/\tau_k}\right)$$

A média de $\xi_{\max}$ pode ser calculada integrando $\int_0^\infty t \cdot F'_{\xi_{\max}}(t)\, dt$.

Uma interpretação da lei de Poisson. Sejam $\xi_1, \xi_2, ...$ variáveis independentes com lei $\exp(\tau)$, e seja η 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 η é

$$F_\eta(k) = \mathbb{P}\{\eta \leq k\} = \sum_{i=0}^{k} \frac{(t/\tau)^i}{i!} e^{-t/\tau}$$

e portanto η tem lei Poisson (t/τ).

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"

$$\rho = \frac{N}{\text{vol}\left(B^3\left(R\right)\right)}$$

onde $\text{vol}\left(B^3\left(R\right)\right)$ 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 $\xi : \Omega \to \mathbb{R}$ 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

$$\Phi\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Se ξ tem lei normal $N(0,1)$, então a variável aleatória $\eta = \sigma\xi + m$, onde $m, \sigma \in \mathbb{R}$ 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 η com lei $N(m, \sigma^2)$ tem $\mathbb{E}\eta = m$ e $\mathbb{V}\eta = \sigma^2$.

Uma soma de variáveis normais e independentes é uma variável normal. De facto, é possível mostrar que 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 $S_n = \xi_1 + \xi_2 + ... + \xi_n$ tem lei $N\left(m, \sigma^2\right)$ com média $m = m_1 + m_2 + ... + m_n$ e variância $\sigma^2 = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2$.

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 η com lei $N(m, \sigma^2)$ tem $\mathbb{E}\eta = m$ e $\mathbb{V}\eta = \sigma^2$.

b. Seja ξ uma variável com lei normal $N\left(m, \sigma^2\right)$. Determine intervalos simétricos $m \pm \varepsilon \cdot \sigma$ tais que a probabilidade

$$\mathbb{P}\left(m - \varepsilon \cdot \sigma \leq \xi \leq m + \varepsilon \cdot \sigma\right)$$

seja $\geq 90\%$ ou 95% ou 99%.

Uma interpretação geométrica da lei normal. Seja $B_{\sqrt{n}}^{n+1}$ a bola de raio $\sqrt{n}$ centrada na origem de $\mathbb{R}^{n+1}$, seja $\mathbb{P}_n$ a probabilidade uniforme em $B_{\sqrt{n}}^{n+1}$, e seja $\pi : \mathbb{R}^{n+1} \to \mathbb{R}$ a projeção $x = (x_0, x_1, ..., x_n) \mapsto x_0$. A medida imagem $\pi\mathbb{P}_n$ tem suporte $[-\sqrt{n}, \sqrt{n}]$ e densidade

$$\gamma_n \cdot \left(\sqrt{1 - \frac{x_0^2}{n}} \right)^n$$

onde γ_n é um fator de normalização. No limite quando $n \to \infty$ a densidade da medida $\pi\mathbb{P}_n$ 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 $S_{\sqrt{E}}^{3n-1} = \left\{ x \in \mathbb{R}^{3n} \text{ t.q. } |x|^2 = E \right\}$. Fixar uma "energia média por partícula", por exemplo 1, e fazer crescer o número de partículas equivale a estudar as esferas $S_{\sqrt{n}}^{3n-1}$ no limite termodinâmico quando $n \to \infty$. Seja $\mathbb{P}_n$ a probabilidade uniforme em $S_{\sqrt{n}}^{3n-1}$, e seja $\pi : \mathbb{R}^{3n} \to \mathbb{R}^3$ a projeção $x = (x_1, ..., x_n) \mapsto x_1$, onde x_1 é o momento da primeira partícula. A medida imagem $\pi\mathbb{P}_n$ 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 $\mathbb{R}^3$. 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 $\xi : \Omega \to \mathbb{R}$ 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 α 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 $\mathbb{E}\xi = \alpha/\lambda$ e $\mathbb{V}\xi = \alpha/\lambda^2$.

A lei $\Gamma(1, \lambda)$ é a lei exponencial $\exp(1/\lambda)$.

Se ξ tem lei $N(0, \sigma)$ então ξ^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 $\alpha = \alpha_1 + \alpha_2 + ... + \alpha_n$.

Qui-quadrado. Se $\xi_1, \xi_2, ..., \xi_n$ são independentes e têm lei $N(0,1)$, então a variável

$$\chi^2_n = \xi_1^2 + \xi_2^2 + ... + \xi_n^2$$

tem lei $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, dita lei *qui-quadrado*. Em particular $\mathbb{E}\chi^2_n = n$ e $\mathbb{V}\chi^2_n = 2n$.

Se n é grande, a lei de $\sqrt{2\chi^2_n}$ é muito bem aproximada pela lei normal $N\left(\sqrt{2n-1}, 1\right)$.

T

de Student. A *lei de Student* t_n é a lei da variável

$$\varsigma = \frac{\xi}{\sqrt{\eta/n}}$$

onde ξ e η são independentes, ξ tem lei $N(0,1)$ e η tem lei χ^2_n.

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 α um número entre 0 e 1, uma probabilidade. O *quantil de ordem* α da variável aleatória ξ é o maior dos valores x tais que $\mathbb{P}\{\xi \leq x\} \leq \alpha$, ou seja

$$q_\alpha = \sup\{x \text{ t.q. } \mathbb{P}\{\xi \leq x\} \leq \alpha\}$$

Observe que, se a densidade de ξ é estrictamente positiva, então q_α é o único valor tal que $\mathbb{P}\{\xi \leq q_\alpha\} = \alpha$. Os livros de estatística costumam ter tabelas dos quantis das leis normal, qui-quadrado, T de Student e outras.

Quantis da lei normal. Seja ϕ_α o quantil de ordem α 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 $\xi \sim N(0,1)$,

$$\mathbb{P}\left\{|\xi| \leq \phi_{1-\alpha/2}\right\} = 1 - \alpha$$

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,

$$\mathbb{P}\{|\xi| \leq 1.64\} \simeq 0.90 \quad \mathbb{P}\{|\xi| \leq 1.96\} \simeq 0.95 \quad \mathbb{P}\{|\xi| \leq 2.58\} \simeq 0.99$$

Outra observação útil é que, se η tem lei $N(m, \sigma^2)$, então os seus quantis q_α são simplesmente $q_\alpha = \sigma\phi_\alpha + m$. Também observamos que

$$\mathbb{P}\{|\eta - m| \leq \sigma\} \simeq 0.683 \quad \mathbb{P}\{|\eta - m| \leq 2\sigma\} \simeq 0.954 \quad \mathbb{P}\{|\eta - m| \leq 3\sigma\} \simeq 0.997$$

10 Convergência e aproximação

Convergências. Seja (ξ_n) uma sucessão de variáveis aleatórias com valores reais definidas no espaço de probabilidades $(\Omega, \mathcal{E}, \mathbb{P})$. 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 S_n é o número de sucessos em n provas de Bernoulli, a sucessão (S_n) não converge nunca.

A sucessão (ξ_n) converge para ξ *quase certamente* (ou em *quase todo ponto*), notação $\xi_n \to_{\text{qtp}} \xi$, se

$$\mathbb{P}\left(\omega \in \Omega \text{ t.q. } \lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)\right) = 1$$

A sucessão (ξ_n) converge para ξ *em probabilidades*, notação $\xi_n \to_{\mathbb{P}} \xi$, se para todo $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\omega \in \Omega \text{ t.q. } |\xi_n(\omega) - \xi(\omega)| < \varepsilon\right) = 1$$

É importante ter uma noção de convergência para uma sucessão de variáveis aleatórias definidas em espaços de probabilidades distintos. A sucessão (ξ_n) converge para ξ *em lei* (ou *em distribuição*), notação $\xi_n \to_{\text{d}} \xi$, se para toda função real contínua e limitada φ acontece que

$$\mathbb{E}\varphi(\xi_n) \to \mathbb{E}\varphi(\xi)$$

Uma definição equivalente é: $\xi_n \to_{\text{d}} \xi$ sse para todo ponto de continuidade x da função de repartição F_ξ de ξ, acontece que

$$\lim_{n\to\infty} F_{\xi_n}(x) = F_\xi(x)$$

onde F_{ξ_n} são as funções de repartição das ξ_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 ξ_n. Se as ξ_n são independentes e identicamente distribuídas e têm variância finita, então a desigualdade de Chebyshev implica que

$$\frac{S_n}{n} \to_{\mathbb{P}} m$$

onde m é o valor médio das ξ_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 $\mathbb{E}\xi_1 = m$, 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" (J.L. Doob, *Measure Theory*, Springer-Verlag, New York 1994).

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"

$$S_n^* = \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathbb{V}S_n}}$$

que têm valor médio 0 e variância 1. O primeiro resultado nesse sentido foi obtido por De Moivre e Laplace, no caso de variáveis ξ_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 $\mathbb{E}\xi_n = m$ e $\mathbb{V}\xi_n = \sigma^2 > 0$. 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 ξ é definida por $\phi_\xi(\theta) = \mathbb{E}e^{i\theta\xi}$. 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 $\theta \in \mathbb{R}$.

Seja ϕ 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 $\mathbb{E}\eta_k = 0$ e $\mathbb{V}\eta_k = 1$. 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 $\mathbb{P}\{a < S_n^* < b\}$ pode ser aproximada por $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$, no sentido em que

$$\mathbb{P}\{a < S_n^* < b\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

quando $n \to \infty$.

A velocidade de convergência depende, obviamente, das leis das variáveis ξ_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

$$\text{erro}_n = \sup_{-\infty<x<\infty} \left| \mathbb{P}\{S_n^* < x\} - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \right|$$

é $\text{erro}_n \leq \text{const}/\sqrt{n}$, 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_ξ. Sejam $[a, b]$ um intervalo da recta e η_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 $\varphi : [0, 1] \to \mathbb{R}$ uma função limitada e integrável. Então as variáveis $\varphi(\xi_k)$ têm valor médio

$$\mathbb{E}\varphi(\xi_k) = \int_0^1 \varphi(t) dt$$

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 φ 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 ξ tem lei uniforme em $[0, 1]$ então $a + (b - a)\xi$ tem lei uniforme em $[a, b]$.

b. Se ξ 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 τ, então η, 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 $\text{area}\,(A)$.

11 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" Δx.

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 "desvios" 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 "desvios 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*, ou *standard uncertainty),* 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.

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...

Dígitos significativos. 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 Δx é uma medida da "sensibilidade" dos instrumentos do laboratório, ou melhor da reproduzibilidade das experiências.

Histogramas. Se o número n de observações é suficientemente grande, um histograma pode sugerir um modelo para a distribuição dos dados $x_1, x_2, ..., x_n$.

Modelo das observações: hipótese gaussiana. 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 ξ com esperança $\mathbb{E}\xi = m$ (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 ξ é mesmo variável, e tem uma certa lei. Não podemos saber a lei de ξ, logo podemos supôr que tem uma certa variância $\mathbb{V}\xi = \sigma^2$. 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

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

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 podem ser muito parecidos com o gráfico de uma distrubuição normal. Neste caso uma hipótese de trabalho pode ser que ξ 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 ξ 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 ξ seja igual a m mais uma soma de muitos "erros aleatórios" $\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_k$ 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 ξ é bem aproximada por uma lei normal com média m.

A variância σ^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 σ. Como estimar σ? Os estatísticos chamam

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

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 $\mathbb{E}S^2=\sigma^2$, portanto S^2 é uma estimação razoável de σ^2. Também útil é saber que a variância de S^2 é $\mathbb{V}S^2=\frac{2\sigma^2}{n-1}$, e portanto se n é grande a variância amostral tem boa probabilidade de estar perto da variância de ξ (lei dos grandes números). O modelo diz que a variável

$$\frac{n-1}{\sigma^2}\cdot S^2$$

tem lei χ^2_{n-1}, mas esta variável ainda contém a variância desconhecida σ^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 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 t_{n-1}. 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 $\overline{x}-3\frac{S}{\sqrt{n}}$ e $\overline{x}+3\frac{S}{\sqrt{n}}$ com probabilidade $\geq 99\%$", e , julgando esta uma probabilidade mesmo grande, escrever

$$\text{o valor do observável } x \text{ é igual a } \overline{x}\pm 3\cdot\frac{S}{\sqrt{n}}$$

Exemplo: a agulha de Buffon. O chão tem linhas paralelas a distância ℓ. Lanço n vezes uma agulha de comprimento ℓ 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 π é igual a $2/p$, portanto um físico que quer estimar π escreve

$$\pi=\frac{2}{f}\pm\frac{2}{f^2}\Delta f$$

Em 1901, o senhor Lazzarini lançou 3408 agulhas obtendo $2/f=3.1415929$. Quantos dígitos desta estimação são confiáveis?

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 número 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

$$b'/n \pm 1/\sqrt{n}$$

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...

Dados não gaussianos, mediana. Pode acontecer que um histograma dos $x_1, x_2, ..., x_n$ não mostra um padrão conhecido. Neste caso a hipótese normal não é credível, e pode ser mais honesto utilizar a mediana e portanto o desvio médio da mediana como estimadores do valor do observável e do erro na sua estimação.

A *mediana* x_{med} dos resultados das observações é o valor "central" dos dados. Se ordenamos os dados de forma a ter $x_1 \leq x_2 \leq ... \leq x_n$, então

$$x_{\text{med}} = \begin{cases} x_{(n+1)/2} & \text{se } n \text{ é impar} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{se } n \text{ é par} \end{cases}$$

O *desvio médio da mediana* é

$$D_{\text{med}} = \frac{1}{n}\left(|x_1 - x_{\text{med}}| + |x_2 - x_{\text{med}}| + ... + |x_n - x_{\text{med}}|\right)$$

Observem que a mediana e o desvio médio (da mediana) são menos sensível do que a média e o desvio padrão aos valores extremos das observações.

Outlets. Pode aconteçer que um (ou alguns) dos valores $x_1, x_2, ..., x_n$, por exemplo x_{13}, esteja muito afastado dos outros. O que fazer?

Alguns físicos desenvolveram uma série de "testes", receitas para decidir quando aceitar ou rejeitar o dado x_{13} (o "teste Q", o "critério de Chauvenet", ...) baseados em conjecturas acerca da distribuição dos erros e em considerações estatísticas. Antes de aplicar cegamente um dos testes, é boa norma verificar as condições do laboratório em que foi obtido o valor excepcional (famosa é a história do balde de agua escondido de noite pela "sora Cesarina" por baixo da mesa do laboratório de Bruno Pontecorvo em via Panisperna, Roma, episódio que levou a equipa de Enrico Fermi a uma descoberta científica devastante, e dez anos depois à explosão da primeira bomba nuclear em Alamogordo, New Mexico). Também ha físicos que simplesmente decidem não fazer nada, ou, se podem, repetir muitas mais vezes as observações até que x_{13} não influencie significativamente os valores de $\overline{x}$ e S. Para ser honestos, é importante decidir uma "estratégia" antes de obter os dados.

Estimadores. A teoria do físico é mesmo um modelo probabilístico, ou seja uma variável aleatória ξ. 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 ξ. A lei de ξ contém parâmetros, por exemplo a média m, ou a variância σ^2. Os observáveis são os parâmetros da lei de ξ. Esta é a situação mais geral: o modelo físico contém o modelo das experiências.

Estimar o parâmetro θ 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 θ. A coisa mais natural é procurar t de modo que a sua esperança seja $\mathbb{E}_\theta t(\xi_1, \xi_2, ..., \xi_n) = \theta$, onde $\mathbb{E}_\theta$ denota a esperança com respeito a lei determinada pelo valor θ 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 $\mathbb{E}_\theta t = \theta$. No espírito da lei dos grandes números, um estimador t é dito *consistente* se para todo $\varepsilon > 0$

$$\mathbb{P}_\theta \left(|t - \theta| \geq \varepsilon \right) \to 0$$

quando $n \to \infty$.

Exemplos. Se ξ 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 σ^2/n, logo é consistente (pela desigualdade de Chebyshev).

Também, a variância amostral S^2 é um bom estimador da variância: tem esperança σ^2, logo é centrado, e variância $\frac{2\sigma^2}{n-1}$, logo é consistente.

Desigualdade de Rao-Cramér. Uma medida da bondade do estimador centrado t é a sua variância $\mathbb{V}_\theta t$. O estimador centrado t^* do parâmetro θ é dito *eficiente* se

$$\mathbb{V}_\theta t^* = \inf_t \mathbb{V}_\theta t$$

onde o ínfimo é sobre todos os estimadores centrados de θ.

Seja p_θ a densidade discreta das variáveis $\xi_1, \xi_2, ..., \xi_n$. A probabilidade de observar os valores $x_1, x_2, ..., x_n$ é

$$\mathbb{P}_\theta (\omega) = \mathbb{P}_\theta \left(\xi_1 = x_1, \xi_2 = x_2, ..., \xi_n = x_3 \right) = \prod_{i=1}^{n} p_\theta (x_i)$$

e a probabilidade do evento certo é

$$1 = \sum_\omega \mathbb{P}_\theta (\omega)$$

Assumindo que seja possível derivar em ordem a θ a igualdade acima e que $\mathbb{P}_\theta (\omega) > 0$ para todo ω, temos que

$$0 = \sum_\omega \mathbb{P}_\theta (\omega) \cdot \frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) \right)$$

Se t é um estimador centrado de θ, então

$$\theta = \mathbb{E}_\theta t = \sum_\omega t (\omega) \, \mathbb{P}_\theta (\omega)$$

e derivando obtemos

$$1 = \sum_\omega t (\omega) \cdot \mathbb{P}_\theta (\omega) \cdot \frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) = \mathbb{E}_\theta \left(t (\omega) \cdot \frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) \right)$$

Juntando as duas expressões temos que

$$1 = \mathbb{E}_\theta \left((t - \theta) \cdot \frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) \right)$$

e, pela desigualdade de Cauchy-Schwarz,

$$1 \leq \mathbb{E}_\theta (t - \theta)^2 \cdot \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) \right)^2$$

A conclusão é a *desigualdade de Rao-Cramér*

$$\inf_{t \text{ centrado}} \mathbb{V}_\theta t \geq \frac{1}{I_\theta}$$

onde a *informação de Fisher* é definida como

$$I_\theta = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log \mathbb{P}_\theta (\omega) \right)^2$$

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 θ é a seguinte. Dado θ, 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 θ. 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 σ^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 ξ tem lei de Poisson Poisson(λ), o estimador de máxima verosimilhança para λ é a média amostral $\overline{x}$.

b. Se ξ tem lei exponencial $\exp(\tau)$, o estimador de máxima verosimilhança para o tempo característico τ 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

$$\frac{k_a k_e}{k_{ae}}$$

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

$$\mathbb{P}(\xi = k) = (k^n - (k-1)^n) N^{-n}$$

porque $\mathbb{P}(\xi \leq k) = (k/N)^n$. Aproximando a soma com um integral (o que não faz mal se $N \gg 1$), a esperança de ξ é $\mathbb{E}\xi \simeq \frac{n}{n+1}N$. 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 t_{n-1}.

Intervalos para uma probabilidade, ou uma proporção. 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 σ^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 χ^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_α, $q_{1-\alpha/2}$ e $q_{\alpha/2}$ são os quantis da lei χ^2_{n-1}.

Se n é grande (por exemplo $n \geq 30$) a lei de χ^2_{n-1} é bem aproximada pela lei $N\left(\sqrt{2n-1}, 1\right)$, logo podemos estimar

$$q_\alpha \simeq \frac{1}{2}\left(\phi_\alpha + \sqrt{2n-1}\right)$$

onde ϕ_α é o quantil da lei normal $N(0,1)$.

Desvio padrão da média e desvio padrão relativo da média. Se não temos tabelas no laboratório, basta lembrar que intervalos de confiança "generosos" com nível de confiança $\geq 95\%$ ou $\geq 99\%$ ou ≥ 99.73 são da ordem de

$$\overline{x} \pm 2 \cdot \frac{S}{\sqrt{n}} \qquad \overline{x} \pm 2.6 \cdot \frac{S}{\sqrt{n}} \qquad \overline{x} \pm 3 \cdot \frac{S}{\sqrt{n}}$$

(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!

Em terceiro lugar, se n é pequeno, da ordem de poucas unidades (caso em que os quantis da lei de Student diferem significativamente dos quantis da lei normal), não temos informação suficiente para acreditar que os dados experimentais são distribuídos de acordo com a lei de Gauss!

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*

$$S_m = \frac{S}{\sqrt{n}}$$

é uma medida da incerteza na estimação de m, o suposto valor verdadeiro de x, ao utilizar a média $\overline{x}$. Mais significativo do que o desvio padrão é o *desvio padrão relativo* da média (relative standard uncertainty),

$$S_m/\overline{x} = \frac{S}{\overline{x}\sqrt{n}}$$

que diz a quantidade dos dígitos significativos na estimação de m.

Como apresentar os dados. Se o número n de observações é grande e/ou depois de ter visto os histogramas julgamos que a distribuição dos erros é normal, os resultados de uma experiência costumam ser apresentados numa das seguintes maneiras:

• indicando a média e o seu desvio padrão estimado (e.s.d., ou seja "estimated standard deviation" of the mean), ou seja

$$x = \overline{x} \pm S_m \text{ (1 e.s.d. error limit)}$$

- indicando a média e o seu desvio padrão relativo estimado

$$x = \overline{x} \text{ with relative standard deviation } S_m/\overline{x}$$

- indicando um intervalo de confiança, por exemplo

$$x = \overline{x} \pm t_{0.975} \cdot S_m \text{ (95\% confidence, } \nu = n - 1)$$

onde lembramos ao leitor que o quantil $t_{0.975}$ da lei de Student foi obtido com $\nu = n - 1$ graus de liberdade (e portanto que foram feitas n observações).

Os observáveis da física costumam ter uma dimensão (g, erg, eV, atm, ...). O valor de $\overline{x}$ e a sua incerteza S_m devem ser apresentados com o mesmo número de casas decimais, e o número de dígitos significativos da incerteza S_m costuma ser um ou quanto mais dois (por exemplo se o primeiro for pequeno).

Por exemplo, a experiência FIRAS, do satélite COBE da NASA, deu a seguinte estimação da temperatura de radiação cósmica

$$T = (2.725 \pm 0.001) \text{ K}$$

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 observações. Portanto, para reduzir o erro de um factor 10, um físico tem que multiplicar por 100 as observações...

Por exemplo, para estimar π 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 π tenho que lançar 100000000 agulhas!

Propagação dos erros. Se os observáveis $x, y, ..., z$ são estimados ser $\overline{x} \pm \Delta x$, $\overline{y} \pm \Delta y$, ..., $\overline{z} \pm \Delta z$ (utilizando para todos a mesma convenção, que pode ser o desvio padrão da média ou um intervalo de confiança), 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 Δf 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 Δx, Δy, ..., Δz são gaussianos, indepenenes e pequenos.

12 Testes estatísticos

Testes. Às vezes, mais do que estimar uns observáveis x e y, um físico está interessado em testar uma hipótese do tipo

• $x = y$ (por exemplo, verificar se a velocidade da luz não depende da direção relativa às estrelas fixas),

• 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 $H > 0$),

• ou x tem uma certa distribuição (por exemplo, verificar se o decaimento radioactivo é descrito por uma lei exponencial).

Uma maneira ingénua de testar a hipótese "$x = y$" consiste em calcular uns intervalos de confiança $\overline{x} \pm \varepsilon S_x/\sqrt{n}$ e $\overline{y} \pm \varepsilon S_y/\sqrt{m}$ de nível suficientemente grande (i.e. $\varepsilon \simeq 2$ ou 3) para os dois observáveis, e aceitar a hipótese se estes intervalos têm pontos comuns (embora mais inteligente seria medir "directamente" $x - y$, como fizeram Michelson e Morley na famosa experiência do interferómetro). Da mesma forma, parece razoável aceitar a hipótese "$x > a$" se o intervalo de confiança $\overline{x} \pm \varepsilon S_x/\sqrt{n}$ está à direita de a. Esta ideia é a base do procedimento formal (só aparentemente mais complicado) que os estatísticos chamam testes de hipótese paramétricos.

Uma maneira ingénua de testar a hipótese "x tem valores $z_1, z_2, ..., z_m$ com densidade discreta $p_1, p_2, ..., p_m$" consiste em comparar o histograma das n observações, $N_1, N_2, ..., N_m$ (ou seja, N_k é o número de vezes que observamos o valor z_k), com o istograma dos valores esperados $np_1, np_2, ..., np_m$ e verificar que os N_k não diferem muito dos np_k. Se os desvios quadráticos observados $(N_k - np_k)^2$ são da ordem das variâncias esperadas $np_k(1 - p_k)$, ou seja se a soma $T = \sum_{k=1}^{m} (N_k - np_k)^2 / np_k$ é da ordem de m, e não muito maior!; podemos razoavelmente acreditar na nossa hipótese. Esta é a ideia do que os estatísticos chamam teste do qui-quadrado.

O que falta é quantificar a nossa confiança na resposta obtida.

Exemplo: as moedas com spin inteiro. Há livros 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.

Exemplo: os dados honestos. Uma aluna lançou um dado (para ser honestos, lançou quatro dados ...) uma centena de vezes obtendo

$$T_n = \sum_{k=1}^{6} \frac{(N_k - n/6)^2}{n/6} \simeq 2.2$$

Este é um valor aceitável para acreditar que o dado é honesto (ou seja mostra 1, 2, ... ou 6 pintas com probabilidades 1/6).

Testes paramétricos. 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* α do teste é a maior das probabilidades

$$\mathbb{P}(\text{rejeitar a hipótese} \mid \text{a hipótese é verdadeira})$$

(os estatístico chamam esta "a probabilidade de fazer um erro do primeiro tipo").

No nosso modelo, os valores observados $x_1, x_2, ..., x_n$ são distribuidos de acordo com uma certa lei $\mathbb{P}_\theta$ que depende de um parâmetro $\theta \in \Theta$ (o caso típico é x_k com lei normal $N(m, \sigma^2)$, cujos

parâmetros são m e σ^2). A hipótese determina uma lei, ou uma família de leis, da variável observada: $\theta \in \Theta_{\rm h}$ corresponde à "hipótese", e $\theta \notin \Theta_{\rm h}$, ou seja $\theta \in \Theta_{\rm a} = \Theta \backslash \Theta_{\rm h}$, corresponde à "alternativa". O resultado de n observações é uma variável aleatória z (que os estatísticos chamam "estatística do teste"), função dos resultados experimentais $x_1, x_2, ..., x_n$ (por exemplo a média amostral $\overline{x}$, o a variância amostral S^2, ou $\frac{\overline{x}-a}{S/\sqrt{n}}$, ...). 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 α 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 α, tipicamente 10%, 5% ou 1%.

Pode acontecer que as duas hipóteses alternativas sejam igualmente razoáveis. Neste caso também é significativo o parâmetro β, definido como a maior das probabilidades

$$\mathbb{P}\left(\text{aceitar a hipótese } \middle|\ \text{a hipótese é falsa}\right)$$

(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 $\alpha + \beta$.

Os estatísticos chamam *potência* do teste a função $\pi_R : \Theta \to [0,1]$ definida por

$$\begin{aligned} \pi_R(\theta) &= \mathbb{P}\left(\text{rejeitar a hipótese } \middle|\ \text{a lei das } x_k \text{ é } \mathbb{P}_\theta\right) \\ &= \mathbb{P}_\theta\left(z \in R\right) \end{aligned}$$

Ou seja, se a hipótese é verdadeira, i.e. se $\theta \in \Theta_{\rm h}$, então $\pi_R(\theta)$ é a probabilide de fazer um erro do primeiro tipo, e se a hipótese é falsa, i.e. se $\theta \in \Theta_{\rm a}$, então $\pi_R(\theta)$ é a probabilidade de rejeitar com razão a hipótese.

Testes sobre médias. Uma estratégia para testar a hipótese $x > a$ é assim. Obtidos os resultados $x_1, x_2, ..., x_n$ das experiências, podemos calcular

$$z = \frac{\overline{x} - a}{S/\sqrt{n}}$$

Se acreditamos que a distribuição dos x_k é gaussiana, com média m, o modelo nos diz que a variável $t = \frac{\overline{x}-m}{S/\sqrt{n}}$ tem lei de Student t_{n-1}. Podemos ver, nas tabelas, o valor t_α tal que $\mathbb{P}(t \le t_\alpha) = \alpha$. Se a hipótese é verdadeira, i.e. se $m > a$, então

$$\begin{aligned} \mathbb{P}(z \le t_\alpha) &= \mathbb{P}\left(t + \frac{m-a}{S/\sqrt{n}} \le t_\alpha\right) \\ &\le \mathbb{P}(t \le t_\alpha) = \alpha \end{aligned}$$

Portanto,

$$R = \{z \in \mathbb{R} \text{ t.q. } z < t_\alpha\}$$

é uma região crítica de um teste com nível de significância α. A receita é: aceitamos a hipótese se $z > t_\alpha$ e rejeitamos a hipótese se $z < t_\alpha$.

Se a hipótese é $x = a$, uma região crítica de um teste com nível de significância α é

$$R = \left\{z \in \mathbb{R} \text{ t.q. } |z| > t_{1-\alpha/2}\right\}$$

onde $t_{1-\alpha/2}$ é o quantil $1 - \alpha/2$ da lei de Student t_{n-1}. A receita é: aceitamos a hipótese se $|z| < t_{1-\alpha/2}$ e rejeitamos a hipótese se $|z| > t_{1-\alpha/2}$.

Comparação de dados. Outro problema é testar a hipótese $x = y$ a partir das observações $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_m$ de dois observáveis. Se acreditamos que as distribuições dos x_k e dos y_k são gaussianas, com médias m_x e m_y e variâncias σ_x^2 e σ_y^2, respectivamente, então a variável

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

tem lei normal $N(0,1)$ na hipótese de que $m_x = m_y$. Fixado um nível de significância α, as tabelas indicam-nos o valor $\phi_{1-\alpha/2}$ tal que $\mathbb{P}\left(|z| \geq \phi_{1-\alpha/2}\right) = \alpha$. Uma região crítica de um teste com nível de significância α é

$$R = \left\{z \in \mathbb{R} \text{ t.q. } |z| > \phi_{1-\alpha/2}\right\}$$

A receita é: aceitamos a hipótese se $|z| < \phi_{1-\alpha/2}$ e rejeitamos 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

$$S_{\text{tot}}^2 = \frac{1}{n+m-2}\left((n-1)S_x^2 + (m-1)S_y^2\right)$$

denota a variância total (a média aritmética entre as duas variâncias) então a variável

$$z = \frac{\overline{x} - \overline{y}}{S_{\text{tot}}\sqrt{1/n + 1/m}}$$

tem lei de Student t_{n+m-2} na hipótese de que $m_x = m_y$. Fixado um nível de significância α, as tabelas indicam-nos o valor $t_{1-\alpha/2}$ tal que $\mathbb{P}\left(|z| \geq t_{1-\alpha/2}\right) = \alpha$. Uma região crítica de um teste com nível de significância α é

$$R = \left\{z \in \mathbb{R} \text{ t.q. } |z| > t_{1-\alpha/2}\right\}$$

A receita é: aceitamos a hipótese se $|z| < t_{1-\alpha/2}$ e rejeitamos 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 os x_k têm 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^2. O modelo diz que a variável $q = (n-1)\frac{S^2}{\sigma^2}$ tem lei χ_{n-1}^2. Fixado um nível de significância α, uma tabela fornece-nos o valor $q_{1-\alpha}$ tal que $\mathbb{P}\left(q < q_{1-\alpha}\right) = 1 - \alpha$. Se

$$z = (n-1)\frac{S^2}{b^2}$$

na hipótese $\sigma^2 \leq b^2$ temos

$$\alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} > q_{1-\alpha}\right) \geq \mathbb{P}\left(z > q_{1-\alpha}\right)$$

Uma região crítica de um teste com nível de significância α é

$$R = \left\{z \in \mathbb{R} \text{ t.q. } z > q_{1-\alpha}\right\}$$

A receita é: aceitamos a hipótese se $z < q_{1-\alpha}$ e rejeitamos a hipótese se $z > q_{1-\alpha}$.

Se a hipótese for $\sigma^2 = b^2$, um argumento análogo dá-nos uma região crítica

$$R = \left\{z \in \mathbb{R} \text{ t.q. } |z| > q_{1-\alpha/2}\right\}$$

Teste do qui-quadrado, ou de Pearson. O problema é testar um modelo probabilístico (por exemplo, decidir 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 ξ, e queremos testar a nossa conjectura acerca da lei de ξ. Se a variável ξ é simples, enumeramos os seus valores $z_1, z_2, ..., z_m$ e calculamos a densidade discreta $k \mapsto p_k = \mathbb{P}(\xi = z_k)$. Se a variável ξ assume uma quantidade infinita de valores, convém dividir a recta real em intervalos disjuntos $]-\infty, z_1] \cup]z_1, z_2] \cup ... \cup]z_{m-1}, \infty[$ e depois calcular as probabilidades $p_k = \mathbb{P}(z_{k-1} < \xi \leq z_k)$. Isto corresponde a aproximar ξ com uma variável simples, e se os intervalos $]z_k, z_{k+1}]$ e/ou as respectivas probabilidades p_k são suficientemente pequenos não representa uma significativa perda de informação.

Sejam $f_1, f_2, ..., f_m$ as frequências empíricas em n observações, ou seja

$$f_k = \frac{1}{n}\text{card}\{i = 1, 2, ..., n \text{ tais que } z_{k-1} < x_i \leq z_k\}$$

A lei dos grandes números sugere que $f_k \sim p_k$ quando n é grande. O teorema limite central sugere que as flutuações quadráticas médias dos nf_k à volta dos valores esperados np_k sejam da ordem de $np_k(1 - p_k) \simeq np_k$. Uma medida global das flutuações observadas é

$$\begin{aligned} T_n &= \sum_{k=1}^{m} \frac{(nf_k - np_k)^2}{np_k} \\ &= n\sum_{k=1}^{m} \frac{(f_k - p_k)^2}{p_k} \end{aligned}$$

Valores aceitáveis de T_n no caso da hipótese ser verdadeira são portanto da ordem de m. Um *teorema de Pearson* diz que, quando $n \to \infty$, as variáveis T_n convergem em lei para uma variável qui-quadrado χ^2_{m-1}.

Fixado um nível de significância α, uma tabela fornece-nos o valor $q_{1-\alpha}$ tal que $\mathbb{P}\left\{\chi^2_{m-1} < q_{1-\alpha}\right\} = 1 - \alpha$. Uma região crítica de um teste com nivel de significância α é portanto

$$R = \{T_n \in \mathbb{R} \text{ t.q. } T_n > q_{1-\alpha}\}$$

Os estatísticos concordam em dizer que a aproximação de Pearson começa a ser boa (e portanto o teste é significativo) desde que os números esperados np_k de observações em cada um dos intervalos sejam ≥ 5.

13 Modelização

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

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

onde y, x e $a = (a_1, a_2, ..., a_m)$ são certos observáveis. Uma experiência típica consiste em observar os valores $y_1, y_2, ..., y_n$ correspondentes a um certo número de valores $x_1, x_2, ..., x_n$ de x, considerada como variável independente. Objectivos da experiência podem ser

- conjecturar a lei, ou seja a forma da função f,
- estimar os valores α dos "parâmetros livres" a que mais concordam com as observações,
- decidir se a lei $y = f(x, \alpha)$ descreve bem os resultados da experiência,
- fazer previsões.

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 ...

Exemplo. Um diagrama da luminosidade absoluta versus a cor, levou Hertzsprung e Russell a descobrir que uma grande parte das estrelas estão concentradas numa curva (quase uma recta), hoje em dia conhecida como "sequência principal".

Exemplo. Para medir a temperatura de radiação cósmica os físicos conjecturam valida a lei de Planck

$$S_\lambda(T) = \frac{8\pi hc}{\lambda^5} \cdot \frac{1}{e^{hc/\lambda kT} - 1}$$

observam as (energia)/(volume)×(comprimento de onda) S_{λ_k} em correspondência de um certo número de comprimentos de onda λ_k, e depois estimam a temperatura T escolhendo a curva $\lambda \mapsto S_\lambda(T)$ que melhor ajusta os dados.

Método dos mínimos quadrados. Numa experiência ideal temos um bom controlo, e possivelmente nenhum erro significativo, do observável x. Em correspondência de cada valor x_k temos muitas observações de y, e portanto uma estimação da média $\overline{y_k}$ e do desvio padrão S_{y_k}. Um diagrama dos x_k versus $\overline{y_k} \pm S_{y_k}$ pode sugerir, ou não!, uma correlação entre x e y, e possivelmente a forma da lei.

Seja

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

a nossa conjectura.

O *método dos mínimos quadrados* (*least-square fitting*) é uma receita que consiste em escolher os estimadores α para os parâmetros a de maneira tal que a soma dos "erros quadráticos"/"variâncias"

$$Q_a^2 = \sum_{k=1}^{n} \frac{\left(\overline{y_k} - f(x_k, a)\right)^2}{S_{y_k}^2}$$

seja a menor possível, ou seja

$$\sum_{k=1}^{n} \frac{\left(\overline{y_k} - f(x_k, \alpha)\right)^2}{S_{y_k}^2} = \min_a \sum_{k=1}^{n} \frac{\left(\overline{y_k} - f(x_k, a)\right)^2}{S_{y_k}^2}$$

Observe que cada erro $(\overline{y_k} - f(x_k, a))$ é pesado com um factor inversamente proporcional à incerteza S_{y_k}, e que portanto se n é grande um dado incerto, com $S_{y_{13}}$ muito maior que os outros, não influencia significativamente a estimação.

Uma hipótese de trabalho razoável é a hipótese gaussiana: cada y_k tem lei normal com esperança $f(x_k, a)$ e variância $S_{y_k}^2$. Neste caso, a densidade de probabilidade de obter o resultado $\overline{y_k}$ é

$$p(y_k) = \frac{1}{S_{y_k}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\overline{y_k} - f(x_k,a))^2}{S_{y_k}^2}}$$

Na hipótese de que as diferentes observações são independentes, a densidade de probabilidade de obter os resultados $\overline{y_1}, \overline{y_2}, ..., \overline{y_n}$ é proporcional a

$$\exp\left(-\frac{1}{2}\sum_{k=1}^{n}\frac{(\overline{y_k} - f(x_k,a))^2}{S_{y_k}^2}\right)$$

O método dos mínimos quadrados portanto maximiza a densidade de probabilidade, e por esta razão é também chamado *princípio da máxima verosimilhança.*

Em teoria, desde que a função f seja diferenciável, os valores de α são obtidos calculando as derivadas parciais $\partial Q_a^2/\partial a_j$ e resolvendo o sistema de m equações $\partial Q_a^2/\partial a_j = 0$ com $j = 1, 2, ..., m$. 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. A propagação dos erros permite também estimar as incertezas nos parâmetros, na forma $a = \alpha \pm S_\alpha$.

O método dos mínimos quadrados estima os parâmetros e portanto produz a conjectura

$$y = f(x, \alpha)$$

que os estatísticos chamam *curva de regressão.*

O problema é que o método funciona sempre, independentemente da forma de f e dos valores das observações! A posteriori convém avaliar a qualidade do ajuste, com base no bom senso e na honestidade do cientista. O ajuste pode ser considerado bom se os valores de $f(x_k, \alpha)$ estão todos nos intervalos $\overline{y_k} \pm S_{y_k}$, ou se pelo menos não se afastam dos $\overline{y_k}$ por mais de que multiplos pequenos de S_{y_k}. Também, é boa norma verificar que a sequência dos sinais dos erros $\overline{y_k} - f(x_k, \alpha)$ não mostra um padrão suspeito.

Esta consideração ingénua pode ser quantificada, e dá origem ao

Teste do qui-quadrado. O valor de

$$Q_\alpha^2 = \sum_{k=1}^{n}\frac{(\overline{y_k} - f(x_k,\alpha))^2}{S_{y_k}^2} = \min_a \sum_{k=1}^{n}\frac{(\overline{y_k} - f(x_k,a))^2}{S_{y_k}^2}$$

é uma medida da qualidade do ajuste. De facto, se α são os valores verdadeiros dos a, então a hipótese gaussiana implica que Q_α^2 tem lei qui-quadrado χ_{n-m}^2. Uma tabela fornece a probabilidade

$$q = \mathbb{P}\left(Q^2 > Q_\alpha^2\right)$$

onde Q^2 é uma variável com lei χ_{n-m}^2. A interpretação é: q é a probabilidade de observar um qui-quadrado maior do que foi observado na hipótese "$y = f(x, \alpha)$". Se acreditamos que os dados são uma sequência típica de valores das variáveis y_k com lei $N\left(f(x_k, \alpha), S_{y_k}^2\right)$, a probabilidade q não deve ser demasiado pequena.

Os físicos consideram aceitáveis valores $q \geq 0.1$ (esta regra é equivalente a aceitar a hipótese "a lei $y = f(x, \alpha)$ é verdadeira" com nível de significância da ordem de 10%, um valor típico de um teste acerca de uma hipótese conservadora). Se as variâncias $S_{y_k}^2$ foram subestimadas, o se os dados não são gaussianos, pode até acontecer que bons modelos levem a valores $q \simeq 0.001$. Por outro lado, valores grandes de Q_α^2, por exemplo $q \ll 0.001$, são fortes indícios de que a conjectura $y = f(x, \alpha)$ não é uma lei que descreve bem os dados observados.

Se só temos uma observação de y_k para cada valor x_k, não temos uma estimação credível das variâncias $S_{y_k}^2$. O que os físicos fazem nesse caso é pôr as variâncias iguais a 1 nas fórmulas acima, portanto minimizar

$$\sum_{k=1}^{n}(y_k - f(x_k,\alpha))^2 = \min_a \sum_{k=1}^{n}(y_k - f(x_k,a))^2$$

e depois estimar

$$S_{y_k}^2 \simeq \frac{1}{n-m}\sum_{k=1}^{n}\left(y_k - f\left(x_k,\alpha\right)\right)^2$$

(o que significa fazer a hipótese de que as $S_{y_k}^2$ são todas iguais). A partir destas variâncias é possível, usando a fórmula da propagação dos erros, estimar os erros nos parâmetros α.

Regressão linear. Modelos que são tratáveis analiticamente são os modelos lineares, tais que $f\left(x,a\right)$ depende linearmente dos parâmetros a, porque minimizar os desvios quadráticos é equivalente a resolver um sistema de equações lineares.

Um exemplo simples é uma lei linear

$$y = a + bx$$

entre os observáveis x e y.

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 x-y, e observar se estão todos perto de uma recta.

Um modelo da experiência é a hipótese gaussiana: 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 suposto 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 os estimadores α e β para os parâmetros a e b de maneira tal que a soma dos "erros quadráticos"

$$Q_{a,b}^2 = \sum_{K=1}^{n}\text{erro}_k^2 = \sum_{k=1}^{n}\left(a + bx_k - y_k\right)^2$$

seja a menor possível. Resolvendo o sistema de equações

$$0 = \partial Q_{a,b}^2/\partial a \qquad \Rightarrow \qquad 0 = n\overline{y} - n\alpha - \beta\overline{x}$$

$$0 = \partial Q_{a,b}^2/\partial b \qquad \Rightarrow \qquad 0 = \sum_{k=1}^{n} y_k x_k - n\alpha\overline{x} - \beta\sum_{k=1}^{n} x_k x_k$$

obtemos a resposta

$$\beta = \frac{S_{xy}^2}{S_{xx}^2} \qquad \alpha = \overline{y} - \beta\overline{x}$$

onde

$$S_{xy}^2 = \sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}) \quad \text{e} \quad S_{xx}^2 = \sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})$$

A recta estimada

$$y = \alpha + \beta x$$

é chamada *recta de regressão*.

Na hipótese gaussiana, α e β são bons estimadores de a e b respectivamente, porque α tem lei $N\left(a, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}^2}\right)\right)$ e β tem lei $N\left(b, \sigma^2/S_{xx}^2\right)$. Naturalmente não sabemos o valor de σ^2, mas um seu estimador é a *variância residual*

$$S_{\text{res}}^2 = \frac{1}{n-2}\sum\left(\alpha + \beta x_k - y_k\right)^2$$

A variável $\frac{n-2}{\sigma^2}\cdot S_{\text{res}}^2$ tem lei qui-quadrado χ_{n-2}^2. Se definimos

$$S_\alpha = S_{\text{res}}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}^2}} \qquad \text{e} \qquad S_\beta = \frac{S_{\text{res}}}{S_{xx}}$$

o resultado final é que

$$\frac{\alpha - a}{S_\alpha} \quad \text{e} \quad \frac{\beta - b}{S_\beta}$$

têm lei de Student t_{n-2}. Intervalos de confiança de nível $1-\varepsilon$ pelos parâmetros da lei são portanto

$$a = \alpha \pm t_{1-\varepsilon/2} \cdot S_\alpha \qquad \text{e} \qquad b = \beta \pm t_{1-\varepsilon/2} \cdot S_\beta$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student t_{n-2}.

Teste sobre independência (linear). Como decidir que $y = a + bx$ é mesmo uma lei, ou seja que o observável y depende, e linearmente, do observável x? Uma primeira ideia é testar a hipótese $b = 0$, ou seja a hipótese conservadora de que "não há evidência experimental de dependência linear entre x e y". Fixado um nível de significância ε, a região crítica do teste é

$$|\beta / S_\beta| > t_{1-\varepsilon/2}$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student t_{n-2}. Portanto, admitimos que a variável y depende (e linearmente) de x se encontramos um valor $\beta > t_{1-\varepsilon/2} \cdot S_\beta$. Para valores típicos do nível de significância, 10% ou 5%, este limite é da ordem de duas ou três vezes $S_\beta = S_{\mathrm{res}}/S_{xx}$, a razão entre as incertezas nas variáveis $(y_k - \alpha + \beta x_k)$ e x_k, o que é muito razoável.

Qualidade do ajuste. 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 os estimadores de b' e a' são

$$\beta' = \frac{S_{xy}^2}{S_{yy}^2} \qquad \text{e} \qquad \alpha' = \overline{x} - \beta' \overline{y}$$

onde

$$S_{yy}^2 = \sum_{k=1}^{n} (y_k - \overline{y})(y_k - \overline{y})$$

A relação teórica entre b e b' é $bb' = 1$. O produto

$$R^2 = \beta\beta' = \frac{S_{xy}^4}{S_{xx}^2 S_{yy}^2}$$

é dito *coeficiente de determinação*, e toma valores $0 \leq R^2 \leq 1$. A qualidade do ajuste pode ser considerada boa se R^2 está próximo de 1. Por outro lado, é razoável suspeitar que observar um R^2 pequeno, mais perto de 0 que de 1, é indício de que alguma coisa não está a correr bem.

Os estatísticos também dizem que R^2 é "a proporção de variabilidade de y explicada pela regressão", pois é a razão entre a "variabilidade explicada"

$$\begin{aligned} \sum_{k=1}^{n} (\alpha + \beta x_k - \overline{y})^2 &= \sum_{k=1}^{n} (\overline{y} - \beta\overline{x} + \beta x_k - \overline{y})^2 \\ &= \beta^2 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \frac{S_{xy}^4}{S_{xx}^4} S_{xx}^2 = \frac{S_{xy}^4}{S_{xx}^2} \end{aligned}$$

e a "variabilidade total" S_{yy}^2. Isto é mais um argumento para esperar que $R^2 \simeq 1$. A raiz de R^2 é dita coeficiente de correlação empírico.

Correlação. Uma medida empírica da "correlação linear" entre x e y é o *coeficiente de correlação (empírico)*

$$R = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{\sum_k (x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum_k (x_k - \overline{x})^2}\sqrt{\sum_k (y_k - \overline{y})^2}}$$

que toma valores $-1 \leq R \leq 1$. O seu quadrado é o coeficiente de determinação, e o seu sinal é o sinal de β. Mais importante é observar que R é um estimador do coefficiente de correlação $\rho(x, y)$ entre as variáveis aleatórias x e y.

Um valor de R próximo de ± 1 é indício de correlação linear efectiva entre as variáveis. Um valor de R próximo de 0 é indício de que as variáveis podem ser independentes (e portanto $y = a + bx$ não é uma lei da física!). Desenhando no plano os pontos de coordenadas

$$\left(\frac{(x_k - \overline{x})}{S_x}, \frac{(y_k - \overline{y})}{S_y} \right)$$

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 conservadora "o observável y não é linearmente dependente do observável x". Os livros de estatística contêm tabelas das probabilidades p de duas amostras $\xi_1, \xi_2, ..., \xi_n$ e $\eta_1, \eta_2, ..., \eta_n$ de variáveis aleatórias independentes e identicamente distribuídas com lei normal $N(0, 1)$ ter coeficiente de correlação

$$\rho = \frac{\sum_k (\xi_k - \overline{\xi})(\eta_k - \overline{\eta})}{\sqrt{\sum_k \left(\xi_k - \overline{\xi}\right)^2}\sqrt{\sum_k (\eta_k - \overline{\eta})^2}}$$

de módulo superior a $|R|$. Os físicos consideram significativa a correlação quando $p \leq 0.05$ ou 0.01, o que é equivalente a rejeitar a hipótese de independência num teste com nível de significância 5% ou 1%.

A seguinte tabela mostra o limite inferior da região crítica $|R| > r$ deste teste para níveis de significância 5% e 1% em função de $n = 10, 20, 30, 40, 60, 80, 100$

	10	20	30	40	60	80	100
5%	0.63	0.44	0.36	0.31	0.26	0.22	0.20
1%	0.76	0.56	0.46	0.40	0.34	0.29	0.26

Por exemplo, se $n = 10$ as tabelas dizem que $\mathbb{P}(|\rho| \geq 0.76) \simeq 0.01$. Portanto, a hipótese é rejeitada, logo a correlação é considerada efectiva, com nível de significância 5% se é observado um coeficiente de correlação $|R| > 0.63$. Se $n = 100$, a correlação é considerada efectiva a partir de $|R| > 0.20$.

Cuidado: é importante lembrar que este teste só avalia a "correlação linear" entre as variáveis! Por exemplo, se as variáveis x e y verificam a identidade $x^2 + y^2 =$constante então o coeficiente de correlação esperado é 0, embora as variáveis não sejam independentes...

Previsão. A regressão linear estima os valores "mais prováveis" de a e b, e portanto a lei na forma da recta de regressão

$$y = \alpha + \beta x$$

Pode ser utilizada para fazer uma previsão do valor de y em correspondência de um certo valor x da variável independente, desde que o valor x não se afaste muito do intervalo onde fizemos as experiências. E' importante não esquecer que não temos evidência de correlação linear entre as variáveis fora do intervalo observado (a validade das lei de Kepler para sistemas planetários não garante que a força gravitacional seja proporcional a $1/r^2$ a distâncias subatómicas ou extragalácticas!).

A hipótese gaussiana implica que a a variância de $y - \alpha - \beta x$ é igual a

$$\sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}^2} \right)$$

e portanto que a variável

$$\frac{y-\alpha-\beta x}{S_{\text{res}}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}^2}}}$$

tem lei de Student t_{n-2}. Um intervalo de confiança de nível $1-\varepsilon$ para o valor $y=a+bx$ é portanto

$$y=\alpha+\beta x \pm t_{1-\varepsilon/2}\cdot S_{\text{res}}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}^2}}$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student $T(n-2)$. Observe que, como esperado, o intervalo cresce quando x se afasta da média $\overline{x}$ dos valores utilizados na regressão.

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 $y_k=ae^{bx_k}+\text{erro}_k$, não sejam identicamente distribuídos, e modelos mais cuidadosos são necessários para estimar os parâmetros de regressão.

14 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 ξ 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 ξ 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 ξ 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 $\sup_{0\le t\le 1} |B(t)|$ pode ser calculada (embora de uma maneira não elementar), e em particular

$$\mathbb{P}\left(\sup_{0\le t\le 1} |B(t)| > d\right) = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2d^2}$$

Na prática isto quer dizer que temos uma estimação dos valores d_ε 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 ξ" com nível de significância ε: uma região crítica é

$$\sqrt{n}D_n > d_\varepsilon$$

Os limites inferiores d_ε 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

$$F_n(t) = \frac{1}{n}\text{card}\{k = 1, 2, ..., n \text{ t.q. } x_k \le t\} \qquad \text{e} \qquad G_m(t) = \frac{1}{m}\text{card}\{k = 1, 2, ..., m \text{ t.q. } y_k \le t\}$$

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 $\sup_{0\le t\le 1} |B(t)|$. Portanto, a região crítica de um teste sobre a nossa hipótese com nível de significância ε é

$$\sqrt{\frac{nm}{n+m}}D_{n,m} > d_\varepsilon$$

Observe que este teste não utiliza nenhuma hipótese acerca da lei.

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-Verlagr, 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.