

Coletânea de artigos sobre probabilidade básica



Alexandre Kawano

2022

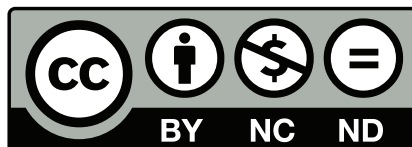
Coletânea de artigos sobre probabilidade básica

Alexandre Kawano

ORCID <https://orcid.org/0000-0002-2248-6422>

2022

“É permitida a reprodução parcial ou total desta obra, desde que citada a fonte e autoria, proibindo qualquer uso para fins comerciais.”



Este trabalho está licenciado com uma Licença Creative Commons - Atribuição-Não Comercial-SemDerivações 4.0 Internacional.

UNIVERSIDADE DE SÃO PAULO -USP

Reitor: Prof. Dr. Carlos Gilberto Carlotti Junior

Vice-reitora: Profa. Dra. Maria Armanda do Nascimento Arruda

ESCOLA POLITÉCNICA -EP

Diretor: Prof. Dr. Reinaldo Giudici

Vice-diretor: Prof. Dr. Sílvio Ikuyo Nabeta

Catálogo-na-publicação (CIP). Serviço de Biblioteca e Documentação
Escola Politécnica da Universidade de São Paulo. Maria Cristina Olaio
Villela CRB 1338/8^a

Kawano, Alexandre

Coletânea de artigos sobre probabilidade básica / A.Kawano. –
São Paulo : Epusp, 2022.

1651 kb.

ISBN 978-65-89190-15-8

DOI 10.11606/9786589190158

1. Probabilidade 2. Modelos Probabilísticos 3. Textos Didáticos.

CDU 519.21

Conteúdo

<i>Prefácio</i>	2
<i>Agradecimentos</i>	84
Cap.1. Enumerabilidade e uma ligação entre as distribuições Uniforme e Geométrica	4
Cap.2. Valor esperado, distribuições Geométrica e Exponencial	23
Cap.3. A variância, a distribuição Normal e a marcha reversa no tempo	41
Cap.4. Independência e o problema das três portas	58
Cap.5. A Lei dos Grandes Números e a representação binária de números no intervalo $[0, 1[$	76

Prefácio

Este livro contém quatro artigos escritos no período que chamarei de reclusão, durante a pandemia de Covid-19 em 2020 e mais um sobre a uma conexão entre a Lei dos Grandes Números e a distribuição de “0”s e “1”s na representação binária de números reais, escrito no ano do retorno presencial, 2022.

As aulas presenciais na Universidade de São Paulo foram suspensas e todos passaram ao modo *on-line*. Muitos de nós, docentes, produzimos vídeos e continuamos a dar aulas remotamente. Entretanto, senti que aulas à distância são frias, e principalmente para um curso de Probabilidades que há uma certa componente de show de mágica, fiquei sentindo um vazio.

Então comecei a escrever algumas notas sobre o conteúdo do curso, desde a sua base fundamental. Essas notas estão em formato de artigos, com foco em um público formado pelos meus próprios alunos, que já tiveram alguma exposição à teoria, enquanto as aulas presenciais não haviam sido interrompidas.

Cada um dos artigos tem foco em um grande fundamento da teoria. No primeiro, falo da enumerabilidade, que é crucial para se entender distribuições de probabilidade e a distinção entre variáveis aleatórias discretas e contínuas. No segundo analiso o conceito de valor esperado, que não é trivial, apesar de intuitivo. No terceiro, o foco é sobre a variância e a origem da distribuição Normal. No quarto analiso o problema clássico das três portas, e aproveito para discutir os conceitos de independência e de correlação. No último artigo, nos debruçamos sobre a Lei dos Grandes Números e a distribuição de “0”s e “1”s na representação binária de números reais no intervalo $[0, 1[$.

A escolha do formato, de agrupar os capítulos em artigos, se deve ao desejo de estabelecer conexões entre ideias apresentadas separadamente em cursos tradicionais. Assim, por exemplo, no primeiro capítulo/artigo são colocados lado a lado a distribuição Uniforme com a Geométrica para explorar o conceito de enumerabilidade.

O conteúdo não cobre o de um curso completo, e nem foi a minha intenção. Já há ótimos livros nas bibliotecas. Aqui focalizo a atenção em alguns tópicos que acho interessantes, quando vistos pela segunda vez. Quis reproduzir a sensação agradável quando revisito uma língua estrangeira através de um autor que lança luz em tópicos que nos fazem pensar de novo. Um exemplo é o livro “Gone Fishin’: New Angles on Perennial Problems” de Jay Rubin, que revisita o japonês.

Evitei introduzir assuntos novos, fora de um curso padrão para alunos de um curso de Engenharia, então não falo explicitamente de medida, nem de σ -álgebra, nem integral de Lebesgue, embora trabalhemos exatamente com esses elementos

no texto. O único tema novo, que normalmente não é visto na graduação é o da entropia no sentido de Claude Shannon, que foi incluído somente porque achei que caísse bem no gosto do público alvo.

O espírito em cada um dos artigos foi o de procurar uma sensação de “encantamento” nos leitores. Por isso, este pequeno livro, que coleta algumas notas para um curso de probabilidades, é dedicado ao Professor José Augusto Pentado Aranha, que sempre foi uma inspiração nesse sentido. Gostaria que fosse também o meu lema encantar pela ciência e pelo conhecimento acima de tudo.

Capítulo 1

Enumerabilidade e uma ligação entre as distribuições Uniforme e Geométrica*

Alexandre Kawano
Universidade de São Paulo

Abril de 2020

Resumo

Neste trabalho mostramos que duas interpretações para o experimento subjacente ao espaço probabilístico ligado à distribuição Geométrica, uma quando o experimento pára imediatamente após o primeiro sucesso, e outra quando o experimento prossegue interruptamente, induzem espaços amostrais, espaços de eventos e funções probabilidade qualitativamente diferentes, uma com espaço amostral enumerável e outro com um não enumerável. O caso não enumerável leva aos números binários e ao cálculo da distribuição de probabilidades da variável aleatória geométrica por meio da distribuição uniforme. Aproveitamos para fazer uma grande revisão dos tópicos fundamentais da Teoria das Probabilidades básica. O público deste trabalho é o dos alunos de graduação com alguma exposição prévia aos conceitos probabilísticos.

1 Introdução

As distribuições Uniforme e Geométrica fazem parte de qualquer curso básico sobre probabilidades em nível universitário, inclusive naquele que é oferecido aos alunos do terceiro semestre dos cursos da Escola Politécnica da Universidade de São Paulo.

Neste pequeno texto em que trazemos uma ligação entre essas duas distribuições de probabilidade tem por público alvo alunos de graduação de cursos de Engenharia, mas pode ser útil também para alunos dos cursos de Matemática, Física e Economia. Faremos aqui uma grande incursão às bases dessa Teoria ligada aos nomes de Fermat (1601–1665), Pascal (1623–1662), Jacob Bernoulli (1654–1705), Laplace (1749 –1827) e Kolmogorov (1903 –1987), entre outros.

*Alexandre Kawano, *Coletânea de artigos sobre probabilidade básica*, Epusp, ISBN 978-65-89190-15-8, DOI 10.11606/9786589190158

Constata-se na apresentação dos tópicos da disciplina “Probabilidades” a mesma ruptura que aparece entre a base matemática e os assuntos mais aplicados. Esse é o caso, por exemplo, de Mecânica dos Fluidos que deveria usar os teoremas já apresentados nos cursos anteriores de Cálculo, mas não o faz, preferindo mostrar aos alunos versões próprias, cujas derivações fazem apelo à intuição física. A mesma atitude acontece com os cursos clássicos de Sistemas Dinâmicos para Engenharia ao usar a transformada de Laplace, Teoria das Estruturas quando não faz uso dos teoremas da Álgebra Linear, para citar apenas alguns exemplos.

Essa tendência à ruptura na apresentação de conceitos matemáticos e aplicados ocorre de forma aguda dentro das disciplinas de “Probabilidades para Engenharia”. Depois de os conceitos de Espaço Amostral Ω , de Eventos \mathcal{F} e de Probabilidades P serem apresentados, ao se apresentar variáveis aleatórias com diversas distribuições, não são feitas as devidas conexões com a base já construída. Isto é, o conceito de espaço probabilístico (Ω, \mathcal{F}, P) cai em um limbo.

No presente texto, tentamos reconectar as distribuições de probabilidade com suas bases, passando por uma discussão sobre a enumerabilidade de conjuntos para visualizarmos bem a diferença entre distribuições discretas e contínuas.

As discussões sobre a natureza da variável aleatória ser discreta ou contínua e sobre as bases do espaço probabilístico (Ω, \mathcal{F}, P) levam naturalmente à uma ligação entre as variáveis aleatórias com distribuições Uniforme e Geométrica, a primeira sendo contínua e a segunda discreta.

A apresentação assume que o leitor já foi exposto a algum curso de probabilidades, conhecendo variáveis aleatórias e distribuições de probabilidade em um nível intuitivo. À medida que o texto progride, vamos revisando e aprofundando os conhecimentos sobre a Teoria. Por essa razão, aparecem termos como “probabilidade”, “variáveis aleatórias”, etc. antes das suas respectivas apresentações mais precisas. Usamos aqui a axiomatização devida à Kolmogorov [4, 2].

2 As distribuições Uniforme e Geométrica

Começamos apresentando as variáveis aleatórias Uniforme e Geométrica de maneira rápida e sem muitos detalhes para depois irmos formalizando as ideias ao longo do texto.

Dizemos [2] que uma variável aleatória U tem distribuição Uniforme no intervalo $[a, b] \subset \mathbb{R}$, $a < b$, se a sua densidade de probabilidade for dada por

$$f_U(x) = \frac{1}{b-a}, \quad (1)$$

isto é, se a probabilidade $P(U \in [c, d])$ com $[c, d] \subset [a, b]$ puder ser calculada por

$$P(U \in [c, d]) = \int_c^d f_U(x) dx. \quad (2)$$

Adiante vamos precisar melhor os termos, inclusive o que entendemos pelo sinal de integração, pois há aspectos importantes a serem considerados.

A razão do nome “Uniforme” é clara. A distribuição Uniforme está associada a experimentos em que se deseja fazer a observação em um experimento probabilístico de uma característica que tem ligada a ela uma descrição numérica, e tal resultado pode cair entre dois extremos $a < b \in \mathbb{R}$. Além disso, a probabilidade de se obter um número em um intervalo $[c, d] \subset [a, b]$ é proporcional ao comprimento $d - c$, não importando a localização desse intervalo dentro de $[a, b]$. O que importa é apenas o comprimento do intervalo.

É muito comum dizer-se que ela surge quando não há nenhuma informação sobre a variável aleatória U a não ser que ela pode assumir valores no intervalo $[a, b]$. Isso é verdade do ponto de vista da teoria da informação de Shannon (apresentamos e a usamos em [6] [7]), mas deve-se ter cuidado ao fazer essa afirmação. Se não há informação sobre U , também não haveria informação sobre U^2 . Assim, se, por exemplo, ladrilhos quadrados são produzidos em uma fábrica, não havendo informação sobre os seus lados, a não ser que tem comprimento entre dois valores extremos, digamos entre $11.9cm$ e $12.1cm$, então, uma distribuição uniforme para o lado implicaria uma distribuição não uniforme para a área dos ajulejos, para a qual também não temos informação. Essa é uma variante do paradoxo de Bertrand (veja por exemplo [3]).

A distribuição Geométrica é bem conhecida. De modo seco, dizemos que uma variável aleatória X tem distribuição Geométrica se ela assume somente valores em \mathbb{Z}_+ e sua distribuição de probabilidades é dada por

$$P(X = n) = (1 - p)^n p,$$

para $p \in [0, 1]$.

As designação “Geométrica”, vem do fato que a sequência $P(X = 0)$, $P(X = 1)$, ... forma uma progressão geométrica. Uma variável aleatória com distribuição Geométrica ocorre naturalmente ao considerarmos um experimento probabilístico que consiste em uma sucessão de ensaios de Bernoulli independentes, em que a probabilidade de sucesso é $p \in [0, 1]$. Neste caso, tal variável aleatória fornece o número de tentativas antes do primeiro sucesso.

2.1 Duas variantes para o experimento que leva à distribuição Geométrica

Podemos imaginar duas variantes para o experimento probabilístico subjacente à variável aleatória com distribuição Geométrica:

- Após a ocorrência do primeiro sucesso, o experimento cessa imediatamente.
- Os ensaios de Bernoulli continuam indefinidamente, sem nunca parar.

A própria natureza distinta de como o experimento é realizado nos dá uma oportunidade aprofundarmos vários aspectos da Teoria de Probabilidades.

Em primeiro lugar, estamos tratando de uma Variável Aleatória *discreta* que é dita ter a distribuição Geométrica.

3 Fundamentos

Para prosseguirmos realizamos aqui uma pequena recordação sobre os fundamentos da Teoria das Probabilidades.

Os elementos basilares são

- Espaço Amostral Ω dos resultados possíveis do experimento.
- Espaço dos Eventos $\mathcal{F} \subset \Omega$ de subconjuntos para os quais serão atribuídas probabilidades.
- Função Probabilidade $P : \mathcal{F} \rightarrow [0, 1]$.
- Variável Aleatória.

Analisemos cada um dos itens para o caso de uma variável aleatória ter com distribuição Geométrica.

3.1 Espaço amostral

Nesta seção analisaremos o espaço amostral associado à distribuição Geométrica. Pode-se fazer uma análise análoga para a variável aleatória com distribuição Uniforme. Brevemente, o experimento probabilístico associado à distribuição uniforme é tipificado pelo sorteio de um número em um intervalo não vazio $[a, b] \subset \mathbb{R}$. Dessa forma, o espaço amostral da Uniforme é simplesmente o intervalo $\Omega_U = [a, b]$.

As duas variantes para o experimento que leva à distribuição Geométrica, na primeira cessando os ensaios de Bernoulli após o primeiro sucesso, e nunca parando na segunda, levam a dois espaços amostrais, respectivamente Ω_1 e Ω_2 de natureza qualitativa diferente.

Em ambas as variantes, vamos supor que os ensaios de Bernoulli que compõe o experimento podem produzir os resultados 1 e 0. O sucesso é representado pelo algarismo 0.

O espaço amostral Ω_1 é formado por seqüências *finitas* do tipo 111...10. Por outro lado, o espaço amostral Ω_2 é composto por seqüências *infinitas* do tipo 1010110010.... Formalmente, temos

$$\Omega_1 = \{ \underbrace{111 \dots 1}_n 0 \mid n \in \mathbb{Z}_+ \},$$

n algarismos

$$\Omega_2 = \{ \text{Seqüências infinitas de 0's e 1's} \},$$

Note que qualitativamente, Ω_1 e Ω_2 apesar de semelhantes à primeira vista, são muito diferentes. Enquanto Ω_1 é enumerável, Ω_2 não é.

Podemos enumerar Ω_1 , usando o número de 1's que precedem o 0 final em cada elemento desse conjunto. Por outro lado, podemos ver que Ω_2 não é enumerável usando os números binários. Podemos colocar um "0." na frente de toda seqüência pertencente a Ω_2 e encarar o resultado como um número real pertencente ao intervalo $[0, 1]$.

Por exemplo, tomando $\omega = 1000110\dots \in \Omega_2$, temos o número binário $0.101011\dots$ construído a partir de ω . Tal número binário corresponde ao número 0.671875 na representação decimal (base 10).

$$1 \times \frac{1}{2^1} + 0 \times \frac{1}{2^2} + 1 \times \frac{1}{2^3} + 0 \times \frac{1}{2^4} + 1 \times \frac{1}{2^5} + 1 \times \frac{1}{2^6} = 0.671875.$$

Devemos tomar cuidado ao considerar as representações binárias que representam o mesmo número real no intervalo $[0, 1]$. Por exemplo, o número $0.5 = \frac{1}{2}$ pode ser representado de duas formas diferentes: $0.1000000000\dots$ ou $0.0111111111\dots$. Esse fenômeno não deve espantar o estudante, pois mesmo na familiar representação decimal, temos o mesmo fenômeno. O número decimal $0.999999\dots$ é outra representação do número 1. Todo número no conjunto $[0, 1] \setminus \{0\} =]0, 1[$ tem duas representações binárias.

Podemos ver que o conjunto $[0, 1]$ não é enumerável usando o famoso Argumento da Diagonal de Cantor. Caso fosse enumerável, então poderíamos colocar os todos números em uma lista, conforme o esquema abaixo.

$$\begin{aligned} a_1 &= 0.a_{11}a_{12}a_{13}\dots \\ a_2 &= 0.a_{21}a_{22}a_{23}\dots \\ a_3 &= 0.a_{31}a_{32}a_{33}\dots \\ &\vdots = \quad \quad \quad \vdots \end{aligned}$$

em que $a_{ij} \in \{0, 1\}$.

Agora, construimos um número \tilde{a} cujo primeiro dígito não é a_{11} , o segundo dígito não é a_{22} , e assim por diante. O número \tilde{a} , que certamente pertence a $[0, 1]$, não está na lista apresentada, que supostamente conteria todos os números do intervalo $[0, 1]$. Logo, o conjunto $[0, 1]$, apesar de conter infinitos números, não é enumerável.

Como o conjunto $]0, 1[$ não é enumerável, o conjunto Ω_2 também não é.

Observe que os elementos do espaço amostral existem irrespectivamente de tempo, passado, presente ou futuro. Simplesmente existem no conjunto dos resultados experimentais possíveis. Fazemos uma rápida discussão sobre a questão do tempo na construção de espaços amostrais em [5], quando discutimos o fenômeno do condicionamento em probabilidades.

3.2 Eventos

O espaço de eventos \mathcal{F} é formado por subconjuntos do espaço amostral Ω que devem obedecer as seguintes regras:

- E.1) $\Omega \in \mathcal{F}$,
- E.2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
- E.3) $\{A_n\}_{n \in \mathbb{N}} \subset \mathcal{F} \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$.

Chamamos a atenção para o fato de que a união em E.3 é feita sobre uma quantidade enumerável de conjuntos.

\mathcal{F} deve ser grande o suficiente para que ele contenha todas as possíveis situações para as quais estamos interessados calcular probabilidades. É importante frisar que são os eventos, isto é, elementos de \mathcal{F} é que tem probabilidades, e não os elementos do espaço amostral.

Para a distribuição Uniforme definida no intervalo real $\Omega_U = [a, b]$, o conjunto dos eventos \mathcal{F}_U são os subconjuntos obtidos pela seguinte construção. Seja a interseção de todos os conjuntos de subconjuntos de Ω que satisfazem as regras 1, 2, 3. Essa interseção é o espaço de eventos \mathcal{F}_U . Ele é o menor conjunto de subconjuntos de Ω_U que obedece as regras E.1, E.2 e E.3 que contém todos os subconjuntos de Ω_U que podem ser escritos como uniões, interseções e complementações de uma quantidade enumerável de intervalos ordinários de $[a, b]$. Intuitivamente, podemos imaginar que colocamos em \mathcal{F}_U o próprio Ω_U , e todos os intervalos do tipo $[a, c] \subset [a, b]$, para depois aplicar as regras E.2 e E.3. O resultado seria finalmente \mathcal{F}_U . A nuance é que no caso intuitivo, o processo pára em algum ponto e no caso abstrato, quando tomamos a interseção de conjuntos, estamos já lidando com infinitas operações potencialmente em quantidade não enumerável.

O comprimento de $[c, d] \in \mathcal{F}_U$ é simplesmente $L([c, d]) = d - c$, o mesmo que de $]c, d]$, $[c, d[$ e $]c, d[$. Para um elemento A qualquer de \mathcal{F}_U , definimos seu comprimento por

$$\text{Compr}(A) = \inf \left\{ \sum_{n=1}^{+\infty} L(I_n) \mid A \subset \cup_{n=1}^{+\infty} I_n, I_n \text{ é um intervalo} \right\}.$$

Com a noção de comprimento, podemos dar um sentido para a integral que aparece em (2). Para o caso da distribuição Uniforme, podemos calcular a integral de f_U , definida em (1), sobre qualquer evento de $A \in \mathcal{F}_U$ da seguinte forma.

$$\int_A f_U(x) dx = \frac{\text{Compr}(A)}{b - a}. \quad (3)$$

Para subconjuntos de $[a, b]$ que não são membros de \mathcal{F}_U , essa integral não está definida.

Conjuntos enumeráveis tem comprimento nulo. Qualquer subconjunto enumerável $C = \{c_1, c_2, \dots\}$ de $[a, b]$ tem comprimento zero. Isso porque dado qualquer $\epsilon > 0$, podemos colocar cada c_n , $n \in \mathbb{N}$ dentro de um intervalo de comprimento $\frac{1}{2}\epsilon\frac{1}{2^n}$, por exemplo, $c_n \in]c_n - \frac{\epsilon}{4}\frac{1}{2^n}, c_n + \frac{\epsilon}{4}\frac{1}{2^n}[$. O comprimento total da cobertura é simplesmente dado por

$$\frac{1}{2}\epsilon \sum_{n=1}^{+\infty} \frac{1}{2^n} = \epsilon.$$

Como $\epsilon > 0$ é arbitrário, a conclusão é inevitável.

Lembramos que o conjunto dos números racionais \mathbb{Q} é enumerável. Há muitas maneiras de ver isso, inclusive aquelas mostradas tipicamente em salas de aula. Para nossos propósitos, basta vermos que o conjunto $\mathbb{Q} \cap [a, b]$ é enumerável. Ele o será uma vez que $\mathbb{Q} \cap [0, 1]$ também o for, pois um é obtido por escalamento e translação a partir do outro. Agora, os racionais em $[0, 1]$ são enumeráveis pois eles podem ser todos representados com número finito de dígitos, ou com por um número com dízima periódica. Cada número com dízima periódica corresponde a um número com número finito de dígitos. Por exemplo (usando base dez): $0.434378345\bar{1}$ corresponde ao número 0.4343783451 com número finito de dígitos. Logo, se os números com número finito de dígitos for enumerável, o conjunto dos racionais em $[0, 1]$ também será. Basta associar os números com dízima periódica com os com número finito de dígitos. E o conjunto dos números com quantidade de dígitos finita é claramente enumerável, porque podemos ordenar em primeiro lugar os números com um dígito, depois os com dois dígitos, e assim por diante.

Isso significa que o conjunto dos racionais em $[a, b]$ tem comprimento nulo, e assim, o conjunto dos irracionais, que não é enumerável, nesse intervalo tem comprimento $b - a$.

Em termos probabilísticos, voltando à variável aleatória U , distribuída uniformemente no intervalo $[a, b]$, a probabilidade de sortear-se um número irracional é 1 e a de um racional é *exatamente* 0.

Exemplo de subconjunto de Ω_U que não está em \mathcal{F}_U . Apesar de ser enorme, \mathcal{F}_U não é todo o conjunto das partes de Ω_U (o conjunto das partes de um conjunto A é o conjunto formado por todos os subconjuntos de A).

O estudante poderia desejar ver um subconjunto de Ω_U que não está em \mathcal{F}_U .

Se tomássemos o conjunto das partes de Ω_U como sendo o espaço de eventos \mathcal{F}_U , a existência da distribuição uniforme seria impossível. Para ver isso, precisamos tomar alguns passos.

Vamos dizer que dois números $e_1, e_2 \in [a, b]$ são equivalentes se $e_1 - e_2$ for um número racional, e escrevemos $e_1 \sim e_2$. É claro que para um número qualquer $e \in [a, b]$ vale $e \sim e$, $e_1 \sim e_2 \Rightarrow e_2 \sim e_1$ e $e_1 \sim e_2, e_2 \sim e_3 \Rightarrow e_1 \sim e_3$. Isso justifica que \sim é mesmo uma relação de equivalência. Perceba que isso seria falso se houvésssemos definido tentativamente uma relação não com a diferença entre e_1 e e_2 sendo um racional (como fizemos), mas sendo um irracional. A implicação $e_1 \sim e_2, e_2 \sim e_3 \Rightarrow e_1 \sim e_3$ seria falsa.

Particionamos $[a, b]$ em subconjuntos $A_\alpha \subset \mathcal{F}_U$, tais que $\cup A_\alpha = [a, b]$ e $A_\alpha \cap A_\beta = \emptyset$ se $\alpha \neq \beta$. Até esse ponto, cada $A_\alpha \in \mathcal{F}_U$, pois cada um deles pode ser construído a partir de uma quantidade enumerável de operações de uniões, intersecções e complementações enumeráveis de intervalos em $[a, b]$. Montamos agora um novo conjunto A tomando um elemento de cada A_α . Observe que a quantidade de conjuntos A_α não é enumerável, pois essencialmente cada A_α é caracterizado por um irracional. Assim, o novo conjunto A é não enumerável, e não está em \mathcal{F}_U , como será visto logo após o próximo parágrafo. Ele também não é o conjunto de todos os irracionais pertencentes a $[a, b]$, que estaria em \mathcal{F}_U ,

pois nesse caso ele seria o complemento de um conjunto claramente em \mathcal{F}_U .

Construímos agora os conjuntos $A_{r_n} = r_n + A$, em que $r_n \in \mathbb{Q} \cap [a, b]$. Se $m \neq n$, então $A_{r_n} \cap A_{r_m} = \emptyset$, pois se $r_n + \alpha = r_m + \beta$, então $\alpha - \beta$ seria um número racional, o que é impossível, pois A é formado tomando-se apenas um elemento de cada A_α .

Agora, $[a, b] \subset \cup_{n \in \mathbb{N}} A_{r_n}$, pois dado $x \in [a, b]$, $x \in A_\alpha$ para algum α , e daí existe um elemento $y \in A$ e $m \in \mathbb{N}$ tais que $x = r_m + y$. Além disso, o comprimento de A_{r_n} para todo $n \in \mathbb{N}$ é o mesmo, pois eles são translações do mesmo conjunto. O conjunto $[2a, 2b]$ contém $\cup_{n \in \mathbb{N}} A_{r_n}$. Logo, o comprimento de $\cup_{n \in \mathbb{N}} A_{r_n}$ deve ser maior que $b - a$ e menor que $2(b - a)$. Mas isso é impossível, pois $\cup_{n \in \mathbb{N}} A_{r_n}$ é uma união com uma quantidade enumerável infinita de conjuntos com o mesmo comprimento (iguais ao de A) e disjuntos dois a dois!

Isso significa que ao passo que certamente faz sentido o cálculo de $\int_A f_U(x) dx$ pela fórmula (3), quando $A \in \mathcal{F}_U$, ele não faria nenhum sentido se $A \notin \mathcal{F}_U$.

Espaços amostrais da Geométrica. No caso da Geométrica, como temos dois espaços amostrais, Ω_1, Ω_2 , cada um advindo de concepções distintas para o experimento probabilístico associado à distribuição Geométrica, temos também dois espaços amostrais distintos.

No caso de Ω_1 , podemos tomamos o espaço de eventos \mathcal{F}_1 como sendo igual ao conjunto das partes de Ω_1 , isto é, \mathcal{F}_1 é formado por todos os subconjuntos de Ω_1 .

Como vimos, nem sempre que podemos tomar indiscriminadamente o conjunto das partes de Ω , para qualquer espaço amostral Ω , como sendo conjunto de eventos, por uma questão de compatibilidade com a função probabilidade, que em última análise tem a ver com nossos objetivos relacionados ao modelo probabilístico. No caso de \mathcal{F}_1 , podemos tomá-lo como sendo o conjunto das partes de Ω_1 , porque ele é enumerável.

Para o espaço amostral Ω_2 , consideramos o espaço de eventos \mathcal{F}_2 obtido pela seguinte construção. Colocamos em \mathcal{F}_2 o elemento Ω_2 . Em seguida, colocamos no espaço de eventos todos os subconjuntos de Ω_2 formados por seqüências do tipo 111110010..., com os n primeiros algarismos iguais a 1, para todo $n \in \mathbb{Z}_+$, seguido do algarismo 0 (seguido de infinitos 0's e 1's). A partir daí aplicam-se as regras E.2 e E.3.

Perceba que \mathcal{F}_2 assim construído é exatamente o mesmo que \mathcal{F}_U (mais precisamente, são isomorfos), quando $[a, b] = [0, 1]$, devido à representação binária dos números em $[0, 1]$, conforme ilustra a figura 1.

3.3 Probabilidade

A função probabilidade P é definida sobre o espaço de eventos, tem por contra-domínio o conjunto $[0, 1]$. $P : \mathcal{F} \rightarrow [0, 1]$ deve obedecer

- $P(\Omega) = 1$

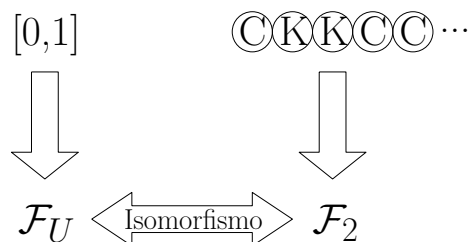


Figura 1: Isomorfismo entre os espaços de eventos da distribuição Uniforme e da distribuição Geométrica

- Se $\{A_n\}_{n \in \mathbb{N}} \subset \mathcal{F}$ são eventos mutuamente excludentes, então

$$P(\cup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} P(A_n).$$

Como temos dois espaços amostrais, Ω_1 e Ω_2 , com os respectivos espaços de eventos \mathcal{F}_1 e \mathcal{F}_2 , devemos ter também duas funções probabilidade.

$$P_1 : \mathcal{F}_1 \rightarrow [0, 1], \quad P_2 : \mathcal{F}_2 \rightarrow [0, 1].$$

O resultado da aplicação da função Probabilidade P é um número no intervalo $[0, 1]$, que está sujeito a várias interpretações. Há pelo menos três delas em nível elementar: A clássica, a frequentista, e a subjetiva [2]. Apenas para ilustrar a complexidade do tema, citamos um trecho retirado da *Stanford Encyclopedia of Philosophy* <https://plato.stanford.edu/entries/probability-interpret/>, no verbete “Interpretations of Probability” (acessado em abril de 2020).

Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means. —Bertrand Russell, 1929 Lecture (cited in [1], 587)

Não nos preocupamos com o que disse Bertrand Russell. Nós adotamos os axiomas vistos acima e apenas seguimos as suas consequências.

3.4 Variáveis aleatórias

Assim como a Probabilidade, uma variável aleatória é também uma função. Portanto, a denominação “variável aleatória” não é feliz, pois ela não é uma variável e nem aleatória. Trata-se mesmo de uma função com domínio bem definido. A sua utilidade principal reside exatamente no fato que ela é uma função, além de transformar resultados experimentais em valores numéricos.

Formalmente, dentro do contexto de um espaço probabilístico em que o espaço amostral é Ω e o espaço de eventos é \mathcal{F} , uma variável aleatória V tem por domínio o espaço amostral Ω e, neste texto, contradomínio real.

$$V : \Omega \rightarrow \mathbb{R}.$$

Devemos além disso, exigir que a imagem inversa $V^{-1}(] - \infty, a])$ seja um elemento do espaço de eventos \mathcal{F} , para todo $a \in \mathbb{R}$, para que V seja uma variável aleatória. Dessa forma faz sentido falar em probabilidades como $P(\{\omega \in \Omega \mid V(\omega) < a\})$, que escreveremos da forma abreviada $P(V < a)$. Devido à natureza do espaço de eventos, faz sentido também falar em $P(V > a)$, que na sua forma por extenso significa $P(\{\omega \in \Omega \mid V(\omega) > a\})$, ou ainda $P(V = a)$.

A variável aleatória como função permite traduzir conceitos associados a eventos, como o da independência para independência de variáveis aleatórias, que operacionalmente é mais fácil de ser verificada.

Variáveis aleatórias discretas e contínuas. Por razões didáticas, as variáveis aleatórias são apresentadas em dois tipos principais: as discretas e as contínuas. As definições desses tipos não são simétricas. As discretas são aquelas cuja imagem é um conjunto enumerável e as contínuas V são aquelas tais que existe uma função $f : \Omega \rightarrow [0, +\infty[$, denominada “Função Densidade de Probabilidade”, tal que $\int_{\Omega} f(x) dx = 1$, e a probabilidade $P(V < a)$ é dada por $P(V < a) = \int_{]-\infty, a[\cap \Omega} f(v) dv$, para todo $a \in \mathbb{R}$. Um exemplo de variável aleatória contínua é aquela associada à distribuição Uniforme. A integral que aparece aqui é sempre interpretada como sendo aquela em que se $x \mapsto f(x) \in \mathbb{R}$ for uma função constante $f(x) = K$, então $\int_{]-\infty, a[\cap \Omega} f(x) dx = K \text{ Compr}(] - \infty, a[\cap \Omega)$, para todo $a \in \mathbb{R}$. Observe que $(] - \infty, a[\cap \Omega) \in \mathcal{F}$, isto é, $] - \infty, a[\cap \Omega$ é um evento. Com essa interpretação para a integral, uma variável contínua jamais é discreta, pois a integral seria sempre nula.

Frisamos que a classificação entre discretas e contínuas é apenas didática, porque existem variáveis aleatórias que não são nem uma e nem outra. Além disso, a função densidade de probabilidade $f_X : \Omega \rightarrow [0, +\infty[$ de uma variável aleatória contínua não precisa ser contínua. Por exemplo, podemos tomar $\Omega = [0, 1]$, e f_X definida por

$$f_X(x) = \begin{cases} 0, & x \in \Omega \cap \mathbb{Q}, \\ 1, & x \in \Omega \setminus \mathbb{Q}, \end{cases}$$

que definitivamente não é contínua. Nesse caso,

$$\int_{\Omega} f_X(x) dx = 0 \times \text{Compr}(\Omega \cap \mathbb{Q}) + 1 \times \text{Compr}(\Omega \setminus \mathbb{Q}) = 1$$

e para $c \in]0, 1[$,

$$P(X < c) = \int_{\Omega \cap]-\infty, c[} f_X(x) dx = 1 \times \text{Compr}([0, c]) = c.$$

A designação “contínua” vem do fato que a função $x \mapsto \int_{-\infty}^x f(x) dx$ de uma função densidade de probabilidade $f : \mathbb{R} \rightarrow [0, +\infty[$ tem a propriedade de ser absolutamente contínua.

A questão da enumerabilidade. A enumerabilidade do espaço amostral é crucial quando discutimos variáveis aleatórias discretas e contínuas. No caso das discretas, é possível especificar-se o conjunto dos eventos como sendo o conjunto das partes do espaço amostral, e a também a probabilidade para cada evento constituído de apenas um elemento do espaço amostral. Por exemplo, se Ω é enumerável, $\omega_i \in \Omega$, $i = 1, 2, \dots$ e $\cup_{i \in \mathbb{N}} \{\omega_i\} = \Omega$, podemos, em princípio fazer $P(\omega_i) = \frac{1}{2^i}$. Entretanto, isso não é possível no caso das variáveis aleatórias contínuas. Vamos ver por quê.

Em primeiro lugar, é preciso lembrar que a união enumerável de conjuntos enumeráveis é um conjunto também enumerável. Tome Ω um conjunto. Se ele for vazio, não precisamos fazer nada. Então suponha que $\Omega \neq \emptyset$ e suponha que atribuímos uma probabilidade não nula a cada um dos seus eventos formados por apenas um elemento, isto é, se $\omega \in \Omega$, então $P(\{\omega\}) > 0$. Seja D_n o subconjunto dos elementos $\omega \in \Omega$ tais que $P(\{\omega\}) > \frac{1}{n}$. D_n deve ser finito, e portanto enumerável, pois caso contrário $\sum_{\omega \in D_n} P(\{\omega\}) > 1$. Como $\cup_{n \in \mathbb{N}} D_n = \Omega$, temos que Ω é necessariamente enumerável, o que é uma contradição.

Variável aleatória no ensino de probabilidades. A introdução do conceito de variável aleatória é muito importante, porém há um efeito colateral didático negativo. De certa forma, ela blinda todos os conceitos anteriores sobre espaço amostral, espaço de eventos e sobre função probabilidade das aplicações da teoria das probabilidades, tratando uma *função* variável aleatória como sendo de fato um resultado aleatório. Após a introdução das variáveis aleatórias, tudo se passa como se o resultado numérico da variável aleatória fosse um evento cuja probabilidade é intuitivamente calculada. A nosso ver, essa é a principal razão daquela ruptura aventada na introdução deste artigo.

Para consubstanciar essa visão, oferecemos o exemplo clássico do lançamento de um dado não viciado. Temos formalmente a descrição do experimento, espaço amostral, o espaço de eventos, a função probabilidade e a variável aleatória. Porém, para a grande maioria dos estudantes, tudo se passa como se o resultado de um lançamento, digamos o número 5 fosse o resultado de uma variável realmente aleatória, tratada como se fosse um elemento do espaço amostral, confundido com um evento, que tem probabilidade $P(V = 5) = \frac{1}{6}$. Nesse caso, tudo o que é anterior ao conceito de variável aleatória é blindado e esquecido no limbo.

4 Variável aleatória associada à distribuição Geométrica

Vamos designar a variável aleatória associada à distribuição Geométrica simplesmente por “variável aleatória Geométrica”.

Como exposto na seção 3.1, temos dois espaços amostrais Ω_1 e Ω_2 induzidos por duas concepções diferentes de experimentos, um em que os ensaios de Bernoulli param assim que ocorre o primeiro sucesso, e outro em que os ensaios são realizados sem nunca parar. Por essa razão, temos duas variáveis aleatórias,

$X_1 : \Omega_1 \rightarrow \mathbb{Z}_+$ e $X_2 : \Omega_2 \rightarrow \mathbb{Z}_+$. Ambas funções retornam o número de não sucessos antes da ocorrência do primeiro sucesso. Por exemplo, temos:

$$X_1(1110) = 3, \quad X_2(1110101010000101010\dots) = 3.$$

Note que do modo como definimos os espaços de eventos \mathcal{F}_1 e \mathcal{F}_2 , as imagens inversas $X_1^{-1}(\{a < n\})$ e $X_2^{-1}(\{a < n\})$ são sempre elementos dos respectivos espaços de eventos para todo $a \in \mathbb{R}$ (lembre-se que \emptyset é sempre um evento). Assim, faz sentido as expressões $P_1(X_1 < 5)$ e $P_2(X_2 < 2)$, por exemplo, pois o domínio da função probabilidade é o conjunto de eventos.

Quanto a ser discreta ou contínua, a variável aleatória Geométrica é discreta, qualquer que seja o experimento considerado, isto é, tanto X_1 como X_2 são ambas discretas. Lembre-se que enquanto Ω_1 é enumerável, Ω_2 não o é; ou seja, ser discreta não depende do espaço amostral, mas sim da imagem da função variável aleatória.

4.1 Cálculo da distribuição de probabilidades da variável aleatória Geométrica usando a distribuição Uniforme

Agora, calculamos as probabilidades $P_1(X_1 = n)$ e $P_2(X_2 = m)$, para $m, n \in \mathbb{Z}_+$, que são formas abreviadas de se escrever, mais corretamente, $P_1(X_1^{-1}(n))$ e $P_2(X_2^{-1}(m))$, respectivamente. Para simplificar a notação, vamos usar as formas abreviadas apenas, mas deve-se ter sempre em mente o que elas realmente significam.

As duas probabilidades $P_1(X_1 = n)$ e $P_2(X_2 = m)$ devem ter o mesmo valor caso $m = n$, o que é esperado, pois se referem essencialmente ao mesmo experimento probabilístico, embora Ω_1 e Ω_2 tenham naturezas qualitativas diferentes. Afinal de contas, deseja-se saber apenas a probabilidade da ocorrência de n não sucessos antes do primeiro sucesso, sabendo-se que cada ensaio individual tem probabilidade de sucesso igual a $p \in [0, 1]$.

4.1.1 O caso do experimento que pára assim que ocorre o primeiro sucesso (cálculo clássico sem o uso da distribuição Uniforme)

$P_1(X_1 = n)$ é a probabilidade que ocorra n não sucessos antes do primeiro sucesso. Formalmente,

$$P_1(X_1^{-1}(n)) = P_1(\underbrace{\{\omega \in \Omega_1 \mid \omega = \underbrace{111\dots 10}_{n \text{ dígitos}}\}}_{\in \mathcal{F}_1}).$$

Essa probabilidade é facilmente calculada por $P_1(X_1 = n) = (1 - p)^n p$.

4.1.2 O caso do experimento que nunca pára (cálculo com o uso da distribuição Uniforme)

Por outro lado, a probabilidade $P_2(X_2 = m)$ é dada formalmente por

$$P_2(X_2^{-1}(m)) = \underbrace{P_2(\{\omega \in \Omega | \omega = \underbrace{11 \dots 1}_m 0 n_{m+2} n_{m+3} \dots, \text{ com } n_{m+2}, n_{m+3}, \dots \in \{0, 1\}\})}_{\in \mathcal{F}_2}.$$

Observe que sabendo-se que a probabilidade de ocorrer um sucesso em um ensaio individual que compõe o experimento é $p \in]0, 1[$, a probabilidade de qualquer ponto $\omega \in \Omega$ é exatamente zero. Por exemplo,

$$P_2(100010001001\dots) = (1 - p)ppp(1 - p)ppp(1 - p)pp(1 - p)\dots = 0.$$

Vamos ilustrar o cálculo de $P_2(X_2 = m)$ por meio de um exemplo. Calculemos $P_2(X_2 = 2)$.

$$P_2(X_2^{-1}(2)) = P_2(\underbrace{\{\omega \in \Omega | \omega = 110n_4n_5 \dots, \text{ com } n_4, n_5, \dots \in \{0, 1\}\}}_{\in \mathcal{F}}).$$

Fazendo-se a identificação de cada ponto de Ω_2 com um número no intervalo real $[0, 1]$, vemos que $P_2(X_2 = 2)$ é exatamente a probabilidade de se obter um número no intervalo que vai de

$$0.1100000000\dots(\text{binário}) = \frac{3}{4}(\text{decimal})$$

ao número

$$0.110111111111\dots(\text{binário}) = \frac{7}{8}(\text{decimal})$$

Para o caso em que $p = \frac{1}{2}$, essa probabilidade é bem fácil de ser calculada, e além disso bela.

Como cada algarismo 0 ou 1 pode sair com igual probabilidade em um ensaio de Bernoulli individual ($p = \frac{1}{2}$), a probabilidade de um número ser sorteado em um intervalo (contido em $[0, 1]$) de comprimento α é exatamente igual a α (distribuição uniforme!). Uma discussão mais aprofundada sobre esse fato é feita em [8].

Pois bem, a probabilidade $P_2(X_2 = 2)$ é igual à probabilidade de sortearmos um número entre $\frac{3}{4}$ e $\frac{7}{8}$, considerando-se uma distribuição uniforme no intervalo $[0, 1]$.

Dáí:

$$P_2(X_2 = 2) = \left(\frac{7}{8} - \frac{3}{4} \right) = \frac{1}{8} = \frac{1}{2^3},$$

que é exatamente o resultado clássico, dado por $P_1(X_1 = 2)$.

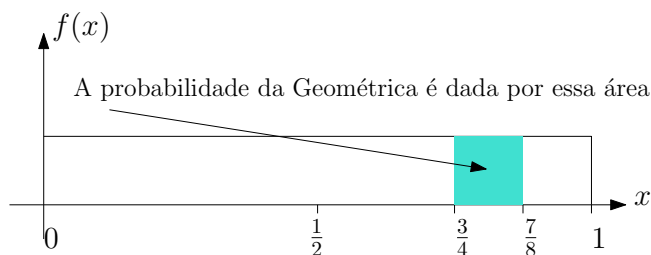


Figura 2: Distribuição uniforme.

4.2 O caso geral $p \in [0, 1]$

Calculamos agora $P_2(X_2 = 2)$ quando $p \in [0, 1]$. A probabilidade $P_2(X_2 = 2)$ permanece sendo aquela de se extrair um número no intervalo $[\frac{3}{4}, \frac{7}{8}] \subset [0, 1]$, mas a distribuição de probabilidades contínua para o sorteio de números reais dentro desse intervalo não é mais uniforme, pois na representação binária de números reais, o bit 1 sai com mais ou menos frequência que o bit 0 dependendo da probabilidade p que é atribuída dentro do contexto do experimento da variável aleatória Geométrica.

Aqui calculamos a probabilidade $P([\frac{3}{4}, \frac{7}{8}] \subset [0, 1])$ por complementação, e na seção seguinte mostramos uma outra abordagem mais apropriada para implementação algorítmica.

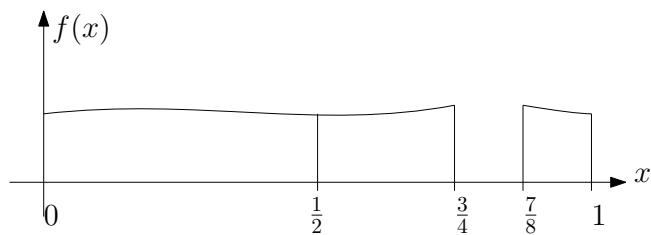


Figura 3: Cálculo por complementação

$$P(x \in [\frac{3}{4}, \frac{7}{8}]) = 1 - (P(x \in [0, \frac{3}{4}]) + P(x \in [\frac{7}{8}, 1])).$$

$$P(x \in [0, \frac{3}{4}]) = P(x \in [0, \frac{1}{2}]) + P(x \in [\frac{1}{2}, \frac{3}{4}]).$$

A probabilidade $P(x \in [0, \frac{1}{2}])$ é calculada percebendo-se que ela é igual à se obter números binários

de 0.000000.... até 0.0111111111....

Essa probabilidade é exatamente p .

A probabilidade $P(x \in [\frac{1}{2}, \frac{3}{4}])$ é igual à se obter números binários

de 0.1000000.... até 0.1011111111....

Essa probabilidade é exatamente $(1 - p)p$.

Analogamente, $P(x \in [\frac{7}{8}, 1])$, é a probabilidade de se obter números binários entre

de 0.111000000... até 0.11111111....

Essa probabilidade é exatamente $(1 - p)^3$.

Finalmente, combinamos os cálculos.

$$\begin{aligned} P(x \in [\frac{3}{4}, \frac{7}{8}]) &= 1 - (P(x \in [0, \frac{3}{4}]) + P(x \in [\frac{7}{8}, 1])) \\ &= 1 - (P(x \in [0, \frac{1}{2}]) + P(x \in [\frac{1}{2}, \frac{3}{4}]) + P(x \in [\frac{7}{8}, 1])). \end{aligned}$$

$$P(x \in [\frac{3}{4}, \frac{7}{8}]) = 1 - (p + (1 - p)p + (1 - p)^3).$$

O que fornece, novamente, a resposta clássica:

$$P(X = 2) = P(x \in [\frac{3}{4}, \frac{7}{8}]) = (1 - p)^2 p.$$

4.3 Abordagem algorítmica para o cálculo da função distribuição acumulada

Lembramos que a função distribuição de probabilidade acumulada F_X de uma variável aleatória é simplesmente a função $F_X : \mathbb{R} \rightarrow [0, 1]$ definida por

$$F_X(x) = P(X \leq x).$$

em que $f_X : \Omega \rightarrow [0, +\infty[$ é a função densidade de probabilidade de X .

Consideremos o problema de se determinar a função distribuição de probabilidade acumulada $F_{\tilde{X}}$ para a distribuição de números reais no intervalo $[0, 1]$ na situação em que quando $x \in [0, 1]$ é expresso na base 2 e as probabilidades para os dígitos 0 e 1 são respectivamente p e $(1 - p)$, conforme o experimento da distribuição Geométrica que dá origem ao espaço amostral Ω_2 .

Temos agora o seguinte experimento probabilístico, que consiste em se fazer um sorteio de um número binário $0.b_1 b_2 \dots b_n$ com n dígitos. O espaço amostral é o conjunto

$$\Omega_3 = \{(b_1 b_2 \dots b_n), b_j \in \{0, 1\}, j = 1, \dots, n\}.$$

O espaço de eventos é formado pelo conjunto das partes de Ω_3 . A variável aleatória discreta a ser considerada é a função $X_3 : \Omega_3 \rightarrow [0, 1]$ definida por

$$X_3(b) = \sum_{j=1}^n b_j \frac{1}{2^j}, \quad b = (b_1 b_2 \dots b_n).$$

A probabilidade $P(X_3 = b)$, com $b = (b_1 b_2 \dots b_n)$, é calculada por

$$P(X_3 = b) = \prod_{j=1}^n (-1)^{b_j} (p - b_j) = \left(\frac{1-p}{p} \right)^S p^n,$$

em que $S = \sum_{j=1}^n b_j$. A probabilidade de se ocorrer $b_j = 0$ é p e de ocorrer $b_j = 1$ é $(1-p)$, para $p \in [0, 1]$. A sequência $(b_j)_{j=1}^n$ é o resultado de ensaios de Bernoulli independentes.

A função distribuição de probabilidade acumulada $F_3(x) = P(X_3 \leq x)$ pode ser calculada tomando-se a representação de números reais usando-se n dígitos binários (isto é, $0.b_1 b_2 \dots b_n$ com $b_j \in \{0, 1\}$, $j = 1, \dots, n$). A função F_3 fornece uma aproximação da função distribuição de probabilidade acumulada $F_{\tilde{X}}$.

Finalmente calculamos a função distribuição de probabilidade acumulada da variável aleatória discreta X_3 por meio de

$$F_{X_3}(x) = \sum_{b \in B(x)} P(X_3 = b),$$

em que $B(x)$ é o conjunto

$$B(x) = \left\{ (b_1 b_2 \dots b_n), b_j \in \{0, 1\}, j = 1, \dots, n \mid \sum_{j=1}^n \frac{b_j}{2^j} \leq x \right\}.$$

Os resultados numéricos são para $p = \frac{1}{2}$, $p = \frac{1}{3}$ e $p = \frac{2}{3}$ são mostrados nas figuras 4, 5 e 6 respectivamente, para quando o número de dígitos n é 8.

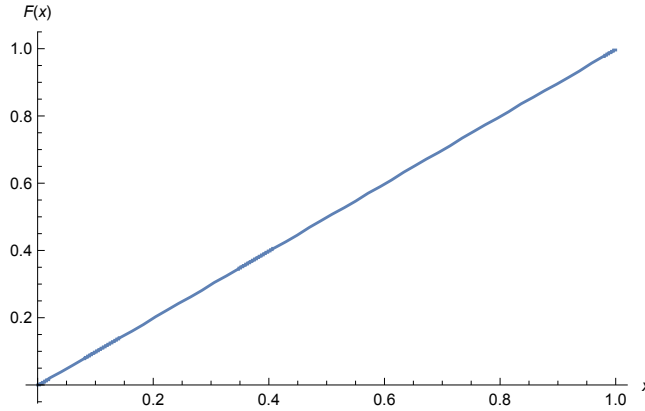


Figura 4: Função distribuição de probabilidade acumulada de X_3 para $p = \frac{1}{2}$.

Com tais curvas é possível se calcular aproximadamente, por exemplo, a probabilidade de se obter, para o experimento da distribuição geométrica, a probabilidade de se obter 4 não sucessos antes do primeiro sucesso, isto é, $P(X_2 = 4)$,

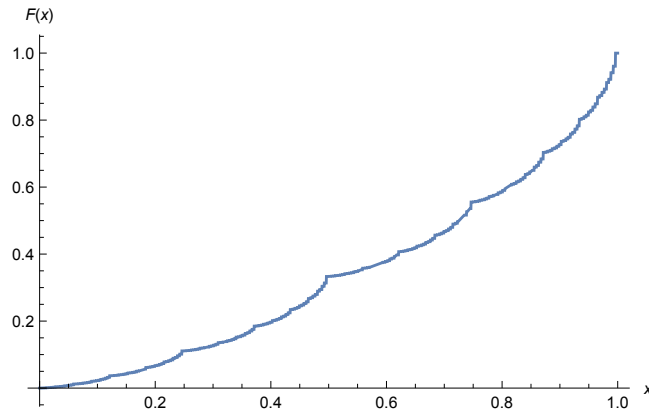


Figura 5: Função distribuição de probabilidade acumulada de X_3 para $p = \frac{1}{3}$.

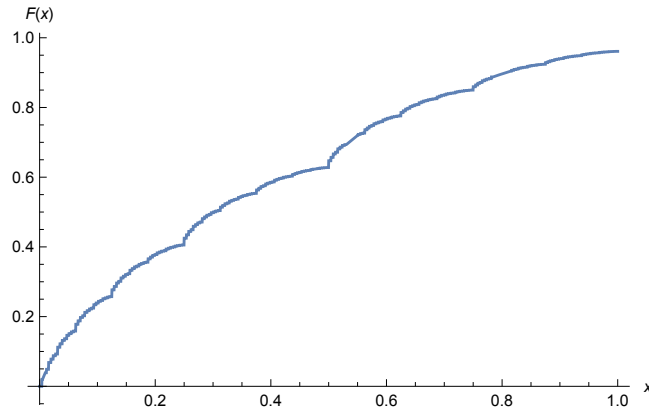


Figura 6: Função distribuição de probabilidade acumulada de X_3 para $p = \frac{2}{3}$.

quando a probabilidade de sucesso em um ensaio de Bernoulli individual é $p = \frac{1}{3}$. Nesse caso,

$$P(X_2 = 4) = F_{X_3}\left(\underbrace{\frac{31}{32}}_{0.111101111\dots}\right) - F_{X_3}\left(\underbrace{\frac{15}{16}}_{0.111100000\dots}\right).$$

O resultado é 0.065843625987932140000416389, quando se usam $n = 20$ dígitos para a representação binária. O resultado exato é como sabemos,

$$(1 - p)^4 p = 0.\overline{065843621399176954732510288} \text{ (periódica)}$$

É claro que o mérito do cálculo aproximado não é facilitar nenhuma conta, pois o cálculo direto é imediato. O importante é a discussão gerada para se

calcular as probabilidades de uma variável aleatória discreta por meio de uma contínua, e aproveitar a ocasião para se visitar pontos importantes da Teoria das probabilidades.

Para finalizar, fechamos com algumas conclusões dessa apresentação.

5 Conclusões

Neste trabalho calculamos a distribuição de probabilidades da distribuição Geométrica de duas formas diferentes, correspondentes a duas interpretações distintas para o experimento probabilístico subjacente. Uma das maneiras de se obter a distribuição de probabilidades usa a distribuição uniforme, cuja aplicação se torna natural.

Aproveitamos a oportunidade para fazermos uma grande revisão dos principais conceitos da Teoria de Probabilidades básica. Neste momento final, gostaríamos de deixar registrados alguns pontos que consideramos importantes para o estudante que eventualmente ler este texto.

- A distribuição Geométrica é discreta, não porque o seu espaço amostral é enumerável, mas porque sua *imagem* o é.
- Uma variável aleatória é mesmo uma função, com domínio no espaço amostral.
- Os elementos (pontos) do espaço amostral existem irrespectivamente de tempo, passado, presente ou futuro. Simplesmente existem em Ω .
- O cálculo das probabilidades é feito sobre elementos do espaço de eventos. Esse cálculo é fortemente influenciado pelos elementos do espaço de eventos.

Referências

- [1] E.T. Bell. *The Development of Mathematics*. McGraw-Hill Book Company, 2nd. edition, 1945.
- [2] Carlos Alberto Barbosa Dantas. *Probabilidade – Um Curso Introdutório*. EDUSP, 3rd. edition, 2013.
- [3] Laurence C. Evans. *An Introduction to Stochastic Differential Equations*. American Mathematical Society, 2013.
- [4] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 3rd. edition, 2001.
- [5] Alexandre Kawano. Independência e o problema das três portas (artigo presente nesta coletânea).

- [6] Alexandre Kawano. Valor esperado, distribuições geométrica e exponencial (artigo presente nesta coletânea).
- [7] Alexandre Kawano. A variância, a distribuição normal e a marcha reversa no tempo (artigo presente nesta coletânea).
- [8] Alexandre Kawano. A Lei dos Grandes Números e a representação binária de números no intervalo $[0, 1[$ (artigo presente nesta coletânea).

Capítulo 2

Valor esperado, distribuições Geométrica e Exponencial*

Alexandre Kawano
Universidade de São Paulo

Maio de 2020

Resumo

Neste artigo¹ exploramos o conceito de Valor esperado, e vemos como ele aparece na conexão entre a distribuição Exponencial e Geométrica. A Exponencial é obtida através de uma sequência convergente em distribuição de variáveis aleatórias Geométricas. A discussão sobre o significado do conceito de Valor Esperado leva naturalmente à interpretação que a Exponencial é aquela distribuição que contém a menor quantidade informação intrínseca, quando se conhece apenas seu o valor esperado. Aproveitamos a ocasião para explorar as bases da Teoria de Probabilidades. O texto tem por público alvo alunos de graduação que já tiveram alguma exposição a um curso de Probabilidades.

1 Introdução

De uma maneira rápida, mas não rigorosa, podemos obter a distribuição exponencial a partir da distribuição geométrica.

Suponha tenhamos um fenômeno que ocorra em média S vezes durante um intervalo de tempo I . Então temos em média $\frac{S}{I}$ sucessos por unidade de tempo.

Vamos agora imaginar que para saber se o fenômeno ocorre ou não ocorre, devemos observar sub-intervalos do intervalo I . Vamos dividir I em N sub-intervalos.

O número médio de sucessos dentro de cada subintervalo será

$$\frac{S}{I} \frac{I}{N} = \frac{S}{N}.$$

*Alexandre Kawano, *Coletânea de artigos sobre probabilidade básica*, Epusp, ISBN 978-65-89190-15-8, DOI 10.11606/9786589190158

¹A ideia deste artigo partiu de uma discussão com o Prof. Fábio Cozman da Escola Politécnica sobre a probabilidade de se extrair um número par do conjunto dos naturais.

Supomos que o sucesso por unidade de tempo seja constante.

$$\lambda = \frac{S}{I}.$$

Isto, é o número médio de sucessos por sub-intervalo é $\frac{S}{N} = \lambda \frac{I}{N}$. Quando N é grande, podemos interpretar esse número

$$\lambda \frac{I}{N} \doteq p_N \tag{1}$$

como sendo a probabilidade de termos um sucesso (encontrar o fenômeno) em um sub-intervalo de comprimento I/N .

Agora, chegamos à distribuição Geométrica. Tome um intervalo de tempo I . Subdivida-o em N sub-intervalos. Cada sub-intervalo tem comprimento I/N , em termos de tempo. Seja K_N o número de sub-intervalos antes do primeiro sucesso nesse arranjo de N subdivisões do intervalo I . Isso equivale a um intervalo de tempo igual a

$$T_N \doteq K_N \frac{I}{N}, \tag{2}$$

isto é, ao comprimento total de K_N blocos de comprimento I/N cada. Note que aqui, temos uma conexão com o mundo real, com unidades físicas. Em geral, devemos tomar I como sendo um dia, uma hora, um segundo, etc.

Do modo como toda a situação foi posta, K_N pode significar também o número de subdivisões (sem nenhum sucesso) entre um sucesso e o próximo, e T_N o tempo correspondente.

Perguntamos agora qual é a probabilidade de termos $T_N > t$.

$$P(T_N > t) = P(K_N > t \frac{N}{I}) = (1 - p_N)^{t \frac{N}{I}} = \left(1 - \lambda \frac{I}{N}\right)^{\frac{N}{I}t},$$

Subdividindo cada vez mais nosso intervalo de tempo I , temos:

$$\lim_{N \rightarrow +\infty} P(T_N > t) = \lim_{N \rightarrow +\infty} \left(1 - \lambda \frac{I}{N}\right)^{\frac{N}{I}t} = e^{-\lambda t}.$$

Quando o intervalo de tempo I for finamente subdividido, podemos inferir que, se T denota o instante de tempo do primeiro sucesso, a probabilidade que o primeiro sucesso ocorra após o tempo t é $P(T > t) = e^{-\lambda t}$. Então, sendo T uma variável aleatória, sua função de probabilidade acumulada é

$$F_T(t) = P(T \leq t) = 1 - e^{-\lambda t}.$$

A variável aleatória T tem distribuição conhecida como Exponencial.

A partir de F_T , obtemos a função densidade de probabilidade $f_T(t) = \frac{dF_T}{dt}(t) = \lambda e^{-\lambda t}$

Valor esperado e variância da Geométrica. Exploramos o conceito de valor esperado de uma variável aleatória X , mais adiante neste texto e o de variância em [4]. Por ora, basta lembrarmos que para uma variável aleatória contínua com função densidade e probabilidade f_X , o valor esperado é calculado por

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx, \quad (3)$$

e para o caso de uma variável aleatória discreta Y , o valor esperado é calculado pela fórmula

$$E[Y] = \sum_{n \in \mathbb{N}} y_n P(Y = y_n), \quad (4)$$

quando as funções $x \mapsto x f_X(x)$ e $n \mapsto y_n P(Y = y_n)$ forem respectivamente integrável e absolutamente somável.

As definições para as variâncias podem ser vistas em qualquer livro que traz a teoria básica de probabilidades, como por exemplo [1].

Observe que podemos obter o valor esperado e a variância de T a partir das fórmulas para a distribuição geométrica. Do fato que K_N é a variável aleatória da Geométrica com parâmetro p_N , dada por (1), temos que o número de não sucessos antes do primeiro sucesso tem valor esperado e variância dadas por [1]

$$E[K_N] = \frac{1 - p_N}{p_N}, \quad V[K_N] = \frac{1 - p_N}{p_N^2}.$$

Agora transformamos K_N em termos de tempo, usando (2), obtendo

$$E[T_N] = \left(\frac{I}{N}\right) \frac{1 - p_N}{p_N}, \quad V[T_N] = \left(\frac{I}{N}\right)^2 \frac{1 - p_N}{p_N^2}. \quad (5)$$

Passando ao limite $N \rightarrow +\infty$, obtemos

$$E[T] = \lambda \quad e \quad V[T] = \frac{1}{\lambda^2}. \quad (6)$$

Propriedade da falta de memória da Geométrica. A propriedade da *falta de memória* da variável aleatória exponencial é consequência direta da Geométrica, pois evidentemente a Geométrica tem essa propriedade, sendo criada a partir de uma sequência de ensaios de Bernoulli independentes.

Intuitivamente a propriedade da falta de memória da Geométrica é clara, pois o que ocorre até o n -ésimo resultado não interfere no que ocorre nos resultados subsequentes. Do ponto de vista matemático fica assim:

$$\begin{aligned} P(T_N > t + t_0 | T_N > t_0) &= \frac{P(T_N > t + t_0)}{P(T_N > t_0)} = \frac{(1 - p_N)^{(N/I)(t+t_0)} p_N}{(1 - p_N)^{(N/I)t_0} p_N} \\ &= (1 - p_N)^{(N/I)t} = P(T_N > t). \end{aligned}$$

Passando ao limite, sabendo que $T_N \rightarrow T$ em distribuição, isto é que $P(T_N \leq t) \rightarrow P(T \leq t)$ (só. Trata-se de modo de convergência muito fraco), obtemos a propriedade de falta de memória também para a Exponencial.

$$P(T > t + t_0 | T > t_0) = P(T > t).$$

A propriedade de falta de memória reforça a interpretação que T pode ser encarado também como a variável aleatória que dá o tempo entre um sucesso e outro.

As ideias que usamos para obter a exponencial a partir da Geométrica são certamente válidas. Elas tem um sabor de um estilo anterior ao formalismo matemático que a humanidade alcançou após o início do século vinte.

Na apresentação há pontos que são sujeitos à questionamentos que vão além da questão formal, que em muitas vezes recai simplesmente a questões de pedantismo. Aproveitamos a situação para aprofundarmos os conceitos básicos da Teoria de Probabilidades.

2 A distribuição Exponencial

Começamos a apresentação formal pela definição de uma variável aleatória que segue a distribuição exponencial com parâmetro $\lambda > 0$. Para isso, devemos especificar, como já fizemos no artigo [3], o espaço amostral Ω_E , o espaço de eventos \mathcal{F}_E e a função probabilidade $P : \mathcal{F}_E \rightarrow [0, 1]$.

O espaço amostral é o conjunto $\Omega_E = \mathbb{R}_+$ dos números reais positivos. O espaço dos eventos \mathcal{F}_E é o mesmo construído no artigo já citado [3], para a variável aleatória uniforme. Coloque em \mathcal{F}_E o próprio Ω_E , e todos os intervalos do tipo $]0, c[\subset \Omega_E$. Em seguida, aplique as regras: $A \in \mathcal{F}_E \Rightarrow A^c \in \mathcal{F}_E$ e $\{A_n\}_{n \in \mathbb{N}} \subset \mathcal{F}_E \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}_E$, para obter finalmente \mathcal{F}_E . O resultado é que \mathcal{F}_E contém todos os subconjuntos de Ω_E que podem ser escritos como uniões, interseções e complementações de uma quantidade enumerável de intervalos ordinários de Ω_E .

Definição 2.1. A variável aleatória T de distribuição Exponencial com parâmetro $\lambda > 0$ é a função $T : \Omega_E \rightarrow \mathbb{R}_+$ que tem a seguinte distribuição de probabilidades $P : \Omega_E \rightarrow [0, 1]$:

$$P(A) = \int_A \lambda e^{-\lambda t} dt.$$

Da definição, obtemos a função de probabilidade acumulada,

$$F_T(t) = P([0, t]) = \int_0^t \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t},$$

que é a mesma vista na introdução deste texto.

O valor esperado $E[T]$ e a variância $V[T]$, que são dadas respectivamente por

$$E[T] = \frac{1}{\lambda} \quad e \quad V[T] = \frac{1}{\lambda^2},$$

que são as mesmas vistas em (6).

3 Discussão sobre as bases da Teoria

Todas as propriedades básicas da variável aleatória Exponencial, vistas em cursos elementares de probabilidades, foram mostradas. Mas resta-nos esclarecer pontos cruciais da manipulação feita na Introdução. Aproveitamos a ocasião para aprofundarmo-nos na Teoria.

3.1 A questão crucial do valor esperado

O primeiro ponto crucial é sobre a interpretação do parâmetro $\lambda > 0$ dada na Introdução. Lá, ele foi interpretado com a fração de sucessos por unidade de tempo, e como consequência, a probabilidade de sucesso em um ensaio de Bernoulli foi computada como $p_N = \lambda \frac{1}{N}$, que é a quantidade média de sucessos por sub-intervalo de tempo.

E de fato, com essa interpretação, sem fazer nenhuma conta, podemos intuir que o valor esperado para a variável aleatória T , que é intuitivamente o tempo médio até o primeiro sucesso, seja exatamente $1/\lambda$: Se a média do número de sucessos em cada intervalo de tempo for igual a λ , então a intuição aponta que devemos esperar em média $1/\lambda$ intervalos de tempo até o primeiro sucesso, que é justamente o que nos informa a fórmula (6)!

Mas exatamente aqui está o ponto crucial da *interpretação do que seja número médio de sucessos por intervalo de tempo* e associá-lo a uma probabilidade. Consideremos, por exemplo, a questão de se saber qual é a probabilidade de se extrair completamente ao acaso um número par do conjunto dos números naturais:

$$1, 2, 3, 4, 5, 6, 7, 8, 9, \dots$$

O estudante poderia ser tentado a responder que essa probabilidade é $\frac{1}{2}$, pois o número de sucessos (o número ser par) por unidade de comprimento, tomado como quantidade de números, é meio, pois aparentemente, em média, números pares ocorrem na metade das vezes, ou ainda, quando pegamos um conjunto de N números completamente ao acaso, como que com um balde do caldeirão de números naturais, metade, isto é, $N/2$, será em média, par.

Para sabermos o número de sucessos por unidade de comprimento, devemos olhar para comprimentos longos. Veja o caso não periódico de sucessos e não sucessos (sucesso marcado com um “0” e um não sucesso por um “1”):

$$1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, \dots$$

A taxa de sucessos deve ser computada com o maior comprimento possível da *string* de números. Olhando os valores disponíveis, a taxa de sucesso da sequência acima seria aproximadamente $\frac{6}{16}$, que seria uma melhor aproximação que $\frac{1}{4}$, quando olhamos os quatro primeiros dígitos.

Voltando à questão dos números pares, devemos tomar o maior comprimento possível para a nossa *string*. Qual seria esse comprimento? Infinito. E qual seria a taxa média de sucessos? Aí está a questão crucial.

Sendo o conjunto dos naturais \mathbb{N} enumerável, mas infinito, podemos fazer um reordenamento deles da seguinte forma:

$$1, 3, 2, 5, 7, 4, 9, 11, 6, 13, 15, 8, \dots$$

$$1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, \dots \text{ ("0": sucesso, "1": não sucesso)}$$

em que nenhum número é deixado de fora da lista.

Agora, a resposta para a mesma questão sobre o número de sucessos (ser par) por unidade de comprimento é outra. Agora esse número passa a ser $\frac{1}{3}$, assim como sortear-se um número par de maneira completamente aleatória dos números naturais. E note que esse número pode ser qualquer outro!

Séries absolutamente convergentes. O mesmo fenômeno apontado acima ocorre quando computamos o valor de $\log_e 2$ usando a série de Taylor para a função $x \mapsto \log_e(x + 1)$. A série é dada por

$$\log_e(x + 1) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \dots$$

Então,

$$\log_e(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \dots, \quad (7)$$

que o estudante pode comprovar facilmente com uma calculadora.

Como consequência,

$$\frac{1}{2} \log_e(2) = \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \frac{1}{10} - \frac{1}{12} + \dots,$$

Ao somarmos as duas séries, obtemos sem mistério

$$\log_e(2) + \frac{1}{2} \log_e(2) = 1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \dots. \quad (8)$$

Essa série tem exatamente os mesmos termos da série (7), mas com um rearranjo na ordem deles. Aliás é o mesmo reordenamento de (7). Então (8) deve representar o mesmo número $\log_e(2)$. Ocorre que isso evidentemente é uma contradição, pois

$$\log_e(2) + \frac{1}{2} \log_e(2) = \log_e(2),$$

o que é um absurdo.

Na verdade, a série reordenada (8) representa tão somente o número $\log_e(2) + \frac{1}{2} \log_e(2)$ e não $\log_e(2)$. Esse exemplo mostra que não se pode reordenar séries convergentes à vontade.

Sabemos dos cursos de Cálculo, em especial do Cálculo IV, que quando a série $\sum_{n=1}^{+\infty} a_n$, $a_n \in \mathbb{R}$, $\forall n \in \mathbb{N}$, for absolutamente convergente, isto é, quando

$$\left| \sum_{n=1}^{+\infty} a_n \right| < +\infty \Rightarrow \sum_{n=1}^{+\infty} |a_n| < +\infty,$$

então $\sum_{n=1}^{+\infty} a_n$ pode ser reordenada, sempre dando o mesmo valor para qualquer ordem usada para se fazer a soma. Quando a série não for absolutamente convergente, o resultado pode depender da reordenamento dos termos da série.

Nem toda variável aleatória tem valor esperado. O que vimos acima tem uma consequência muito importante para a Teoria das Probabilidades: *Não é toda variável aleatória que tem valor esperado*, no sentido que ao calcularmos o valor esperado de uma variável aleatória, devemos exigir que o resultado não dependa da ordem em que a “média” foi calculada.

Por exemplo, a variável aleatória Y discreta que pode assumir valores

$$y_n = (-1)^{(n-1)} \frac{2^n}{n}$$

com probabilidade $P(Y = y_n) = \frac{1}{2^n}$, $n \in \mathbb{N}$, não tem valor esperado. De fato, se ele existisse, teríamos algo como

$$\sum_{n=1}^{+\infty} y_n P(Y = y_n) = \sum_{n=1}^{+\infty} (-1)^{n-1} \frac{2^n}{n} \frac{1}{2^n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \dots = \log_e(2),$$

mas esse valor depende da ordem em que a soma é computada!

Por isso, na definição de valor esperado (4) há o cuidado de se exigir que a soma tenha sentido, isto é, que não dependa da ordem em que ela é feita. Afinal, quando calculamos uma “média”, não temos nenhum critério *a priori* para ordenar os elementos do espaço amostral.

Por isso, vemos que é perigoso associar o conceito de valor esperado ao de média. Esse segundo conceito vem da ciência Estatística, em que trabalhamos com amostras e inferências sobre parâmetros probabilísticos de uma população. As amostras têm tamanho finito, e por isso, falamos em média. Na teoria de probabilidades, é mais apropriado falar em valores esperados, pois lidamos com variáveis aleatórias cujos domínios Ω tem potencialmente infinitos elementos.

Valor esperado de uma variável aleatória com distribuição Geométrica e um ordenamento *a priori*. Agora voltamos à nossa variável aleatória Geométrica subjacente à Exponencial vista na Introdução. O que significa então “termos em média S sucessos em um intervalo de tempo I ”? Aqui há o imperativo de se dar um sentido a essa frase.

Quando dizemos que temos uma média de S sucessos por intervalo de tempo I , deve haver uma outra informação latente, subentendida, que implique um certo ordenamento *a priori*, para que evitemos situações com a vista acima com

os números naturais pares. Como na Exponencial, que é uma variável aleatória contínua, não há sequências discretas, temos que usar um artifício.

Em primeiro lugar, considere uma lista $\Lambda = (t_n)_{n \in \mathbb{N}} \subset \mathbb{R}_+$, que suporemos ordenada em ordem estritamente crescente. Isso é possível porque estamos tratando com números reais. O fato de estarmos tratando com tempos, mais especificamente com o tempo até o *primeiro*² sucesso no caso da Geométrica, está implícito um ordenamento a priori. Esse não é um fato banal. Tome por exemplo o conjunto dos números complexos. Ele não admite um ordenamento a priori, pois se quisermos impor algo como $0 < i$, ou $i < 0$, em que i é o número imaginário tal que $i^2 = -1$, rapidamente chegaremos a um absurdo:

$$\begin{aligned} 0 < i &\Rightarrow 0 < -1, & (\text{absurdo}) \\ 0 > i &\Rightarrow 0 < -i \Rightarrow 0 < -1, & (\text{outro absurdo}). \end{aligned}$$

Evidentemente $0 = i$ é impossível.

A lista Λ simboliza os tempos em que acontecem os sucessos. Definimos agora a função ρ da seguinte forma.

Para cada intervalo de tempo $I =]0, L[$, seja

$$\rho(\Lambda) = \lim_{L \rightarrow +\infty} \frac{\sharp(\Lambda \cap]0, L[)}{L}, \quad (9)$$

em que $\sharp(A)$ devolve o número de elementos do conjunto A . Observe que $\rho(\Lambda)$ pode ser $+\infty$ e que Λ pode apresentar flutuações enormes, Λ pode ter ser muitas coisas, por exemplo, um sinal em código Morse, um sinal extra-terrestre, um processo de Poisson gerado por um professor de probabilidade & estatística, etc. Λ pode conter só ruído ou muita informação. $\sharp(A)$ nos dá somente uma noção da média de número de sucessos por unidade de tempo, no longo prazo, da sequência de números positivos $\Lambda = (t_n)_{n \in \mathbb{N}} \subset \mathbb{R}_+$.

Vamos dizer que duas sequências ordenadas $\Lambda_1, \Lambda_2 \subset \mathbb{R}_+$ são equivalentes se $\rho(\Lambda_1) = \rho(\Lambda_2)$. Quando isso ocorre, denotamos a situação por $\Lambda_1 \sim \Lambda_2$. Com isso, temos classes de equivalência para sequências reais positivas ordenadas. Cada uma dessas classes simboliza o que seja o número médio de sucessos por unidade de tempo. Note que dizer que uma sequência ordenada pertence a uma dessas classes não contém outra informação que somente a do número médio de sucessos por unidade de tempo.

3.2 A sequência de variáveis aleatórias com distribuição Geométrica subjacente à distribuição Exponencial

Na Introdução, usamos uma sequência de variáveis aleatórias Geométricas. Entretanto, antes mesmo de chegarmos a ideia de variáveis aleatórias, devemos apresentar um experimento probabilístico com seu espaço amostral, espaço de eventos e função de probabilidade.

²Aqui certamente devemos ter um ordenamento subjacente, pois falamos sobre o primeiro, segundo, terceiro (etc.) sucesso.

A cada variável aleatória geométrica T_N vista na Introdução, daremos um corresponde experimento probabilístico subjacente, o espaço amostral Ω_N decorrente, um espaço de eventos \mathcal{F}_N e uma função de probabilidade $P_N : \Omega_N \rightarrow [0, 1]$.

O experimento probabilístico consiste em se executar uma sequência de ensaios de Bernoulli. Aqui já aparece uma dificuldade: O experimento pára após o primeiro sucesso, ou nunca pára? Essas duas possibilidades são analisadas em [3], com consequências interessantes. No presente caso, o experimento não pode parar, pelo simples motivo que o tempo do primeiro sucesso do qual trata a variável aleatória Exponencial não é limitado por nada. Em princípio, esse tempo pode ser qualquer valor real positivo. Então, os experimentos de Bernoulli que levam às Geométricas T_N , que devem convergir (em distribuição) à Exponencial T , nunca param.

O espaço amostral Ω_N , $N \in \mathbb{N}$, correspondente a esse experimento consiste então de seqüências do tipo

$$1010100001101010101 \dots,$$

em que 0 significa, para nós, sucesso e 1 não sucesso. Os dígitos são o resultado de uma seqüência de ensaios de Bernoulli independentes.

O espaço de eventos \mathcal{F}_N , $N \in \mathbb{N}$, é obtido por construção. Colocamos em \mathcal{F}_N o elemento Ω_N . Em seguida, colocamos no espaço de eventos todos os subconjuntos de Ω_N formados por seqüências do tipo 111110010..., com os n primeiros algarismos iguais a 1, para todo $n \in \mathbb{Z}_+$, seguido do algarismo 0 (seguido de infinitos 0's e 1's). A partir daí aplicam-se as regras $A \in \mathcal{F}_N \Rightarrow A^c \in \mathcal{F}_N$, $\{A_n\}_{n \in \mathbb{N}} \subset \mathcal{F}_N \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}_N$.

A função de probabilidade $P_N : \mathcal{F}_N \rightarrow [0, 1]$ é obtida como no artigo [3]. Transformamos cada elemento $\omega \in \Omega_N$, em um número no intervalo $[0, 1]$ representado usando-se números binários. Isso é feito prefixando-se um "0." antes de ω . Por exemplo, $\omega = 10101000011$ é transformado em 0.10101000011. Cada ensaio de Bernoulli que produz a seqüência ω , gera 0 (sucesso) com probabilidade $p_N = \lambda \frac{I}{N}$, em que λ é o "número de sucessos por intervalo de tempo" I , como discutido na seção 3.1. Isto é, ele se refere a uma classe de equivalência. Assim, a informação que é passada pelo conhecimento de λ é extremamente limitada.

Podemos calcular a probabilidade de qualquer elemento em \mathcal{F}_N , usando uma função distribuição contínua, como discutido em [3]. Note que não se trata de calcular a probabilidade individual de cada elemento $\omega \in \Omega_N$ para depois fazer uma somatória, pois evidentemente $P_N(\omega) = 0$, para todo $\omega \in \Omega_N$.

No caso dos resultados que nos interessam, temos simplesmente:

$$P_N(\{\omega \in \Omega_N | \omega = 1110n_5n_6n_7 \dots, n_5, n_6, n_7, \dots \in \{0, 1\}\}) = (1 - p_N)^4 p,$$

que é o resultado clássico para a distribuição Geométrica, mas esse é uma ilusão de simplicidade, pois estamos tratando com seqüências infinitas de ensaios de Bernoulli, o que não acontece com o caso clássico.

Chegamos finalmente à variável aleatória Geométrica $T_N : \Omega_N \rightarrow \mathbb{Z}_+$. T_N fornece o número K_N de não sucessos antes do primeiro sucesso multiplicado

pelo fator $\frac{I}{N}$, como visto na Introdução, equação (2). Dai segue:

$$P(T_N > t) = P(K_N > t \frac{N}{I}) = (1 - p_N)^{t \frac{N}{I}} = \left(1 - \lambda \frac{I}{N}\right)^{\frac{N}{I}t}.$$

Esse resultado implica na convergência $T_N \rightarrow T$ em distribuição de T_N , $N \in \mathbb{N}$ a uma variável aleatória Exponencial T , pois

$$\begin{aligned} \lim_{N \rightarrow +\infty} F_{T_N}(t) &= \lim_{N \rightarrow +\infty} (1 - P(T_N > t)) \\ &= 1 - \lim_{N \rightarrow +\infty} \left(1 - \lambda \frac{I}{N}\right)^{\frac{N}{I}t} = 1 - e^{-\lambda t} = F_T(t), \end{aligned}$$

que é a função distribuição de probabilidade acumulada de uma variável aleatória Geométrica com parâmetro λ , isto é, $E[T] = \frac{1}{\lambda}$.

4 A distribuição Exponencial como aquela de mínima informação - Conexão com a distribuição Geométrica

A variável aleatória Exponencial T foi obtida da convergência em distribuição de uma sequência de variáveis aleatórias Geométricas $(T_N)_{N \in \mathbb{N}}$. Cada T_N tem a propriedade da falta de memória, e como consequência, isso ocorre também com T , como vimos na Introdução.

O parâmetro $\lambda > 0$, número de sucessos por unidade de tempo I é comum entre todas as T_N e finalmente T . Vimos que tal parâmetro contém muito pouca informação, sendo um número que quantifica uma classe de equivalência inteira para o que significa “número de sucessos por unidade de tempo”s. Aliás, essa é a única informação que temos para a construção dos experimentos probabilísticos que levam à sequência de variáveis aleatórias Geométricas.

Como os ensaios de Bernoulli, que são ordenados, são todos independentes, o primeiro sucesso, marcado com o bit 0, pode acontecer em qualquer posição de um ponto $\omega \in \Omega_N$ do espaço amostral.

$$1111111 \underbrace{0}_{\text{sucesso}} 1111010111011101010111001 \dots$$

Isso significa que temos quase total ignorância de quando ocorre o primeiro sucesso, a não ser a que número de sucessos por unidade de tempo deve ser λ .

Se a sequência acima, gerada por uma sequência de ensaios de Bernoulli independentes, fosse captada do espaço sideral, de um suposto ser extra-terrestre, a informação nela contida seria nula, pois trata-se somente de sequência de números completamente aleatórios. A fonte que emite tal sinal emite somente ruído.

Passando ao limite, a sequência T_N converge em distribuição a T Exponencial. De (5), vemos que

$$E[T_N] = \frac{1}{\lambda} - \frac{I}{N},$$

que converge a $\frac{1}{\lambda}$ quando $N \rightarrow +\infty$. Ou seja, $\frac{1}{\lambda}$ dá uma medida do tempo médio até o primeiro sucesso, como já era de se esperar. $\frac{1}{\lambda}$ é também uma medida do tempo médio entre sucessos.

Agora mudaremos o ponto de vista.

Desejamos obter a distribuição de probabilidades para uma variável aleatória contínua que assume somente valores reais positivos, dado que sabe-se somente seu valor esperado.

Para a obtenção de tal distribuição o conceito de informação e Entropia serão fundamentais.

4.1 Entropia

O funcional Entropia, que atua sobre funções densidade de probabilidade, definido por

$$H(f) = - \int_{-\infty}^{+\infty} f(t) \log_e(f(t)) dt$$

fornece uma medida da informação contida na função densidade de probabilidade f . Quanto maior $H(f)$, maior é imprevisibilidade média sobre o valor gerado por uma realização da variável aleatória associada a ela.

Sobre a Entropia e a teoria matemática da informação, o trabalho de Shannon [5] é fundamental. A questão básica é como medir a quantidade de informação de uma sequência de sinais. Por exemplo, qual seria a informação em uma sequência 1000111010101011? O argumento de Shannon é que ao conhecermos um dado símbolo, ganhamos alguma informação, uma quantidade positiva. Quanto maior a surpresa ao se tomar conhecimento do símbolo, maior a quantidade de informação ganha no momento em que o símbolo é mostrado. Supondo que a informação seja dada por uma função f suave, e supondo que ela seja função da probabilidade do símbolo ser emitido, teríamos para essa função $h : [0, 1] \rightarrow \mathbb{R}$, a restrição $h(1) = 0$, significando que ao tomarmos conhecimento de um fato já dado como certo antes de sua emissão, a informação ganha é nula.

Além disso, o conhecimento de dois símbolos produzidos independentemente deve produzir uma quantidade de informação igual à soma das informações obtidas do conhecimento de cada símbolo individualmente: $h(pq) = h(p) + h(q)$.

Resta-nos obter $f : [0, 1] \rightarrow \mathbb{R}$, mas isso é fácil, depois que Shannon deu todas as pistas. Basta montar uma equação diferencial que f deve satisfazer.

$$h(pq) = h(p) + h(q) \Rightarrow q h'(pq) = h'(p) \Rightarrow h'(pq) + q p h''(pq) = 0.$$

Daí, $(x h'(x))' = 0$, então $x h'(x)$ é uma constante $c \in \mathbb{R}$. A conclusão é inevitável: $h'(x) = \frac{c}{x}$, e

$$h(x) = c \log_e(c_2 x), \tag{10}$$

Para alguma constante $c_2 \in \mathbb{R}$, mas como $h(1) = 0$, devemos ter $c_2 = 1$.

Tomamos $c = -1$, mas poderia ser qualquer outro valor negativo (para que h seja uma função que retorna um valor positivo), inclusive $c = -\log_2(e)$, que trocaria a base e para a base 2 na fórmula (10).

Observe que $h(x)$ dá a informação ganha no momento em que um símbolo desconhecido, que aparece com probabilidade x , mostra sua identidade.

Agora podemos falar em informação de uma fonte que produz uma sequência de símbolos. Se ela produz só os símbolos -1 e $+1$ com probabilidades p e $(1-p)$ respectivamente, associamos o valor esperado H_2 da informação a ser ganha no momento em que cada símbolo é descoberto:

$$H_2 = -(p \log_e p + (1-p) \log_e(1-p)). \quad (11)$$

Se a fonte produzir n símbolos com probabilidades p_i , $i = 1, \dots, n$, então esse valor esperado é

$$H_n = -\sum_{i=1}^n p_i \log_e p_i. \quad (12)$$

Tratamos as variáveis aleatórias como fontes de informação. Para as discretas, usamos exatamente as fórmulas acima para quantificar a informação associada a elas. Para as variáveis aleatórias contínuas, usamos integrais no lugar de somatórias. Mais precisamente, no caso de uma variável aleatória contínua X , cuja distribuição de probabilidade é calculada usando-se a função densidade de probabilidade f_X , temos para a quantidade de informação de Shannon:

$$H(f_X) = -\int_{-\infty}^{+\infty} f_X(x) \log_e(f_X(x)) dx. \quad (13)$$

Essa informação média é a respeito da realização da variável aleatória em si. De certa forma, ela mede a surpresa média ao se saber do valor da realização da variável aleatória. Esse ponto é importante: A informação não está na função densidade de probabilidade f_X , mas sim no momento da realização da variável aleatória X , e essa informação se refere à própria realização de X .

O funcional H , que mede a informação ou surpresa média de uma fonte, no sentido de Shannon, recebe o nome de *Entropia*. Reproduzo aqui uma história interessante, que envolve o grande Matemático John von Neumann, contada em [6], sobre o batismo dessa designação:

What's in a name? In the case of Shannon's measure the naming was not accidental. In 1961 one of us (Tribus) asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under

that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.’ ”

O estudante deve ter em mente que quanto maior $H(f_X)$, também é maior a informação a ser obtida em média, quando se conhece uma realização de X , a respeito dessa própria realização, ou seja, menos se sabia a respeito de X *a priori*. *Grosso modo*, $H(f_X)$ dá uma medida da desinformação contida em f_X , mas de informação no momento da realização de um valor particular de X .

Considere duas variáveis aleatórias, uma uniforme U distribuída no intervalo $[-1, 1]$ e outra Y com distribuição $f_Y(y) = \frac{3}{2}y^2$, se $y \in [-1, 1]$ e $f_Y(y) = 0$, caso contrário. Devemos esperar que $H(f_U) > H(f_Y)$, pois em média, a “surpresa” ao obter uma realização de U será maior. De fato, $H(f_U) = 0.693147$ e $H(f_Y) = 0.261202$.

Um princípio usado para encontrar-se distribuições de probabilidade que satisfazem restrições dadas é buscar aquela que tem a maior entropia, dentro de uma família onde se faz a busca. A encontrada por esse critério é designada *a de Máxima Entropia*. Suponha, por exemplo, que saibamos que uma variável aleatória W é contínua e que sua imagem é o intervalo $[-1, 1]$, isto é, W assume valores entre -1 e $+1$. Sabendo somente este fato, qual seria a distribuição de probabilidades que melhor representaria esse estado de informações? Observe que temos infinitas possibilidades para essa resposta. Por exemplo, a distribuição f_Y acima é uma candidata.

Respondemos a pergunta posta buscando a distribuição suportada no intervalo $[-1, 1]$ que tem a maior Entropia, pois ela representa a situação em que se geraria a maior surpresa média ao obter-se um valor da variável aleatória correspondente. Em outros termos, ela corresponderá àquela com *menor* informação que já temos a priori sobre ela.

4.1.1 Exemplo: obtenção da distribuição Uniforme

Concretamente, sabendo-se somente que a variável aleatória W assume valores entre $[-1, 1]$, vamos buscar a função densidade de probabilidade f que maximiza

$$H(f) = - \int_{-1}^{+1} f(x) \log(f(x)) dx,$$

com a restrição $\int_{-1}^{+1} f(x) dx = 1$, pois f deve ser uma função densidade de probabilidade.

Esse problema de maximização é familiar aos estudantes de Mecânica Analítica. Basta aplicar o método dos multiplicadores de Lagrange aliado à equação de Euler-Lagrange (ver, por exemplo, [2]). O funcional a ser maximizado é

$$f \mapsto - \int_{-1}^{+1} f(x) \log(f(x)) dx + k \left(\int_{-1}^{+1} f(x) dx - 1 \right).$$

A equação de Euler-Lagrange é

$$\frac{\partial L}{\partial f} - \frac{d}{dx} \left(\frac{\partial L}{\partial f'} \right) = 0, \quad (14)$$

com

$$L(f, f', x) = -f \log_e f + k f.$$

A equação (14) fornece $-\log_e f - 1 + k = 0$, isto é, que

$$f(x) = e^{k-1},$$

que é uma constante. Da condição $\int_{-1}^{+1} f(x) dx = 1$, obtemos $e^{k-1} = \frac{1}{2}$.

Então a função densidade de probabilidade com maior entropia que tem suporte em $[-1, 1]$ é a uniforme. Veja que esse era um resultado já esperado intuitivamente.

4.1.2 Exemplo: obtenção da distribuição suportada em $[-1, +1]$ de uma variável aleatória X , quando se sabe seu valor esperado $E[X]$

Suponha agora que saibamos *a priori* que X seja uma variável aleatória contínua que tenha imagem em $[-1, +1]$ e que $E[X] = m \in]-1, +1[$.

Aplicamos o mesmo método descrito acima, o da Máxima Entropia, para obter f_X , contínua em $[-1, +1]$, quando se conhece apenas o seu suporte $[-1, +1]$ e $E[X] = m$. O funcional a ser maximizado é

$$\begin{aligned} f_X \mapsto & - \int_{-1}^{+1} f_X(x) \log(f_X(x)) dx + k_1 \left(\int_{-1}^{+1} f_X(x) dx - 1 \right) \\ & + k_2 \left(\int_{-1}^{+1} x f_X(x) dx - m \right). \end{aligned}$$

A equação de Euler-Lagrange é a mesma (14), em que agora

$$L(f_X, f'_X, x) = -f_X \log_e f_X + k_1 f_X + k_2 x f_X.$$

Substituindo L na equação de Euler-Lagrange, chegamos à

$$-\log_e f_X - 1 + k_1 + k_2 x = 0,$$

que nos fornece

$$f_X(x) = e^{k_1-1} e^{k_2 x}.$$

As constantes k_1 e k_2 são obtidas das restrições

$$\int_{-1}^{+1} f_X(x) dx = 1 \quad \text{e} \quad E[x] = \int_{-1}^{+1} x f_X(x) dx = m.$$

Na figura 1 estão mostradas as soluções f_X para diversos valores de $m \in]-1, +1[$. Note que valor $m = 0$ leva à distribuição Uniforme.

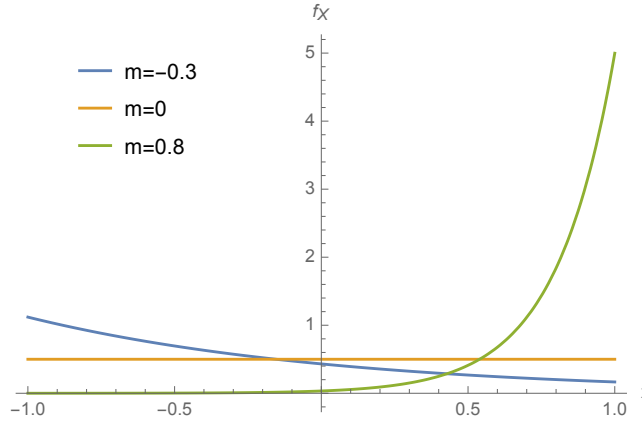


Figura 1: Solução do Método da Máxima Entropia. É conhecido o valor esperado e o suporte.

4.1.3 A distribuição Exponencial via maximização da entropia

Agora desejamos obter a distribuição de probabilidades para uma variável aleatória contínua \tilde{T} que assume somente valores reais positivos, não em um intervalo limitado, mas no intervalo $[0, +\infty[$, dado que sabe-se somente seu valor esperado.

Vamos obter a tal distribuição de \tilde{T} , sabendo-se *somente* que essa variável aleatória assume valores em \mathbb{R}_+ e tem valor esperado $E[\tilde{T}] = \frac{1}{\lambda}$.

Maximizaremos a entropia $H(f_{\tilde{T}})$ para a função densidade de probabilidade associada à variável aleatória \tilde{T} . As restrições são

$$\int_0^{+\infty} f_{\tilde{T}}(t) dt = 1 \quad \text{e} \quad E[\tilde{T}] = \int_0^{+\infty} t f_{\tilde{T}}(t) dt = \frac{1}{\lambda}. \quad (15)$$

O funcional a ser maximizado é

$$Q(f_{\tilde{T}}) = - \int_0^{+\infty} f_{\tilde{T}}(t) \log_e(f_{\tilde{T}}(t)) dt + k_1 \left(\int_0^{+\infty} f_{\tilde{T}}(t) dt - 1 \right) + k_2 \left(\int_0^{+\infty} t f_{\tilde{T}}(t) dt - \frac{1}{\lambda} \right).$$

Para obter f que maximiza $Q(f_{\tilde{T}})$, devemos resolver a equação de Euler-Lagrange

$$\frac{\partial L}{\partial f_{\tilde{T}}} - \frac{d}{dt} \left(\frac{\partial L}{\partial f'_{\tilde{T}}} \right) = 0, \quad (16)$$

em que

$$L(f_{\tilde{T}}, f'_{\tilde{T}}, t) = -f_{\tilde{T}} \log_e f_{\tilde{T}} + k_1 f_{\tilde{T}} + k_2 t f_{\tilde{T}}.$$

A equação (16) fornece

$$-\log_e f_{\bar{T}} - \frac{f_{\bar{T}}}{f_{\bar{T}}} + k_1 + k_2 t = 0,$$

ou seja,

$$f_{\bar{T}}(t) = e^{k_1 - 1} e^{k_2 t}, \quad (17)$$

em que $k_2 < 0$, pois por (15), a possibilidade $k_2 \geq 0$ é impossível.

Aplicando-se as restrições (15), obtemos

$$k_2 = -e^{k_1 - 1} \quad \text{e} \quad k_2 = -\lambda.$$

Isto é, de (17), temos

$$f_{\bar{T}}(t) = \lambda e^{-\lambda t},$$

que é justamente a função densidade de probabilidade de uma variável aleatória Exponencial!

Isso fecha o nosso ciclo, quando começamos a questionar o que significa “número médio de sucessos por unidade de tempo”, para analisar o conceito de Valor Esperado, ver que tipo de informação está subentendido na sequência de variáveis aleatórias geométricas, que converge em distribuição a uma Exponencial.

5 Exemplo numérico

Apenas para ilustrar, trazemos aqui um exemplo numérico, que pode ser repetido pelos interessados. Ele simula uma situação em que geramos sequências de ensaios de Bernoulli com parâmetro p . A probabilidade de cada ensaio gerar 1 é p e de gerar 0 é $(1 - p)$. Temos portanto um experimento probabilístico, como aquele de nos dá a Geométrica, com sucessivos ensaios de Bernoulli que nunca cessam.

Usando a notação da Introdução, para a simulação desta seção tomamos como parâmetros $I = 1$, $N = 30$ e $p = 0.3$. Mostramos abaixo três sequências típicas.

1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1
 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1
 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0

Para cada sequência, observamos a posição do primeiro sucesso (dígito 0) e gravamos o resultado. Após gerar $M = 500$ sequências contamos o número de vezes em que o primeiro sucesso aconteceu em cada posição (há $N = 30$ posições). Cada posição corresponde a um instante de tempo. A posição i corresponde ao tempo $T_i = \frac{i}{N}$, conforme a equação (2). O gráfico do número de sucessos em cada instante t_i deve ser normalizado para que ele represente pontos sobre o gráfico de uma função densidade de probabilidade.

Com os dados coletados, e feita a normalização, podemos colocar tudo em um gráfico, que está mostrado na Figura 2. O parâmetro da distribuição Exponencial é $\lambda = p \frac{N}{T}$.

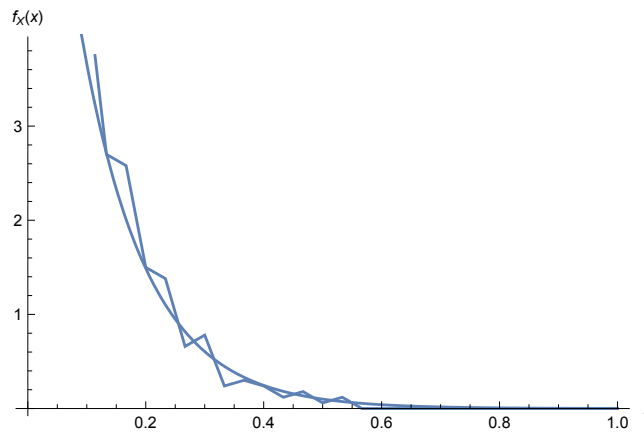


Figura 2: Comparação entre o resultado da simulação e a função densidade de probabilidade da Exponencial.

6 Conclusões

Neste texto, revisitamos principalmente o conceito de Valor Esperado, para dele tirarmos uma ligação entre a distribuição Geométrica e a Exponencial. Durante o percurso, fizemos uma grande revisão de pontos importantes da Teoria de Probabilidades.

Agora, apenas gostaríamos de deixar registrados alguns pontos que consideramos importantes para o estudante que eventualmente ler este texto.

- Os conceitos de Média e de Valor Esperado, apesar de ligados, são qualitativamente diferentes. É preciso cuidado ao se fazer a ligação entre elas.
- Nem toda variável aleatória tem valor esperado.
- A distribuição Exponencial pode ser obtida de uma sequência de Geométricas que convergem em distribuição.
- Essencialmente, a Exponencial aparece quando há uma situação de pouca informação sobre a variável aleatória. Mais precisamente, ela aparece quando conhecemos somente o seu valor esperado.

Referências

- [1] Carlos Alberto Barbosa Dantas. *Probabilidade – Um Curso Introdutório*. EDUSP, 3rd. edition, 2013.
- [2] L.N. F. França and A. Z. Matsura. *Mecânica Geral*. Edgar Blucher, 3ed. edition, 2011.

- [3] Alexandre Kawano. *Enumerabilidade e uma Ligação entre as Distribuições Uniforme e Geométrica*. (Artigo presente nesta coletânea)
- [4] Alexandre Kawano. *A variância, a distribuição normal e a marcha reversa no tempo*. (Artigo presente nesta coletânea)
- [5] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(4):623–656, 1948.
- [6] Myron Tribus and Edward C. McIrvine. Energy and Information. *Scientific American*, 225(3):179–190, 1971.

Capítulo 3

A variância, a distribuição Normal e a marcha reversa no tempo*

Alexandre Kawano
Universidade de São Paulo

Maio de 2020

Resumo

Neste artigo, que tem por público alvo alunos de graduação que já tiveram alguma exposição a um curso de Probabilidades, analisamos o sentido que podemos dar ao conceito de Variância de um ponto de vista da conservação de energia, e a partir dele, chegamos de forma natural à distribuição Normal, que é ligada ao processo de difusão. Vemos no texto também como a Normal justifica o uso do Método dos Mínimos Quadrados. Através da construção de uma sequência de variáveis aleatórias que converge em distribuição à Normal chegamos à uma variável aleatória discreta que usamos para analisar a reversibilidade no tempo de processos de difusão. Aproveitamos a oportunidade para explorar as bases da Teoria de Probabilidades.

1 Introdução

A distribuição Normal, ou de Gauss ocupa, ou deveria ocupar, grande parte do tempo na programação de uma disciplina de Probabilidades. Isso acontece porque ela aparece em muitas áreas do conhecimento, como na teoria dos erros, na termodinâmica estatística, na equação da difusão, na transformação de Weierstrass, entre outras.

Neste texto, que tem por público alvo alunos de graduação, iniciamos recordando a distribuição Normal, para depois ver como ela é usada para se justificar o Método dos Mínimos Quadrados. Aliás esse foi o ponto de partida histórico, quando Gauss o desenvolveu para resolver um problema astronômico de se determinar a órbita do cometa Ceres em 1802 [6].

A distribuição Normal destoa um pouco das demais apresentadas aos alunos nos cursos básicos de probabilidades. De fato, nenhuma outra variável aleatória contínua tem função densidade de probabilidade com um termo ao quadrado

*Alexandre Kawano, *Coletânea de artigos sobre probabilidade básica*, Epusp, ISBN 978-65-89190-15-8, DOI 10.11606/9786589190158

em seus parâmetros. O aparecimento do termo quadrático não é natural. Ele está relacionado ao conceito de Variância, que tipicamente é apresentado ao estudante bem no início do curso, como um ente que mede a dispersão dos valores que um fenômeno assume. Entretanto, essa interpretação só é possível para distribuições com formas bem definidas. Assim, neste texto apresentamos uma visão alternativa para a introdução da variância, usando o conceito de energia cinética. Tendo obtida a variância, chegamos à distribuição Normal como aquela que contém a menor informação, conhecidas apenas o seu valor esperado e a sua variância.

No texto vemos com o a Normal está ligada à equação do calor, que também rege fenômenos difusivos. Construímos uma sequência de variáveis aleatórias discretas que converge em distribuição à Normal. Usando uma variável aleatória discreta da sequência, analisamos a questão da reversibilidade termodinâmica de processos de difusão. As ideias são ao final aplicadas ao tratamento de imagens.

Como nos textos anteriores [4] e [5], aproveitamos para fazer um aprofundamento da teoria das Probabilidades, que como já afirmamos, é cheia de nuances.

2 A distribuição Normal

A variável aleatória X com distribuição Normal com valor esperado $E[X] = \mu$ e variância $V[X] = \sigma^2$ é contínua, e tem associada a ela a seguinte função densidade de probabilidade

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}. \quad (1)$$

Denotamos $X \sim N(\mu, \sigma^2)$.

Sabemos do Teorema do Limite Central, que dada uma sequência de variáveis aleatórias independentes e identicamente distribuídas $(X_n)_{n \in \mathbb{N}}$, todas com $E[X_n] = \mu$ e $V[X_n] = \sigma$, vale

$$Y_j \doteq \frac{\sum_{n=1}^j X_n - n\mu}{\sqrt{j\sigma^2}} \xrightarrow{D} Y.$$

Isto é Y_j converge *em distribuição* para $Y \sim N(0, 1)$.

A distribuição Normal é tão importante que a sua função distribuição acumulada F_X encontra-se tabelada. Uma tabela da Normal dita padronizada, com valor esperado μ e variância $\sigma^2 = 1$ pode ser encontrada, por exemplo, em [2].

Para usar-se a tabela com $X \sim N(\mu, \sigma^2)$ transformamos X em $Y \sim N(0, 1)$ através da transformação

$$Y = \frac{X - \mu}{\sigma}.$$

A função densidade de probabilidade (1) também aparece ligada à solução da equação do calor

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) - \frac{\sigma^2}{2} \frac{\partial^2 u}{\partial x^2}(t, x) = 0, & \forall (t, x) \in]0, +\infty) \times \mathbb{R}, \\ u(0, x) = \delta_\mu, \end{cases} \quad (2)$$

em que δ_0 é a distribuição Delta de Dirac [1].

A solução de (2) é a função

$$u(t, x) = \frac{1}{\sqrt{2\pi\sigma\sqrt{t}}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2 t}}, \quad (3)$$

que é a mesma função densidade de probabilidade (1), mas com média $\mu = 0$ e variância $\sigma^2 t$, que agora depende do tempo. A equação do calor é a mesma que rege fenômenos de difusão simples, e por isso ela é também conhecida como a Equação da Difusão.

3 Método dos Mínimos Quadrados (MMQ)

O Método dos Mínimos Quadrados está intimamente ligada à distribuição de Gauss, e de fato aí está a sua origem histórica, quando Gauss estudou o problema de se estimar a órbita do cometa Ceres em 1801 [6]. Aqui não vamos seguir os passos de Gauss, mas ver como as ideias da distribuição que leva seu nome e o MMQ estão ligadas.

Suponha que estamos diante de dados experimentais de um fenômeno cujo modelo prediz para uma entrada $x \in \mathbb{R}$, o valor $y = g_{a_1, \dots, a_r}(x)$. Por exemplo, x poderia ser pressão e y a temperatura. A função g depende dos parâmetros a_1, \dots, a_r . Os dados experimentais consistem de um conjunto de pares ordenados $D = \{(x_i, y_u) \in \mathbb{R}^2 \mid i = 1, 2, \dots, n\}$, que estão contaminados por erros. Desejamos encontrar os parâmetros $a_1, \dots, a_r \in \mathbb{R}$ tais que g_{a_1, \dots, a_r} aproxima da melhor forma possível os pontos em D .

A abordagem usada nos laboratórios estudantis é encontrar os parâmetros a_1, \dots, a_r pelo Método dos Mínimos Quadrados, que consiste em encontrá-los tais que a soma

$$S(a_1, \dots, a_r) = \sum_{i=1}^n (y_i - g_{a_1, \dots, a_r}(x_i))^2$$

seja mínima.

Ecoando uma dúvida geral entre os bons alunos, perguntamos: Por que minimização de *quadrados*? Não poderia ser a minimização da diferença $(y_i - (ax_i + b))^m$ qualquer número m par? A resposta é que para $m = 2$ há uma boa justificativa dada pela distribuição Normal.

Supomos que o erro $E_i = y_i - g_{a_1, \dots, a_r}(x_i)$ tenha distribuição Normal com média 0 e variância ϵ^2 . Isso quer dizer que y_i é a realização experimental de uma variável aleatória Y_i normalmente distribuída com média $g_{a_1, \dots, a_r}(x_i)$ e variância ϵ^2 .

Se os erros entre as medidas forem independentes, então, a função densidade de probabilidade conjunta de (Y_1, \dots, Y_n) é dada pelo produto de n funções do tipo (1):

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) &= f_{Y_1}(y_1) \dots f_{Y_n}(y_n) \\ &= \frac{1}{(\sqrt{2\pi\epsilon})^n} e^{-\frac{1}{2\epsilon^2} \sum_{i=1}^n (y_i - g_{a_1, \dots, a_r}(x_i))^2}. \end{aligned} \quad (4)$$

A escolha pela minimização de $S(a_1, \dots, a_r) = \sum_{i=1}^n (y_i - g_{a_1, \dots, a_r}(x_i))^2$ leva à maximização da verossimilhança, ou “probabilidade” que de fato tenhamos observado os dados experimentais D , medida pela mesma equação (4). Note as aspas em “probabilidade”. Não se trata de probabilidade, mas uma interpretação para a plausibilidade de se ter obtido os dados realmente observados.

A conclusão é que as suposições de que os erros experimentais seguem uma distribuição Normal, e que cada medida é independente uma da outra levam naturalmente ao critério de minimização da soma dos quadrados dos erros, e por conseguinte ao MMQ. Caso a distribuição para os erros fosse outra, teríamos um outro método para se encontrar g_{a_1, \dots, a_r} .

4 Variância

Se atentarmos para o Método dos Mínimos Quadrados, notamos a grande importância do quadrado na fórmula da função de densidade de probabilidade f_X dada por (1), porque é por causa dele é que nasce o MMQ.

Hoje sabemos que esse quadrado está relacionado ao conceito de variância, mas é interessante notar que o termo Variância é relativamente recente, datando do início do século XX, cunhado por Fisher [3].

4.1 A variância como consequência mecânica

Na época de Gauss, já havia a primeira forma do Teorema do Limite Central (de Moivre, 1718), tratando da convergência da soma de valores aleatórios independentes a uma função com forma exponencial. Gauss desenvolveu a distribuição de probabilidades que porta seu nome (1801), que essencialmente é uma exponencial com dois parâmetros, visando a solução do problema de ajuste de curvas. A descoberta da equação de calor de Fourier, que é mostrada na sua forma moderna em (2) ocorreu somente depois, por volta de 1822, data da publicação de seu trabalho *Théorie Analytique de la Chaleur*.

Podemos dar uma possível justificativa para o uso da soma de quadrados através da Mecânica lagrangeana (1788), reformulação da newtoniana em termos de energia, que já estava disponível a Gauss.

Para apelar à intuição física, e aos conhecimentos dos cursos de Mecânica Geral que o estudante que lê este trabalho certamente já tem, considere um sistema de N partículas de massas p_i , tais que $\sum_{i=1}^N p_i = 1$, se movendo com velocidades u_i sobre a reta real. Se não há forças externas atuando sobre esse sistema, a sua quantidade de movimento é constante:

$$u_B = \sum_{i=1}^N p_i u_i = cte. \quad (5)$$

Coincidentemente, neste nosso arranjo, a velocidade do baricentro do sistema é numericamente igual à quantidade de movimento do sistema. Por isso, adotamos o símbolo u_B , que é a velocidade média das partículas.

Para esse sistema isolado suponha que a sua energia seja somente a soma das energias cinéticas das suas partículas componentes, isto é, não há energia potencial envolvida. Então a energia

$$E_0 = \sum_{i=1}^N \frac{1}{2} p_i u_i^2 \quad (6)$$

é constante no tempo.

Agora, lembre-se que para um sistema de partículas isolado se movendo em uma reta, as únicas quantidades conservadas são a quantidade de movimento e a energia. Em outras palavras, u_B e E_0 descrevem em termos gerais o sistema de partículas. Se quisermos criar uma função que descrevesse o sistema em termos gerais, que usasse o menor número de parâmetros, seria natural usar u_B e E_0 . Porém, essas grandezas do modo como foram apresentadas por (5) e (6) têm um pequeno problema. Elas não são completamente independentes, pois o conhecimento de E_0 , digamos $E_0 = 0$, leva ao conhecimento de u_B ($E_0 = 0 \Rightarrow u_B = 0$). Mas esse problema é contornado sabendo-se que a energia depende de um referencial.

Podemos obter a energia cinética do sistema de partículas medindo-se as velocidades com relação a um referencial inercial que se move com a mesma velocidade u_B do baricentro do sistema, que no nosso caso, é numericamente igual a Q . Neste novo referencial, a energia passa a ser

$$E = \sum_{i=1}^N \frac{1}{2} p_i (u_i - u_B)^2. \quad (7)$$

Nesse caso, E e u_B são completamente independentes do ponto de vista da passagem de informação. Fisicamente, podemos pensar em uma porção de gás dentro de um laboratório. A energia cinética de suas partículas medida no em um referencial fixo no laboratório essencialmente é a sua temperatura. O conhecimento dessa temperatura não dá nenhuma informação sobre a velocidade com que o laboratório se move, e vice-versa.

Note que como $\sum_{i=1}^N p_i (u_i - u_B)^2 = \sum_{i=1}^N p_i u_i^2 - \left(\sum_{i=1}^N p_i u_i \right)^2$, que nada mais é que a conhecida fórmula $V[U] = E[U^2] - (E[U])^2$ (se as velocidades das partículas forem tratadas como realizações de uma variável aleatória U), temos a equivalência

$$\{u_B, E_0\} \iff \{u_B, E\}.$$

É importante observar que a energia cinética E é uma soma dos quadrados das diferenças das velocidades das partículas em relação ao seu valor esperado. Nesse sentido, E ocupa o sentido que damos à Variância na teoria de probabilidades.

Voltando à questão da especificação de uma distribuição para as velocidades u_i , conhecendo-se apenas a sua velocidade média u_B , e a sua energia E , que são ambas constantes por causa das leis de conservação da quantidade de movimento linear e da energia, respectivamente, podemos indagar qual seria aquela que

fornece contém a menor quantidade de informação intrínseca, no sentido da Informação de Shannon. Uma introdução às ideias principais relacionadas a ela pode ser encontrada em [5], que foi escrita pensando nos estudantes interessados em probabilidades.

Aqui, imaginamos que a quantidade de partículas N é muito grande, como em um mol de gás, por exemplo, e daí podemos obter uma boa aproximação quando trocamos as somatórias acima por integrais e a distribuição de probabilidades $\{p_i\}_{i=1}^N$ por uma função densidade de probabilidades f_U .

A Entropia de Shannon de uma função densidade de probabilidade f_U é dada por

$$H(f_U) = - \int_{-\infty}^{+\infty} f_U(u) \log_e(f_U(u)) du.$$

Vamos procurar f_U que maximiza a Entropia $H(f_U)$ sujeita às restrições

$$\begin{aligned} \int_{-\infty}^{+\infty} u f_U(u) du &= 1, \quad \int_{-\infty}^{+\infty} u f_U(u) du = u_B, \\ \int_{-\infty}^{+\infty} (u - u_B)^2 f_U(u) du &= E. \end{aligned} \tag{8}$$

O método natural para fazermos essa maximização é usar o método dos multiplicadores de Lagrange com a equação de Euler-Lagrange, como é feito na Mecânica. Consideramos o funcional

$$\begin{aligned} \tilde{H}(f_U, k_1, k_2) &= - \int_{-\infty}^{+\infty} f_U(u) \log_e(f_U(u)) du + k_1 \left(\int_{-\infty}^{+\infty} f_U(u) du - 1 \right) \\ &+ k_2 \left(\int_{-\infty}^{+\infty} u f_U(u) du - u_B \right) + k_3 \left(\int_{-\infty}^{+\infty} f_U(u)(u - u_B)^2 du - E \right). \end{aligned}$$

Buscamos os extremos do funcional acima resolvendo-se a equação de Euler-Lagrange.

$$\frac{\partial L}{\partial f_U} - \frac{d}{du} \left(\frac{\partial L}{\partial f'_U} \right) = 0, \tag{9}$$

em que

$$L(f_U, f'_U, u) = -f_U \log_e f_U + k_1 f_U + k_2 u f_U + k_3 (u - u_B)^2 f_U$$

A equação (9) fornece

$$-\log_e f_U - 1 + k_1 + k_2 u + k_3 (u - u_B)^2 = 0,$$

isto é, que

$$f_U(u) = \underbrace{e^{1-k_1-k_2u}}_{\text{Se } k_2 = 0, \text{ então isso é só uma constante}} \times e^{-k_3(u-u_B)^2}. \tag{10}$$

Agora basta determinarmos as constantes k_1, k_2, k_3 para chegarmos à forma final de f_U . Para nós, basta ver que (10) já tem a forma da função densidade de probabilidade gaussiana, a menos da presença do incomodo $-k_2u$ no expoente. Mas $k_2 = 0$, pois caso contrário, $\int_{-\infty}^{+\infty} f_U(u) du = +\infty$.

Aplicando-se as restrições (8), chegamos à fórmula (1).

A conclusão que nos interessa é que f_U obtida pela maximização da Entropia de Shannon, conhecendo-se a velocidade média v_B do sistema de partículas, que é numericamente igual a sua quantidade de movimento linear, e a energia E , chegamos à função densidade de probabilidade Normal.

Sob essa luz, a distribuição Normal é aquela que melhor representa o estado de conhecimento sobre uma distribuição de velocidades de um conjunto de partículas quando se conhece apenas sua velocidade média e a sua energia, ambos conceitos objetivos bem ancorados na Física. Fazendo uma conexão da energia com a variância, podemos transpor esse resultado para a situação geral: A distribuição de Gauss é a melhor que representa o estado de conhecimento sobre uma varável aleatória, quando se conhece apenas suas média e variância, essa como abstração do conceito de energia.

4.2 A variância como medida de dispersão

Até este ponto, não mencionamos a variância como medida de dispersão, pois esse conceito não nos parece natural. Ao considerarmos duas distribuições como as da figura 1, não se pode afirmar qual das duas tem mais “dispersão”.

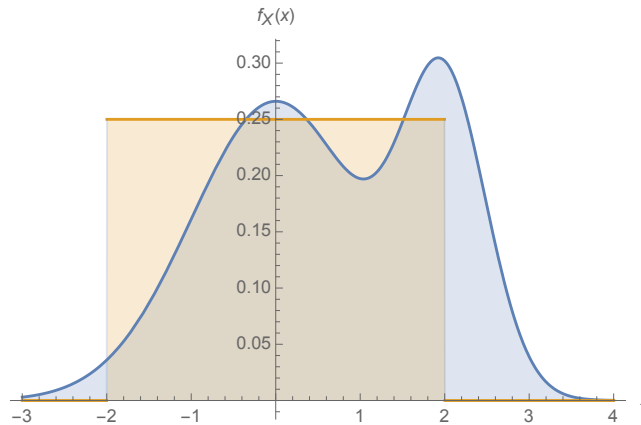


Figura 1: Comparação entre dispersões

Na figura 1, temos uma distribuição uniforme com variância igual a 2.08333 e outra que é obtida pela soma de duas Normais, cuja variância é 1.63889.

A caracterização de dispersão por meio da variância se torna natural somente após o contato com distribuições com no máximo um extremo local, como a Normal.

A partir da interpretação da variância como medida da dispersão em distribuições como a Normal, a palavra variância passou ao nosso vocabulário como tendo também sentidos de dispersão e incerteza, como por exemplo na frase “Princípio da Incerteza de Heisenberg”, que trata de variâncias.

5 A obtenção da distribuição normal a partir de uma sequência de ensaios de Bernoulli via equação do calor/difusão

Vamos construir uma sequência de variáveis aleatórias como foi feito em [4] e [5], mas desta vez vai ser um pouco mais complexo.

Consideramos um experimento que consiste em realizar-se N ensaios de Bernoulli com probabilidade $p = 1/2$, dando os resultados $+1$ e -1 com igual probabilidade, e registrar o resultado. O espaço amostral Ω_N é então o conjunto formado pelas sequências *finitas* do tipo $+1 - 1 - 1 + 1 + 1 + 1 + 1 \dots$. O espaço de eventos \mathcal{F}_N é o conjunto das partes de Ω_N , sem grandes complicações. A função probabilidade $P_N : \mathcal{F}_N \rightarrow [0, 1]$ retorna a probabilidade dos eventos de \mathcal{F}_N . O exemplo típico é P_N retornar a probabilidade de que em um experimento ocorra n vezes o número $(+1)$ e $(N - n)$ o número (-1) . Naturalmente, essa probabilidade é dada pela distribuição Binomial: $P_N(\{\text{ocorrer } n \text{ vezes } +1 \text{ e } (N - n) \text{ vezes } -1\}) = \binom{N}{n}(p)^n(1 - p)^{N-n}$.

Seja a variável aleatória $X_N : \Omega_N \rightarrow \mathbb{Z}$. $X_N(\omega)$ fornece a soma dos números que compõe a sequência ω . Por exemplo, se $\omega = +1 - 1 - 1 + 1 + 1 + 1 + 1$, então $X_7 = 3$. A variável aleatória X_N é resultante de um processo conhecido como *passeio aleatório*. Vamos relacionar as variáveis aleatórias X_N com um contexto físico, para então chegarmos à equação do calor/difusão.

Tome a reta real, mas considere apenas os números inteiros nela. Suponha que seja colocada uma partícula na posição $x_0 = 0$ no instante $t = 0$. Fixe $N \in \mathbb{N}$ e o tempo $T_N > 0$ durante o qual o todo o experimento ocorrerá. T_N deve ser escolhido convenientemente, como veremos adiante. A cada intervalo de tempo $\Delta t = T_N/N$, isto é, nos instantes de tempo $t_i = i\Delta t$, realizamos um ensaio de Bernoulli, como descrito acima. Se o resultado for $(+1)$, então a partícula avança à direita percorrendo uma distância Δx , e se o resultado for -1 , ela retrocede para a esquerda $-\Delta x$, conforme a figura 2. Então, como em $t_0 = 0$, a posição inicial era $x = 0$, vemos que a posição da partícula no instante $t = t_N$ é dada por $x = (\Delta x)X_N$. A distribuição de probabilidades da variável aleatória $(\Delta x)X_N$ não é difícil de ser obtida por métodos que são vistos normalmente em sala de aula, mas aqui adotaremos uma outra abordagem mais física, para que possamos fazer uma conexão com a equação do calor/difusão.

Vamos denotar por $u_{i,n}\Delta x$ a probabilidade da partícula estar na posição $x_n = n\Delta x$ no instante de tempo $t = i\Delta t$, $i \geq 0$, $n \in \mathbb{Z}$.

Como a partícula nunca permanece em seu lugar, a probabilidade que ela se encontre em $u_{i+1,n}$ deve contemplar somente duas possibilidades:

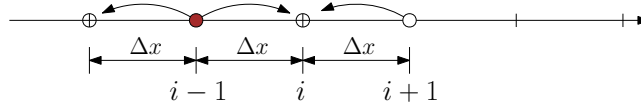


Figura 2: Movimento da partícula.

- i. Ela estava em na posição $x_{n-1} = (n-1)\Delta x$ no instante $t_i = (i)\Delta t$ e passou para o ponto $x_n = n\Delta x$;
- ii. Ela estava em na posição $x_{n+1} = (n+1)\Delta x$ no instante $t_i = (i)\Delta t$ e passou para o ponto $x_n = n\Delta x$.

Então, podemos afirmar que $u_{i+1,n}\Delta x = \frac{1}{2}u_{i,n-1}\Delta x + \frac{1}{2}u_{i,n+1}\Delta x$. Cancelando o fator Δx , temos

$$u_{i+1,n} = \frac{1}{2}u_{i,n-1} + \frac{1}{2}u_{i,n+1}. \quad (11)$$

Escrevemos a mesma fórmula para $u_{i,n-1}$ e $u_{i,n+1}$:

$$\begin{cases} u_{i,n-1} = \frac{1}{2}u_{i-1,n-2} + \frac{1}{2}u_{i-1,n}, \\ u_{i,n+1} = \frac{1}{2}u_{i-1,n} + \frac{1}{2}u_{i-1,n+2}. \end{cases} \quad (12)$$

Fazendo a substituição de (12) em (11), obteremos

$$u_{i+1,n} = \frac{1}{4}(u_{i-1,n-2} + 2u_{i-1,n} + u_{i-1,n+2}). \quad (13)$$

Subtraímos $u_{i-1,n}$ de ambos os termos de (13), para obter

$$u_{i+1,n} - u_{i-1,n} = \frac{1}{4}(u_{i-1,n-2} - 2u_{i-1,n} + u_{i-1,n+2}). \quad (14)$$

Agora dividimos e multiplicamos por $2\Delta t$ e Δx para obter aproximações de derivadas:

$$\frac{u_{i+1,n} - u_{i-1,n}}{2\Delta t} = \frac{1}{4} \frac{u_{i-1,n-2} - 2u_{i-1,n} + u_{i-1,n+2}}{(\Delta x)^2} \frac{(\Delta x)^2}{2\Delta t}. \quad (15)$$

Mas $\Delta t = \frac{T_N}{N}$. Então

$$\frac{u_{i+1,n} - u_{i-1,n}}{2\Delta t} = \left[\frac{1}{2} \frac{N(\Delta x)^2}{4T_N} \right] \frac{u_{i-1,n-2} - 2u_{i-1,n} + u_{i-1,n+2}}{(\Delta x)^2}. \quad (16)$$

Escolhemos $T_N = \frac{N(\Delta x)^2}{4\sigma^2}$. Daí,

$$\frac{u_{i+1,n} - u_{i-1,n}}{2\Delta t} = \frac{\sigma^2}{2} \frac{u_{i-1,n-2} - 2u_{i-1,n} + u_{i-1,n+2}}{(\Delta x)^2}, \quad (17)$$

lembrando que $\Delta t = \frac{t}{N}$.

Escolha $T_N = t$. Com isso,

$$\Delta t = \frac{t}{N} \quad \text{e} \quad \Delta x = \frac{2\sigma\sqrt{t}}{\sqrt{N}}.$$

Agora passe ao limite (16) fazendo $N \rightarrow +\infty$. Com isso, obteremos a equação diferencial parcial linear

$$\frac{\partial u}{\partial t}(t, x) - \frac{\sigma^2}{2} \frac{\partial^2 u}{\partial x^2}(t, x) = 0. \quad (18)$$

Em resumo, a nossa variável aleatória $U_N = (\Delta x)X_N = \frac{2\sigma\sqrt{t}}{\sqrt{N}}X_N$, obtida pelo trânsito da partícula em sobre pontos equidistantes, com $\Delta x = \frac{2\sigma\sqrt{t}}{\sqrt{N}}$, pulando de posição a cada $\Delta t = \frac{t}{N}$, converge a uma variável aleatória contínua U , que dá a posição da partícula sobre a reta real no instante de tempo $t > 0$.

Para cada $t > 0$ fixo, U tem função densidade de probabilidade $x \mapsto u(t, x)$. Por outro lado, $(t, x) \mapsto u(t, x)$ é solução da equação diferencial (18), quando a condição inicial é $u(t=0) = \delta_{x_0}$, um Delta de Dirac aplicado no ponto x_0 no instante $t = t_0$. A equação (18) com essa condição inicial é exatamente o mesmo problema (2), cuja solução é (3), que é uma função densidade de probabilidade de Gauss! Temos então o resultado: $U_N \xrightarrow{D} U$, com $U \sim N(0, \sigma\sqrt{t})$.

Um aspecto importante dessa nossa derivação é que como sub-produto obtemos um método de se aproximar o processo de condução do calor, ou da difusão de um soluto. Esse método será usado na seção seguinte para indagarmos sobre a reversibilidade no tempo de processos desse tipo.

Note que espacialmente, a variável aleatória U_N , que é uma função com domínio Ω_N , tem imagem no conjunto $[-2\sigma\sqrt{Nt}, +2\sigma\sqrt{Nt}]$.

O estudante deve prestar atenção no fato que em uma simulação da convergência $U_N \xrightarrow{D} U$, fixamos inicialmente t . A variável aleatória corresponde à distribuição de uma partícula ser encontrada em pontos do espaço, *no instante de tempo t , fixo previamente*. Isso é compatível com a nossa escolha para o espaço amostral Ω_N , componente essencial do nosso experimento probabilístico.

6 A reversibilidade no tempo

Agora que temos uma aproximação discreta para o processo de condução de calor, ou de difusão de solutos, dada pela variável aleatória U_N construída na seção 5, podemos fazer algumas simulações e discutir a questão da reversibilidade no tempo de processos de difusão.

Para cada simulação, fixamos $t > 0$, $N \in \mathbb{N}$, $\Delta t = \frac{t}{N}$ e $\Delta x = \frac{2\sigma\sqrt{t}}{\sqrt{N}}$. Dividimos o intervalo $[-2\sigma\sqrt{Nt}, +2\sigma\sqrt{Nt}]$ em $2N$ trechos. Os pontos extremos desses trechos definem os pontos, onde as partículas estacionam a cada passo de tempo Δt .

Para um ponto $i\Delta x$ no interior de $I = [-2\sigma\sqrt{Nt}, +2\sigma\sqrt{Nt}]$, supomos que a probabilidade de a partícula passar para o ponto $(i+1)\Delta x$ é $\frac{1}{2}$, igual à probabilidade de ir para a posição $(i-1)\Delta x$. Os pontos da fronteira I do intervalo não são absorventes. Mais precisamente, nos pontos da fronteira, a partícula permanece nesses pontos com probabilidade $\frac{1}{2}$ e retorna para o ponto mais próximo também com probabilidade $\frac{1}{2}$.

A simulação é facilitada ao implementarmos se trabalharmos com matrizes e vetores. A fração da quantidade de partículas em cada ponto onde elas se estacionam é codificada por um vetor $[w]$ de dimensão $1 \times (2N+1)$. A passagem de tempo é feita através da pós-multiplicação de $[w]$ por uma matriz de transição $[T]_N$ do seguinte tipo:

$$[T]_N = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

que é uma matriz de transição de Марков.

O estado, ou a distribuição de probabilidades $[w]_m$ após m passos é dado pela aplicação da matriz $[T]_N^m$ ao vetor $[w]_0$. Assim, após m passos, o estado inicial $[w]_0$ será transformado no estado $[w]_m = [w]_0[T]_N^m$.

A distribuição de probabilidades $[w]_N$ para a variável aleatória U_N é dada pela aplicação de N vezes a matriz de transição sobre o vetor inicial $[w]_0$ que tem todas as entradas nulas, exceto a $N+1$ -ésima, que vale 1 (é um vetor nulo, com um “1” bem na sua posição central). Isto é, a distribuição de probabilidades de U_N é dada por $[w]_N = [w]_0[T]_N^N$. A partir dessa distribuição de probabilidades discreta, obtemos uma aproximação da função densidade de probabilidade $x \mapsto u(t, x)$, para $t > 0$ fixo, que satisfaz a equação (18), através da fórmula

$$u(t, \cdot) \approx \frac{[w]_N}{\Delta x}, \text{ para } t = N\Delta t.$$

Na figura 3 está mostrada a aproximação de $x \mapsto u(t, x)$, $t = N\Delta t$, para $N = 4, 5, 6$.

Uma outra maneira de se interpretar a obtenção da distribuição de probabilidades $[w]_N$ para U_N é encarar que ela é o resultado de N processos de difusão, cada um com duração $\Delta t = \frac{t}{N}$, partindo de uma condição inicial que é fornecida pelo resultado do processo anterior. Em símbolos: $[w]_N = ((([w]_0[T]_N)[T]_N) \cdots [T]_N)$.

Assim, o operador $[T]_N$ fornece um meio de passar de um estado para outro em apenas um intervalo de tempo Δt . Na figura 4 está mostrada $[w]_N$ obtida com essa estratégia de se atingir o tempo final t com uma sucessão de passos de tamanho $\Delta t = \frac{t}{N}$. Os dados são $N = 3$, $t = 6s$, $\sigma^2 = 1.0 \times 10^{-2}m^2/s$. Nesse caso, $\Delta t = 2s$, $\Delta x = 2.82 \times 10^{-2}m$.

Da segunda lei da termodinâmica, esperamos que o processo seja irreversível. Afinal, ninguém nunca viu uma mancha se “desdifundir”, ou voltar no tempo. Pensando nisso, calculamos os determinantes da matriz de transição $[T]_N$ em função de sua ordem N . Quanto maior N , mais a variável aleatória U_N , se aproxima da normal, e o processo discreto se aproxima da evolução previsto pela equação do calor/difusão. Um cálculo simples revela que

$$\det([T]_N) = \frac{(-1)^n}{2^{2n}},$$

que decresce em módulo muito rapidamente. Então, embora $[T]_N$ seja inversível para todo N , as instabilidades numéricas na inversão de $[T]_N$ vão aumentando rapidamente à medida que N cresce.

Para o caso particular de $t = 0$, quando a condição inicial é extrema, a fórmula (3) prevê que ao voltarmos no tempo, isto é, quando $t < 0$, devemos ter probabilidades imaginárias!

Não podemos, entretanto, dizer que a inversão no tempo é impossível. A única coisa que podemos afirmar é que

$$\lim_{N \rightarrow +\infty} \det([T]_N) = 0,$$

e portanto, para $N \rightarrow +\infty$, o método aqui proposto não permite o retorno no tempo. Na verdade, por mais incrível que pareça, podemos identificar a condição inicial, a partir do estado em um tempo $t > 0$. Esse é um resultado da teoria dos problemas inversos. O que ocorre é que há uma condição muito forte para isso acontecer. Devemos conhecer com precisão absoluta a condição no tempo $t > 0$. Se houver algum erro, ou seja se o dado inicial a ser invertido não pertencer à imagem de um processo difusivo, o retorno à condição inicial retornará apenas ruído.

6.1 Uma aplicação no processamento de imagens

Podemos tomar uma foto e fazer com que a tinta que a compõe se difunda no papel. Vamos fazer isso digitalmente, considerando uma foto como sendo

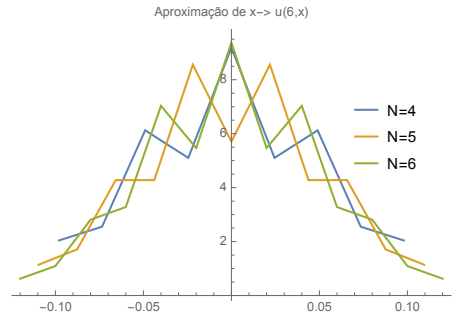


Figura 3: Aproximação de $x \mapsto u(t = 6, x)$.

uma matriz de números de ordem $M \times (2N + 1)$ que dá a tonalidade de cinza de cada um de seus pixels. Para cada linha dessa matriz, aplicamos o nosso operador de difusão $[T]_N$. O resultado está na figura 5. As figuras tem largura de $N = 567$ pixels, e portanto $N = (587 - 1)/2 = 283$. O determinante de $[T]_N$ é -4.14×10^{-171} .

Vamos imaginar um grande número de partículas envolvidas. Então, pela Lei Fraca dos Grandes Números, cujo enunciado é dado abaixo, e é conhecido dos alunos do curso de Probabilidades da Escola Politécnica da Universidade de São Paulo, a distribuição de probabilidades para U_N fornece uma boa aproximação para a fração da (enorme) quantidade de partículas em cada ponto do espaço no instante de tempo t .

Teorema 6.1 (Jacob Bernoulli (1713)). *Seja X_1, \dots uma sequência de variáveis aleatórias independentes e identicamente distribuídas com média μ e variância $\sigma^2 < \infty$. Seja $S_n = \sum_{j=1}^n X_j$. Então dado qualquer $\epsilon > 0$,*

$$\lim_{n \rightarrow +\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0.$$

Conseguimos aplicar a transformação inversa de $[T]_N$ e fazer com que a imagem borrada volte a ficar nítida? A resposta é sim, se o número de vezes m que o operador de difusão $[T]_N$ é aplicado for baixo. A figura 6 mostra o resultado para $m = 5$.

Quando o número m cresce, o “desborramento” não é mais possível. A sequência de figuras 7 mostra o resultado catastrófico.

Pela discussão no final da seção 6, o que está acontecendo é o seguinte: quando m é pequeno, a imagem obtida pela aplicação de $[T]_M^m$ não envolve erros numéricos significativos. Entretanto, quando m cresce, crescem também esses erros. Quando $[T]_M^{-1}$ é aplicado, aqueles erros são amplificados. O resultado é apenas ruído no resultado para a imagem inversa.

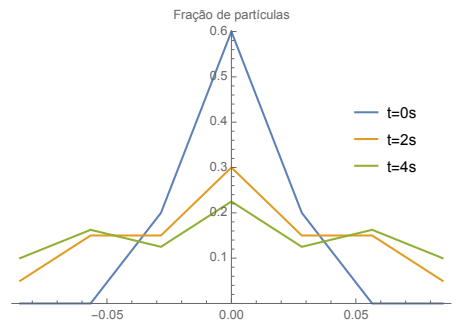
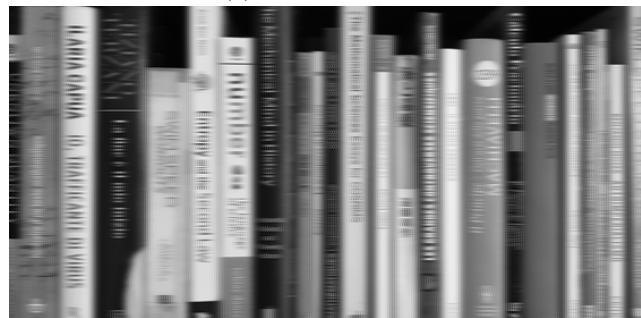


Figura 4: Evolução no tempo. Difusão.

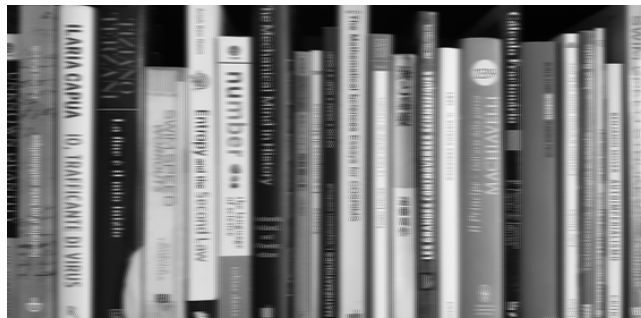


(a) Imagem original.



(b) Imagem com aplicação de $[T]_N^{10}$.

Figura 5: Borramento de imagem.



(a) Imagem original borrada.

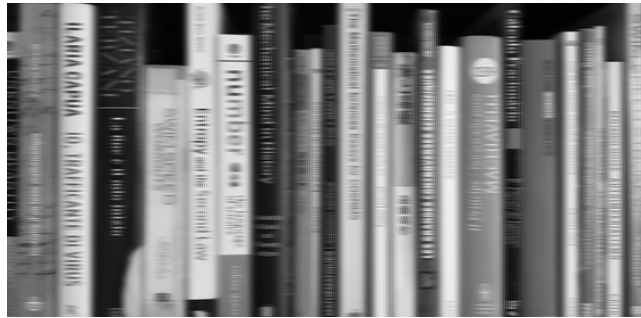


(b) Imagem com aplicação de $[T]_N^{-5}$.

Figura 6: Quando a imagem volta a ser nítida.



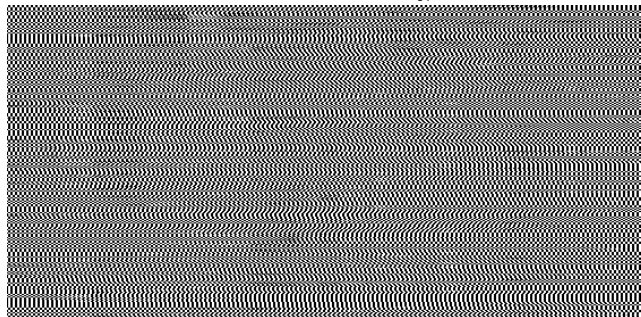
(a) Imagem original borrada após a aplicação de $[T]_N^{10}$.



(b) Retorno com $[T]_N^{-2}$.



(c) Retorno com $[T]_N^{-3}$.



(d) Retorno com $[T]_N^{-4}$.

Figura 7: Quando a imagem não volta a ser nítida.

7 Conclusões

Neste texto, revisitamos principalmente o conceito de Variância. De sua análise, chegamos à distribuição Normal via maximização da Entropia de Shannon. A função densidade de probabilidade gaussiana é aquela que tem a maior entropia, dado o conhecimento da sua média e da sua variância.

Mostramos uma ligação da distribuição de Gauss com a equação do calor/difusão, e a partir dela, fizemos algumas simulações que envolvem a reversibilidade no tempo de processos difusivos.

Agora, apenas gostaríamos de deixar registrados alguns pontos que consideramos importantes para o estudante que eventualmente ler este texto.

- A Variância não é um conceito naturalmente ligado à dispersão. Após a apresentação da distribuição Normal, a ligação entre variância e dispersão se torna natural. Hoje as palavras variância e dispersão são usadas no linguajar comum quase como sinônimos.
- A distribuição Normal pode ser obtida da convergência de variáveis aleatórias discretas.
- De todas as funções densidade de probabilidade, a gaussiana é aquela que, dadas média e variância fixas, contém tem a maior entropia no sentido de Shannon.

Referências

- [1] Paulo Domingos Cordaro and Alexandre Kawano. *O Delta de Dirac*. Livraria da Física, 2002.
- [2] Carlos Alberto Barbosa Dantas. *Probabilidade – Um Curso Introdutório*. EDUSP, 3rd. edition, 2013.
- [3] Ronald A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, page 399-433, 1918.
- [4] Alexandre Kawano. *Enumerabilidade e uma Ligação entre as Distribuições Uniforme e Geométrica*. (Artigo presente nesta coletânea)
- [5] Alexandre Kawano. *Valor esperado, distribuições geométrica e exponencial*. (Artigo presente nesta coletânea)
- [6] R. L. Plackett. Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares. *Biometrika*, 59(2):239–251, 1972.

Capítulo 4

Independência e o problema das três portas*

Alexandre Kawano
Universidade de São Paulo

Maio de 2020

Resumo

Discutimos neste trabalho a probabilidade condicionada, o conceito de independência e três problemas: o do casal com dois filhos, o de se extrair um número par do conjunto dos naturais e o famoso problema das três portas. O público alvo é o de alunos de graduação que já tenham sido expostos a algum curso de probabilidades.

1 Introdução

No problema das três portas tem o seguinte *script*. No palco, há três portas fechadas. Atrás de apenas uma delas há um prêmio. Um apresentador mostra as três portas fechadas a uma pessoa, e pede a ela para escolher uma delas, para que ganhe o prêmio que ela esconde. Depois da escolha, o apresentador abre uma outra, que naturalmente não tem o prêmio, pois ele sabe de antemão onde ele está. Em seguida, o apresentador faz a pergunta à pessoa: “você quer trocar a sua escolha inicial”?

Existe um paradoxo nesse problema. Antes da abertura de uma delas, mostrando que ela não contém o prêmio, a probabilidade da pessoa ganhar o jogo é $\frac{1}{3}$. Após a abertura da porta, como só restam duas, a probabilidade passa a ser $\frac{1}{2}$. Esse valor pode ser entendido da seguinte forma: depois que uma porta é eliminada, o apresentador faz uma pergunta que pode ser interpretada como um pedido para que uma nova escolha envolvendo apenas duas portas seja feita.

Por outro lado, há também o argumento de que *antes* da abertura da porta, já se sabia que não haveria nenhum prêmio atrás dela. Então, abrindo ou não uma porta, o resultado já era conhecido. Então, a pessoa poderia ter se virado de costas, ou ter pedido ao apresentador para que ele pulasse a parte de abrir a porta e o resultado seria o mesmo, pois já se sabia que ele mostraria um quarto vazio. Isso induz a pensar que a probabilidade da pessoa ganhar

*Alexandre Kawano, *Coletânea de artigos sobre probabilidade básica*, Epusp, ISBN 978-65-89190-15-8, DOI 10.11606/9786589190158

o prêmio continua sendo $\frac{1}{3}$, se manter a escolha inicial, e conseqüentemente a probabilidade do prêmio estar em C deve ser $\frac{2}{3}$. A assimetria pode parecer estranha, mas ela acontece porque o problema não é simétrico. Afinal, a porta A foi escolhida, e não C . O valor $\frac{1}{3}$ pode ser compreendido também facilmente se pensarmos na frequência das vezes que, ao se manter a decisão até o fim, o prêmio é ganho, o que é equivalente a acertar onde ele está já na primeira tentativa.

Suponha que a porta escolhida tenha sido a “A” e que o apresentador tenha aberto a “B”. A probabilidade que o prêmio está atrás da “B” é $\frac{1}{2}$ ou $\frac{1}{3}$?

Se a pessoa deve trocar de porta uma boa discussão. Vamos aproveitar este problema para revisar os conceitos de probabilidade condicionada, e independência de eventos e variáveis aleatórias, passando pela ideia de correlação linear.

Para dar uma resposta simples, vamos analisar as duas estratégias disponíveis, a de se manter ou de se trocar a primeira escolha. Elas estão mostradas na figura 1. As bolas escuras estão os pontos onde uma decisão é tomada. Na primeira, a pessoa escolhe uma porta. Na segunda, as possibilidades de decisão são de seguir a estratégia E_1 de manter firme a escolha, ou a estratégia E_2 , de trocá-la. Entre as duas bolas escuras, o apresentador abre uma porta, que não é aquela que foi escolhida, e que, naturalmente, não contém o prêmio.

As bolas brancas representam a “escolha” da Natureza. No ramo da estratégia E_1 , há duas possibilidades. Ou o prêmio está na porta escolhida (que ocorre com probabilidade $\frac{1}{3}$), ou não está. Da mesma maneira, o ramo E_2 se divide em duas possibilidades, pelo mesmo motivo. Note que para o ramo E_2 , se a pessoa tiver escolhido uma porta sem o prêmio inicialmente (que ocorre com probabilidade $\frac{2}{3}$), depois que ela trocou de opção, o prêmio só pode estar na última porta escolhida, pois o apresentador já abriu uma porta, que não contém nada.

Suponha que sua estratégia seja a E_1 . Se o prêmio Q estava atrás da escolhida, então o valor esperado do ganho para essa estratégia é

$$\frac{1}{3} \times Q + \frac{2}{3} \times 0 = \frac{Q}{3}.$$

Por outro lado, a estratégia E_2 de se trocar a escolha leva ao valor esperado

$$\frac{1}{3} \times 0 + \frac{2}{3} \times Q = \frac{2Q}{3}.$$

Por esta análise das estratégias, vemos que vale a pena trocar a escolha da porta, pois a estratégia E_2 leva a um valor esperado superior. Porém veremos ao longo do texto que o problema guarda nuances muito interessantes.

2 Probabilidade condicionada

Como já dito na introdução, este texto é escrito pensando em estudantes que já tenham sido expostos a um curso de probabilidades. Para revermos o conceito

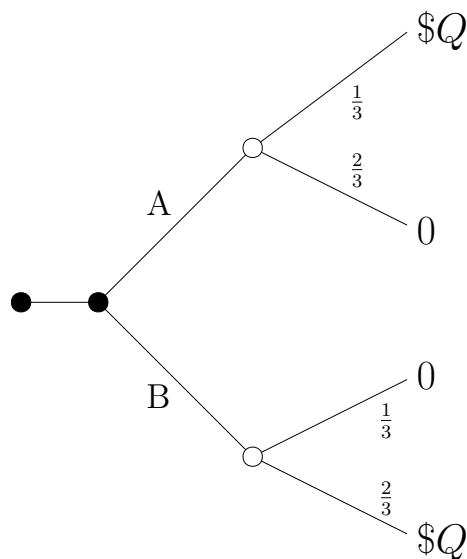


Figura 1: Árvore de decisão.

de probabilidade condicionada, começamos com o enunciado, também clássico, como o das três portas, do problema dos dois filhos.

Problema dos dois filhos. Suponha que um casal tenha dois filhos. Sabendo-se que um deles é menino, qual é a probabilidade que o outro também seja menino? Em geral, tem-se a tendência de responder-se $\frac{1}{2}$, porém essa resposta não está correta.

Para ver o por quê, consideramos uma outra questão. “Qual é a probabilidade de o segundo filho ser menino?” Para responder isso, vamos explicitar o espaço amostral Ω , o espaço de eventos $\mathcal{F} \subset \Omega$ e a função probabilidade $P : \mathcal{F} \rightarrow \mathbb{R}$.

O espaço amostral é o conjunto $\Omega = \{(O, O), (O, A), (A, O), (A, A)\}$. O espaço de eventos \mathcal{F} é o conjunto das partes de Ω , que tem ao todo $2^4 = 16$ elementos. Explicitamente,

$$\begin{aligned} \mathcal{F} = & \{\emptyset, \Omega, \{(O, O)\}, \{(O, A)\}, \{(A, O)\}, \{(A, A)\}, \{(O, O), (O, A)\}, \\ & \{(O, O), (A, O)\}, \{(O, O), (A, A)\}, \{(O, A), (A, O)\}, \{(O, A), (A, A)\}, \\ & \{(A, O), (A, A)\}, \{(O, O), (O, A), (A, O)\}, \\ & \{(O, A), (A, O), (A, A)\}, \{(O, O), (A, O), (A, A)\}, \{(O, O), (O, A), (A, A)\}\}. \end{aligned}$$

A probabilidade de cada evento de \mathcal{F} é calculada por técnicas elementares. Por exemplo, $P(\{(O, A)\}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Um cálculo mais trabalhoso, mas

elementar é

$$\begin{aligned} P(\{(O, O), (A, O)\}) &= P(\{(O, O)\}) + P(\{(A, O)\}) - P(\{(O, O)\} \cap \{(A, O)\}) \\ &= P(\{(O, O)\}) + P(\{(A, O)\}) - P(\emptyset) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

Na linguagem corrente, $P(\{(O, O), (A, O)\})$ é a probabilidade que o segundo filho seja um menino. O estudante deve perceber que aqui não entra *nada* de condicionamento. É um cálculo direto e básico.

Voltamos à questão de se saber qual é a probabilidade de que os dois filhos sejam meninos, sabendo-se que um deles é menino. Se soubermos que o menino mencionado é o primeiro, ou o segundo, então de fato $\frac{1}{2}$ é a resposta à pergunta: “Sabendo-se que um deles é menino, digamos o segundo, qual é a probabilidade que o primeiro também seja menino?”. Isso ocorre porque a informação dada não passa nada a respeito do sexo do outro filho.

Matematicamente podemos ver isso analisando mais de perto o espaço amostral e o espaço de eventos para a situação. Sabendo-se que o segundo filho é menino, temos *a priori* somente os seguintes elementos do espaço amostral original a serem considerados: $\{(A, O), (O, O)\}$. Podemos considerar esse conjunto como o novo espaço amostral Ω_1 . O espaço de eventos é o conjunto das partes de Ω_1 . A função de probabilidade deve obedecer $P_1(\Omega_1) = 1$. Como (A, O) e (O, O) são equiprováveis, é natural atribuir $P_1(\{(O, O)\}) = P_1(\{(A, O)\}) = \frac{1}{2}$. Aí está.

Entretanto, quando não se fornece nenhuma informação sobre a ordem, se o menino a que se refere a pergunta é o primeiro ou o segundo, a resposta não é $\frac{1}{2}$, mas sim $\frac{1}{3}$. Para entender isso, vamos olhar atentamente para o novo espaço amostral Ω_N . Dada aquela informação, consideramos

$$\Omega_N = \{(O, O), (O, A), (A, O)\}.$$

O espaço de eventos \mathcal{F}_N correspondente é o seu conjunto das partes.

Os eventos $\{(O, O)\}$, $\{(O, A)\}$, $\{(A, O)\}$ tem a mesma probabilidade no espaço probabilístico (Ω, \mathcal{F}, P) indicado no início da seção. De fato, cada uma dessas probabilidades vale $\frac{1}{4}$. São portanto equiprováveis do ponto de vista matemático, porque construímos $P : \mathcal{F} \rightarrow [0, 1]$ para que isso acontecesse e refletisse uma realidade física, ou biológica.

No novo espaço probabilístico $(\Omega_N, \mathcal{F}_N, P_N)$, a resposta à pergunta feita é expressa por $P_N(\{(O, O)\})$. Qualquer definição razoável de $P_N : \Omega_N \rightarrow [0, 1]$ deve fornecer $P_N(\{(O, O)\}) = \frac{1}{3}$, pois $\{(O, O)\}$ é uma possibilidade dentro do conjunto equiprovável $\{\{(O, O)\}, \{(O, A)\}, \{(A, O)\}\}$. Note que como $\Omega_N = \{(O, O), (O, A), (A, O)\}$, necessariamente devemos ter $P(\{(O, O)\}) + P(\{(O, A)\}) + P(\{(A, O)\}) = 1$.

Condicionamento e a questão do tempo. Em um espaço probabilístico (Ω, \mathcal{F}, P) geral, o espaço amostral contém os resultados experimentais possíveis,

sem entrar elementos temporais, como passado, presente e futuro. Ω simplesmente existe. Por outro lado, ao trabalharmos com probabilidades condicionadas, a questão temporal entra. Na notação $P(A|B)$, o evento $B \subset \mathcal{F}$ é pré condição, ou é anterior a A . O que ocorre na realidade é que dado o evento B , inicia-se um novo experimento probabilístico, com um novo espaço amostral $\Omega_B = \Omega \cap B$, um novo espaço de eventos $\mathcal{F}_B = \{Q \cap B | Q \in \mathcal{F}\}$ e uma nova função Probabilidade $P_B : \mathcal{F}_B \rightarrow [0, 1]$, que satisfaz $P_B(\Omega_B) = 1$.

Para continuar trabalhando normalmente com o espaço probabilístico original, como se nada tivesse acontecido com a introdução da informação dada por B , estendemos a função $P_B : \mathcal{F}_B \rightarrow [0, 1]$ ao espaço de eventos original \mathcal{F} . Essa extensão é a função $P(\cdot|B) : \mathcal{F} \rightarrow [0, 1]$ dada por

$$P(A|B) = P_B(A \cap B), \forall A \in \mathcal{F}.$$

Resta ainda o problema de se especificar $P_B : \mathcal{F}_B \rightarrow [0, 1]$. Definimos

$$P_B(A \cap B) = \frac{P(A \cap B)}{P(B)}.$$

Assim, a probabilidade condicionada de $A \in \mathcal{F}$ ao evento $B \in \mathcal{F}$ (ver por exemplo [1]) fica definida por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1)$$

Essa definição conduz ao resultado 1/3 que vimos para o problema dos dois filhos.

O estudante deve atentar ao fato que $P(\cdot|B)$ e $P(\cdot)$ são duas funções Probabilidade diferentes, sendo a nova $P(\cdot|B)$ definida em termos da básica $P(\cdot)$.

É também muito importante que A e B sejam eventos do mesmo espaço de eventos \mathcal{F} . Isso significa que frases como “ $P(\text{Sair cara}|\text{Choveu ontem})$ ” devem ser tratadas com cuidado. Em princípio não estão erradas, mas deve-se ter o cuidado de verificar-se se os eventos “Sair cara” e “Choveu ontem” pertencem ao mesmo espaço probabilístico. Isso em geral não é problemático. Basta especificar-se bem o experimento e o espaço amostral correspondente.

Da equação (1), vem de maneira elementar que

$$P(A \cap B) = P(A|B)P(B),$$

valendo para todos eventos $A, B \in \mathcal{F}$.

Teorema da Probabilidade Total. Tome um conjunto de eventos B_j , $j \in \mathbb{N}$ tais que $\cup_{j \in \mathbb{N}} B_j = \Omega$ e $i \neq j \Rightarrow B_i \cap B_j = \emptyset$. Isto é, o subconjunto de eventos $\{B_j\}_{j \in \mathbb{N}} \subset \mathcal{F}$ forma uma partição do espaço amostral Ω . Como para qualquer evento $A \in \mathcal{F}$ e qualquer B_j , $j \in \mathbb{N}$, temos $A \cap B_j \in \mathcal{F}$ e $A = \cup_{j \in \mathbb{N}} (A \cap B_j)$, temos

$$P(A) = \sum_{j \in \mathbb{N}} P(A \cap B_j) = \sum_{j \in \mathbb{N}} P(A|B_j)P(B_j). \quad (2)$$

Esse resultado, $P(A) = \sum_{j \in \mathbb{N}} P(A|B_j)P(B_j)$, é conhecido como *Teorema da Probabilidade Total*.

O problema do sorteio dos números pares. Usando a relação (2), vamos tentar resolver a questão de se saber qual é a probabilidade de se sortear “completamente ao acaso” um número par dentro do conjunto dos números naturais. Intuitivamente a resposta pareceria ser $\frac{1}{2}$, mas não é, como já mostrado em [3]. Revisitamos agora esse problema.

Para tentarmos justificar porquê a probabilidade de se sortear “completamente ao acaso” um número par dentro do conjunto dos números naturais seria $\frac{1}{2}$, poderíamos subdividir os naturais em subconjuntos de 10 elementos cada: $\mathbb{N} = \{1, 2, \dots, 10\} \cup \{11, 22, \dots, 20\} \cup \dots$. O conjunto desses subconjuntos forma uma partição de \mathbb{N} . Chame

$$B_j = \{10(j-1) + 1, 10(j-1) + 2, \dots, 10j\}.$$

Assim, $B_1 = \{1, 2, \dots, 10\}$, $B_2 = \{11, 22, \dots, 20\}$...

Denotamos por $\mathbb{A} = \{2, 4, \dots\}$ o conjunto dos números pares positivos.

Usando o Teorema da Probabilidade Total, podemos afirmar que

$$P(\mathbb{A}) = \sum_{j \in \mathbb{N}} P(\mathbb{A}|B_j)P(B_j).$$

Como $P(\mathbb{A}|B_j) = \frac{1}{2}$, para quaisquer valores que atribuamos a $P(B_j)$, $j = 1, 2, \dots$, lembrando que $\sum_{j \in \mathbb{N}} P(B_j) = 1$, devemos ter

$$P(\mathbb{A}) = \frac{1}{2}.$$

Entretanto, o argumento não está correto, pois poderíamos ter escolhido outra partição no lugar de $\{B_j\}_{j \in \mathbb{N}}$. Por exemplo,

$$C_j = \{12(j-1) + 1 + 2(n-1) | n = 1, 2, \dots, 6\} \cup \{6(j-1) + 2n | n = 1, 2, 3\}, \forall j \in \mathbb{N}.$$

$\{C_j\}_{j \in \mathbb{N}}$ é uma partição de \mathbb{N} . Exemplos: $C_1 = \{1, 3, 2, 5, 7, 4, 9, 11, 6\}$, $C_2 = \{13, 15, 8, 17, 19, 10, 21, 23, 12\}$. Como o experimento é “completamente ao acaso”, somos levados a afirmar, por simetria, que

$$P(\mathbb{A}|C_j) = \frac{1}{3}, \forall j \in \mathbb{N}.$$

Daí teríamos que concluir do Teorema da Probabilidade Total que

$$P(\mathbb{A}) = \frac{1}{3},$$

pois $P(\mathbb{A}) = \sum_{j \in \mathbb{N}} P(\mathbb{A}|C_j)P(C_j) = \sum_{j \in \mathbb{N}} \frac{1}{3}P(C_j) = \frac{1}{3} \sum_{j \in \mathbb{N}} P(C_j) = \frac{1}{3}$.

O absurdo não está no valor $\frac{1}{3}$, mas no fato que esse é um resultado diferente do anterior. E onde está o problema? Ele está na falta de rigor na definição

da função probabilidade $P : \Omega \rightarrow [0, 1]$. De fato, a disciplina a ser seguida é especificar o espaço amostral Ω , o espaço de eventos \mathcal{F} e a função Probabilidade, para depois começarmos a fazer os cálculos. No exemplo acima, como as probabilidades não foram definidas previamente, elas estão fluídas, mudando ao sabor da ilusão da simetria que queremos impor. Em resumo, a frase “completamente ao acaso” é muito fraca como especificação.

Especificação de probabilidades para eventos discretos. Devemos nos convencer que é impossível especificar um único valor de probabilidade para todos os números naturais, de forma que o sorteio deles seja equiprovável. De fato, se $\epsilon > 0$ é esse valor, então como há uma infinidade de números naturais, teríamos $\sum_{j \in \mathbb{N}} P(j) = \sum_{j \in \mathbb{N}} \epsilon$, que não é igual a 1. Então, se é assim, somos forçados a especificarmos o valor de $P(j)$, a probabilidade de se sortear o número natural j , para cada j .

Nessa especificação, entram nossos preconceitos, ilusões e informações prévias. Para um dado não viciado, costumamos atribuir probabilidade igual ($\frac{1}{6}$) para cada resultado possível de um lançamento, por simetria. A mesma coisa acontece com moedas e sexo dos filhos, embora nesse último caso não exista verdadeira simetria. Também por simetria e uma certa teimosia tendemos a dizer que o sorteio de um número natural “completamente ao acaso” ocorre com probabilidade $\frac{1}{2}$. Como já vimos, não é possível se especificar uma mesma probabilidade para todo número natural. Então, na especificação da probabilidade, poderíamos atribuir valores cada vez menores a números maiores. Por exemplo, $P(10^6) < P(2)$. Mas, devemos convir, não há nenhuma razão *a priori* para fazer-se isso.

Ao descrevermos o experimento, o espaço amostral e o espaço de eventos ficam definidos, e quando especificamos probabilidades no início da análise do problema, o “paradoxo” cessa. Por exemplo, podemos, para salvar o desejo de manter $P(\mathbb{A}) = \frac{1}{2}$, estabelecer

$$\begin{cases} P(2j) = \frac{1}{2^{(j+1)}}, \\ P(2j - 1) = \frac{1}{2^{(j+1)}}. \end{cases}$$

Nesse caso, $P(\mathbb{A}) = \frac{1}{2}$, mas note que $P(\mathbb{A}|C_j) \neq \frac{1}{3}$. Na verdade essa probabilidade condicionada varia. Os primeiros valores são mostrados na tabela 1.

Tabela 1: $P(\mathbb{A}|C_j)$

	j=1	j=2	j=3	j=4	j=5
$P(\text{Par} C_j)$	0.47058	0.87671	0.98273	0.99781	0.99972
$P(C_j)$	0.92969	0.06238	0.006956	8.56×10^{-4}	1.07×10^{-4}

2.1 Independência

No exemplo do sorteio de um número par visto acima, podemos calcular a probabilidade de um número ser par se soubermos que ele pertence à B_j . A fórmula a ser usada é a (1).

$$P(\mathbb{A}|B_j) = \frac{P(B_j \cap \mathbb{A})}{P(B_j)} = \frac{1}{2} = P(\mathbb{A}),$$

ou seja, saber que B_j ocorre, qualquer que seja $j \in \mathbb{N}$, não altera a probabilidade da ocorrência do evento “ser par”. Isso leva à noção de independência.

Independência de eventos. De maneira geral, dizemos, por definição, que dois eventos A e B são independentes quando

$$P(A|B) = P(A),$$

que aliada à fórmula (1), leva à

$$P(A \cap B) = P(A)P(B).$$

Entretanto, apesar da definição fazer sentido intuitivo, deve-se ter sempre em mente que o condicionamento altera o espaço probabilístico, e esse fato indica que a independência no sentido *hard* não existe.

A discussão é baseada em termos de probabilidades. Por isso, podem ocorrer situações que fogem à intuição. Dois casos extremos são intuitivamente aceitáveis. Suponha que A, B são eventos de um espaço probabilístico e $A \cap B = \emptyset$, com $P(A) \neq 0$ e $P(B) \neq 0$. Então eles não são independentes. De fato, $P(A \cap B) = P(\emptyset) = 0 \neq P(A)P(B)$. Isso é compreendido pensando que se um evento acontece, o outro não pode ocorrer e vice-versa. Quando $A = B$, $P(A) \neq 1$, eles também não são independentes: $P(A|A) = 1 \neq P(A)$. Realmente, se sabemos que A ocorre, não se pode dizer que A não ocorre...

Quando $A \cap B \neq \emptyset$, A e B podem ou não ser independentes. Por exemplo, seja a variável aleatória X distribuída uniformemente entre $[0, 1]$ (variáveis aleatórias uniformemente distribuídas são especialmente discutidas em [2]). Sejam $A = [a_1, a_2]$, $B = [b_1, b_2]$, contidos em $\Omega = [0, 1]$ e pertencentes ao espaço de eventos \mathcal{F} . Suponha que $A \cap B \neq \emptyset$, $a_2 > b_1$. Temos liberdade suficiente para escolher a_1, a_2, b_1, b_2 tais que

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{a_2 - b_1}{b_2 - b_1} = P(A) = a_2 - a_1,$$

como, por exemplo, como na figura 2. Jogando com os números, A e B podem ser ou deixar de ser independentes facilmente.

Na figura 3, está mostrado um caso em que os eventos A e C são independentes, B e C são independentes, mas A e B não são independentes.

Exemplos como o da figura 3 motivam uma definição mais forte para independência quando estão envolvidos mais de dois eventos. Dizemos, por definição,

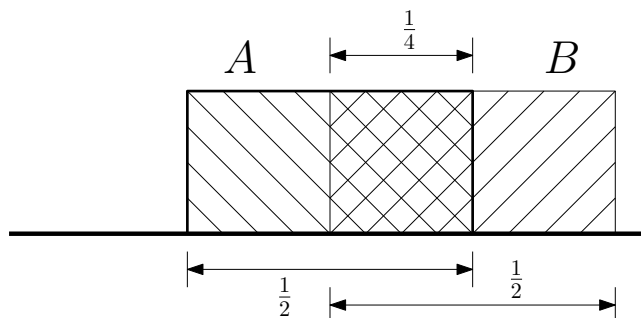


Figura 2: Eventos A e B independentes.

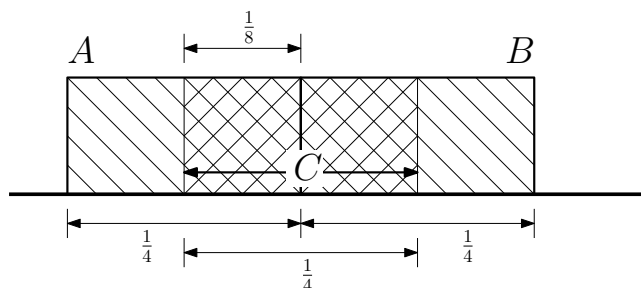


Figura 3: Eventos A e C são independentes. B e C são independentes. A e B não são independentes.

que a família $\{A_j\}_{j \in J}$, é independente se

$$P\left(\bigcap_{j \in I} A_j\right) = \prod_{j \in I} P(A_j), \quad (3)$$

para todo subconjunto *finito* I do conjunto de índices J .

Desta definição para um conjunto de eventos, vemos que o conjunto de eventos $\{A, B, C\}$ da figura 3 não é independente.

Vamos analisar agora o problema de um número ser par, quando usamos a partição $\{C_j\}_{j \in \mathbb{N}}$ vista acima. Agora, os resultados não são intuitivos. $P(\mathbb{A}|C_5) = 0.99972$, ao passo que $P(\mathbb{A}|C_1) = 0.47058$. Mas se olharmos os conjuntos C_5 e C_1 , $C_1 = \{1, 2, 3, 4, 5, 6, 7, 9, 11\}$ e $C_5 = \{26, 28, 30, 49, 51, 53, 55, 57, 59\}$, que têm a mesma quantidade de pares e ímpares. Entretanto, a probabilidade de ter sorteado um número par sabendo-se que ele pertence a C_5 é quase 1!

Evidentemente, o fato de ser um número par e pertencer ao conjunto C_5 não são eventos independentes. Isso é escrito matematicamente como $P(\mathbb{A}|C_5) \neq P(\mathbb{A})$. Já que esses eventos não são independentes, podemos perguntar qual é a

probabilidade de ter sido sorteado um número em C_5 sabendo-se que ele é par.

$$\begin{aligned} P(C_5|\mathbb{A}) &= \frac{P(\mathbb{A}|C_5)P(C_5)}{P(\mathbb{A})} \\ &= \frac{0.99972 \times 1.07 \times 10^{-4}}{1/2} = 2.136 \times 10^{-4}. \end{aligned}$$

Por outro lado, um cálculo simples revela que $P(C_1|\mathbb{A}) = 0.875$. Isso significa que saber que o número sorteado é par indica que a probabilidade dele pertencer a C_1 é mais de 4 mil vezes maior que pertencer a C_5 , mesmo que C_1 e C_5 contenham cada um três números pares e seis ímpares!

Até aqui discutimos a independência de eventos. Essa discussão se estende naturalmente à independência de variáveis aleatórias.

Independência de variáveis aleatórias. Já sabemos [2] que variáveis aleatórias são funções $X : \Omega \rightarrow \mathbb{R}$ tais que a imagem inversa de $] - \infty, a]$ é um evento, para todo $a \in \mathbb{R}$, isto é, $X^{-1}(] - \infty, a]) \in \mathcal{F}$.

Por definição, duas variáveis aleatórias X e Y definidas em um mesmo espaço amostral Ω , são independentes se os eventos $\{X \leq a\}$ e $\{Y \leq b\}$ são independentes para todos $a, b \in \mathbb{R}$. Da mesma forma, um conjunto de variáveis aleatórias $\{X_i\}_{i \in I}$ é independente se o conjunto de eventos $\{\{X_i \leq b_i\}_{i \in J}\}$, definidos em um mesmo espaço amostral Ω , para todo subconjunto $J \subset I$ finito e todos $b_i \in \mathbb{R}$. Isso significa, por exemplo, que as variáveis aleatórias X, Y e Z são independentes se

$$F_{X,Y,Z}(x, y, z) \doteq P(\{X \leq x\} \cap \{Y \leq y\} \cap \{Z \leq z\}) = F_X(x)F_Y(y)F_Z(z), \quad (4)$$

para todos $x, y, z \in \mathbb{R}$, em que $F_X : \mathbb{R} \rightarrow [0, 1]$ é a função distribuição de probabilidade acumulada da variável aleatória X . A notação é padrão (ver por exemplo [2], [1]).

É claro que o teste para a independência de variáveis aleatórias faz todas as combinações requeridas em (3), pois $\lim_{x \rightarrow +\infty} F_X(x) = 1$.

Do ponto de vista operacional, é muito mais fácil verificarmos (4) que (3), porque em geral temos fórmulas bem definidas para todos os elementos que aparecem em (4), e isso pode dar a impressão aos alunos que a independência de variáveis aleatórias é mais fácil que a de eventos, mas isso é apenas uma ilusão do ponto de vista conceitual.

Correlação de variáveis aleatórias. Aproveitamos a oportunidade para comentar o caso de variáveis aleatórias correlacionadas. Esse termo só se aplica a variáveis aleatórias e não a eventos em geral. A Covariância de duas variáveis aleatórias X e Y definidas em um mesmo espaço probabilístico (Ω, \mathcal{F}, P) é definida por

$$COV[X, Y] = E[X \times Y] - E[X]E[Y],$$

de modo que quando $X = Y$, $COV[X, X] = V[X]$.

Tabela 2: Distribuição de probabilidades para a V.A. X .

X	$P(X)$
-1	1/3
0	1/3
1	1/3

Quando X e Y são independentes, o conhecimento do valor assumido por uma não interfere na distribuição de probabilidades da outra. Assim, não há dúvida que

$$E[XY] = E[X]E[Y].$$

Isso significa que quando X e Y são independentes, sempre ocorre $COV[X, Y] = 0$. Por outro lado, o fato $COV[X, Y] = 0$ não implica que X e Y são independentes. Um exemplo basta para esclarecer isso. Suponha que X e Y sejam relacionadas por $Y = X^2$ e que X possa assumir os valores $-1, 0$ e 1 com as probabilidades mostradas na tabela 2.1.

A distribuição de probabilidades de X induz as de Y e de XY .

$$\begin{aligned} P(Y = -1) &= 0; \\ P(Y = 0) &= 1/3, \\ P(Y = 1) &= 2/3, \\ P(XY = -1) &= P(X^3 = -1) = 1/3, \\ P(XY = 0) &= P(X^3 = 0) = 1/3, \\ P(XY = 1) &= P(X^3 = 1) = 1/3. \end{aligned}$$

Com esses valores, calculamos

$$E[X] = 0, \quad E[Y] = \frac{2}{3}, \quad E[XY] = 0.$$

Então $COV[X, Y] = E[XY] - E[X]E[Y] = 0$. A covariância de X e Y é nula, mas X e Y não são independentes, pois $P(X = 1|Y = 0) = 0$, que é diferente de $P(X = 1) = 1/3$. Ocorre que a o Coeficiente de Correlação $\rho(X, Y)$ entre duas variáveis aleatórias X e Y é definida por (ver por exemplo [1]) por

$$\rho(X, Y) = \frac{COV[X, Y]}{\sqrt{V[X]V[Y]}},$$

e quando $\rho(X, Y) = 0$, dizemos que X e Y são *não correlacionadas*.

No caso do nosso exemplo, com $Y = X^2$, $\rho(X, Y) = 0$, indicando que X e Y não são correlacionadas. Mas, como sabemos essas variáveis aleatórias são independentes.

O caso da distribuição Normal. Independência implica não correlação, mas a não existência de correlação não implica independência. Entretanto, há um caso que merece menção, que em geral é omitido nos livros elementares de probabilidades. Merece atenção, porque envolve a distribuição Normal, que é tida por muitos como a mais importante de todas (sobre a Normal, ver por exemplo [1] ou [4]).

No caso da distribuição Normal, não correlação implica independência. Vejamos porquê isso acontece. A função densidade de probabilidade f_X da normal $X \sim N(\mu, \sigma^2)$, é dada por

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

A função densidade de probabilidade conjunta de (X, Y) normalmente distribuída é dada por

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{|C[X, Y]|}} e^{-\frac{1}{2}([x-\mu_X, y-\mu_Y]^t C[X, Y]^{-1} [x-\mu_X, y-\mu_Y])},$$

em que $C[X, Y]$ é a matriz de covariância dada por

$$C[X, Y] = \begin{bmatrix} V[X] & \rho(X, Y)\sigma_X\sigma_Y \\ \rho(X, Y)\sigma_X\sigma_Y & V[Y] \end{bmatrix}.$$

Quando $\rho(X, Y) = 0$,

$$([x - \mu_X, y - \mu_Y]^t C[X, Y]^{-1} [x - \mu_X, y - \mu_Y]) = V[X](x - \mu_X)^2 + V[Y](y - \mu_Y)^2$$

e $|C[X, Y]| = V[X]V[Y]$. Então, a função densidade de probabilidade conjunta $f_{X,Y}$ pode ser escrita pelo produto

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

A consequência imediata é que $E[XY] = E[X]E[Y]$. Conclusão inevitável: para variáveis aleatórias normalmente distribuídas não correlação implica independência.

3 Análise do problema das três portas

Retomamos o enunciado do problema das três portas feito na Introdução deste artigo, mas agora apreciando algumas versões.

Alguém que chega atrasado. Suponha que um menino tenha chegado atrasado e não saiba nada a respeito do jogo. Ele vê apenas duas portas fechadas, digamos A e C , uma porta aberta, e escuta a pergunta sobre em qual porta está o prêmio. Para o menino, a probabilidade de ele estar em A é $\frac{1}{2}$, assim como em C .

Você que chegou quando ouviu a pergunta do apresentador. Consideremos agora uma outra situação. Suponha que você, que conhece as regras do jogo, chegou no exato momento em que uma porta está escancarada aberta, o apresentador e a pessoa estão no palco e a pergunta “...qual das duas portas você escolhe?” acaba de ser formulada. Digamos a que a porta do meio, a porta B esteja aberta. Do seu ponto de vista, qual é a probabilidade que o prêmio esteja em A ?

Do seu ponto de vista, que acaba de chegar, e não sabe o que aconteceu até este ponto, a não ser que havia três portas fechadas e um prêmio atrás de uma delas, o espaço amostral é $\Omega = \{A, B, C\}$, indicando que o prêmio está em A , B ou C . O espaço de eventos é necessariamente o conjunto das partes de Ω , isto é, $\mathcal{F} = \{\{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}, \emptyset\}$. A função de probabilidade $P : \mathcal{F} \rightarrow [0, 1]$ especifica $F(\{A\}) = F(\{B\}) = F(\{C\}) = \frac{1}{3}$, $F(\Omega) = 1$, $F(\emptyset) = 0$, $\{A\}$, $\{B\}$ e $\{C\}$ sendo mutuamente excludentes, as probabilidades para os demais elementos de \mathcal{F} são facilmente calculadas.

Como você sabia que havia três portas e um prêmio escondido, o espaço probabilístico inicial era (Ω, \mathcal{F}, P) , mas quando você chegou e viu a porta aberta, esse espaço probabilístico foi alterado. A questão temporal entra através do condicionamento, como discutido na seção 2. Ao ver que o prêmio não está atrás de B , ele está atrás de A ou C . Matematicamente, a probabilidade de o prêmio estar em A passa de $P(A)$ para $P(A|\{A, C\})$. O novo espaço probabilístico é $(\Omega_B, \mathcal{F}_B, P(\cdot|\{A, C\}))$, em que $\Omega_B = \Omega \cap \{A, C\}$, $\mathcal{F}_B = \{f \cap \{A, C\} | f \in \mathcal{F}\}$. Concretamente,

$$\Omega_B = \{A, C\}, \mathcal{F}_B = \{\{A\}, \{B\}, \{C\}, \{A, C\}, \emptyset\}.$$

A função Probabilidade é dada por

$$P(\cdot|\{A, C\}) = \frac{P(\cdot \cap \{A, C\})}{P(\{A, C\})}.$$

Temos em particular $P(\{B\}|\{A, C\}) = 0$ e

$$P(\{A\}|\{A, C\}) = \frac{P(\{A\} \cap \{A, C\})}{P(\{A, C\})} = \frac{P(\{A\})}{P(\{A, C\})} = \frac{1/3}{2/3} = \frac{1}{2}.$$

Do seu ponto de vista, essa é a probabilidade que o prêmio esteja atrás da porta A , que é a mesma de estar atrás da outra porta fechada, pois as probabilidades devem somar 1.

Vamos relacionar esse resultado com a probabilidade de $\frac{1}{3}$ e $\frac{2}{3}$, respectivamente, para as estratégias de se manter e trocar a escolha inicial. Para quem está assistindo o jogo desde o início, essas são as probabilidades, pois ele/a sabe qual é a escolha inicial. Você não sabe qual era a escolha inicial, então para você resta tomar o valor esperado de $\frac{1}{3}$ e $\frac{2}{3}$ com igual distribuição, pois há total simetria. O resultado é o seu $\frac{1}{2}$.

Aqui fazemos um comentário para que o estudante tenha consciência clara sobre isso. Como o problema probabilístico depende do ponto de vista, na vida real todos as probabilidades são condicionadas.

Para alguém que sabe que uma das portas tem um problema. Tome-mos agora o ponto de vista de uma membro do *staff* do programa que participou da montagem das portas, sabe que há um pequeno problema na dobradiça da porta B , mas não sabe onde está o prêmio. Ela está na plateia desde o início do show. Ela viu que a pessoa escolheu a porta A e o apresentador abriu a B , justamente aquela que tinha um problema na dobradiça! Para ela, qual é a probabilidade de que o prêmio está na porta escolhida A ?

Para responder a essa pergunta, vamos explicitar a experiência e o espaço probabilístico. Vamos fixar a porta A como a escolhida. Por simetria, essa escolha é inócua. No experimento, não há um sorteio, o da porta onde se esconde o prêmio, mas *dois*. Há também o da porta mostrada! Denotando por (X_1, X_2) o resultado do experimento, em que X_1 é o resultado do primeiro sorteio, isto é, onde está o prêmio, e X_2 o do segundo, ou seja, qual porta é mostrada, o espaço amostral é

$$\Omega = \{(A, B), (A, C), (B, C), (C, B)\}.$$

Note que o resultado (A, A) não é possível, pois o apresentador nunca abre a porta escolhida. (B, B) ou (C, C) também não são possíveis, porque ele também não abre a porta que esconde o prêmio.

O espaço de eventos é necessariamente o conjunto das partes de Ω .

$$\begin{aligned} \mathcal{F} = & \{\emptyset, \{(A, B)\}, \{(A, C)\}, \{(B, C)\}, \{(C, B)\}, \{(A, B), (A, C)\}, \\ & \{(A, B), (B, C)\}, \{(A, B), (C, B)\}, \{(A, C), (B, C)\}, \{(A, C), (C, B)\}, \\ & \{(B, C), (C, B)\}, \{(A, B), (B, C), (C, B)\}, \{(A, C), (B, C), (C, B)\}, \\ & \{(A, B), (A, C), (C, B)\}, \{(A, B), (A, C), (B, C)\}, \Omega\}. \end{aligned}$$

A função probabilidade deve ser também especificada. Podemos atribuir a maioria dos valores com o argumento da simetria, ou simplesmente pelas propriedades da função $P\mathcal{F} \rightarrow [0, 1]$. Assim,

- $P(\emptyset) = 0, P(\Omega) = 1,$
- $P(\{(B, C)\}) = \frac{1}{3}$ (pois se o prêmio está em B , a única possibilidade é a abertura de C , e a probabilidade de estar em B é $\frac{1}{3}$),
- $P(\{(C, B)\}) = \frac{1}{3},$
- $P(\{(A, B), (A, C)\}) = \frac{1}{3}$ (pois em linguagem corrente $P(\{(A, B), (A, C)\})$ é a probabilidade do prêmio estar em A , lembrando que (A, A) é impossível),
- $P(\{(B, C), (C, B)\}) = \frac{2}{3}$, pela soma de dois eventos mutuamente exclusivos,
- $P(\{(A, B), (A, C), (C, B)\}) = \frac{2}{3}$, pois $P(\{(A, B), (A, C), (C, B)\}) = P(\{(A, B), (A, C)\}) + P(\{(C, B)\}) = \frac{1}{3} + \frac{1}{3}.$
- $P(\{(A, B), (A, C), (B, C)\}) = \frac{2}{3}.$

Para os demais valores da função probabilidade há liberdade de escolha. Vamos especificar

$$P(\{(A, B)\}) = \frac{p}{3}.$$

Agora podemos completar a definição de $P\mathcal{F} \rightarrow [0, 1]$.

- $P(\{(A, C)\}) = \frac{(1-p)}{3},$
- $P(\{(A, B), (B, C)\}) = \frac{p}{3} + \frac{1}{3},$
- $P(\{(A, B), (C, B)\}) = \frac{p}{3} + \frac{1}{3},$
- $P(\{(A, C), (B, C)\}) = \frac{(1-p)}{3} + \frac{1}{3},$
- $P(\{(A, C), (C, B)\}) = \frac{(1-p)}{3} + \frac{1}{3},$
- $P(\{(A, B), (B, C), (C, B)\}) = \frac{p}{3} + \frac{2}{3},$
- $P(\{(A, C), (B, C), (C, B)\}) = \frac{(1-p)}{3} + \frac{2}{3}.$

Após a abertura da porta, o espaço probabilístico é alterado, como já discutido na seção 2. A membro do *staff* viu que o apresentador abriu a B , ou seja ocorreu o evento $\{(A, B), (C, B)\}$. O espaço amostral passa a ser

$$\Omega_B = \Omega \cap \{(A, B), (C, B)\} = \{(A, B), (C, B)\}.$$

O espaço de eventos passa a ser $\mathcal{F}_B = \{x \cap \{(A, B), (C, B)\} | x \in \mathcal{F}\}$. Então

$$\mathcal{F}_B = \{\emptyset, \{(A, B)\}, \{(C, B)\}, \{(A, B), (C, B)\}, \Omega\}.$$

A função probabilidade condicionada, que vamos denotar por $P(\cdot|B)$, e não por $P(\cdot|\{(A, B), (C, B)\})$, para não carregar a notação, é dada por

$$P(\cdot|B) = \frac{P(\cdot)}{P(\{(A, B), (C, B)\})} = \frac{P(\cdot \cap \{(A, B), (C, B)\})}{\frac{p}{3} + \frac{1}{3}}.$$

Para a membro do *staff* do programa, quando ela vê que a porta B é aberta, imediatamente, para ela a probabilidade que o prêmio está atrás da porta A é

$$\begin{aligned} P(\{(A, B), (A, C)|B\}) &= \frac{P(\{(A, B), (A, C)\} \cap \{(A, B), (C, B)\})}{\frac{p}{3} + \frac{1}{3}} \\ &= \frac{P(\{(A, B)\})}{\frac{p}{3} + \frac{1}{3}} = \frac{\frac{p}{3}}{\frac{p}{3} + \frac{1}{3}} = \frac{p}{1+p}. \end{aligned}$$

Antes de perguntar sobre o valor de p , vemos que como o resultado é sempre menor ou igual a $\frac{1}{2}$, para qualquer valor de p no intervalo $[0, 1]$, nunca é vantajoso trocar a porta. Um simples gráfico (figura 4) de $p \mapsto \frac{p}{1+p}$ revela que $P(\{(A, B), (A, C)|B\})$, isto é, a probabilidade do prêmio estar atrás da porta A depois da abertura da B , pode assumir inclusive os valores $\frac{1}{3}$ e $\frac{1}{2}$ das análises anteriores, para $p = \frac{1}{2}$ e $p = 1$, respectivamente.

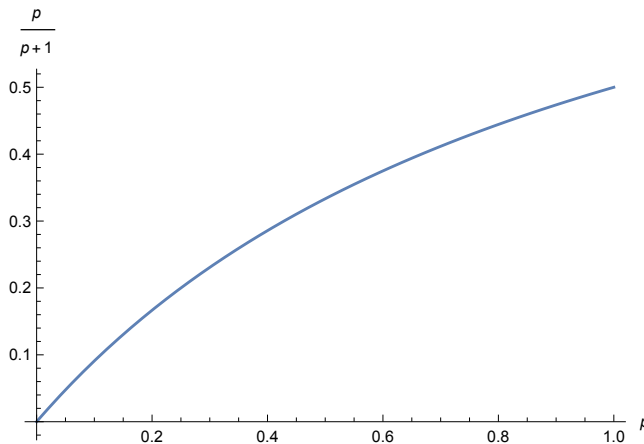


Figura 4: Gráfico da probabilidade que o prêmio está atrás da porta A , após a abertura da porta B .

E quanto vale p ? Para a membro, que sabe que a dobradiça da porta B estava com problema, p deve ser um valor bem pequeno. Para ela, intuitivamente o prêmio deve estar em C com grande chance, pois como a porta A foi escolhida, o apresentador tinha duas possibilidades: ou abrir B ou abrir C . Mas B tinha um problema, e apesar disso ela foi aberta. Logo, muito provavelmente a C esconde o prêmio. Digamos que para ela, $p = 0.1$. Então, a troca de opção leva a uma probabilidade de quase 0.91 da pessoa sair feliz do programa.

Para um senhor que nada sabe. Para um senhor na plateia que nada sabe, e pensa que quando o apresentador tiver liberdade de escolha ele vai escolher uma entre as opções “completamente ao acaso” (como jogando uma moeda não viciada), o valor de p é $p = \frac{1}{2}$. Nesse caso, a abertura da porta não deve alterar o a probabilidade do prêmio estar na porta originalmente escolhida, pois o ato de se abrir uma porta é inconsciente (decisão tomada pelo lançamento de uma moeda). De fato, quando $p = \frac{1}{2}$, a probabilidade do prêmio estar atrás de A continua sendo $\frac{1/2}{1+1/2} = \frac{1}{3}$.

A independência e a ilusão. O estudante deve observar que como P (“prêmio estar atrás a porta A ”) e P (“prêmio estar atrás a porta A ”|“apresentador abrir

a porta B ”) são iguais, isto é, que

$$P(\{(A, B), (A, C)|B\}) = P(\{(A, B), (A, C)\}) = \frac{1}{3},$$

então os eventos “prêmio estar atrás a porta A ” e “apresentador abrir a porta B ” são *independentes*. Entretanto isso não significa que o dado condicionante “apresentador abrir a porta B ” não tenha influência sobre o problema. Pelo contrário. Ele altera profundamente sua estrutura, pois altera o seu espaço probabilístico Ω, \mathcal{F}, P).

Fechamos um ciclo, mas o problema não termina aí.

Podemos nos perguntar o que aconteceria se a porta aberta fosse a C . Nesse caso, o espaço probabilístico seria $\Omega_C = \{(A, C), (B, C)\}$, $\mathcal{F}_C = \{\emptyset, \Omega_C, \{(A, C)\}, \{(B, C)\}\}$ e a função probabilidade, que denotaremos por $P(\cdot|C)$ no lugar de $P(\cdot|\{(A, C), (B, C)\})$, fornece

$$\begin{aligned} P(\{(A, B), (A, C)|C\}) &= \frac{P(\{(A, B), (A, C)\} \cap \{(A, C), (B, C)\})}{\frac{(1-p)}{3} + \frac{1}{3}} \\ &= \frac{P(\{(A, C)\})}{\frac{2-p}{3}} = \frac{\frac{(1-p)}{3}}{\frac{2-p}{3}} = \frac{1-p}{2-p}. \end{aligned}$$

Essa é a probabilidade, do ponto de vista da membro do *staff*, do prêmio estar em A após a porta C ter sido aberta. Para ela, p é um valor baixo. Assumindo $p = 0.1$, $P(\{(A, B), (A, C)|C\}) = 0.473684$. Pode parecer que esse resultado seja muito diferente daquele que o senhor da plateia, que nada sabia, para quem $p = \frac{1}{2}$, e por isso $P(\{(A, B), (A, C)|C\}) = \frac{1}{3}$, mas em um certo sentido, não é bem assim. Vejamos em que sentido esses resultados são muito parecidos, ou até mesmo idênticos.

Do ponto de vista da membro do *staff*, calculemos o valor esperado para a fração das vezes em que o prêmio está atrás da porta A (a escolhida) depois que uma porta (ou *Bou C*) sem o prêmio é aberta. Esse valor esperado é calculado por

$$\begin{aligned} E[P(\text{“Prêmio em } A\text{”})] &= P(\text{“Prêmio em } A\text{”}|\text{“Abre } C\text{”})P(\text{“Abre } C\text{”}) \\ &\quad + P(\text{“Prêmio em } A\text{”}|\text{“Abre } B\text{”})P(\text{“Abre } B\text{”}) \\ &= P(\{(A, B), (A, C)|C\})P(\{(A, C), (B, C)\}) \\ &\quad + P(\{(A, B), (A, C)|B\})P(\{(A, B), (C, B)\}) \\ &= \frac{1-p}{2-p} \left(\frac{(1-p)}{3} + \frac{1}{3} \right) + \frac{p}{1+p} \left(\frac{p}{3} + \frac{1}{3} \right) \\ &= \frac{1}{3}. \end{aligned}$$

Como digo nas minhas aulas presenciais, um curso de Probabilidades e de Mágica são muito próximos. Em probabilidades há muita arte da prestidigitação, e é claro, muito prazer e encantamento. *You can do magic!*

4 Conclusões

Neste texto, revisitamos principalmente o conceito de probabilidade condicionada e o de independência quando falamos de eventos e variáveis aleatórias. Não se trata de um assunto fácil. De fato, usando as ideias de condicionamento, podemos construir vários problemas cujas soluções parecem paradoxais. Entre muitos paradoxos, um dos mais conhecidos é o das três portas, que discutimos neste texto.

Agora, gostaríamos de deixar registrados alguns pontos que consideramos importantes para o estudante que eventualmente ler este texto.

- Com o condicionamento, é introduzida a questão da pré-condição ou da temporalidade na teoria das probabilidades.
- O condicionamento cria um novo espaço probabilístico com seu próprio espaço amostral, espaço de eventos e função probabilidade.
- Na vida real, podemos dizer que todas as probabilidades são condicionadas.
- Independência implica não correlação, mas o inverso é em geral falso. A exceção importante é o caso de duas variáveis aleatórias normalmente distribuídas. Nesse caso particular, independência é equivalente a não correlação.
- É sempre muito importante explicitar o experimento, o espaço amostral, o espaço de eventos e a função probabilidade.

O autor agradece toda a equipe da disciplina de Probabilidades que é oferecida aos alunos do terceiro semestre dos cursos de Engenharia da Escola Politécnica da Universidade de São Paulo.

Referências

- [1] Carlos Alberto Barbosa Dantas. *Probabilidade – Um Curso Introdutório*. EDUSP, 3rd. edition, 2013.
- [2] Alexandre Kawano. *Enumerabilidade e uma Ligação entre as Distribuições Uniforme e Geométrica*. (Artigo presente nesta coletânea)
- [3] Alexandre Kawano. *Valor esperado, distribuições geométrica e exponencial*. (Artigo presente nesta coletânea)
- [4] Alexandre Kawano. *A variância, a distribuição normal e a marcha reversa no tempo*. (Artigo presente nesta coletânea)

Capítulo 5

A Lei dos Grandes Números e a representação binária de números no intervalo $[0, 1[$ *

Alexandre Kawano
Universidade de São Paulo

Junho de 2022

Resumo

Neste artigo, que tem por público alvo alunos de graduação que já tiveram alguma exposição a um curso de Probabilidades, mostramos uma relação entre a distribuição uniforme, a lei dos grandes números, independência de variáveis aleatórias e a representação binária de números reais no intervalo $[0, 1[$.

1 Introdução

Neste artigo consideramos um experimento probabilístico que consiste em se sortear um número real do intervalo $\Omega = [0, 1[$. Nós assumimos que o resultado do sorteio é dado pela variável aleatória $X : \Omega \rightarrow \mathbb{R}$ distribuída de acordo com a distribuição uniforme no intervalo $[0, 1[$. Formalmente, $\Omega = [0, 1[$ é o espaço amostral, \mathcal{F} é o espaço dos eventos, como definido em [1] e $P : \mathcal{F} \rightarrow [0, 1]$ é a função probabilidade. Só para reforçar, se $a, b \in [0, 1[$, então $P(X \in [a, b]) = b - a$.

Sabemos que cada número $\omega \in \Omega$ (aqui fazemos a identificação entre o elemento do espaço amostral $\omega \in \Omega$ e o valor que variável aleatória $X(\omega) = x$ assume) pode ser representado usando a base 2, isto é, ω pode ser representado por $\omega = x = 0.x_1x_2 \dots x_n \dots$, em que $x_i \in \{0, 1\}$ é um dígito da representação. Vale a relação

$$\omega = \sum_{j=1}^{+\infty} \frac{x_j}{2^j}.$$

Existem números $\omega \in \Omega$ que tem duas representações, uma que termina (após uma determinada posição, só há zeros à direita) e outra que não termina.

*Alexandre Kawano, *Coletânea de artigos sobre probabilidade básica*, Epusp, ISBN 978-65-89190-15-8, DOI 10.11606/9786589190158

Isso ocorre porque

$$\sum_{j=n_0}^{+\infty} \frac{1}{2^j} = \frac{1}{2^{n_0-1}}.$$

Por exemplo, $\frac{1}{2} = 0.1$ e ao mesmo tempo $\frac{1}{2} = 0.011111\dots$. No sistema decimal, isso é manifesto, por exemplo no clássico exemplo $1 = 0.99999\dots$

Quando há duas representações binárias, descartamos aquela que não termina. O leitor pode se perguntar qual seria o tamanho do conjunto dos números envolvidos no descarte. Os números envolvidos formam um conjunto desprezível no seguinte sentido. Seja o conjunto

$$D \doteq \{\omega \in \Omega : X(\omega) \text{ tem duas representações}\}.$$

Os elementos de D são aqueles que na representação binária tem um número finito de dígitos significativos (só tem zeros à direita a partir de uma determinada posição). Eles formam um subconjunto próprio dos racionais, e portanto, como já vimos em [1], D tem probabilidade nula.

Nessa situação em que agora garantimos que há apenas uma representação binária para cada elemento de $[0, 1[$, consideramos as variáveis aleatórias $X_n(\omega) = x_n$. Em outras palavras, a variável aleatória $X_n : \Omega \rightarrow \mathbb{R}$ retorna a n -ésima casa da representação decimal de ω .

Supomos que a variável aleatória X , que dá o resultado do sorteio é uniformemente distribuída. As questões que levantamos aqui é sobre a distribuição dos dígitos 0 e 1 da representação binária de ω . Especificamente, desejamos saber se

- Q1.** Os zeros e uns são igualmente distribuídos, isto é, se $E[X_j] = \frac{1}{2}$, para todo $j \in \mathbb{N}$.
- Q2.** As variáveis aleatórias X_i e X_j são independentes, para todos $i, j \in \mathbb{N}$, $i \neq j$.
- Q3.** Para quais $\omega \in \Omega$ a frequência de zeros e de uns na sua representação binária é igual.

A análise das questões nos permitirá ver uma conexão entre a Lei dos Grandes Números e a distribuição Uniforme, passando por considerações a respeito da independência de variáveis aleatórias.

2 A Lei dos Grandes Números

Suponha um experimento probabilístico com espaço amostral Ω , espaço de eventos \mathcal{F} e uma função probabilidade $P : \mathcal{F} \rightarrow [0, 1]$. Seja uma sequência de variáveis aleatórias $(W_n)_{n \in \mathbb{N}}$ independentes associada ao experimento probabilístico, tal que $W_n = 1$ com probabilidade p e $W_n = 0$ com probabilidade $(1 - p)$. *Grosso modo* a Lei dos Grandes Números diz que a sequência $\sum_{n=1}^m \frac{W_n}{m}$ converge para p . Intuitivamente esse é um resultado razoável, tão intuitivo que

serviu como base para a interpretação frequentista do que seria a probabilidade da ocorrência de um evento. Entretanto, muitas ideias são subjacentes a esse resultado.

Em primeiro lugar, precisamos analisar como interpretar a convergência

$$S_m \doteq \frac{\sum_{n=1}^m W_n}{m} \rightarrow p.$$

A resposta simples seria fazer uma analogia com o Cálculo e dizer que dado $\epsilon > 0$, existe m_0 tal que $m \geq m_0 \Rightarrow |S_m - p| < \epsilon$. Essa resposta é parcialmente correta, pois ela deixa de lado o fato fundamental que as variáveis aleatórias são funções definidas em um espaço amostral Ω . Uma forma explícita e correta seria escrever

$$S_m(\omega) = \frac{\sum_{n=1}^m W_n(\omega)}{m} \rightarrow p, \quad (1)$$

em que $\omega \in \Omega$ aparece explicitamente. Note que ω é o mesmo para todas as variáveis aleatórias. Não realizamos m experimentos para obter S_m . O experimento é realizado uma única vez, e as m variáveis aleatórias são calculadas. O que pode acontecer é que o experimento pode ser constituído de uma sucessão de ensaios, como lançamentos de uma moeda, por exemplo. Entretanto, como é claro no caso dos dígitos binários da representação de $\omega \in [0, 1[$, nem sempre trata-se de uma sucessão de ensaios, mas verdadeiramente *um* sorteio. Nesse caso $W_n(\omega)$ sendo, por exemplo o n -ésimo dígito de ω , temos uma sucessão de variáveis aleatórias baseadas em um único resultado experimental.

Voltemos a (1). Após a discussão acima, percebemos que falta especificar para quais $\omega \in \Omega$ a convergência acontece. Ao especificar isso, teremos as versões fraca e forte da Lei dos Grandes Números.

2.1 A Lei Fraca dos Grandes Números

Na Lei Fraca dos Grandes Números, a convergência em (1) ocorre *em probabilidade*. Isso quer dizer que dado qualquer $\epsilon > 0$, temos

$$\lim_{m \rightarrow +\infty} P(|S_m - p| > \epsilon) = 0.$$

Note que a função probabilidade P é exatamente aquela apresentada no início da seção 2. Apesar do nome “Lei”, a Lei Fraca dos Grandes Números é na verdade um teorema.

Vamos formalizar o seu enunciado e após apresentarmos a desigualdade de Chebychov (Чебышев) damos uma prova do teorema, que é simples e instrutiva.

Teorema 2.1 (Lei Fraca dos Grandes Números (Bernoulli)). *Suponha um experimento probabilístico com espaço amostral Ω , espaço de eventos \mathcal{F} e uma função probabilidade $P : \mathcal{F} \rightarrow [0, 1]$. Seja uma sequência de variáveis aleatórias $(W_n)_{n \in \mathbb{N}}$ independentes e identicamente distribuídas associada ao experimento probabilístico, tal que $W_n = 1$ com probabilidade p e $W_n = 0$ com probabilidade $(1 - p)$. Então para todo $\epsilon > 0$,*

$$\lim_{m \rightarrow +\infty} P(|S_m - p| > \epsilon) = 0.$$

Antes de apresentarmos a prova do teorema acima, vemos a desigualdade de Chebychov (Чебышев).

Teorema 2.2 (Chebychov). *Seja um espaço probabilístico com espaço amostral A , espaço de eventos T e função de probabilidades $\text{Prob} : T \rightarrow [0, 1]$. Seja uma variável aleatória $Y : A \rightarrow \mathbb{R}$ que possui valor esperado $E[Y] \in \mathbb{R}$ e variância $V[Y]$. Então*

$$\text{Prob}(|Y - E[Y]| \geq \epsilon) \leq \frac{V[Y]}{\epsilon^2}.$$

O resultado é extraordinário, pois para se estabelecer a desigualdade não precisamos de quase nenhuma informação sobre a variável aleatória.

Demonstração do teorema 2.1.

$$\begin{aligned} V[Y] &= E[(Y - E[Y])^2] = \int_{\Omega} (Y(\omega) - E[Y])^2 d\text{Prob}(\omega) \\ &\geq \int_{\{\omega \in \Omega : |Y(\omega) - E[Y]| \geq \epsilon\}} (Y(\omega) - E[Y])^2 d\text{Prob}(\omega) \\ &\geq \int_{\{\omega \in \Omega : |Y(\omega) - E[Y]| \geq \epsilon\}} \epsilon^2 d\text{Prob}(\omega) \\ &= \epsilon^2 \text{Prob}(|Y - E[Y]| \geq \epsilon). \end{aligned}$$

Daí chegamos à tese. □

Demonstração do teorema 2.1. Trata-se de uma aplicação direta da Desigualdade de Chebychov. De fato,

$$V[S_m] = p \quad \text{e} \quad V[S_m] = \frac{\sum_{n=1}^m V[X_n]}{m^2} = \frac{V[X_1]}{m}.$$

Então

$$P(|S_m - p| > \epsilon) \leq \frac{V[X_1]}{m\epsilon^2},$$

e daí

$$\lim_{m \rightarrow +\infty} P(|S_m - p| > \epsilon) = 0. \tag{2}$$

□

2.2 A Lei Forte dos Grandes Números

A Lei Fraca dos Grandes Números é aparentemente semelhante à sua variante fraca.

Para entendermos a diferença, começamos observando que em (2), o subconjunto $\{\omega \in \Omega : |S_m(\omega) - p| > \epsilon\} \subset \Omega$ depende de m , isto é, pode ser que para cada m tenhamos um subconjunto diferente.

Na Lei Forte dos Grandes Números, existe um evento $C \in \mathcal{F}$ com probabilidade nula, $P(C) = 0$ tal que

$$S_m(\omega) \rightarrow p$$

para todo $\omega \in \Omega \setminus C$. Em palavras: a convergência acontece para todo elemento de Ω , exceto para ω pertencente a um subconjunto desprezível de probabilidade nula. Quando isso acontece, dizemos que a convergência ocorre *quase certamente*.

Note a diferença entre as leis forte e fraca. Enquanto na lei fraca os subconjuntos de Ω ficam mudando, na lei forte $\omega \in \Omega$ é fixo. Nesse sentido, a lei forte implica a fraca, mas não vice-versa.

Formalizamos a Lei Forte dos Grandes Números com o enunciado abaixo.

Teorema 2.3 (Lei Forte dos Grandes Números (Колмогоров)). *Suponha um experimento probabilístico com espaço amostral Ω , espaço de eventos \mathcal{F} e uma função probabilidade $P : \mathcal{F} \rightarrow [0, 1]$. Seja uma sequência de variáveis aleatórias $(W_n)_{\mathbb{N}}$ independentes e identicamente distribuídas associada ao experimento probabilístico, tal que $W_n = 1$ com probabilidade p e $W_n = 0$ com probabilidade $(1 - p)$. Então,*

$$S_m = \frac{\sum_{n=1}^m W_n}{m} \rightarrow p$$

quase certamente.

3 Independência das variáveis aleatórias envolvidas na representação binária de um número no intervalo $[0, 1[$

Na introdução definimos as variáveis aleatórias X_n , $n \in \mathbb{N}$ que fornecem os dígitos da representação binária de um número $\omega \in \Omega$, $\Omega = [0, 1[$. Lembramos que X_n pode assumir somente os valores $x_n = 0$ ou $x_n = 1$ e que

$$\omega = \sum_{j=1}^{+\infty} \frac{x_j}{2^j}.$$

Os valores de X_n estão associados à posição de ω em $\Omega = [0, 1[$. Como dissemos na introdução, quando ω tem duas representações binárias, consideramos apenas aquela que termina, isto é, aquela que tem um número finito de dígitos. Como vimos, esse descarte de representações ocorre em um evento de probabilidade desprezível, precisamente nula.

Se $x_1 = 0$ então $\omega \in [0, \frac{1}{2}[$ e se $x_1 = 1$ então $\omega \in [\frac{1}{2}, 1[$. Se $x_2 = 0$ então $\omega \in [0, \frac{1}{2^1} - \frac{1}{2^2}[\cup [\frac{1}{2^1}, 1 - \frac{1}{2^2}[$ e se $x_2 = 1$ então $\omega \in [0 + \frac{1}{2^2}, \frac{1}{2^1}[\cup [\frac{1}{2^1} + \frac{1}{2^2}, 1[$. Para x_3 , temos

$$x_1 = \begin{cases} 0, & \text{se } \omega \in [0, \frac{1}{2}[\\ 1, & \text{se } \omega \in [\frac{1}{2}, 1[\end{cases}$$

$$x_2 = \begin{cases} 0, & \text{se } \omega \in [0, \frac{1}{2^1} - \frac{1}{2^2} [\cup [\frac{1}{2^1}, 1 - \frac{1}{2^2} [\\ 1, & \text{se } \omega \in [0 + \frac{1}{2^2}, \frac{1}{2^1} [\cup [\frac{1}{2^1} + \frac{1}{2^2}, 1[\end{cases}$$

$$x_3 = \begin{cases} 0, & \text{se } \omega \in [0, \frac{1}{2^1} - \frac{1}{2^2} - \frac{1}{2^3} [\cup [\frac{1}{2^1}, 1 - \frac{1}{2^2} - \frac{1}{2^3} [& (3) \\ & \cup [0 + \frac{1}{2^2}, \frac{1}{2^1} - \frac{1}{2^3} [\cup [\frac{1}{2^1} + \frac{1}{2^2}, 1 - \frac{1}{2^3} [& (4) \\ 1, & \text{se } \omega \in [0 + \frac{1}{2^3}, \frac{1}{2^1} - \frac{1}{2^2} [\cup [\frac{1}{2^1} + \frac{1}{2^3}, 1 - \frac{1}{2^2} [& (5) \\ & [0 + \frac{1}{2^2} + \frac{1}{2^3}, \frac{1}{2^1} [\cup [\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3}, 1[& (6) \end{cases}$$

e assim sucessivamente. Podemos ilustrar a situação com uma figura. Na figura 1, os primeiros 4 dígitos dos números $a, b, c \in [0, 1[$ são dados por

$$\begin{cases} a \rightarrow x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 0. \\ b \rightarrow x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1. \\ c \rightarrow x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 1. \end{cases}$$

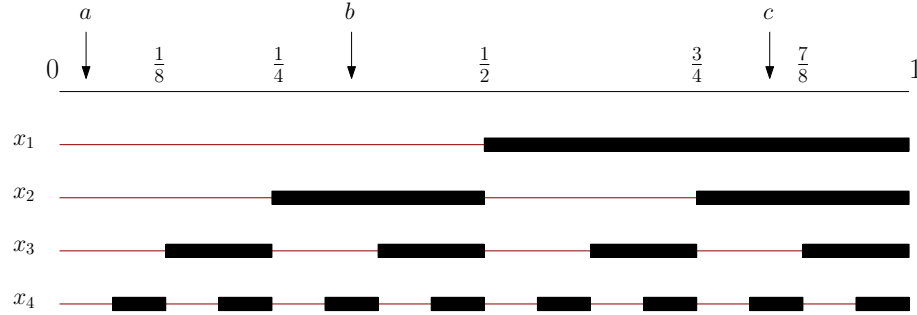


Figura 1: Dígitos binários em função da posição do número no intervalo Ω .

Como todo número no intervalo $\Omega = [0, 1[$ tem no máximo duas representações binárias, uma que termina (tem um número finito de dígitos) e outra que não termina, e convencionamos descartar a representação que não termina, temos uma relação biunívoca entre Ω e a representação binária dos elementos desse conjunto. Isto é, a função $B : \Omega \rightarrow \{(x_n)_{n \in \mathbb{N}} : x_n \in \{0, 1\}\}$ definida por

$$B(\omega) = (x_1, x_2, \dots)$$

é bijetora. Observe que B fornece os dígitos após a vírgula na representação binária $\omega = 0.x_1x_2x_3 \dots$

Suponha que a variável aleatória X de um experimento probabilístico seja uniformemente distribuída no intervalo $\Omega = [0, 1[$. Recordamos que a variável aleatória X_n retorna o n -ésimo dígito de X .

Imediatamente, temos

$$P[X_j = 0] = P[X_j = 1] = \frac{1}{2}, \forall j \in \mathbb{N},$$

pois X é uniformemente distribuída em $[0, 1[$, isto é as probabilidades $P[X_j = 0]$ e $P[X_j = 1]$ são dadas pelos comprimentos correspondentes na figura 1. Isso responde à questão Q1 posta na introdução.

Agora desejamos saber se X_i e X_j , para $i \neq j$ são independentes. A resposta é sim, pelo argumento a seguir.

X_i e X_j , $i \neq j$, serão independentes se e somente se

$$\begin{aligned} P[X_i = 0, X_j = 0] &= P[X_i = 0]P[X_j = 0] \\ P[X_i = 1, X_j = 0] &= P[X_i = 1]P[X_j = 0] \\ P[X_i = 0, X_j = 1] &= P[X_i = 0]P[X_j = 1] \\ P[X_i = 1, X_j = 1] &= P[X_i = 1]P[X_j = 1]. \end{aligned}$$

Como todas as verificações acima são todas análogas, verificaremos apenas a última. Poderíamos fazer a verificação algebricamente, mas é mais fácil observar a figura 1 e perceber que $P[X_i = 1, X_j = 1]$ vale sempre $\frac{1}{4}$, pois devemos tomar o equivalente a metade do comprimento do intervalo $[0, 1[$ e depois metade dele. E como $P[X_i = 1] = P[X_j = 1] = \frac{1}{2}$, para todos os $i, j \in \mathbb{N}$, imediatamente vemos que X_i e X_j satisfazem (7). As demais verificações são praticamente idênticas. A conclusão é que X_i e X_j , $i \neq j$, são de fato independentes.

A independência de X_i e X_j , $i \neq j$, diz em particular, que não há relação entre os dígitos sucessivos da representação binária de um número $\omega \in \Omega$.

Com isso, respondemos à pergunta Q2.

4 Distribuição de zeros e uns na expansão binomial de um número no intervalo $[0, 1[$

Finalmente podemos responder à questão Q3, com a ajuda da Lei Forte dos Grandes Números, sabendo que X_i e X_j , $i \neq j$, são independentes e que $P[X_j = 0] = P[X_j = 1] = \frac{1}{2}$, $\forall j \in \mathbb{N}$. Afinal, para que números $\omega \in \Omega$ há uma quantidade maior de uns ou de zeros na sua representação binária?

A razão

$$\frac{\sum_{n=1}^m X_n(\omega)}{m}$$

fornece a frequência relativa de uns na representação binária de $\omega \in \Omega$. Sabemos pela Lei Forte dos Grandes Números, teorema 2.3, que para ω em um evento com probabilidade nula, temos

$$\frac{\sum_{n=1}^m X_n(\omega)}{m} \rightarrow p = \frac{1}{2},$$

pois para todos $j \in \mathbb{N}$, $p = \frac{1}{2}$ é a probabilidade de ocorrer $X_j = 1$.

A conclusão é que para um evento de probabilidade zero, ω sorteado de acordo com a distribuição uniforme, tem representação binária com dígitos 1 e 0 com igual frequência. A nosso ver, esse é um resultado inusitado e interessante.

Referências

- [1] Alexandre Kawano. *Enumerabilidade e uma Ligação entre as Distribuições Uniforme e Geométrica*. (Artigo presente nesta coletânea)

Capítulo 6

Agradecimentos

O autor agradece toda a equipe da disciplina de Probabilidades que é oferecida aos alunos do terceiro semestre dos cursos de Engenharia da Escola Politécnica da Universidade de São Paulo e aos próprios alunos, que foram a verdadeira motivação para a produção dos artigos deste livro.