UnB | Departamento de Matemática

Desigualdades de Markov e Chebyshev e o lema de Borel-Cantelli

L. Cioletti

Outubro de 2026

Resumo

Nestas notas, estudamos as desigualdades de Markov, Chebyshev, Jensen e Lyapunov e o lema de Borel-Cantelli. Começamos com as desigualdades de Markov e Chebyshev, que fornecem cotas para probabilidades a partir da esperança e da variância. Introduzimos também a mediana, comparamos seu valor com a média de uma variável não negativa e aplicamos essas desigualdades ao problema dos aniversários, identificando a escala $\sqrt{N}$ para a ocorrência de coincidências quando há N dias possíveis. Em seguida, estudamos funções convexas, apresentamos algumas de suas caracterizações e demonstramos a desigualdade de Jensen a partir da existência de retas suporte. Caracterizamos o caso de igualdade para funções estritamente convexas e obtemos, como aplicação, a desigualdade de Lyapunov, que compara momentos absolutos de ordens distintas. Mostramos ainda que a finitude de um momento absoluto implica a finitude dos de ordens menores. Em um jogo de capital multiplicativo, mostramos que a esperança do capital pode permanecer constante enquanto a probabilidade de terminar com ganho tende a zero. Por fim, introduzimos os limites superior e inferior de eventos e demonstramos as duas partes do lema de Borel-Cantelli, obtendo a lei 0-1 de Borel para eventos independentes. Aplicamos esses resultados ao comprimento da maior sequência de caras nos primeiros n lançamentos independentes de uma moeda honesta, mostrando que sua razão com $\log_2 n$ tende a 1 quase certamente. O texto inclui exemplos e uma lista de exercícios ao final. A exposição das desigualdades de Markov e Chebyshev e do material de convexidade segue de perto a Seção 7.5 do Capítulo 7, os problemas desse capítulo e a Seção 8.2 do Capítulo 8 da referência [1].

1. Introdução

Nas notas anteriores, estudamos a esperança, a variância e os momentos de uma variável aleatória, além de sua função geradora de momentos. Para calcular essas grandezas, partimos da distribuição da variável. Suponha agora que conhecemos apenas sua média e sua variância. O que podemos dizer sobre a probabilidade de a variável assumir valores em um dado intervalo? Em geral, essas informações não permitem calcular a probabilidade procurada, pois distribuições diferentes podem ter a mesma média e a mesma variância. Ainda assim, podemos usá-las para obter estimativas que valem para todas essas distribuições.

CC BY-NC-SA 4.0- Attribution-NonCommercial-ShareAlike 4.0 International

Para obter uma primeira estimativa, vamos considerar uma variável aleatória X não negativa e com esperança finita. Se X assume valores maiores ou iguais a $t > 0$ com probabilidade p, esses valores contribuem com pelo menos tp para a esperança. Como os demais valores são não negativos, devemos ter $tp \leqslant \mathbb{E}[X]$. Portanto, $\mathbb{P}(X \geqslant t) \leqslant \mathbb{E}[X]/t$. Esta é a desigualdade de Markov, que estudaremos na Seção 2. Note que o argumento usa apenas a não negatividade de X e o valor de sua esperança. Podemos, assim, aplicá-lo a outras variáveis não negativas, em particular ao quadrado do desvio de uma variável em relação à sua média.

O quadrado desse desvio tem esperança igual à variância. Aplicando Markov a $(X - \mathbb{E}[X])^2$, obteremos a desigualdade de Chebyshev, que permite estimar a probabilidade de X se afastar de sua média quando a variância é finita. Veremos uma aplicação dessas duas desigualdades no problema dos aniversários. Se há N dias possíveis para os aniversários, supostos independentes e uniformemente distribuídos, queremos saber quantas pessoas são necessárias para que uma coincidência se torne provável. Contando o número de pares que fazem aniversário no mesmo dia e calculando sua média e sua variância, mostraremos que a mudança de comportamento ocorre quando o número de pessoas é da ordem de $\sqrt{N}$. Esse exemplo mostra como a escolha da variável aleatória pode tornar úteis informações sobre apenas dois de seus momentos.

O uso dos momentos nessas estimativas também nos leva a perguntar como momentos de ordens distintas se relacionam. Por exemplo, sabemos que $(\mathbb{E}[X])^2 \leqslant \mathbb{E}[X^2]$, pois a variância é não negativa. Podemos interpretar essa desigualdade como uma comparação entre elevar a média ao quadrado e tomar a média dos quadrados. Na Seção 3, veremos que essa comparação se estende a uma classe de funções que inclui a função quadrática: as funções convexas. Para entender essa extensão, começaremos pela geometria de seus gráficos. Estudaremos as cordas e as inclinações das secantes até obter uma reta de suporte, isto é, uma reta que passa por um ponto do gráfico e fica abaixo dele. Essa reta nos permitirá demonstrar a desigualdade de Jensen, $g(\mathbb{E}[X]) \leqslant \mathbb{E}[g(X)]$, sob as hipóteses apropriadas.

Aplicando Jensen a funções potência, obteremos na Seção 4 a desigualdade de Lyapunov, que compara as quantidades $(\mathbb{E}[|X|^r])^{1/r}$ para diferentes valores de $r > 0$. Em particular, veremos que a finitude de um momento absoluto implica a finitude dos momentos absolutos de ordens menores. A comparação entre a função da média e a média da função também será útil para estudar um jogo em que o capital é multiplicado, a cada rodada, por um fator aleatório de média igual a 1. Embora a esperança do capital permaneça constante, mostraremos que a probabilidade de ele ficar abaixo de qualquer valor positivo fixado tende a 1 com o número de rodadas. O logaritmo transforma os produtos em somas, e a desigualdade de Jensen ajuda a entender por que a esperança do logaritmo do capital decresce mesmo quando a esperança do capital não muda.

Ao acompanhar o capital ao longo das rodadas, passamos a considerar uma sequência de variáveis aleatórias. Nesse tipo de situação, além de estimar uma probabilidade em cada etapa, podemos querer saber o que acontece ao longo de toda a sequência. Por exemplo, se A_n é o evento de ocorrer um desvio na etapa n, será que esses desvios deixam de ocorrer a partir de algum momento ou continuam ocorrendo infinitas vezes? Mesmo que $\mathbb{P}(A_n)$ tenda a zero, essa informação, sozinha, não decide a questão. Precisamos levar em conta o conjunto das probabilidades e a dependência entre os eventos.

É dessa questão que trata o lema de Borel-Cantelli, estudado na Seção 5. Se a série $\sum_n \mathbb{P}(A_n)$ converge, então, com probabilidade 1, somente um número finito dos eventos A_n ocorre. Se os eventos são independentes e a série diverge, então, com probabilidade 1, infi-

nitos deles ocorrem. Para formular e demonstrar essas afirmações, introduziremos os limites superior e inferior de eventos e recordaremos a continuidade das medidas de probabilidade. Ao final, aplicaremos essas ferramentas à maior sequência de caras em n lançamentos independentes de uma moeda honesta, mostrando que seu comprimento, dividido por $\log_2 n$, tende a 1 quase certamente. A demonstração partirá de estimativas para as probabilidades de sequências de caras muito longas ou muito curtas. Comecemos, portanto, pelas ferramentas que permitem obter estimativas de probabilidades a partir de esperanças.

2. As Desigualdades de Markov e Chebyshev

2.1. A Desigualdade de Markov

A pergunta que orienta esta seção é a seguinte: se sabemos que $\mathbb{E}[X]$ é finito e que X não assume valores negativos, o que podemos dizer sobre a probabilidade de X ser grande? A cauda direita de X é a função $t \longmapsto \mathbb{P}(X \geqslant t)$, definida para $t > 0$. A desigualdade de Markov diz que essa função não pode decair mais devagar do que $1/t$.

Teorema 1 (Desigualdade de Markov)**.** Seja X uma variável aleatória não negativa. Então, para cada $t > 0$,

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}[X]}{t}. \tag{1}$$

Prova. Fixe $t > 0$. Os eventos $\{X \geqslant t\}$ e $\{X < t\}$ são disjuntos e a união deles é Ω, de modo que formam uma partição de Ω. A esperança é aditiva nessa decomposição, e a forma de escrever isso é multiplicar X pela indicadora de cada uma das partes:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\Big[X\Big(\mathbb{1}_{\{X\geqslant t\}} + \mathbb{1}_{\{X<t\}}\Big)\Big] \\ &= \mathbb{E}\Big[X\,\mathbb{1}_{\{X\geqslant t\}}\Big] + \mathbb{E}\Big[X\,\mathbb{1}_{\{X<t\}}\Big].\end{aligned}$$

A segunda parcela é não negativa, porque $X \geqslant 0$. Descartá-la só pode diminuir o lado direito, logo

$$\mathbb{E}[X] \geqslant \mathbb{E}\Big[X\,\mathbb{1}_{\{X\geqslant t\}}\Big].$$

No evento $\{X \geqslant t\}$, a variável X é limitada inferiormente por t. Substituir X por t nessa parcela só pode diminuir a esperança, donde

$$\mathbb{E}\Big[X\,\mathbb{1}_{\{X\geqslant t\}}\Big] \geqslant \mathbb{E}\Big[t\,\mathbb{1}_{\{X\geqslant t\}}\Big] = t\,\mathbb{P}(X \geqslant t).$$

Reunindo as duas desigualdades, obtemos $\mathbb{E}[X] \geqslant t\,\mathbb{P}(X \geqslant t)$. Como $t > 0$, podemos dividir ambos lados da desigualdade por t e concluir a validade de (1). ■

A cota é ótima no sentido de que não se pode melhorar a constante. Se X assume o valor t com probabilidade p e o valor 0 com probabilidade $1-p$, então $\mathbb{E}[X] = pt$ e $\mathbb{P}(X \geqslant t) = p$, de modo que (1) vale com igualdade. O que a desigualdade não controla é a forma da cauda: duas variáveis com a mesma média podem ter caudas muito diferentes, e Markov dá a mesma cota para ambas.

A prova não usou praticamente nenhuma propriedade específica da distribuição da variável aleatória X, apenas a positividade e a linearidade da esperança. Por isso ela se estende

sem mudança a qualquer potência ou função crescente de X. Se $g\colon [0,\infty) \to [0,\infty)$ é não decrescente e $g(t) > 0$, então o evento $\{X \geqslant t\}$ está contido em $\{g(X) \geqslant g(t)\}$, e a desigualdade de Markov aplicada a $g(X)$ dá

$$\mathbb{P}(X \geqslant t) \leqslant \mathbb{P}\Big(g(X) \geqslant g(t)\Big) \leqslant \frac{\mathbb{E}[g(X)]}{g(t)}. \tag{2}$$

A desigualdade de Chebyshev é o caso $g(x) = x^2$, e é essa escolha que introduz a variância.

2.2. A Desigualdade de Chebyshev

Se X tem média μ e variância finita, então a variável aleatória $(X-\mu)^2$ é não negativa e sua esperança é $\mathrm{Var}(X)$. Aplicando a desigualdade de Markov a essa variável, obtemos uma cota para a probabilidade de X se afastar de μ.

Teorema 2 (Desigualdade de Chebyshev)**.** Seja X uma variável aleatória com $\mathbb{E}[X^2] < \infty$. Então, para cada $\varepsilon > 0$,

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geqslant \varepsilon\Big) \leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2}. \tag{3}$$

Prova. Seja $\mu = \mathbb{E}[X]$. A função $u \longmapsto u^2$ é crescente em $[0,\infty)$ e ambos os membros na desigualdade $|X-\mu| \geqslant \varepsilon$ são não negativos, logo

$$\{|X-\mu| \geqslant \varepsilon\} = \{(X-\mu)^2 \geqslant \varepsilon^2\}.$$

Aplicando o Teorema 1 à variável não negativa $(X-\mu)^2$, com $t = \varepsilon^2$, obtemos

$$\mathbb{P}\Big((X-\mu)^2 \geqslant \varepsilon^2\Big) \leqslant \frac{\mathbb{E}\Big[(X-\mu)^2\Big]}{\varepsilon^2} = \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Isso prova (3). ∎

Equivalentemente, tomando o complementar,

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| < \varepsilon\Big) \geqslant 1 - \frac{\mathrm{Var}(X)}{\varepsilon^2}. \tag{4}$$

A forma (4) é a que costuma ser usada: ela diz que, com probabilidade alta, X fica em um intervalo centrado na média, e que o comprimento desse intervalo precisa crescer como o desvio padrão para que a probabilidade se aproxime de 1.

A mesma prova, com $|X-\mu|^p$ no lugar de $(X-\mu)^2$, dá

$$\mathbb{P}\Big(|X-\mu| \geqslant \varepsilon\Big) \leqslant \frac{\mathbb{E}\Big[|X-\mu|^p\Big]}{\varepsilon^p},$$

para todo $p > 0$ tal que o momento exista. Quanto maior o momento disponível, mais rápido é o decaimento na cota superior. A Seção 4 mostra que os momentos de ordens maiores são os mais exigentes, e que a existência deles é uma condição mais forte.

Observação 3. A desigualdade de Chebyshev é geral, e é justamente por isso que ela não é ótima em exemplos concretos. Para $X \sim \mathrm{Bin}(n,p)$, por exemplo, a cota para $\mathbb{P}(|X-np| \geqslant \varepsilon)$ é da ordem de $1/\varepsilon^2$, enquanto a distribuição real decai de forma gaussiana. O valor da desigualdade está em não exigir nada sobre a forma da distribuição.

2.3. A Mediana e uma Cota para a Média

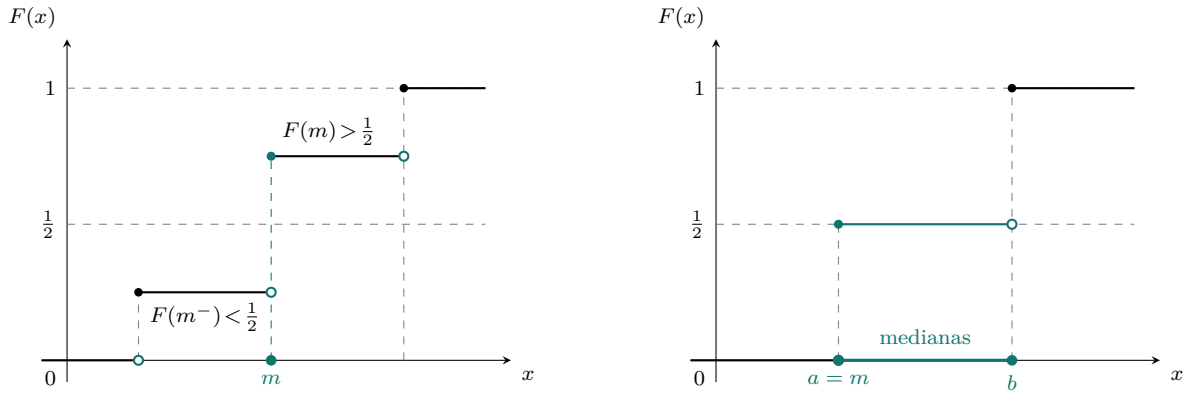
A desigualdade de Markov também compara a média com a mediana, e essa comparação é útil porque a mediana sempre existe, mesmo quando a média ou valor esperado não está bem-definido.

Definição 4 (Mediana)**.** Seja X uma variável aleatória. Um número real m é uma *mediana* de X se

$$\mathbb{P}(X < m) \leqslant \frac{1}{2} \leqslant \mathbb{P}(X \leqslant m).$$

Toda variável aleatória possui pelo menos uma mediana. De fato, seja F a função de distribuição de X e defina

$$m = \inf\{x \in \mathbb{R} : F(x) \geqslant 1/2\}.$$



(a) Mediana única

(b) Intervalo de medianas

Figura 1: A mediana separa a distribuição de modo que a probabilidade de ficar estritamente de cada lado dela é no máximo $1/2$. À esquerda, somente o número m satisfaz essa condição: $F(m^-) = \mathbb{P}(X < m) < 1/2 < F(m)$. À direita, X assume os valores a e b com probabilidade $1/2$ cada, e todo ponto de $[a, b]$ é mediana. O extremo b também está incluído, pois $\mathbb{P}(X < b) = 1/2$, embora $F(b) = 1$. A definição por ínfimo seleciona a menor mediana, que nesse caso é $m = a$.

Vamos justificar com cuidado por que esse ínfimo existe e por que ele fornece uma mediana. Para simplificar a notação, escrevemos

$$S = \{x \in \mathbb{R} : F(x) \geqslant 1/2\}.$$

O conjunto S reúne os pontos nos quais a probabilidade acumulada já alcançou pelo menos $1/2$. Como $F(x)$ tende a 1 quando x tende a $+\infty$, podemos escolher um número b suficientemente grande para que $F(b) > 1/2$. Assim, $b \in S$ e S não é vazio. Na outra direção, $F(x)$ tende a 0 quando x tende a $-\infty$. Podemos, portanto, escolher um número a tal que $F(a) < 1/2$. Pela monotonicidade de F,

$$x \leqslant a \quad \Longrightarrow \quad F(x) \leqslant F(a) < 1/2.$$

Nenhum ponto à esquerda de a pertence a S. Logo a é uma cota inferior de S, isto é, um número menor ou igual a todos os elementos de S. Como S é não vazio e possui uma cota inferior, a propriedade de completude dos números reais garante que $m = \inf S$ é um número real.

O ínfimo é a *maior das cotas inferiores* de S. Vamos traduzir essa descrição em duas informações que usaremos na prova. Primeiro, como m é uma cota inferior, todo $y \in S$ satisfaz $m \leqslant y$. Em particular, se $x < m$, então $x \notin S$, ou seja,

$$x < m \quad \Longrightarrow \quad F(x) < 1/2.$$

Segundo, como m é a maior cota inferior, nenhum número maior que m pode ser uma cota inferior de S. Assim, dado qualquer $\varepsilon > 0$, existe algum $y \in S$ tal que

$$m \leqslant y < m + \varepsilon.$$

De fato, se não existisse tal y, todos os elementos de S seriam maiores ou iguais a $m + \varepsilon$. Esse número seria uma cota inferior de S maior que m, contradizendo a escolha de m.

Essa segunda informação permite controlar F à direita de m. Como $y \in S$ e $y < m + \varepsilon$, a monotonicidade dá

$$F(m + \varepsilon) \geqslant F(y) \geqslant 1/2.$$

Portanto, à esquerda de m temos $F(x) < 1/2$ e à direita temos $F(x) \geqslant 1/2$. Falta verificar o que acontece no próprio ponto m. Essa verificação é necessária: o ínfimo de um conjunto nem sempre pertence ao conjunto. Por exemplo, o intervalo $(0, 1)$ tem ínfimo 0, mas não contém 0.

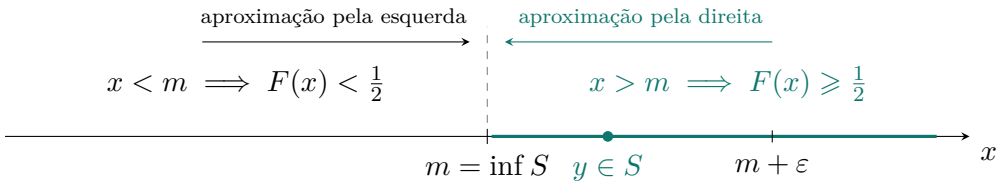


Figura 2: As duas propriedades do ínfimo localizam os pontos em relação a m. Nenhum ponto menor que m pertence a S; para cada $\varepsilon > 0$, há um ponto $y \in S$ entre m e $m + \varepsilon$ (possivelmente $y = m$). A monotonicidade de F garante então que todos os pontos maiores que m pertencem a S. A inclusão do próprio m será justificada usando a continuidade à direita.

Para decidir o que acontece em m, vamos aproximá-lo por pontos à sua direita, onde já sabemos que F é pelo menos $1/2$. Tomando $\varepsilon = 1/n$, obtemos

$$F(m + 1/n) \geqslant 1/2 \qquad \text{para todo } n \geqslant 1.$$

Quando n cresce, $m+1/n$ se aproxima de m pela direita. É justamente para essa aproximação que podemos usar a continuidade à direita da função de distribuição:

$$\mathbb{P}(X \leqslant m) = F(m) = \lim_{n \to \infty} F(m + 1/n) \geqslant 1/2.$$

Obtivemos, assim, uma das desigualdades exigidas na definição de mediana.

Para obter a outra, precisamos calcular $\mathbb{P}(X < m)$. Agora a aproximação deve ser feita pela esquerda, pois queremos excluir o valor m. Cada ponto $m - 1/n$ é menor que m e, portanto,

$$\mathbb{P}(X \leqslant m - 1/n) = F(m - 1/n) < 1/2.$$

Os eventos $\{X \leqslant m-1/n\}$ crescem com n e sua união é $\{X < m\}$. Para entender a igualdade da união, note que um valor estritamente menor que m tem uma distância positiva até m. Quando $1/n$ fica menor que essa distância, esse valor já pertence ao evento $\{X \leqslant m - 1/n\}$. Por outro lado, o valor m não pertence a nenhum desses eventos. Pela continuidade da probabilidade para sequências crescentes, que recordaremos na Seção 5.2,

$$\mathbb{P}(X < m) = \lim_{n\to\infty} \mathbb{P}(X \leqslant m - 1/n) = \lim_{n\to\infty} F(m - 1/n) \leqslant 1/2.$$

Note que a desigualdade no limite pode deixar de ser estrita: números menores que $1/2$ podem se aproximar de $1/2$. Reunindo as duas conclusões,

$$\mathbb{P}(X < m) \leqslant \frac{1}{2} \leqslant \mathbb{P}(X \leqslant m).$$

Assim, o número m escolhido é uma mediana de X. A aproximação pela direita garantiu que a probabilidade acumulada em m já alcançou $1/2$; a aproximação pela esquerda garantiu que, antes de m, ela não ultrapassa $1/2$.

Proposição 5. Seja X uma variável aleatória não negativa com $\mathbb{E}[X] < \infty$. Se m é uma mediana de X, então $m \leqslant 2\mathbb{E}[X]$.

Prova. Como m é mediana, $\mathbb{P}(X \geqslant m) \geqslant 1/2$. Pelo Teorema 1,

$$\frac{1}{2} \leqslant \mathbb{P}(X \geqslant m) \leqslant \frac{\mathbb{E}[X]}{m}.$$

Se $m > 0$, basta multiplicar a última desigualdade por $2m$. Se $m \leqslant 0$, a conclusão é imediata, porque $\mathbb{E}[X] \geqslant 0$. ■

A proposição diz que a mediana de uma variável não negativa não pode estar muito à direita da média. A recíproca não vale: a média pode estar muito à direita da mediana, e é isso que acontece com variáveis de cauda pesada. O Exercício 1 pede o exemplo em que a cota é atingida.

2.4. Aplicação: O Limiar do Problema dos Aniversários

Em uma sala com n pessoas, qual é a probabilidade de que duas façam aniversário no mesmo dia? A resposta exata depende de uma contagem, mas a ordem de grandeza do número de pessoas necessário para que a coincidência se torne provável pode ser obtida com as desigualdades de Markov e Chebyshev, sem calcular explicitamente a distribuição. Nosso objetivo será determinar essa escala: queremos entender como o número de pessoas precisa crescer quando o número N de dias possíveis aumenta. Esse é um exemplo de uma pergunta assintótica, na qual procuramos identificar o comportamento do problema para valores grandes dos parâmetros.

Suponha que os aniversários sejam independentes e uniformemente distribuídos em N dias, e que N seja grande. Vamos supor $n \geqslant 2$. Para detectar uma coincidência, basta que

algum par de pessoas faça aniversário no mesmo dia. Isso sugere contar os pares coincidentes. Numeramos as pessoas de 1 a n e, para cada par $\{i, j\}$ com $i < j$, definimos

$$I_{ij} = \mathbb{1}_{\{i \text{ e } j \text{ fazem aniversário no mesmo dia}\}}, \qquad X = \sum_{1 \leqslant i < j \leqslant n} I_{ij}.$$

Assim X é a variável aleatória que conta o número de pares de pessoas, na sala, que fazem aniversário no mesmo dia. Não estamos contando pessoas: se três delas fazem aniversário no mesmo dia, por exemplo, elas contribuem com três pares para X. O que importa é que existe tal par se, e somente se, $X \geqslant 1$. Logo o evento de interesse é $\{X \geqslant 1\}$. Para cada par de pessoas, qualquer que seja o dia do aniversário da primeira, a segunda tem probabilidade $1/N$ de fazer aniversário nesse mesmo dia, pela independência e pela distribuição uniforme dos aniversários. Assim, $\mathbb{E}[I_{ij}] = 1/N$ e, pela linearidade da esperança, somando sobre os $\binom{n}{2}$ pares, obtemos

$$\mathbb{E}[X] = \binom{n}{2}\frac{1}{N}.$$

Pela desigualdade de Markov,

$$\mathbb{P}(X \geqslant 1) \leqslant \frac{\mathbb{E}[X]}{1} = \binom{n}{2}\frac{1}{N} = \frac{n(n-1)}{2}\frac{1}{N} \approx \frac{n^2}{2N}. \tag{5}$$

O termo n^2/N já sugere a escala $\sqrt{N}$: ele é pequeno quando n é pequeno em relação a $\sqrt{N}$. Mais precisamente, se fazemos N crescer e escolhemos $n = n(N)$ de modo que $n/\sqrt{N} \to 0$, então

$$\mathbb{P}(X \geqslant 1) \leqslant \frac{n(n-1)}{2N} \leqslant \frac{1}{2}\left(\frac{n}{\sqrt{N}}\right)^2 \longrightarrow 0.$$

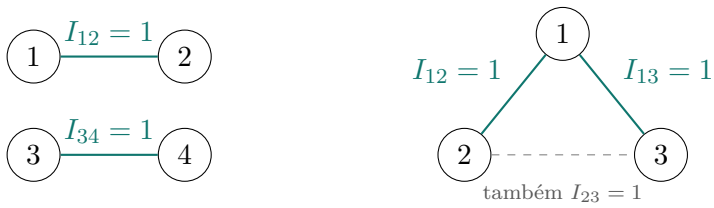
É esse o sentido da notação $n \ll \sqrt{N}$. Para concluir que $\sqrt{N}$ é a escala da mudança de comportamento, falta estudar o outro lado: o que acontece quando n é grande em relação a $\sqrt{N}$?

Nesse caso, esperamos muitas coincidências, pois $\mathbb{E}[X]$ é grande. Uma esperança grande, sozinha, não garante que X seja positivo com alta probabilidade: ela poderia ser sustentada por valores muito grandes e muito raros de X. Para afastar essa possibilidade, vamos calcular a variância e usar Chebyshev para controlar os desvios de X em relação à sua média.

Comecemos pela variância de cada indicadora. Como I_{ij} assume apenas os valores 0 e 1, temos $I_{ij}^2 = I_{ij}$ e, portanto,

$$\operatorname{Var}(I_{ij}) = \mathbb{E}[I_{ij}^2] - (\mathbb{E}[I_{ij}])^2 = \frac{1}{N} - \frac{1}{N^2} = \frac{1}{N}\left(1 - \frac{1}{N}\right).$$

Para obter a variância da soma, precisamos também das covariâncias entre indicadoras distintas. Não podemos simplesmente supor que elas sejam independentes: dois pares podem compartilhar uma pessoa. Vamos verificar os dois casos, representados na Figura 3.



Pares sem pessoas em comum Pares com uma pessoa em comum

Figura 3: Cada círculo representa uma pessoa e cada segmento contínuo exige que seus extremos façam aniversário no mesmo dia. À esquerda, as duas coincidências envolvem quatro pessoas. À direita, exigir as duas coincidências faz com que as três pessoas tenham o mesmo aniversário; a terceira coincidência, tracejada, é uma consequência das outras duas.

Se os pares não têm pessoas em comum, como $\{1, 2\}$ e $\{3, 4\}$, as duas coincidências dependem de grupos distintos de aniversários independentes. Logo

$$\mathbb{P}(I_{12} = 1,\, I_{34} = 1) = \mathbb{P}(I_{12} = 1)\mathbb{P}(I_{34} = 1) = \frac{1}{N^2}.$$

Se os pares compartilham uma pessoa, como $\{1, 2\}$ e $\{1, 3\}$, as duas coincidências ocorrem exatamente quando as três pessoas fazem aniversário no mesmo dia. Para cada dia fixado, essa probabilidade é $(1/N)^3$, pela independência dos três aniversários. Como há N dias possíveis e esses casos são disjuntos,

$$\mathbb{P}(I_{12} = 1,\, I_{13} = 1) = N\left(\frac{1}{N}\right)^3 = \frac{1}{N^2} = \mathbb{P}(I_{12} = 1)\mathbb{P}(I_{13} = 1).$$

Assim, nos dois casos a probabilidade conjunta é o produto das probabilidades. Como as variáveis são indicadoras, isso mostra que são duas a duas independentes. Em particular, para quaisquer dois pares distintos, o produto das indicadoras tem esperança $1/N^2$ e sua covariância é $1/N^2 - (1/N)(1/N) = 0$. A fórmula da variância de uma soma fica então reduzida à soma das variâncias:

$$\operatorname{Var}(X) = \sum_{i<j} \operatorname{Var}(I_{ij}) = \binom{n}{2}\frac{1}{N}\Big(1 - \frac{1}{N}\Big).$$

Vale observar que a independência é apenas duas a duas. Por exemplo, $I_{12} = I_{13} = 1$ força $I_{23} = 1$. Isso não atrapalha o cálculo da variância, no qual aparecem somente covariâncias entre duas variáveis.

Agora queremos usar essa variância para limitar a probabilidade de não haver pares de pessoas fazendo aniversário no mesmo dia. Por questão de simplicidade denote por $\mu \equiv \mathbb{E}[X] = \binom{n}{2}/N > 0$. Se $X = 0$, então $|X - \mu| = \mu$: a ausência de coincidências corresponde a um desvio em relação à média de tamanho igual à própria média. Portanto, $\{X = 0\} \subseteq \{|X - \mu| \geqslant \mu\}$. Essa inclusão indica a escolha $\varepsilon = \mu$ na desigualdade de Chebyshev que vale a seguinte desigualdade

$$\mathbb{P}(X = 0) \leqslant \mathbb{P}\Big(|X - \mathbb{E}[X]| \geqslant \mathbb{E}[X]\Big) \leqslant \frac{\operatorname{Var}(X)}{(\mathbb{E}[X])^2}.$$

Substituindo a média e a variância que calculamos, a razão acima fica

$$\frac{\operatorname{Var}(X)}{(\mathbb{E}[X])^2} = \frac{\binom{n}{2}\frac{1}{N}\left(1-\frac{1}{N}\right)}{\left(\binom{n}{2}\frac{1}{N}\right)^2} = \frac{N\left(1-\frac{1}{N}\right)}{\binom{n}{2}} = \frac{N-1}{\binom{n}{2}}.$$

Tomando o complementar do evento $\{X = 0\}$, obtemos

$$\mathbb{P}(X \geqslant 1) \geqslant 1 - \frac{N-1}{\binom{n}{2}} \approx 1 - \frac{2N}{n^2}. \tag{6}$$

Se $n/\sqrt{N} \to \infty$, então $n \to \infty$ e

$$\frac{N-1}{\binom{n}{2}} = \frac{2(N-1)}{n(n-1)} \longrightarrow 0.$$

Logo $\mathbb{P}(X \geqslant 1) \to 1$. Escrevemos $n \gg \sqrt{N}$ para indicar esse regime.

Reunindo (5) e (6), vemos que o número de pessoas necessário para mudar o comportamento do problema é da ordem de $\sqrt{N}$. Abaixo dessa escala, no sentido $n/\sqrt{N} \to 0$, a probabilidade de coincidência tende a 0; acima dela, no sentido $n/\sqrt{N} \to \infty$, tende a 1. Esse é o ganho de determinar a escala: mesmo sem uma fórmula exata, sabemos em quais regimes o evento é raro e em quais é muito provável. Sabemos também onde procurar uma transição entre os dois comportamentos, isto é, quando n é comparável a $\sqrt{N}$. As estimativas ainda não descrevem precisamente essa transição, mas identificam a relação entre os parâmetros que governa o problema.

Para avaliar a precisão das cotas em um caso concreto, tome $N = 365$. Com $n = 23$, Markov fornece a cota superior $\binom{23}{2}/365 = 253/365 \approx 0{,}69$, enquanto a probabilidade exata é aproximadamente 0,51. Com $n = 41$, a cota inferior de (6) dá

$$\mathbb{P}(X \geqslant 1) \geqslant 1 - \frac{364}{820} \approx 0{,}556,$$

enquanto a probabilidade exata é aproximadamente 0,903. As cotas localizam a escala, mas podem ser pouco precisas para valores específicos de n e N.

Para comparar com as cotas, vamos obter a fórmula exata. Se os aniversários das primeiras k pessoas são distintos, eles ocupam k dos N dias. A probabilidade de a pessoa seguinte evitar esses dias é $(N-k)/N$. Para multiplicar essas probabilidades condicionais, usamos a *regra da multiplicação*,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A), \qquad \mathbb{P}(A) > 0.$$

Sua aplicação sucessiva fornece a seguinte igualdade

$$\mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) = \mathbb{P}(A_1) \prod_{j=2}^{n} \mathbb{P}\left(A_j \,\middle|\, \bigcap_{i=1}^{j-1} A_i\right),$$

desde que os eventos nos quais condicionamos tenham probabilidade positiva. Essa fórmula foi demonstrada na Proposição 7 da Subseção 2.3, "A Regra da Multiplicação", das notas *Probabilidade Condicional e Independência*.

Aqui, se D_k é o evento de que os aniversários das primeiras k pessoas são distintos, temos $D_{k+1} \subseteq D_k$. Como mencionado acima segue da regra da multiplicação que

$$\mathbb{P}(D_{k+1}) = \mathbb{P}(D_k)\mathbb{P}(D_{k+1} \mid D_k),$$

com $\mathbb{P}(D_1) = 1$ e $\mathbb{P}(D_{k+1} \mid D_k) = (N-k)/N$. Aplicando essa identidade sucessivamente, obtemos, para $n \leqslant N$,

$$\mathbb{P}(X = 0) = 1 \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} \cdots \frac{N-n+1}{N} = \prod_{k=0}^{n-1} \Big(1 - \frac{k}{N}\Big).$$

Se $n > N$, necessariamente há uma coincidência. Para $n \leqslant N$, a probabilidade procurada é 1 menos o produto acima.

O produto também permite enxergar a escala $\sqrt{N}$ por outro caminho. Quando n é pequeno em relação a N, cada k/N é pequeno. Usando $\log(1-u) \approx -u$, obtemos

$$\log \mathbb{P}(X = 0) = \sum_{k=0}^{n-1} \log\left(1 - \frac{k}{N}\right) \approx -\frac{1}{N} \sum_{k=0}^{n-1} k = -\frac{n(n-1)}{2N}.$$

Assim, na região da transição,

$$\mathbb{P}(X \geqslant 1) \approx 1 - \exp\left(-\frac{n(n-1)}{2N}\right).$$

Para essa probabilidade ficar próxima de $1/2$, devemos ter $\exp(-n(n-1)/(2N)) \approx 1/2$, ou seja, $n(n-1) \approx 2N \log 2$. Para N grande, isso corresponde a $n \approx \sqrt{2 \log 2}\,\sqrt{N}$. A contagem exata permite, portanto, refinar a escala com uma constante para o nível de probabilidade $1/2$. Markov e Chebyshev já haviam identificado a escala $\sqrt{N}$ usando apenas a média e a variância do número de pares coincidentes.

3. Convexidade e a Desigualdade de Jensen

Na seção anterior, usamos a média e a variância para estimar probabilidades. Vamos agora estudar uma relação entre os próprios momentos. Já sabemos que $(\mathbb{E}[X])^2 \leqslant \mathbb{E}[X^2]$, pois a variância é não negativa. Nessa desigualdade, comparamos duas operações: primeiro tomar a média e depois elevar ao quadrado, ou primeiro elevar ao quadrado e depois tomar a média. Que outras funções, no lugar do quadrado, permitem uma comparação semelhante?

Para responder, vamos olhar para o gráfico da função quadrática. Entre dois pontos quaisquer de seu gráfico, a corda que os une fica acima do gráfico. Essa propriedade define as funções convexas, como ilustra a Figura 5. Partindo dessa geometria, chegaremos à desigualdade de Jensen, que compara $g(\mathbb{E}[X])$ com $\mathbb{E}[g(X)]$ quando g é convexa. A comparação entre momentos de ordens distintas, dada pela desigualdade de Lyapunov, será uma de suas aplicações.

3.1. Funções Convexas e Algumas de Suas Caracterizações

Definição 6 (Função convexa)**.** Sejam $-\infty \leqslant a < b \leqslant \infty$. Uma função $g\colon (a,b) \to \mathbb{R}$ é *convexa* se

$$g\Big((1-t)u + tv\Big) \leqslant (1-t)g(u) + tg(v) \tag{7}$$

para todos $t \in [0,1]$ e $u, v \in (a,b)$. Se a desigualdade é estrita sempre que $u \neq v$ e $t \in (0,1)$, dizemos que g é *estritamente convexa*.

O argumento da função g em (7) é ponto $(1-t)u+tv$. Para verificar que o lado esquerdo de (7) está bem-definido, basta observar que para qualquer valor de $t \in [0,1]$, temos que este ponto pertence ao intervalo fechado $[u,v]$. Mais precisamente, quando o parâmetro t varia em $[0,1]$ este ponto percorre todo segmento de reta entre os pontos u e v, veja figura abaixo

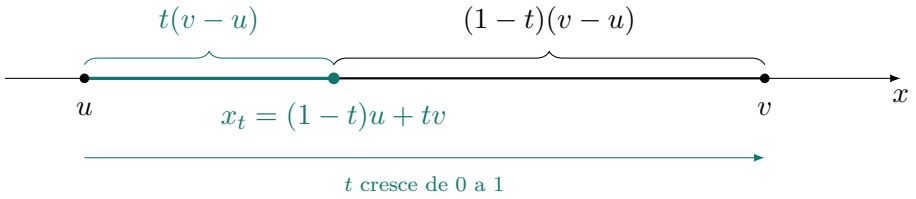


Figura 4: Para $u < v$, escrever $x_t = (1-t)u + tv = u + t(v-u)$ mostra que t é a fração do caminho de u até v já percorrida. Quando $t = 0$, temos $x_t = u$; quando $t = 1$, temos $x_t = v$. Para $0 < t < 1$, o ponto fica entre os extremos; em particular, $t = 1/2$ fornece o ponto médio.

O lado direito de (7) é a altura do ponto correspondente na corda que une $(u, g(u))$ e $(v, g(v))$. A definição diz, portanto, que a corda fica acima do gráfico. A Figura 5 mostra a situação e destaca a folga entre a corda e o gráfico.

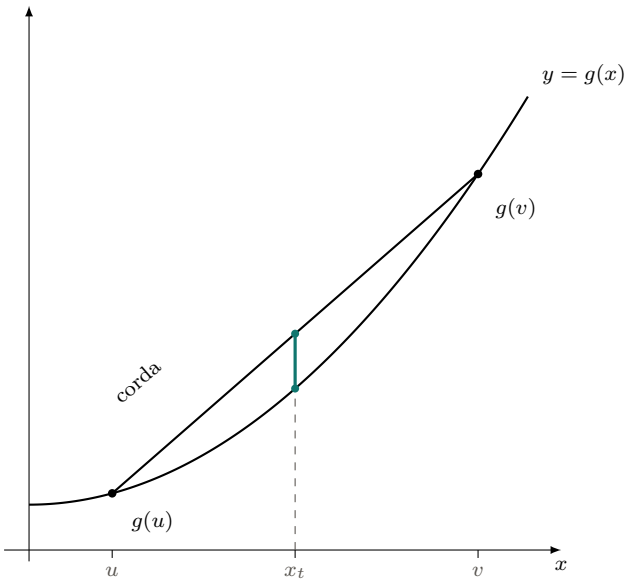


Figura 5: A função é convexa quando a corda que une dois pontos quaisquer do gráfico fica acima do gráfico. O segmento vertical destacado tem comprimento $(1-t)g(u)+tg(v)-g(x_t)$, onde o ponto $x_t = (1-t)u + tv$.

Observação 7. Também é comum escrever a combinação convexa de dois pontos na forma $\lambda u+(1-\lambda)v$, com $\lambda \in [0,1]$, o que equivale a tomar $\lambda = 1-t$. No texto usaremos $(1-t)u+tv$ para dois pontos e reservaremos a notação $\lambda_1, \ldots, \lambda_n$ para combinações convexas de n pontos.

A desigualdade (7) da Definição 6 se estende a qualquer número finito de pontos. Se $g\colon (a,b) \to \mathbb{R}$ é uma função convexa, se $x_1, \ldots, x_n \in (a,b)$ e se $\lambda_1, \ldots, \lambda_n$ são números reais não negativos com soma 1, então

$$g\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) \leqslant \sum_{i=1}^{n} \lambda_i g(x_i). \tag{8}$$

A definição de convexidade compara apenas dois pontos. Para usá-la em uma combinação de vários pontos, a ideia é agrupar alguns deles e tratar o grupo como um único ponto. Por exemplo, podemos escrever

$$\frac{1}{4}x_1 + \frac{1}{4}x_2 + \frac{1}{2}x_3 = \frac{1}{2}\underbrace{\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right)}_{y} + \frac{1}{2}x_3.$$

Primeiro formamos o ponto médio y de x_1 e x_2; depois tomamos o ponto médio de y e x_3. Os pesos $1/4$ aparecem porque o peso $1/2$ do grupo é dividido igualmente entre seus dois pontos. Assim, uma combinação de três pontos é construída por duas combinações de dois pontos, como mostra a Figura 6.

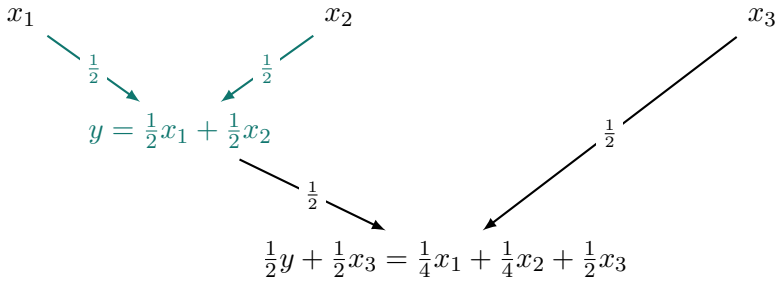


Figura 6: Uma combinação de três pontos construída em duas etapas. Os números nas setas são os pesos usados em cada etapa. Cada um dos pontos x_1 e x_2 recebe metade do peso de y, ficando com peso final $(1/2)(1/2) = 1/4$. O ponto x_3 recebe diretamente peso $1/2$. O diagrama representa o agrupamento dos pontos, sem pressupor uma ordem entre seus valores na reta.

A convexidade pode acompanhar essas duas etapas. Aplicando a definição primeiro ao par y, x_3 e depois ao par x_1, x_2, obtemos

$$\begin{aligned} g\left(\frac{1}{4}x_1 + \frac{1}{4}x_2 + \frac{1}{2}x_3\right) &\leqslant \frac{1}{2}g(y) + \frac{1}{2}g(x_3) \\ &\leqslant \frac{1}{2}\left(\frac{1}{2}g(x_1) + \frac{1}{2}g(x_2)\right) + \frac{1}{2}g(x_3) \\ &= \frac{1}{4}g(x_1) + \frac{1}{4}g(x_2) + \frac{1}{2}g(x_3). \end{aligned}$$

Para pesos quaisquer, usaremos a mesma ideia em uma prova por indução. O caso $n = 2$ é a própria definição de convexidade. Suponha que a desigualdade já seja válida para

combinações de $n-1$ pontos e consideremos uma combinação de n pontos. Vamos agrupar os primeiros $n-1$ pontos, deixando x_n separado. O peso total do grupo é

$$\lambda_1 + \cdots + \lambda_{n-1} = 1 - \lambda_n.$$

Se $\lambda_n = 1$, todos os outros pesos são zero e os dois lados de (8) são iguais a $g(x_n)$. Basta, portanto, tratar o caso $\lambda_n < 1$, no qual o peso do grupo é positivo.

Para representar esse grupo por uma combinação convexa, precisamos de pesos cuja soma seja 1. Os pesos originais somam $1-\lambda_n$; dividindo cada um deles por essa soma, obtemos

$$\alpha_i = \frac{\lambda_i}{1-\lambda_n}, \qquad 1 \leqslant i \leqslant n-1.$$

Esses novos pesos são não negativos e satisfazem

$$\sum_{i=1}^{n-1} \alpha_i = \frac{\lambda_1 + \cdots + \lambda_{n-1}}{1-\lambda_n} = 1.$$

Podemos, assim, definir o ponto que representa o grupo por

$$y = \sum_{i=1}^{n-1} \alpha_i x_i.$$

Como y é uma média ponderada dos primeiros $n-1$ pontos, ele fica entre o menor e o maior deles e, em particular, pertence a (a,b). Ao atribuir ao grupo seu peso total $1-\lambda_n$, recuperamos os pesos originais:

$$(1-\lambda_n)y + \lambda_n x_n = \sum_{i=1}^{n-1} (1-\lambda_n)\alpha_i x_i + \lambda_n x_n = \sum_{i=1}^{n} \lambda_i x_i.$$

Agora temos uma combinação dos dois pontos y e x_n, à qual podemos aplicar a definição de convexidade:

$$g\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leqslant (1-\lambda_n) g(y) + \lambda_n g(x_n).$$

Resta estimar $g(y)$. Como y é uma combinação convexa de $n-1$ pontos, a hipótese de indução fornece

$$g(y) \leqslant \sum_{i=1}^{n-1} \alpha_i g(x_i).$$

Substituindo essa estimativa na desigualdade anterior e usando $(1-\lambda_n)\alpha_i = \lambda_i$, concluímos que

$$g\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leqslant (1-\lambda_n) \sum_{i=1}^{n-1} \alpha_i g(x_i) + \lambda_n g(x_n) = \sum_{i=1}^{n} \lambda_i g(x_i).$$

Isso completa o passo de indução e demonstra (8).

Essa desigualdade já tem uma interpretação probabilística. Se X assume um número finito de valores $x_1, \ldots, x_n$ e $p_i = \mathbb{P}(X = x_i)$, então os p_i são não negativos e somam 1. Portanto, podemos usá-los como pesos em (8), obtendo

$$g\Big(\mathbb{E}[X]\Big) = g\left(\sum_{i=1}^{n} p_i x_i\right) \leqslant \sum_{i=1}^{n} p_i g(x_i) = \mathbb{E}[g(X)].$$

Assim, para variáveis com um número finito de valores, a comparação entre a função da média e a média da função decorre diretamente da convexidade. Essa é a desigualdade de Jensen nesse caso particular. Para tratar também variáveis com infinitos valores ou com densidade, desenvolveremos a geometria da convexidade até obter uma reta de suporte. Ela permitirá demonstrar Jensen sob as hipóteses de integrabilidade da Seção 3.5, sem precisar reduzir cada distribuição a uma soma finita. Começaremos estudando como variam as inclinações das secantes ao gráfico de uma função convexa.

3.2. As Inclinações das Retas Secantes

Até aqui, usamos a convexidade para comparar a altura do gráfico com a altura de uma corda. Vamos agora examinar a inclinação das retas que passam por dois pontos do gráfico, chamadas de *retas secantes*. Se as abscissas desses pontos são $u < v$, a inclinação da reta secante associada a estes pontos é dada por

$$\frac{g(v) - g(u)}{v - u}.$$

O numerador mede a variação na altura e o denominador, o deslocamento horizontal. Essa razão é também a taxa média de variação de g entre u e v.

Para entender o que a convexidade diz sobre essa taxa, vamos considerar um ponto intermediário w, com $u < w < v$. Como $(w, g(w))$ fica abaixo da corda entre $(u, g(u))$ e $(v, g(v))$, a secante de u a w tem inclinação menor ou igual à de u a v. Pelo mesmo motivo, a secante de w a v tem inclinação maior ou igual à de u a v. O lema a seguir traduz essa observação em desigualdades, ilustradas na Figura 7.

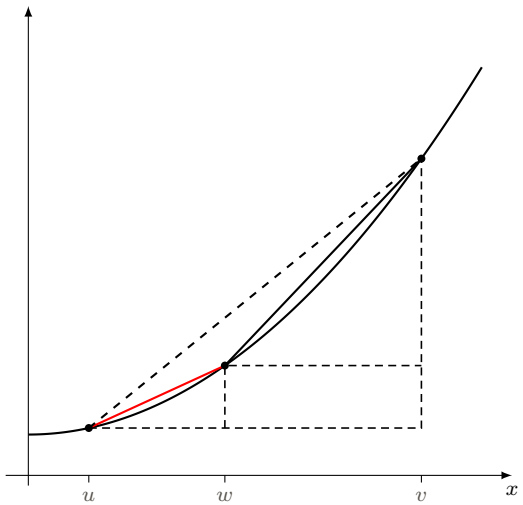


Figura 7: Para três pontos $u < w < v$, a inclinação s_1 da reta secante sobre $[u, w]$ não excede a inclinação s_{uv} da reta secante sobre $[u, v]$, que por sua vez não excede a inclinação s_2 da reta secante ao longo do subintervalo $[w, v]$.

Essa ordem entre as inclinações será útil para construir uma reta que passe por $(w, g(w))$ e fique abaixo de todo o gráfico: sua inclinação deverá ser pelo menos tão grande quanto as das secantes à esquerda de w e não maior que as das secantes à direita.

Lema 8 (Monotonicidade das secantes)**.** Seja $g\colon (a,b) \to \mathbb{R}$ convexa e sejam $u < w < v$ em (a,b). Então

$$\frac{g(w)-g(u)}{w-u} \leqslant \frac{g(v)-g(u)}{v-u} \leqslant \frac{g(v)-g(w)}{v-w}. \tag{9}$$

Prova. Primeiro, vamos representar o ponto w como uma combinação convexa de u e v, isto é, $w = (1-t)u + tv$, onde

$$t = \frac{w-u}{v-u} \in (0,1).$$

Pela convexidade,

$$g(w) \leqslant (1-t)g(u) + tg(v). \tag{10}$$

Subtraindo $g(u)$ dos dois lados de (10), obtemos

$$g(w) - g(u) \leqslant t\Big(g(v) - g(u)\Big),$$

e, dividindo por $w-u>0$ e usando $t = (w-u)/(v-u)$, segue que

$$\frac{g(w)-g(u)}{w-u} \leqslant \frac{g(v)-g(u)}{v-u}.$$

Isso prova a primeira desigualdade.

Para a segunda, queremos comparar as variações de g entre u e v e entre w e v. Por isso, subtraímos $g(v)$ dos dois lados de (10), obtendo

$$g(w) - g(v) \leqslant (1-t)\Big(g(u) - g(v)\Big).$$

Multiplicando por (-1), ambos os lados da desigualdade acima ficamos com

$$g(v) - g(w) \geqslant (1-t)\Big(g(v) - g(u)\Big) = \frac{v-w}{v-u}\Big(g(v) - g(u)\Big),$$

onde na última igualdade, usamos

$$1 - t = 1 - \frac{w-u}{v-u} = \frac{v-w}{v-u}.$$

Assim, ao dividir por $v-w>0$, o fator $v-w$ do lado direito se cancela, e obtemos a segunda desigualdade:

$$\frac{g(v)-g(u)}{v-u} \leqslant \frac{g(v)-g(w)}{v-w}.$$

■

Para $w \in (a,b)$ fixado vamos ver como variam as inclinações das secantes que têm uma extremidade no ponto $(w, g(w))$. À esquerda de w, isto é, ao longo do intervalo (a,w) a função

$$u \longmapsto \frac{g(w)-g(u)}{w-u}$$

é não-decrescente. De fato, se $a < u_1 < u_2 < w$, a segunda desigualdade de (9), aplicada ao trio $u_1 < u_2 < w$, dá

$$\frac{g(w)-g(u_1)}{w-u_1} \leqslant \frac{g(w)-g(u_2)}{w-u_2}.$$

Similarmente, à direita de w, isto é, no intervalo (w, b) a função

$$v \longmapsto \frac{g(v) - g(w)}{v - w}$$

é não-decrescente. De fato, se $w < v_1 < v_2 < b$, a primeira desigualdade de (9), aplicada ao trio $w < v_1 < v_2$, fornece

$$\frac{g(v_1) - g(w)}{v_1 - w} \leqslant \frac{g(v_2) - g(w)}{v_2 - w}.$$

Agora precisamos verificar que essas inclinações não divergem quando as variáveis se aproximam de w. Para isso, usamos outra consequência de (9): toda secante à esquerda tem inclinação menor ou igual à de qualquer secante à direita, isto é,

$$\frac{g(w) - g(u)}{w - u} \leqslant \frac{g(v) - g(w)}{v - w} \qquad (a < u < w < v < b). \tag{11}$$

Fixando $v = v_0 \in (w, b)$, obtemos uma cota superior finita para todas as inclinações à esquerda. Fixando $u = u_0 \in (a, w)$, obtemos uma cota inferior finita para todas as inclinações à direita. Portanto, os números

$$L_w = \sup_{u \in (a,w)} \frac{g(w) - g(u)}{w - u}, \qquad R_w = \inf_{v \in (w,b)} \frac{g(v) - g(w)}{v - w}$$

são finitos. Pela monotonicidade que acabamos de provar, eles são também os limites laterais

$$L_w = \lim_{u \to w^-} \frac{g(w) - g(u)}{w - u}, \qquad R_w = \lim_{v \to w^+} \frac{g(v) - g(w)}{v - w}.$$

À esquerda, aproximar u de w faz as inclinações crescerem até seu supremo; à direita, aproximar v de w faz as inclinações decrescerem até seu ínfimo.

Além disso, tomando primeiro o supremo com $u \in (a, w)$ em (11), obtemos

$$L_w \equiv \sup_{u \in (a,w)} \frac{g(w) - g(u)}{w - u} \leqslant \sup_{u \in (a,w)} \frac{g(v) - g(w)}{v - w} = \frac{g(v) - g(w)}{v - w},$$

para todo $v \in (w, b)$. Tomando agora, na desigualdade acima, o ínfimo com $v \in (w, b)$ concluímos que

$$L_w = \inf_{v \in (w,b)} L_w \leqslant \inf_{v \in (w,b)} \frac{g(v) - g(w)}{v - w} \equiv R_w.$$

Assim, existe pelo menos um número α que é maior ou igual a todas as inclinações à esquerda e menor ou igual a todas as inclinações à direita. Na próxima subseção, veremos que os valores $\alpha \in [L_w, R_w]$ são precisamente as inclinações das retas que passam por $(w, g(w))$ e ficam abaixo de todo o gráfico de g.

3.3. A Reta Suporte

Seja $g\colon (a, b) \to \mathbb{R}$ uma função convexa. Uma *reta de suporte* ao seu gráfico em $w \in (a, b)$ é uma reta que passa por $(w, g(w))$ e cuja altura, em cada abscissa do domínio, é menor ou igual à altura do gráfico. O próximo teorema mostra que a convexidade garante a existência

de pelo menos uma reta assim em cada ponto do gráfico. Se além de ser uma função convexa g é também diferenciável em w, então existe uma única reta suporte nesse ponto que é a reta tangente ao gráfico neste ponto. É importante observar que convexidade garante a existência de uma reta suporte mesmo quando a derivada não existe. No caso de não diferenciabilidade podemos ter a existência de várias retas suporte com inclinações distintas.

Teorema 9 (Reta suporte)**.** Seja $g\colon (a,b) \to \mathbb{R}$ convexa e seja $w \in (a,b)$. Então existe $\alpha \in \mathbb{R}$ tal que

$$g(w) + \alpha(x-w) \leqslant g(x) \qquad \text{para todo } x \in (a,b). \tag{12}$$

Prova. Sejam L_w e R_w como acima. Como $L_w \leqslant R_w$, escolha $\alpha \in [L_w, R_w]$. Se $x < w$, então

$$\frac{g(w)-g(x)}{w-x} \leqslant L_w \leqslant \alpha,$$

e multiplicando por $w-x>0$ obtemos $g(w)-g(x) \leqslant \alpha(w-x)$, que implica na desigualdade (12) para $x<w$. Se $x>w$, então segue da definição de R_w que

$$\alpha \leqslant R_w \leqslant \frac{g(x)-g(w)}{x-w},$$

e multiplicando por $x-w>0$ obtemos (12) para $x>w$. Para $x=w$ a desigualdade é uma igualdade. ■

A Figura 8 mostra dois casos. À esquerda, a função é diferenciável e a reta de suporte é a tangente, com $\alpha = g'(w)$. À direita, a função tem um bico e existem várias retas de suporte, uma para cada $\alpha \in [L_w, R_w]$. É por isso que a convexidade não exige derivada.

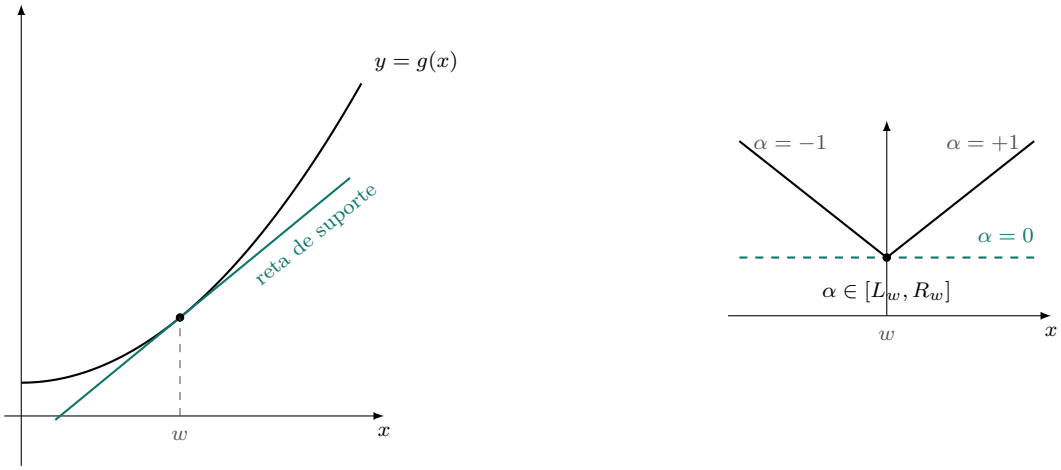


Figura 8: Toda função convexa admite em cada ponto uma reta suporte. À esquerda, o caso diferenciável, em que a reta suporte é a tangente, de inclinação $\alpha = g'(w)$. À direita, uma função com bico: qualquer reta de inclinação $\alpha \in [L_w, R_w]$ que passa pelo bico é uma reta suporte, e por isso a convexidade não exige derivada.

Observação 10. Se g é diferenciável em w, então $L_w = R_w = g'(w)$ e a reta de suporte é única. De fato, o quociente $(g(w)-g(u))/(w-u)$ tende a $g'(w)$ quando $u \to w^-$ e $(g(v)-g(w))/(v-w)$ tende a $g'(w)$ quando $v \to w^+$. Como as duas funções são monótonas e limitadas, os supremo e ínfimo coincidem com os limites laterais.

3.4. Caracterizações Equivalentes de Convexidade

Reunimos agora as três descrições equivalentes do conceito de convexidade e acrescentamos a caracterização via derivada segunda para o caso de funções convexas suaves que aparecem frequentemente nos exemplos.

Proposição 11. Seja $g\colon (a,b)\to\mathbb{R}$. As afirmações seguintes são equivalentes.

(i) g é convexa.

(ii) Para todos $u<w<v$ em (a,b), vale (9).

(iii) Para cada $w\in(a,b)$ existe $\alpha\in\mathbb{R}$ tal que vale (12).

Se g é duas vezes continuamente diferenciável, as três são ainda equivalentes a

(iv) $g''(x)\geqslant 0$ para todo $x\in(a,b)$.

Prova. A implicação (i)⇒(ii) é o Lema 8, e (ii)⇒(iii) é o Teorema 9, cuja prova usa apenas a existência de L_w e R_w e a desigualdade $L_w\leqslant R_w$, que por sua vez segue de (ii).

Para provar (iii)⇒(i), fixe $u,v\in(a,b)$ e $t\in[0,1]$ e considere $w=(1-t)u+tv$. A ideia é usar a reta suporte em w para comparar $g(w)$ com a combinação convexa $(1-t)g(u)+tg(v)$. Por (iii), existe $\alpha\in\mathbb{R}$ tal que

$$g(u)\geqslant g(w)+\alpha(u-w)\quad\text{e}\quad g(v)\geqslant g(w)+\alpha(v-w).$$

Multiplicando essas desigualdades pelos escalares não-negativos $1-t$ e t, respectivamente, e somando, obtemos

$$(1-t)g(u)+tg(v)\geqslant g(w)+\alpha\Big((1-t)(u-w)+t(v-w)\Big)=g(w),$$

pois $(1-t)(u-w)+t(v-w)=(1-t)u+tv-w=0$, pela escolha de w. Como $w=(1-t)u+tv$, essa é precisamente a desigualdade de convexidade (7).

Suponha agora que g seja duas vezes continuamente diferenciável. Como (i), (ii) e (iii) já são equivalentes, basta provar as implicações (iv)⇒(ii) e (i)⇒(iv).

Para provar (iv)⇒(ii), suponha que $g''\geqslant 0$. Então g' é não decrescente. Fixe $u<w<v$ em (a,b). Pelo teorema do valor médio, a inclinação da secante sobre $[u,w]$ é $g'(\xi_1)$ para algum $\xi_1\in(u,w)$ e a inclinação sobre $[w,v]$ é $g'(\xi_2)$ para algum $\xi_2\in(w,v)$. Como $\xi_1<\xi_2$ e g' é não decrescente, vale (9), o que prova (ii).

Para provar (i)⇒(iv), suponha que g seja convexa. Seja $x_0\in(a,b)$ e suponha, por absurdo, que $g''(x_0)<0$. Pela continuidade de g'', existe $h>0$ tal que $g''(\xi)<0$, para todo $\xi\in(x_0-h,x_0+h)$. A fórmula de Taylor com resto de Lagrange, aplicada a x_0+h e a x_0-h, dá

$$g(x_0+h)+g(x_0-h)-2g(x_0)=\frac{h^2}{2}\Big(g''(\xi_+)+g''(\xi_-)\Big)$$

para certos $\xi_+\in(x_0,x_0+h)$ e $\xi_-\in(x_0-h,x_0)$. O lado direito é negativo, logo

$$g(x_0)>\frac{g(x_0+h)+g(x_0-h)}{2},$$

o que contradiz a convexidade no par $x_0-h<x_0+h$ com $t=1/2$. ∎

Observação 12. A proposição acima fornece um critério rápido para verificar convexidade. Por exemplo, como consequência deste critério, verificamos imediatamente que são convexas as funções $g(x)=x^2$, $g(x)=e^{tx}$, para cada $t\in\mathbb{R}$ fixado, $g(x)=x^p$, para cada $p\geqslant 1$ fixado, $g(x)=1/x$ em $(0,\infty)$ e $g(x)=-\log x$ em $(0,\infty)$. A última tem derivada segunda $1/x^2>0$ e este é o exemplo que produz a desigualdade entre as médias geométrica e aritmética.

3.5. A Desigualdade de Jensen

Usando o resultado de existência da reta suporte, a prova da Desigualdade de Jensen é curta. Basta aplicar a desigualdade (12) ponto a ponto e tomar esperanças.

Teorema 13 (Desigualdade de Jensen)**.** Seja X uma variável aleatória assumindo valores em um intervalo (a, b), com $\mathbb{E}[X]$ finita, e seja $g\colon (a, b) \to \mathbb{R}$ convexa tal que $\mathbb{E}\Big[|g(X)|\Big] < \infty$. Então

$$\mathbb{E}\Big[g(X)\Big] \geqslant g\Big(\mathbb{E}[X]\Big). \tag{13}$$

Prova. Seja $\mu = \mathbb{E}[X]$. Como $a < X < b$ temos da monotonicidade da esperança que $a < \mathbb{E}[X] < b$. Logo $\mu \in (a, b)$. Aplicando o Teorema 9, com $w = \mu$, podemos afirmar que existe $\alpha \in \mathbb{R}$ tal que

$$g(x) \geqslant g(\mu) + \alpha(x - \mu) \qquad \text{para todo } x \in (a, b).$$

Substituindo x por X e tomando esperanças, obtemos

$$\mathbb{E}\Big[g(X)\Big] \geqslant g(\mu) + \alpha\Big(\mathbb{E}[X] - \mu\Big) = g(\mu) = g\Big(\mathbb{E}[X]\Big),$$

o que prova a validade da desigualdade (13). ■

A mesma demonstração se estende a qualquer intervalo $I \subseteq \mathbb{R}$ (aberto, fechado ou semi-aberto) no qual g seja convexa e contínua nos extremos pertencentes a I: se $\mu = \mathbb{E}[X]$ está no interior de I, a desigualdade (12) vale no interior e se estende aos extremos por continuidade; se μ coincide com um extremo de I, então $X = \mu$ quase certamente e (13) vale com igualdade.

A Figura 9 interpreta a desigualdade em termos de um centro de massa. Suponha que X assuma os valores u e v, onde $u < v$, com probabilidades p e $1 - p$, respectivamente, onde $0 < p < 1$. Atribuindo essas massas aos pontos $P_1 = (u, g(u))$ e $P_2 = (v, g(v))$, o centro de massa é o ponto em $\mathbb{R}^2$ dado pela expressão

$$A = pP_1 + (1 - p)P_2 = \Big(\mathbb{E}[X], \mathbb{E}[g(X)]\Big).$$

Esse ponto pertence à corda que une P_1 a P_2. Já o ponto $B = \Big(\mathbb{E}[X], g(\mathbb{E}[X])\Big)$ pertence ao gráfico de g e tem a mesma abscissa que A. Pela convexidade, B está abaixo de A ou coincide com ele. Assim, a desigualdade de Jensen afirma que a diferença

$$\Delta = \mathbb{E}[g(X)] - g(\mathbb{E}[X])$$

é não negativa. Na figura abaixo, Δ é exatamente o comprimento do segmento de reta vertical que une os ponto B e A.

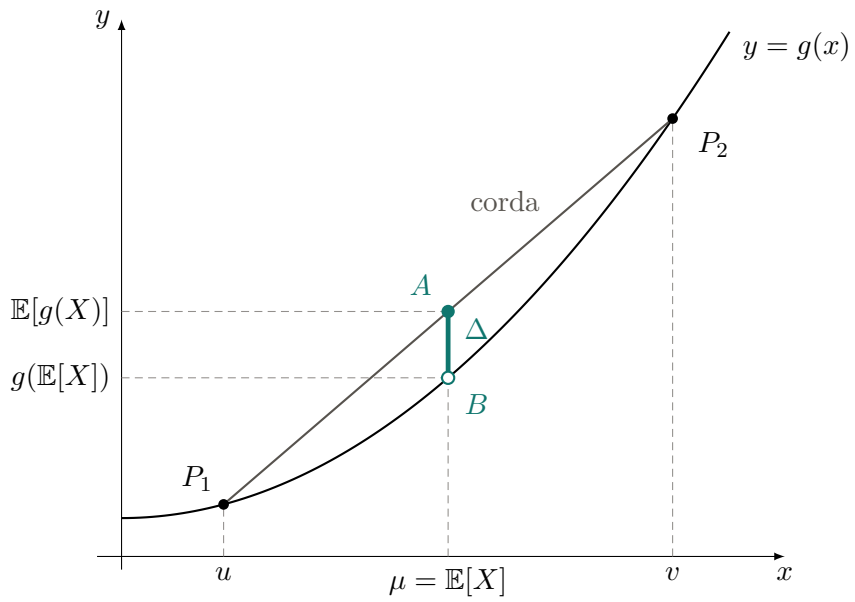


Figura 9: Interpretação geométrica de Jensen. O centro de massa A das massas p e $1-p$ está na corda; o ponto B está no gráfico e tem a mesma abscissa $\mu = \mathbb{E}[X]$. O segmento destacado tem comprimento $\Delta = \mathbb{E}[g(X)] - g(\mathbb{E}[X]) \geqslant 0$.

Observação 14. Há duas maneiras de combinar a função g com a esperança: podemos aplicar g aos valores de X e depois tomar a média, obtendo $\mathbb{E}[g(X)]$, ou tomar primeiro a média de X e depois aplicar g, obtendo $g(\mathbb{E}[X])$. Quando g é convexa, a desigualdade de Jensen garante que o primeiro resultado é maior ou igual ao segundo.

A diferença entre duas expressões que aparecem na Desigualdade de Jensen quantifica, assim, o efeito de substituir X por sua média antes de aplicar g. Esse efeito depende tanto da distribuição de X quanto da função escolhida. Por exemplo, para $g(x) = x^2$, a diferença é exatamente a variância de X, quando o segundo momento é finito. Já para uma função afim, ela é sempre nula, mesmo que X não seja constante.

3.6. O Caso de Igualdade

Para descobrir quando a desigualdade de Jensen é saturada, isto é, quando ocorre a igualdade, comparamos $g(X)$ com a reta de suporte ao gráfico de g no ponto $\mu = \mathbb{E}[X]$. A diferença entre esses valores é sempre não negativa, e sua esperança é exatamente a diferença entre os dois lados de (13). Se g é estritamente convexa, essa reta toca o gráfico apenas em μ; assim, a igualdade só pode ocorrer quando $X = \mu$ quase certamente.

Proposição 15. Nas hipóteses do Teorema 13, suponha que g seja estritamente convexa. Então vale a igualdade em (13) se, e somente se, $X = \mathbb{E}[X]$ quase certamente.

Prova. Se $X = \mu$ quase certamente, então $\mathbb{E}[g(X)] = g(\mu) = g\Big(\mathbb{E}[X]\Big)$ e a igualdade vale. Suponha agora que valha a igualdade. Seja α o coeficiente dado pela reta de suporte em μ e defina

$$Z = g(X) - g(\mu) - \alpha(X - \mu).$$

Por (12), $Z \geqslant 0$. Como estamos assumindo que vale a igualdade $\mathbb{E}[g(X)] = g\Big(\mathbb{E}[X]\Big)$ então temos que $\mathbb{E}[Z] = 0$. Como mostrado anteriormente sabemos que se uma variável aleatória é

não-negativa e possui esperança nula então esta variável aleatória é nula quase certamente, logo

$$g(X) = g(\mu) + \alpha(X - \mu) \qquad \text{quase certamente.}$$

No evento $\{X \neq \mu\}$, a igualdade acima diria que a reta de suporte toca o gráfico em dois pontos distintos, X e μ. Mas, pela convexidade estrita, o ponto médio entre eles estaria estritamente abaixo da corda, e a corda coincide com a reta de suporte, o que contradiz (12). Logo $\{X \neq \mu\}$ tem probabilidade nula. ■

3.7. Exemplos

O primeiro exemplo recupera a não negatividade da variância.

Exemplo 16 (Não negatividade da variância)**.** Tome $g(x) = x^2$, que é estritamente convexa. O Teorema 13 dá

$$\mathbb{E}[X^2] \geqslant \Big(\mathbb{E}[X]\Big)^2,$$

isto é, $\mathrm{Var}(X) \geqslant 0$. A igualdade vale se, e somente se, X é constante quase certamente. Esse é o mesmo conteúdo da desigualdade de Cauchy-Schwarz para X e 1, que já foi provada nas notas sobre momentos, por outro argumento. Aqui ela aparece como um caso particular de convexidade.

O segundo exemplo é a desigualdade entre as médias geométrica e aritmética.

Exemplo 17 (Média geométrica e média aritmética)**.** A função $g(x) = -\log x$ é estritamente convexa em $(0, \infty)$, pois $g''(x) = 1/x^2 > 0$. Seja X uma variável aleatória positiva com $\mathbb{E}[X] < \infty$ e $\mathbb{E}\Big[|\log X|\Big] < \infty$. Então, segue da Desigualdade de Jensen que

$$\mathbb{E}[\log X] \leqslant \log \mathbb{E}[X], \tag{14}$$

ou, exponenciando,

$$\exp\Big(\mathbb{E}[\log X]\Big) \leqslant \mathbb{E}[X].$$

Suponha agora que X assume os valores positivos $x_1, \ldots, x_n$ com a mesma probabilidade. Então

$$\mathbb{E}[\log X] = \frac{1}{n}\sum_{i=1}^{n} \log x_i = \log \gamma, \qquad \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

onde

$$\gamma = \left(\prod_{i=1}^{n} x_i\right)^{1/n}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

são a média geométrica e a média aritmética dos x_i. Por (14), $\log \gamma \leqslant \log \bar{x}$, e como o logaritmo é crescente,

$$\gamma \leqslant \bar{x}.$$

Além disso, a igualdade vale se, e somente se, todos os x_i são iguais.

O terceiro exemplo compara a média harmônica com a geométrica.

Exemplo 18 (Média harmônica)**.** A média harmônica de números não-nulos $x_1, \ldots, x_n$ é definida por

$$\frac{1}{\eta} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}.$$

No caso em que $x_i > 0$ podemos aplicar o Exemplo 17 aos números $1/x_1, \ldots, 1/x_n$, obtemos

$$\Big(\prod_{i=1}^{n} \frac{1}{x_i}\Big)^{1/n} \leqslant \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i},$$

isto é, $1/\gamma \leqslant 1/\eta$, ou seja, $\eta \leqslant \gamma$. A média harmônica não excede a geométrica, que por sua vez não excede a aritmética.

O quarto exemplo conecta Jensen com a função geradora de momentos.

Exemplo 19 (A função geradora de momentos)**.** Para cada $t \in \mathbb{R} \setminus \{0\}$, a função $g(x) = e^{tx}$ é estritamente convexa (para $t = 0$ é constante). Portanto segue da Desigualdade de Jensen que

$$\mathbb{E}\Big[e^{tX}\Big] \geqslant e^{t\mathbb{E}[X]},$$

isto é, $M_X(t) \geqslant e^{t\mathbb{E}[X]}$ para todo t tal que a geradora exista, com igualdade para $t \neq 0$ se, e somente se, X é constante quase certamente.

4. A Desigualdade de Lyapunov

A convexidade também organiza a comparação entre momentos de ordens distintas. A pergunta é a seguinte: se $\mathbb{E}|X|^r$ é finito para algum $r > 0$, o que isso diz sobre $\mathbb{E}|X|^s$ para outros valores de s? A resposta tem duas partes. Primeiro, a finitude se propaga para baixo. Depois, as normas $(\mathbb{E}|X|^r)^{1/r}$ crescem com r. A segunda parte é a desigualdade de Lyapunov.

4.1. Finitude para Ordens Menores

O ponto de partida é uma desigualdade pontual entre potências.

Lema 20. Sejam $0 < s \leqslant r$ e $x \in \mathbb{R}$. Então

$$|x|^s \leqslant 1 + |x|^r. \tag{15}$$

Prova. Se $|x| \leqslant 1$, então $|x|^s \leqslant 1 \leqslant 1 + |x|^r$. Se $|x| > 1$, então $|x|^s \leqslant |x|^r \leqslant 1 + |x|^r$, porque $s \leqslant r$ e a função $p \longmapsto |x|^p$ é crescente em $(0, \infty)$ quando $|x| > 1$. ■

Tomando esperanças em (15), obtemos

$$\mathbb{E}|X|^s \leqslant 1 + \mathbb{E}|X|^r.$$

Em particular, se $\mathbb{E}|X|^r < \infty$, então $\mathbb{E}|X|^s < \infty$. Isso mostra que o conjunto

$$I = \Big\{r > 0 : \mathbb{E}|X|^r < \infty\Big\}$$

é um intervalo com extremo esquerdo 0: se $r \in I$, então $(0, r] \subseteq I$. O intervalo pode ser $(0, \infty)$, como no caso gaussiano, em que todos os momentos são finitos. Pode ser $(0, R)$ para algum $R > 0$, como no caso da distribuição de Cauchy, em que $\mathbb{E}|X|^r$ é finito para $0 < r < 1$ e infinito para $r \geqslant 1$. E pode ser $(0, R]$, se o momento de ordem R for finito mas os de ordem maior não forem. Em qualquer caso, a existência de um momento de ordem alta implica a existência de todos os momentos de ordem menor.

4.2. Enunciado e Prova da Desigualdade de Lyapunov

A desigualdade de Lyapunov compara as normas em ordens diferentes. A prova é uma aplicação de Jensen à variável Z^s e à função $y \longmapsto y^{r/s}$, que é convexa porque $r/s \geqslant 1$.

Teorema 21 (Desigualdade de Lyapunov)**.** Seja Z uma variável aleatória não negativa e sejam $0 < s \leqslant r$. Então

$$\Big(\mathbb{E}[Z^s]\Big)^{1/s} \leqslant \Big(\mathbb{E}[Z^r]\Big)^{1/r} \tag{16}$$

Prova. Se $\mathbb{E}[Z^r] = \infty$, a desigualdade é imediata. Suponha que $\mathbb{E}[Z^r] < \infty$. Pelo Lema 20, aplicado a Z, $\mathbb{E}[Z^s] < \infty$. Se $\mathbb{E}[Z^s] = 0$, então $Z = 0$ quase certamente e a desigualdade é trivial. Suponha $\mathbb{E}[Z^s] > 0$. Seja $Y = Z^s$ e $g(y) = y^{r/s}$, definida para $y \geqslant 0$. Como $r/s \geqslant 1$, a função g é convexa em $[0, \infty)$. De fato,

$$g'(y) = \frac{r}{s}\, y^{\frac{r}{s}-1} \qquad \text{e} \qquad g''(y) = \frac{r}{s}\left(\frac{r}{s} - 1\right) y^{\frac{r}{s}-2} > 0.$$

Usando a Desigualdade de Jensen para Y e g concluímos que

$$\mathbb{E}\Big[Z^r\Big] = \mathbb{E}\Big[Y^{r/s}\Big] = \mathbb{E}\Big[g(Y)\Big] \geqslant g\Big(\mathbb{E}[Y]\Big) = \Big(\mathbb{E}[Y]\Big)^{r/s} = \Big(\mathbb{E}[Z^s]\Big)^{r/s}.$$

Extraindo a raiz r-ésima dos dois lados, que é uma função crescente em $[0, \infty)$, obtemos finalmente a desigualdade (16). ■

A ordem das duas últimas subseções não é acidental. Para aplicar Jensen a $Y = Z^s$, precisamos saber que $\mathbb{E}[Z^s]$ é finita, e o Lema 20 fornece essa informação a partir de $\mathbb{E}[Z^r]$ sem usar Lyapunov. Se tentássemos provar a finitude por Lyapunov, usaríamos a desigualdade que queremos demonstrar.

Corolário 22. Seja X uma variável aleatória e sejam $0 < s \leqslant r$. Então

$$\Big(\mathbb{E}|X|^s\Big)^{1/s} \leqslant \Big(\mathbb{E}|X|^r\Big)^{1/r}.$$

Em particular, $\mathbb{E}|X| \leqslant \Big(\mathbb{E}[X^2]\Big)^{1/2}$.

Prova. Aplique o Teorema 21 a $Z = |X|$. ■

A desigualdade de Lyapunov permite comparar as cotas obtidas aplicando a desigualdade de Markov a diferentes potências de $|X|$. Suponha que $\mathbb{E}|X|^r < \infty$, e escreva $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ para $0 < q \leqslant r$. Para cada $s \in (0, r]$ e $t > 0$, a desigualdade de Markov dá

$$\mathbb{P}(|X| \geqslant t) = \mathbb{P}(|X|^s \geqslant t^s) \leqslant \frac{\mathbb{E}|X|^s}{t^s} = \left(\frac{\|X\|_s}{t}\right)^s.$$

Pelo Corolário 22, $\|X\|_s \leqslant \|X\|_r$; logo, a cota acima também é limitada por $(\|X\|_r/t)^s$. Quando $t > \|X\|_r$, a razão $\|X\|_r/t$ está entre 0 e 1, de modo que, para $s < r$,

$$\left(\frac{\|X\|_r}{t}\right)^r < \left(\frac{\|X\|_r}{t}\right)^s.$$

Assim, entre essas cotas comparadas usando a mesma quantidade $\|X\|_r$, a que usa o maior expoente fornece o menor limite para desvios acima de $\|X\|_r$. Em particular, usando diretamente o momento de ordem r, obtemos

$$\mathbb{P}(|X| \geqslant t) \leqslant \frac{\mathbb{E}|X|^r}{t^r},$$

uma cota que decai como t^{-r} quando $t \to \infty$. Portanto, quanto maior for a ordem do momento finito disponível, mais rápido pode ser o decaimento polinomial garantido para a cauda de $|X|$.

4.3. A Curva das Normas

O Teorema 21 diz que a função $r \longmapsto (\mathbb{E}|X|^r)^{1/r}$ é não decrescente. Há uma segunda estrutura por trás dela: o logaritmo dessa função é convexo. A prova usa a desigualdade de Hölder, que é a forma geral da desigualdade de Cauchy-Schwarz e que também sai da convexidade.

Lema 23 (Desigualdade de Young)**.** Sejam $p, q > 1$ com $1/p + 1/q = 1$ e sejam $a, b \geqslant 0$. Então

$$ab \leqslant \frac{a^p}{p} + \frac{b^q}{q}.$$

Prova. Se $a = 0$ ou $b = 0$, a desigualdade é imediata. Caso contrário, a função $t \longmapsto e^t$ é convexa, logo

$$ab = \exp\Big(\frac{1}{p}\log a^p + \frac{1}{q}\log b^q\Big) \leqslant \frac{1}{p}e^{\log a^p} + \frac{1}{q}e^{\log b^q} = \frac{a^p}{p} + \frac{b^q}{q}.$$

■

Teorema 24 (Desigualdade de Hölder)**.** Sejam $p, q > 1$ com $1/p + 1/q = 1$ e sejam X, Y variáveis aleatórias tais que $\mathbb{E}|X|^p < \infty$ e $\mathbb{E}|Y|^q < \infty$. Então

$$\mathbb{E}|XY| \leqslant \Big(\mathbb{E}|X|^p\Big)^{1/p}\Big(\mathbb{E}|Y|^q\Big)^{1/q}.$$

Prova. Se $\mathbb{E}|X|^p = 0$, então $X = 0$ quase certamente e a desigualdade é imediata. O mesmo vale para Y. Suponha então que os dois momentos sejam positivos e defina

$$\tilde{X} = \frac{X}{(\mathbb{E}|X|^p)^{1/p}}, \qquad \tilde{Y} = \frac{Y}{(\mathbb{E}|Y|^q)^{1/q}}.$$

Então $\mathbb{E}|\tilde{X}|^p = \mathbb{E}|\tilde{Y}|^q = 1$. Aplicando o Lema 23 a $|\tilde{X}|$ e $|\tilde{Y}|$, obtemos

$$|\tilde{X}\tilde{Y}| \leqslant \frac{|\tilde{X}|^p}{p} + \frac{|\tilde{Y}|^q}{q}.$$

Tomando esperanças,

$$\mathbb{E}|\tilde{X}\tilde{Y}| \leqslant \frac{1}{p} + \frac{1}{q} = 1,$$

que é a desigualdade para X e Y. ■

Proposição 25. Seja X uma variável aleatória com $\mathbb{P}(X \neq 0) > 0$ e seja $I = \{r > 0 : \mathbb{E}|X|^r < \infty\}$. Então a função $r \longmapsto \log \mathbb{E}|X|^r$ é convexa em I.

Prova. Sejam $r_1 < r_2$ em I e $t \in (0,1)$. Escreva $r = (1-t)r_1 + tr_2$. Queremos mostrar que

$$\log \mathbb{E}|X|^r \leqslant (1-t)\log \mathbb{E}|X|^{r_1} + t \log \mathbb{E}|X|^{r_2}.$$

Equivalentemente, depois de exponenciar,

$$\mathbb{E}|X|^r \leqslant \Big(\mathbb{E}|X|^{r_1}\Big)^{1-t}\Big(\mathbb{E}|X|^{r_2}\Big)^t.$$

Escrevemos $|X|^r = |X|^{(1-t)r_1}|X|^{tr_2}$ e aplicamos a desigualdade de Hölder com

$$p = \frac{1}{1-t}, \qquad q = \frac{1}{t},$$

de modo que $1/p + 1/q = 1$. Tomando $X_1 = |X|^{(1-t)r_1}$ e $X_2 = |X|^{tr_2}$, Hölder dá

$$\mathbb{E}|X|^r \leqslant \Big(\mathbb{E}|X|^{r_1}\Big)^{1-t}\Big(\mathbb{E}|X|^{r_2}\Big)^t,$$

como queríamos. ■

4.4. Aplicação: O Jogo Honesto e a Ruína

Considere o seguinte jogo. O jogador começa com capital $W_0 = 1$. Em cada rodada, o capital é multiplicado por 3/2 ou por 1/2, com probabilidade 1/2 cada, independentemente das rodadas anteriores. O jogo é honesto no sentido de que a taxa de retorno esperado, em cada uma das rodadas é dada por cópias independentes da variável aleatória R_1 cuja a distribuição é dada por

$$\mathbb{P}\left(R_1 = \frac{3}{2}\right) = \frac{1}{2} = \mathbb{P}\left(R_1 = \frac{1}{2}\right)$$

e seu valor esperado é dado por

$$\mathbb{E}[R_1] = \frac{1}{2}\cdot\frac{3}{2} + \frac{1}{2}\cdot\frac{1}{2} = 1.$$

Pela hipótese de independência, sabemos que taxa de retorno esperada deste jogo após n rodadas é dada pela esperança da variável aleatória $W_n \equiv R_1 \cdot R_2 \cdot \ldots \cdot R_n$ que é precisamente

$$\mathbb{E}[W_n] = \mathbb{E}[R_1 \cdot R_2 \cdot \ldots \cdot R_n] = \mathbb{E}[R_1] \cdot \mathbb{E}[R_2] \cdot \ldots \cdot \mathbb{E}[R_n] = 1.$$

A pergunta é se o capital típico se mantém, cresce ou decai. Para entender melhor este problema observe que o capital (aqui o capital inicial é de uma unidade) depois de n rodadas é dado por

$$W_n = R_1 R_2 \cdots R_n,$$

com os R_i independentes e identicamente distribuídos. Como $W_n > 0$ sempre, podemos tomar logaritmos. A convexidade de $-\log$ ou, equivalentemente, a concavidade de log, dá

$$\mathbb{E}[\log W_n] \leqslant \log \mathbb{E}[W_n] = 0.$$

A desigualdade é estrita, porque W_n não é constante. De fato,

$$\mathbb{E}[\log R_1] = \frac{1}{2}\log\frac{3}{2} + \frac{1}{2}\log\frac{1}{2} = \frac{1}{2}\log\frac{3}{4} = -\frac{1}{2}\log\frac{4}{3} \approx -0{,}14384,$$

logo

$$\mathbb{E}[\log W_n] = \sum_{i=1}^{n} \mathbb{E}[\log R_i] = -\frac{n}{2}\log\frac{4}{3} \xrightarrow[n\to\infty]{} -\infty. \tag{17}$$

Enquanto o logaritmo da esperança do capital é zero, a esperança do logaritmo tende a $-\infty$. A diferença entre as duas quantidades se manifesta como sendo a diferença dos dois termos que aparecem na Desigualdade de Jensen. Neste caso vimos que esta diferença cresce linearmente em n.

Vale registrar a leitura pela desigualdade de Lyapunov. Como os R_i são independentes,

$$\mathbb{E}[W_n^2] = \prod_{i=1}^{n} \mathbb{E}[R_i^2] = \left(\frac{5}{4}\right)^n,$$

e o Corolário 22 dá

$$\mathbb{E}[W_n] \leqslant \left(\mathbb{E}[W_n^2]\right)^{1/2} = \left(\frac{5}{4}\right)^{n/2}.$$

A cota cresce exponencialmente, enquanto a esperança é constante. A distância entre os dois valores médios apresenta também uma certa assimetria analogamente ao que vimos pela aplicação da Desigualdade de Jensen.

Para transformar (17) em uma afirmação sobre o capital, sem usar a lei dos grandes números, basta conhecer a distribuição exata. Seja K_n o número de rodadas em que o capital foi multiplicado por $3/2$. Então $K_n \sim \mathrm{Bin}(n, 1/2)$ e

$$W_n = \left(\frac{3}{2}\right)^{K_n}\left(\frac{1}{2}\right)^{n-K_n} = \frac{3^{K_n}}{2^n}.$$

Se $K_n \leqslant n/2$, então

$$W_n \leqslant \frac{3^{n/2}}{2^n} = \left(\frac{\sqrt{3}}{2}\right)^n.$$

Como a distribuição de K_n é simétrica em torno de $n/2$, o evento $\{K_n \leqslant n/2\}$ tem probabilidade pelo menos $1/2$.

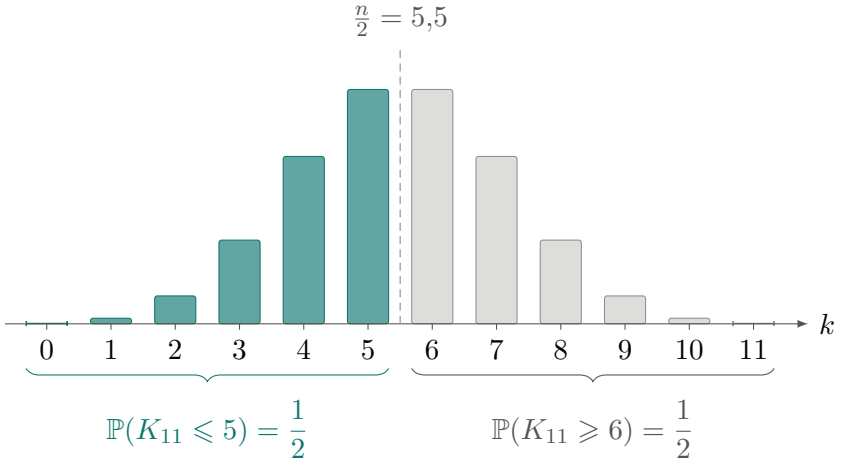


Figura 10: Função de probabilidade de uma v.a com distribuição $\mathrm{Bin}(11, \frac{1}{2})$

Logo

$$\mathbb{P}\Big(W_n \leqslant \Big(\frac{\sqrt{3}}{2}\Big)^n\Big) \geqslant \frac{1}{2}. \tag{18}$$

Com probabilidade pelo menos um meio, o capital decai a uma taxa geométrica de $\sqrt{3}/2 \approx 0{,}866$ por rodada, isto é, pelo menos 13,4% ao longo de cada rodada, em termos geométricos. Quando n é par, o valor mais provável de W_n ocorre em $K_n = n/2$ e vale exatamente $(\sqrt{3}/2)^n$. Quando n é ímpar, os dois valores mais prováveis correspondem a $K_n = (n-1)/2$ e $K_n = (n+1)/2$.

A desigualdade de Chebyshev dá uma segunda leitura, agora sobre a probabilidade de o jogador terminar acima do capital inicial. O evento $\{W_n > 1\}$ é o mesmo que $\{3^{K_n} > 2^n\}$, ou seja,

$$K_n > n\frac{\log 2}{\log 3}.$$

Como $\log 2/\log 3 \approx 0{,}63093$, o evento é $\{K_n - n/2 > 0{,}13093\, n\}$. A variância de K_n é $n/4$, logo Chebyshev dá

$$\mathbb{P}(W_n > 1) \leqslant \frac{n/4}{(0{,}13093\, n)^2} = \frac{14{,}58}{n}. \tag{19}$$

Assim a probabilidade de terminar com mais do que se começou tende a zero, embora a esperança do capital seja 1 para todo n.

Os números para $n = 20$ resumem o fenômeno. Temos $\mathbb{E}[W_{20}] = 1$ e $\mathbb{E}[W_{20}^2] = (5/4)^{20} \approx 86{,}7$, de modo que o desvio padrão de W_{20} é aproximadamente 9,3. A cota (18) diz que, com probabilidade pelo menos $1/2$, o capital é no máximo $(0{,}866)^{20} \approx 0{,}056$. O valor mais provável é exatamente esse, com probabilidade $\binom{20}{10}2^{-20} \approx 0{,}176$. Em outras palavras, o capital mais provável é cerca de 5,6% do inicial, enquanto a média é 100% do inicial. A média é sustentada por uma minoria de trajetórias com ganhos muito grandes.

Como os fatores $R_1, \ldots, R_n$ são independentes, a esperança do produto é o produto das esperanças e, neste jogo, $\mathbb{E}[W_n] = 1$. Essa quantidade descreve o capital médio, mas não basta para descrever como o capital evolui na maioria das trajetórias. Para estudar o crescimento multiplicativo típico é mais eficiente olhar para logaritmos $\log W_n = \sum_{j=1}^{n} \log R_j$. Assim, $\mathbb{E}[\log R_1]$ é o incremento médio por rodada na escala logarítmica; como esse valor é negativo, os cálculos com a distribuição de K_n acima mostram que o capital costuma decair, embora sua esperança permaneça igual a 1. A desigualdade de Jensen expressa a diferença entre essas duas quantidades:

$$\mathbb{E}[\log W_n] < \log \mathbb{E}[W_n] = 0.$$

Portanto, a esperança do capital e o crescimento na escala logarítmica dão informações diferentes: a esperança do capital é sustentada pelas trajetórias raras em que ele cresce muito, enquanto o crescimento na escala logarítmica registra o efeito médio dos fatores multiplicativos e, neste jogo, reflete o decaimento típico do capital.

5. O Lema de Borel-Cantelli

Até aqui, estimamos a probabilidade de um evento. Nesta seção, a pergunta é diferente: dada uma sequência de eventos $A_1, A_2, \ldots$, queremos saber se eles ocorrem infinitas vezes. A soma $\sum_n \mathbb{P}(A_n)$ é o número esperado de ocorrências, e o lema de Borel-Cantelli diz que o

comportamento dessa série decide a questão, desde que os eventos sejam independentes na direção difícil.

5.1. Limites Superior e Inferior de Eventos

Para formular a pergunta, precisamos de uma noção de limite para sequências de eventos.

Definição 26 (Limite superior)**.** Seja $(A_n)_{n\geqslant 1}$ uma sequência de eventos em um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$. O *limite superior* de (A_n) é o evento

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Escrevemos também $\{A_n$ infinitas vezes$\}$ ou, abreviadamente, $\{A_n$ i.v.$\}$ para esse evento.

A definição merece uma leitura em palavras. Um ponto ω pertence a $\bigcup_{k\geqslant n} A_k$ se, e somente se, ω pertence a algum A_k com $k \geqslant n$. Se ω pertence à interseção sobre n, então, para cada n, existe $k \geqslant n$ com $\omega \in A_k$. Isso significa que os índices k para os quais $\omega \in A_k$ não são limitados, ou seja, ω pertence a infinitos dos eventos A_n. Reciprocamente, se ω pertence a infinitos A_n, então para cada n existe $k \geqslant n$ com $\omega \in A_k$, e ω está no limite superior. Assim

$$\limsup_{n\to\infty} A_n = \{\omega \in \Omega : \omega \in A_n \text{ para infinitos índices } n\}.$$

O limite inferior tem definição dual.

Definição 27 (Limite inferior)**.** Nas mesmas hipóteses, o *limite inferior* de (A_n) é o evento

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

Escrevemos também $\{A_n$ eventualmente$\}$.

Aqui, ω pertence a $\bigcap_{k\geqslant n} A_k$ se, e somente se, $\omega \in A_k$ para todo $k \geqslant n$. Se ω pertence à união sobre n, então existe n tal que $\omega \in A_k$ para todo $k \geqslant n$, isto é, ω pertence a todos os A_n a partir de algum índice. Logo

$$\liminf_{n\to\infty} A_n = \{\omega \in \Omega : \omega \in A_n \text{ para todo } n \text{ suficientemente grande}\}.$$

Os dois eventos se relacionam por complementação. Como o complementar troca uniões por interseções e inverte o sentido das inclusões,

$$\left(\limsup_{n\to\infty} A_n\right)^c = \liminf_{n\to\infty} A_n^c. \tag{20}$$

Além disso, todo ponto que pertence a todos os A_n a partir de um índice pertence a infinitos deles, logo

$$\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n.$$

A inclusão pode ser estrita: no lançamento de uma moeda, o evento "sai cara infinitas vezes" contém o evento "sai cara a partir de algum lançamento", e a inclusão é estrita quase certamente.

Observação 28. A definição se traduz em uma identidade entre indicadoras:

$$\mathbb{1}_{\limsup_n A_n} = \limsup_{n\to\infty} \mathbb{1}_{A_n}, \qquad \mathbb{1}_{\liminf_n A_n} = \liminf_{n\to\infty} \mathbb{1}_{A_n}.$$

Em particular, $\omega \in \limsup_n A_n$ se, e somente se,

$$\sum_{n=1}^{\infty} \mathbb{1}_{A_n}(\omega) = \infty.$$

Essa leitura é útil porque transforma a pergunta sobre eventos em uma pergunta sobre séries de números não-negativos.

5.2. A Continuidade das Medidas de Probabilidade

O instrumento técnico que liga os limites de eventos às probabilidades é a continuidade das medidas, que já foi demonstrada na nota de propriedades de probabilidade. Recordamos o enunciado.

Teorema 29 (Continuidade para sequências crescentes)**.** Seja $(\Omega, \mathcal{F}, \mathbb{P})$ um espaço de probabilidade e seja $(A_n)_{n\geqslant 1}$ uma sequência crescente de eventos, com $A = \bigcup_{n\geqslant 1} A_n$. Então

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Corolário 30 (Continuidade para sequências decrescentes)**.** Seja $(B_n)_{n\geqslant 1}$ uma sequência decrescente de eventos, com $B = \bigcap_{n\geqslant 1} B_n$. Então

$$\mathbb{P}(B) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

A prova do teorema está na nota de propriedades de probabilidade, onde a união é decomposta em anéis disjuntos. O corolário segue aplicando o teorema aos complementares, que formam uma sequência crescente de limite $\Omega \setminus B$.

No que segue, aplicaremos o corolário à sequência

$$C_n = \bigcup_{k=n}^{\infty} A_k.$$

Essa sequência é decrescente, pois $C_{n+1} \subseteq C_n$, e sua interseção é exatamente $\limsup_n A_n$. Logo

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = \lim_{n\to\infty} \mathbb{P}(C_n). \tag{21}$$

A identidade (21) é a ponte entre o lema de Borel-Cantelli e a continuidade.

5.3. A Desigualdade de Boole

Falta controlar a probabilidade de uma união enumerável. A desigualdade de Boole faz isso, e sua versão finita já apareceu na nota de propriedades de probabilidade.

Lema 31 (Desigualdade de Boole)**.** Seja $(A_n)_{n\geqslant 1}$ uma sequência de eventos. Então

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Prova. Defina $B_1 = A_1$ e, para $n \geqslant 2$,

$$B_n = A_n \setminus \bigcup_{k=1}^{n-1} A_k.$$

Os eventos B_n são dois a dois disjuntos e $\bigcup_{n\geqslant 1} B_n = \bigcup_{n\geqslant 1} A_n$. Além disso, $B_n \subseteq A_n$, logo $\mathbb{P}(B_n) \leqslant \mathbb{P}(A_n)$. Pela σ-aditividade,

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

■

5.4. O Primeiro Lema de Borel-Cantelli

Vermos a seguir que a continuidade da medida e a desigualdade de Boole é praticamente tudo que precisamos para a prova da primeira parte do Lema de Borel-Cantelli.

Teorema 32 (Primeiro lema de Borel-Cantelli)**.** Seja $(A_n)_{n\geqslant 1}$ uma sequência de eventos tal que $\sum_{n\geqslant 1} \mathbb{P}(A_n) < \infty$. Então

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = 0.$$

Em outras palavras, com probabilidade 1 ocorrem apenas finitos dos eventos A_n.

Prova. Já que a sequência de eventos $C_n \equiv \bigcup_{k\geqslant n} A_k$ é uma sequência decrescente de eventos, segue da continuidade da medida que

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = \lim_{n\to\infty} \mathbb{P}(C_n).$$

Portanto segue da Desigualdade de Boole Lema 31 que

$$\mathbb{P}(C_n) \leqslant \sum_{k=n}^{\infty} \mathbb{P}(A_k).$$

Como a série $\sum_k \mathbb{P}(A_k)$ converge, sua cauda $\sum_{k\geqslant n} \mathbb{P}(A_k) \to 0$, quando $n \to \infty$. Logo $\mathbb{P}(C_n) \to 0$ e, portanto, $\mathbb{P}(\limsup_n A_n) = 0$. ■

Note que a prova não usou independência. O primeiro lema vale para qualquer sequência de eventos.

5.5. O Segundo Lema de Borel-Cantelli

A recíproca do primeiro lema é falsa em geral: se $A_n = A$ para todo n, com $0 < \mathbb{P}(A) < 1$, então $\sum_n \mathbb{P}(A_n) = \infty$ e $\limsup_n A_n = A$, cuja probabilidade não é 1. O segundo lema mostra que a recíproca vale quando os eventos são independentes.

Usaremos a desigualdade elementar

$$1 - x \leqslant e^{-x} \qquad \text{para } x \in \mathbb{R}, \tag{22}$$

que segue da convexidade de $x \longmapsto e^x$, ou da comparação das séries de Taylor.

Teorema 33 (Segundo lema de Borel-Cantelli)**.** Seja $(A_n)_{n\geqslant 1}$ uma sequência de eventos independentes tal que $\sum_{n\geqslant 1}\mathbb{P}(A_n)=\infty$. Então

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big)=1.$$

Prova. Seja $p_n=\mathbb{P}(A_n)$. Pela identidade (20),

$$\Big(\limsup_{n\to\infty} A_n\Big)^c=\liminf_{n\to\infty} A_n^c=\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^c.$$

Pela desigualdade de Boole, basta mostrar que $\mathbb{P}\big(\bigcap_{n\geqslant m} A_n^c\big)=0$ para cada m.

Fixe m. Para $N\geqslant m$, os eventos $A_m^c,\dots,A_N^c$ são independentes, logo

$$\mathbb{P}\Big(\bigcap_{n=m}^{N} A_n^c\Big)=\prod_{n=m}^{N}(1-p_n)\leqslant \exp\Big(-\sum_{n=m}^{N} p_n\Big),$$

onde usamos a desigualdade elementar (22) em cada fator. A sequência $\bigcap_{n=m}^{N} A_n^c$ é decrescente em N, com interseção $\bigcap_{n\geqslant m} A_n^c$. Pela propriedade de continuidade da medida Corolário 30, podemos afirmar que

$$\mathbb{P}\Big(\bigcap_{n=m}^{\infty} A_n^c\Big)=\lim_{N\to\infty}\mathbb{P}\Big(\bigcap_{n=m}^{N} A_n^c\Big)\leqslant \lim_{N\to\infty}\exp\Big(-\sum_{n=m}^{N} p_n\Big)=0,$$

onde a última igualdade segue da hipótese da série $\sum_n p_n$ ser divergente. Como m é arbitrário, todos os eventos $\bigcap_{n\geqslant m} A_n^c$ têm probabilidade nula, e a união enumerável deles também. Logo $\mathbb{P}\big((\limsup_n A_n)^c\big)=0$, isto é, $\mathbb{P}(\limsup_n A_n)=1$. ∎

Observação 34. A hipótese de independência é indispensável no segundo lema. O exemplo $A_n=A$, com $0<\mathbb{P}(A)<1$, mostra que sem ela a conclusão falha. Um segundo exemplo, que mostra que a conclusão pode falhar mesmo quando os eventos são distintos, é $A_n=[0,1/n]$ no espaço $([0,1],\mathcal{B}([0,1]),\lambda)$, onde λ é a medida uniforme: a soma das probabilidades diverge, mas a interseção dos A_n é $\{0\}$, de modo que $\limsup_n A_n=\{0\}$ e tem probabilidade nula.

5.6. A Lei 0-1 de Borel

Reunindo as duas partes, obtemos um critério completo para eventos independentes.

Corolário 35 (Lei 0-1 de Borel)**.** Seja $(A_n)_{n\geqslant 1}$ uma sequência de eventos independentes. Então

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big)\in\{0,1\}.$$

Mais precisamente, a probabilidade é 0 se $\sum_n \mathbb{P}(A_n)<\infty$ e igual a 1 se $\sum_n \mathbb{P}(A_n)=\infty$.

Prova. Se a série converge, o Teorema 32 dá probabilidade 0. Se diverge, o Teorema 33 dá probabilidade 1. ∎

5.7. Exemplos

Exemplo 36 (Caras em lançamentos sucessivos)**.** Lance uma moeda honesta infinitas vezes e seja A_n o evento "sai cara no lançamento n". Os eventos são independentes, $\mathbb{P}(A_n) = 1/2$ e $\sum_n \mathbb{P}(A_n) = \infty$. Pelo Teorema 33, saem caras infinitas vezes com probabilidade 1.

Exemplo 37 (Blocos iniciais de caras)**.** No mesmo espaço, seja B_n o evento "os primeiros n lançamentos são todos caras". Então $\mathbb{P}(B_n) = 2^{-n}$ e $\sum_n 2^{-n} < \infty$. Pelo Teorema 32, com probabilidade 1 apenas finitos dos B_n ocorrem. Isso **não** contradiz o exemplo anterior: o que se repete infinitas vezes são caras em posições isoladas, não blocos de caras que começam na posição 1.

5.8. Aplicação: A Maior Sequência de Caras

Considere uma moeda honesta lançada infinitas vezes e seja L_n o comprimento da maior sequência de caras consecutivas inteiramente contida nos primeiros n lançamentos. Quanto vale L_n para n grande? A intuição sugere que sequências longas da mesma face são raras, entretanto o número de posições onde elas podem começar cresce com n, e o equilíbrio entre os dois efeitos produz a escala $\log_2 n$. O lema de Borel-Cantelli determina a constante.

Seja A_i o evento "sai cara no lançamento i". Para $k \geqslant 1$ e $j \geqslant 1$, o evento

$$B_{j,k} = A_j \cap A_{j+1} \cap \cdots \cap A_{j+k-1}$$

tem probabilidade 2^{-k}, e eventos desse tipo são independentes quando os blocos de índices são disjuntos.

Limite superior. Fixe $\varepsilon > 0$ e, para cada $m \geqslant 2$, ponha $k_m = \lceil (1+\varepsilon)\log_2 m \rceil$. Para $m \geqslant k_m$, seja $E_m = B_{m-k_m+1,k_m}$ o evento de que uma sequência de k_m caras consecutivas termine exatamente no lançamento m. Então

$$\mathbb{P}(E_m) = 2^{-k_m} \leqslant m^{-(1+\varepsilon)}.$$

Como $1+\varepsilon > 1$, a série $\sum_m m^{-(1+\varepsilon)}$ converge, e o Teorema 32 garante que, quase certamente, apenas finitos dos eventos E_m ocorrem. Existe, portanto, um índice aleatório N_0 tal que E_m não ocorre para nenhum $m \geqslant N_0$.

Considere agora $n \geqslant N_0$ e qualquer sequência de caras consecutivas contida em $\{1, \ldots, n\}$, digamos terminando na posição $m \leqslant n$. Se $m < N_0$, seu comprimento é no máximo $m < N_0$; se $N_0 \leqslant m \leqslant n$, como E_m não ocorre, seu comprimento é estritamente menor do que $k_m \leqslant k_n$. Logo $L_n \leqslant \max\{N_0, k_n - 1\}$. Para n suficientemente grande temos $k_n > N_0$, donde $L_n \leqslant k_n - 1 < (1+\varepsilon)\log_2 n$. Em particular,

$$\limsup_{n\to\infty} \frac{L_n}{\log_2 n} \leqslant 1 \qquad \text{quase certamente.}$$

Limite inferior. Fixe $\varepsilon \in (0,1)$ e ponha $k_n = \lfloor (1-\varepsilon)\log_2 n \rfloor$. Dividimos os primeiros n lançamentos em $m_n = \lfloor n/k_n \rfloor$ blocos disjuntos de k_n posições consecutivas. Se $L_n < k_n$, nenhum desses blocos é todo de caras. Como os blocos são disjuntos, os eventos correspondentes são independentes, e cada um tem probabilidade 2^{-k_n} de ser todo de caras. Logo

$$\mathbb{P}(L_n < k_n) \leqslant \left(1 - 2^{-k_n}\right)^{m_n} \leqslant \exp\!\left(-m_n 2^{-k_n}\right).$$

Agora $2^{-k_n} \geqslant n^{-(1-\varepsilon)}$ e, para n grande, $m_n \geqslant n/(2k_n)$. Assim

$$m_n 2^{-k_n} \geqslant \frac{n^{\varepsilon}}{2k_n} \geqslant \frac{n^{\varepsilon}}{2(1+\log_2 n)} \xrightarrow[n\to\infty]{} \infty.$$

Segue que $\mathbb{P}(L_n < k_n) \leqslant \exp\bigl(-n^{\varepsilon}/(2k_n)\bigr)$, e a série dessas probabilidades converge, porque a exponencial decai mais rápido do que qualquer potência de n. Pelo Teorema 32,

$$\mathbb{P}\Bigl(L_n < (1-\varepsilon)\log_2 n \text{ para infinitos } n\Bigr) = 0,$$

isto é, quase certamente $L_n \geqslant k_n$ para todo n suficientemente grande. Logo

$$\liminf_{n\to\infty} \frac{L_n}{\log_2 n} \geqslant 1 \qquad \text{quase certamente.}$$

Conclusão. Reunindo os dois limites,

$$\frac{L_n}{\log_2 n} \xrightarrow[n\to\infty]{} 1 \qquad \text{quase certamente.} \tag{23}$$

Mais precisamente, para cada $\varepsilon > 0$ e quase todo ω, vale

$$(1-\varepsilon)\log_2 n - 1 \leqslant L_n \leqslant (1+\varepsilon)\log_2 n$$

para todo n suficientemente grande.

Vale observar o papel das duas direções. O limite superior controla a probabilidade de uma sequência longa terminar no instante m e usa o primeiro lema para concluir que isso só ocorre um número finito de vezes. O limite inferior usa a independência de blocos disjuntos para estimar a probabilidade de não haver nenhum bloco de caras, e depois aplica novamente o primeiro lema. A segunda parte do lema não é usada, o que mostra que ela não é a única forma de obter conclusões de probabilidade 1.

Os números tornam a escala concreta. Como $\log_2(10^6) \approx 19{,}93$, em um milhão de lançamentos a maior sequência tem cerca de 20 caras. Para ver 30 caras seguidas é preciso da ordem de $2^{30} \approx 10^9$ lançamentos, e para 40 caras, da ordem de $2^{40} \approx 10^{12}$. A probabilidade de 20 caras em 20 lançamentos é $2^{-20} \approx 9{,}5 \times 10^{-7}$, mas há cerca de 10^6 posições onde a sequência pode começar, e o produto é da ordem de 1.

6. Exercícios Propostos

1. Determine quais distribuições em $[0,\infty)$, com média μ, têm 2μ como mediana. Isto é, determine quando a cota da Proposição 5 é atingida.

2. Use a desigualdade de Chebyshev para mostrar que, em n lançamentos de um dado honesto, a probabilidade de o número de seis estar entre $\frac{1}{6}n - \sqrt{n}$ e $\frac{1}{6}n + \sqrt{n}$ é pelo menos $\frac{31}{36}$.

∗**3 Desigualdade de Cantelli.** Seja X uma variável aleatória com média μ e variância $\sigma^2 < \infty$. Mostre que, para todo $a > 0$,

$$\mathbb{P}(X - \mu \geqslant a) \leqslant \frac{\sigma^2}{\sigma^2 + a^2}.$$

Sugestão: Para $c > 0$, observe que $X - \mu \geqslant a$ implica $X - \mu + c \geqslant a + c$, aplique a desigualdade de Markov a $(X - \mu + c)^2$ e minimize a cota em c.

4. Seja I um intervalo aberto da reta e seja $f\colon I \to \mathbb{R}$ duas vezes diferenciável, com $f''(x) > 0$ para todo $x \in I$. Usando a fórmula de Taylor com resto de Lagrange, mostre que f é estritamente convexa em I. Dê um exemplo mostrando que a recíproca não vale.

5 Entropia de Shannon. Sejam $p_1, \ldots, p_n > 0$ tais que $\sum_{i=1}^n p_i = 1$. Usando a desigualdade de Jensen, mostre que

$$-\sum_{i=1}^{n} p_i \log p_i \leqslant \log n,$$

e determine quando vale a igualdade.

6. Sejam $x_1, \ldots, x_n > 0$ e seja η a média harmônica definida no Exemplo 18. Aplicando diretamente a desigualdade de Jensen à função estritamente convexa $g(x) = 1/x$ em $(0, \infty)$, mostre que $\eta \leqslant \bar{x}$, onde $\bar{x}$ é a média aritmética, e determine quando vale a igualdade.

7. Seja $p \geqslant 1$. Usando a desigualdade de Jensen, encontre o menor valor da constante c_p tal que

$$(a+b)^p \leqslant c_p(a^p + b^p) \qquad \text{para todos } a, b \geqslant 0.$$

8. Seja X uma variável aleatória limitada e defina $M = \inf\{c \geqslant 0 : \mathbb{P}(|X| \leqslant c) = 1\}$. Mostre que

$$\lim_{r\to\infty} \big(\mathbb{E}|X|^r\big)^{1/r} = M.$$

9. Sejam $A_1, A_2, \ldots$ eventos independentes com $\mathbb{P}(A_n) = p_n \in [0, 1)$ para todo n. Mostre que

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} A_n^c\Big) > 0$$

se, e somente se, $\sum_n p_n < \infty$.

∗ **10.** Mostre que a conclusão do Teorema 33 continua válida se a hipótese de independência for substituída por independência dois a dois. Isto é, mostre que, se (A_n) é uma sequência de eventos dois a dois independentes com $\sum_n \mathbb{P}(A_n) = \infty$, então $\mathbb{P}(\limsup_n A_n) = 1$. *Sugestão:* Aplique a desigualdade de Chebyshev a $S_{m,N} = \sum_{n=m}^{N} \mathbb{1}_{A_n}$.

11. Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes com distribuição exponencial de parâmetro 1 e seja $M_n = \max\{X_1, \ldots, X_n\}$. Mostre que

$$\frac{M_n}{\log n} \xrightarrow[n\to\infty]{} 1 \qquad \text{quase certamente.}$$

Referências

[1] Geoffrey Grimmett and Dominic Welsh. *Probability—an introduction*. Oxford University Press, Oxford, second edition, 2014.

[2] Leonardo T. Rolla and Bernardo N. B. de Lima. *Probabilidade*. Projeto Euclides. IMPA, Rio de Janeiro, 2026.

[3] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.