Estatística Geral I

Eduardo Horta

1 Variáveis Aleatórias Contínuas (sem Cálculo!)

Embora o conceito de uma variável aleatória discreta seja, em geral, apreendido facilmente, e em que pese sua ampla utilidade, existem muitas situações nas quais essa noção é inadequada para descrever fenômenos de interesse. Vamos começar com um exemplo.

Exemplo 1.1 (Morettin e Bussab, Exemplo 7.1)**.** Consideremos um relógio analógico cujo ponteiro dos segundos 'salta' de 6 em 6 graus. Esse relógio é alimentado por uma pilha AA que pode se esgotar a qualquer momento, e o instante exato em que o relógio para é, para todos os efeitos, aleatório. Seja X a variável aleatória "ângulo formado pelo ponteiro dos segundos no instante em que o relógio parar". Então X é uma v.a. discreta cujos possíveis valores são $0°, 6°, 12°, \ldots, 348°$ e $354°$. Vamos chamar esses possíveis valores de **sítios** ocupados pelo ponteiro. Qual é a função massa de probabilidade de X? Se considerarmos que o momento em que a pilha se esgota 'não privilegia' nenhum instante em particular, então a probabilidade de o ponteiro se encontrar em qualquer dos sítios deve ser a mesma: X é variável aleatória Uniforme discreta, com função massa de probabilidade

$$p_X(x) = \begin{cases} \frac{1}{60} & \text{se } x = 0°, 6°, \ldots, 354° \\ 0 & \text{demais valores de } x. \end{cases}$$

Consideremos agora um relógio cujo ponteiro dos segundos se move continuamente. Tal relógio pode ser pensado como um mecanismo ideal que mede o passar do tempo com *precisão arbitrária*. Para ilustrar, vamos considerar primeiramente um relógio de alta precisão: por exemplo, um relógio que mede o tempo em nanosegundos (10^{-9} segundos). Nesse caso, já que 1 segundo $= 10^9$ nanosegundos, o número total de sítios que o ponteiro pode ocupar é de $60 \cdot 10^9$. De fato, o ponteiro inicia no sítio $0°$, 'salta' para o sítio $1° \cdot 6 \cdot 10^{-9}$, depois para o sítio $2° \cdot 6 \cdot 10^{-9}$ e assim sucessivamente. Isto é, os possíveis sítios ocupados pelo ponteiro são dados por

$$x_k = k \cdot \frac{360°}{60 \cdot 10^9}, \qquad k = 0, 1, 2, \ldots, 60{\cdot}10^9 - 1$$

(Para obter essa fórmula, basta pensar que dividimos os $360°$ do relógio em $60 \cdot 10^9$ partes) Portanto, ainda assumindo o modelo Uniforme discreto, temos que a probabilidade de encontrarmos, no momento da falha, tal ponteiro em um sítio fixado qualquer é de apenas $1/(60 \cdot 10^9)$. Em tal situação, visto que é extremamente improvável que encontremos o ponteiro em um sítio específico, talvez seja mais razoável nos indagarmos sobre a probabilidade de o encontrarmos em determinadas *regiões* do relógio; por exemplo, qual é a probabilidade de o ponteiro se encontrar no primeiro quadrante (quer dizer, entre $0°$ e $90°$), no momento da falha? Como o relógio tem quatro quadrantes, uma resposta razoável é que a probabilidade requerida deva ser igual a $1/4$. Vejamos como obter essa resposta de maneira rigorosa. Considere dois sítios x_i e x_j quaisquer, com $0° \leq x_i < x_j < 360°$. Então, vale a igualdade

$$\mathbb{P}[x_i < X \leq x_j] = \frac{x_j - x_i}{360}. \tag{1.1}$$

De fato, é fácil verificar que (faça a conta!)

$$\frac{x_j - x_i}{360} = \frac{j - i}{60 \cdot 10^9}$$

e que essa quantidade corresponde, precisamente, ao número de sítios compreendidos entre x_i (exclusive) e x_j (inclusive), dividido pela quantidade total de sítios. Assim,

$$\mathbb{P}[0° < X \leq 90°] = \frac{1}{4}.$$

Note ainda que $\mathbb{P}[0° \leq X \leq 90°] = 1/4 + 1/(60 \cdot 10^9)$, e consequentemente incluir ou não o sítio $0°$ na nossa definição de "primeiro quadrante" faz pouca diferença: as probabilidades obtidas diferem por somente $1/(60 \cdot 10^9)$. É importante perceber que a equação (1.1) vale independentemente da precisão que estamos considerando. Quer dizer, a relação ali expressa vale para qualquer relógio cujo ponteiro dos segundos 'salta' entre sítios equiespaçados $x_0, x_1, \ldots, x_N$.

Passemos agora ao caso do relógio cujo ponteiro dos segundos se desloca continuamente. Uma primeira questão a ser colocada seria a seguinte: quais são os possíveis sítios ocupados pelo ponteiro? Uma resposta razoável é que os possíveis valores assumidos por X são quaisquer ângulos maiores ou iguais a $0°$ e menores do que $360°$. Isto é, o conjunto dos possíveis sítios deve ser o intervalo $[0°, 360°)$. Agora, um resultado matemático profundo nos diz que *é impossível enumerar todos os elementos de um intervalo de números reais!* Em particular, se formarmos uma lista *arbitrária* de números reais pertencentes ao intervalo $[0°, 360°)$, digamos $\{a_1, a_2, a_3 \ldots\}$, então algum número b em $[0°, 360°)$ não estará nessa lista! Quer dizer, $b \neq a_1$, $b \neq a_2$, e assim por diante[1]. Esse fato demonstra a inadequação do conceito de variável

[1] Você pode se perguntar o que ocorreria se adicionássemos o número b à nossa lista original. Não adiantaria: existiria algum número c no intervalo $[0°, 360°)$ que estaria fora dessa nova lista. Esse procedimento nunca acabaria!

aleatória discreta para descrever fenômenos como aquele que propusemos aqui.

De que maneira, então, podemos propor um modelo para descrever as probabilidades associadas à variável aleatória X quando o ponteiro dos relógios se move continuamente? O ponto de partida será a equação (1.1). Dados quaisquer números a e b, com $0° \leq a < b < 360°$, podemos *definir*

$$\mathbb{P}[a < X \leq b] = \frac{b-a}{360}. \tag{1.2}$$

Note que, nesse caso, dado qualquer número x no intervalo $[0°, 360°)$, temos que $\mathbb{P}[X = x] = 0$. Isto é, a probabilidade de encontrarmos o ponteiro, no instante em que o relógio falha, em um sítio específico qualquer, é nula! Isso pode soar estranho à primeira vista, mas lembre-se que nosso relógio mede o tempo com precisão *arbitrária*. Note também que dessa forma conseguimos incorporar a ideia inicial de que o momento em que a pilha se esgota 'não privilegia' nenhum instante em particular: de fato, a probabilidade acima depende apenas do tamanho do arco que vai de a até b, mas não da posição particular desse arco sobre o círculo: se transladarmos ambos os números, a e b, pela mesma quantidade, a probabilidade na equação (1.2) permanece inalterada. Por exemplo,

$$\mathbb{P}[7° < X \leq 97°] = \frac{97-7}{360} = \frac{90-0}{360} = \mathbb{P}[0° < X \leq 90°].$$

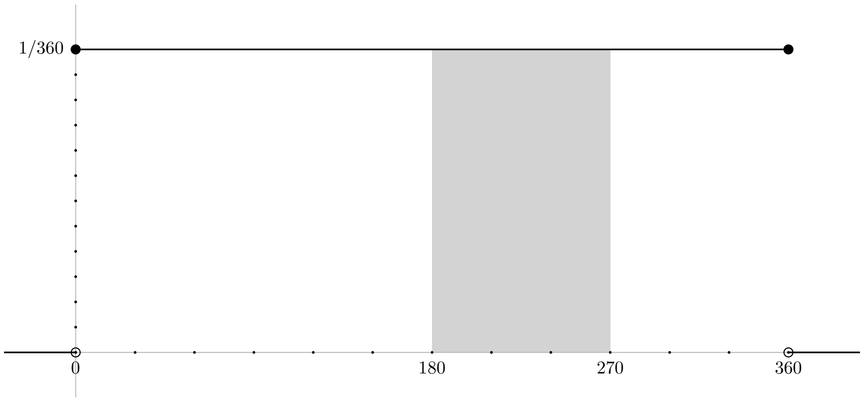
Vamos agora definir uma função real de uma variável real, que denotaremos por f_X e chamaremos de **função densidade de probabilidade de X**. Essa função será utilizada para calcular probabilidades associadas à variável aleatória X. No presente exemplo, f_X é assim definida:

$$f_X(x) = \begin{cases} \frac{1}{360} & \text{se } 0 \leq x \leq 360 \\ 0 & \text{se } x < 0 \text{ ou } x > 360. \end{cases} \tag{1.3}$$

A Figura 1 mostra o gráfico da função f_X. Em cinza vemos a região compreendida abaixo do gráfico de f_X e acima do eixo das abscissas, entre os números $a = 180$ e $b = 270$. Evidentemente, a área dessa região é igual a $(270-180) \times 1/360 = 1/4$, a qual corresponde precisamente a $\mathbb{P}[180 < X \leq 270]$. De fato, o mesmo procedimento pode ser usado para calcular $\mathbb{P}[a < X \leq b]$ para valores arbitrários de a e b (desde que $0 \leq a < b \leq 360$). Como veremos em seguida, a propriedade de que $\mathbb{P}[a < X \leq b]$ corresponde a uma área abaixo do gráfico de *certa* função f_X é precisamente o que caracteriza as variáveis aleatórias contínuas.//

Variáveis Aleatórias Contínuas. Vamos agora introduzir o conceito de variável aleatória contínua. Para isso, precisaremos de um pouco mais de notação: considere

Figura 1: Gráfico de f_X.



uma função não negativa de uma variável real, digamos f (isto é, f é uma função cujo domínio e contradomínio são $\mathbb{R}$, e é tal que $f(x) \geq 0$ para todo x real). Se a e b são dois números reais quaisquer tais que $a < b$,

> **utilizamos o símbolo $\int_a^b f(x)\,dx$ para denotar a área da região compreendida abaixo do gráfico de f e acima do eixo das abscissas, entre os números a e b.**

Em um curso de Cálculo, você irá aprender que esse símbolo é chamado de *integral da função f no intervalo* $[a, b]$. Você também vai aprender a calcular essa área quando a expressão algébrica de f estiver dada. Por exemplo, se $f(x) = x^2$, então você irá aprender a determinar que $\int_0^1 f(x)\,dx = 1/3$. Note que você já sabe calcular esse valor em alguns casos especiais. Por exemplo, para a função f_X da equação (1.3), temos $\int_a^b f_X(x)\,dx = (b-a)/360$, desde que $0 \leq a < b \leq 360$. Você também já sabe calcular essas áreas quando a função f é linear. Como um **exercício**, tente determinar $\int_0^1 f(x)\,dx$ no caso $f(x) = 1+x$. Uma observação importante aqui é que permitimos que $a = -\infty$ e $b = +\infty$: então, por exemplo, escrever $\int_{-\infty}^{\infty} f(x)\,dx = 1$ significa que a área *total* compreendida abaixo do gráfico da função f e acima do eixo das abscissas, é igual a 1.

Podemos agora passar à definição formal de uma variável aleatória contínua. Dizemos que X é uma **variável aleatória contínua** se X é uma função real cujo domínio é o espaço amostral Ω, tal que *para alguma* função não negativa f_X satisfazendo $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$, valha a seguinte igualdade:

$$\mathbb{P}[a < X \leq b] = \int_a^b f_X(x)\,dx, \tag{1.4}$$

para quaisquer números reais a e b tais que $a < b$. *Formalismos à parte*, podemos pensar uma variável aleatória contínua, digamos X, da seguinte maneira:

1. X é um característico numérico de um experimento aleatório;

2. somente temos interesse em calcular probabilidades do tipo $\mathbb{P}[a < X \leq b]$, onde $a < b$ são números reais;

3. podemos calcular essas probabilidades utilizando *alguma* função f_X, através da área $\int_a^b f_X(x)\,dx$.

A função f_X é chamada a **função densidade de probabilidade** de X. Variáveis aleatórias contínuas, portanto, são aquelas variáveis aleatórias para as quais utilizamos sua função densidade de probabilidade para calcular probabilidades de interesse, através da equação (1.4). Os diferentes modelos probabilísticos para variáveis aleatórias contínuas se resumem, consequentemente, a diferentes escolhas de f_X. Se você 'desenhar' uma função não negativa, tal que a área total abaixo de seu gráfico seja igual a 1, então podemos dizer que essa função recém desenhada representa *algum* modelo probabilístico de uma v.a. contínua X. É importante ressaltar mais uma vez que, quando estudamos variáveis aleatórias contínuas, o tipo de probabilidade que estaremos interessados em calcular serão como as acima: a probabilidade de a variável em questão se situar em alguma *região* (intervalos, uniões de intervalos, etc). Essa mudança de enfoque é necessária porque variáveis aleatórias contínuas possuem massa zero em qualquer ponto x. De fato, se x é *qualquer* número real, então $\mathbb{P}[X = x]$ é igual à área de uma figura geométrica cuja base tem comprimento zero, e portanto $\mathbb{P}[X = x] = 0$.

2 Alguns Modelos Probabilísticos Contínuos

Quando estudamos variáveis aleatórias *discretas*, alguns modelos – devido à sua importância – recebem um nome especial e merecem ser estudados com maior detalhe. Por exemplo, aprendemos sobre os modelos Uniforme discreto, Bernoulli, Binomial, Poisson, etc. Em cada um desses casos, determinar o modelo se resumia a determinar a função massa de probabilidade das variáveis aleatórias em questão. Isto é, cada modelo probabilístico discreto corresponde a uma expressão algébrica específica para a função massa de probabilidade. Por exemplo, uma variável aleatória discreta X é dita Poisson com parâmetro $\lambda > 0$ se sua FMP é dada por[2]

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots \tag{2.1}$$

Utilizamos a notação

$$X \sim \text{Poisson}(\lambda)$$

[2]Lembre que $e \cong 2{,}71828$.

para resumir esse fato; quer dizer, a expressão acima deve ser lida como "X é uma variável aleatória Poisson com parâmetro λ". Agora note que, na verdade, a equação (2.1) determina uma família de modelos: cada escolha para o parâmetro λ determina uma diferente FMP. Dito de outra forma, se sabemos que X é uma variável aleatória Poisson com parâmetro λ e quisermos calcular, digamos, $\mathbb{P}[X = 1]$, então para que obtenhamos o valor *numérico* correspondente a essa probabilidade é preciso atribuir um valor numérico específico para o parâmetro λ. Por exemplo, se $X \sim \text{Poisson}(2)$, então essa probabilidade é igual a 0,2706706.

No caso de variáveis aleatórias *contínuas*, essa estrutura de apresentação irá se repetir. A diferença é que, agora, os diferentes modelos probabilísticos correspondem a diferentes expressões algébricas para a função *densidade* de probabilidade. Aqui, como lá, essas expressões vão depender de certos parâmetros. Consequentemente, cada um dos modelos abaixo apresentados corresponde, na verdade, a uma família *paramétrica* de modelos.

Um último comentário sobre semelhanças e diferenças entre modelos discretos e contínuos: muitas vezes, modelos discretos podem ser deduzidos diretamente da estrutura do problema proposto. Por exemplo, se considerarmos a variável aleatória X = "número de caras em 10 lançamentos de uma moeda honesta", então com algum esforço e noções básicas de análise combinatória obteremos a função massa de probabilidade correspondente ao modelo Binomial($n = 10$, $p = 1/2$). No caso de variáveis aleatórias contínuas, os modelos são construções puramente teóricas (as densidades de probabilidade são meramente funções para as quais se tem uma fórmula dada). A adequação desses modelos ao estudo de fenômenos concretos decorre, por outro lado, de um conhecimento empírico: por exemplo, sabe-se que a altura de indivíduos em uma população é bem descrita por um modelo Normal; sabe-se que o tempo de duração de uma lâmpada é bem descrito por um modelo Exponencial, etc. Ressaltamos ainda que, embora não tenhamos introduzido a noção de valor esperado e variância para variáveis aleatórias contínuas (precisaríamos de ferramentas do Cálculo para isso), a interpretação dessas quantidades é análoga ao caso discreto: $\mathbb{E}(X)$ é um parâmetro de posição da variável aleatória X, e $\text{Var}(X)$ é um parâmetro de dispersão dessa variável.

2.1 O Modelo Normal

Dizemos que uma variável aleatória contínua X é **Normal com parâmetros μ e $\sigma^2 > 0$** se sua função densidade de probabilidade é dada por

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.2}$$

Notação:

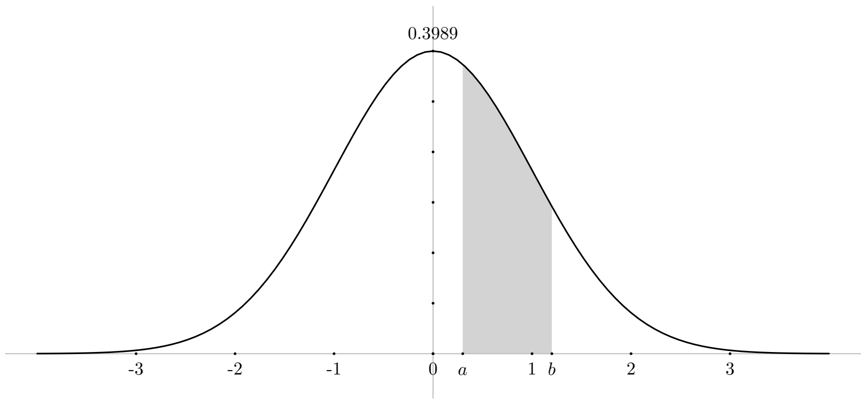
$$X \sim N\Big(\mu,\, \sigma^2\Big)$$

Quando $\mu = 0$ e $\sigma^2 = 1$, dizemos que X é **Normal padrão**. A Figura 2 mostra o gráfico da função densidade de probabilidade f_X de uma v.a. Normal padrão. O gráfico também mostra, em cinza, a região compreendida abaixo do gráfico de f_X e acima do eixo das abscissas, entre os números $a = 0{,}3$ e $b = 1{,}2$. A área dessa região é igual[3] a 0,2670189, isto é,

$$\int_{0,3}^{1,2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 0{,}2670189.$$

Portanto, se X é uma v.a. Normal padrão, podemos usar a equação (1.4) para concluir que $\mathbb{P}[0{,}3 < X \leq 1{,}2] = 0{,}2670189$. Alterando-se os valores de a e b, podemos assim obter o valor numérico de $\mathbb{P}[a < X \leq b]$ em qualquer caso de nosso interesse. Por fim, indicamos (sem demonstrar) que, se X é uma variável aleatória Normal com parâmetros μ e σ^2, então $\mathbb{E}(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$.

Figura 2: A densidade Normal padrão.



2.2 O Modelo Exponencial

Dizemos que uma variável aleatória contínua X é **Exponencial com parâmetro** $\lambda > 0$ se sua função densidade de probabilidade é dada por

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0. \end{cases} \tag{2.3}$$

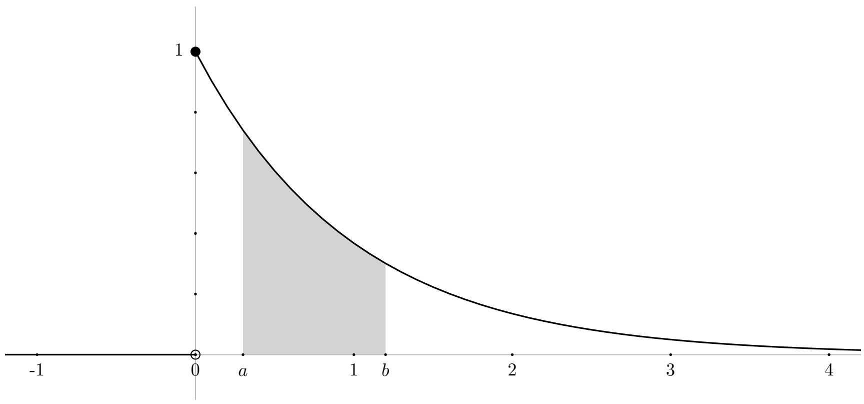
Notação:

$$X \sim \text{Exponencial}(\lambda)$$

[3]Na verdade, esse número é uma aproximação. O cálculo dá área é feito por um programa de computador, através de uma aproximação do gráfico de f_X por um polígono.

A Figura 3 mostra o gráfico da função densidade de probabilidade f_X de uma v.a. $X \sim$ Exponencial(1). O gráfico também mostra, em cinza, a região compreendida abaixo do gráfico de f_X e acima do eixo das abscissas, entre os números $a = 0{,}3$ e $b = 1{,}2$. A área dessa região é igual a 0,439624. Portanto, $\mathbb{P}[0{,}3 < X \leq 1{,}2] = 0{,}439624$. Compare com o exemplo semelhante dado no modelo Normal padrão. Por fim, indicamos (sem demonstrar) que, se X é uma variável aleatória Exponencial com parâmetro λ, então $\mathbb{E}(X) = 1/\lambda$ e $\mathrm{Var}(X) = 1/\lambda^2$.

Figura 3: A densidade Exponencial(1).



2.3 O Modelo Uniforme

Dizemos que uma variável aleatória contínua X é **Uniforme com parâmetros** α **e** β (onde α e β são números reais com $\alpha < \beta$) se sua função densidade de probabilidade é dada por

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{se } \alpha \leq x \leq \beta \\ 0 & \text{se } x < \alpha \text{ ou } x > \beta. \end{cases} \tag{2.4}$$

Notação:

$$X \sim \text{Uniforme}(\alpha,\, \beta)$$

No Exemplo 1.1, quando consideramos o relógio cujo ponteiro se desloca continuamente, a variável aleatória X = "ângulo formado pelo ponteiro dos segundos no instante em que o relógio parar" é, portanto, uma v.a. Uniforme com parâmetros $\alpha = 0$ e $\beta = 360$. Na Figura 1 temos o gráfico da função densidade de probabilidade correspondente. Por fim, indicamos (sem demonstrar) que, se X é uma variável aleatória Uniforme com parâmetros α e β, então $\mathbb{E}(X) = (\alpha + \beta)/2$ e $\mathrm{Var}(X) = (\beta - \alpha)^2/12$.