

Da densidade à probabilidade e à máxima verossimilhança

Este material explica por que multiplicamos densidades na máxima verossimilhança e como esse produto se relaciona com a probabilidade de observar valores próximos aos dados. Usaremos uma distribuição normal e gráficos gerados em Python.

1. Primeiro: separar variável aleatória e valor observado

Antes de medir, o resultado é incerto: representamos essa quantidade por uma variável aleatória, como X_i . Depois de medir, obtemos um número, representado por x_i .

Símbolo	Significado	Exemplo
X_i	Variável aleatória da i -ésima medida	Resultado ainda desconhecido
x_i	Valor observado dessa medida	$x_1 = 8$
$\mathbf{X} = (X_1, \dots, X_n)$	Vetor de variáveis aleatórias	Três medidas a realizar
$\mathbf{x} = (x_1, \dots, x_n)$	Vetor dos dados observados	$(8, 10, 12)$
μ	Parâmetro do modelo	Média populacional desconhecida
$f(x \mid \mu)$	Densidade avaliada em x , para uma média candidata μ	Altura da curva naquele ponto

Escrever $f(x \mid \mu)$ significa avaliar o modelo para o valor candidato μ . Isso não exige tratar μ como variável aleatória: na abordagem de máxima verossimilhança, ele é um parâmetro fixo, mas desconhecido.

2. Densidade é altura; probabilidade é área

Para uma variável com distribuição contínua descrita por uma densidade:

$$P(a \leq X \leq b \mid \mu) = \int_a^b f(t \mid \mu) dt.$$

A função f indica a concentração de probabilidade. O valor $f(x)$, sozinho, não é a probabilidade de observar exatamente x .

$P(X = x \mid \mu) = 0$

Isso acontece porque um único ponto tem largura zero e, portanto, área zero sob a curva. Probabilidade zero não significa que o valor esteja fora dos resultados possíveis: alguma medida será observada, embora cada valor exato tenha probabilidade zero no modelo contínuo.

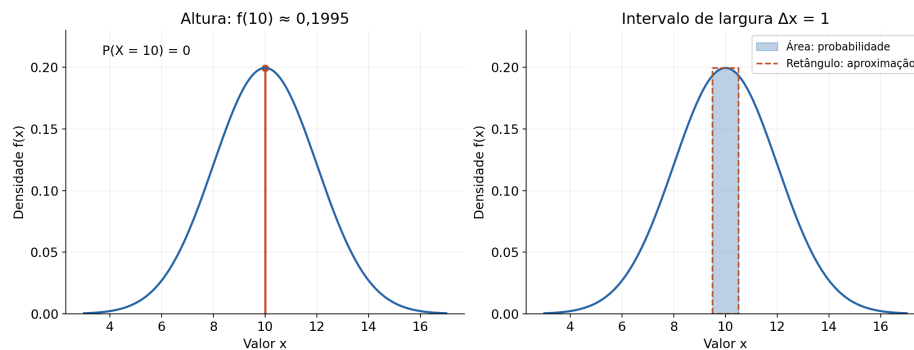


Figure 1: Altura da densidade em 10 e área de um intervalo ao redor de 10.

Como ler a figura: à esquerda, o segmento vertical representa a altura $f(10)$. À direita, a região azul representa uma probabilidade. O retângulo tracejado aproxima essa área.

Para $X \sim \mathcal{N}(10, 2^2)$:

$$f(10) = \frac{1}{2\sqrt{2\pi}} \approx 0,19947.$$

Esse número não significa “19,947% de probabilidade de obter exatamente 10”.

Uma densidade pode superar 1. Por exemplo, na distribuição uniforme em $[0, 0,5]$, a densidade é 2 dentro desse intervalo, mas a área total é $2 \times 0,5 = 1$. Uma probabilidade sempre fica entre 0 e 1.

Se a variável representa uma medida em centímetros, a densidade tem unidade cm^{-1} . Multiplicá-la por uma largura em centímetros produz uma quantidade sem unidade, como uma probabilidade deve ser.

3. Um intervalo pequeno transforma altura em probabilidade aproximada

Considere um intervalo centrado em x , de largura total Δx :

$$I_x = \left[x - \frac{\Delta x}{2}, x + \frac{\Delta x}{2} \right].$$

Se a densidade varia pouco nesse intervalo, podemos aproximar a área por um retângulo:

$$P(X \in I_x \mid \mu) \approx f(x \mid \mu) \Delta x.$$

É a regra geométrica: **área aproximada = altura $\times$ largura**.

Para a normal de média 10 e desvio padrão 2:

Largura Δx	Intervalo	Probabilidade exata	Aproximação $f(10)\Delta x$
1	[9,5; 10,5]	0,19741265	0,19947114
0,1	[9,95; 10,05]	0,01994504	0,01994711
0,01	[9,995; 10,005]	0,0019947093	0,0019947114

A aproximação melhora conforme a largura diminui neste exemplo. Para largura 0,1, a probabilidade é aproximadamente 0,019945, ou **1,9945%**.

Não se deve usar essa aproximação indiscriminadamente para intervalos largos ou regiões em que a densidade varia muito. A integral fornece o valor exato.

4. Várias observações: onde entra a independência?

Suponha que observamos:

$$\mathbf{x} = (8, 10, 12).$$

Assumimos que X_1, X_2, X_3 são independentes sob o modelo, para cada valor fixado de μ , e têm distribuição normal com desvio padrão conhecido $\sigma = 2$.

São as variáveis aleatórias que assumimos independentes; as densidades são funções.

Defina um pequeno intervalo I_i ao redor de cada x_i . Pela independência:

$$P(X_1 \in I_1, \dots, X_n \in I_n \mid \mu) = \prod_{i=1}^n P(X_i \in I_i \mid \mu).$$

Essa igualdade é exata. Aplicando a aproximação do retângulo em cada intervalo de mesma largura Δx :

$$P(X_1 \in I_1, \dots, X_n \in I_n \mid \mu) \approx \prod_{i=1}^n [f(x_i \mid \mu) \Delta x].$$

Como há uma largura multiplicada para cada observação:

$$P(X_1 \in I_1, \dots, X_n \in I_n \mid \mu) \approx \left[\prod_{i=1}^n f(x_i \mid \mu) \right] (\Delta x)^n.$$

Para três observações e $\Delta x = 0,1$, o fator de largura é $(0,1)^3 = 0,001$. Ele representa o volume da pequena região no espaço dos três resultados.

5. A verossimilhança é a densidade conjunta nos dados observados

Com independência, a densidade conjunta se fatoriza exatamente:

$$f_{\mathbf{X}}(\mathbf{x} \mid \mu) = \prod_{i=1}^n f(x_i \mid \mu).$$

Ao manter $\mathbf{x}$ fixo e considerar essa expressão como função do parâmetro, temos a verossimilhança:

$$L(\mu; \mathbf{x}) = \prod_{i=1}^n f(x_i \mid \mu).$$

Expressão	O que representa?
$P(\mathbf{X} = \mathbf{x} \mid \mu)$	Probabilidade do vetor exato: zero neste modelo contínuo
$L(\mu; \mathbf{x})$	Densidade conjunta avaliada nos dados, vista como função de μ
$L(\mu; \mathbf{x})(\Delta x)^n$	Aproximação da probabilidade de todos os valores caírem nos respectivos intervalos

A definição de verossimilhança usa a densidade conjunta diretamente. A aproximação por intervalos fornece uma interpretação intuitiva dessa definição.

6. Por que a largura não altera o máximo da expressão aproximada?

Fixamos intervalos de mesma largura positiva, que não dependem de μ . Assim, $(\Delta x)^n$ é uma constante em relação ao parâmetro.

Multiplicar uma função por uma constante positiva muda sua escala vertical, mas não o ponto onde ela atinge o máximo:

$$\arg \max_{\mu} [L(\mu)(\Delta x)^n] = \arg \max_{\mu} L(\mu).$$

Aqui, “arg max” devolve o **valor do parâmetro**, e não a altura máxima da curva.

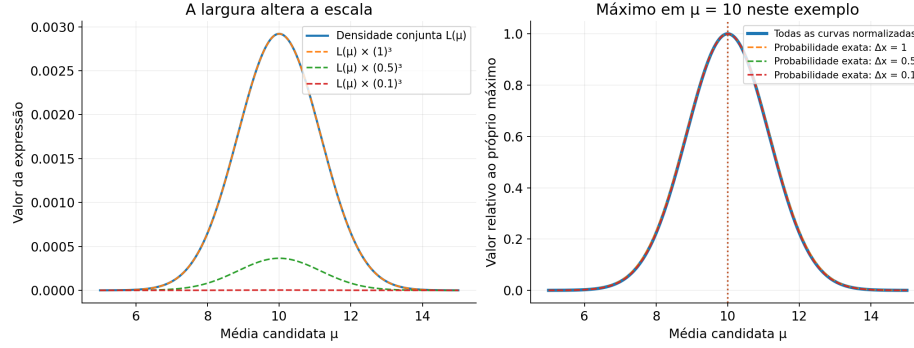


Figure 2: Efeito da largura na escala das curvas e comparação com probabilidades exatas de intervalos.

Como ler a figura: à esquerda, fatores de largura diferentes reduzem as alturas, mas mantêm o pico no mesmo lugar. À direita, as curvas foram divididas pelos próprios máximos para facilitar a comparação. As probabilidades exatas dos intervalos também têm máximo em 10 neste exemplo simétrico.

Distinção importante: a igualdade dos máximos de $L(\mu)$ e $L(\mu)(\Delta x)^n$ é exata. Já a probabilidade dos intervalos é apenas aproximada por esse produto. Para intervalos de largura finita, maximizar a probabilidade exata dos intervalos pode produzir uma estimativa diferente da máxima verossimilhança baseada nas densidades, em outros modelos ou dados.

Se as larguras forem diferentes, mas fixas em relação ao parâmetro, o fator será $\prod_i \Delta x_i$ e continuará sem alterar o máximo da expressão aproximada. Se as larguras dependerem do parâmetro, esse argumento de constância não se aplica.

7. Exemplo numérico: dados 8, 10 e 12

Para a normal com $\sigma = 2$, a densidade é:

$$f(x | \mu) = \frac{1}{2\sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{8} \right].$$

Logo:

$$L(\mu) = f(8 | \mu)f(10 | \mu)f(12 | \mu).$$

Com intervalos de largura $\Delta x = 0,1$:

Média candidata μ	Densidade conjunta $L(\mu)$	Probabilidade aproximada $L(\mu)(0,1)^3$
6	$7,23734 \times 10^{-6}$	$7,23734 \times 10^{-9}$
10	$2,91975 \times 10^{-3}$	$2,91975 \times 10^{-6}$
14	$7,23734 \times 10^{-6}$	$7,23734 \times 10^{-9}$

O fator 0,001 reduz todos os valores na mesma proporção. Portanto, a comparação entre candidatos é preservada.

A probabilidade conjunta pode ser pequena porque exigimos que **todas as três medidas** caiam simultaneamente em faixas estreitas específicas. Isso não impede que ela seja usada para comparar modelos.

8. Derivação: por que a média amostral maximiza a verossimilhança normal?

Para n observações independentes normais, com σ conhecido:

$$L(\mu) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right].$$

Definindo

$$k(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

temos $k(\mu) \leq 0$, pois a soma dos quadrados é não negativa. Assim,

$$e^{k(\mu)} = e^{-|k(\mu)|} = \frac{1}{e^{|k(\mu)|}}.$$

Quanto menor for $|k(\mu)|$, maior será essa expressão. Como $\sigma > 0$ é conhecido e fixo, minimizar $|k(\mu)|$ equivale a minimizar a soma dos quadrados das distâncias entre as observações e a média candidata. Vamos definir

$$S(\mu) = \sum_{i=1}^n (x_i - \mu)^2.$$

Portanto, maximizar a verossimilhança equivale a minimizar $S(\mu)$. Em termos do expoente, buscamos **maximizar** $k(\mu)$, **tornando-o o mais próximo possível de zero**.

Derivando:

$$S'(\mu) = 2 \sum_{i=1}^n (\mu - x_i) = 2 \left(n\mu - \sum_{i=1}^n x_i \right).$$

Igualando a zero:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Como $S''(\mu) = 2n > 0$, esse ponto é um mínimo de S e, consequentemente, um máximo de L .

Para os dados do exemplo:

$$\hat{\mu} = \frac{8 + 10 + 12}{3} = 10.$$

O expoente máximo precisa ser zero?

Não. Tornar o expoente o mais próximo possível de zero significa encontrar seu maior valor permitido pelos dados. Esse valor pode continuar sendo negativo.

Para os dados 8, 10, 12, com $\sigma = 2$, a soma dos quadrados é mínima em $\mu = 10$:

$$S(10) = (8 - 10)^2 + (10 - 10)^2 + (12 - 10)^2 = 4 + 0 + 4 = 8.$$

Assim,

$$k(10) = -\frac{S(10)}{2\sigma^2} = -\frac{8}{2 \cdot 2^2} = -1.$$

Portanto, **o maior valor possível do expoente é -1 neste exemplo**, e a maior exponencial é $e^{-1} \approx 0,36788$. Esse valor corresponde apenas ao fator exponencial da verossimilhança: o fator constante $(1/(\sigma\sqrt{2\pi}))^n$ também faz parte de $L(\mu)$.

O expoente seria zero somente se $S(\mu) = 0$, isto é, se **todas as observações fossem iguais à média candidata**. Como 8, 10 e 12 são diferentes entre si, nenhuma escolha de μ consegue zerar todas as distâncias simultaneamente.

9. E se as observações forem dependentes?

Ainda usamos a densidade conjunta para definir a verossimilhança, mas ela não pode, em geral, ser substituída pelo produto das densidades individuais.

Pela regra da cadeia, para três observações:

$$f(x_1, x_2, x_3 \mid \mu) = f(x_1 \mid \mu) f(x_2 \mid x_1, \mu) f(x_3 \mid x_1, x_2, \mu).$$

A independência permite retirar os valores anteriores desses condicionamentos. Sem ela, precisamos modelar a dependência, por exemplo por uma matriz de covariâncias em um modelo normal multivariado.

A aproximação local continua tendo a forma “densidade conjunta $\times$ volume”, mas a densidade conjunta pode não se fatorizar em densidades individuais.

10. Python: comparar densidade, aproximação e probabilidade exata

Instale as bibliotecas, se necessário:

```
pip install numpy scipy matplotlib
```

```
import numpy as np
from scipy.stats import norm

# Dados observados e características do modelo
dados = np.array([8.0, 10.0, 12.0])
sigma = 2.0
delta_x = 0.1

for mu in [6.0, 10.0, 14.0]:
    # Verossimilhança: densidade conjunta para dados independentes
    densidades = norm.pdf(dados, loc=mu, scale=sigma)
    L = np.prod(densidades)

    # Probabilidade aproximada para as pequenas faixas
    prob_aproximada = L * delta_x**len(dados)

    # Probabilidade exata de cada intervalo: diferença de CDFs
    inferiores = dados - delta_x / 2
    superiores = dados + delta_x / 2
    probs_intervalos = (
        norm.cdf(superiores, loc=mu, scale=sigma)
        - norm.cdf(inferiores, loc=mu, scale=sigma)
    )

    # Produto exato graças à independência
```

```

prob_exata = np.prod(probs_intervalos)

print(f"Média candidata: {mu:.1f}")
print(f"Verossimilhança:      {L:.8e}")
print(f"Probabilidade aprox.: {prob_aproximada:.8e}")
print(f"Probabilidade exata:   {prob_exata:.8e}")
print()

print("Estimativa de máxima verossimilhança:", dados.mean())

norm.pdf avalia a densidade. norm.cdf calcula a probabilidade acumulada até
um valor. A diferença de duas CDFs fornece a probabilidade de um intervalo.

Para amostras grandes, prefira somar log-densidades, usando norm.logpdf,
porque produtos de muitos números pequenos podem sofrer arredondamento
para zero no computador:

mu = 10.0
log_L = np.sum(norm.logpdf(dados, loc=mu, scale=sigma))

```

11. A formulação correta da ideia central

Para observações independentes, a probabilidade de todos os resultados caírem em pequenos intervalos ao redor dos dados é aproximadamente o produto das densidades avaliadas nos dados, multiplicado pelo produto das larguras. Quando essas larguras são fixas em relação ao parâmetro, elas não alteram o máximo da expressão aproximada. Por isso, a máxima verossimilhança maximiza a densidade conjunta avaliada nos dados observados.

Isso evita três confusões: densidade não é probabilidade; o vetor exato tem probabilidade zero no modelo contínuo; e quem é independente são as variáveis aleatórias.

Arquivos e visualização

O arquivo gerar_figuras.py permite reproduzir as figuras com NumPy e Matplotlib.