Prática em R - Aula 04

A Distribuição Normal como modelo generativo de alturas

Prof. Fabio Cop (*fcferreira@unifesp.br*)

Instituto do Mar - Unifesp

2026-03-29

Conteúdo

Curso: Bacharelado Interdisciplinar em Ciências do Mar
Unidade Curricular (UC): Probabilidade e Estatística
Atividade: Prática no Laboratório de Informática

1 A Distribuição Normal — FDP e parâmetros

Nas aulas anteriores, a função massa de probabilidade (FMP) atribuía uma probabilidade pontual $P(X = k)$ a cada valor inteiro k. Em variáveis contínuas, essa abordagem não é possível: a probabilidade de qualquer valor exato é zero. A ferramenta adequada é a função densidade de probabilidade (FDP), $f(x)$, que organiza a probabilidade em termos de área sob a curva.

A Distribuição Normal é a FDP central para variáveis contínuas simétricas. Sua FDP é:

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

A função `dnorm(x, mean, sd)` calcula $f(x)$ em R. O valor retornado é a densidade no ponto x e pode ser maior que 1, pois densidade não é probabilidade. A probabilidade está sempre na área sob a curva, calculada entre dois limites.

Código

```
# Grade de valores no eixo x para uma Normal(160, 10) - alturas em cm
x <- seq(120, 200, length.out = 400)
fx <- dnorm(x, mean = 160, sd = 10)

# Gráfico da FDP
plot(x, fx, type = "l", lwd = 2, col = "steelblue",
     xlab = "Altura (cm)", ylab = "Densidade f(x)",
     main = "Função densidade de probabilidade - Normal(160, 10)")
abline(v = 160, lty = 2, col = "gray50")   # ponto de maxima densidade

# Densidade no ponto de máximo (x = mu)
dnorm(160, mean = 160, sd = 10)

# Densidade com sigma muito pequeno: o valor pode ultrapassar 1
# A área total sob a curva continua sendo exatamente 1
dnorm(160, mean = 160, sd = 0.5)
```

O que observar

- Observe o valor retornado por `dnorm(160, mean = 160, sd = 10)` e o valor retornado por `dnorm(160, mean = 160, sd = 0.5)`. Onde cada um desses valores aparece no gráfico plotado?
- Qual propriedade matemática da função densidade de probabilidade garante que a área total sob a curva permaneça igual a 1, mesmo quando o valor de $f(x)$ em determinado ponto é muito pequeno ou muito grande?

1.1 Efeito de μ e σ na curva

O parâmetro μ controla a localização da curva: alterá-lo desloca toda a curva horizontalmente sem mudar sua forma. O parâmetro σ controla a dispersão: valores menores produzem curvas mais estreitas e altas, valores maiores produzem curvas mais largas e baixas. Em todos os casos, a área total sob a curva permanece igual a 1.

Código

```r
# Grade de valores
x <- seq(110, 220, length.out = 500)

# Distribuição de referência: Normal(160, 10)
fx_ref <- dnorm(x, mean = 160, sd = 10)

# Efeito de mu: deslocar para mu = 175, mantendo sigma = 10
fx_mu2 <- dnorm(x, mean = 175, sd = 10)

# Efeito de sigma: sigma = 5, 10 e 20, com mu = 160 fixo
fx_s1 <- dnorm(x, mean = 160, sd = 5)
fx_s2 <- dnorm(x, mean = 160, sd = 20)

par(mfrow = c(1, 2))

# Painel esquerdo: efeito de mu (sigma fixo = 10)
plot(x, fx_ref, type = "l", lwd = 2, col = "steelblue",
     ylim = c(0, 0.085),
     xlab = "Altura (cm)", ylab = "Densidade",
     main = expression("Efeito de " * mu ~ "(sigma == 10)"))
lines(x, fx_mu2, col = "orange", lwd = 2)
legend("topright",
       legend = c(expression(mu == 160), expression(mu == 175)),
       col = c("steelblue", "orange"), lwd = 2, bty = "n")

# Painel direito: efeito de sigma (mu fixo = 160)
plot(x, fx_ref, type = "l", lwd = 2, col = "steelblue",
     ylim = c(0, 0.085),
     xlab = "Altura (cm)", ylab = "Densidade",
     main = expression("Efeito de " * sigma ~ "(mu == 160)"))
lines(x, fx_s1, col = "darkgreen", lwd = 2)
lines(x, fx_s2, col = "red", lwd = 2)
legend("topright",
       legend = c(expression(sigma == 10),
                  expression(sigma == 5),
                  expression(sigma == 20)),
       col = c("steelblue", "darkgreen", "red"), lwd = 2, bty = "n")

par(mfrow = c(1, 1))
```

O que explorar

- Altere μ para 145 cm e 180 cm. O que muda na curva além da posição horizontal?
- Compare os três valores de σ (5, 10, 20): como a altura máxima da curva e a sua largura se relacionam? A área total se altera?

---

2 Probabilidades de intervalo com `pnorm()` e quantis com `qnorm()`

A probabilidade de X pertencer ao intervalo $[a, b]$ corresponde à área sob a curva entre a e b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = \texttt{pnorm}(b,\, \mu,\, \sigma) - \texttt{pnorm}(a,\, \mu,\, \sigma)$$

A função `pnorm(q, mean, sd)` calcula a probabilidade acumulada $P(X \leq q)$. A função `qnorm(p, mean,`

`sd`) resolve o problema inverso: dado $P(X \leq q) = p$, qual o valor de q?

Código

```r
# Modelo: altura de adultos X ~ Normal(160, 10) em centímetros

# Probabilidade de altura entre 150 e 180 cm
pnorm(180, mean = 160, sd = 10) - pnorm(150, mean = 160, sd = 10)

# Probabilidade de altura acima de 185 cm
1 - pnorm(185, mean = 160, sd = 10)

# Probabilidade de altura abaixo de 135 cm
pnorm(135, mean = 160, sd = 10)

# Visualização: FDP com área de interesse sombreada
x <- seq(120, 200, length.out = 400)
fx <- dnorm(x, mean = 160, sd = 10)

plot(x, fx, type = "l", lwd = 2,
     xlab = "Altura (cm)", ylab = "Densidade",
     main = "Normal(160, 10) - probabilidade de intervalo")

# Sombrear a área entre 150 e 180 cm
x_int <- seq(150, 180, length.out = 200)
polygon(c(150, x_int, 180),
        c(0, dnorm(x_int, mean = 160, sd = 10), 0),
        col = rgb(0.2, 0.5, 0.8, 0.3), border = NA)
abline(v = c(150, 180), lty = 2, col = "gray50")
```

O que observar

- Qual é a probabilidade de altura entre 150 e 180 cm? Que proporção do total essa área representa?
- Compare a probabilidade de altura acima de 185 cm com a probabilidade de altura abaixo de 135 cm. Como a simetria da Distribuição Normal se manifesta nessa comparação?

2.1 A regra empírica 68-95-99,7

A simetria da Distribuição Normal implica uma distribuição característica de probabilidade ao redor de μ: aproximadamente 68% da probabilidade está no intervalo $[\mu - \sigma, \mu + \sigma]$, 95% no intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ e 99,7% no intervalo $[\mu - 3\sigma, \mu + 3\sigma]$. O código abaixo verifica esses valores numericamente com `pnorm()` e identifica os limites dos intervalos com `qnorm()`.

Código

```r
# Verificar a regra empírica para X ~ Normal(160, 10)
mu    <- 160
sigma <- 10

# Probabilidade em ±1, ±2 e ±3 desvios padrões
pnorm(mu + sigma,     mu, sigma) - pnorm(mu - sigma,     mu, sigma)
pnorm(mu + 2 * sigma, mu, sigma) - pnorm(mu - 2 * sigma, mu, sigma)
pnorm(mu + 3 * sigma, mu, sigma) - pnorm(mu - 3 * sigma, mu, sigma)

# Quantis: limites que delimitam 90%, 95% e 99% da distribuição
qnorm(0.90, mu, sigma)   # 90% abaixo deste valor
qnorm(0.95, mu, sigma)   # 95% abaixo deste valor
```

```r
qnorm(0.975, mu, sigma)  # limite superior do intervalo central de 95%

# Intervalo central de 95%: limites inferior e superior
qnorm(c(0.025, 0.975), mu, sigma)
```

O que observar

- Os valores calculados por `pnorm()` para ± 1, ± 2 e ± 3 σ correspondem aos 68%, 95% e 99,7% da regra empírica?
- Os limites retornados por `qnorm(c(0.025, 0.975), mu, sigma)` delimitam exatamente quais alturas? Como esses limites se expressam em função de μ e σ?

---

3 As distribuições a priori para os parâmetros

Na abordagem bayesiana, os parâmetros μ e σ do modelo para alturas não são valores fixos desconhecidos. São variáveis aleatórias com suas próprias distribuições de probabilidade, chamadas distribuições a priori. O modelo generativo completo é:

$$Y \sim \text{Normal}(\mu,\, \sigma)$$
$$\mu \sim \text{Normal}(160,\, 20)$$
$$\sigma \sim \text{Exponencial}(0{,}1)$$

A distribuição a priori Normal$(160, 20)$ para μ concentra a probabilidade em alturas médias biologicamente razoáveis para adultos, cobrindo aproximadamente $[120, 200]$ cm com probabilidade substancial. A distribuição a priori Exponencial(0,1) para σ garante que o parâmetro permaneça estritamente positivo e concentra a probabilidade abaixo de 30 cm, com média de 10 cm.

Código

```r
par(mfrow = c(1, 2))

# Distribuição a priori para mu: Normal(160, 20)
mu_vals <- seq(80, 240, length.out = 400)
plot(mu_vals, dnorm(mu_vals, mean = 160, sd = 20),
     type = "l", lwd = 2, col = "steelblue",
     xlab = expression(mu ~ "(cm)"),
     ylab = "Densidade a priori",
     main = expression("Priori para " ~ mu ~ ": Normal(160, 20)"))
# Faixa de ±2 desvios padrões da priori (abrange ~95% da probabilidade)
abline(v = c(160 - 2 * 20, 160 + 2 * 20), lty = 2, col = "gray50")

# Distribuição a priori para sigma: Exponencial com taxa 0.1
sigma_vals <- seq(0, 60, length.out = 400)
plot(sigma_vals, dexp(sigma_vals, rate = 0.1),
     type = "l", lwd = 2, col = "steelblue",
     xlab = expression(sigma ~ "(cm)"),
     ylab = "Densidade a priori",
     main = expression("Priori para " ~ sigma ~ ": Exponencial(0.1)"))
# Média da distribuição Exponencial: 1/taxa = 10 cm
abline(v = 1 / 0.1, lty = 2, col = "gray50")

par(mfrow = c(1, 1))
```

O que observar

- Para a distribuição a priori de μ: que faixa de valores recebe probabilidade substancial? Os limites marcados correspondem a alturas médias biologicamente possíveis?
- Para a distribuição a priori de σ: a distribuição Exponencial permite valores negativos? Como isso se relaciona com a natureza do parâmetro σ?

3.1 Comparando diferentes escolhas de distribuição a priori para μ

Uma distribuição a priori mais dispersa pode parecer "menos informativa" sobre μ. Visualizar as diferentes opções ajuda a entender o que cada escolha implica no espaço dos parâmetros.

Código

```
# Grade de valores para mu
mu_vals <- seq(-100, 420, length.out = 600)

# Três escolhas de distribuição a priori para mu
prior_20  <- dnorm(mu_vals, mean = 160, sd = 20)
prior_50  <- dnorm(mu_vals, mean = 160, sd = 50)
prior_100 <- dnorm(mu_vals, mean = 160, sd = 100)

# Comparação das três prioris no espaço de mu
plot(mu_vals, prior_100, type = "l", lwd = 2, col = "red",
     xlab = expression(mu ~ "(cm)"), ylab = "Densidade a priori",
     main = expression("Comparação de distribuições a priori para " ~ mu))
lines(mu_vals, prior_50,  col = "orange", lwd = 2)
lines(mu_vals, prior_20,  col = "steelblue", lwd = 2)
abline(v = 0, lty = 3, col = "gray60")    # altura zero
abline(v = 300, lty = 3, col = "gray60")  # 3 metros de altura
legend("topright",
       legend = c(expression(sigma[0] == 100),
                  expression(sigma[0] == 50),
                  expression(sigma[0] == 20)),
       col = c("red", "orange", "steelblue"), lwd = 2, bty = "n")
```

O que explorar

- As linhas verticais marcam $\mu = 0$ e $\mu = 300$. Para cada distribuição a priori, qual a probabilidade de μ cair fora do intervalo $[0, 300]$?
- Qual das três prioris concentra mais probabilidade em valores de μ biologicamente plausíveis para alturas médias adultas?

---

4 Distribuição preditiva a priori

A distribuição preditiva a priori é a distribuição dos valores de Y que o modelo espera gerar antes de observar qualquer dado. Ela responde à pergunta: "Se as distribuições a priori para μ e σ estão corretas, que alturas devo esperar observar?"

Para obtê-la por simulação, repete-se o seguinte algoritmo muitas vezes:

1. Sortear um valor de μ da distribuição a priori Normal$(160, 20)$.
2. Sortear um valor de σ da distribuição a priori Exponencial(0,1).
3. Gerar uma altura Y da distribuição Normal(μ, σ).

O conjunto de valores Y obtidos constitui uma amostra da distribuição preditiva a priori.

Código

```r
set.seed(2026)
N <- 2000

# Passo 1: sortear mu da distribuição a priori Normal(160, 20)
mu_sim <- rnorm(N, mean = 160, sd = 20)

# Passo 2: sortear sigma da distribuição a priori Exponencial(0.1)
sigma_sim <- rexp(N, rate = 0.1)

# Passo 3: gerar uma altura para cada par (mu, sigma)
y_sim <- rnorm(N, mean = mu_sim, sd = sigma_sim)

# Histograma da distribuição preditiva a priori
hist(y_sim, freq = FALSE, breaks = 50,
     xlab = "Altura simulada (cm)", ylab = "Densidade",
     main = "Distribuição preditiva a priori - alturas de adultos",
     col = "lightblue", border = "white")

# Amplitude e casos fora da faixa biologicamente razoável
range(y_sim)
sum(y_sim < 0)     # alturas negativas
sum(y_sim > 300)   # alturas acima de 3 metros
```

O que observar

- Qual é a amplitude dos valores simulados? O modelo gera alturas fora de uma faixa biologicamente razoável para adultos?
- Em que proporção do total surgem valores impossíveis (negativos ou acima de 300 cm)?

4.1 O perigo das distribuições a priori aparentemente não informativas

Uma distribuição a priori Normal$(160, 100)$ para μ parece ampla e pouco restritiva. A distribuição preditiva a priori revela as consequências dessa escolha no espaço dos dados: alturas simuladas completamente fora do intervalo biologicamente possível. A checagem preditiva a priori torna esse problema visível antes de qualquer dado ser analisado.

Código

```r
# Distribuição a priori mais dispersa para mu: Normal(160, 100)
mu_sim_amplo <- rnorm(N, mean = 160, sd = 100)
y_sim_amplo  <- rnorm(N, mean = mu_sim_amplo, sd = sigma_sim)

par(mfrow = c(1, 2))

# Distribuição preditiva a priori - priori original sigma_0 = 20
hist(y_sim, freq = FALSE, breaks = 50,
     xlab = "Altura (cm)", ylab = "Densidade",
     main = expression("Priori: " ~ mu %~% Normal(160, 20)),
     col = "lightblue", border = "white")

# Distribuição preditiva a priori - priori mais dispersa sigma_0 = 100
hist(y_sim_amplo, freq = FALSE, breaks = 50,
     xlab = "Altura (cm)", ylab = "Densidade",
     main = expression("Priori: " ~ mu %~% Normal(160, 100)),
     col = "lightyellow", border = "white")
```

```r
par(mfrow = c(1, 1))

# Contagem de alturas impossíveis com a priori mais dispersa
sum(y_sim_amplo < 0)
sum(y_sim_amplo > 300)

# Proporção de alturas impossíveis em cada cenário
mean(y_sim < 0 | y_sim > 300)
mean(y_sim_amplo < 0 | y_sim_amplo > 300)
```

O que explorar

- Compare os dois histogramas: qual das distribuições preditivas a priori concentra mais alturas em faixas biologicamente razoáveis?
- Compare as proporções de alturas impossíveis nos dois cenários. O que a priori mais dispersa implica sobre o que o modelo "espera" observar?
- Experimente uma priori intermediária, como $\mu \sim \text{Normal}(160, 50)$. Como a distribuição preditiva a priori se situa entre os dois extremos já explorados?