Prática em R - Aula 03

Modelos generativos e a aproximação por grade

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 Verossimilhança binomial com `dbinom()` e `pbinom()`

O experimento da atirada de globo consiste em lançar um globo terrestre ao ar e registrar se o dedo indicador pousa sobre água (**W**) ou terra (**T**) ao capturá-lo (McElreath 2020). O objetivo é estimar a proporção p de oceano na superfície terrestre. Um experimento com nove lançamentos produziu a sequência:

W, T, W, W, W, T, W, T, W

O total é $k = 6$ águas em $n = 9$ lançamentos. Nesta seção, usamos `dbinom()` para calcular a verossimilhança binomial em diferentes valores de p e construir a curva de verossimilhança. Em seguida, usamos `pbinom()` para calcular probabilidades acumuladas.

Código

```r
# Dados do experimento
k <- 6   # águas observadas
n <- 9   # total de lançamentos

# Verossimilhança em um único ponto: P(k=6 | n=9, p=0.70)
dbinom(6, size = 9, prob = 0.70)

# Grade de valores de p em [0, 1]
p_grid <- seq(0, 1, length.out = 100)

# Verossimilhança em cada ponto da grade
likelihood <- dbinom(k, size = n, prob = p_grid)

# Gráfico da curva de verossimilhança
plot(p_grid, likelihood, type = "l", lwd = 2,
     xlab = "p (proporção de oceano)", ylab = "Verossimilhança",
     main = "Verossimilhança binomial - k=6, n=9")
```

O que observar

- Identifique onde a curva de verossimilhança atinge o valor máximo. Que relação esse valor tem com os dados do experimento (k e n)?
- Observe o comportamento da curva nos extremos ($p = 0$ e $p = 1$). Como você interpreta esses valores em termos de compatibilidade com os dados?

Código

```r
# P(K >= 6 | n=9, p=0.70): probabilidade de observar 6 ou mais águas com p=0,70
1 - pbinom(5, size = 9, prob = 0.70)

# O mesmo resultado com p=0,50: seria o valor observado incomum?
1 - pbinom(5, size = 9, prob = 0.50)
```

O que observar

- Com $p = 0{,}70$, calcule a probabilidade de observar seis ou mais águas em nove lançamentos.
- Com $p = 0{,}50$, repita o cálculo. Compare os dois resultados: o valor observado ($k = 6$) seria

incomum se a proporção de oceano fosse apenas 50%?

2 O modelo como gerador de dados com `rbinom()`

A função `rbinom()` usa o modelo binomial como um modelo generativo: dado um valor de p, ela simula realizações do experimento. Nesta seção, iremos gerar mil experimentos de nove lançamentos com $p = 0{,}70$ e comparar a distribuição das contagens simuladas com o valor observado $k = 6$.

Código

```
# Simular 1000 experimentos de 9 lançamentos com p fixo em 0,70
k_simulado <- rbinom(1000, size = 9, prob = 0.70)

# Histograma das contagens simuladas
hist(k_simulado, breaks = -0.5 + (0:10),
     xlab = "Número de águas em 9 lançamentos",
     ylab = "Frequência",
     main = "Simulação: 1000 experimentos com p = 0,70")
abline(v = 6, col = "red", lwd = 2)   # valor observado
```

O que explorar

- Localize a linha vermelha ($k = 6$) no histograma: ela está na região central ou periférica da distribuição?
- Altere `prob = 0.70` para `prob = 0.30` e execute o código novamente. O valor observado $k = 6$ continua igualmente compatível com o modelo?
- O que acontece com a posição e a forma da distribuição simulada quando p muda?

Código

```
# Comparar histogramas com diferentes valores de p
par(mfrow = c(1, 3))

for (p_val in c(0.30, 0.50, 0.70)) {
  k_sim <- rbinom(1000, size = 9, prob = p_val)
  hist(k_sim, breaks = -0.5 + (0:10),
       xlab = "Número de águas",
       ylab = "Frequência",
       main = paste("p =", p_val),
       xlim = c(0, 9))
  abline(v = 6, col = "red", lwd = 2)
}
par(mfrow = c(1, 1))
```

O que observar

- Para cada valor de p, identifique em torno de qual contagem o histograma concentra a maior frequência. Como esse valor se relaciona com $p \times n$?
- Para qual dos três valores de p a linha vermelha ($k = 6$) está na região de maior frequência? Para qual ela está na cauda da distribuição?

3 Checagem preditiva a priori

A checagem preditiva a priori combina simulação com a distribuição a priori sobre p: em vez de fixar p em um único valor, amostramos p da distribuição a priori e, para cada valor amostrado, simulamos um experimento de nove lançamentos. O resultado é uma distribuição das contagens que o modelo espera ver antes de observar qualquer dado real.

Código

```r
# Checagem preditiva a priori com distribuição a priori uniforme
p_amostrado <- runif(1000, min = 0, max = 1)   # amostrar p da priori uniforme
k_pred_unif <- rbinom(1000, size = 9, prob = p_amostrado)

hist(k_pred_unif, breaks = -0.5 + (0:10),
     xlab = "Número de águas simuladas",
     ylab = "Frequência",
     main = "Checagem preditiva a priori - distribuição a priori uniforme")
abline(v = 6, col = "red", lwd = 2)   # valor observado
```

O que observar

- Como se comporta o histograma das contagens simuladas entre 0 e 9 águas quando a distribuição a priori para p é uniforme no intervalo $[0, 1]$.
- O que isso indica sobre o modelo, antes de ver os dados.
- Onde se encontra a linha vermelha ($k = 6$) no histograma.

Código

```r
# Checagem preditiva a priori com distribuição a priori informativa
p_amostrado_info <- runif(1000, min = 0.60, max = 0.80)
k_pred_info      <- rbinom(1000, size = 9, prob = p_amostrado_info)

par(mfrow = c(1, 2))

hist(k_pred_unif, breaks = -0.5 + (0:10),
     xlab = "Número de águas simuladas", ylab = "Frequência",
     main = "Priori uniforme [0, 1]", xlim = c(0, 9))
abline(v = 6, col = "red", lwd = 2)

hist(k_pred_info, breaks = -0.5 + (0:10),
     xlab = "Número de águas simuladas", ylab = "Frequência",
     main = "Priori informativa [0.60, 0.80]", xlim = c(0, 9))
abline(v = 6, col = "red", lwd = 2)

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

O que observar

- Compare os dois histogramas. Como a distribuição das contagens simuladas muda quando a distribuição a priori para p muda de uniforme em $[0, 1]$ para uniforme em [0,60, 0,80]?
- O que esse resultado indica sobre a relação entre a distribuição a priori adotada e a distribuição preditiva antes de observar os dados?

4 Aproximação por grade

A aproximação por grade substitui o intervalo contínuo $p \in [0, 1]$ por um conjunto finito de pontos igualmente espaçados. Em cada ponto da grade, calcula-se a verossimilhança e o produto verossimilhança $\times$ distribuição a priori. A normalização desse produto gera uma distribuição discreta que aproxima a distribuição a posteriori contínua.

4.1 Distribuição a priori uniforme

Código

```
# Grade com 100 pontos em [0, 1]
p_grid <- seq(0, 1, length.out = 100)
k <- 6; n <- 9

# Passo 1: verossimilhança em cada ponto
likelihood <- dbinom(k, size = n, prob = p_grid)

# Passo 2: distribuição a priori uniforme e produto
prior_unif <- rep(1, 100)          # todos os pontos com o mesmo peso
peso_unif  <- prior_unif * likelihood

# Passo 3: normalização
posterior_unif <- peso_unif / sum(peso_unif)

# Gráfico da distribuição a posteriori
plot(p_grid, posterior_unif, type = "l", lwd = 2,
     xlab = "p (proporção de oceano)",
     ylab = "Probabilidade a posteriori",
     main = "Distribuição a posteriori - distribuição a priori uniforme")
abline(v = k / n, lty = 2, col = "gray40")
```

O que observar

- A distribuição a posteriori tem pico próximo a $p = 0{,}67$, que coincide com a proporção observada.
- Com a distribuição a priori uniforme, a forma da distribuição a posteriori é idêntica à da verossimilhança, com diferença apenas na escala do eixo y, pois a distribuição a posteriori foi normalizada para somar 1.

4.2 Efeito da resolução da grade

Código

```
# Comparação da distribuição a posteriori com diferentes resoluções de grade
par(mfrow = c(2, 2))

for (n_pts in c(5, 10, 20, 100)) {
  pg   <- seq(0, 1, length.out = n_pts)
  lik  <- dbinom(k, size = n, prob = pg)
  post <- lik / sum(lik)   # distribuição a priori uniforme: post proporcional a lik
  plot(pg, post, type = "b", pch = 19,
       main = paste("Grade com", n_pts, "pontos"),
       xlab = "p", ylab = "Probabilidade a posteriori")
}

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

O que observar

- Com 5 pontos, quais valores de p são avaliados? Como a aparência da distribuição a posteriori muda em relação ao caso com 100 pontos?
- Observe como a distribuição a posteriori evolui à medida que o número de pontos aumenta. O que você conclui sobre a relação entre a resolução da grade e a qualidade da aproximação?

4.3 Distribuição a priori uniforme versus informativa

Código

```r
p_grid <- seq(0, 1, length.out = 100)
k <- 6; n <- 9

# Verossimilhança
likelihood <- dbinom(k, size = n, prob = p_grid)

# Distribuição a priori uniforme
prior_unif     <- rep(1, 100)
posterior_unif <- prior_unif * likelihood
posterior_unif <- posterior_unif / sum(posterior_unif)

# Distribuição a priori informativa: zero abaixo de p = 0,50
prior_info     <- ifelse(p_grid < 0.5, 0, 1)
posterior_info <- prior_info * likelihood
posterior_info <- posterior_info / sum(posterior_info)

# Gráfico comparativo
plot(p_grid, posterior_unif, type = "l", lwd = 2, col = "steelblue",
     xlab = "p (proporção de oceano)",
     ylab = "Probabilidade a posteriori",
     main = "Distribuição a priori uniforme vs. informativa",
     ylim = c(0, max(posterior_unif, posterior_info)))
lines(p_grid, posterior_info, col = "red", lwd = 2)
abline(v = k / n, lty = 2, col = "gray40")
legend("topleft",
       legend = c("Priori uniforme", "Priori informativa (p >= 0,5)"),
       col = c("steelblue", "red"), lwd = 2, bty = "n")
```

O que observar

- Localize o pico de cada curva. As duas distribuições a posteriori têm o pico no mesmo valor de p? Se não, qual delas tem o pico mais à direita?
- Compare a largura das duas curvas. A distribuição a priori informativa altera apenas a posição do pico ou também a forma da distribuição a posteriori?
- A linha tracejada vertical marca a proporção observada $k/n \approx 0{,}67$. Para qual das duas distribuições a posteriori esse valor está mais próximo do pico?

O que explorar

- Modifique os valores de `k` e `n` no código acima (por exemplo, use $k = 60$ e $n = 90$, mantendo a mesma proporção observada) e execute novamente.
- Compare a largura das distribuições a posteriori obtidas com os dois tamanhos de amostra.
- Observe como o papel da distribuição a priori informativa muda quando o volume de dados aumenta.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* 2º ed. Chapman; Hall/CRC.