Prática em R - Aula 05

Distribuição a posteriori no modelo Normal

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 Exploração inicial do conjunto de dados

O dataset Howell1 (McElreath 2020) registra dados antropométricos de 544 indivíduos !Kung San do deserto de Kalahari, coletados pelo antropólogo Nancy Howell. As variáveis disponíveis são `height` (altura em cm), `weight` (peso em kg), `age` (idade em anos) e `male` (indicador de sexo). Neste laboratório, o foco recai sobre a variável `height` dos 352 adultos com 18 anos ou mais.

O objetivo inicial é examinar a distribuição das alturas antes de qualquer ajuste de modelo, identificar as características da variável e avaliar se a Distribuição Normal é um modelo generativo plausível para esses dados.

Código

```
library(ggplot2)

# Carregar os dados
dados <- read.csv("https://raw.githubusercontent.com/FCopf/prob-est-2026/refs/heads/main/_bibliograp
                  sep = ";")

# Subset de adultos (age >= 18)
adultos <- dados[dados$age >= 18, ]

# Resumo descritivo das variáveis
str(adultos)
summary(adultos[, c("height", "weight", "age")])

# Estatísticas da variável de interesse
cat("n       =", nrow(adultos), "\n")
cat("Média   =", round(mean(adultos$height), 2), "cm\n")
cat("Mediana =", round(median(adultos$height), 2), "cm\n")
cat("DP      =", round(sd(adultos$height), 2), "cm\n")
cat("Mín.    =", round(min(adultos$height), 2), "cm\n")
cat("Máx.    =", round(max(adultos$height), 2), "cm\n")
```

O que observar

- Examine a diferença entre a média e a mediana de `height`. O que essa diferença (ou semelhança) sugere sobre a assimetria da distribuição?
- Em que faixa de alturas se concentra a maioria dos adultos? Como essa faixa se compara com o desvio padrão calculado?

1.1 Visualização da distribuição

Código

```
# Histograma com curva de densidade e curva Normal teórica sobreposta
media_h <- mean(adultos$height)
dp_h    <- sd(adultos$height)

ggplot(adultos, aes(x = height)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30,
                 fill = "steelblue", color = "white",
                 linewidth = 0.2, alpha = 0.7) +
```

```r
  geom_density(color = "steelblue4", linewidth = 1) +
  stat_function(fun  = dnorm,
                args = list(mean = media_h, sd = dp_h),
                color = "#e6a073", linewidth = 1, linetype = "dashed") +
  labs(x = "Altura (cm)", y = "Densidade",
       title = "Distribuição das alturas - adultos !Kung San",
       subtitle = paste0("n = ", nrow(adultos),
                         "  |  média = ", round(media_h, 1),
                         " cm  |  dp = ", round(dp_h, 1), " cm")) +
  theme_minimal(base_size = 12)
```

O que observar

- A curva de densidade empírica (azul escuro) e a curva Normal teórica (laranja tracejada) se sobrepõem bem? Em que regiões da distribuição há maior divergência?
- A forma do histograma justifica o uso da Distribuição Normal como modelo? Identifique características que apóiam ou questionam essa escolha.

---

2 Distribuições a priori e checagem preditiva a priori

Antes de observar os dados como evidência para estimar os parâmetros, é necessário especificar as distribuições a priori para μ e σ. As distribuições a priori expressam o conhecimento disponível sobre os parâmetros antes de qualquer dado ser incorporado ao modelo.

O modelo generativo para alturas de adultos 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 uma faixa de alturas biologicamente razoável para adultos humanos, cobrindo com probabilidade substancial o intervalo de 120 a 200 cm. A distribuição a priori Exponencial(0,1) para σ garante que o desvio padrão permaneça estritamente positivo e concentra a probabilidade abaixo de 30 cm, com valor esperado de 10 cm.

A checagem preditiva a priori verifica se as distribuições a priori, tomadas em conjunto, geram dados com características compatíveis com o fenômeno estudado. Para isso, ajusta-se o modelo sem incorporar os dados observados como evidência, amostrando exclusivamente das distribuições a priori especificadas.

Código

```r
library(brms)

# Definição explícita das distribuições a priori
prioris <- c(
  prior(normal(160, 20),    class = Intercept),  # priori para mu
  prior(exponential(0.1),   class = sigma)        # priori para sigma
)

# Ajuste com sample_prior = "only": amostra apenas das distribuições a priori
fit_prior <- brm(
  formula = height ~ 1,
  family  = gaussian(),
  prior   = prioris,
  data    = adultos,
```

```
  sample_prior = "only",
  chains  = 4,
  iter    = 1000,
  warmup  = 500,
  refresh = 0
)
```

O que observar

- Analise o papel do argumento `sample_prior = "only"` no comportamento do ajuste. Em que aspecto fundamental esse ajuste difere de um `brm()` padrão que incorpora os dados?
- Com base nas distribuições a priori especificadas, que faixa de alturas o modelo considera plausível antes de ver os dados?

2.1 Visualização da distribuição preditiva a priori

Código

```
# Checagem preditiva a priori: distribuição simulada vs. dados observados
pp_check(fit_prior, ndraws = 100) +
  labs(title = "Checagem preditiva a priori",
       x = "Altura (cm)", y = "Densidade") +
  xlim(-50, 400) +
  theme_minimal(base_size = 12)
```

O que explorar

- O intervalo de alturas gerado pela distribuição preditiva a priori engloba valores biologicamente impossíveis (alturas negativas ou extremamente elevadas)? O que isso indica sobre o grau de informatividade das distribuições a priori escolhidas?
- Altere o desvio padrão da distribuição a priori de μ de 20 para 100 (`normal(160, 100)`) e reajuste o modelo com `sample_prior = "only"`. Como a distribuição preditiva a priori se altera?

---

3 Ajuste do modelo com brms

Com as distribuições a priori verificadas, o modelo é ajustado incorporando os dados observados. O `brms` compila o modelo em C++ e utiliza HMC para amostrar da distribuição a posteriori dos parâmetros. O resultado é um conjunto de amostras da distribuição a posteriori que resume o que o modelo aprendeu sobre μ e σ à luz dos dados.

Código

```
# Ajuste do modelo com os dados observados
fit <- brm(
  formula = height ~ 1,
  family  = gaussian(),
  prior   = prioris,
  data    = adultos,
  chains  = 4,
  iter    = 1000,
  warmup  = 500,
  refresh = 0
)
```

```r
# Visualizar o sumário do ajuste
summary(fit)
```

O que observar

- O sumário apresenta as estimativas para `b_Intercept` (correspondente a μ) e `sigma`. Quais são os valores centrais estimados para cada parâmetro?
- A coluna `l-89%` e `u-89%` do sumário delimita o intervalo de credibilidade de 89%. O que esse intervalo indica sobre a precisão com que os dados determinam μ?

---

4 Distribuições a posteriori dos parâmetros

A distribuição a posteriori é o produto central da estimação bayesiana. Ela expressa a incerteza residual sobre cada parâmetro após combinar as distribuições a priori com as evidências dos dados. Nesta seção, os parâmetros μ e σ são explorados por meio de resumos numéricos e visualizações.

4.1 Resumo numérico dos parâmetros

Código

```r
library(posterior)

# Extrair amostras da distribuição a posteriori em formato tidy
amostras <- as_draws_df(fit)

# Resumo detalhado: média, dp e intervalos de credibilidade de 89%
summarise_draws(
  amostras,
  mean, median, sd,
  ~quantile(.x, c(0.055, 0.945)),
  .filter = ~grepl("b_Intercept|sigma", .x)
)

# Intervalo de credibilidade de 89% via posterior_interval()
posterior_interval(fit, prob = 0.89,
                   pars = c("b_Intercept", "sigma"))
```

O que observar

- Examine a amplitude do intervalo de credibilidade de 89% para `b_Intercept` e para `sigma`. O que cada intervalo informa sobre o respectivo parâmetro e o que ele descreve no contexto do modelo?
- O desvio padrão da distribuição a posteriori de `b_Intercept` é consideravelmente menor do que o valor estimado de `sigma`. Que diferença conceitual entre os dois parâmetros explica essa discrepância de magnitude?

4.2 Visualização das distribuições marginais a posteriori

Código

```r
library(bayesplot)
library(ggplot2)
library(patchwork)

# Distribuições marginais a posteriori para mu e sigma
p_mu <- ggplot(amostras, aes(x = b_Intercept)) +
```

```
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = "steelblue", color = "white",
                 linewidth = 0.2, alpha = 0.7) +
  geom_density(color = "steelblue4", linewidth = 1) +
  geom_vline(xintercept = posterior_interval(fit, prob = 0.89,
                                             pars = "b_Intercept"),
             linetype = "dashed", color = "gray40") +
  labs(x = expression(mu ~ "(cm)"), y = "Densidade",
       title = expression("Distribuição a posteriori de " ~ mu)) +
  theme_minimal(base_size = 12)

p_sigma <- ggplot(amostras, aes(x = sigma)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = "orange", color = "white",
                 linewidth = 0.2, alpha = 0.7) +
  geom_density(color = "darkorange4", linewidth = 1) +
  geom_vline(xintercept = posterior_interval(fit, prob = 0.89,
                                             pars = "sigma"),
             linetype = "dashed", color = "gray40") +
  labs(x = expression(sigma ~ "(cm)"), y = "Densidade",
       title = expression("Distribuição a posteriori de " ~ sigma)) +
  theme_minimal(base_size = 12)

p_mu + p_sigma
```

O que observar

- As linhas tracejadas delimitam o intervalo de credibilidade de 89%. Que fração das amostras da distribuição a posteriori está contida entre essas linhas?
- A distribuição a posteriori de μ é consideravelmente mais estreita do que a distribuição a priori Normal$(160, 20)$. O que esse estreitamento indica sobre o efeito dos dados no conhecimento sobre μ?

4.3 Distribuição a posteriori conjunta

Código

```
# Gráfico de dispersão das amostras conjuntas (mu, sigma)
ggplot(amostras, aes(x = b_Intercept, y = sigma)) +
  geom_point(alpha = 0.15, color = "#1a9988", size = 0.8) +
  geom_density_2d(color = "gray30", linewidth = 0.4) +
  labs(x = expression(mu ~ "(cm)"), y = expression(sigma ~ "(cm)"),
       title = "Distribuição a posteriori conjunta de μ e  ") +
  theme_minimal(base_size = 12)
```

O que observar

- As amostras de μ e σ formam uma nuvem de pontos com estrutura visível. Existe correlação entre os dois parâmetros a posteriori? O que isso indicaria sobre a relação entre a estimativa da média e a estimativa da variabilidade?
- Compare a região ocupada pelas amostras com o espaço total de valores plausíveis para (μ, σ). Como $n = 352$ observações delimitam essa região?

---

5 Distribuição preditiva da posteriori

A distribuição a posteriori dos parâmetros descreve o que o modelo aprendeu sobre μ e σ. A distribuição preditiva da posteriori vai além: ela descreve a distribuição de novas alturas $\tilde{y}$, propagando simultaneamente a variabilidade individual das alturas e a incerteza residual sobre os parâmetros.

A checagem preditiva da posteriori verifica se os dados que o modelo consegue gerar são compatíveis com os dados que foram realmente observados. Essa comparação é uma das ferramentas centrais para avaliar a adequação do modelo.

5.1 Checagem preditiva da posteriori

Código

```
# Checagem preditiva da posteriori: sobreposição de distribuições simuladas
# (linhas azul-claras) sobre os dados observados (linha escura)
pp_check(fit, ndraws = 100) +
  labs(title = "Checagem preditiva da posteriori - modelo Normal",
       x = "Altura (cm)", y = "Densidade") +
  theme_minimal(base_size = 12)

# Checagem alternativa: comparação de estatísticas resumo
pp_check(fit, type = "stat", stat = "mean",  ndraws = 1000) +
  labs(title = "Distribuição preditiva da média a posteriori",
       x = "Média simulada (cm)") +
  theme_minimal(base_size = 12)

pp_check(fit, type = "stat", stat = "sd", ndraws = 1000) +
  labs(title = "Distribuição preditiva do DP a posteriori",
       x = "Desvio padrão simulado (cm)") +
  theme_minimal(base_size = 12)
```

O que observar

- No primeiro gráfico, as distribuições simuladas (linhas azul-claras) cobrem bem a distribuição dos dados observados (linha escura)? Em que regiões da distribuição o modelo apresenta maior ou menor ajuste?
- Nos gráficos de estatísticas resumo, onde se posiciona o valor observado (linha vertical) em relação à distribuição de valores simulados? O modelo reproduz adequadamente a média e o desvio padrão dos dados?

5.2 Predição para novas observações

Código

```
# Gerar amostras da distribuição preditiva da posteriori
y_pred <- posterior_predict(fit, ndraws = 4000)

# Converter para vetor (todas as predições)
y_pred_vec <- as.vector(y_pred)

# Comparar dados observados com predições
dados_comp <- data.frame(
  valor = c(adultos$height, y_pred_vec),
  fonte = c(rep("Observado", nrow(adultos)),
            rep("Preditiva da posteriori", length(y_pred_vec)))
)

ggplot(dados_comp, aes(x = valor, fill = fonte)) +
```

```r
  geom_histogram(aes(y = after_stat(density)), bins = 40,
                 alpha = 0.5, color = "white",
                 linewidth = 0.2, position = "identity") +
  scale_fill_manual(values = c("Observado"                = "steelblue",
                               "Preditiva da posteriori" = "#e6a073")) +
  labs(x = "Altura (cm)", y = "Densidade", fill = NULL,
       title = "Dados observados e distribuição preditiva da posteriori") +
  theme_minimal(base_size = 12) +
  theme(legend.position = "top")
```

O que observar

- A distribuição preditiva da posteriori é mais larga do que a distribuição dos dados observados. Quais são as duas fontes de variação que tornam a distribuição preditiva mais larga do que os dados isolados?
- Se o modelo não estivesse capturando bem a variabilidade dos dados, como a distribuição preditiva da posteriori se diferenciaria da distribuição observada?

---

6 Alteração das distribuições a priori e sensibilidade da posteriori

A inferência bayesiana combina o conhecimento prévio (distribuições a priori) com as evidências dos dados (verossimilhança) para produzir a distribuição a posteriori. Com amostras grandes, os dados tendem a dominar as distribuições a priori. Com amostras pequenas ou distribuições a priori fortemente informativas, a influência das distribuições a priori é mais perceptível.

Esta seção explora como a escolha das distribuições a priori afeta a distribuição a posteriori. Três modelos são comparados:

- **Modelo fracamente informativo:** $\mu \sim \text{Normal}(160, 20)$ (distribuição a priori original)
- **Modelo fortemente informativo:** $\mu \sim \text{Normal}(160, 1)$ (distribuição a priori concentrada em 160 cm)
- **Modelo com priori deslocada:** $\mu \sim \text{Normal}(185, 5)$ (distribuição a priori incompatível com os dados)

Código

```r
# Modelo com distribuição a priori fortemente informativa (muito estreita)
fit_info <- update(
  fit,
  prior = c(
    prior(normal(160, 1),   class = Intercept),
    prior(exponential(0.1), class = sigma)
  ),
  refresh = 0
)

# Modelo com distribuição a priori deslocada (incompatível com os dados)
fit_deslocada <- update(
  fit,
  prior = c(
    prior(normal(185, 5),   class = Intercept),
    prior(exponential(0.1), class = sigma)
  ),
  refresh = 0
)
```

```r
# Resumo dos três modelos
cat("=== Distribuição a priori fracamente informativa ===\n")
posterior_interval(fit,         prob = 0.89, pars = "b_Intercept")

cat("=== Distribuição a priori fortemente informativa ===\n")
posterior_interval(fit_info,    prob = 0.89, pars = "b_Intercept")

cat("=== Distribuição a priori deslocada ===\n")
posterior_interval(fit_deslocada, prob = 0.89, pars = "b_Intercept")
```

O que observar

- Compare os intervalos de credibilidade de 89% para μ nos três modelos. Qual dos modelos apresenta maior deslocamento em relação à média amostral `mean(adultos$height)`?
- A distribuição a priori deslocada (`Normal(185, 5)`) está em conflito com os dados. Com $n = 352$, os dados conseguem superar essa distribuição a priori e direcionar a distribuição a posteriori para a região dos dados? O que isso indica sobre o peso relativo dos dados frente às distribuições a priori?

6.1 Comparação visual das distribuições a posteriori

Código

```r
library(posterior)

# Extrair amostras dos três modelos
am_fraca     <- as_draws_df(fit)$b_Intercept
am_info      <- as_draws_df(fit_info)$b_Intercept
am_deslocada <- as_draws_df(fit_deslocada)$b_Intercept

# Organizar em um único data frame
comp_post <- data.frame(
  mu    = c(am_fraca, am_info, am_deslocada),
  model = rep(c("Priori fraca\nNormal(160, 20)",
                "Priori informativa\nNormal(160, 1)",
                "Priori deslocada\nNormal(185, 5)"),
              each = length(am_fraca))
)

ggplot(comp_post, aes(x = mu, fill = model, color = model)) +
  geom_density(alpha = 0.35, linewidth = 0.9) +
  geom_vline(xintercept = mean(adultos$height),
             linetype = "dashed", color = "black", linewidth = 0.8) +
  scale_fill_manual(values  = c("#1a9988", "#2c5f7c", "#e6a073")) +
  scale_color_manual(values = c("#1a9988", "#2c5f7c", "#e6a073")) +
  labs(x = expression(mu ~ "(cm)"), y = "Densidade",
       fill = "Modelo", color = "Modelo",
       title = "Comparação das distribuições a posteriori de μ",
       subtitle = "Linha tracejada = média amostral dos dados") +
  theme_minimal(base_size = 12) +
  theme(legend.position = "right")
```

O que explorar

- A distribuição a posteriori do modelo com priori deslocada (`Normal(185, 5)`) converge para a região dos dados ou permanece próxima do valor da distribuição a priori? Como o tamanho amostral $n = 352$ influencia esse resultado?
- Substitua os 352 adultos por uma amostra aleatória de 10 observações (`adultos_pequeno <-`

`adultos[sample(nrow(adultos), 10), ]`) e re-ajuste os três modelos. Como a sensibilidade às distribuições a priori se altera com amostras menores?

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