Introdução à inferência bayesiana

Conceitos e aplicação

LAPEGS, junho de 2026

Guilherme D. Garcia | gdgarcia.ca

Bayes, Laplace, Fisher?

- A estatística bayesiana existe há séculos (Bayes, 1763; Laplace, 1774; 1812; McGrayne, 2011)
- A limitação sempre esteve na sua **implementação** (computacionalmente intensa)
- Por isso, temos a sensação de que a estatística frequentista veio "antes" (Fisher, 1925)

Nossos dados

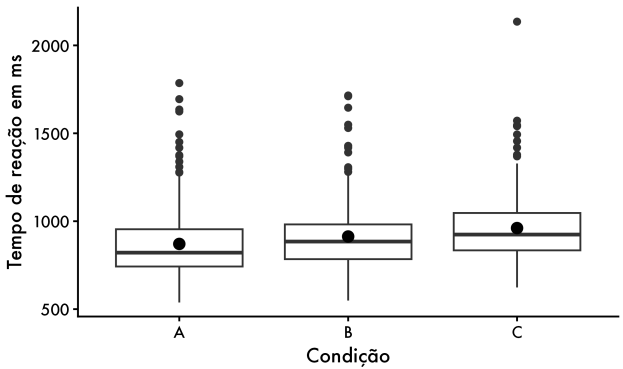


Figura 1: Dados para análise

- Média e desvio-padrão de cada condição:

```
# Versão em ms:
# # A tibble: 3 × 3
#   Condition  mean     s
#   <fct>     <dbl> <dbl>
# 1 A          871.  211.
# 2 B          913.  201.
# 3 C          961.  199.

# Versão em log(ms):
# # A tibble: 3 × 3
#   Condition  mean     s
#   <fct>     <dbl> <dbl>
# 1 A         6.74 0.217
# 2 B         6.80 0.202
# 3 C         6.85 0.186
```

Ver o output ↓

1 Ideia geral

Objetivo para hoje

Construir uma intuição para **inferência bayesiana**

- O foco será em **modelos de regressão**, mas a ideia independe do *tipo* de modelo utilizado
- A pergunta principal será: **como quantificamos incerteza?**
- Se tivermos tempo: um pouco de código em R

Mensagem central: Um modelo bayesiano não dá apenas uma estimativa. Ele dá uma distribuição de possibilidades compatíveis com os dados e com as nossas suposições.

Nosso itinerário

1. Comparação básica entre abordagens frequentista e bayesiana + resultados com R
2. A importância da distribuição *a priori* e como estimamos parâmetros com Bayes
3. Ferramentas úteis para rodar modelos e visualizar resultados

O problema estatístico

Queremos aprender algo sobre um processo que não observamos perfeitamente.

O que observamos

- Dados com ruído
- Amostras finitas
- Medidas imperfeitas
- Variação entre indivíduos, lugares ou momentos

O que queremos

- Relações entre variáveis
- Tamanho de efeitos
- Previsões
- Incerteza sobre tudo isso

Estatística: uma forma quantificável de pensar sob incerteza.

Por que regressão?

Muitas perguntas aplicadas têm esta forma:

$$y = \alpha + \beta x + \varepsilon$$

- y é o resultado que queremos entender
- x é uma variável explicativa
- α e β descrevem a relação
- ε representa a variação que o modelo não explicou

Em uma análise bayesiana: não perguntamos apenas "qual é o melhor valor de $\hat{\beta}$?" Perguntamos: "quais valores de $\hat{\beta}$ são plausíveis considerando o que observamos e o nosso conhecimento prévio?"

Regressão como exemplo

Modelo

```
y ~ x ...
```

Perguntas

- A relação é positiva ou negativa?
- A relação é grande ou pequena?
- Quanta incerteza ainda temos?
- O modelo prevê dados parecidos com os observados?

Leitura: quando `x` muda, esperamos que `y` mude. Essa mudança pode ser teoricamente motivada ou espúria. Por isso, uma análise excelente ≠ um estudo excelente: **especialmente** na era pós-IA...

2 Frequência e Bayes

Duas leituras de probabilidade

Frequentista

Probabilidade descreve o comportamento de procedimentos em muitas **repetições**.

Os parâmetros são fixos.

Bayesiana

Probabilidade descreve incerteza sobre quantidades desconhecidas.

Os parâmetros têm distribuições.

> A diferença não é "objetivo" contra "subjetivo". A diferença é **o que** a probabilidade representa.

Duas leituras de probabilidade: exemplo frequentista

Frequentista

- Estudamos o tempo de reação de participantes com base em uma condição de 3 níveis
- Coletamos dados e verificamos a probabilidade desses dados sob uma hipótese nula
- Ou seja, calculamos algo como $P(\text{dados} \mid H_0)$
- O resultado nos dá uma probabilidade **dos dados** sob essa hipótese, ou seja, o nosso valor-p
- Perceba que os parâmetros aparecem como estimativas pontuais na saída abaixo:

```
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  6.74456    0.01384 487.409  < 2e-16 ***
# ConditionB   0.05099    0.01971   2.587   0.0099 **
# ConditionC   0.10513    0.02023   5.196 2.78e-07 ***

# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Duas leituras de probabilidade: exemplo frequentista

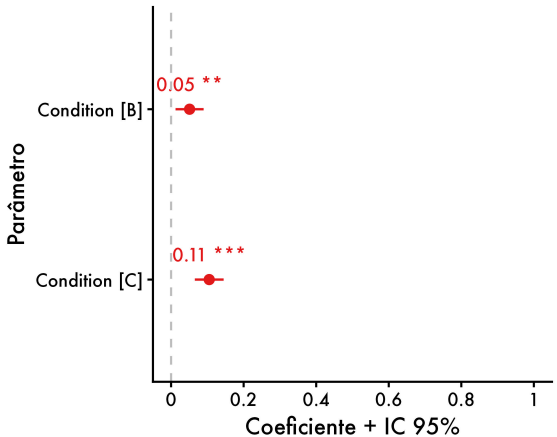


Figura 2: Resultado do modelo frequentista

- Estimativa é apenas um “ponto”
- IC não é uma distribuição
- Conclusão categórica: IC inclui zero?

Duas leituras de probabilidade: exemplo bayesiano

- Aqui, os **dados** são fixos e nosso parâmetro tem distribuições *a posteriori*
- Os valores β aqui representam apenas a média de uma distribuição de valores plausíveis
- Não vemos valores P, porque estamos calculando $P(\beta|\text{dados})$, e não mais $P(\text{dados}|\beta)$

```
# Regression Coefficients:
#             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# Intercept       6.74      0.01     6.72     6.77 1.00     3390     2936
# ConditionB      0.05      0.02     0.01     0.09 1.00     3693     2800
# ConditionC      0.11      0.02     0.07     0.14 1.00     3394     3196
```

Duas leituras de probabilidade: exemplo bayesiano

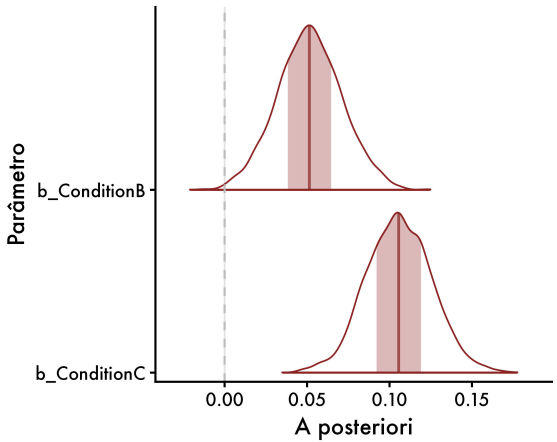


Figura 3: Resultado do modelo bayesiano

- Estimativa é uma **distribuição** de valores prováveis considerando os dados modelados
- IC é um intervalo de credibilidade
- Conclusão com nuance: onde está o zero?
- No gráfico, área sombreada = 50%

Intervalos: cuidado com a interpretação

Intervalo de confiança (frequentista)

Em repetições hipotéticas, o procedimento cobre o valor verdadeiro em certa proporção dos casos.

→ É uma propriedade do **procedimento**.

Intervalo de credibilidade (bayesiano)

Dado o modelo e os dados, certa proporção da distribuição posterior está dentro do intervalo.

→ É uma afirmação sobre o **parâmetro**.

O que muda na prática?

Em muitos casos simples, os resultados podem ser parecidos (exemplo acima).

Mas a formulação bayesiana facilita:

1. Incorporar suposições de modo explícito através de distribuições *a priori* (a seguir)
2. Estimar modelos complexos com a mesma lógica básica (e que muitas vezes não convergiriam)
3. Falar diretamente sobre incerteza
4. Lidar com heteroscedasticidade e com modelos personalizados
5. Fazer previsões a partir da distribuição *a posteriori* (a seguir)
6. Comparações múltiplas sem a necessidade de correções típicas (sem erro *family-wise*)
7. Personalização total de modelos, com uma gama de distribuições disponíveis para diferentes cenários

☞ A vantagem é menos "ganhar significância" e mais **modelar incerteza de forma coerente**.

3 Atualização bayesiana

A regra de Bayes

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

`P(theta)`	*a priori*: o que era plausível antes dos dados (o que a literatura nos diz)
`P(y | theta)`	*likelihood*: o que os dados dizem para cada valor
`P(theta | y)`	*a posteriori*: o que é plausível depois dos dados (nossa "conclusão")
`P(y)`	normalização: faz a distribuição posterior somar/integrar 1

☞ Bayes é uma regra de atualização.

Priori × dados ⟶ posterior

Ideia: nossa conclusão sobre algo é uma combinação (**ponderada**) de dois elementos:

- o nosso conhecimento prévio (= a expectativa, ou *a priori*)
- os dados que observamos (= a evidência, ou *likelihood*)

Cenários hipotéticos

☞ Uma moeda desconhecida que acreditamos ser normal. "Normal" $\rightarrow P(\text{cara}) = P(\text{coroa}) = 0.5$

1. A evidência concorda com nossa crença
2. A evidência discorda da nossa crença, e somos parcialmente flexíveis
3. A evidência discorda da nossa crença, mas somos ortodoxos

- Chamaremos o nosso parâmetro de θ. Em uma moeda normal, $\theta = 0.5$

Cenário 1

- Se a evidência (*likelihood*) está de acordo com o *a priori*, nossa conclusão (*a posteriori*) também estará:

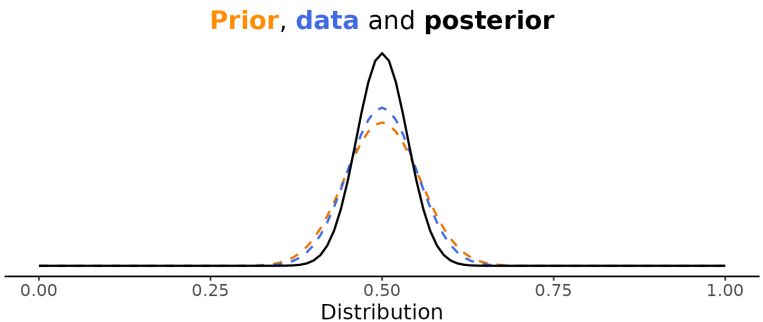


Figura 4: Quando expectativa e evidência estão em acordo.

Cenário 2

- Contradição entre *a priori* e a evidência: os dados sugerem que $\theta \approx 0.85$
- Como nossa crença não é tão extrema, nossa conclusão é um "meio-termo":

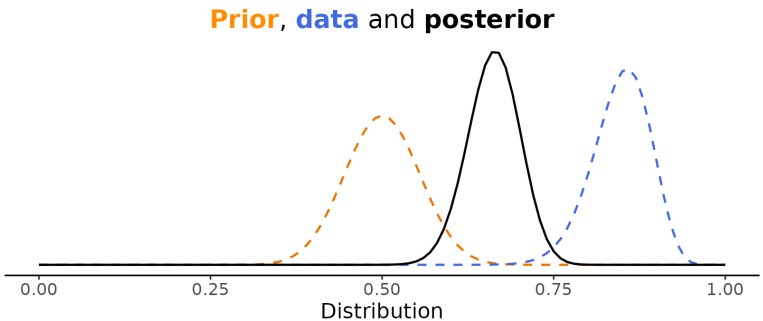


Figura 5: Um meio-termo entre crença e evidência.

Cenário 3

- Se o nosso *a priori* for extremo (distribuição estreita), nossa conclusão ficará mais "cega" às evidências:

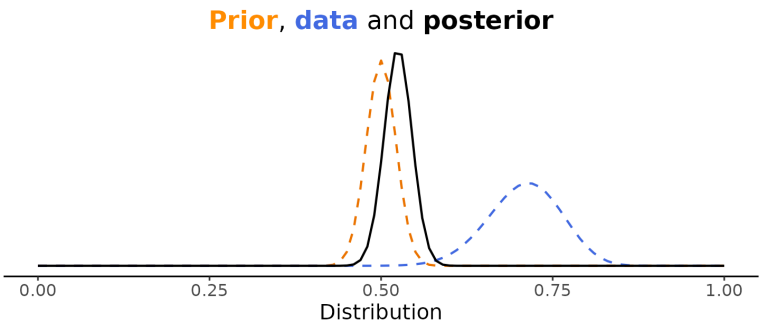


Figura 6: Uma conclusão que ignora as evidências.

☞ Mais simulações interativas **aqui**

Priori não é um "chute"

Uma distribuição *a priori* é uma suposição explícita sobre valores plausíveis (com base na **literatura**)

Boas funções da distribuição *a priori*

- Evitar valores absurdos
- Regularizar estimativas
- Tornar suposições visíveis

Riscos

- Escala mal pensada
- Distribuição *a priori* forte demais (cenário 3)
- Suposição escondida no modelo

A pergunta não é se temos suposições. A pergunta é se elas estão explícitas e se fazem sentido.

Como o modelo chega ao *a posteriori*?

- Em exemplos realistas, **não** conseguimos aplicar Bayes analiticamente
- O obstáculo está no denominador da regra de Bayes:

$$P(y) = \int P(y \mid \theta)\, P(\theta)\, d\theta$$

- Essa integral percorre todos os valores possíveis de θ
- Em modelos com vários parâmetros, ela não tem solução fechada

A saída: em vez de **calcular** a distribuição *a posteriori*, tiramos **amostras** dela

Amostrar em vez de calcular

Não precisamos da fórmula exata do *a posteriori*. Basta uma **coleção grande de amostras** vinda dele.

Analogia

Pesquisa de opinião: não entrevistamos toda a população, mas uma amostra grande o suficiente já descreve bem o todo.

Com as amostras, podemos:

- desenhar o histograma (a "forma" do *a posteriori*)
- calcular médias e medianas
- obter intervalos de credibilidade

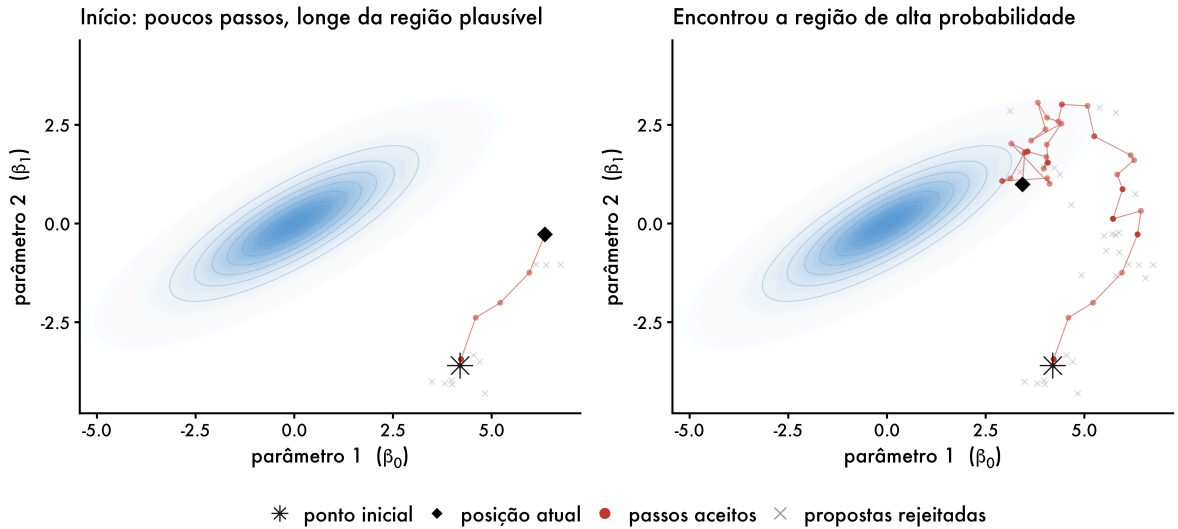
☞ Obter muitas amostras do *a posteriori* $\approx$ conhecer o *a posteriori*

Uma caminhada pelo espaço de parâmetros

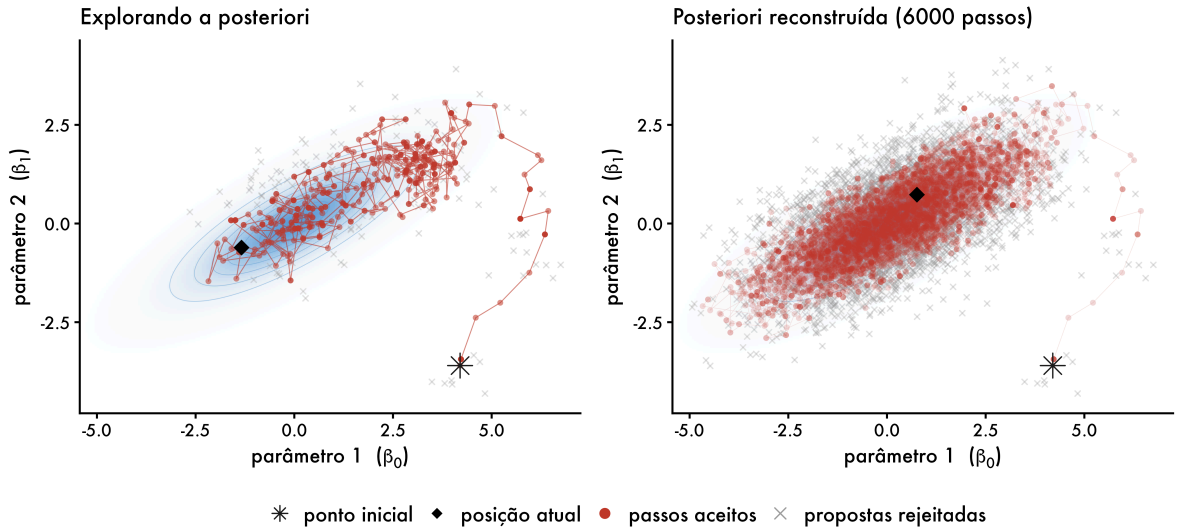
Como amostrar de uma distribuição que sequer conhecemos por completo? Soltamos um “explorador” para caminhar pelo espaço dos parâmetros (MCMC).

- A cada passo, ele propõe um valor novo e decide se o **aceita** ou **rejeita**
- Tende a passar mais tempo onde o *a posteriori* é **mais denso**
- As posições visitadas **são** a nossa amostra

Uma caminhada pelo espaço de parâmetros



Uma caminhada pelo espaço de parâmetros



A caminhada convergiu?

Antes de confiar nas amostras, olhamos a trajetória de cada parâmetro ao longo das iterações

- O começo (*burn-in*) é descartado: o explorador ainda estava se localizando
- Depois, a cadeia oscila numa faixa estável
- É isso que `Rhat` ($\hat{R} \approx 1$ e o `ESS` resumem em números

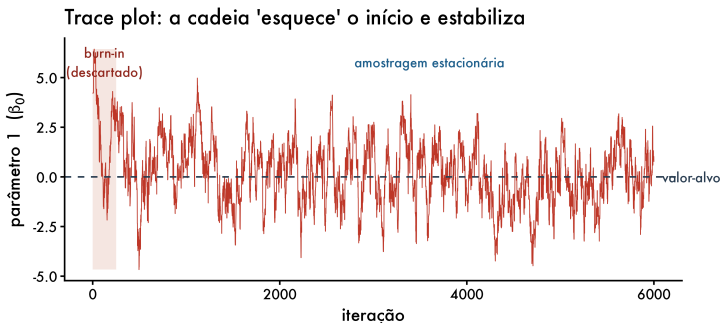


Figura 9: *Trace plot* para verificar convergência de cadeias

Nem toda caminhada é igual

Random walk

Passos curtos e aleatórios. Explora devagar e trava em modelos com muitos parâmetros.

HMC (o que `brms` /Stan usa)

Usa a **inclinação** do *a posteriori* para "deslizar" e explorar muito mais rápido.

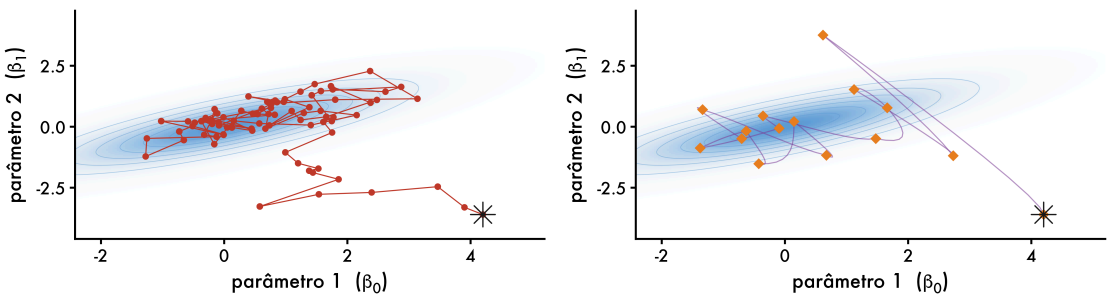


Figura 10: Random walk vs HMC

4 Regressão bayesiana

Modelo linear bayesiano (A)

Podemos escrever uma regressão simples assim:

$$y_i \sim \text{normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$

E completamos o modelo com distribuições *a priori*: perceba que podemos estimar σ também

$$\alpha \sim \text{normal}(0, 5)$$
$$\beta \sim \text{normal}(0, 2)$$
$$\sigma \sim \text{exponential}(1)$$

- Note que criaremos **um** *a posteriori* para σ: ainda estamos lidando com homoscedasticidade

O mesmo modelo em `brms`

```r
library(brms)

fit_lm <- brm(
  LogRT ~ Condition,
  data = dan,
  family = gaussian(),
  prior = c(
    prior(normal(0, 5), class = Intercept),
    prior(normal(0, 2), class = b),
    prior(exponential(1), class = sigma)
  )
)
```

☞ A fórmula parece familiar; o que muda é que estimamos distribuições posteriores.

Output

- Coeficientes em escala log (`LogRT` era a nossa variável de resposta original)
- Perceba que `sigma` [1] também é um parâmetro estimado pelo modelo (uma vantagem importante)

```
# Regression Coefficients:
#           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# Intercept     6.74      0.01     6.72     6.77 1.00     3044     2693
# ConditionB    0.05      0.02     0.01     0.09 1.00     3227     3066
# ConditionC    0.11      0.02     0.06     0.14 1.00     3142     2819
#
# Further Distributional Parameters:
#       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# sigma     0.20      0.01     0.19     0.21 1.00     3860     2845
```

- Mas o modelo supõe que as condições compartilham o mesmo σ (como em modelos frequentistas)
- Isso, contudo, muitas vezes *não é apropriado*

[1] Este σ é o desvio-padrão **residual**, então ele não será idêntico ao desvio-padrão marginal da variável `LogRT`.

Modelo linear bayesiano (B)

Vamos agora considerar uma regressão que supõe heteroscedasticidade e que usa tempos de reação em ms

Modelo

$$\begin{aligned} \text{rt}_i &\sim \text{lognormal}(\mu_i, \sigma_i) \\ \mu_i &= \alpha + \beta x_i \\ \log \sigma_i &= \gamma + \delta x_i \end{aligned}$$

Distribuições *a priori*

$$\begin{aligned} \alpha &\sim \text{normal}(6.5, 0.5) \\ \beta &\sim \text{normal}(0, 0.2) \\ \gamma &\sim \text{normal}(0, 1) \\ \delta &\sim \text{normal}(0, 0.5) \end{aligned}$$

- A diferença para o modelo anterior é só uma linha: $\log \sigma_i = \gamma + \delta x_i$
- O σ deixa de ser constante e passa a variar com x

O mesmo modelo em `brms`

```r
library(brms)

fit_het <- brm(
  bf(
    rt ~ Condition,
    sigma ~ Condition # <- sigma agora varia por condição
  ),
  data = dan,
  family = lognormal(), # <- agora usamos ms e modelamos a variável em escala log
  prior = c(
    prior(normal(6.5, 0.5), class = Intercept),
    prior(normal(0, 0.2), class = b),
    prior(normal(0, 1), class = Intercept, dpar = sigma),
    prior(normal(0, 0.5), class = b, dpar = sigma)
  )
)
```

☞ A fórmula não muda tanto, mas é preciso adicionar `bf()` (fórmula do `brms`) dentro de `brm()`

Output

- Perceba que `sigma` também é um parâmetro estimado pelo modelo, mas agora **para cada condição**
- O `brms` automaticamente aplica um link log a `sigma`, então os efeitos abaixo estão na escala log

```
# Regression Coefficients:
#                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# Intercept            6.74      0.01     6.72     6.77 1.00     4024     2730
# sigma_Intercept     -1.53      0.05    -1.62    -1.43 1.00     4491     2784
# ConditionB           0.05      0.02     0.01     0.09 1.00     4428     3235
# ConditionC           0.11      0.02     0.07     0.14 1.00     4446     3273
# sigma_ConditionB    -0.07      0.07    -0.20     0.07 1.00     4143     3064
# sigma_ConditionC    -0.15      0.07    -0.29    -0.01 1.00     4402     3108
```

- Exemplo: condição A $\rightarrow$ média estimada em 6.74 e desvio-padrão estimado em 0.22 ($e^{-1.53}$)
- Exemplo: condição B $\rightarrow$ média estimada em 6.74+0.05 e desvio-padrão estimado em 0.20 ($e^{-1.53-0.07}$)

Ver os dados ↑

Conclusão e ideias finais

Três pontos importantes:

1. Modelos bayesianos nos dão uma visão mais **intuitiva** de como modelamos nossos dados
2. Distribuições *a priori* nos permitem conectar teoria e análise
3. A interpretação de nossos modelos se torna mais **direta** e **nuançada**

☞ Para aprofundar: McElreath, Kruschke e a documentação/ecossistema de `brms` (Bürkner, 2017; Kruschke, 2015; McElreath, 2020). Para comparações múltiplas, ver Garcia (2025).

Referências

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by M r. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Bürkner, P.-C. (2017). `brms`: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver, Boyd.

Garcia, G. D. (2025). Bayesian Estimation in Multiple Comparisons. *Studies in Second Language Acquisition*, *47*(3). https://doi.org/10.1017/s0272263125100922

Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and S tan* (2.º ed.). Academic Press.

Laplace, P.-S. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences de Paris (Savants étrangers)*, *6*, 621–656.

Laplace, P.-S. (1812). *Théorie analytique des probabilités*. Courcier.

McElreath, R. (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan* (2.º ed.). Chapman & Hall/CRC.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

5 Apêndice

Mas onde *está* o *a posteriori*?

A dúvida: se não conseguimos **calcular** o *a posteriori*, como amostramos de algo que não temos?

Estamos confundindo duas operações bem diferentes:

Avaliar num ponto → fácil

Para um θ específico, sei a "altura" do *a posteriori*:

$$P(y \mid \theta) \times P(\theta)$$

(*likelihood* × priori — fórmulas que já vimos)

Normalizar → difícil

Dividir por

$$P(y) = \int P(y \mid \theta)\, P(\theta)\, d\theta$$

a integral sobre todos os valores possíveis de θ.

☞ Sabemos a **altura** do *a posteriori* em qualquer ponto; só não sabemos a sua **área total**.

O que não sabemos calcular não faz falta

O amostrador nunca usa a altura absoluta — ele só **compara** dois pontos. A decisão de ir de θ para θ' depende de uma razão:

$$\frac{P(\theta' \mid y)}{P(\theta \mid y)} = \frac{P(y \mid \theta')\,P(\theta')}{P(y \mid \theta)\,P(\theta)}$$

- O denominador $P(y)$ aparece em cima *e* embaixo da razão — e se cancela
- Ou seja: a única coisa que não conseguimos calcular é justamente a que **não faz falta**

☞ O histograma das amostras se normaliza sozinho: basta contar a frequência das visitas.

Mapeando a montanha no nevoeiro

O *a posteriori* não está guardado em lugar nenhum: é uma **superfície implícita** (priori × *likelihood*). Não vemos o mapa inteiro, mas sentimos o chão onde pisamos.

A analogia

- Você é largado numa serra com nevoeiro
- Não enxerga o mapa (= a integral $P(y)$)
- Mas um altímetro dá a altitude onde você pisa (= *a priori* × *likelihood*)
- Você anda, compara alturas, sobe mais do que desce e fica mais tempo no alto

☞ Essas pegadas **são** a nossa amostra do *a posteriori*.

O resultado

Depois de milhares de passos, suas **pegadas** desenham a serra — mais densas onde ela é mais alta.

Você mapeou a montanha sem nunca ter visto o mapa, usando só a altitude sob os pés.