Bayesian statistics

Các khái niệm cơ bản

Khương Quỳnh Long
Hà Nội, 08/2019
https://gitlab.com/LongKhuong/adhere-bayesian-statistics

Nội dung

1. Thống kê bayes là gì?
2. Định nghĩa xác suất bayes
3. Confidence interval vs credible interval
4. Định lý Bayes

1. Thống kê Bayes là gì?

Thống kê Bayes là gì?

Frequentist vs. Bayesian

Objective vs. Subjective

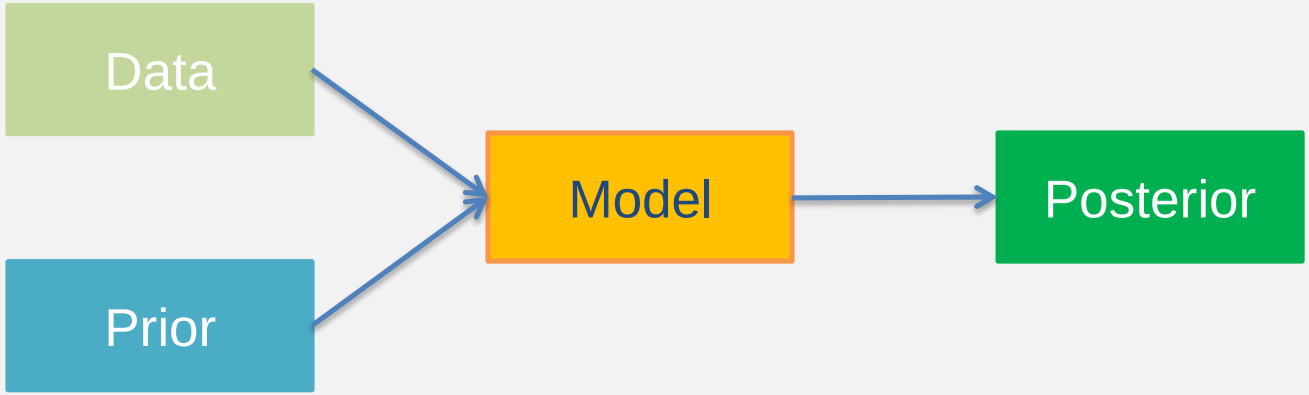
P values vs. Posterior probabilities

Thống kê Bayes là gì?

- Một “trường phái” (lĩnh vực) thống kê dựa vào suy luận Bayes về xác suất
- (Không phải một mô hình hay một phép kiểm cụ thể)
- Sử dụng xác suất để diễn tả mức độ chắc chắn/không chắc chắn (uncertainty) cho toàn bộ thành phần của mô hình

Thống kê Bayes là gì?

- Tính toán, cập nhật (update) thông tin hậu định (posterior) dựa vào dữ liệu thu thập và thông tin tiền định (prior)



Định nghĩa xác suất theo Frequentist và Bayesian

Xác suất xuất hiện mặt sấp khi tung đồng xu là 50%

Xác suất thuốc A tốt hơn thuốc B là 70%

Nghĩa là?

Xác suất theo Frequentist

- Tần số tương đối "về lâu về dài" (long-run) của các giá trị của biến số
- Ví dụ: tung đồng xu 500 lần, số lần xuất hiện mặt sấp ~ 250 lần (50%)

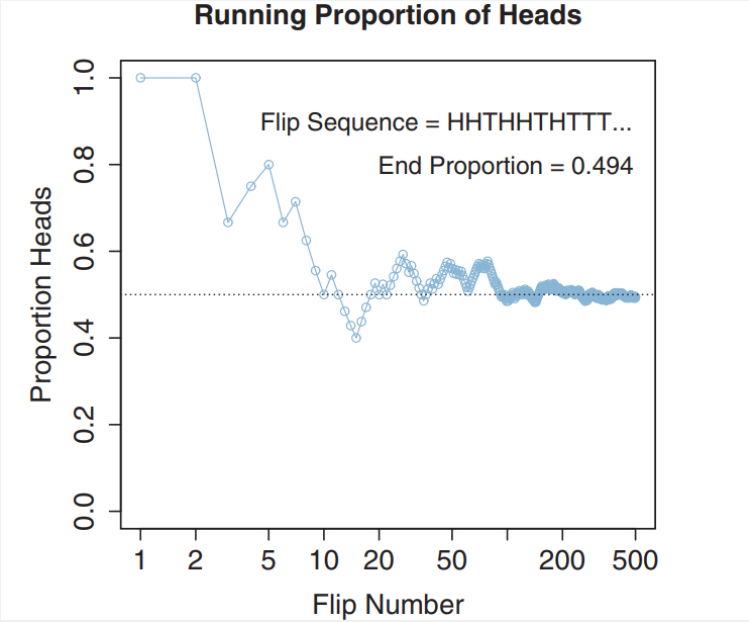
$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

- Trong đó:

n là số lần biến cố A xảy ra trong N lần thử

➔ Suy luận theo tần số mẫu

➔ Không "cá thể hóa"



Xác suất theo Bayesian

- Liên quan đến mức độ tin tưởng (degree of belief)
- Ví dụ: xác suất suất hiện mặt sấp khi tung đồng xu là 50% → xác suất xuất hiện mặt sấp trong 1 lần tung đồng xu là 50%!
- P(A) = P
- Gần với thực tế, “cá thể hóa”.
- “subjective”

95% Confidence interval
Vs. 95% Credible interval

95% Confidence interval

Nghiên cứu trên 10,000 người,
tỉ lệ ung thư là 10% (KTC 95% = 8% – 12%)

Nghĩa là?

95% Confidence interval

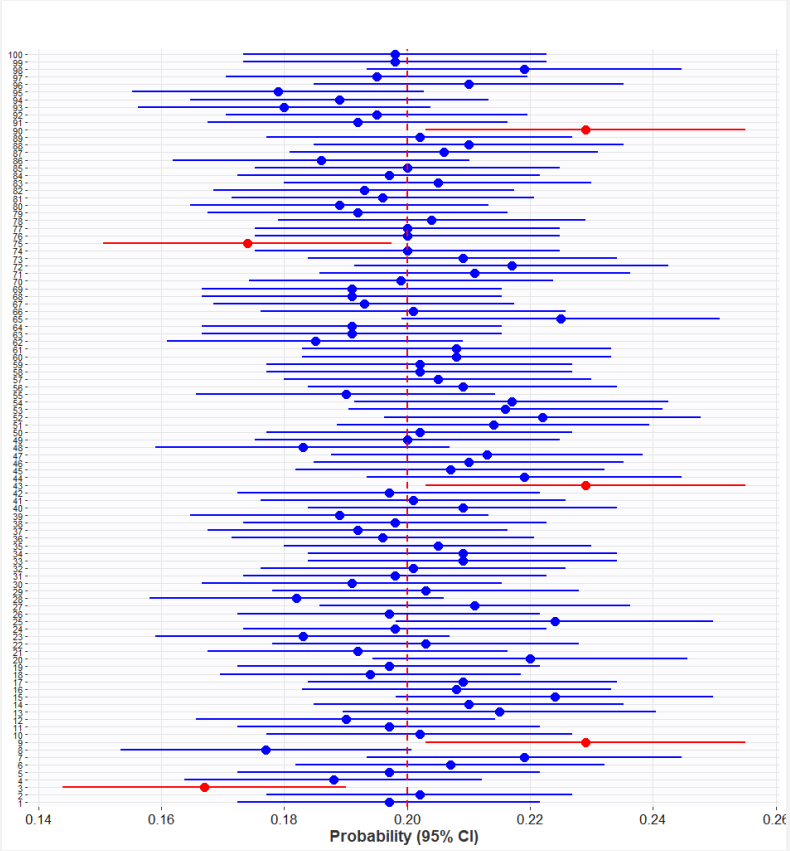
- 95% khoảng tin cậy sẽ chứa giá trị thật của dân số?
- Xác suất để giá trị thật của dân số nằm trong khoảng này là 95% ?

95% Confidence interval

- Theo Frequentist, giá trị thật của dân số là thật, nhưng không biết (unknown) và cố định (fixed).
- Nếu lặp lại nghiên cứu tương tự n lần, mỗi lần tính 1 CI, trung bình 95% số n lần tạo thành khoảng CI (từ … đến …) chứa giá trị thật của dân số.
- 95%CI là một quá trình "long-run" (95% của n lần), không cụ thể cho một nghiên cứu nào.

➔ Cho một nghiên cứu cụ thể, giá trị thật của dân số có thể nằm trong 95%CI hoặc không !

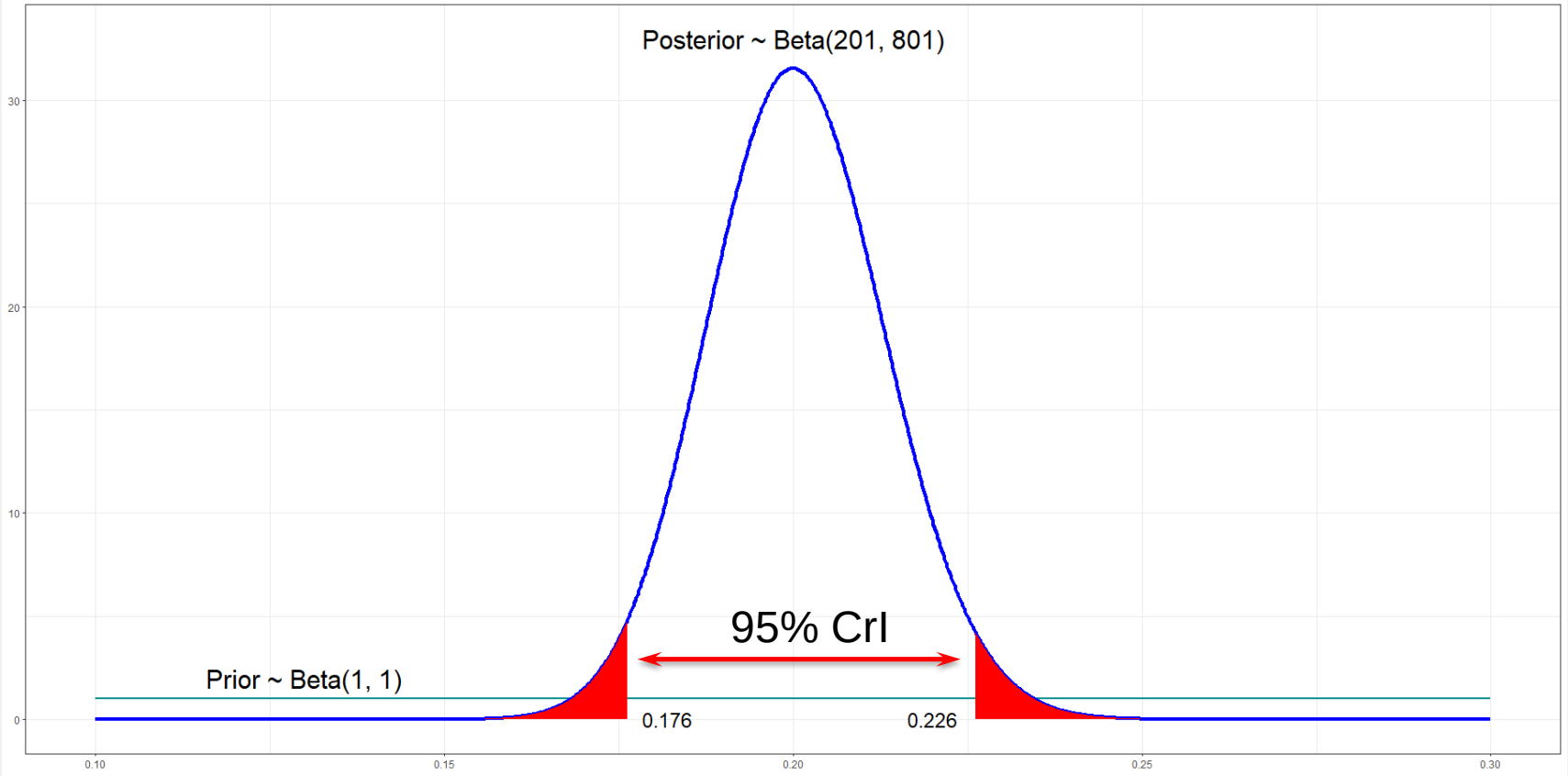
95% Confidence interval



95% Credible interval

- Theo Bayesian, giá trị thật của dân số là không biết (unknown) do đó không cố định → tuân theo một phân phối. Phân phối này được ước lượng từ prior và dữ liệu.
- Xác suất Bayes phản ánh “degree of belief” do đó với 95% credible interval có thể nói 95% giá trị thật của dân số nằm trong khoảng này

95% Credible interval



Định lý Bayes
(Bayes' Rule)

Xác suất có điều kiện

	Nhóm tuổi				
Bệnh	<18	18–49 B	50–64	65+	**Tổng**
Có A	50	60	80	100	290
Không	200	200	250	250	900
Tổng	250	260	330	350	1190

Conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of event A occured and event B occured

Probability of event A given B has occured

Probability of event B

- Xác suất người mắc bệnh trong độ tuổi từ 18-49 = 60/260 = 23.1%
- hay P(Bệnh | Tuổi18-49) = P(Bệnh và Tuổi18-49) / P(Tuổi18-49)

Định lý Bayes

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \rightarrow \quad P(A \mid B)\,P(B) = P(A \cap B)$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad \rightarrow \quad P(B \mid A)\,P(A) = P(A \cap B)$$

$$P(A \mid B)\,P(B) = P(B \mid A)\,P(A)$$

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Định lý Bayes

Likelihood
Độ khả dĩ của dữ liệu: Xác suất thu thập được dữ liệu dưới điều kiện giả thuyết H đúng (Đơn giản là dữ liệu thu thập)

Prior
Xác suất tiền định: Xác suất giả thuyết H mà chúng ta tin là xảy ra (đúng) trước khi thu thập dữ liệu

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)}$$

Posterior
Xác xuất hậu định: Cần tìm Xác suất giả thuyết H đúng cho bởi dữ liệu thu thập

Marginal likelihood (normalizing constant)
Hằng số: Xác suất của dữ liệu (tổng tất cả likelihood*prior của tất cả các giả thuyết)

Định lý Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)}$$

$$P(\theta|D) = \frac{P(D|\theta) * P(\theta)}{P(D)}$$

$$P(model_1|D) = \frac{P(D|model_1) * P(model_1)}{P(D)}$$

Thank you !