Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2023.234

ĐÁNH GIÁ CỠ MẪU CHO ƯỚC LƯỢNG THAM SỐ TRONG NHỮNG MÔ HÌNH CẤU TRÚC GIAI ĐOẠN

Phạm Thị Thu Hoa* và Phạm Thị Thu Hường
Khoa Sư phạm, Trường Đại học An Giang, Đại học Quốc gia Thành phố Hồ Chí Minh
**Tác giả liên hệ (Corresponding author): ptthoa@agu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 20/08/2023
Sửa bài (Revised): 21/09/2023
Duyệt đăng (Accepted): 19/10/2023

Title: *Evaluation of sample size for estimation in stage structured development models*

Author(s): *Pham Thi Thu Hoa* and Pham Thi Thu Huong*

Affiliation(s): *Vietnam National University, Ho Chi Minh City*

TÓM TẮT

Mô hình cấu trúc giai đoạn nghiên cứu quá trình phát triển được phân chia theo từng giai đoạn. Mô hình này rất phổ biến trong nghiên cứu sự phát triển của các loại bệnh và sự phát triển sinh học của thực vật và động vật. Cách tiếp cận Bayes nhúng phép biến đổi tham số vào thuật toán Metropolis-Hastings được sử dụng để ước lượng các tham số cho các mô hình này cho đến nay được đánh giá là cách tiếp cận chính xác với các nghiên cứu thực nghiệm. Mục đích chính của bài viết là áp dụng phương pháp tiếp cận Bayes ước lượng tham số trong mô hình cấu trúc giai đoạn không xuất hiện tỷ lệ chết cho các nghiên cứu mô phỏng để xác định cỡ mẫu thích hợp cho mô hình cấu trúc với các giai đoạn cho trước. Kết quả của việc đánh giá cỡ mẫu này được áp dụng cho dữ liệu thời gian ủ bệnh của COVID-19. Nghiên cứu trên dữ liệu này được xem là sự tiếp nối của các nghiên cứu trước và có ý nghĩa trong công tác phòng chống đại dịch.

Từ khoá: *Dữ liệu phát triển theo giai đoạn, dữ liệu tần số theo giai đoạn, giai đoạn ủ bệnh của COVID-19, mô hình cấu trúc giai đoạn, phân tích Bayes*

ABSTRACT

Stage-structured models consider development processes which are divided by different developmental stages. These models are common statistical models in disease progressions and biological development of plants and animals. A Bayesian approach based on deterministic transformations in the Metropolis-Hastings algorithm is considered the most accurate estimate to estimate parameters for these stage-structured models. The main purpose of this paper is to evaluate the appropriate sample size for the given stage-structured model by applying the Bayesian approach to estimate parameters. The results of the proposed sample size assessment method are very useful in finding out the sample size at each sampling time and the appropriate number of sampling times for designing experiments. The results of this sample size assessment are also applied to COVID-19 incubation period data. This study on the COVID-19 incubation period is a continuation of previous studies and has implications for pandemic prevention.

Keywords: *Bayesian analysis, incubation period of COVID-19, multi-stage models, stage duration data, stage frequency data*

1. GIỚI THIỆU

Mô hình cấu trúc giai đoạn được áp dụng cho các sự phát triển được phân ra thành các giai đoạn phát triển khác nhau. Một cá thể sẽ phát triển qua từng giai đoạn và không bỏ qua một giai đoạn nào. Có nhiều quá trình trong tự nhiên được thể hiện bằng mô hình cấu trúc giai đoạn như sự phát triển của các loại vi khuẩn hay sự phát triển của các loại động vật và thực vật. Đặc biệt, các mô hình cấu trúc giai đoạn rất phổ biến trong các nghiên cứu ở các phòng thí nghiệm cho các quá trình phát triển có cấu trúc giai đoạn (Read & Ashford, 1968; Schuh & Tweedie, 1979; Hoeting et al., 2003; Knape et al., 2014; De Valpine & Knape, 2015; Knape & De Valpine, 2016). Dữ liệu của mô hình là số lượng cá thể trong từng giai đoạn phát triển tại mỗi thời điểm lấy mẫu. Trong các mô hình cấu trúc giai đoạn, thời gian chuyển từ giai đoạn này đến giai đoạn kế tiếp là không thể xác định được. Do đó xác suất để một cá thể sống sót tại thời gian t ở giai đoạn nào đó trong quá trình phát triển là tích chập của xác suất sống sót ở các giai đoạn trước đó trong khoảng thời gian t (Pham & Branford, 2016; Pham et al., 2019; Pham & Pham, 2019).

Một số nghiên cứu trước đây đã đưa ra các phương pháp ước lượng tham số cho mô hình cấu trúc giai đoạn (Knape et al., 2014; Knape & De Valpine, 2016; Pham & Branford, 2016; Pham et al., 2019; Pham & Pham, 2019). Trong đó có cách tiếp cận Bayes nhúng phép biến đổi tham số vào thuật toán Metropolis-Hastings (MH) được sử dụng để ước lượng các tham số trong các mô hình cấu trúc giai đoạn (Pham et al., 2019; Pham & Pham, 2019). Các phương pháp ước lượng này cho đến nay được xem là có cách tiếp cận chính xác hơn vì trong mô hình thời gian chuyển từ giai đoạn này đến giai đoạn kế tiếp được mô hình giống với thực nghiệm là không thể xác định được. Trong bài báo này, các phương pháp ước lượng này được áp dụng để đánh giá cỡ mẫu cho mô hình cấu trúc giai đoạn không xuất hiện tỷ lệ tử vong có 3 giai đoạn, 5 giai đoạn và 7 giai đoạn. Các nghiên cứu mô phỏng được áp dụng để chọn ra cỡ mẫu thích hợp cho mô hình cấu trúc có 3 giai đoạn, 5 giai đoạn và 7 giai đoạn. Việc này có ý nghĩa trong việc chọn cỡ mẫu tại các lần lấy mẫu và số thời điểm lấy mẫu thích hợp cho mô hình thực nghiệm trước khi tiến hành thí nghiệm.

Kết quả của việc đánh giá cỡ mẫu của các mô hình cấu trúc giai đoạn được áp dụng cho dữ liệu thời gian ủ bệnh của COVID-19 (Backer et al., 2020; Yin et al., 2021) và áp dụng cách tiếp cận Bayes để ước lượng tham số cho giai đoạn ủ bệnh của COVID-19. Việc biết được quá trình phát triển của giai đoạn ủ bệnh của virus COVID-19 có ý nghĩa quan trọng trong việc điều tra dịch tễ học để phòng chống sự lan rộng của dịch bệnh. Hiểu rõ thời gian ủ bệnh giúp giám sát tích cực những người có mức độ phơi nhiễm cao và giúp xác định thời gian cách ly đối với các bệnh nhân nhiễm bệnh COVID-19. Kiến thức về quá trình phát triển của giai đoạn ủ bệnh rất cần thiết trong việc ước tính khả năng lây truyền của dịch bệnh và cải thiện công tác phòng chống dịch bệnh (Backer et al., 2020; Goel & Kumar, 2020; Mingyue et al., 2020; Wang et al., 2020; Rai et al., 2021).

Dữ liệu thời gian ủ bệnh của COVID-19 được mô hình là dữ liệu cấu trúc gồm 2 giai đoạn bao gồm giai đoạn tiếp xúc với mầm bệnh và giai đoạn ủ bệnh. Thời gian chuyển từ giai đoạn tiếp xúc với mầm bệnh đến giai đoạn ủ bệnh cũng như từ giai đoạn ủ bệnh đến giai đoạn tiếp theo của COVID-19 là không xác định được. Từ dữ liệu gốc, dữ liệu 98 bệnh nhân nhiễm COVID-19 được quan sát trong 22 ngày. Dữ liệu được thu thập bao gồm thông tin của giai đoạn tiếp xúc với mầm bệnh và giai đoạn ủ bệnh nằm trong một khoảng thời gian được các bệnh nhân cung cấp. Mỗi ngày số bệnh nhân hết thời gian phơi nhiễm và xuất hiện triệu chứng được thu thập lại. Cần lưu ý rằng, các phương pháp và cách đánh giá cỡ mẫu được đề xuất trong bài báo có thể được áp dụng để phân tích thêm các giai đoạn khác của virus COVID-19 cũng như các giai đoạn phát triển của các loại virus gây bệnh khác.

2. PHƯƠNG PHÁP TIẾP CẬN BAYES CHO ƯỚC LƯỢNG THAM SỐ TRONG MÔ HÌNH CẤU TRÚC GIAI ĐOẠN

Các mô hình cấu trúc giai đoạn được đề cập là mô hình không xuất hiện tỷ lệ tử vong trong đó điểm khởi đầu quá trình phát triển của các cá thể được kiểm soát (Pham et al., 2019; Pham & Pham, 2019). Trong các mô hình cấu trúc giai đoạn này, quá trình phát triển của mỗi cá thể được biến đổi qua I giai đoạn được phân chia trước. Thời gian để một cá thể chuyển từ một giai đoạn sang giai đoạn tiếp theo là không thể xác định được. Trong lần lấy mẫu đầu tiên, chúng ta cho rằng tất cả các cá thể bắt đầu ở giai đoạn 1. Ở mỗi lần lấy mẫu, dữ liệu được thu thập là số lượng các cá thể sống sót ở các giai đoạn khác nhau. Dữ liệu mô phỏng mô hình cấu trúc 3 giai đoạn, có 15 thời điểm lấy mẫu, cỡ mẫu là 50 cá thể ở mỗi thời điểm lấy mẫu được trình bày ở Bảng 1.

Bảng 1. Dữ liệu mô hình có 3 giai đoạn với 50 cá thể ở mỗi thời gian thu mẫu

t	Giai đoạn 1	Giai đoạn 2	Giai đoạn 3	Giai đoạn cuối
0,1	50	0	0	0
0,5	30	20	0	0
0,9	23	21	6	0
1,4	12	28	9	1
1,8	9	24	16	1
2,2	1	22	18	9
2,6	3	14	21	12
3,1	2	10	21	17
3,5	0	9	19	22
3,9	0	8	15	27
4,3	0	1	9	40
4,7	0	1	9	40
5,2	0	0	3	47
5,6	0	1	1	48
6,0	0	0	4	46

2.1. Mô hình không xuất hiện tỷ lệ tử vong

Phân phối của giai đoạn thứ $j, j=1,...,I$, trong quá trình phát triển được mô hình theo phân phối Gamma với tham số hình dạng và vị trí (a_j, λ_j). Đặt $\theta=(a_1,\lambda_1,...,a_I,\lambda_I)$ là tập hợp các tham số trong I giai đoạn phát triển. Tại thời điểm lấy mẫu $t_k, k=1,...,K$, số lượng các cá thể sống sót ở giai đoạn thứ j là $N_j(t_k)=N_{kj}$, $j=1,...,I$ và tổng số lượng cá thể sống sót ở thời gian $t_k, k=1,...,K$ được định nghĩa là $N_k=\sum_{j=1}^{I+1}N_j(t_k)$. Tại thời điểm lấy mẫu $t_k, k=1,...,K$, dữ liệu được quan sát là số lượng các cá thể sống sót ở các giai đoạn được ký hiệu là $y_k=(N_1(t_k),N_2(t_k),...,N_{I+1}(t_k))$.

Do thời gian chuyển giữa các giai đoạn là không xác định nên xác suất để một cá thể sống sót ở giai đoạn $j, j=1,...,I$ ở thời gian t được xác định bởi các tích chập như sau:

$$p_j(t)=h_1*h_2*...*h_{j-1}*H_j(t), \quad j=1,...,I \quad (1)$$

ta có trong mô hình cấu trúc giai đoạn không xuất hiện tỷ lệ tử vong hàm mật độ của một cá thể sống sót ở thời điểm *t*, giai đoạn thứ *i* $h_i(t)$ được mô hình là phân phối Gamma $g_i(t)$ như sau:

$$h_i(t)=g_i(t)=t^{a_i-1}\exp(-\lambda_i t)\lambda_i^{a_i}/\Gamma(a_i), i=1,...,j-1$$

là hàm mật độ phát triển ở giai đoạn thứ *i* và $H_j(t)=\int_t^\infty g_j(x)dx$.

Tại thời điểm lấy mẫu $t_k, k=1,...,K$, y_k có phân phối đa thức như sau:

$$\begin{aligned}&(y_k \mid N_k,\theta)\sim \text{Multinomial}\\&(N_k, p_1(t_k), p_2(t_k),..., p_{I+1}(t_k))\end{aligned} \quad (2)$$

trong đó $p_{I+1}(t_k)=1-\sum_{j=1}^{I}p_j(t_k)$. Hàm hợp lý của tất cả các thông tin được quan sát có dạng

$$f(y\mid\theta)=\prod_{k=1}^{K}\text{p}(N_k, p_1(t_k), p_2(t_k),..., p_{I+1}(t_k)). \quad (3)$$

2.2. Thuật toán Bayes

Phương pháp nhúng phép biến đổi tham số vào thuật toán MH (Read & Ashford, 1968; Robert & Casella, 2009; Pham et al., 2019; Pham & Pham, 2019) được sử dụng để ước lượng các tham số trong các giai đoạn. Giả sử rằng các tham số trong mô hình $\theta=(a_1,\lambda_1,...,a_I,\lambda_I)$ là không có thông tin tiên nghiệm, do đó phân phối tiên nghiệm của các tham số được chọn là có phân phối đều. Ta có hàm phân phối hậu nghiệm có dạng:

$$\begin{aligned}\pi(\theta\mid y)\propto\ & f(y\mid\theta)p(\theta)\\ \propto\ & \prod_{k=1}^{K}f(y_k\mid\theta)\prod_{m=1}^{2I}p(\theta_m)\\ \propto\ & \prod_{k=1}^{K}\text{p}(N_k, p_1(t_k), p_2(t_k),...,\\ & p_{I+1}(t_k))\prod_{m=1}^{2I}p(\theta_m),\end{aligned} \quad (4)$$

trong đó giá trị của các hàm $p_j(t)$, $j=1,...,I$ tính được bằng cách lấy hàm ngược của hàm Laplace như sau:

$$\begin{aligned}\mathcal{L}(p_j(t))&=\int_0^\infty p_j(t)\exp(-st)dt\\&=\left(\frac{\lambda_1}{\lambda_1+s}\right)^{a_1}....\left(\frac{\lambda_{j-1}}{\lambda_{j-1}+s}\right)^{a_{j-1}}\frac{1-\beta_j(s)}{s},\end{aligned} \quad (5)$$

tham số *s* được xác định là giá trị trung bình của giai đoạn phát triển đang đề cập và

$\beta_j(s) = \mathcal{L}(h_j) = \int_0^\infty g_j(t)\exp(-st)dt$. Xác suất chấp nhận trong thuật toán MH được tính bằng công thức

$$\alpha \propto \min\left(1, \frac{\pi\left(a_j^{(*)}, \lambda_j^{(*)} \mid y_i\right) q\left(a_j^{(t)} \mid a_j^{(*)}\right)}{\pi\left(a_j^{(t)}, \lambda_j^{(t)} \mid y_i\right) q\left(a_j^{(*)} \mid a_j^{(t)}\right)}\right), \quad (6)$$

trong đó phân phối $q\left(a_j^{(t)} \mid a_j^{(*)}\right)$ được chọn từ những bước ngẫu nhiên của thuật toán MH. Mối liên hệ giữa tham số a_j và λ_j được xác định như sau (Pham et al., 2019)

$$a_j^{(*)}(s) = \frac{\log \hat{\beta}_j(s)}{\log[\hat{\lambda}_j^{(*)} / (\hat{\lambda}_j^{(*)} + s)]}. \quad (7)$$

Các tham số của mô hình được cập nhật từ giai đoạn 1 đến giai đoạn *I* qua thuật toán MH được trình bày trong Bảng 2.

Bảng 2. Thuật toán MH để cập nhật các tham số ở giai đoạn thứ *j*

Bước 1	Chọn giá trị ban đầu cho các tham số. Vì các tham số trong mô hình là không có thông tin tiên nghiệm, nên các giá trị ban đầu này được chọn tùy ý.
Bước 2	Chọn số vòng lặp T của thuật toán, vòng lặp được chạy từ $t = 1$ cho đến T.
Bước 3	Giả sử giá trị đang có tại bước t là $a^{(t)}$, đề xuất giá trị mới là $a^{(*)} \sim q\left(a^{(*)} \mid a^{(t)}\right)$.
Bước 4	Với 2 giá trị $a^{(t)}$ và $a^{(*)}$ tính giá trị $\lambda^{(t)}$ và $\lambda^{(*)}$ sử dụng công thức (7).
Bước 5	Tính xác suất chấp nhận α sử dụng công thức (6).
Bước 6	Đặt $\left(a^{(t+1)}, \lambda^{(t+1)}\right) = \left(a^{(*)}, \lambda^{(*)}\right)$ với xác suất α, ngược lại ta đặt $\left(a^{(t+1)}, \lambda^{(t+1)}\right) = \left(a^{(t)}, \lambda^{(t)}\right)$.
Bước 7	Kết thúc vòng lặp.

3. NGHIÊN CỨU MÔ PHỎNG

Nghiên cứu mô phỏng được tiến hành để tìm ra cỡ mẫu thích hợp cho mô hình cấu trúc với các giai đoạn cho trước. Phương pháp nhúng phép biến đổi tham số vào thuật toán MH ở phần 2 được sử dụng để ước lượng tham số cho các mô hình cấu trúc không có tỷ lệ tử vong có 3 giai đoạn, 5 giai đoạn và 7 giai đoạn. Các giai đoạn phát triển trong mô hình được mô phỏng theo phân phối Gamma. Các mô hình được mô phỏng với các cỡ mẫu khác nhau. Các cỡ mẫu được giữ lại là các cỡ mẫu có hiệu suất tin cậy lớn hơn 90%, tức là tính tỷ lệ của các giá trị trung bình ước lượng nằm trong khoảng tin cậy từ 2,5% đến 97,5%. Các tỷ lệ phần trăm được tính toán này là lớn hơn 90%, đặc biệt ở các giai đoạn cuối.

Mỗi dãy Markov theo phương pháp Monte Carlo (MCMC) với cách tiếp cận Bayes ở phần 2, được chạy với số vòng lặp là 100.000 và được loại bỏ 10.000 vòng lặp ban đầu để tăng tính hội tụ của dãy MCMC. Biểu đồ vết và bảng tương quan ở các dãy MCMC được đánh giá để xét tính hội tụ của các dãy đó. Ngoài ra kiểm định Gelman and Rubin cũng được sử dụng để đánh giá sự hội tụ của các dãy này. Giá trị ước lượng các tham số trong mô hình là giá trị trung bình của 50 lần mô phỏng.

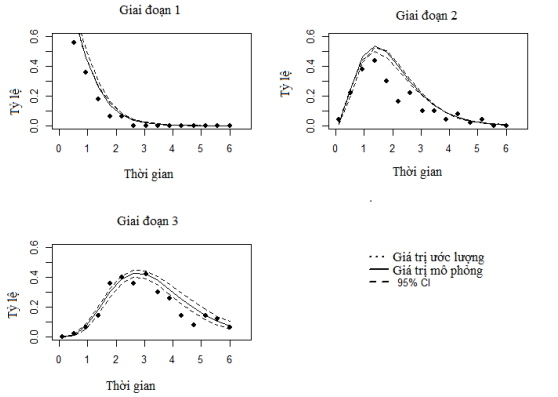
3.1. Mô hình có 3 giai đoạn

Mô hình có 3 giai đoạn phát triển được mô phỏng theo phân phối Gamma với tham số tham số hình dạng và vị trí $(a_1, \lambda_1) = (2,2)$, $(a_2, \lambda_2) = (2,1.7)$ và $(a_3, \lambda_3) = (2,1.4)$ lần lượt ở 3 giai đoạn. Mô hình được mô phỏng với 15 thời điểm lấy mẫu được chia đều từ 0 đến 6 (Bảng 1). Có 10 mô phỏng được tiến hành với các cỡ mẫu 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 cá thể ở mỗi thời điểm lấy mẫu. Sau khi ước lượng so sánh hiệu suất tin cậy cỡ mẫu mô phỏng với 30 cá thể ở mỗi thời điểm lấy mẫu được chọn được có hiệu suất tin cậy ở cả 3 giai đoạn đều lớn hơn 90%. Tuy nhiên hiệu suất tin cậy đối với mô phỏng có cỡ mẫu 50 cá thể ở mỗi thời điểm lấy mẫu có giá trị lớn hơn 98% ở cả 3 giai đoạn.

Kết quả ước lượng tham số từ 50 dữ liệu mô phỏng cho mô hình 3 giai đoạn với 50 cá thể ở mỗi thời điểm lấy mẫu được trình bày ở Bảng 3. Giá trị ước lượng là giá trị trung bình của 50 dữ liệu mô phỏng. Các tham số ước lượng ở 3 giai đoạn có giá trị gần với giá trị thực. Hiệu suất tin cậy với cỡ mẫu 50 cá thể ở mỗi thời điểm lấy mẫu có giá trị rất đáng tin cậy. So sánh tỷ lệ ước lượng và tỷ lệ mô phỏng ở 3 giai đoạn được trình bày ở Hình 1. Tại 15 thời điểm lấy mẫu, tỷ lệ ước lượng rất gần với tỷ lệ mô phỏng.

Bảng 3. Kết quả ước lượng tham số của mô hình có 3 giai đoạn

Tham số	Giá trị thực	Giá trị ước lượng	Hiệu suất tin cậy
a_1	2,0	2,1	99%
λ_1	2,0	2,0	99%
a_2	2,0	2,2	99%
λ_2	1,7	1,7	98%
a_3	2,0	2,2	97%
λ_3	1,4	1,5	98%



Hình 1. Tỷ lệ thực và tỷ lệ ước lượng của mô hình có 3 giai đoạn

3.2. Mô hình có 5 giai đoạn

Mô hình có 5 giai đoạn phát triển được mô phỏng theo phân phối Gamma với tham số tham số hình dạng và vị trí $(a_1,\lambda_1)=(2,2)$, $(a_2,\lambda_2)=(2,1.7)$, $(a_3,\lambda_3)=(2,1.4)$, $(a_4,\lambda_4)=(2,1.3)$ và $(a_5,\lambda_5)=(2,1.1)$ lần lượt ở 5 giai đoạn. Mô hình được mô phỏng với 15 thời điểm lấy mẫu được chia đều từ 0 đến 6. Có 10 mô phỏng được tiến hành với các cỡ mẫu 50, 100, 150, 200, 250, 300, 350, 400, 450, 500 cá thể ở mỗi thời điểm lấy mẫu. Sau khi ước lượng so sánh hiệu suất tin cậy cỡ mẫu mô phỏng với 100 cá thể ở mỗi thời điểm lấy mẫu được chọn có hiệu suất tin cậy ở cả 3 giai đoạn đều lớn hơn 90%.

Kết quả ước lượng tham số từ 50 dữ liệu mô phỏng cho mô hình 5 giai đoạn với 100 cá thể ở mỗi thời điểm lấy mẫu được trình bày ở Bảng 4. Giá trị ước lượng là giá trị trung bình của 50 dữ liệu mô phỏng. Các tham số ước lượng ở 3 giai đoạn có giá trị gần với giá trị thực. Hiệu suất tin cậy với cỡ mẫu 50 cá thể ở mỗi thời điểm lấy mẫu có giá trị rất đáng tin cậy. Tỷ lệ ước lượng và tỷ lệ mô phỏng ở 5 giai đoạn được so sánh qua Hình 2. Tại 15 thời điểm lấy mẫu, tỷ lệ ước lượng rất gần với tỷ lệ mô phỏng.

Bảng 4. Kết quả ước lượng tham số của mô hình có 5 giai đoạn

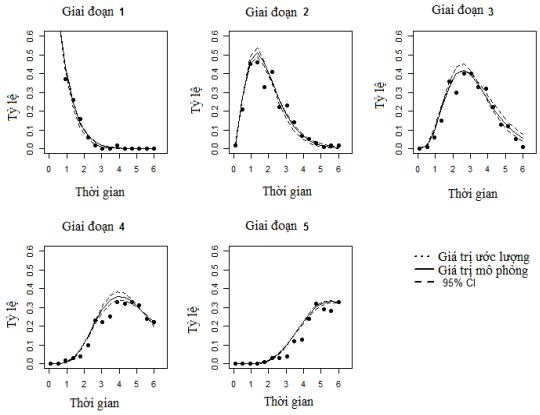
Tham số	Giá trị thực	Giá trị ước lượng	Hiệu suất tin cậy
a_1	2,0	1,9	99%
λ_1	2,0	1,9	99%
a_2	2,0	2,1	98%
λ_2	1,7	1,7	99%
a_3	2,0	2,2	98%
λ_3	1,4	1,5	95%
a_4	2,0	1,9	95%
λ_4	1,3	1,2	97%
a_5	2,0	2,0	94%
λ_5	1,1	1,2	95%

3.3. Mô hình có 7 giai đoạn

Mô hình có 7 giai đoạn phát triển được mô phỏng theo phân phối Gamma với tham số tham số hình dạng và vị trí $(a_1,\lambda_1)=(2,2)$, $(a_2,\lambda_2)=(2,1.7)$, $(a_3,\lambda_3)=(2,1.4)$, $(a_4,\lambda_4)=(2,1.3)$, $(a_5,\lambda_5)=(2,1.1)$, $(a_6,\lambda_6)=(2,1.0)$ và $(a_7,\lambda_7)=(2,0.9)$ lần lượt ở 7 giai đoạn. Mô hình được mô phỏng với 15 thời điểm lấy mẫu được chia đều từ 0 đến 6. Có 10 mô phỏng được tiến hành với các cỡ mẫu 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 cá thể ở mỗi thời điểm lấy mẫu. Sau khi ước lượng so sánh hiệu suất tin cậy cỡ mẫu mô phỏng với 500 cá thể ở mỗi thời điểm lấy mẫu được chọn được có hiệu suất tin cậy ở cả 3 giai đoạn đều lớn hơn 90%.

Kết quả ước lượng tham số từ 50 dữ liệu mô phỏng cho mô hình 7 giai đoạn với 500 cá thể ở mỗi thời điểm lấy mẫu được trình bày ở Bảng 5. Giá trị ước lượng là giá trị trung bình của 50 dữ liệu mô phỏng. Các tham số ước lượng ở 7 giai đoạn có giá trị gần với giá trị thực. Hiệu suất tin cậy với cỡ mẫu 50 cá thể ở mỗi thời điểm lấy mẫu có giá trị lớn hơn 90%. So sánh tỷ lệ ước lượng và tỷ lệ mô phỏng ở 7 giai đoạn được trình bày ở Hình 3. Tại 15 thời điểm

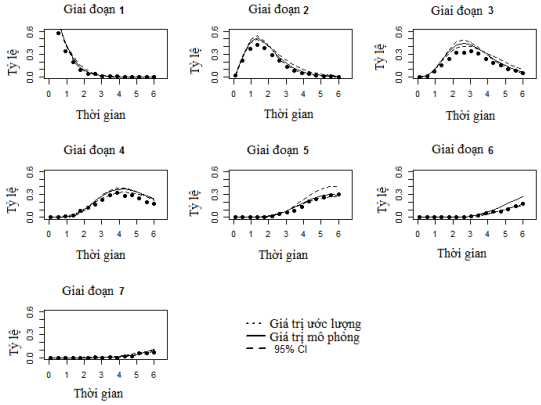
lấy mẫu, tỷ lệ ước lượng tương đối gần với tỷ lệ mô phỏng.



Hình 2. Tỷ lệ thực và tỷ lệ ước lượng của mô hình có 5 giai đoạn

Bảng 5. Kết quả ước lượng tham số của mô hình có 7 giai đoạn

Tham số	Giá trị thực	Giá trị ước lượng	Hiệu suất tin cậy
a_1	2,0	2,1	99%
λ_1	2,0	2,1	99%
a_2	2,0	1,9	98%
λ_2	1,7	1,7	99%
a_3	2,0	2,0	97%
λ_3	1,4	1,4	98%
a_4	2,0	2,0	95%
λ_4	1,3	1,3	97%
a_5	2,0	2,0	94%
λ_5	1,1	1,1	95%
a_6	2,0	2,0	93%
λ_6	1,0	1,0	96%
a_7	2,0	2,0	91%
λ_7	0,9	0,9	92%



Hình 3. Tỷ lệ thực và tỷ lệ ước lượng của mô hình có 7 giai đoạn

4. ÁP DỤNG TRÊN SỐ LIỆU THỜI GIAN Ủ BỆNH CỦA COVID-19

Thời gian ủ bệnh COVID-19 là khoảng thời gian giữa lần tiếp xúc đầu tiên với virus COVID-19 đến thời gian cơ thể có những triệu chứng khởi phát (Backer et al., 2020; Goel & Kumar, 2020; Mingyue et al., 2020; Rai et al., 2021; Wang et al., 2020; Yin et al., 2021). Trong thời gian ủ bệnh, người nhiễm có thể lây bệnh cho người khác, việc lây nhiễm xảy ra trước khi xuất hiện triệu chứng đầu tiên. Việc xác định thời gian ủ bệnh của COVID-19 được xem là chìa khóa để kiểm soát sự lây lan của dịch bệnh. Việc xác định thời gian ủ bệnh này giúp theo dõi những bệnh nhân có mức độ phơi nhiễm cao và giúp xác định thời gian cách ly.

Dữ liệu thời gian ủ bệnh của bài báo được trích từ dữ liệu trực tuyến thu thập từ những bệnh nhân ở Vũ Hán, Trung Quốc 01/2020 (Backer et al., 2020; Yin et al., 2021). Từ dữ liệu trên, dữ liệu của 98 bệnh nhân được trích xuất có thông tin về thời bắt đầu tiếp xúc với nguồn bệnh, thời gian kết thúc tiếp xúc với nguồn bệnh, thời gian xuất hiện triệu chứng. Thời gian giữa lần tiếp xúc đầu tiên với virus COVID-19 và thời gian cơ thể có những triệu chứng khởi phát thì không thể xác định chính xác được. Theo thông tin thu thập thì các thời gian này chỉ được xác định nằm trong khoảng thời gian mà người bệnh có thể xác định được.

Như vậy, theo dữ liệu thu thập được, thời gian ủ bệnh COVID-19 được mô hình theo 2 giai đoạn, giai đoạn tiếp xúc với mầm bệnh và giai đoạn ủ bệnh. Thời gian chuyển từ giai đoạn tiếp xúc với mầm bệnh đến giai đoạn ủ bệnh cũng như từ giai đoạn ủ bệnh đến giai đoạn tiếp theo của COVID-19 là không xác định được. 98 bệnh nhân nhiễm COVID-

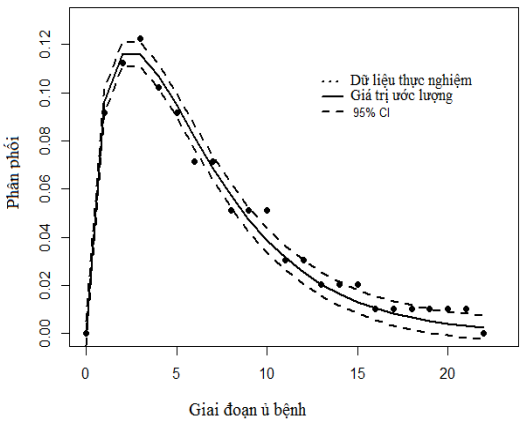
19 được quan sát trong 22 ngày. Mỗi ngày số bệnh nhân hết thời gian phơi nhiễm và xuất hiện triệu chứng được thu thập lại.

Thời gian tiếp xúc với mầm bệnh của các bệnh nhân được đặt ở thời điểm 0. Giai đoạn ủ bệnh COVID-19 được mô hình theo phân phối Gamma. Với cỡ mẫu 98 cho mỗi thời điểm lấy mẫu, áp dụng nghiên cứu mô phỏng ở phần 3 cho mô hình cấu trúc 2 giai đoạn, kết quả ước lượng tham số có độ tin cậy rất cao. Sau khi tiến hành nghiên cứu mô phỏng, cách tiếp cận Bayes được trình bày ở phần 2 được áp dụng ước tham số cho giai đoạn ủ bệnh của COVID-19. Kết quả cho thấy thuật toán hội tụ rất nhanh với độ chính xác cao. Tham số của giai đoạn ủ bệnh là giá trị trung bình của 50 lần ước lượng.

Giá trị ước lượng tham số của giai đoạn ủ bệnh COVID-19 được trình bày ở Bảng 6. Giai đoạn ủ bệnh COVID-19 có phân phối Gamma (1,7, 0,3) có giá trị trung bình là 6,1 ngày. Thời gian ủ bệnh COVID-19 kéo dài 14 ngày có xác suất là là 0,93. Như vậy, bệnh nhân COVID-19 trung bình biểu hiện các triệu chứng là 6,1 ngày sau khi nhiễm virus COVID-19. Dữ liệu thực nghiệm rất gần với phân phối ước lượng (Hình 4). Khi ước lượng tham số của giai đoạn ủ bệnh COVID-19, dữ liệu thời gian chuyển đổi giữa các giai đoạn là không xác định chính xác, nên cách tiếp cận Bayes được đề xuất kết quả có tính chính xác cao.

Bảng 6. Kết quả ước lượng tham số của giai đoạn ủ bệnh COVID-19

Tham số	Giá trị ước lượng	Phương sai
a	1,7	0,08
λ	0,3	0,02



Hình 4. Phân phối thời gian ủ bệnh COVID-19

5. KẾT LUẬN VÀ NHỮNG KHUYẾN NGHỊ

Các mô hình cấu trúc giai đoạn rất phổ biến trong các nghiên cứu về sinh học cũng như các nghiên cứu cho những dữ liệu có sự phát triển cấu trúc theo từng giai đoạn trong các phòng thí nghiệm. Ví dụ như quá trình phát triển của các loại bệnh, các loại côn trùng hay quá trình phát triển của các giống cây trồng. Trong các mô hình này, thời gian chuyển tiếp giữa các giai đoạn là không xác định được, dữ liệu thu thập được chỉ bao gồm số lượng cá thể ở các giai đoạn tại các thời điểm lấy mẫu khác nhau. Quy mô của nghiên cứu thực nghiệm bao gồm cỡ mẫu ở mỗi thời điểm lấy mẫu và số thời điểm lấy mẫu thích hợp cho mô hình cần được xác định trước khi tiến hành các nghiên cứu thực nghiệm để việc ước lượng tham số có kết quả chính xác hơn.

Với mô hình cấu trúc giai đoạn cho trước, các nhà nghiên cứu có thể sử dụng phương pháp nghiên cứu mô phỏng ở phần 3 để tìm ra cỡ mẫu tại các lần lấy mẫu và số thời điểm lấy mẫu thích hợp cho mô hình thực nghiệm trước khi tiến hành thí nghiệm. Với nghiên cứu mô phỏng, dữ liệu mô phỏng là chính xác từ phân phối được chọn của mô hình. Tuy nhiên, với dữ liệu thực tế của mô hình cấu trúc giai đoạn chứa những sai số của dữ liệu thực tế. Cho nên, khi tiến hành thí nghiệm với cùng các phân phối của nghiên cứu mô phỏng, thì cỡ mẫu của nghiên cứu thực tế nên có cỡ mẫu nhiều hơn so với nghiên cứu mô phỏng. Do đó kết quả phương pháp đánh giá cỡ mẫu được đề xuất rất hữu ích trong việc tìm ra cỡ mẫu tại các thời điểm lấy mẫu và số thời điểm lấy mẫu thích hợp cho mô hình thực nghiệm trước khi tiến hành thí nghiệm.

Các phương pháp đề xuất được triển khai ước lượng tham số trên dữ liệu thời gian ủ bệnh COVID-19 khi thời gian chuyển giữa các giai đoạn là không xác định được. Kết quả ước lượng với cách tiếp cận Bayes cho thấy với cỡ mẫu của dữ liệu thực nghiệm thuật toán hội tụ rất nhanh với độ chính xác cao. Kết quả của nghiên cứu tương ứng với các nghiên cứu về thời gian ủ bệnh trước đây và được đánh giá là kết quả nghiên cứu có các giá trị nối tiếp trong công tác phòng chống dịch bệnh. Kết quả áp dụng cách tiếp cận Bayes được đề xuất cho thấy thời gian cách ly 14 ngày có thể ngăn chặn phần lớn sự lây truyền của COVID-19. Ngoài ra, phương pháp đánh giá cỡ mẫu được đề xuất có thể được áp dụng để phân tích thêm các giai đoạn khác của virus COVID-19 cũng như các giai đoạn phát triển của các loại virus gây bệnh khác.

TÀI LIỆU THAM KHẢO

Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation period of 2019 novel coronavirus infections among travellers from Wuhan, China. *Eurosurveillance*, *25*(5), 20-28. https://doi.org/10.1101/2020.01.27.20018986

De Valpine, P., & Knape, J. (2015). Estimation of general multistage models from cohort data. *Journal of Agricultural, Biological, and Environmental Statistics*, *20*(1), 140-155. https://doi.org/10.1007/s13253-014-0189-7

Goel, K., & Kumar, A. (2020). Nonlinear dynamics of a time-delayed epidemic model with two explicit aware classes, saturated incidences, and treatment. *Nonlinear Dynamics*, 101, 1693-1715. https://doi.org/10.1007/s11071-020-05762-9

Hoeting, J., Tweedie, R., & Olver, C. (2003). Transform estimation of parameters for stagefrequency data. *Journal of American Statistical Association*, 98, 503-514. https://doi.org/10.1198/016214503000000288

Knape, J., Daane, K., & De Valpine, P. (2014). Estimation of stage duration distributions and mortality under repeated cohort censuses. *Biometrics*, *70*(2), 346-355. https://doi.org/10.1111/biom.12138

Knape, J., & De Valpine, P. (2016). Monte Carlo estimation of stage structured development from cohort data. *Ecology*, *97*(4), 992-1002. https://doi.org/10.1890/15-0942.1

Mingyue, Q. I. U., Tao, H. U., & Hengjian, C. U. I. (2020). Parametric estimation for the incubation period distribution of Covid-19 under doubly interval censoring. *Acta Mathematicae Applicatae Sinica*, *43*(2), 200-210.

Pham, H., & Branford, A. (2016). Exploring parameter relations for multi-stage models in stagewise constant and time dependent hazard rates *Australian & New Zealand Journal of Statistics*, *58*(3), 357-376. https://doi.org/10.1111/anzs.12164

Pham, H., Nur, D., Pham, H. T. T., & Branford, A. (2019). A Bayesian approach for parameter estimation in multi-stage models. *Communications in Statistics-Theory and Methods*, *48*(10), 2459-2482. https://doi.org/10.1080/03610926.2018.1465090

Pham, H., & Pham, H. T. T. (2019). A Bayesian approach for multi-stage models with linear time-dependent hazard rate. *Monte Carlo Methods and Applications*, *25*(4), 307-316. https://doi.org/10.1515/mcma-2019-2051

Rai, B., Shukla, A., & Dwivedi, L. K. (2021). Incubation period for Covid-19: a systematic review and meta-analysis. *Journal of Public Health*, 30, 1-8. https://doi.org/10.1007/s10389-021-01478-1

Read, K., & Ashford, J. (1968). A system of models for the life cycle of a biological organism. *Biometrika*, *55*(1), 211-221. https://doi.org/10.1093/biomet/55.1.211

Robert, C., & Casella, G. (2009). *Introducing Monte Carlo Methods with R*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-1576-4

Schuh, H., & Tweedie, R. (1979). Parameter estimation using transform estimation in time-evolving models. *Mathematical Biosciences*, *45*(1), 37-67. https://doi.org/10.1016/0025-5564(79)90095-6

Wang, Y., Wei, Z., & Cao, J. (2020). Epidemic dynamics of influenza-like diseases spreading in complex networks. *Nonlinear Dynamics*, 101, 1801-1820. https://doi.org/10.1007/s11071-020-05867-1

Yin, M. Z., Zhu, Q. W., & La, X. (2021). Parameter estimation of the incubation period of covid-19 based on the doubly interval-censored data model, *Nonlinear Dynamics*, *106*(2), 1347-1358. https://doi.org/10.1007/s11071-021-06587-w