Mô phỏng Monte Carlo bằng phần mềm R trong giảng dạy Xác suất Thống kê ở bậc đại học

Lê Thị Kim Anh

Email: anhltk@buh.edu.vn
Trường Đại học Ngân hàng
56 Hoàng Diệu 2, phường Linh Chiểu,
Thành phố Thủ Đức, Thành phố Hồ Chí Minh,
Việt Nam

TÓM TẮT: *Bài viết đề xuất sử dụng phần mềm R để thực hiện mô phỏng theo phương pháp Monte Carlo các khái niệm, định lí quan trọng trong môn học Xác suất Thống kê ở bậc đại học. Qua kinh nghiệm giảng dạy và hiểu biết của tác giả, các giáo trình Xác suất Thống kê được sử dụng trong đa số các trường đại học ở Việt Nam chưa chú trọng các phương pháp mô phỏng khi trình bày các khái niệm của môn học. Điều này dẫn đến việc học và hiểu của sinh viên còn nhiều hạn chế, đặc biệt là các khái niệm khó như khái niệm khoảng tin cậy, định lí giới hạn trung tâm hay công thức xác suất Bayes. Dùng phương pháp mô phỏng Monte Carlo trong giảng dạy Xác suất Thống kê có thể giúp sinh viên hiểu kiến thức của môn học vừa trực quan vừa đúng bản chất.*

TỪ KHÓA: **Phương pháp Monte Carlo, Xác suất Thống kê.**

➔ Nhận bài 17/3/2022 ➔ Nhận bài đã chỉnh sửa 11/4/2022 ➔ Duyệt đăng 15/9/2022.

DOI: https://doi.org/10.15625/2615-8957/12210904

1. Đặt vấn đề

Tại Việt Nam, đa số các trường đại học nói chung giảng dạy Xác suất Thống kê cho sinh viên khối ngành Kinh tế kĩ thuật theo kiểu thiên về thực hành giải toán với điểm chung là dựa vào các giáo trình xuất bản trong nước hoặc tài liệu lưu hành nội bộ. Với sự hiểu biết của chúng tôi và qua khảo sát một số đầu sách Xác suất Thống kê có mặt trên thị trường thì ở Việt Nam phương pháp Monte Carlo chưa được đề cập cũng như gợi ý sử dụng nhằm hỗ trợ cho việc dạy học các khái niệm khó tiếp cận và hay hiểu sai trong thống kê. Điều này khiến cho sinh viên không học chuyên ngành Toán ở bậc đại học hiểu không đúng bản chất các khái niệm, định lí được phát biểu trong chương trình học.

Ở các nước phát triển, phương pháp mô phỏng Monte Carlo cũng được nghiên cứu áp dụng vào giảng dạy Xác suất Thống kê cũng như các sách viết về Xác suất Thống kê [1], [2]. Một số nghiên cứu còn đi xa hơn bằng việc viết các Shiny App (trong R) hoặc giao dự án cho sinh viên viết các Shiny App mô phỏng cho các nội dung học trong chương trình môn học [3], [4]. Trong bài viết này, tác giả lựa chọn khoảng tin cậy của ước lượng, định lí giới hạn trung tâm để thực hiện mô phỏng Monte Carlo nhằm cung cấp cái nhìn cụ thể hơn cũng như làm tài liệu tham khảo cho các giảng viên muốn áp dụng.

2. Nội dung nghiên cứu

2.1. Mô phỏng Monte Carlo và ngôn ngữ R

Phương pháp Monte Carlo là phương pháp mô phỏng nhờ vào máy tính với các dữ liệu tạo ra bằng các hàm tạo số ngẫu nhiên có sẵn. Sử dụng phương pháp Monte Carlo ta có thể mô phỏng một số khái niệm của Xác suất Thống kê do ta có thể thực hiện được đủ lâu và đủ nhiều trên máy tính mà không cần phải làm rất nhiều thử nghiệm thật sự trong thế giới thực. Ví dụ sau đây mô tả cách xấp xỉ số p theo phương pháp mô phỏng Monte Carlo:

Dùng hàm tạo số ngẫu nhiên trong một ngôn ngữ lập trình cụ thể (ở đây chúng tôi dùng R và dùng hàm runif(n,a,b) để xuất ngẫu nhiên n giá trị có phân phối đều trên khoảng (a, b)) để tạo ra n = 100 điểm ngẫu nhiên nằm trong hình vuông tâm tại (0, 0) và độ dài cạnh là 2 đơn vị trên hệ trục toạ độ Oxy.

Đếm số điểm nằm bên trong hình tròn tâm (0,0), bán kính 1. Giả sử có r điểm như vậy.

Về mặt xác suất, nếu các điểm có phân bố đều trong hình vuông thì

$$\frac{\text{diện tích hình tròn}}{\text{diện tích hình vuông}} = \frac{\pi}{4} \approx \frac{r}{n}\frac{\text{diện tích hình tròn}}{\text{diện tích hình vuông}} = \frac{\pi}{4} \approx \frac{r}{n}.$$

Khi n càng lớn, tỉ số r/n càng tiến về số p/4. Điều này cho phép ta xấp xỉ p bởi 4r/n khi n đủ lớn (xem Bảng 1).

Các lệnh trong R có thể như sau:

```
set.seed(123)
n<-1000
x<-runif(n,-1,1)
y<-runif(n,-1,1)
r<-length(x[x^2+y^2<=1])
pi_sim<-4*r/n
```

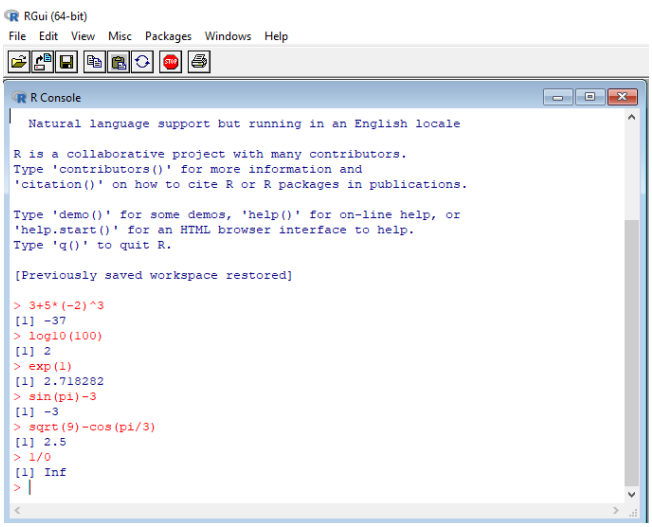
Để thực hiện mô phỏng, nhiều ngôn ngữ lập trình có thể được sử dụng như R, Matlab, Python, C, … Như ví dụ trên, chúng tôi dùng ngôn ngữ R (còn gọi là phần mềm R). Đây là ngôn ngữ được thiết kế và sử dụng trong cộng đồng các nhà thống kê và không ngừng phát triển [6].

Bảng 1: Kết quả xấp xỉ số pi qua mô phỏng Monte Carlo

n	100	1000	10000	100000
pi_sim	3.4	3.2	3.1576	3.14632

Với R ta có thể tính toán số học đơn giản (+, -, *, /, căn bậc hai) cũng như các hàm số phức tạp khác như logarit, lượng giác, mũ,... Ngoài ra, R còn là một phần mềm tích hợp để thao tác dữ liệu, tính toán và trình bày đồ họa. Một số ưu điểm của R có thể kể đến [5]:

- Lưu trữ và xử lí dữ liệu hiệu quả.
- Tính toán hiệu quả trên các mảng, đặc biệt là các ma trận.
- Có một bộ sưu tập lớn, chặt chẽ, tích hợp các công cụ trung gian để phân tích dữ liệu.
- Mã nguồn mở với nhiều gói lệnh chuyên dụng được tạo ra bởi cộng đồng sử dụng lớn.
- Miễn phí.

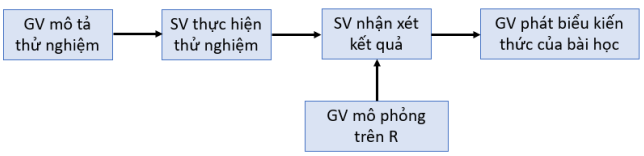


Hình 1: Một phần giao diện R và tính toán đơn giản trong R

2.2. Mô phỏng Monte Carlo một số khái niệm, định lí trong môn học Xác suất Thống kê

Trong bài viết này, chúng tôi thiết kế hướng tiếp cận giảng dạy các khái niệm quan trọng trong thống kê học sử dụng mô phỏng Monte Carlo. Việc làm này không thể thiếu các công cụ hỗ trợ và R là một trong số các ngôn ngữ lập trình được chúng tôi sử dụng vì tính đơn giản và miễn phí của nó. Việc cài đặt R cũng như Rstudio không thuộc phạm vi của bài viết này.

Điểm chung trong tất cả các thiết kế dạy học có thể được nhìn thấy như sơ đồ bên dưới. Trong đó, trước hết giảng viên yêu cầu sinh viên thực hiện mô phỏng thủ công một thử nghiệm đơn giản dễ thực hiện nhằm để người học có cái nhìn ban đầu về thử nghiệm sẽ được mô phỏng trên máy tính sau đó. Người học sau đó đưa ra nhận xét ban đầu về các kết quả thu được. Tiếp đó, giảng viên dẫn dắt để đi đến mô phỏng lặp lại thử nghiệm với số lần tương đối nhiều trên R, ví dụ 1000 lần hoặc hơn. Kết quả của thử nghiệm sau đó được người học nhận xét trước khi được giảng viên mở rộng và tổng quát và phát biểu thành các khái niệm định lí hay kết quả liên quan (xem Hình 2).



Hình 2: Sơ đồ thiết kế hoạt động dạy học sử dụng mô phỏng Monte Carlo.

Chúng tôi sẽ chỉ mô phỏng trên R một số khái niệm như trình bày bên dưới, các bước còn lại nằm trong hoạt động dạy học của giảng viên như mô tả thử nghiệm, phát biểu kiến thức, nhận xét của sinh viên, … có thể được thiết kế phù hợp với phương pháp giảng dạy cũng như mục tiêu dạy học cụ thể chúng tôi không đề cập ở đây.

2.2.1. Định lí giới hạn trung tâm

Trong thống kê, định lí giới hạn trung tâm được phát biểu như sau:

Định lí [8]. Nếu dãy $X_1, X_2, \ldots, X_n$ là mẫu ngẫu nhiên kích thước n được lấy ra từ quần thể có trung bình m và phương sai hữu hạn s^2, thì:

$$\left(\frac{\overline{X}-\mu}{\sigma}\right)\sqrt{n} \xrightarrow{F} N(0,1),$$

trong đó $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ là trung bình mẫu.

Nghĩa là: $\lim_{n\to\infty} F_{S_n}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$

Định lí giới hạn trung tâm là một định lí quan trọng làm nền tảng cho nhiều lập luận và phương pháp của thống suy diễn nhưng khó hiểu với sinh viên khi được phát biểu dưới góc độ toán học thuần tuý. Định lí phát biểu rằng, nếu mẫu ngẫu nhiên kích thước n được lấy ra từ quần thể có trung bình m và độ lệch chuẩn s thì phân phối của trung bình của mẫu ngẫu nhiên là xấp xỉ chuẩn N(m, s^2/n) khi kích thước mẫu n lớn. Khi đó, nếu $Z = \frac{\overline{X}-\mu}{\sigma}$ thì Z có phân phối xấp xỉ chuẩn tắc với n tương đối lớn. Để tiếp cận nội dung của định lí này về bản chất, chúng tôi thực hiện mô phỏng sau trên phần mềm R:

Gieo con xúc sắc sáu mặt 5 lần (kích thước mẫu, n = 5) lần và ghi nhận trung bình cộng (trung bình mẫu, Xbar). Do con xúc sắc 6 mặt là bình thường có

phân phối đều rời rạc nên số chấm xuất hiện có trung bình $\mu = 3.5$ với phương sai $\sigma^2 = 35/12$.

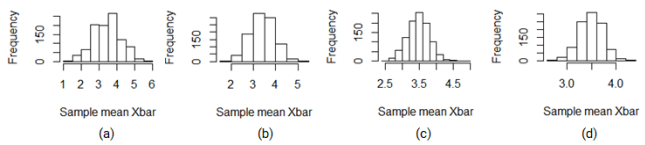
Vẽ biểu đồ histogram của 1000 lần lấy mẫu ngẫu nhiên để quan sát phân phối của trung bình mẫu (xem Hình 3).

Tăng kích thước mẫu lên n = 10, n = 50,… để thấy phân phối của trung bình mẫu là xấp xỉ chuẩn.

Tính trung bình mean(Xbar) của 1000 mẫu cũng như độ lệch chuẩn sd(Xbar).

Các dòng lệnh cụ thể trên R như sau:

```
set.seed(123)
n<-5
sim<-replicate(1000,mean(sample(1:6,n,replace=T)))
hist(sim,xlab="Sample mean Xbar",main="Histogram",breaks=10)
mean(sim)
sd(sim)
```



Hình 3: Biểu đồ Histogram cho 1000 lần lấy mẫu ngẫu nhiên kích thước n. (a) n = 5, (b) n = 10, (c) n = 30, (d) n = 50.

So sánh kết quả mô phỏng với phát biểu của định lí giới hạn trung tâm: khi n càng lớn mean(Xbar) tiến về μ và sd(Xbar) tiến về $\frac{\sigma}{\sqrt{n}}$. (xem Bảng 2).

2.2.2. Khoảng tin cậy và độ tin cậy trong bài toán ước lượng

Trong thống kê suy diễn, xây dựng một ước lượng khoảng cho tham số của quần thể là một bài toán cơ bản nhưng không phải sinh viên nào cũng hiểu đúng thế nào là "khoảng tin cậy 95% cho trung bình μ của quần thể".

Sử dụng R để mô phỏng, giảng viên có thể giúp sinh viên hiểu đúng bản chất của khái niệm này. Trong phần này, chúng tôi chọn tham số trung bình quần thể μ để xây dựng khoảng tin cậy với độ tin cậy cho trước (ví dụ 95%).

Các bước mô phỏng trong R có thể thực hiện theo ý tưởng sau:

Lấy 1000 mẫu kích thước n = 25 từ quần thể có phân phối chuẩn chuẩn tắc với trung bình $\mu = 0$ và độ lệch chuẩn $\sigma = 1$.

Tính trung bình mẫu của tất cả các mẫu.

Tính khoảng ước lượng cho m với độ tin cậy 95%.

Tính tỉ lệ các mẫu mà khoảng tin cậy thật sự chứa trung bình quần thể $\mu = 0$.

Các lệnh trong R:

```
set.seed(123)
conf<-function(n,alpha){
m<-mean(rnorm(n))
se<-1/sqrt(n)
za2<-qnorm(1-alpha/2)
ci<-c(m-se*za2,m+se*za2)
if (ci[1]>0 || ci[2]<0){return(0)}
else{return(1)}
}
sum(replicate(1000,conf(25,0.05)))
```

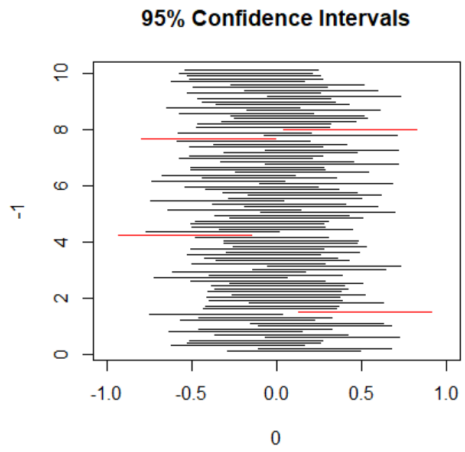
Bảng 3 cho kết quả các trường hợp chúng tôi mô phỏng sử dụng số lượng mẫu khác nhau (1000, 10000, 50000), mỗi mẫu kích thước 25 với độ tin cậy khác nhau (95%, 97%).

Bảng 3: Một số kết quả mô phỏng

Số lần lấy mẫu	Số mẫu có khoảng ước lượng chứa trung bình m	Tỉ lệ	Độ tin cậy
1000	955	0.955	95%
10000	9531	0.9531	95%
50000	47493	0.94986	95%
1000	978	0.978	97%
10000	9704	0.9704	97%
50000	48460	0.9692	97%

Bảng 2: So sánh kết quả mô phỏng với định lí giới hạn trung tâm

Kích thước mẫu (n)	mean(Xbar)	Trung bình quần thể (m)	sd(Xbar)	$\frac{\sigma}{\sqrt{n}}$
5	3.4870	3.5	0.7798	0.7638
10	3.3697		0.5541	0.5401
30	3.4920		0.3141	0.3118
50	3.5008		0.2383	0.2415



Hình 4: Mô phỏng 100 khoảng tin cậy 95% cho 100 mẫu kích thước 25. Có 4 mẫu trong đó khoảng tin cậy (màu đỏ) không thật sự chứa trung bình quần thể (m = 0).

Nhận xét từ kết quả mô phỏng, giảng viên nhấn mạnh ý nghĩa của cái gọi là khoảng ước lượng độ tin cậy 95%: Nếu chúng ta thực hiện lấy mẫu và tính toán khoảng tin cậy 95% thì khoảng tin cậy tính ra có thể chứa hoặc không chứa trung bình m. Nhưng về lâu dài, tức số mẫu nhiều đến vô hạn, có 95% số khoảng tin cậy thật sự chứa trung bình quần thể m. Điều này giúp sinh viên không hiểu sai về khoảng tin cậy. Một trong những cách hiểu sai là cho rằng, xác suất để m nằm trong một khoảng tin cậy cụ thể nào đó là 0.95 (xem Hình 4).

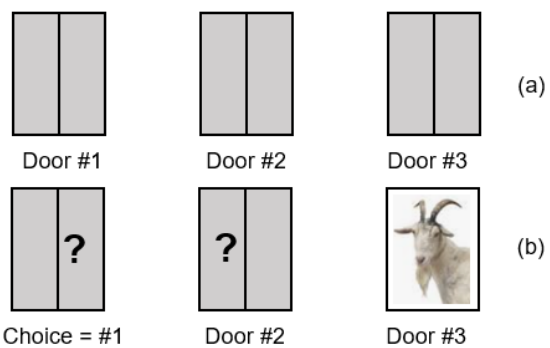
Bên cạnh đó, giảng viên còn có thể minh hoạ trực quan hơn qua mô phỏng trên R với các lệnh sau:

```
set.seed(123)
n<-25
e<-qnorm(0.975)/sqrt(n)
x1<-x2<-c()
for (i in 1:100){
xbar<-mean(rnorm(n))
x1[i]<-xbar-e
x2[i]<-xbar+e
}
plot(0,-1,ylim=c(0.2,10),main="95% Confidence Intervals")
y1<-y2<-seq(0.1,10.1,length.out=100)
for (i in 1:100){
if (x1[i]>0 || x2[i]<0){
segments(x1[i],y1[i],x2[i],y2[i],col='red')
}
else{
segments(x1[i],y1[i],x2[i],y2[i])
}
}
```

Xác suất có điều kiện và bài toán Monty Hall

Bài toán (hay nghịch lí) Monty Hall lần đầu tiên xuất hiện trên tạp chí trên tạp chí Scientific American năm 1959 trong phần "Trò chơi toán học" và sau đó trở nên nổi tiếng trong chương trình truyền hình của Monty Hall có tên Let's Make a Deal. Bài toán còn được gọi là "bài toán ba cánh cửa" trong đó người chơi phải đối mặt với ba cánh cửa giống hệt nhau. Một cửa giấu một giải thưởng có giá trị, thường là một chiếc ô tô. Hai cửa còn lại cất giấu các giải thưởng vô giá trị, chẳng hạn như con dê. Sau khi khách lựa chọn ban đầu cho một cửa, người dẫn chương trình, người biết rõ vị trí của giải thưởng sẽ mở một cửa không được người chơi chọn và cũng là cửa không có giải thưởng. Tiếp theo, người chơi được hỏi liệu anh ta muốn giữ lại lựa chọn ban đầu hay muốn chuyển sang cửa còn lại chưa mở (xem Hình 5). Theo suy nghĩ thông thường, việc đổi sang cửa mới hay giữ lại cửa ban đầu có xác suất 50%-50% vì sau cùng chỉ còn lại hai cánh cửa và chỉ một có phần thưởng. Tuy nhiên, bằng cách áp dụng định lí Bayes cũng như công thức xác suất toàn phần, xác suất có phần thưởng khi người chơi đổi sang cửa mới là 2/3 thay vì 1/2. Vì kết luận này phản ánh trực giác ban đầu nên bài toán còn được gọi là một nghịch lí. Nghịch lí có thể được sử dụng trong giảng dạy, đặc biệt là dùng để kích thích và tạo động lực cho sinh viên tìm hiểu các công thức tính có điều kiện trong đó xác suất phụ thuộc vào thông tin mà người tính xác suất có được [9]. Giải thích nghịch lí này như là một cơ hội để giáo viên giới thiệu các công thức này cách tự nhiên và hiệu quả. Để cho ví dụ được sinh động hơn, giáo viên có thể sử dụng R để mô phỏng nhiều lượt chơi khác nhau và ghi nhận kết quả với các dòng lệnh sau:

```
set.seed(123)
monty<-function(){
rand_door<-sample(1:3,1)
choice<-sample(1:3,1)
ifelse(choice==rand_door,'goat','car')
}
n<-1000
trials<-replicate(n,monty())
table(trials)/n
```



Hình 5: (a) 3 cánh cửa trong trò chơi Monty Hall; (b) Người chơi chọn cửa #1, cửa #3 được mở và người chơi có quyền chọn lại cửa #2 hoặc vẫn giữ cửa #1.

Bảng sau cho thấy một số kết quả mô phỏng với các giá trị khác nhau của n.

Bảng 4: Một số kết quả mô phỏng

N	1000		10000		100000	
Tỉ lệ	Có quà	Không	Có quà	Không	Có quà	Không
	0.635	0.365	0.668	0.332	0.66796	0.33204

Khi n càng lớn, tỉ lệ có quà càng dần về 2/3, cũng là giá trị xác suất tính theo công thức Bayes.

3. Kết luận

Phương pháp mô phỏng Monte Carlo trong dạy học Xác suất Thống kê nói chung chưa được chú trọng ở bậc đại học Việt Nam. Qua bài viết, tôi đưa ra các gợi ý sử dụng mô phỏng Monte Carlo thiết kế dạy học một số vấn đề của môn học như công thức xác suất Bayes qua bài toán Monty Hall, khái niệm của học phần Xác suất Thống kê như khoảng tin cậy, định lí giới hạn trung tâm giúp sinh viên hiểu các công thức, khái niệm, định lí một cách trực tiếp và đúng bản chất. Trong các nghiên cứu tiếp theo, chúng tôi sẽ so sánh hiệu quả dạy học của hai cách tiếp cận dạy học: truyền thống (không sử dụng mô phỏng) và có sử dụng mô phỏng Monte Carlo.

Tài liệu tham khảo

[1] Matthew J. Sigal - R. Philip Chalmers, (2016), *Play It Again: Teaching Statistics With Monte Carlo Simulation*, Journal of Statistics Education, 24:3, p.136-156.

[2] Probability and Statistics for Computer Scientists 3rd Edition, Michael Baron, (2019), *Chapman and Hall/CRC*.

[3] Sabrina Luxin Wang - Anna Yinqi Zhang - Samuel Messer - Andrew Wiesner - Dennis K. Pearl, (2021), *Student-Developed Shiny Applications for Teaching Statistics*, Journal of Statistics and Data Science Education, 29:3, p.218-227.

[4] Doi, Jimmy Potter, Gail Wong, Jimmy et al, (2016), *Web Application Teaching Tools for Statistics Using R and Shiny*, Technology Innovations in Statistics Education, 9(1).

[5] W. N. Venables, D. M. Smith and the R Core Team, *An Introduction to R*, Notes on R: A Programming Environment for Data Analysis and Graphics Version 4.1.3 (2022-03-10), https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

[6] Michael J. Crawley, (2014), *Statistics: An Introduction Using R*, 2nd Edition, Wiley.

[7] Reuven Y. Rubinstein, Dirk P. Kroese, (2016), *Simulation and the Monte Carlo Method*, Wiley.

[8] Lê Sĩ Đồng, Giáo trình Xác suất – Thống kê, (2013), NXB Giáo dục Việt Nam.

[9] Bennett, Kevin L., (2018), *Teaching the Monty Hall Dilemma to Explore Decision-Making, Probability, and Regret in Behavioral Science Classrooms*, International Journal for the Scholarship of Teaching and Learning: Vol. 12: No. 2, Article 13.

MONTE CARLO SIMULATION WITH R PROGRAMMING LANGUAGE IN TEACHING PROBABILITY AND STATISTICS AT UNIVERSITY LEVEL

Le Thi Kim Anh

Email: anhltk@buh.edu.vn
Ho Chi Minh University of Banking
56 Hoang Dieu 2 street, Linh Chieu ward,
Thu Duc city, Ho Chi Minh City, Vietnam

ABSTRACT: *The article aims to use R software to perform Monte Carlo simulations of important concepts and theorems in the subject of Statistical Probability. Based on the author's teaching experience and knowledge, the Statistical Probability textbooks used in most schools in Vietnam have not focused on simulation methods when presenting the concepts of this subject. This leads to many limitations in students' learning and understanding, especially difficult concepts such as the concept of confidence intervals, the central limit theorem, and Bayes's theorem. Using the Monte Carlo simulation method in teaching Probability and Statistics can help students understand the subject's knowledge both intuitively and intrinsically.*

KEYWORDS: Monte Carlo methods, probability and statistics.