SỬ DỤNG BOOTSTRAP TRONG VIỆC XÁC ĐỊNH MẬT ĐỘ XƯƠNG CỦA PHỤ NỮ VIỆT NAM

Nguyễn Văn Thu[(1)], Nguyễn Đức Phương[(2)]
(1)Trường Đại học Quốc tế, ĐHQG-HCM
(2) Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
(Bài nhận ngày 12 tháng 03 năm 2008, hoàn chỉnh sửa chữa ngày 24 tháng 04 năm 2008)

TÓM TẮT: *Trong bài báo này, chúng tôi sử dụng phương pháp bootstrap để nghiên cứu độ lệch tiêu chuẩn của mật độ xương tối đa của phụ nữ Việt Nam. Kết quả này có tầm quan trọng trong việc nhận biết mức độ nguy hiểm của căn bệnh loãng xương.*

1. GIỚI THIỆU

Trong thống kê, theo phương pháp mà chúng ta vẫn thường dùng để ước lượng hay kiểm định tham số thống kê là đưa ra các giả định về phân phối của X hoặc giả định về cở mẫu. Dựa vào các giả định này để tìm phân phối của các thống kê mà ta đang xét. Chẳng hạn để ước lượng khoảng cho phương sai trường hợp không biết giá trị của kỳ vọng μ thì người ta xét thống kê

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

khi $X \sim N\left(\mu,\sigma^2\right)$. Nhưng không phải lúc nào giả định của thống kê mà chúng ta đang xét luôn thỏa đáng. Trong trường hợp vi phạm các giả định thống kê thì kết quả của việc phân tích sẽ không có ý nghĩa.

Phương pháp bootstrap đã được xây dựng để giải các vấn đề như thế này. Phương pháp phân tích bootstrap là tập hợp một số kĩ thuật phân tích dựa vào nguyên lí tái chọn mẫu (resampling) để ước tính các thông số mà các phương pháp thống kê truyền thống không có giải đáp. Phương pháp bootstrap do Giáo sư Bradley Efron thuộc Đại học Stanford phát triển từ cuối thập niên 1970s, nhưng mãi đến khi máy tính trở nên thông dụng thì mới thành một phương pháp phổ biến trong phân tích thống kê. Sự ra đời của phương pháp phân tích bootstrap được đánh giá một cuộc cách mạng quan trọng trong thống kê học, vì nó giải quyết nhiều vấn đề mà trước đây tưởng như không thể nào giải được.

2. PHÂN PHỐI BOOTSTRAP

Định nghĩa 1 (Mẫu bootstrap). *Mẫu bootstrap* $x^{\#} = \left(x_1^{\#},\ldots,x_n^{\#}\right)$ *là mẫu ngẫu nhiên cở* n *trong đó mỗi* $x_i^{\#}$ *nhận được với xác suất* $1/n$ *bằng cách lấy mẫu có hoàn lại từ mẫu gốc* $x = \left(x_1,\ldots,x_n\right)$.

Với mẫu ngẫu nhiên $\left(X_1,\ldots,X_n\right)$, hàm phân phối của thống kê $\theta_n = \theta\left(X_1,\ldots,X_n\right)$ được định bởi $G(t) = \mathbb{P}\left(\theta_n < t\right)$.

Định nghĩa 2: *(Phân phối bootstrap). Đặt* $\theta_n^{\#} = \theta^{\#}\left(X_1^{\#},\ldots,X_n^{\#}\right)$ *là thống kê trên mẫu bootstrap.* $G^{\#}\left(t\right) = \mathbb{P}\left(\theta_n^{\#} < t\right)$ *là phân phối của* $\theta_n^{\#}$.

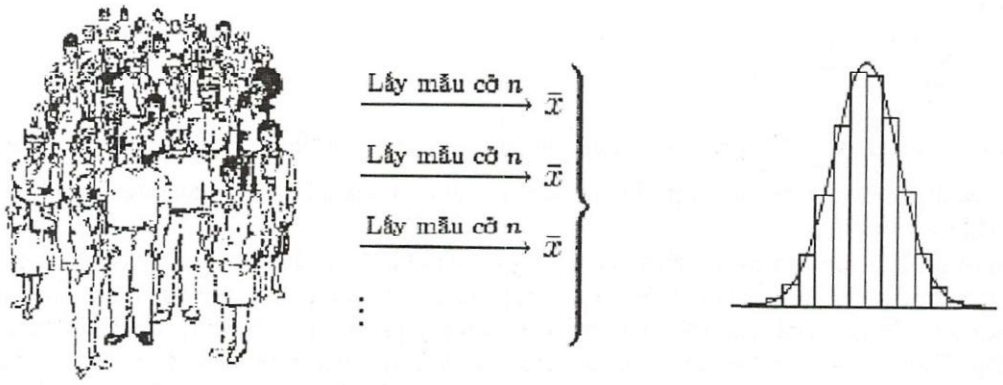
2. SAI SỐ TIÊU CHUẨN

Nguyên lý và mục đích đằng sau của thống kê học là ước tính những thông số của tổng thể. Trong thực tế chúng ta không biết các thông số này, mà chỉ dựa vào những ước tính từ một hay nhiều mẫu để suy luận cho giá trị của tổng thể mà các mẫu được chọn. Nhưng chọn mẫu phải ngẫu nhiên thì mới mang tính đại diện cao. Cứ mỗi lần chọn mẫu, chúng ta có một nhóm đối tượng khác với mẫu thứ i, chúng ta có một giá trị t_n^i mới của thống kê $\theta_n = \theta(X_1,\ldots,X_n)$. Câu hỏi đặt ra là chọn nhiều lần thì các số t_n^i dao động cỡ nào.

Nếu chúng ta chọn mẫu N lần (mỗi lần n đối tượng), thì ta sẽ có N số t_n^i, $(i=1,\ldots,N)$. Độ lệch tiêu chuẩn của N số t_n^i gọi là sai số tiêu chuẩn, ký hiệu

$$se(\theta_n) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(t_n^i - \bar{t}_n\right)^2}$$

Trong đó $\bar{t}_n = \frac{1}{N}\sum_{i=1}^{N} t_n^i$. Do đó, sai số tiêu chuẩn phản ánh độ dao động hay biến thiên của các số t_n^i.



Tổng thể

Phân phối mẫu của $\bar{X}$

Hình 1. Ý tưởng xây dựng phân phối mẫu cho $\bar{X}$.

Ví dụ: Hình 1 minh họa ý tưởng xây dựng phân phối mẫu cho $\bar{X}$. Độ lệch tiêu chuẩn của các giá trị trung bình chính là sai số tiêu chuẩn.

Trong thực hành, vệc chọn mẫu N lần để xác định độ lệch tiêu chuẩn của θ_n không khả thi. Thay vào đó ta chỉ có một mẫu (gọi là mẫu gốc), ta sử dụng phương pháp bootstrap để ước tính độ lệch tiêu chuẩn của θ_n. Ta xem mẫu gốc là tổng thể mới, thực hiện tái lấy mẫu từ mẫu gốc này và tính giá trị các thống kê. Các bước cụ thể như sau:

Bước 1: Tái lấy mẫu từ mẫu gốc ta được các mẫu bootstrap $x^{\#i} = \left(x_1^{\#i},\ldots,x_n^{\#i}\right)$, $(i=1,\ldots,B)$.

Bước 2: Với mỗi mẫu bootstrap có được ở bước 1 ta đi tính giá trị của thống kê $\theta_n^{\#}$.

Bước 3: Sau khi thực hiên hai bước trên ta có được B giá trị của thống kê đang khảo sát $t_n^{\#1},\ldots,t_n^{\#B}$ ta tính độ lệch tiêu chuẩn của B giá trị $t_n^{\#1},\ldots,t_n^{\#B}$. Độ lệch tiêu chuẩn này là ước lượng bootstrap của sai số tiêu chuẩn,

$$se^{\#}(\theta_n)=\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(t_n^{\#i}-\overline{t}_n^{\#}\right)^2}$$

Trong đó $\overline{t}_n^{\#}=\frac{1}{B}\sum_{i=1}^{B}t_n^{\#i}$.

3. KHOẢNG TIN CẬY BOOTSTRAP-T

Gọi θ là tham số không biết của phân phối và $\hat{\theta}$ là ước lượng điểm cho θ, chúng ta xây dựng khoảng ước lượng cho tham số θ với mức độ tin cậy cho trước. Cho α là một số thực lớn hơn 0 và nhỏ hơn 1, thường α nhận giá trị nhỏ như là 0.01, 0.05 hay 0.10. Với độ tin cậy $(1-\alpha)\cdot 100\%$ thì khoảng tin cậy của θ là $\left(\hat{\theta}-z(1-\alpha/2)\cdot\widehat{se};\hat{\theta}-z(\alpha/2)\cdot\widehat{se}\right)$. Trong đó $\widehat{se}$ có thể là bootstrap ước lượng hay là các ước lượng khác cho sai số tiêu chuẩn. $z(1-\alpha/2)$ và $z(\alpha/2)$ là phân vị mức $1-\alpha/2$ và $\alpha/2$ của phân phối của biến ngẫu nhiên $Z=\left(\hat{\theta}-\theta\right)/\widehat{se}$. Chú ý là phân phối của biến ngẫu nhiên Z không yêu cầu phải là phân phối chuẩn.

Ví dụ: Giả sử khi Z có phân phối chuẩn tắc $N(0,1)$ thì giá trị $z(1-\alpha/2)$ và $z(\alpha/2)$ là phân vị chuẩn tắc. Cụ thể, $z(0.975)=1.96$ và $z(0.025)=-1.96$. Do đó khoảng tin cậy 95% của θ là

$$\left(\hat{\theta}-1.96\cdot\widehat{se};\hat{\theta}+1.96\cdot\widehat{se}\right)$$

Khi Z không có phân phối chuẩn hoặc student thì $z(1-\alpha/2)$ và $z(\alpha/2)$ không biết. Tuy nhiên, chúng ta có thể dùng phương pháp bootstrap để xây dựng bảng giá trị mới cho $z(1-\alpha/2)$ và $z(\alpha/2)$. Các bước như sau:

Bước 1: Tạo B mẫu bootstrap $x^{\#1},\ldots,x^{\#B}$.

Bước 2: Với mỗi mẫu bootstrap có được ở bước 1 ta đi tính giá trị của thống kê $Z^{\#i}=\dfrac{\hat{\theta}^{\#i}-\hat{\theta}}{se^{\#i}}$.

Bước 3: Sau khi thực hiện bước 2 ta có B giá trị $Z^{\#i}$. Ta tìm giá trị của $z(1-\alpha/2)$ thỏa

$$\frac{\#\left\{Z^{\#i}<z(1-\alpha/2)\right\}}{B}=1-\frac{\alpha}{2}$$

và giá trị $z(\alpha/2)$ thỏa

$$\frac{\#\left\{Z^{\#i}<z(\alpha/2)\right\}}{B}=\frac{\alpha}{2}.$$

4. KHOẢNG TIN CẬY PHẦN TRĂM (THE PERCENTILE INTERVAL)

Với các giá trị $t_n^{\#i}$ tính được từ mẫu bootstrap, ta xếp chúng theo thứ tự tăng dần. Cận dưới của ước lượng là giá trị $t_n^{\#u}$ ở vị trí $B\cdot\alpha$ và cận trên của ước lượng là giá trị $t_n^{\#b}$ ở vị trí $B\cdot(1-\alpha)$. Các bước thực hiện:

Bước 1: Tạo B mẫu bootstrap $x^{\#1},\ldots,x^{\#B}$.

Bước 2: Với mỗi mẫu bootstrap có được ở bước 1 ta đi tính giá trị của thống kê $\theta^{\#i}=\theta\left(x_n^{\#1},\ldots,x_n^{\#1}\right)$.

Bước 3: Sau khi thực hiện bước 2 ta có B giá trị $\theta^{\#i}$. Giá trị cận dưới của khoảng ước lượng là $\hat{\theta}_l$ thỏa $\dfrac{\#\left\{\theta^{\#i}<\hat{\theta}_l\right\}}{B}=\dfrac{\alpha}{2}$ và cận trên của ước lượng $\hat{\theta}_u$ thỏa $\dfrac{\#\left\{\theta^{\#i}<\hat{\theta}_u\right\}}{B}=1-\dfrac{\alpha}{2}$.

5. HỒI QUI BOOTSTRAP

Mô hình tuyến tính tổng quát $Y=X\beta+\varepsilon$, trong đó $Y=\left(y_1,\ldots,y_p\right)^T$, $\varepsilon=\left(\varepsilon_1,\ldots,\varepsilon_p\right)^T$ và

$$X=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix}$$

Các giả định trong phân tích hồi qui:

Giả định 1: Kỳ vọng của ε bằng không.

Giả định 2: Các ε_i có phương sai bằng nhau.

Giả định 3: Không có tương quan giữa các ε_i.

Giả định 4: Biến giải thích là phi ngẫu nhiên, tức là các giá trị của chúng là các số đã được xác định. Không có quan hệ tuyến tính hoàn toàn giữa các X_i.

Theo định lý Gauss - Markov, với các giả định từ 1 – 4 thì $\hat{\beta}=\left(X^TX\right)^{-1}X^TY$ là ước lượng tuyến tính không chệch có phương sai bé nhất. Để tiến hành ước lượng và kiểm định các hệ số mô hình thì người ta cần đến giả định 5 đó là véctơ sai số có phân phối chuẩn. Như đã trình bày ở phần trước khi dùng phương pháp bootstrap thì ta không cần giả định gì về phân phối. Do đó khi mô hình hồi qui không đáp ứng được giả định 5 thì có thể dùng phương pháp bootstrap để ước lượng hay kiểm định các hệ số.

Bootstrap ước lượng sai số tiêu chuẩn cho hệ số β_i là

$$se^{\#}(\beta_j)=\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\beta_j^{\#i}-\overline{\beta}_j^{\#}\right)^2}$$

Trong đó $\beta_j^{\#i}$ là giá trị ước tính cho β_j của mẫu thứ i và $\overline{\beta}_j^{\#}$ là giá trị trung bình của B giá trị $\beta_j^{\#i}$. Đồng thời chúng ta cũng có thể dùng phương pháp bootstrap để tìm khoảng ước lượng cho β_i.

6. ỨNG DỤNG BOOTSTRAP TRONG VIỆC XÁC ĐỊNH MẬT ĐỘ XƯƠNG CỦA PHỤ NỮ VIỆT NAM

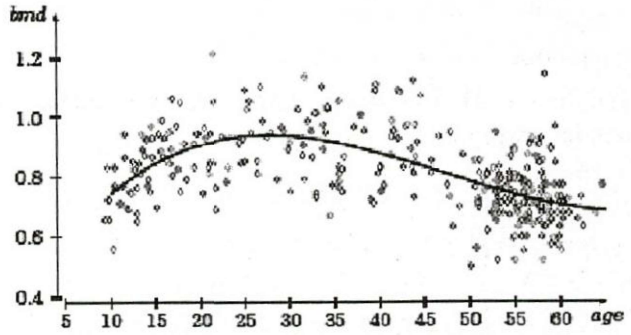
Trong phần này chúng tôi sẽ trình bày một ứng dụng của phương pháp bootstrap trong việc xác định mật độ xương của phụ nữ Việt Nam. Mật độ chất khoáng trong xương (bone mineral density - bmd) là một yếu tố rất quan trọng trong vấn đề tiên lượng mức độ gãy xương ở phụ nữ sau thời kì mãn kinh. Những người có bmd thấp thường có nguy cơ gãy xương cao. Cứ mỗi độ lệch tiêu chuẩn giảm bmd thì nguy cơ gãy xương tăng khoảng 2 đến 3 lần. Ở độ tuổi vị thành niên, bmd tăng nhanh, đạt đến độ cao nhất vào khoảng độ tuổi 18 - 30. Đến thời kỳ sau mãn kinh (tức sau khoảng 50 tuổi), bmd bắt đầu giảm dần dần và dẫn đến nguy cơ gãy xương. Để chẩn đoán bệnh loãng xương, tổ chức y tế thế giới đưa ra chỉ số

$$T = \frac{bmd_A - bmdp}{sd}$$

Ở đây bmd_A là mật độ xương của người A, $bmdp$ là mật độ xương tối đa của một quần thể (một nhóm người hoặc của một dân tộc nào đó) và sd là độ lệch tiêu chuẩn của mật độ xương tối đa. Nếu chỉ số T của một người phụ nữ dưới (-2.5) thì người đó được chẩn đoán bị loãng xương. Vấn đề quan trọng được đặt ra là ước lượng các tham $bmdp$ và sd.

Số liệu sử dụng trong bài báo này là sở hữu của Bác sĩ Nguyễn Thị Thanh Hương (Đại học Y Hà Nội) và Giáo sư Nguyễn Văn Tuấn (Viện nghiên cứu Y khoa Garvan, Úc). Trong giới hạn của bài báo này chúng tôi chỉ nghiên cứu độ lệch tiêu chuẩn của mật độ xương tối đa sd. Mô hình thống kê được dùng để biểu diễn mối quan hệ giữa mật độ xương và độ tuổi là mô hình hồi qui đa thức bậc ba có dạng

$$bmd_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 age_i^3 + \varepsilon_i,\ i = 1,\ldots,n$$



Hình 2. Mô hình quan hệ bmd và age

Với mỗi giá trị $A = age$ (tuổi) ta ước tính $B = bmd$ theo mô hình sau

$$B = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 A^2 + \hat{\beta}_3 A^3,$$

trong đó $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ và $\hat{\beta}_3$ là các hệ số. Độ tuổi đạt mật độ xương tối đa được tính bởi công thức

$$A_{\max} = \frac{-\hat{\beta}_2 - \sqrt{\hat{\beta}_2^2 - 3\hat{\beta}_1\hat{\beta}_3}}{3\hat{\beta}_3}$$

Giá trị mật độ xương tối đa

$B_{\max} = \hat{\beta}_0 + \hat{\beta}_1 A_{\max} + \hat{\beta}_2 A_{\max}^2 + \hat{\beta}_3 A_{\max}^3$.

Phương pháp bootstrap có thể dùng để ước tính các giá trị $A_{\max}^{\#i}$, ($i = 1,\ldots,B$). Hơn nữa, độ lệch tiêu chuẩn của mật độ xương tối đa được tính bởi

$$sd = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(A_{\max}^{\#i} - \overline{A}_{\max}^{\#}\right)^2}.$$

Để hỗ trợ tính toán, chúng tôi sử dụng phần mềm phân tích thống kê R. Sau đây là thuật toán để ước tính độ lệch tiêu chuẩn của mật độ xương tối đa. Dữ liệu được lưu với tên file là data.txt.

```
>setwd("C:/")
>data<-read.table("data.txt",header=TRUE,na.strings=".")
>attach(data)
>n<- length ( age) # xac dinh co mau
>B < -100000 #So lan tai lay mau
#Cac doi tuong de luu cac he so
>beta0 <- numeric (B)
>beta1 <- numeric (B)
>beta2 <- numeric (B)
>beta3 <- numeric (B)
# Thuc hien phep lap tinh cac he so
>for (i in 1:B)
{ Resample <- Data[ sample (1:n, n, replace =T), ]
y <- Resample [, " bmd "]
x <- Resample [, " age "]
fix <- lm(y ~ x+I(x ^2)+I(x ^3))#Uoc tinh cac he so hoi qui
beta0 [i] <- fix$coefficients[1]
beta1 [i] <- fix$coefficients[2]
beta2 [i] <- fix$coefficients[3]
beta3 [i] <- fix$coefficients[4]}
>A.max<- (-beta2-sqrt(beta2^2 - 3*beta3*beta1))/(3*beta3)
>B.max <- beta0 + beta1*A.max + beta2*A.max^2 + beta3*A.max^3
>sd(B.max) #Do lech chuan cua mat do xuong toi da
[1] 0.01299935
```

Kết quả của thuật toán trên được cho bởi [1].

7. KẾT LUẬN

Trong thống kê ứng dụng, ngoài các phương pháp ước lượng thống kê thông thường như ước lượng hợp lý cực đại, ước lượng phi tham số, v.v... ta còn có phương pháp bootstrap. Không những phương pháp bootstrap không cần giả định về phân phối mà nó còn giải quyết những vấn đề mà trước đây tưởng như không giải được.

Trong bài báo này chúng tôi đã trình bày một ứng dụng của phương pháp bootstrap trong việc xác định một chỉ số quan trong của căn bệnh nguy hiểm như loãng xương. Đây là một trong những ưu thế đặc biệt của thống kê bootstrap.

LỜI CẢM ƠN: *Chúng tôi xin cảm ơn hai người chủ trì công trình nghiên cứu y khoa là Bác sĩ Nguyễn Thị Thanh Hương (Đại học Y Hà Nội) và Giáo sư Nguyễn Văn Tuấn (Viện nghiên cứu y khoa Garvan, Úc) đã cung cấp số liệu về bệnh loãng xương ở Việt Nam và có những gợi ý sâu sắc để nghiên cứu này được thực hiện. Cũng xin đồng cảm ơn các thành viên khác cùng tham gia công trình bao gồm Giáo sư Phạm Thị Minh Đức, Lê Hồng Quang, Nguyễn Văn Định, Nguyễn Bá Đức, Nguyễn Huy Bình, Nguyễn Tuấn Anh, Lê Tuấn Thành, và Bo von Schoultz.*

APPLICATION OF BOOTSTRAP IN ESTIMATING THE BONE MINERAL DENSITY OF VIETNAMESES WOMEN

Nguyen Van Thu[(1)], Nguyen Duc Phuong[(2)]
(1)International University, VNU-HCM
(2) University of Natural Sciences, VNU-HCM

ABSTRACT: *In this paper, we apply the bootstrap method to study the standard deviation for bone mineral density of Vietnameses women. This result is important in recognizing seriousness of the osteoporosis*

TÀI LIỆU THAM KHẢO

[1]. Michaelr. Chernick. *Bootstrap Methods:A Guide for Practitioners and Researchers*. A John Wiley & Sons, Inc., Publication. (2007).

[2]. Bradley Efron. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Inc., Publication. (1994).

[3]. Phillip Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Publication. (2004).

[4]. F.M. Dekking and C. Kraikamp. *A Modern Introduction to Probability and Statistics*. Springer Publication. (2007).

[5]. John Bibby and Helge Toutenburg. *Prediction and Improved Estimation In Linear Models*. A John Wiley & Sons, Inc., Publication. (1977).

[6]. Roger W. Johnson. *An Introduction To The Bootstrap*. Teaching Statistics. 2001; 23: 49 - 54. (2001).

[7]. Chris Ricketts and John Berry. *Teaching Statistics Through Resampling*. Center for Teaching Mathematics, University of Plymouth, UK.

[8]. Jason S Haukoos and Roger J Lewis. *Advanced Statistics: Bootstrapping Confidence Intervals For statistics with ``Difficult'' Distributions*. Academic Emergency Medicine. Apr 2005; 12, 4: 360 - 365; ProQuest Medical Library. (2005).

[9]. James Carpenter and John Bithell. *Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians*. Statist. Med; 19:1141 - 1164. (2000).

[10]. Kenneth A. Bollen and Robert Stine. *Direct and Inderect Effects: Classical and Bootstrap Estimates of Varuability*. Sociological Methodology; 20: 115 - 140. (1990).