**17**

**Phương pháp Bootstrap**

Năm 1979, Giáo sư Bradley Efron (1) giới thiệu phương pháp phân tích có tên là “Bootstrap”. Chỉ trong vòng vài năm, phương pháp Bootstrap (2) được phát triển thêm và ứng dụng rất phổ biến, mà có người nói ví von là sự lan rộng của phương pháp này như là một trận … cháy rừng! Trong thực tế, cộng đồng thống kê học xem phương pháp Bootstrap là một cuộc *cách mạng* trong khoa học thống kê. Trong chương này, tôi sẽ giải thích ý tưởng và vài ứng dụng phổ biến của phương pháp Bootstrap. Tôi sẽ không giải thích lí thuyết đằng sau phương pháp Bootstrap, nhưng hi vọng qua vài ứng dụng, bạn đọc sẽ tìm hiểu những cơ sở lí thuyết của phương pháp này.

Một trong những chức năng chính yếu của thống kê là tóm lược một mẫu nghiên cứu (sample) và khái quát hoá dữ liệu (hay phát hiện) ra một quần thể (population) mà mẫu được thu thập. Chúng ta biết rằng các chỉ số liên quan đến mẫu được gọi là *statistic.*  Một số statistic cơ bản là trung bình, trung vị, độ lệch chuẩn, tỉ lệ, v.v. Dĩ nhiên, một statistic hay chỉ số thống kê như số trung bình chẳng hạn sẽ dao động từ mẫu này đến mẫu khác, và chúng ta muốn biết mức độ dao động chung quanh trung bình quần thể. Chúng ta có thể dùng lí thuyết xác suất để ước tính tất cả các giá trị trung bình mẫu khả dĩ, và trình bày trong một phân bố gọi là *phân bố mẫu* (sampling distribution).

Nhưng trong nhiều trường hợp chúng ta không có lí thuyết xác suất để xây dựng một phân bố mẫu. Chẳng hạn như chúng ta không có phương pháp toán nào để xây dựng phân bố mẫu cho trung vị. Trong những trường hợp này, phương pháp Bootstrap có thể giúp chúng ta giải quyết vấn đề một cách lịch thiệp. Thật vậy, với Bootstrap, chúng ta có thể tạo ra phân bố mẫu cho bất cứ chỉ số thống kê nào mà có khi chẳng cần đến lí thuyết xác suất.

**17.1 Lấy mẫu có hoàn lại và không có hoàn lại**

Để hiểu phương pháp Bootstrap, chúng ta phải điểm qua hai cách lấy mẫu: lấy mẫu có hoàn lại (sampling with replacement) và lấy mẫu không có hoàn lại (sampling without replacement). Để minh hoạ cho hai cách lấy mẫu, giả dụ chúng ta có một *quần thể* gồm 10 giá trị như sau:

**population = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)**

*Lấy mẫu không hoàn lại* có nghĩa là mỗi lần lấy một giá trị từ quần thể, giá trị đó không được hoàn lại cho quần thể. Vì thế, giá trị của mẫu luôn luôn đặc thù, không giống nhau. Trong R, chúng ta có hàm sample(x, n, replace=F)để lấy mẫu không hoàn lại. Ví dụ:

**> sample(population, 5, replace=F)**

**[1] 9 1 5 3 8**

Trong hàm trên, chúng ta lấy mẫu với 5 đối tượng từ quần thể population, và không có hoàn lại. Chú ý tất cả các giá trị mẫu không giống nhau. Chúng ta có thể lấy mẫu nhiều lần, và mỗi mẫu có giá trị khác nhau:

**> sample(population, 5, replace=F)**

**[1] 7 8 5 6 10**

**> sample(population, 5, replace=F)**

**[1] 7 2 5 3 1**

**> sample(population, 5, replace=F)**

**[1] 3 2 7 6 9**

*Lấy mẫu có hoàn lại* có nghĩa là lấy một giá trị từ quần thể, rồi bỏ giá trị mới lấy ra vào quần thể đó, và lần hai tiếp tục lấy một giá trị khác (nhưng cũng có thể là giá trị cũ). Trong R, chúng ta có hàm sample(x, n, replace=T) giúp lấy mẫu có hoàn lại:

**> sample(population, 5, replace=T)**

**[1] 1 9 5 4 1**

Trong hàm trên, chúng ta lấy 5 giá trị từ quần thể population, nhưng có hoàn lại. Chú ý giá trị 1 lặp lại 2 lần. Chúng ta có thể lấy nhiều mẫu như thế và kết quả cũng có thể lặp lại hơn 2 lần, hay không lặp lại:

**> sample(population, 5, replace=T)**

**[1] 8 7 7 3 3**

**> sample(population, 5, replace=T)**

**[1] 4 10 10 2 10**

**> sample(population, 5, replace=T)**

**[1] 2 9 4 6 10**

Phương pháp bootstrap dựa vào ý tưởng lấy mẫu có hoàn lại. Ý tưởng chủ đạo của bootstrap là dùng dữ liệu của một mẫu nghiên cứu để xây dựng một quần thể, với mục tiêu là xấp xỉ phân bố mẫu của một chỉ số thống kê.

**17.2 Giải thích bootstrap: xác định khoảng tin cậy 95% cho trung vị**

Chúng ta muốn xác định khoảng tin cậy 95% (KTC95) cho số trung vị. Nhưng không có lí thuyết thống kê nào cung cấp một giải pháp để xác định KTC95 cho số trung vị. Trong trường hợp này, phương pháp bootstrap có thể rất có ích để đáp ứng mục tiêu. Để cụ thể hoá vấn đề, chúng ta có số liệu về đau thấp khớp của 11 bệnh nhân như sau:

0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, và 0.25

hay dùng ngôn ngữ R:

pain=c(0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, 0.25)

(Chú ý chỉ số càng cao, độ đau càng nghiêm trọng). Số trung bình của 11 bệnh nhân là 0.163 và độ lệch chuẩn 0.112. Vì số trung bình thấp hơn 2 lần độ lệch chuẩn, chúng ta có thể kết luận rằng biến số này không tuân theo luật phân bố chuẩn. Số trung vị của pain là 0.15:

> median(pain)

[1] 0.15

Vấn đề bây giờ là xác định khoảng tin cậy 95% của số trung vị. Nói cách khác, nếu nghiên cứu được lặp lại 1000 lần, và mỗi lần chọn 11 đối tượng, thì KTC95 của số trung vị ra sao. Phương pháp bootstrap rất có ích để giải quyết vấn đề. Phương pháp này được tiến hành như sau:

* Bước 1: Bắt đầu bằng mẫu gốc *x*1, *x*2, *x*3, …, *xn*. Trong ví dụ trên:

0.05, 0.05, 0.05, 0.05, 0.10, 0.15, 0.20, 0.25, 0.25, 0.30, 0.35

* Bước 2: Chọn ngẫu nhiên *n* cá nhân từ mẫu gốc với qui trình *lấy mẫu có hoàn lại*. Mỗi lần chọn mẫu, tính số trung vị và tạm gọi số này là *mi*.

* Bước 3: Lặp lại bước hai *N* lần (*N* thường là 1000 hay 10000 hay thậm chí 1 triệu – tùy theo nhu cầu). Trong trường hợp trên, 10 mẫu đầu tiên có thể là:

Mẫu 1: 0.05 0.05 0.10 0.05 0.20 0.20 0.05 0.25 0.10 0.10 0.30 🡪 0.10

Mẫu 2: 0.05 0.25 0.30 0.05 0.30 0.30 0.05 0.05 0.25 0.05 0.35 🡪 0.25

Mẫu 3: 0.35 0.10 0.05 0.25 0.05 0.05 0.20 0.25 0.15 0.25 0.10 🡪 0.15

Mẫu 4: 0.05 0.05 0.10 0.25 0.15 0.05 0.20 0.05 0.10 0.25 0.05 🡪 0.10

Mẫu 5: 0.30 0.25 0.05 0.25 0.25 0.05 0.20 0.05 0.25 0.05 0.05 🡪 0.20

Mẫu 6: 0.05 0.25 0.10 0.05 0.05 0.15 0.25 0.05 0.05 0.05 0.05 🡪 0.05

Mẫu 7: 0.05 0.15 0.25 0.05 0.05 0.30 0.20 0.25 0.30 0.05 0.35 🡪 0.20

Mẫu 8: 0.05 0.05 0.20 0.05 0.10 0.05 0.05 0.10 0.20 0.10 0.05 🡪 0.05

Mẫu 9: 0.05 0.05 0.10 0.25 0.20 0.25 0.25 0.20 0.35 0.25 0.35 🡪 0.25

Mẫu 10: 0.05 0.05 0.05 0.25 0.35 0.25 0.25 0.15 0.20 0.20 0.15 🡪 0.20

v.v.

(Chú ý: số phía sau 🡪 có nghĩa là số trung vị cho từng mẫu)

* Bước 4: Đến đây chúng ta có *N* số trung vị. Sắp xếp *N* số từ thấp đến cao và đánh số: 1, 2, 3, …, *N*. Chọn số trung vị ở hạng 2.5% và 97.5% của *N* số trung vị, và đó chính là khoảng tin cậy 95%. Chẳng hạn như nếu *N* = 1000 lần, thì khoảng tin cậy 95% chính là số trung vị hàng thứ 25 và 975.

Các bước tính toán trên có thể thực hiện bằng ngôn ngữ R (hay một ngôn ngữ hay phần mềm nào mà bạn đọc quen thuộc) rất dễ dàng. Đối với R, các mã sử dụng (và giải thích kèm theo) như sau:

# nhập các số liệu gốc vào một vector có tên là x

pain = c(0.05,0.15,0.35,0.25,0.20,0.05,0.10,0.05,0.30,0.05,0.25)

# bước 2 – xác định xem có bao nhiêu số liệu trong vector x

n = length(pain)

# muốn lấy 10000 mẫu từ số liệu gốc, còn gọi là bootstrap samples

B = 10000

# tạo một vector mới để chứa số trung vị

median = numeric(B)

# bắt đầu lấy B mẫu và mỗi mẫu tính toán số trung vị

for (i in 1:B)

{

bs.sample = sample(pain, n, replace=T)

median[i] = median(bs.sample)

}

# Bước 3: ước tính khoảng tin cậy 95%

quantile(median, probs=c(0.025, 0.975))

Hàm quantile cho chúng ta biết khoảng tin cậy 95% của số trung vị là 0.05 đến 0.25. Chúng ta cũng có thể xem phân bố của 10000 số trung vị mẫu:

> hist(median, col="blue", border="white", main="")



Qua ví dụ trên, chúng ta có thể tóm tắt phương pháp bootstrap như sau: phương pháp bootstrap là một phương pháp mô phỏng dựa vào dữ liệu thực tế để dùng cho suy luận khoa học. Phương pháp này có thể tóm lược trong 3 bước:

1. Từ một mẫu nghiên cứu *n* đối tượng, lấy nhiều mẫu bootstrap [có hoàn lại] và mỗi mẫu có *n* đối tượng;
2. Trong mỗi mẫu bootstrap, tính chỉ số thống kê θ\*;
3. Xem xét phân bố θ\*.

**17.3 Ứng dụng 1: So sánh hai số trung vị**

Số liệu dưới đây là số ngày sống sót của hai nhóm chuột. Nhóm thứ nhất là nhóm chứng (không được điều trị), và nhóm thứ hai là nhóm được điều trị. Thời gian sống sót của nhóm 1 (trung vị) là 46 ngày, và nhóm điều trị là 94 ngày. Câu hỏi đặt ra là điều trị có thật sự ảnh hưởng đến thời gian sống sót:

Nhóm chứng (n = 9): 52, 104, 146, 10, 51, 30, 40, 27, 46

Nhóm điều trị (n = 7): 94, 197, 16, 38, 99, 141, 23

Chúng ta có thể ứng dụng phương pháp bootstrap để giải đáp câu hỏi trên. Chúng ta sẽ dùng R như sau:

# nhập dữ liệu

control = c(52, 104, 146, 10, 51, 30, 40, 27, 46)

treatment = c(94, 197, 16, 38, 99, 141, 23)

# mẫu bootstrap 10000

B = 10000

# định danh hai biến cho 2 nhóm

con = rep(NA, B)

trt = rep(NA, B)

# lấy mẫu bootstrap

for (i in 1:B) {

con[i] = median(sample(control, replace=T))

trt[i] = median(sample(treatment, replace=T))

dif[i] = trt[i]-con[i]

}

# tìm khoảng tin cậy 95% của dif

ci.bs = c(quantile(dif, c(0.025)), quantile(dif, c(0.975)))

# tạo dataframe để vẽ biểu đồ

diff = data.frame(dif)

library(ggplot2)

p = ggplot(data=diff, aes(x=dif))

p = p + geom\_histogram(aes(y=..density..), binwidth=5, color="black", fill="white")

p = p + geom\_density(alpha=0.2, fill="#FF6666")

p = p + geom\_vline(xintercept=0, col="#BB0000", linetype="dashed")

p = p + geom\_vline(xintercept=ci.bs[1], col="#BB0000", linetype="longdash")

p = p + geom\_vline(xintercept=ci.bs[2], col="#BB0000", linetype="longdash")

p = p + xlab("Difference")

p



**> quantile(dif, c(0.025, 0.5, 0.975))**

**2.5% 50% 97.5%**

**-29 48 101**

Chúng ta thấy tính trung bình, hiệu số của 2 trung vị là 48 ngày (tính trung bình, nhóm được điều trị sống lâu hơn nhóm chứng 48 ngày), nhưng khoảng tin cậy 95% dao động từ -29 đến 101 ngày. Nói cách khác, nhóm điều trị có thể có thời gian sống sót thấp hơn nhóm chứng 29 ngày, hoặc cao hơn nhóm chứng 101 ngày. Với khoảng tin cậy 95% như thế, chúng ta không thể kết luận hiệu quả của điều trị về thời gian sống sót.

**17.4 Ứng dụng 2: So sánh hai tỉ lệ**

Bảng số liệu dưới đây tóm lược một nghiên cứu về aspirin và bệnh tim mạch ở người đàn ông trung niên. Nghiên cứu được thiết kế theo mô hình thử nghiệm lâm sàng đối chứng ngẫu nhiên (randomized controlled trial). Theo đó, một nhóm gồm 22071 nam được chia thành 2 nhóm một cách ngẫu nhiên: nhóm thứ nhất bao gồm 11037 được cho uống aspirin mỗi ngày, và nhóm hai bao gồm 11034 người không uống aspirin (còn gọi là nhóm chứng). Các nhà nghiên cứu muốn biết aspirin cho giảm nguy cơ bệnh nhồi máu cơ tim. Kết quả sau 3 năm nghiên cứu có 104 người trong nhóm điều trị và 189 người trong nhóm chứng bị nhồi máu cơ tim:

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhóm** | **Số đối tượng** | **Số bị bệnh (nhồi máu cơ tim)** | **Số không mắc bệnh** |
| Aspirin | 11037 | 104 | 10937 |
| Placebo (chứng) | 11034 | 189 | 10845 |

Câu hỏi nghiên cứu là tỉ lệ (hay nguy cơ) bệnh của hai nhóm có thực sự khác nhau. Một phương pháp đơn giản nhất để trả lời câu hỏi trên là so sánh hai nguy cơ (risk). Trong y khoa, nguy cơ có thể xem như là xác suất mắc bệnh. Dựa vào số liệu trên, chúng ta có thể tính nguy cơ mắc bệnh của nhóm aspirin là

Risk1= 104 / 11037 = 0.0094

và nguy cơ của nhóm chứng:

Risk2 = 189 / 11034 = 0.0171

Tỉ số của hai nguy cơ (Risk1 / Risk2) thường được gọi là *risk ratio (RR).* Nếu RR = 1, thuốc không có hiệu quả; nếu RR < 1, aspirin giảm nguy cơ bệnh; nhưng nếu RR > 1 thì điều đó có nghĩa là aspirin tăng nguy cơ bệnh. Trong trường hợp trên, RR = 0.0094 / 0.0171 = 0.55. Nói cách khác, nguy cơ mắc bệnh trong nhóm aspirin bằng 55% nhóm chứng. Một cách khác để kết luận là aspirin giảm nguy cơ nhồi máu cơ tim 45%.

Nhưng vấn đề là ước tính KTC95 của RR, và qua đó chúng ta có thể suy luận về hiệu quả của aspirin. Trong thực tế, chúng ta có thể ước tính KTC95 bằng lí thuyết xác suất [3], nhưng ở đây, chúng ta sẽ ứng dụng phương pháp bootstrap để ước tính KTC95 cho RR.

# nhập dữ liệu cho 2 nhóm

n = c(11037, 11034)

disease = c(104, 189)

# tạo 2 biến có tên là aspirin và control. Mỗi biến bao gồm

# giá trị 1 (bệnh), 0 (không bệnh). Ví dụ: hàm rep(1, disease[1]) tạo ra

# 104 đối tượng với giá trị 1. Tương tự rep(0, n[1]-disease[1]) tạo ra

# 11037-104 đối tượng với giá trị 0. Hàm c(rep(1, disease[1]), rep(0,

# n[1]-disease[1])) là nhập hai vector thành một và gọi là aspirin.

aspirin = c(rep(1, disease[1]), rep(0, n[1]-disease[1]))

control = c(rep(1, disease[2]), rep(0, n[2]-disease[2]))

# xác định số mẫu bootstrap

B = 10000

# tạo ra hai vector trống cho nhóm aspirin và control

bs1 = rep(NA, B)

bs2 = rep(NA, B)

# lấy mẫu bootstrap

for (i in 1:B) {

resample1 = sample(aspirin, n[1], replace=T)

dis1 = sum(resample1)

resample2 = sample(control, n[2], replace=T)

dis2 = sum(resample2)

bs1[i] = dis1/n[1]

bs2[i] = dis2/n[2]

}

RR = bs1/bs2

quantile(RR, c(0.025, 0.50, 0.975)

Với hàm quantile trên, chúng ta sẽ thấy RR trung bình là 0.55 (tức như chúng ta tính “thủ công”). Hơn nữa, khoảng tin cậy 95% của RR dao động trong khoảng 0.43 đến 0.70. Nói cách khác, aspirin giảm nguy cơ bệnh từ 31% đến 57%. Vì khoảng tin cậy 95% đều dưới 1, chúng ta có chứng cứ để kết luận aspirin có hiệu quả giảm nguy cơ bệnh nhồi máu cơ tim.

Chúng ta có thể dùng ggplot2 để vẽ biểu đồ phân bố của RR. Để vẽ biểu đồ này, trước hết chúng ta phải đưa 10000 giá trị **RR** vào một dataframe tạm gọi là **rr**. Sau đó vẽ biểu đồ phân bố, và vẽ khoảng tin cậy 95%.

# đưa khoảng tin cậy 95% vào biến ci.bs

ci.bs = c(quantile(RR, 0.025), quantile(RR, 0.975))

# tạo ra một dataframe

rr = data.frame(RR)

library(ggplot2)

p = ggplot(data=rr, aes(x=RR))

p = p + geom\_histogram(aes(y=..density..), bindwidth=0.02, col="blue", fill="white")

p = p + geom\_density(alpha=0.2, fill="#FF6666")

p = p + geom\_vline(xintercept=1, col="#BB0000", linetype="dashed")

p = p + geom\_vline(xintercept=ci.bs[1], col="#00AA00", linetype="longdash")

p = p + geom\_vline(xintercept=ci.bs[2], col="#00AA00", linetype="longdash")

p



Dĩ nhiên, với phương pháp bootstrap, chúng ta có thể tính hiệu số (Risk1 – Risk2) thay vì tỉ số nguy cơ (Risk1 / Risk2).

**17.5 Ứng dụng 3: So sánh với phương pháp cổ điển**

Dữ liệu sau đây được thu thập từ một nghiên cứu đánh giá một phương pháp can thiệp lên bệnh nhân mất trí nhớ. Mỗi bệnh nhân phải trải qua 20 test về xử lí tình huống hàng ngày, như tự làm vệ sinh, thay quần áo, mở cửa bằng chìa khoá, v.v. Số liệu dưới đây phản ảnh phần trăm bệnh nhân thành công trong số 20 test.

Nhóm can thiệp (n = 11): 0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, 0.25

Nhóm chứng (n = 8): 0, 0.15, 0, 0.05, 0, 0, 0.05, 0.10

Câu hỏi nghiên cứu là can thiệp có hiệu quả tăng sự độc lập của bệnh nhân. Giả thuyết là phần trăm làm đúng của nhóm can thiệp không khác với nhóm chứng. Chúng ta có thể dùng phương pháp kiểm định t để kiểm định giả thuyết:

**treated = c(0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05,**

**0.30, 0.05, 0.25)**

**control = c(0, 0.15, 0, 0.05, 0, 0, 0.05, 0.10)**

**t.test(treated, control)**

**> t.test(treated, control)**

Welch Two Sample t-test

data: treated and control

t = 3.0583, df = 15.485, **p-value = 0.007736**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

**0.03655926 0.20321347**

sample estimates:

mean of x mean of y

0.1636364 0.0437500

Chúng ta thấy giả thuyết không có hiệu quả có thể bác bỏ vì trị số P = 0.0077, và khoảng tin cậy 95% khác biệt giữa hai nhóm là 0.037 đến 0.203.

Nhưng phương pháp kiểm định t có vấn đề về giả định, vì số liệu không tuân theo luật phân bố chuẩn, và phương sai của hai nhóm cũng không tương đương nhau. Do đó, chúng ta sẽ ứng dụng phương pháp bootstrap để kiểm định giả thuyết. Chúng ta sẽ làm theo các bước:

1. Lấy mẫu từ nhóm điều trị
2. Lấy mẫu từ nhóm chứng
3. Tính hiệu số của 2 số trung bình
4. Lặp lại bước 1-3 B lần
5. Xem xét phân bố để xác định khoảng tin cậy 95%

Chúng ta triển khai các bước trên qua R như sau:

treat = c(0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05,

0.30, 0.05, 0.25)

control = c(0, 0.15, 0, 0.05, 0, 0, 0.05, 0.10)

n1 = length(treat)

n2 = length(control)

B = 10000

difference = numeric(B)

no.effect = 0

for (i in 1:B) {

bs.treat <- sample(treat, n1, replace=T)

bs.control <- sample(control, n2, replace=T)

difference[i] = mean(bs.treat) - mean(bs.control)

if (difference[i] < 0) no.effect = no.effect+1

}

Chúng ta tìm hiểu trong số 10000 mẫu, có bao nhiêu mẫu mà can thiệp có ảnh hưởng:

**> no.effect/10000**

[1] 5e-04

Ở một góc độ đơn giản, chúng ta có thể nói xác suất can thiệp có hiệu quả là 1 – 0.0005 = 0.9995 hay 99.95%.

Khoảng tin cậy 95% giữa 2 nhóm dao động trong khoảng 0.05 đến 0.19:

> quantile(difference, probs=c(0.025, 0.50, 0.975))

2.5% 50% 97.5%

0.04659091 0.11931818 0.19261364

Chúng ta có thể xem qua phân bố của khác biệt giữa hai nhóm:

hist(difference, breaks=20, col="blue", border="white")



Chúng ta chú ý khoảng tin cậy 95% bootstrap của hai chỉ số trung bình giữa hai nhóm (0.05 đến 0.19) không khác nhiều so với phương pháp cổ điển (0.037 đến 0.203). Điều này cho thấy phương pháp cổ điển khá “robust” khi dữ liệu không đáp ứng các giả định phổ biến.

Nói tóm lại, phương pháp phân tích bootstrap dựa vào khái niệm tái chọn mẫu (resampling) và mô phỏng (dựa trên dữ liệu thực tế) để tìm phân bố cho một chỉ số thống kê. Trong thực tế, phương pháp bootstrap đã được ứng dụng cho rất nhiều vấn đề phức tạp mà phương pháp cổ điển không có giải pháp. Cần phải nói rằng phương pháp bootstrap không thể biến dữ liệu xấu thành dữ liệu tốt được. Trong vài trường hợp, phương pháp bootstrap vẫn thất bại khi nghiên cứu được thiết kế quá kém hay không phù hợp. Việc ứng dụng phương pháp bootstrap do đó đòi hỏi suy xét cẩn thận.

**Chú thích và tài liệu tham khảo**

(1) Efron B (1979). Bootstrap methods: Another look at jackknife. Ann. Stat. 7, 1-26.

(2) Danh từ “bootstrap” xuất phát từ cuốn tiểu thuyết “The Adventures of Baron Munchausen” của Rudolph Erich Raspe. Trong tiểu thuyết, vị bá tước bị rớt xuống đáy một cái hồ, ông nghĩ rằng ông có thể tự mình thoát khỏi bằng cách dùng dây giầy (bootstrap) để từ từ kéo mình lên khỏi hồ.

[3] Phương pháp ước tính khoảng tin cậy 95% của một tỉ số nguy cơ RR dựa vào phương pháp Greenwood. Theo đó, RR được hoán chuyển sang đơn vị log; tính phương sai và sai số chuẩn và khoảng tin cậy 95% cho logRR; hoán chuyển ngược lại đơn vị RR. Để dễ theo dõi, tôi in lại kết quả nghiên cứu trong bảng dưới đây:

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhóm** | **Số đối tượng** | **Số bị bệnh (nhồi máu cơ tim)** | **Số không mắc bệnh** |
| Aspirin | 11037 | 104 | 10937 |
| Placebo (chứng) | 11034 | 189 | 10845 |

Bước 1: hoán chuyển từ RR sang logRR. Chúng ta biết rằng RR = 0.55, nên

logRR = log(0.55) = -0.598

Bước 2: tính phương sai, sai số chuẩn và khoảng tin cậy 95% của logRR:

Sai số chuẩn của logRR:

Khoảng tin cậy 95% của logRR:

logRR ± 1.96×SE(logRR) = -0.598 ± 1.96×0.121

= -0.835, -0.361

Bước 3: hoán chuyển ngược lại đơn vị tỉ số:

95%CI RR = exp(-0.835), exp(-0.361)

= 0.43, 0.70