

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN MÔN HỌC  
TOÁN HỌC**

**ĐỀ TÀI:**

**BOOTSTRAP MEDTHOD**

**Giáo viên hướng dẫn: TS. Dương Tôn Đảm**

**Sinh viên thực hiện: CH1901020 - Nguyễn Trung Hiếu**

**Hồ Chí Minh, ngày 10 tháng 01 năm 2020**

**Mục lục**

[**I.** **Giới thiệu** 1](#_Toc61173443)

[**II.** **Khái niệm** 1](#_Toc61173444)

[**III.** **Phương pháp tiếp cận** 1](#_Toc61173445)

[**IV.** **Ưu điểm** 2](#_Toc61173446)

[**V.** **Nhược điểm** 2](#_Toc61173447)

[**VI.** **Một số ứng dụng của Bootstrap** 3](#_Toc61173448)

[**VII.** **Phương pháp nâng cao hiệu quả tính toán** 4](#_Toc61173449)

**PHƯƠNG PHÁP BOOTSTRAP**

# **Giới thiệu**

Năm 1979, Giáo sư Bradley Efron giới thiệu phương pháp phân tích có tên là "Bootstrap". Chỉ trong vòng vài năm, phương pháp Bootstrap được phát triển thêm và ứng dụng phổ biến, mà có người ví von là sự lan rộng của phương pháp này nhưng là một ma trận... cháy runwgf! Trong thực tế cộng đồng thông kê học xem phương pháp Bootstrap là một cách mạng trong khoa học thống kê.

Một trong những chức năng chính yếu của thống kê là tóm lược một mẫu nghiên cứu (sample) và khái quát dư liệu (hay phát hiện) ra một quần thể (population) mà mẫu ddược 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 thông 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 quẩn 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 số phân bố gọi là phân bố maaux (sampling distribution)

Nhưng trong nhiều trường hợp chúng ta không có lý 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. Với Bootstrap, chúng ta có thể tạo ra phân bố mẫu cho bất cứ chỉ sô thống kê nào mà có khi chẳng cần đến lý thuyết xác xuất.

# **Khái niệm**

Bootstrapping là một thủ tục thống kê lấy lại mẫu một tập dữ liệu để tạo ra nhiều mẫu mô phỏng. Quy trình này cho phép tính toán các sai số chuẩn, xây dựng khoảng tin cậy và thực hiện kiểm tra giả thuyết cho nhiều loại thống kê mẫu. Phương pháp Bootstrap là cách tiếp cận thay thế cho việc kiểm tra giả thuyết truyền thống và đáng chú ý là dễ hiểu hơn và hợp lệ với nhiều điều kiện hơn.

# **Phương pháp tiếp cận**

Ý tưởng cơ bản của bootstrapping là suy luận về một tập hợp từ dữ liệu mẫu (mẫu → tổng thể) có thể được mô hình hóa bằng cách lấy mẫu lại dữ liệu mẫu và thực hiện suy luận về một mẫu từ dữ liệu được lấy mẫu lại (resampled → sample). Vì dân số là không xác định, nên sai số thực sự trong thống kê mẫu so với giá trị dân số của nó là không xác định. Trong bootstrap-resamples, 'dân số' thực tế là mẫu, và điều này được biết đến; do đó chất lượng suy luận của mẫu 'đúng' từ dữ liệu được lấy mẫu lại (lấy mẫu lại → mẫu) có thể đo lường được.

Chính thức hơn, bootstrap hoạt động bằng cách xử lý suy luận của phân phối xác suất thực J , với dữ liệu ban đầu, tương tự như suy luận của phân phối thực nghiệm Ĵ , với dữ liệu được lấy mẫu lại. Có thể đánh giá độ chính xác của các suy luận liên quan đến Ĵ sử dụng dữ liệu được lấy mẫu lại vì chúng tôi biết Ĵ . Nếu Ĵ là một xấp xỉ hợp lý đối với J thì có thể suy ra chất lượng của suy luận trên J.

1. **Ưu điểm**

Một ưu điểm lớn của bootstrap là tính đơn giản của nó. Đây là một cách đơn giản để lấy được các ước tính về sai số tiêu chuẩn và khoảng tin cậy cho các ước lượng phức tạp về phân phối, chẳng hạn như điểm phân vị, tỷ lệ, tỷ lệ chênh lệch và hệ số tương quan. Bootstrap cũng là một cách thích hợp để kiểm soát và kiểm tra tính ổn định của kết quả. Mặc dù đối với hầu hết các vấn đề, không thể biết khoảng tin cậy thực sự, nhưng về mặt cơ bản thì bootstrap chính xác hơn các khoảng chuẩn thu được bằng cách sử dụng phương sai mẫu và các giả định về độ chuẩn. [16] Bootstrapping cũng là một phương pháp thuận tiện giúp tránh chi phí lặp lại thử nghiệm để lấy các nhóm dữ liệu mẫu khác.

# **Nhược điểm**

Mặc dù bootstrapping là (trong một số điều kiện) nhất quán về mặt tiệm cận , nó không cung cấp các đảm bảo về mẫu hữu hạn chung. Kết quả có thể phụ thuộc vào mẫu đại diện. Sự đơn giản rõ ràng có thể che giấu thực tế rằng các giả định quan trọng đang được thực hiện khi thực hiện phân tích khởi động (ví dụ: tính độc lập của các mẫu) trong đó chúng sẽ được nêu chính thức hơn trong các cách tiếp cận khác. Ngoài ra, bootstrapping có thể tốn nhiều thời gian.

Số lượng các mẫu bootstrap được đề xuất trong tài liệu đã tăng lên khi khả năng tính toán có sẵn đã tăng lên. Nếu kết quả có thể có hậu quả đáng kể trong thế giới thực, thì người ta nên sử dụng càng nhiều mẫu càng hợp lý, với khả năng và thời gian tính toán có sẵn. Việc tăng số lượng mẫu không thể làm tăng lượng thông tin trong dữ liệu gốc; nó chỉ có thể giảm ảnh hưởng của lỗi lấy mẫu ngẫu nhiên có thể phát sinh từ chính quy trình bootstrap. Hơn nữa, có bằng chứng cho thấy số lượng mẫu lớn hơn 100 dẫn đến những cải tiến không đáng kể trong việc ước lượng sai số tiêu chuẩn. Trên thực tế, theo nhà phát triển ban đầu của phương pháp khởi động, ngay cả việc đặt số lượng mẫu ở mức 50 cũng có thể dẫn đến ước tính sai số chuẩn khá tốt.

Adèr và cộng sự. đề xuất quy trình bootstrap cho các trường hợp sau:

Khi phân phối lý thuyết của một thống kê quan tâm là phức tạp hoặc chưa biết. Vì thủ tục khởi động không phụ thuộc vào phân phối nên nó cung cấp một phương pháp gián tiếp để đánh giá các thuộc tính của phân phối bên dưới mẫu và các tham số quan tâm có được từ phân phối này.

Khi kích thước mẫu không đủ để suy luận thống kê đơn giản. Nếu phân phối cơ bản là nổi tiếng, thì khởi động hệ thống cung cấp một cách để giải thích các biến dạng do mẫu cụ thể có thể không đại diện đầy đủ cho tập hợp gây ra.

Khi tính toán công suất phải được thực hiện và có sẵn một mẫu thử nghiệm nhỏ. Hầu hết các phép tính công suất và kích thước mẫu phụ thuộc nhiều vào độ lệch chuẩn của thống kê quan tâm. Nếu ước tính được sử dụng không chính xác, kích thước mẫu yêu cầu cũng sẽ sai. Một phương pháp để có được ấn tượng về sự thay đổi của thống kê là sử dụng một mẫu thử nghiệm nhỏ và thực hiện khởi động trên nó để có ấn tượng về phương sai.

Tuy nhiên, Athreya đã chỉ ra rằng nếu một người thực hiện một bootstrap ngây thơ trên mẫu có nghĩa là khi tổng thể cơ bản thiếu một phương sai hữu hạn (ví dụ: phân phối luật lũy thừa ), thì phân phối bootstrap sẽ không hội tụ đến cùng một giới hạn như trung bình của mẫu. Do đó, khoảng tin cậy dựa trên mô phỏng Monte Carlo của bootstrap có thể bị sai lệch. Athreya tuyên bố rằng "Trừ khi ai đó chắc chắn một cách hợp lý rằng phân phối cơ bản không nặng về đuôi , người ta nên ngần ngại sử dụng bootstrap ngây thơ".

# **Một số ứng dụng của Bootstrap**

***6.1 Ước tính phân phối trung bình mẫu***

Hãy xem xét một thí nghiệm tung đồng xu. Chúng tôi lật đồng xu và ghi lại xem nó chạm đầu hay sấp. Gọi X = x1 , x2 ,…, x10 là 10 quan sát từ thí nghiệm. x i = 1 nếu lật thứ i chạm đầu và 0 nếu ngược lại. Từ lý thuyết thông thường, chúng ta có thể sử dụng thống kê t để ước tính phân phối của trung bình mẫu,



Thay vào đó, chúng tôi sử dụng bootstrap, cụ thể là lấy mẫu lại trường hợp, để lấy ra phân phối {\ displaystyle {\ bar {x}}}{\ bar {x}}. Đầu tiên chúng tôi lấy mẫu lại dữ liệu để lấy mẫu lại bootstrap . Ví dụ về mẫu đầu tiên có thể trông giống như sau X 1 \* = x 2 , x 1 , x 10 , x 10 , x 3 , x 4 , x 6 , x 7 , x 1 , x 9. Có một số bản sao vì mẫu khởi động lại đến từ lấy mẫu với thay thế từ dữ liệu. Ngoài ra, số điểm dữ liệu trong mẫu khởi động lại bằng số điểm dữ liệu trong các quan sát ban đầu của chúng tôi. Sau đó, chúng tôi tính giá trị trung bình của mẫu lại này và thu được giá trị trung bình của bootstrap đầu tiên : μ 1 \*. Chúng tôi lặp lại quá trình này để có được mẫu thứ hai X 2 \* và tính giá trị trung bình khởi động thứ hai μ 2 \*. Nếu chúng ta lặp lại điều này 100 lần, thì chúng ta có μ 1 \*, μ 2 \*, ..., μ 100 \*. Điều này đại diện cho một phân phối bootstrap theo kinh nghiệmcủa trung bình mẫu. Từ phân phối thực nghiệm này, người ta có thể suy ra khoảng tin cậy bootstrap cho mục đích kiểm tra giả thuyết.

***6.2 Bayesian bootstrap***

Bootstrapping có thể được diễn giải trong một khuôn khổ Bayesian bằng cách sử dụng một lược đồ tạo các tập dữ liệu mới thông qua việc xác định lại trọng số của dữ liệu ban đầu. Đưa ra một bộ N điểm dữ liệu, trọng số được chỉ định cho điểm dữ liệu i trong một tập dữ liệu mới DJ Là Ở đâu xJ là danh sách có thứ tự từ thấp đến cao gồm {\ displaystyle N-1}N-1 số ngẫu nhiên được phân phối đồng đều trên [0,1], đứng trước 0 và kế tiếp là 1. Phân phối của một tham số được suy ra từ việc xem xét nhiều tập dữ liệu như vậy DJ sau đó có thể hiểu là phân phối sau trên tham số đó.

# **Phương pháp nâng cao hiệu quả tính toán**

Bootstrap là một kỹ thuật mạnh mẽ mặc dù có thể yêu cầu tài nguyên máy tính đáng kể cả về thời gian và bộ nhớ. Một số kỹ thuật đã được phát triển để giảm bớt gánh nặng này. Nhìn chung, chúng có thể được kết hợp với nhiều loại lược đồ Bootstrap khác nhau và nhiều lựa chọn thống kê khác nhau.

***7.1 Poisson bootstrap***

Bootstrap thông thường yêu cầu chọn ngẫu nhiên n phần tử từ một danh sách, tương đương với việc vẽ từ một phân phối đa thức. Điều này có thể yêu cầu một số lượng lớn các lần chuyển dữ liệu và là một thách thức để chạy các phép tính này song song. Đối với các giá trị lớn của n, Poisson bootstrap là một phương pháp hiệu quả để tạo các tập dữ liệu bootstrapped. Khi tạo một mẫu bootstrap duy nhất, thay vì lấy ngẫu nhiên từ dữ liệu mẫu có thay thế, mỗi điểm dữ liệu được ấn định một trọng số ngẫu nhiên được phân phối theo phân phối Poisson với lambda = 1. Đối với dữ liệu mẫu lớn, điều này sẽ lấy mẫu ngẫu nhiên gần đúng với thay thế. Điều này là do ước tính sau:



Phương pháp này cũng có lợi cho việc truyền dữ liệu và phát triển các tập dữ liệu, vì không cần biết trước tổng số mẫu khi bắt đầu lấy mẫu bootstrap.

***7.2 Bag of Little Bootstraps***

Đối với các tập dữ liệu lớn, việc giữ tất cả dữ liệu mẫu trong bộ nhớ và lấy lại mẫu từ dữ liệu mẫu thường là một điều cấm đoán về mặt tính toán. Bag of Little Bootstraps (BLB) cung cấp một phương pháp tổng hợp trước dữ liệu trước khi khởi động để giảm các ràng buộc về tính toán. Điều này hoạt động bằng cách phân vùng tập dữ liệu thành{\ displaystyle b}bcác nhóm có kích thước bằng nhau và tổng hợp dữ liệu trong mỗi nhóm. Tập dữ liệu tổng hợp trước này trở thành dữ liệu mẫu mới để lấy mẫu thay thế. Phương pháp này tương tự như Block Bootstrap, nhưng động cơ và định nghĩa của các khối rất khác nhau. Theo các giả định nhất định, phân phối mẫu phải gần đúng với kịch bản khởi động đầy đủ. Một hạn chế là số lượng nhóm{\ displaystyle b = n ^ gamma .Ở đâu gamma in [0,5,1] và các tác giả khuyên bạn nên sử dụng b = n ^ 0,7 như một giải pháp chung.