Bài giảng 3: Dữ liệu và phân phối mẫu

TS. Tô Đức Khánh

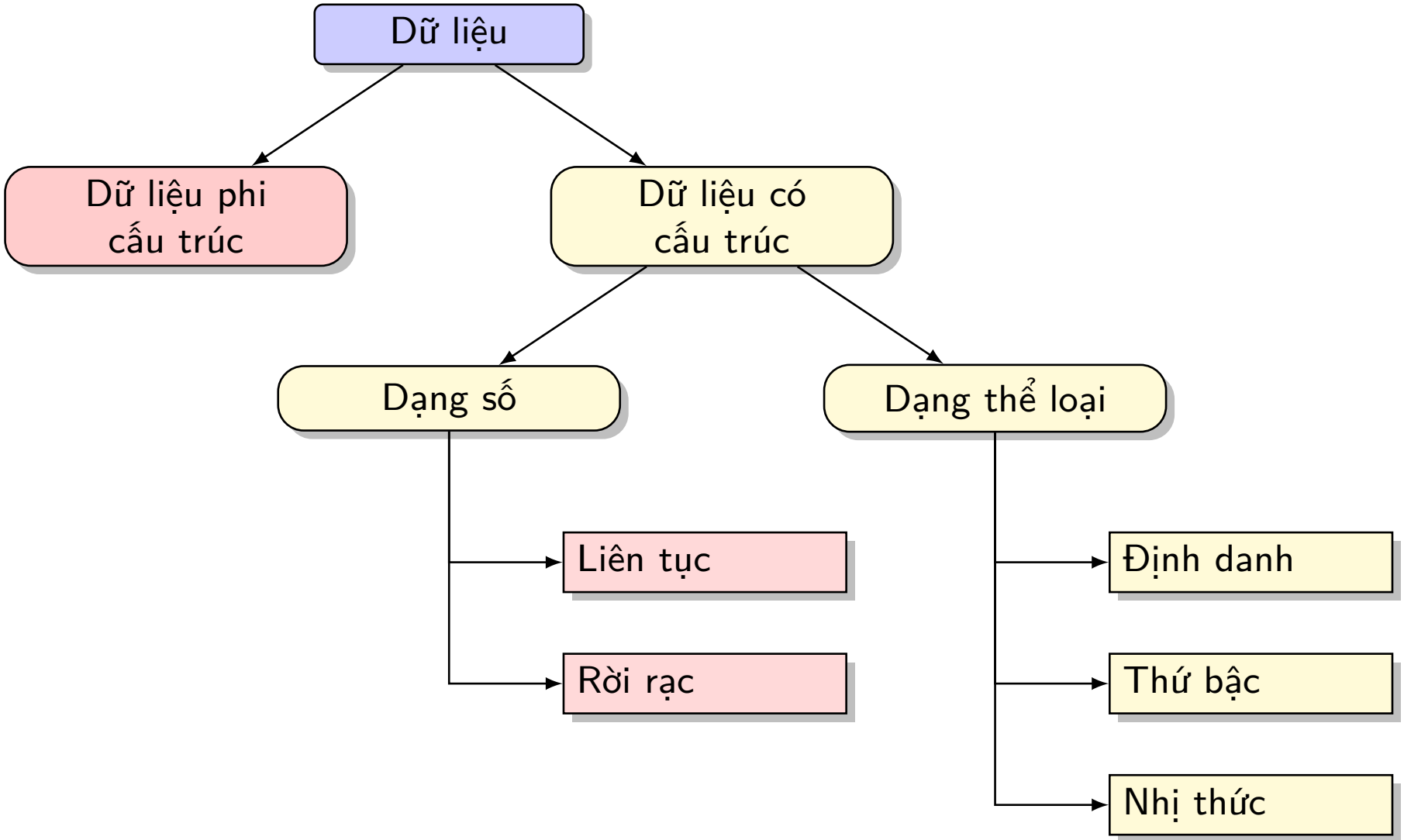
Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

Quần thể vs Dữ liệu

Trong phần trước ta đã học về dữ liệu và các đặc tính liên quan.



Quần thể vs Dữ liệu

Nhưng dữ liệu mà ta có, chỉ là một mẫu được quan sát/thu thập của 1 quần thể dữ liệu lớn hơn.

Quần thể vs Dữ liệu

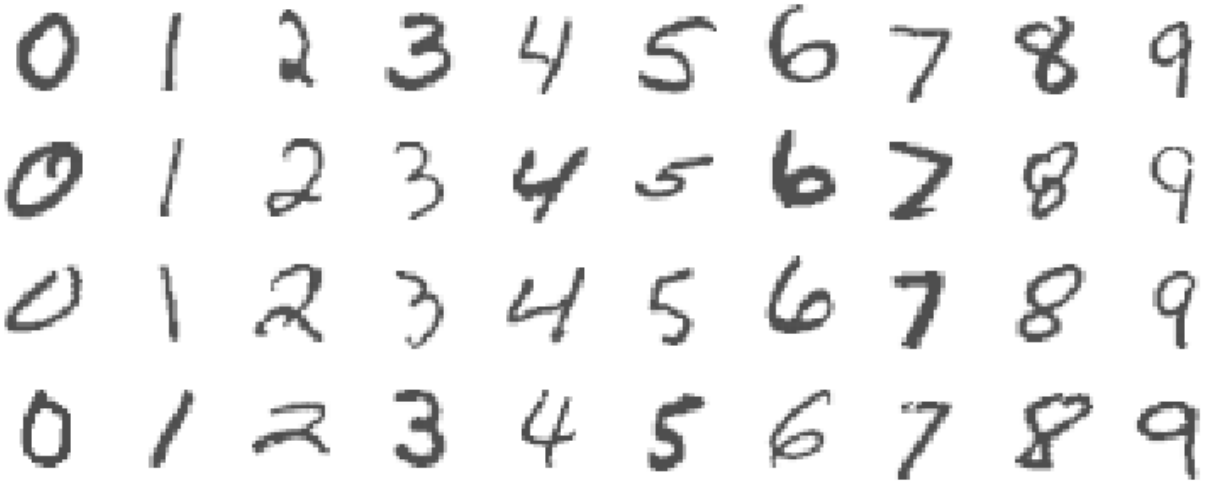
Ví dụ dữ liệu về thời gian cất cánh trễ của máy bay xuất phát ở New York

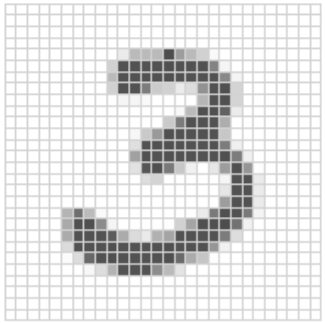
	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time
1	2013	1	1	517	515	2	830
2	2013	1	1	533	529	4	850
3	2013	1	1	542	540	2	923
4	2013	1	1	544	545	-1	1004
5	2013	1	1	554	600	-6	812
6	2013	1	1	554	558	-4	740
7	2013	1	1	555	600	-5	913
8	2013	1	1	557	600	-3	709
9	2013	1	1	557	600	-3	838
10	2013	1	1	558	600	-2	753
11	2013	1	1	558	600	-2	849

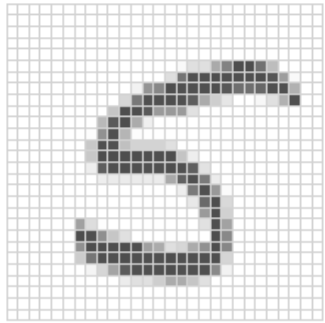
- chỉ trong năm 2013
- chỉ quan sát tại New York

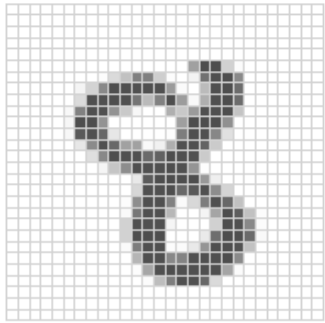
Quần thể vs Dữ liệu

Ví dụ dữ liệu về chữ số viết tay





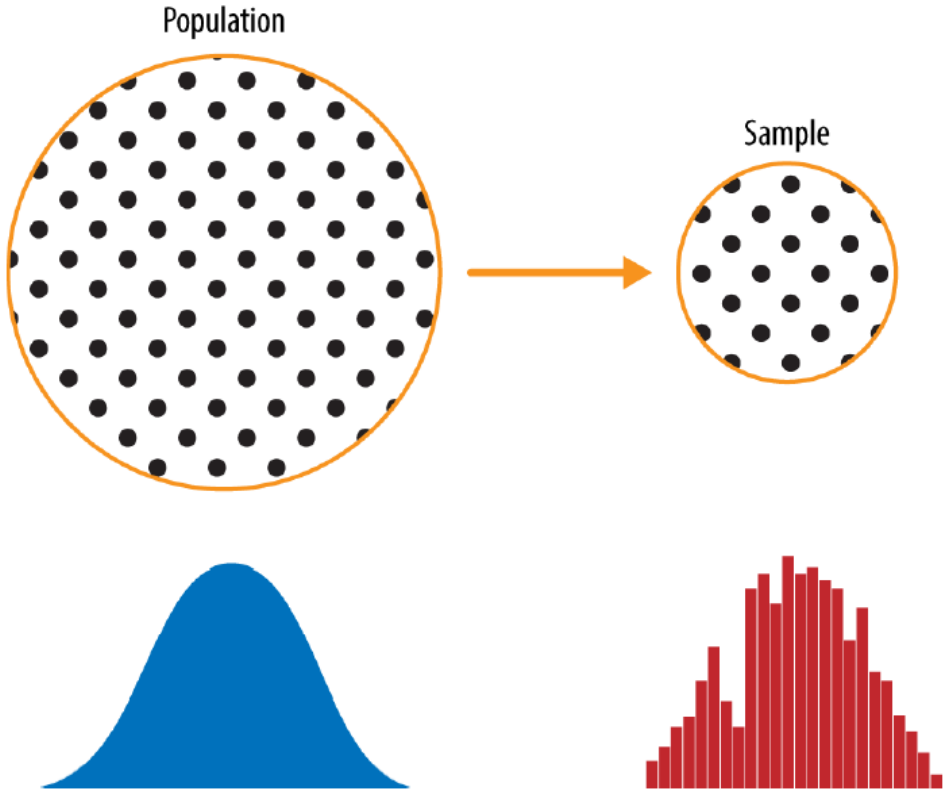




- giới hạn số người thu thập;
- giới hạn khu vực lấy mẫu.

Quần thể vs Dữ liệu

Một mẫu (dữ liệu) là một tập hợp các đơn vị nhỏ hơn (nhưng hy vọng mang tính đại diện) từ một quần thể được sử dụng để xác định sự thật về quần thể đó.



Quần thể vs Dữ liệu

Trong thực tế,

- các tính chất của một biến trong quần thể thường được giả định

  ví dụ, tuân theo một phân phối chưa biết (có thể là phân phối chuẩn);
- tuy nhiên, cái ta có trong tay là dữ liệu và phân phối thực nghiệm.

Để có được mẫu từ một quần thể ta cần một **quá trình lấy mẫu** (sampling procedure).

- Thống kê truyền thống tập trung rất nhiều vào quần thể, sử dụng lý thuyết dựa trên những giả định chắc chắn về quần thể.
- Thống kê hiện đại tập trung vào dữ liệu, nơi mà những giả định về phân phối là không cần thiết.

Lấy mẫu ngẫu nhiên

Các thuật ngữ chính

Mẫu (Sample) một tập con từ một bộ dữ liệu lớn;

Quần thể (Population) một bộ dữ liệu lớn hoặc một dữ liệu lý thuyết;

N (n) Kích cỡ của quần thể (mẫu);

Lấy mẫu ngẫu nhiên (Random sampling) Lấy phần tử một cách ngẫu nhiên từ quần thể;

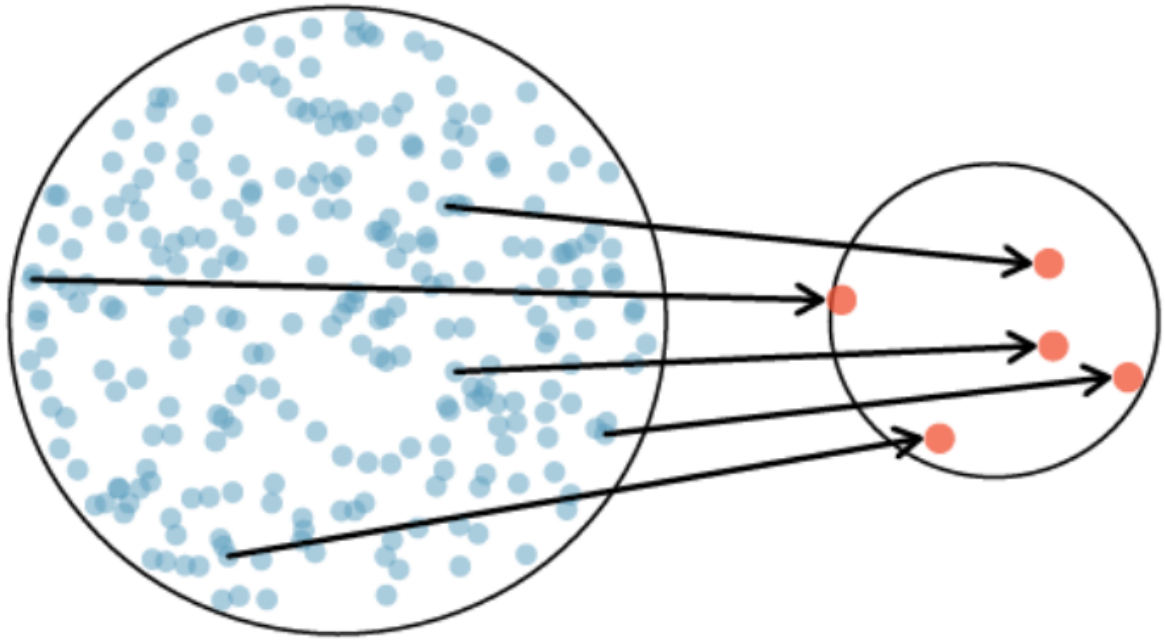
Lấy mẫu phân tầng (Stratified sampling) Phân chia quần thể thành các tấng khác nhau, và thực hiện lấy mẫu ngẫu nhiên ở mỗi tầng;

Mẫu ngẫu nhiên đơn giản (Simple random sample) là một mẫu kết quả của quá trình lấy mẫu ngẫu nhiên đơn giản (không phân chia tầng trong quần thể);

Mẫu thiên vị (Sample bias) là một mẫu trình bày sai về quần thể.

Lấy mẫu ngẫu nhiên

Lấy mẫu ngẫu nhiên (Random sampling)



Ví dụ, ta chọn ngẫu nhiên 20 chuyến bay bất kỳ trong dữ liệu `flights` (gồm 336,776 chuyến bay).

Lấy mẫu ngẫu nhiên

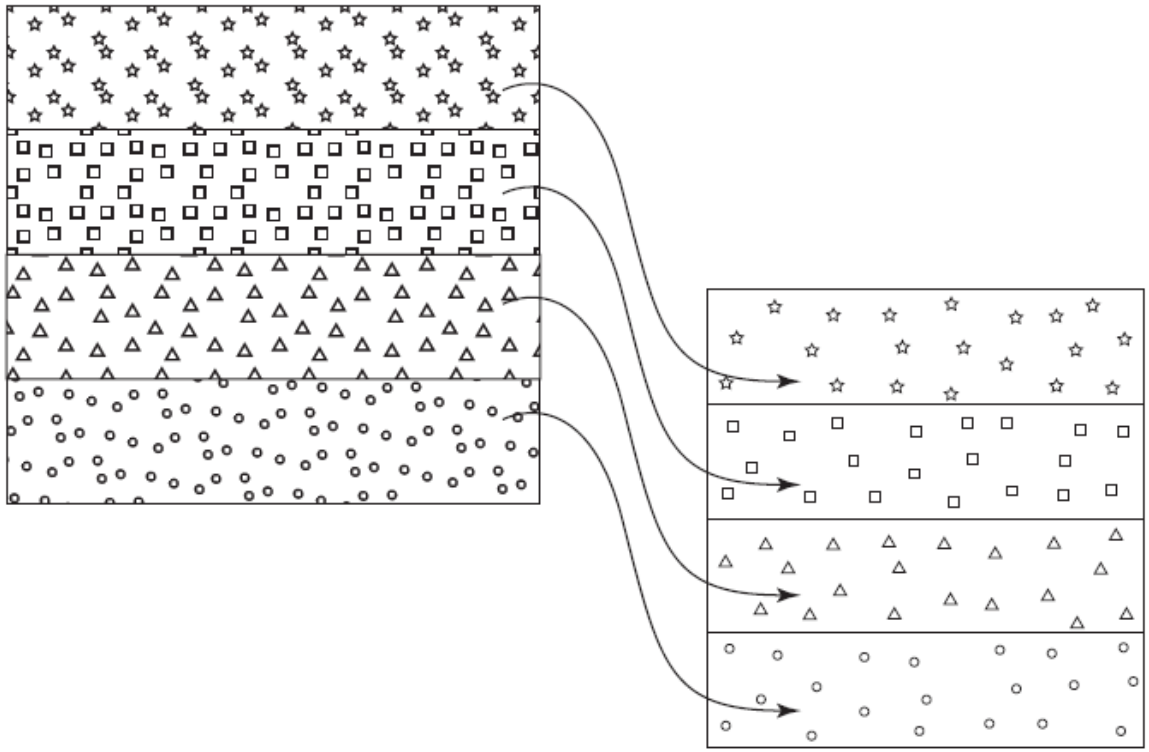
Có hai dạng lấy mẫu ngẫu nhiên:

lấy có hoàn lại (with replacement) tức là các quan sát sẽ được trả lại quần thể sau khi lấy mẫu, và do đó, có thể được lấy mẫu trong tương lai;

lấy không hoàn lại (without replacement) tức là các quan sát chỉ được lấy mẫu một lần duy nhất, và không được trả lại quần thể.

Lấy mẫu ngẫu nhiên

Lấy mẫu phân tầng (Stratified sampling)



Ví dụ, ta chọn ngẫu nhiên 20 chuyến bay bất kỳ trong mỗi tháng (từ 1 tới 12) của dữ liệu `flights` (gồm 336,776 chuyến bay).

Chất lượng dữ liệu và số lượng dữ liệu

Chất lượng dữ liệu - Data quality

Chất lượng dữ liệu thường quan trọng hơn số lượng dữ liệu khi đưa ra ước tính hoặc mô hình dựa trên mẫu.

Chất lượng dữ liệu trong khoa học dữ liệu liên quan đến:

- tính đầy đủ;
- nhất quán về định dạng;
- độ rõ ràng;
- độ chính xác của từng điểm dữ liệu.

$\hookrightarrow$ tính đại diện của dữ liệu cho quần thể.

Chất lượng dữ liệu và số lượng dữ liệu

Số lượng dữ liệu

Số lượng lớn dữ liệu yêu cầu:

- thời gian;
- chi phí;
- công sức.

Ví dụ, theo dõi các dữ liệu cực đoan hoặc bị khuyết:

- chiếm 10% của 1,000,000 quan sát;
- chiếm 10% của 1,000 quan sát.

$\hookrightarrow$ trong một số trường hợp, số lượng lớn dữ liệu không thực tế và kém hiệu quả.

Nhưng trong một vài nhiệm vụ cụ thể, số lượng lớn dữ liệu lại hữu ích trong việc nâng cao độ chính xác.

Ví dụ, truy vấn từ/cụm từ trên Google, những từ/cụm từ "hot" càng cho ra kết quả chính xác.

Sự lựa chọn thiên vị

Sự lựa chọn thiên vị (Selection bias) đề cập đến việc lựa chọn dữ liệu có chọn lọc - có ý thức hoặc vô thức - theo cách dẫn đến một kết luận sai lệch hoặc không vững chắc.

Cụ thể:

- lấy mẫu không ngẫu nhiên - nonrandom sampling;
- cherry-picking data;
- chọn khoảng thời gian làm nổi bật các phân tích thống kê;
- dừng thử nghiệm khi kết quả trông như "hấp dẫn".

- **Săn lùng dữ liệu - Data snooping:** Săn lùng rộng rãi thông qua dữ liệu để tìm kiếm điều gì đó thú vị.
- **Vast search effect:** liên tục chạy các mô hình khác nhau và đặt các câu hỏi khác nhau với một tập dữ liệu lớn, chắc chắn ta sẽ tìm thấy điều gì đó thú vị. Tuy nhiên, kết quả tìm thấy có thực sự thú vị hay chỉ là một kết quả không thực?

Phân phối mẫu của một thống kê

Thống kê - a Statistic

Một thống kê - a statistic, là một tên gọi ám chỉ tới các đại lượng được tính toán từ một dữ liệu, ví dụ:

- trung bình mẫu $\overline{Y}$;
- độ lệch chuẩn mẫu S;
- median;
- thống kê $Z = \dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$;
- thống kê $T = \dfrac{\overline{Y} - \mu}{S/\sqrt{n}}$.

Phân phối mẫu của một thống kê

Thống kê - a Statistic

Một thống kê - a statistic, là một tên gọi ám chỉ tới các đại lượng được tính toán từ một dữ liệu, ví dụ:

- trung bình mẫu $\overline{Y}$;
- độ lệch chuẩn mẫu S;
- median;
- thống kê $Z = \dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$;
- thống kê $T = \dfrac{\overline{Y} - \mu}{S/\sqrt{n}}$.

Bởi vì dữ liệu là ngẫu nhiên, nên các một thống kê cũng là đại lượng ngẫu nhiên.

$\hookrightarrow$ có các đặc trưng của một đại lượng ngẫu nhiên:

- trung bình;
- phương sai, độ lệch chuẩn;
- phân phối.

Phân phối mẫu của một thống kê

Phân phối mẫu

Phân phối mẫu là thuật ngữ nhắm tới phân phối của một thống kê, khi ta thực hiện lấy mẫu nhiều lần từ quần thể.

Độ chệch - Bias

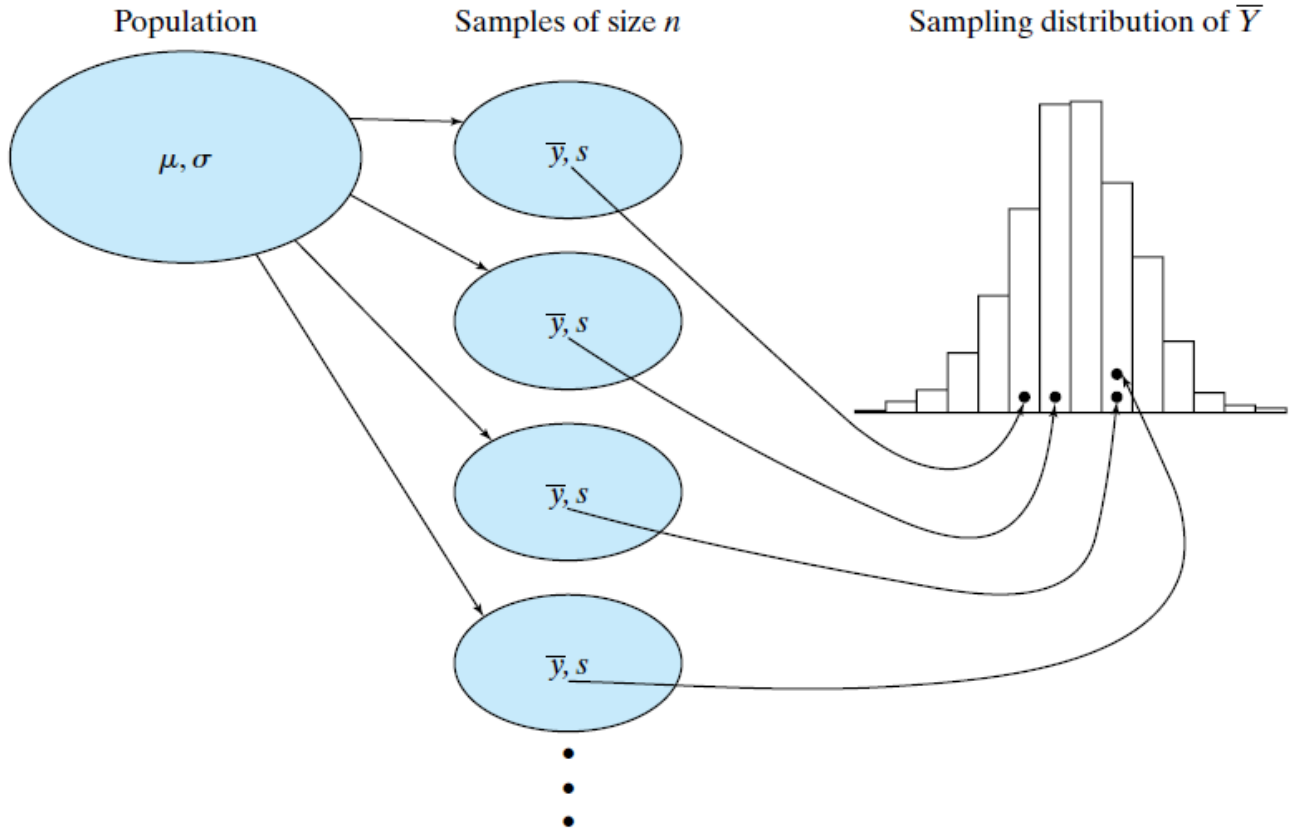
Độ chệch (**Bias**) là thuật ngữ nhắm tới sự chênh lệch của trung bình của một thống kê so với giá trị "chính xác" (dựa trên quần thể/dữ liệu lớn), khi ta thực hiện lấy mẫu nhiều lần từ quần thể.

Sai số chuẩn - Standard error

Sai số chuẩn (**Standard error**) ám chỉ tới độ lệch chuẩn của một thống kê (ước lượng), đo sự biến động của ước lượng ở trên nhiều mẫu.

Phân phối mẫu của một thống kê

Sơ đồ dưới đây minh họa phân phối mẫu của trung bình mẫu $\overline{Y}$.



Ví dụ

Xét dữ liệu về thu nhập theo năm của 50,000 hồ sơ vay tiền gửi tới Lending Club.

Từ dữ liệu lớn này, ta thực hiện như sau:

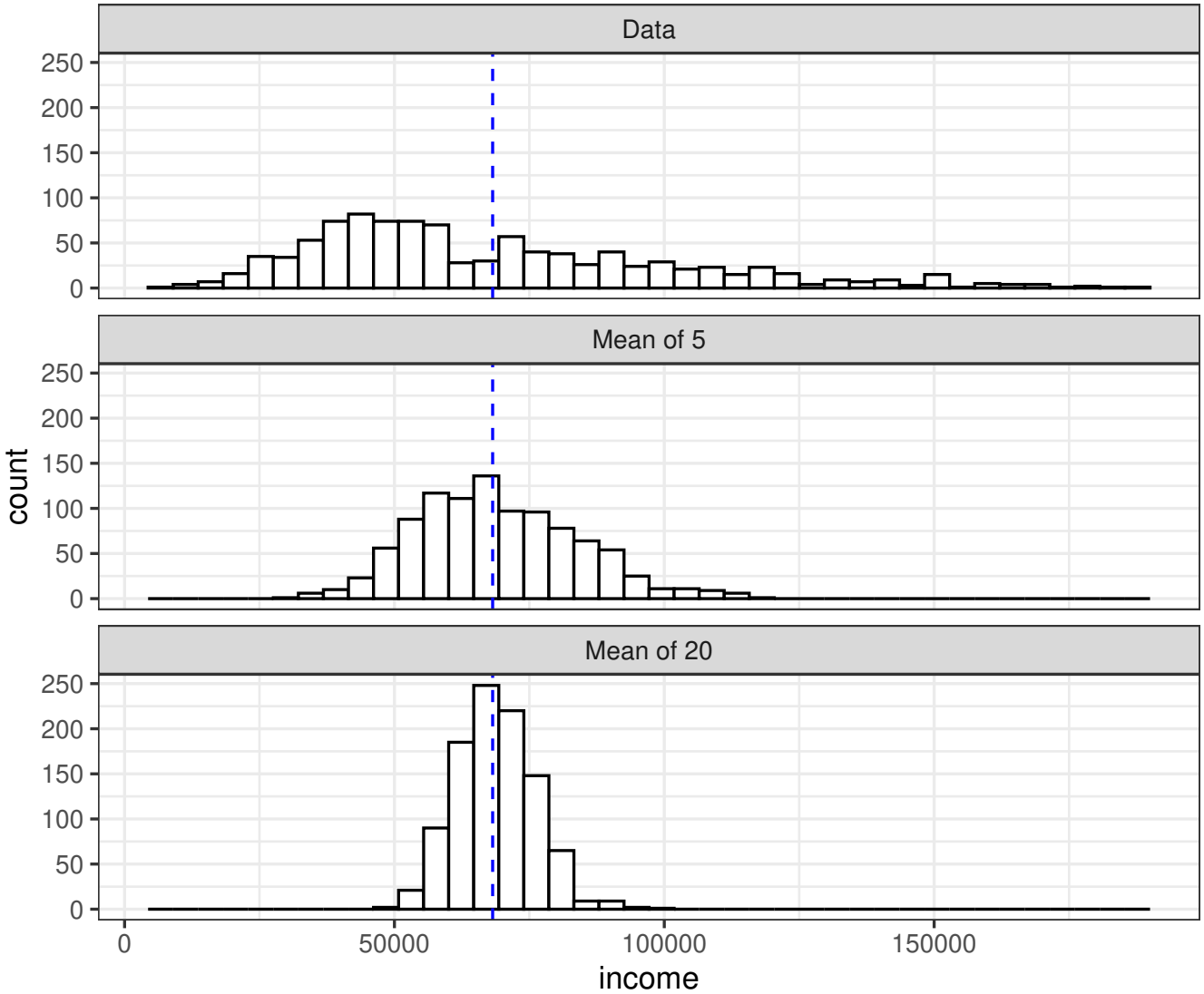
- Chọn ngẫu nhiên 1000 quan sát, coi như một quần thể.
- Từ 1000 quan sát, chọn ngẫu nhiên 1000 mẫu, có cỡ 5, và tính trung bình thu nhập của mỗi mẫu.

  $\hookrightarrow$ ta có một tập hợp 1000 trung bình mẫu của cỡ mẫu 5.
- Từ 1000 quan sát, chọn ngẫu nhiên 1000 mẫu, có cỡ 20, và tính trung bình thu nhập của mỗi mẫu.

  $\hookrightarrow$ ta có một tập hợp 1000 trung bình mẫu của cỡ mẫu 20.

Biểu đồ histogram dưới đây biểu diễn phân phối tương ứng của các dữ liệu, cùng với trung bình mẫu của mẫu gồm 1000 quan sát (đường nét đứt màu xanh).

Ví dụ


Data
Mean of 5
Mean of 20
count
income
0
50
100
150
200
250
0
50000
100000
150000

Ví dụ

Ta có bảng tổng hợp sau

	Trung bình	Độ chệch	SE
Mẫu 1000	68199.66	–	–
Trung bình của mẫu 5	69154.31	954.65	14979.52
Trung bình của mẫu 20	68751.65	551.99	7245.36

trong đó,

- độ chệch bằng trung bình của mẫu gồm 1000 trung bình trừ đi trung bình của mẫu 1000 quan sát;
- SE là sai số chuẩn (standard error) bằng độ lệch chuẩn của mẫu gồm 1000 trung bình.

Định lý giới hạn trung tâm - Central Limit Theorem

Định lý giới hạn trung tâm - Central Limit Theorem là:

- một trong số các kết quả quan trọng trong lý thuyết xác suất-thống kê;
- xương sống cho các suy luận thống kê (khoảng tin cậy, kiểm định thống kê).

Central Limit Theorem
Cho dãy mẫu ngẫu nhiên $Y_1, Y_2, \ldots, Y_n$ độc lập cùng phân phối xác suất, sao cho $\mathbb{E}(Y_i) = \mu$ và $\mathbb{V}\mathrm{ar}(Y_i) = \sigma^2$ với mọi $i = 1, 2, \ldots, n$. Đặt $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ (trung bình mẫu). Khi đó:

$$\overline{Y} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

khi n đủ lớn.

Tức là trung bình mẫu $\overline{Y}$ sẽ xấp xỉ phân phối chuẩn, với trung bình μ, phương sai $\dfrac{\sigma^2}{n}$, khi cỡ mẫu n đủ lớn.

Sai số chuẩn

Theo kết quả của định lý Giới Hạn Trung Tâm, ta có sai số chuẩn cho trung bình mẫu $\overline{Y}$ là

$$\mathrm{SE} = \frac{s}{\sqrt{n}},$$

với

- s là độ lệch chuẩn mẫu;
- n là cỡ mẫu.

Khi cỡ mẫu n tăng lên b lần thì sai số chuẩn giảm xuống với tỷ lệ là $\sqrt{b}$.

Mặt khác, trong ví dụ trên:

- với cỡ mẫu 5, $\mathrm{SE} = 14979.52$;
- với cỡ mẫu 20, $\mathrm{SE} = 7245.36$.

$\hookrightarrow$ cỡ mẫu tăng 4 lần, thì SE giảm xấp xỉ 2 lần.

Nhắc lại, trong ví dụ trên, ta không tính SE theo công thức dựa trên định lý Giới Hạn Trung Tâm.

Sai số chuẩn

Sai số chuẩn SE ở ví dụ trên được tính theo cách thức sau:

1. thu thập một loạt M bộ dữ liệu khác nhau, cùng cỡ mẫu n;
2. với mỗi mẫu, tính giá trị của thống kê (chẳng hạn, trung bình);
3. tính độ lệch chuẩn của tất cả các giá trị thống kê được tính trong bước 2, và sử dụng kết quả thu được như là ước lượng của sai số chuẩn.

Trong thực tế, việc áp dụng phương thức này không thực tế, do phải lấy nhiều mẫu.

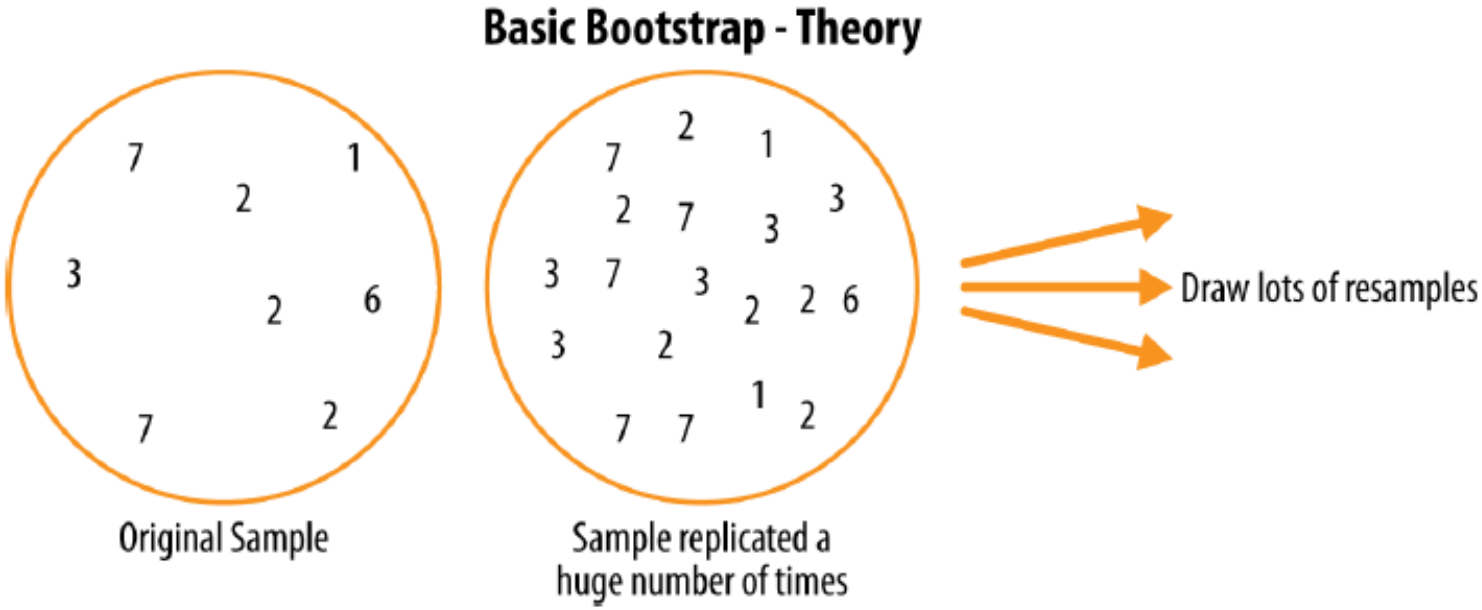
Tuy nhiên, ta có một phương pháp khác hoạt động với cách thức tương tự, có tên gọi là phương pháp bootstrap (bootstrap method).

Phương pháp bootstrap

Phương pháp bootstrap

Phương pháp bootstrap là một phương pháp hiệu quả để ước lượng phân phối mẫu (sampling distribution) của một thống kê, hoặc một mô hình.

Phương pháp bootstrap hoạt động trên nguyên lý lấy lại mẫu (resampling) dựa trên dữ liệu gốc (original data).



Phương pháp bootstrap

Cụ thể, thuật toán của phương pháp bootstrap với một mẫu cỡ n như sau:

1. Tạo một mẫu ngẫu nhiên cỡ n từ dữ liệu gốc, có lặp lại (with replacement);
2. Tính giá trị của thống kê (ví dụ, trung bình) với mẫu vừa tạo;
3. Lặp lại bước 1 và 2 trong R lần (ít nhất 500 lần), và lưu kết quả lại.

Ta có thể sử dụng R giá trị của thống kê để:

- xác định độ lệch chuẩn của chúng $\Longrightarrow$ sai số chuẩn;
- biểu diễn histogram để mô tả phân phối mẫu của thống kê;
- xác định khoảng tin cậy cho thống kê.

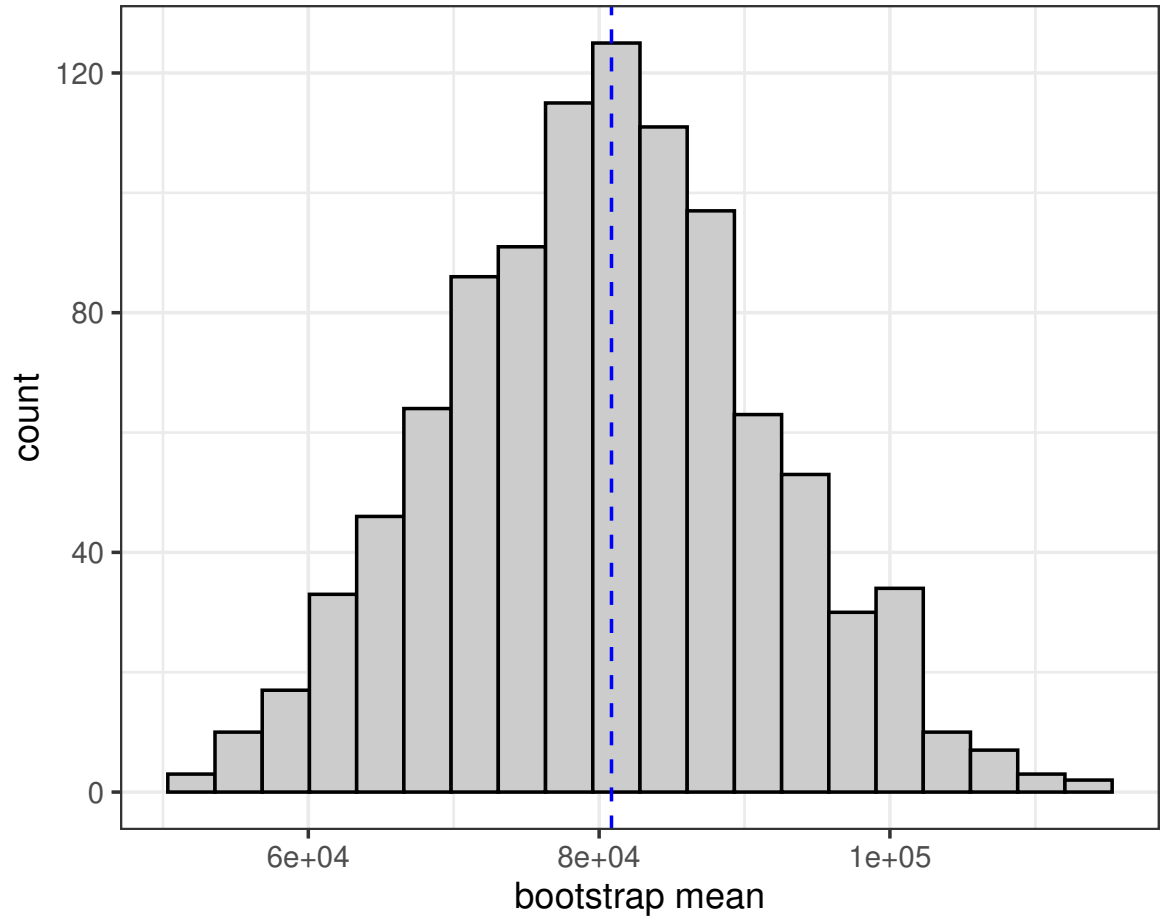
Phương pháp bootstrap

Ví dụ, ta có một bộ dữ liệu gồm thu nhập hàng năm tương ứng 10 người.

data	lần 1	lần 2	...	lần R
86000	66500	53000		100000
40000	86000	105000		66500
55000	55000	128000		140000
35000	55000	100000		55000
53000	100000	100000		86000
140000	86000	55000		35000
128000	66500	140000		140000
105000	128000	40000		105000
66500	66500	100000		140000
100000	53000	100000		86000
$\bar{y} = 80850$	$\bar{y}_1 = 76250$	$\bar{y}_2 = 92100$		$\bar{y}_R = 95350$

Phương pháp bootstrap

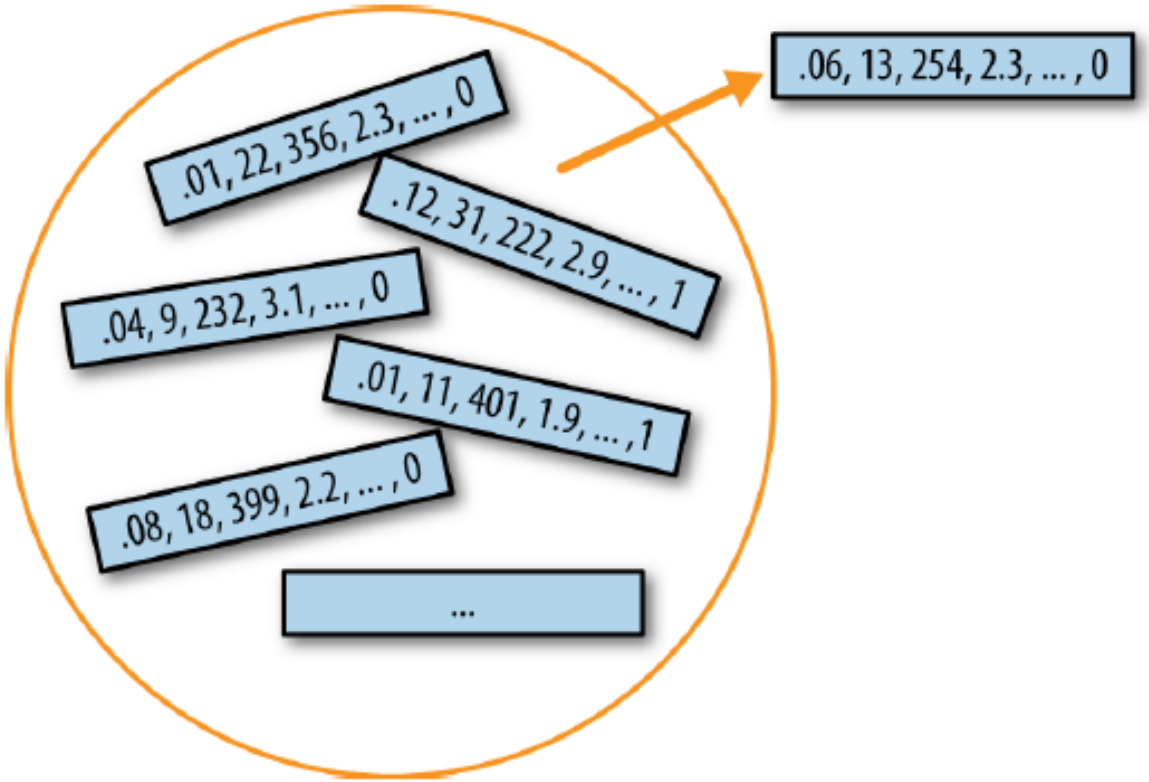
Ta lặp lại quá trình bootstrap 1000 lần.



- độ chệch: -421.35
- sai số chuẩn: 11000.69

Phương pháp bootstrap

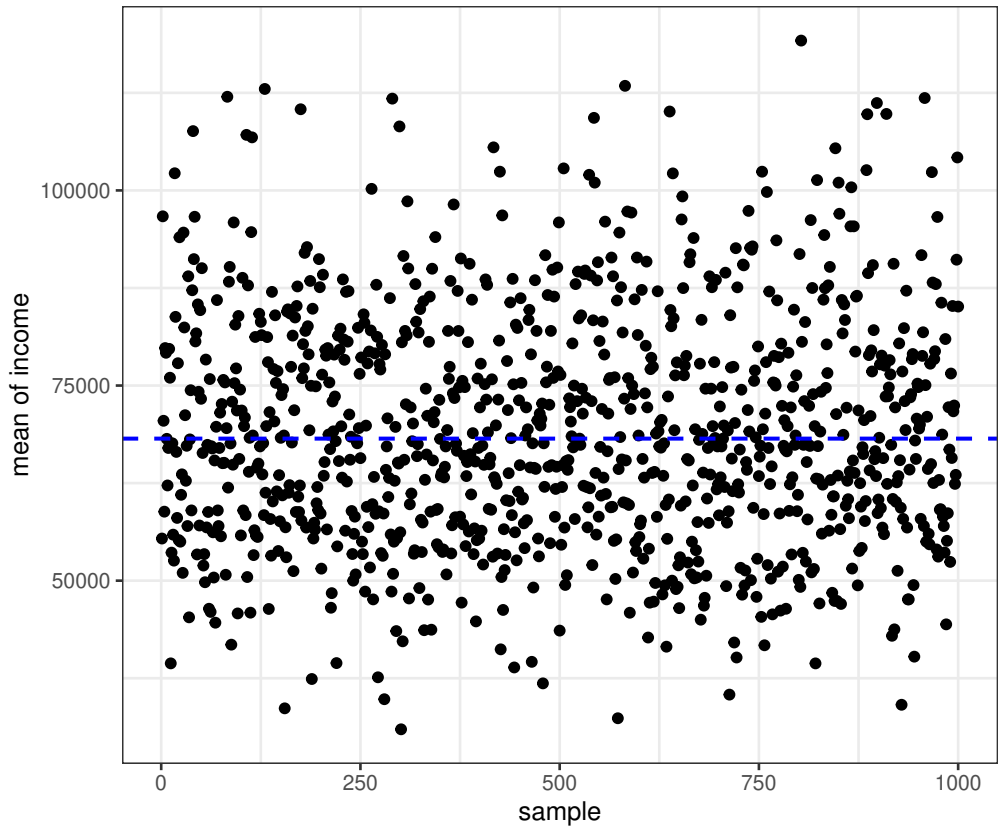
Phương pháp bootstrap còn có thể sử dụng cho dữ liệu nhiều chiều (multivariate data), trong đó, với mỗi lần tạo mẫu, ta sẽ lấy ngẫu nhiên 1 dòng của dữ liệu.



Khoảng tin cậy

Ta xét lại ví dụ ở phần trên:

- dữ liệu lớn gồm 1000 ghi chép thu nhập hàng năm của 1000 người;
- ước lượng trung bình thu nhập trong 1000 mẫu có cỡ 5, được lấy ngẫu nhiên từ dữ liệu lớn (các điểm đen trên hình);
- tính trung bình thu nhập trong dữ liệu lớn (đường nét đứt màu xanh).



Khoảng tin cậy

Khoảng tin cậy

Khoảng tin cậy (**confidence interval**) là kết quả của một quá trình ước lượng khoảng (interval estimation), và cung cấp một khoảng hợp lý mà giá trị "thật sự" của thống kê có thể thuộc vào.

Ví dụ: khoảng tin cậy 95% cho trung bình thu nhập (của quần thể) là $(65000, 82300)$, ta có thể nói rằng "ta tin cậy 95% rằng trung bình thu nhập tổng thể μ nằm trong khoảng từ 65000 tới 82300".

Hơn nữa, ta có thể chọn một số bất kỳ trong khoảng tin cậy làm đại điện cho trung bình quần thể μ.

Khoảng tin cậy

Khoảng tin cậy

Khoảng tin cậy (**confidence interval**) là kết quả của một quá trình ước lượng khoảng (interval estimation), và cung cấp một khoảng hợp lý mà giá trị "thật sự" của thống kê có thể thuộc vào.

Ví dụ: khoảng tin cậy 95% cho trung bình thu nhập (của quần thể) là $(65000, 82300)$, ta có thể nói rằng "ta tin cậy 95% rằng trung bình thu nhập tổng thể μ nằm trong khoảng từ 65000 tới 82300".

Hơn nữa, ta có thể chọn một số bất kỳ trong khoảng tin cậy làm đại điện cho trung bình quần thể μ.

Đối với một nhà khoa học dữ liệu, khoảng tin cậy là một công cụ:

- để biết được kết quả mẫu có thể thay đổi như thế nào;
- để thông báo lỗi tiềm ẩn trong ước tính;
- để tìm hiểu xem liệu có cần một mẫu lớn hơn không.

Khoảng tin cậy cho trung bình

Trong lý thuyết thống kê, ta đã học được rằng, khoảng tin cậy cho trung bình có thể được xây dựng bởi các dạng sau:

- biết phân phối của quần thể là phân phối chuẩn, và biết độ lệch chuẩn σ:

$$\left(\overline{y} - z_{1-\alpha/2}\sigma/\sqrt{n}, \overline{y} + z_{1-\alpha/2}\sigma/\sqrt{n}\right),$$

- biết phân phối của quần thể là phân phối chuẩn, và không biết độ lệch chuẩn σ:

$$\left(\overline{y} - t_{1-\alpha/2,n-1}s/\sqrt{n}, \overline{y} + t_{1-\alpha/2,n-1}s/\sqrt{n}\right),$$

- không biết phân phối của quần thể, và không biết độ lệch chuẩn σ:

$$\left(\overline{y} - z_{1-\alpha/2}s/\sqrt{n}, \overline{y} + z_{1-\alpha/2}s/\sqrt{n}\right),$$

  khi n đủ lớn, đáy là kết quả của việc áp dụng định lý Giới Hạn Trung Tâm.

Khoảng tin cậy cho trung bình

Để áp dụng các cách tính trên, ta cần:

- thông tin của phân phối của quần thể; hoặc
- thông tin của độ lệch chuẩn σ; hoặc
- cỡ mẫu n lớn.

Tuy nhiên, ta chỉ có duy nhất là dữ liệu, và không có bất kỳ thông tin gì về

- phân phối;
- độ lệch chuẩn σ.

$\hookrightarrow$ ta không thể áp dụng các phương pháp trên hoặc áp dụng không hiệu quả.

Khoảng tin cậy bootstrap

Khoảng tin cậy bootstrap là một phương pháp ước lượng khoảng tin cậy dựa trên việc lấy mẫu lại của dữ liệu và không yêu cầu bất kỳ giả định nào về phân phối của quần thể.

Trong lý thuyết thống kê, khoảng tin cậy bootstrap còn được gọi là khoảng tin cậy phi tham số bởi vì nó không yêu cầu:

- phân phối của quần thể;
- độ lệch chuẩn của quần thể.

Khoảng tin cậy bootstrap

Khoảng tin cậy bootstrap percentile

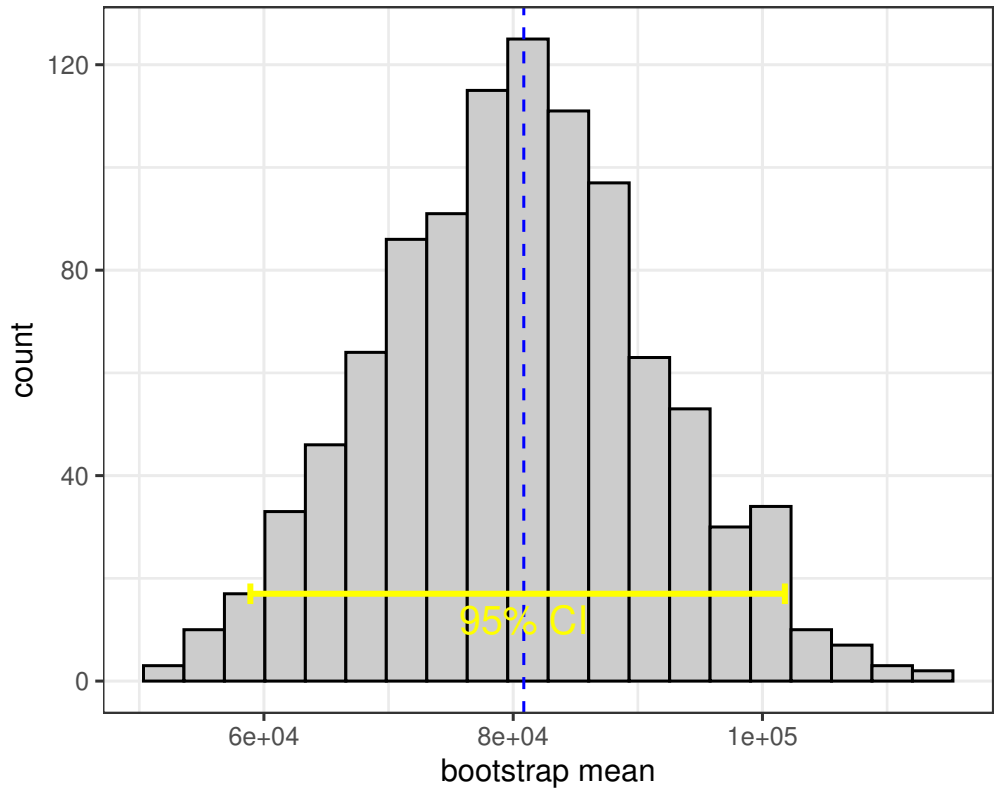
$(1-\alpha)\times 100\%$ Khoảng tin cậy bootstrap percentile được xác định như sau:

1. Tạo một mẫu ngẫu nhiên cỡ n (cùng với cỡ của dữ liệu gốc) từ dữ liệu gốc, có lặp lại (with replacement);
2. Tính giá trị của thống kê (ví dụ, trung bình) với mẫu vừa tạo;
3. Lặp lại bước 1 và 2 trong R lần (ít nhất 1000 lần), và lưu kết quả lại.
4. Sắp xếp tăng dần R giá trị của thống kê.
5. Xác định giá trị ở vị trí thứ $\lfloor R\times\alpha/2\rfloor$ và ở vị trí thứ $\lfloor R\times(1-\alpha/2)\rfloor$ làm hai giá trị giới hạn của khoảng tin cậy.

Khoảng tin cậy bootstrap

Ví dụ, ta xét lại mẫu gồm 10 ghi chép thu nhập hàng năm của 10 người:

- trung bình mẫu: 80850;
- thực hiện lặp bootstrap 1000 lần;
- 95% khoảng tin cậy bootstrap percentile là: (58908.89, 101797.46)



Khoảng tin cậy bootstrap

Một số nhận xét:

- Khi ta tăng cỡ mẫu n, thì khoảng tin cậy sẽ ngắn lại.

  95% khoảng tin cậy với một mẫu cỡ 10 là: (58908.89, 101797.46)

  95% khoảng tin cậy với một mẫu cỡ 20 là: (52754.72, 79122.79)
- Khi ta tăng độ tin cậy, thì khoảng tin cậy sẽ rộng hơn. (Khoảng tin cậy 99% sẽ rộng hơn 95%).
- Đối với khoảng tin cậy có độ tin cậy cao, ta cần lặp lại với số lần R lớn. Chẳng hạn: đối với 95% khoảng tin cậy, $R \geq 1000$; đối với 99%, $R \geq 5000$.