Bài giảng 15: Phương pháp hoán vị

Nguyễn Văn Tuấn
Garvan Institute of Medical Research, Australia
Đại học Tôn Đức Thắng, Việt Nam

Nội dung

- Ý tưởng về hoán vị
- Kiểm định hoán vị (permutation test)
- Phương pháp bootstrap

Ý tưởng về hoán vị

Ôn bài một chút ...

- Hoán vị = permutation
- Có 3 người và 3 cái ghế, bao nhiêu cách sắp xếp?
  - Ghế 1: có 3 cách
  - Ghế 2: có 2 cách (sau khi đã sắp xếp 1 người cho ghế 1)
  - Ghế 3: chỉ còn có 1 cách (sau khi đã sắp xếp 2 người)
- Trả lời **3 x 2 x 1 = 6 (viết tắt 3!)**

đọc 3 factorial

Mở rộng khái niệm hoán vị

- Nhóm 1 có *n* đối tượng
- Nhóm 2 có *m* đối tượng
- Hoán vị cho 2 nhóm là *n.m*

Khái niệm lấy mẫu (sample)

- Sample (động từ) lấy mẫu
- Hai loại lấy mẫu:
  - sampling **with** replacement – lấy mẫu có hoàn lại
  - sampling **without** replacement – lấy mẫu không hoàn lại

Lệnh trong R: **`sample(x, n, replace=T/F)`**

Ví dụ lấy mẫu trong R

```
x = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
```

lấy 5 giá trị có hoàn lại

```
s1 = sample(x, 5, replace=T)
s1 = sample(x, 5, T)
```

lấy 5 giá trị không hoàn lại

```
s1 = sample(x, 5, F)
```

Kiểm định hoán vị

Permutation test

- Có khi còn gọi là *randomization test, re-randomization test, exact test*
- Ý tưởng của R. A. Fisher và E. J. Pitman (thập niên 1930)

Ý tưởng về hoán vị

- Nhóm Rx (n=21): 24, 61, 59, 46, 43, 53, 43, 44, 52, 43, 57, 49, 58, 67, 62, 57, 56, 33, 71, 49, 54
- Nhóm chứng (n=23): 42, 33, 46, 37, 62, 20, 43, 41, 10, 42, 53, 48, 55, 19, 17, 55, 37, 85, 26, 54, 60, 28, 42

Ý tưởng về hoán vị

1. Nhập 21+23 = 44 thành một mẫu
2. Chọn ngẫu nhiên (không thay thế) 21 trong số 44 và xem *như nhóm Rx*; tính trung bình x_1
3. Chọn ngẫu nhiên (không thay thế) 23 trong số 44 và xem *như nhóm chứng*; tính trung bình x_2
4. Tính hiệu số d = $x_1 - x_2$
5. Lặp lại bước 2-4 nhiều lần, có một tập hợp *d*
6. Thẩm định phân bố của *d*

Package "coin"

`coin` package

```
oneway_test(x ~ as.factor(group),
distribution=approximate(B=1000))
```

Sắp xếp dữ liệu theo yêu cầu của *coin*

```
x     group
24    Rx
61    Rx
...
42    Placebo
33    Placebo
...
```

Package "coin"

```
rx = c(24, 61, 59, 46, 43, 53, 43, 44, 52, 43,
57, 49, 58, 67, 62, 57, 56, 33, 71, 49, 54)

placebo = c(42, 33, 46, 37, 62, 20, 43, 41, 10,
42, 53, 48, 55, 19, 17, 55, 37, 85, 26, 54, 60,
28, 42)

x = c(rx, placebo)

group = c(rep(1,21), rep(2,23))

library(coin)

oneway_test(x ~ as.factor(group),
distribution=approximate(B=1000))
```

Kết quả kiểm định hoán vị

```
> oneway_test(x ~ as.factor(group), distribution=approximate(B=1000))
```

```
Asymptotic 2-Sample Permutation Test

data:  x by as.factor(group) (1, 2)
Z = 2.1648, p-value = 0.0304
alternative hypothesis: true mu is not equal to 0
```

Tóm lược: phương pháp phi tham số

- Đối với những biến không theo luật phân phối chuẩn
- Wilcoxon's rank test

```
wilcox.test(g1, g2)
wilcox.test(x ~ group)
```

- Permutation test (*coin* package)

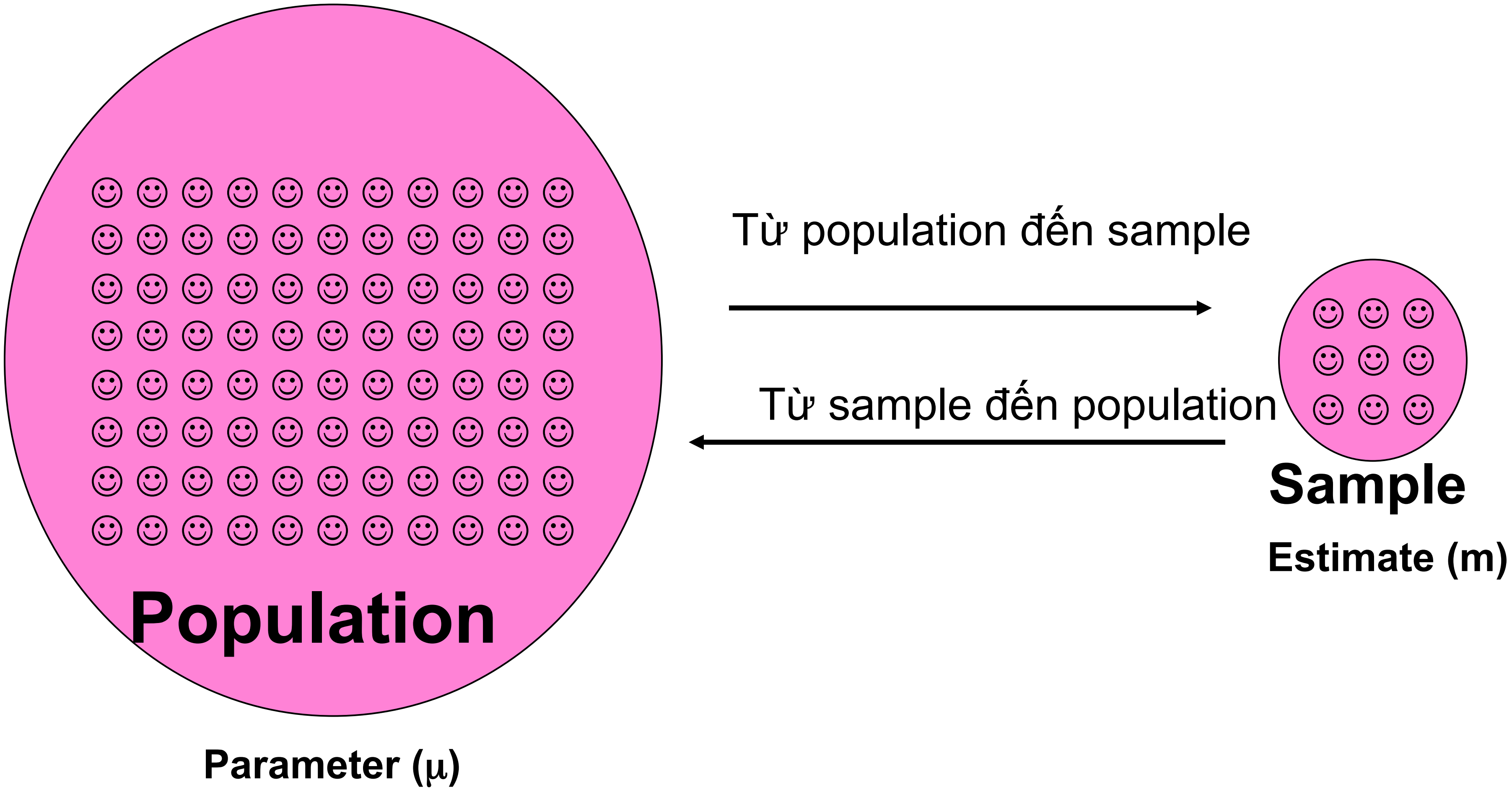
```
oneway_test(x ~ group)
```

Phương pháp bootstrap

Population và sample

- Chiều cao trung bình của người Việt ?
  - lấy mẫu ***ngẫu nhiên*** từ một quần thể
  - Tính chiều cao trung bình từ mẫu
  - Đánh giá mức độ bất định: tính khoảng tin cậy 95%
  - Phát biểu kết luận

Sample và Population



Population và sample: biểu diễn bằng R

- Giả dụ như chúng ta có một POPULATION chỉ 10 người

```
pop.ht <- c(156, 145, 189, 190, 176, 168, 158, 167, 150, 155)
mean(pop.ht)
[1] 165.4
```

- Bây giờ chúng ta lấy mẫu ngẫu nhiên 5 người từ population, tính trung bình

```
sample.ht <- sample(pop.ht, 5)
mean(sample.ht)
[1] 168.8
```

- Lấy mẫu ngẫu nhiên 5 người, một lần nữa, tính trung bình

```
sample.ht <- sample(pop.ht, 5); mean(sample.ht)
[1] 169.2
```

- và một lần nữa

```
sample.ht <- sample(pop.ht, 5); mean(sample.ht)
[1] 162.2
```

Dao động từ mẫu này sang mẫu khác

- Chú ý sự khác biệt về chiều cao trung bình giữa các mẫu
- Chúng ta không biết μ (chiều cao trung bình của population). Chúng ta chỉ có thể suy luận về μ
  - Những giá trị khả dĩ của μ?
- Giải đáp từ phương pháp cổ điển:
  - Lấy một mẫu ngẫu nhiên
  - Tính giá trị trung bình (m), standard deviation (sd), standard error (se)
  - Tính khoảng tin cậy 95% của **μ = m ± 1.96*se**

KTC 95% của một tham số (parameter)

```
pop.ht <- c(156, 145, 189, 190, 176, 168, 158, 167, 150, 155)
# take random sample of 5 and calculate mean, sd, se
random.sample <- sample(pop.ht, 5)
m <- mean(random.sample); std <- sd(random.sample)
[1] 172
[1] 16.68832

# calculate se and 95% CI
se <- std/sqrt(5)
lower.ci95 <- m-1.96*se
upper.ci95 <- m+1.96*se
lower.ci95; upper.ci95
157.37 - 186.63
```

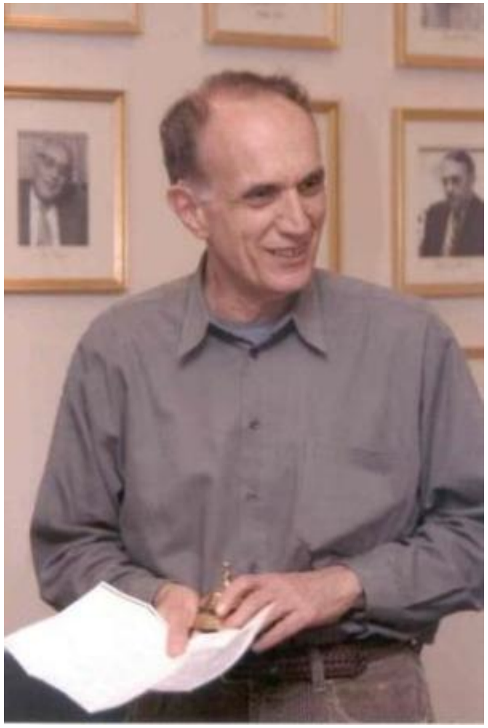
- **Kết luận sơ bộ:** 95% giá trị trung bình mẫu dao động từ 157.4 cm đến 186.6 cm

Vấn đề

- Cở mẫu nhỏ, ước số không ổn định
- Làm sao ước tính KTC95% cho các tham số như *median, variance, standard deviation, regression coefficients, proportion, ratio*, v.v.
- Phương pháp cổ điển không có
  - Không có công thức tính KTC95% của median hay tỉ số của 2 biến ngẫu nhiên
- Bootstrap solution!

Ý tưởng bootstrap

- Tác giả: Gs Bradley Efron (Stanford University), 1979.
- Mẫu gốc thể hiện population
- Lấy mẫu (có hoàn lại) nhiều lần từ mẫu gốc
- Phân bố của tham số từ nhiều mẫu có thể xác định



Sampling with replacement (lấy mẫu có hoàn lại)

- Sampling <u>without replacement</u> (lấy mẫu không hoàn lại): lấy mẫu ngẫu nhiên, nhưng không hoàn lại mẫu gốc.

- Sampling <u>with replacement</u> (lấy mẫu có hoàn lại): sau khi lấy mẫu, hoàn lại mẫu gốc, sau đó tiếp tục lấy mẫu.

Ví dụ lấy mẫu có hoàn lại

Mẫu gốc

3.12, 0.00, 1.57, 19.67, 0.22, 2.20

m = 4.46

Lấy mẫu (còn gọi là "**bootstrap sample**")

1.57	0.00	0.22
0.22	2.20	3.12
19.67	2.20	1.57
0.00	2.20	3.12
0.23	19.67	2.20
3.12	1.57	0.22
m=4.13	**m=4.64**	**m=1.74**

Trong R, chúng ta viết:

```
original.sample <- c(3.12, 0.00, 1.57, 19.67, 0.22, 2.20)
bs.sample <- sample(original.sample, replace=T)
bs.sample
```

Phương pháp bootstrap

- Bước 1: Bắt đầu với mẫu gốc : (x_1, x_2, x_3, ..., x_n);
- Bước 2: Lấy mẫu có hoàn lại → (x_1, x_1, x_2, x_4...) và tính chỉ số thống kê quan tâm, gọi là t;
- Lặp lại bước 2 khoảng B lần (B có thể là 10.000)

  (x_1, x_1, x_2, x_4...) → t_1

  (x_1, x_1, x_2, x_4...) → t_2

  (x_1, x_1, x_2, x_4...) → t_3

  ...

  (x_1, x_1, x_2, x_4...) → t_B
- Thu thập các trị số t
- Xem xét phân bố của t

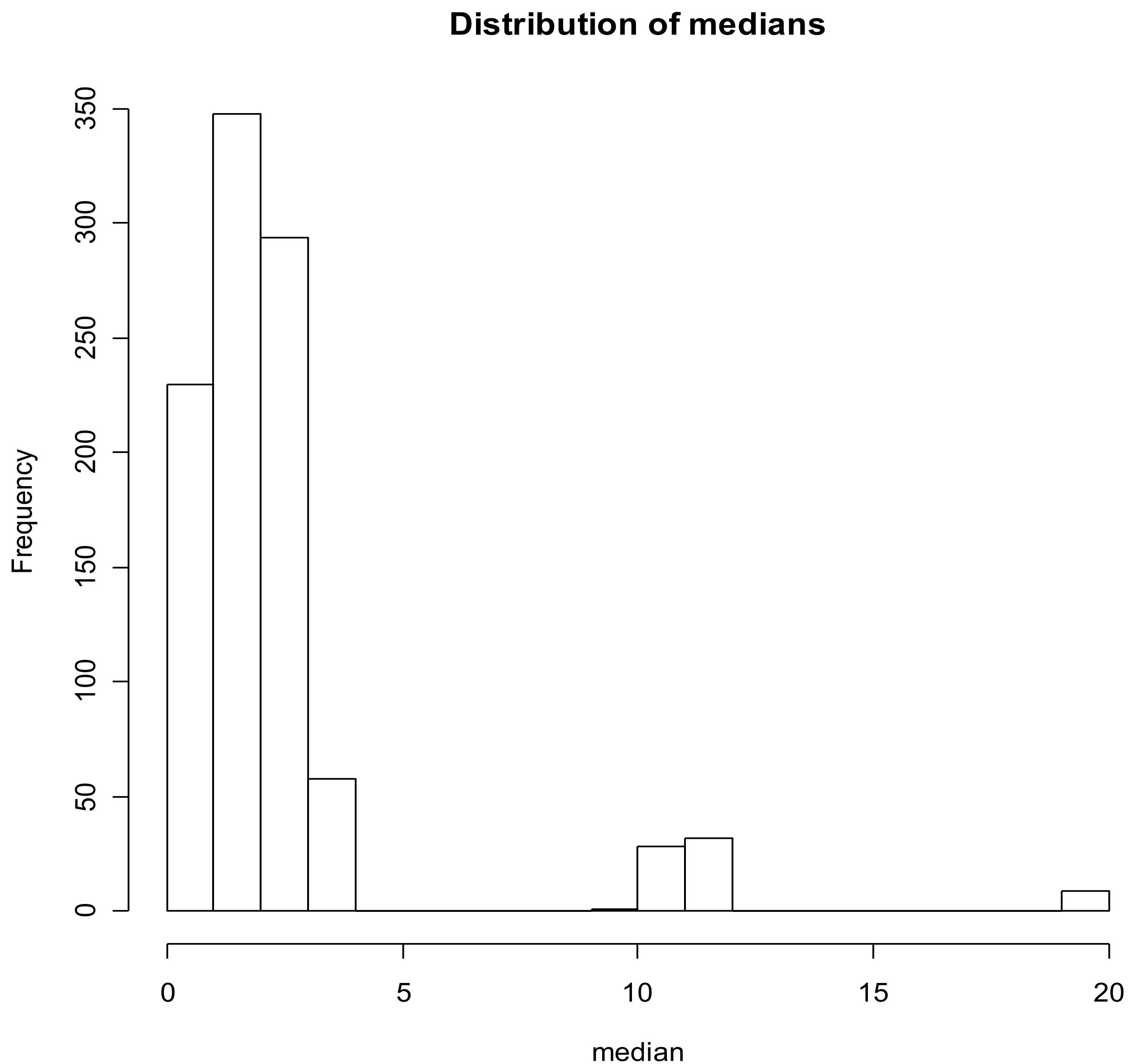
Vì dụ: tìm khoảng tin cậy 95% của median

```
# original sample
original.sample <- c(3.12, 0.00, 1.57, 19.67, 0.22, 2.20)
n = length(original.sample)
# number of bootstrap samples = 1000
B = 1000
# create an empty vector
median = numeric(B)
# start resampling and calculate the median in each
bootstrap sample
for (i in 1:B)
  {
    bs.sample <- sample(original.sample, n, replace=T)
    median[i] = median(bs.sample)
 }

# get a histogram of the medians
hist(median, breaks=20, main="Distribution of medians")
# get median and 95% CI
quantile(median, probs=c(0.025, 0.975))
```

Notes: the above programming can be done more efficiently

```
original.sample <- c(3.12, 0.00, 1.57, 19.67, 0.22, 2.20)
N = length(original.sample)
B = 1000
median <- c()
for (i in 1:B)
median <- c(median, median(sample(original.sample, N,
replace=T)))
quantile(median, c(0.025, 0.50, 0.975))
```




```
> quantile(median, probs=c(0.025, 0.50, 0.975))
   2.5%    50%  97.5%
 0.110  1.885 11.395
```

So sánh phương pháp cổ điển và bootstrap

- Thử xem xét dữ liệu:

```
original.sample <- c(3.12, 0.00, 1.57, 19.67, 0.22, 2.20)
```

- Ước tính bằng PP cổ điển:
  - Mean = 4.463, SD = 7.54, SE = 7.54/sqrt(6) = 3.08
  - 95% CI: -1.57 to 10.50

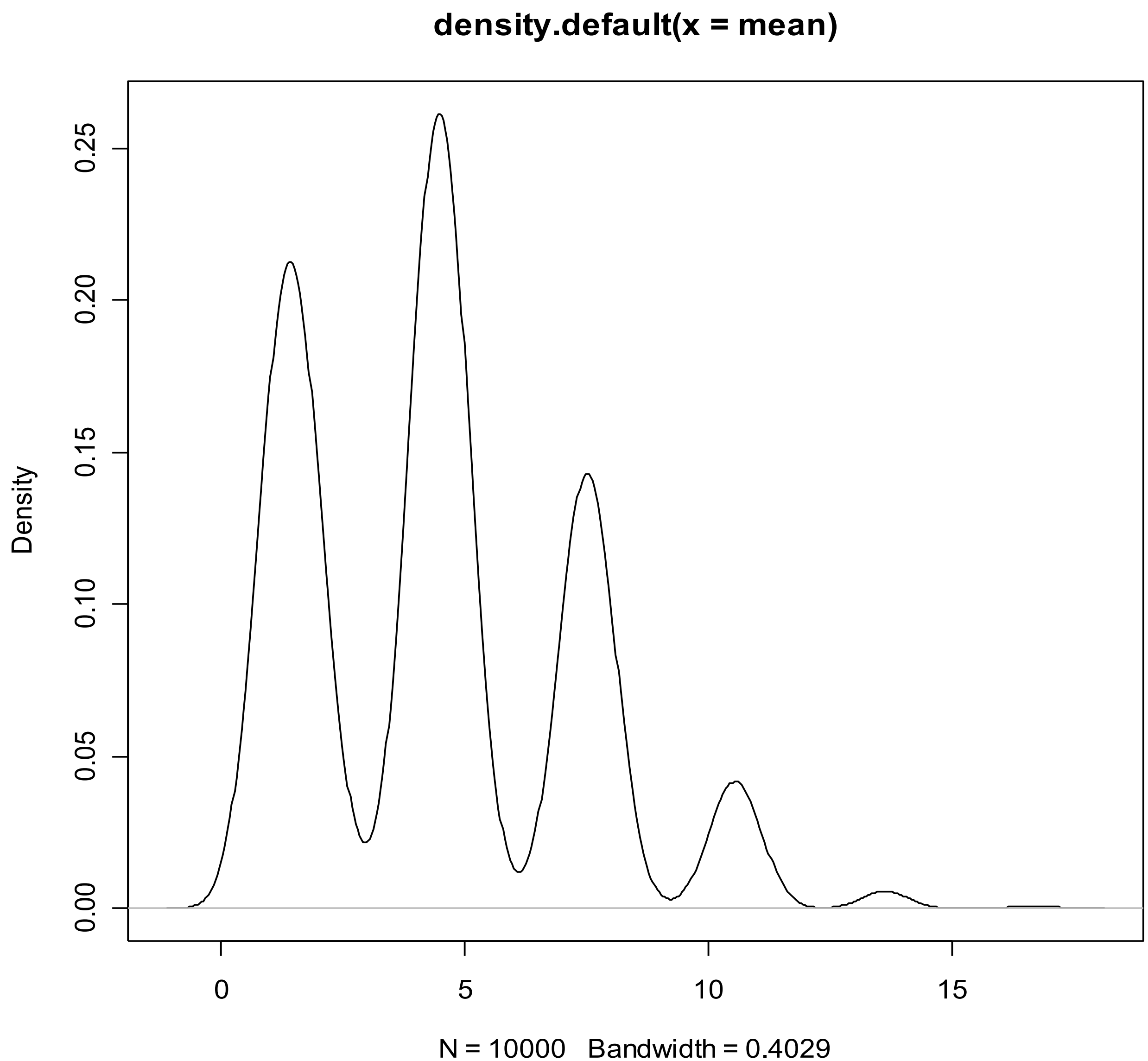
- Khoảng tin cậy này ... vô duyên!

So sánh phương pháp cổ điển và bootstrap

- Bootstrap estimates (based on 1000 resamples)

```
original.sample <- c(3.12, 0.00, 1.57, 19.67, 0.22, 2.20)
n <- length(original.sample)
B <- 1000
mean = numeric(B)
for (i in 1:B)
  {
    bs.sample <- sample(original.sample, n, replace=T)
    mean[i] = mean(bs.sample)
 }
quantile(mean, probs=c(0.025, 0.50, 0.975))
plot(density(mean))
```

```
> quantile(mean, probs=c(0.025, 0.50, 0.975))
    2.5%       50%     97.5%
 0.80575  4.31000 10.72583
```



density.default(x = mean)
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
N = 10000 Bandwidth = 0.4029

Giả định và những vấn đề

- Giả định
  - Các giá trị độc lập với nhau
  - Phương sai tương đương nhau
- PP Bootstrap không
  - có hiệu quả nếu tái chọn mẫu không tốt (eg paired vs two-sample t-test)
  - biến bad data thành good data
  - chỉnh sửa nghiên cứu tồi
- B là bao nhiêu?
  - 200 cho standard error
  - 2000 cho khoảng tin cậy

Phương pháp bootstrap (thay thế t test)

Ứng dụng phương pháp bootstrap

- PP Bootstrap có thể ứng dụng để kiểm định:
  – khác biệt 2 nhóm
  – hệ số tương quan
  – tỉ số 2 biến ngẫu nhiên
  – hồi qui tuyến tính
  – và rất nhiều vấn đề khác

So sánh hai nhóm: vấn đề

- Hai nhóm bệnh nhân dementia (điều trị và nhóm chứng)
- Outcome: daily activity score
- Câu hỏi: có sự khác biệt giữa 2 nhóm?
- Standard deviation (SD) lớn hơn trung bình

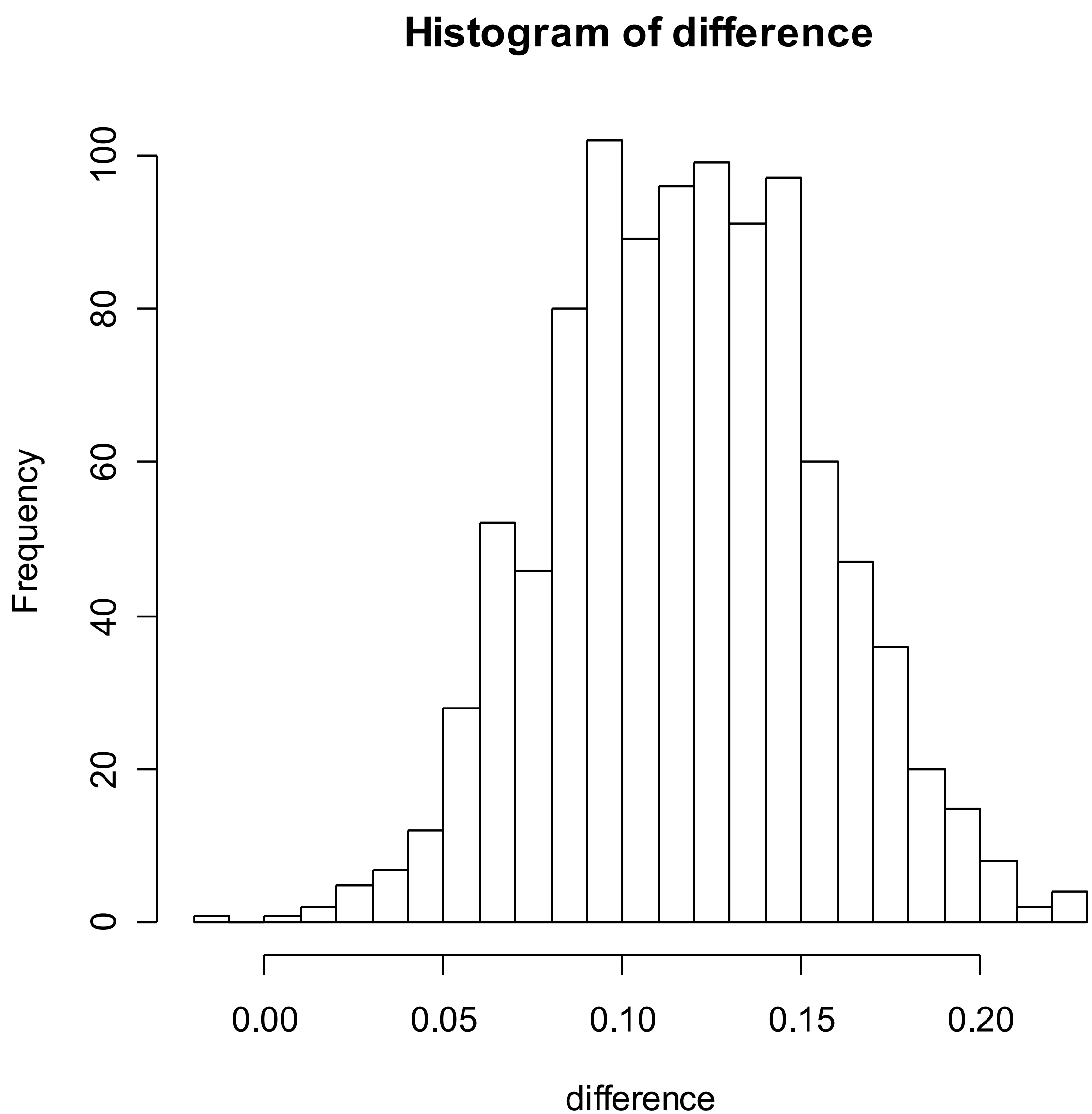
	Treated	Placebo
	0.05	0
	0.15	0.15
	0.35	0
	0.25	0.05
	0.20	0
	0.05	0
	0.10	0.05
	0.05	0.10
	0.30	
	0.05	
	0.25	
N	**11**	**8**
Mean	**0.164**	**0.044**
SD	**0.112**	**0.056**

Phân tích bằng PP Bootstrap

- Giải pháp khả dĩ: t-test trên dữ liệu hoán chuyển
- PP tốt hơn: bootstrap analysis
  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
  5. Xem xét phân bố

Dùng R

```
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)
n <- length(treated)
m <- length(control)
B = 1000
difference <- numeric(B)
no.effect = 0
for (i in 1:B) {
   bs.treated <- sample(treated, n, replace=T)
   bs.control <- sample(control, m, replace=T)
   difference[i] = mean(bs.treated) - mean(bs.control)
   if (difference[i] < 0) no.effect = no.effect+1
  }
hist(difference, breaks=20)
no.effect/1000
quantile(difference, probs=c(0.025, 0.50, 0.975))
      2.5%        50%      97.5%
0.04943182 0.11818182 0.19092330
```



Trong số 1000 mẫu, chỉ có 1 mẫu là không có khác biệt (difference < 0).
Trị số P = 1 / 1000 = 0.001

So sánh với phương pháp cổ điển

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

	Bootstrap results	Classical stats
Mean difference	0.118	0.12
95% CI	0.05 – 0.19	0.04 – 0.20