Bài giảng 13:
Phân tích mô tả biến liên tục

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Trọng lượng của 100 phụ nữ Việt Nam

45.0 38.0 57.0 49.5 57.0 42.0 54.0 54.0 34.0 45.0
42.0 49.0 27.0 57.0 50.0 61.0 49.5 48.0 54.0 49.0
48.5 58.0 52.0 40.0 49.0 59.0 50.0 50.0 47.0 50.0
47.0 45.0 48.0 49.0 50.0 53.0 48.5 53.0 43.0 49.0
53.0 52.0 49.0 50.0 50.0 43.0 60.0 55.0 57.0 58.0
48.0 52.0 50.0 45.0 56.0 53.0 57.0 47.0 51.0 60.0
30.0 35.0 50.0 60.0 42.0 40.0 62.0 54.5 38.0 41.0
47.0 38.5 59.0 69.0 59.0 44.0 66.0 57.0 44.0 50.0
42.0 45.0 44.0 55.0 59.0 53.0 42.0 53.0 47.0 47.0
58.0 55.0 41.0 49.5 49.0 38.0 58.0 58.0 39.5 42.8

Làm sao mô tả và biến những số liệu này thành information?

Mô tả *một* biến liên tục

- Mean (trung bình)
- Standard deviation (độ lệch chuẩn)
- Standard error (sai số chuẩn)
- Coefficient of variation (hệ số biến thiên)
- 95% Confidence Interval (khoảng tin cậy 95%)

Mean – trung bình

- **Nghiên cứu 1**: số liệu thu thập từ 6 người: 6, 7, 8, 4, 5, và 6. Trung bình là:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6+7+8+4+5+6}{6} = \frac{36}{6} = 6$$

- **Nghiên cứu 2**: số liệu của 4 người: 10, 2, 3, và 9. Trung bình:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{10+2+3+9}{4} = \frac{24}{4} = 6$$

Variation – dao động / biến thiên

- Số trung bình không phản ảnh đầy đủ phân bố của dữ liệu. Chúng ta cần chỉ số phản ảnh độ ***biến thiên***:
- Một chỉ số hiển nhiên là ***hiệu số*** từ số trung bình:
- Với nghiên cứu 1, (số liệu: 6, 7, 8, 4, 5, và 6), chúng ta có:

(6-**6**) + (7-**6**) + (8-**6**) + (4-**6**) + (5-**6**) + (6-**6**)

= 0 + 1 + 2 – 2 – 1 + 0

= 0

Không tốt!

Sum of squares – tổng bình phương

- Chúng ta cần phải bình phương hiệu số. Chỉ số này là "***Sum of squares***" (SS)
- Nghiên cứu 1: 6, 7, 8, 4, 5, 6, chúng ta có:

  $SS = (6-\mathbf{6})^2 + (7-\mathbf{6})^2 + (8-\mathbf{6})^2 + (4-\mathbf{6})^2 + (5-\mathbf{6})^2 + (6-\mathbf{6})^2 = 10$
- Nghiên cứu 2: 10, 2, 3, 9, chúng ta có:

  $SS= (10-\mathbf{6})^2 + (2-\mathbf{6})^2 + (3-\mathbf{6})^2 + (9-\mathbf{6})^2 = 50$

Chỉ số này tốt hơn!

Nhưng chưa xem xét đến cỡ mẫu *n*.

Variance – phương sai

- Chúng ta cần chia SS cho cỡ mẫu *n*. Nhưng mỗi lần tính SS, chúng ta mất 1 bậc tự do (degree of freedom). Do đó, mẫu số đúng là *n*-1. Chỉ số này gọi là **phương sai** - ***variance*** (kí hiệu bởi s^2)

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

- Hay viết gọn hơn:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Variance – ví dụ

- **Nghiên cứu 1: 6, 7, 8, 4, 5, và 6, phương sai là:**

$$s^2 = \frac{(6-6)^2 + (7-6)^2 + (8-6)^2 + (5-6)^2 + (6-6)^2}{6-1} = \frac{10}{5} = 2$$

- **Nghiên cứu 2: 10, 2, 3, 9, phương sai là:**

$$s^2 = \frac{(10-6)^2 + (2-6)^2 + (3-6)^2 + (9-6)^2}{4-1} = \frac{50}{3} = 16.7$$

Standard deviation – độ lệch chuẩn

- Vấn đề của phương sai là có đơn vị bình phương, nhưng trung bình là đơn vị gốc. Chúng ta cần hoán chuyển phương sai sang đơn vị gốc.
- Lấy căn số bậc 2 của phương sai = độ lệch chuẩn → ***standard deviation*** (kí hiệu là *s*)

- Nghiên cứu 1, s = sqrt(2) = 1.41

  Nghiên cứu 2, s = sqrt(16.7) = 4.1

Coefficient of variation – hệ số biến thiên

- Trong nhiều nghiên cứu, SD dao động với số trung bình (mean)
- Một chỉ số khác thường được sử dụng để mô tả độ dao động / biến thiên là ***hệ số biến thiên*** - ***coefficient of variation*** (CV).
- CV mô tả SD là phần trăm của số trung bình:

  CV = SD/mean*100

- Nghiên cứu 1: CV = 1.41 / 6 * 100 = 23.5%

  Nghiên cứu 2: CV = 4.1 / 6 * 100 = 68.3%

Tóm lược các chỉ số căn bản

Nghiên cứu	N	Mean	SD	CV (%)
1	6	6.0	1.4	23.5
2	4	6.0	4.1	68.3

Ý nghĩa của SD

- "Trong một quần thể, trọng lượng trung bình là 55 kg với độ lệch chuẩn (SD) Là 8.2 kg".
- Câu này có nghĩa là gì?

- Nếu trọng lượng tuân theo luật phân bố chuẩn (normal distribution), câu trên **có thể** hiểu là xác suất một cá nhân trong quần thể đó có trọng lượng *w* kg là:

$$P(Weight = w) = \frac{1}{s\sqrt{2\pi}} \exp\left[\frac{-(w-\bar{x})^2}{2s^2}\right]$$

Ý nghĩa của SD

- Trong ví dụ trên $x = 55$, $s = 8.2$
- *Xác suất một cá nhân được chọn ngẫu nhiên từ quần thể có trọng lượng 40, 50, 80 kg là:*

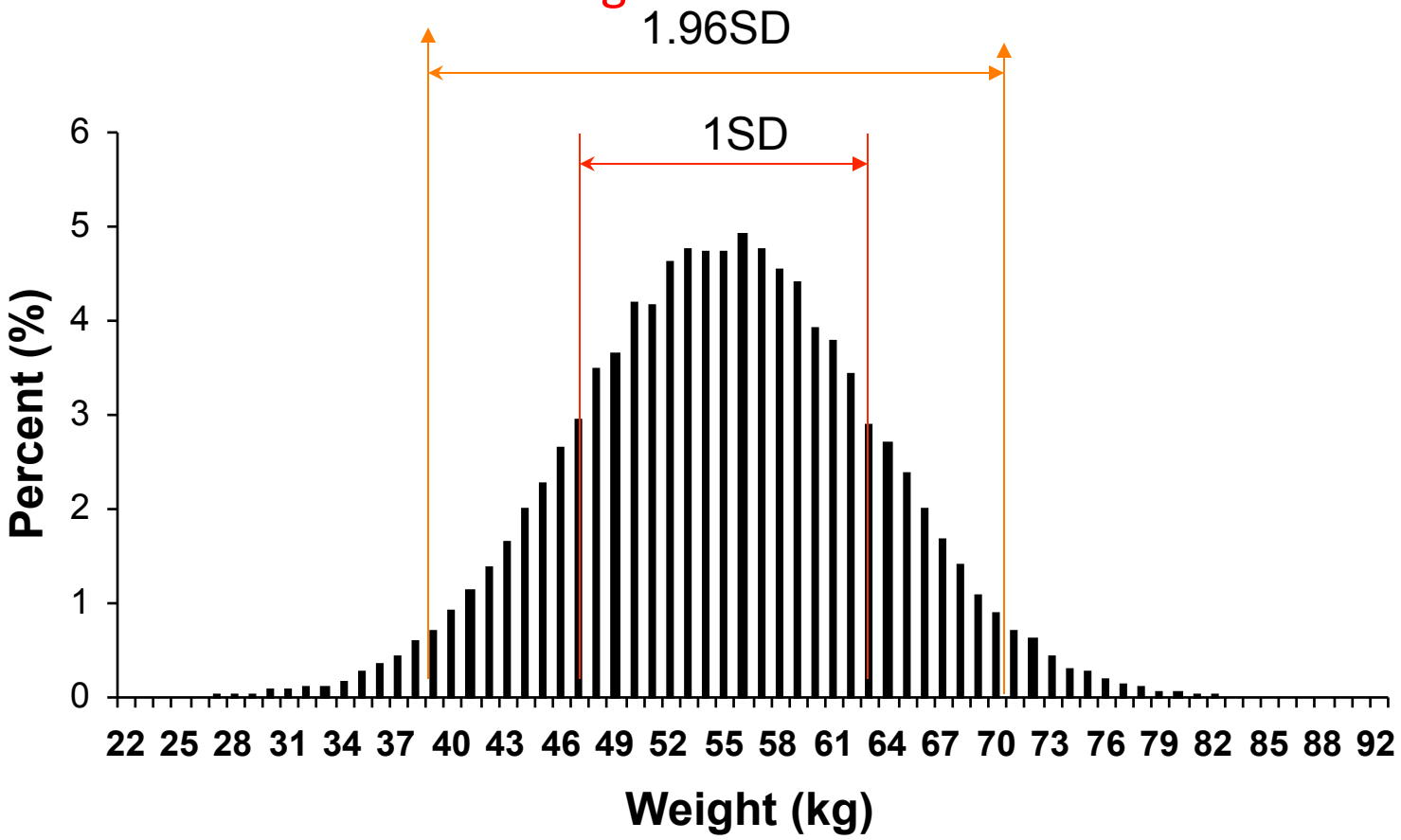
$$P(Weight = 40) = \frac{1}{8.2 \times \sqrt{2 \times 3.1416}} \exp\left[\frac{-(40-55)^2}{2 \times 8.2 \times 8.2}\right] = 0.009$$

$$P(Weight = 50) = \frac{1}{8.2 \times \sqrt{2 \times 3.1416}} \exp\left[\frac{-(50-55)^2}{2 \times 8.2 \times 8.2}\right] = 0.040$$

$$P(Weight = 80) = \frac{1}{8.2 \times \sqrt{2 \times 3.1416}} \exp\left[\frac{-(80-55)^2}{2 \times 8.2 \times 8.2}\right] = 0.0004$$

Mean = 55 kg, SD = 8.2 kg

- 68% các cá nhân trong quần thể có cân nặng dao động trong khoảng 55 +/- 8.2*1 = 46.8 đến 63.2 kg.
- 95% các cá nhân trong quần thể có cân nặng dao động trong khoảng 55 +/- 8.2*1.96 = 38.9 đến 71.1 kg



Khoảng tin cậy 95% = mean ± 1.96xSD

Z-scores – chỉ số z

- Có thể hoán chuyển đơn vị kg sang chỉ số z (z scores)
- z-score là ***số độ lệch chuẩn (SD) từ trung bình***

$$Z = \frac{x - \bar{x}}{s}$$

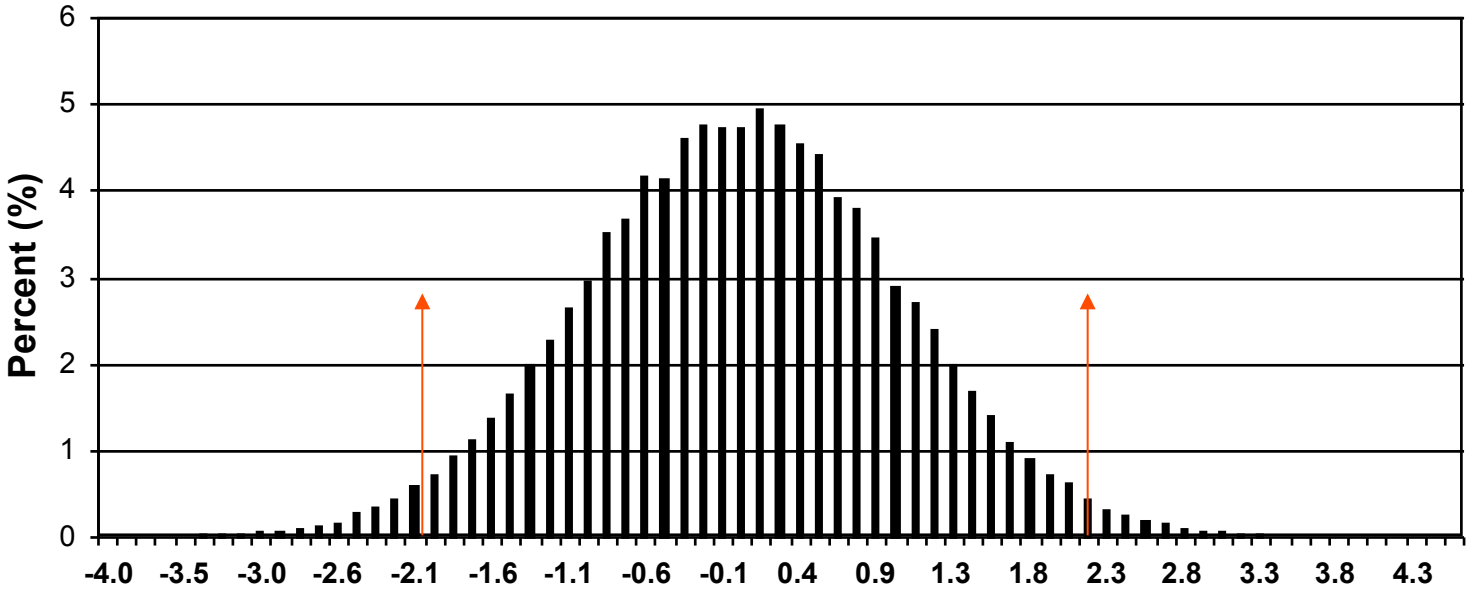
- weight = 55 kg → z=(55-55)/8.2 = 0 SDs
- weight = 40 kg → z=(40-55)/8.2 = -1.8 SDs
- weight = 80 kg → z=(80-55)/8.2 = 3.0 SDs

Z-scores = Standard Normal Distribution

- Z score không có đơn vị đo lường (unitless); cho phép so sánh giữa các biến
- Z-scores có trung bình = 0 và phương sai = 1
- Z-scores → **Standard Normal Distribution**

Z-scores và area under the curve

Z-scores và weight – một cách nhìn khác:



- **Area under the curve for $z \leq -1.96 = 0.025$**
- **Area under the curve for $-1.0 \leq z \leq 1.0 = 0.6828$**
- **Area under the curve for $-2.0 \leq z \leq 2.0 = 0.9544$**
- **Area under the curve for $-3.0 \leq z \leq 3.0 = 0.9972$**

95% confidence interval

- Một mẫu với *n* đo lường từ (x_1, x_2, …, x_n), có mean *x,* và SD *s*
  - 95% các cá nhân có giá trị x_i dao động trong khoảng *x-1.96s* and *x+1.96s*
- Trung bình = 55 kg, SD = 8.2 kg
  - 95% các cá nhân có cân nặng từ 39 đến 71 kg

Standard error – sai số chuẩn

- Định nghĩa của SE:

$$SE = \frac{SD}{\sqrt{n}}$$

  - **Câu hỏi**: nếu **nhiều mẫu ngẫu nhiên** được chọn từ một quần thể, và mỗi lần, chúng ta tính trung bình mẫu (sample means); độ dao động của các số trung bình mẫu là bao nhiêu?
  - Trong thực tế, SE có nghĩa là độ lệch chuẩn của các số trung bình mẫu (*the standard deviation of sample* ***means***)

Một cách hiểu SE

- Thử tưởng tượng chúng ta có 1 quần thể gồm 10 người: 130, 189, 200, 156, 154, 160, 162, 170, 145, 140
- Chúng ta biết rằng số trung bình quần thể là 160.6 cm
- Bây giờ chúng ta lấy mẫu (có hoàn lại) rất nhiều lần từ quần thể; và mỗi lần, chúng ta tính một số trung bình. Kết quả có thể như sau:

Sample 1: 140, 160, 200, 140, 145 → x1 = 157.0
Sample 2: 154, 170, 162, 160, 162 → x2 = 161.6
Sample 3: 145, 140, 156, 140, 156 → x3 = 147.4
Sample 4: 140, 170, 162, 170, 145 → x4 = 157.4
Sample 5: 156, 156, 170, 189, 170 → x5 = 168.2
Sample 6: 130, 170, 170, 170, 170 → x6 = 162.0
Sample 7: 156, 154, 145, 154, 189 → x7 = 159.6
Sample 8: 200, 154, 140, 170, 170 → x8 = 166.8
Sample 9: 140, 170, 145, 162, 160 → x9 = 155.4
Sample 10: 200, 200, 162, 170, 162 → x10 = 178.8

…

Độ lệch chuẩn của các số trung bình x1, x2, x3, …., x_N gọi là <u>standard error</u>

Khoảng tin cậy 95% của trung bình quần thể

- **Xem một quần thể gồm 10 người: 130, 189, 200, 156, 154, 160, 162, 170, 145, 140**
- **Chúng ta biết chính xác rằng trung bình là: 160.6 cm**

- Lấy mẫu lần 1: 140, 160, 200, 140, 145

  Mean = 157.0

  SD = 25.4

  SE = 25.4 / sqrt(5) = 11.36

  Khoảng tin cậy 95% của trung bình quần thể nằm trong khoảng: 157 $\pm$ 1.96*11.36 = (134.7 đến 179.3)

R codes

- R có một số hàm liên quan
  - **`sample`**: lấy mẫu từ một quần thể
  - **`rnorm`**: mô phỏng một biến phân bố chuẩn
  - **`pnorm`**: tính xác suất của một phân bố chuẩn

Cách dùng sample

- Nếu chúng ta có population (quần thể) 10 người: 130, 189, 200, 156, 154, 160, 162, 170, 145, 140; Chúng ta có thể lấy mẫu 5 người (nhiều lần)

```
pop = c(130, 189, 200, 156, 154, 160, 162,
170, 145, 140)

sample1 = sample(pop, 5); sample1

sample2 = sample(pop, 5); sample2

sample3 = sample(pop, 5); sample3
```

Cách dùng sample

```
> sample(pop, 5)
[1] 130 189 170 160 154
> sample(pop, 5)
[1] 160 145 200 170 154
> sample(pop, 5)
[1] 140 130 160 156 189
```

Cách dùng rnorm

- Chúng ta muốn mô phỏng một biến số về chiều cao, biết rằng trong quần thể, trung bình là 160 cm và độ lệch chuẩn là 6.5 cm.
- Chúng ta muốn lấy mẫu 500 người ngẫu nhiên từ quần thể đó

```
rnorm(n, mean, sd)
```

```
rnorm(500, 160, 6.5)
```

Cách dùng rnorm

```
> rnorm(500, 160, 6.5)

[1] 164.4778 161.8173 168.8121 174.9737
166.0209

...

[496] 159.6290 153.2323 160.9172 149.9402
154.4308
```

Chúng ta sẽ có một mẫu gồm 500 người với chiều cao như trên. Chúng ta có thể kiểm tra xem trung bình và SD của mẫu có gần như trung bình và mẫu của quần thể

Cách dùng rnorm

Chúng ta sẽ có một mẫu gồm 500 người với chiều cao như trên. Chúng ta có thể kiểm tra xem trung bình và SD của mẫu có gần như trung bình và mẫu của quần thể

```
height  = rnorm(n=1000, mean=160, sd=6.5)

mean(height); sd(height)

[1] 159.7199

[1] 6.665989
```

Cách dùng pnorm

- **pnorm**: dùng để tính xác suất (hay diện tích dưới đường của hàm số phân bố chuẩn)

```
pnorm(q, mean, sd)
```

- Ví dụ: muốn biết có bao nhiêu người có chiều cao bằng hoặc thấp hơn (<=) 150 cm, nếu biết rằng trong quần thể, chiều cao trung bình là 160 cm và SD 6.5 cm. Đáp số (6.2%) bằng lệnh sau đây:

```
pnorm(150, mean=160, sd=6.5)
```

Cảm nhận rnorm và pnorm

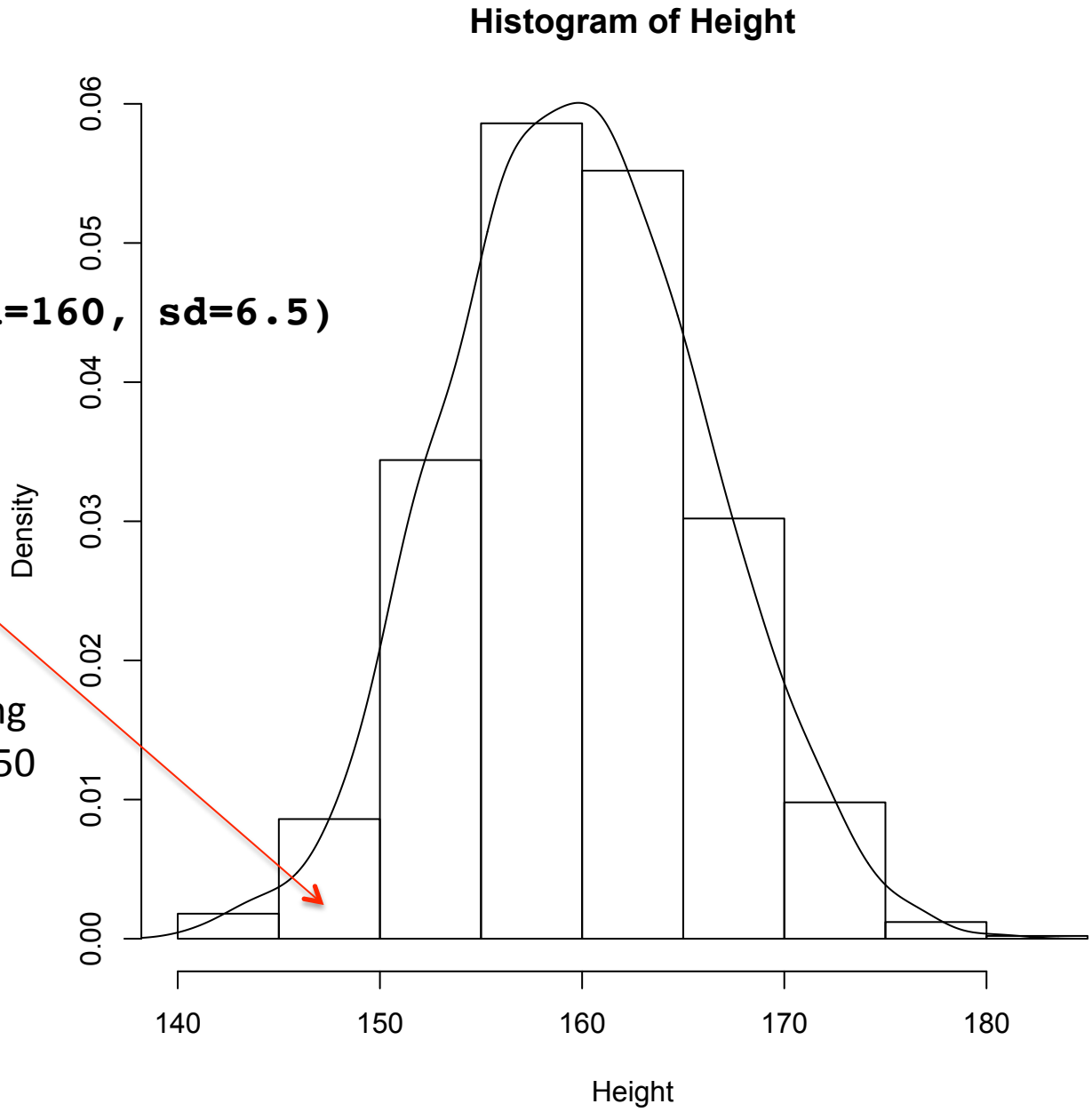
```
Height = rnorm(n=1000, mean=160, sd=6.5)

hist(Height, prob=T); lines(density(Height))

pnorm(150, mean=160, sd=6.5)
```

```
> pnorm(150, mean=160, sd=6.5)
[1] 0.0619679
```

Thể hiện diện tích (xác suất) dưới đường cong tính từ 0 đến 150



Một cách để hiểu SE dựa vào mô phỏng

```
population = c(130, 189, 200, 156, 154, 160, 162, 170,
145, 140) # quần thể gồm 10 người

k=5 #lấy mẫu chỉ 5 người mỗi lần
n = 1000 #lấy 1000 mẫu
mean = numeric(n) #định nghĩa biến mean là numeric
for (i in 1:n) #bắt đầu loop từ 1 đến n=1000
{
sample.i <- sample(pop, k, replace=T) #lấy mẫu
mean[i] = mean(sample.i) #tính trung bình mẫu
} #xong loop n lần

hist(mean, breaks=20, main="Distribution of means")
#xem phân bố 1000 giá trị trung bình
sd(mean) #tính SD của 1000 giá trị trung bình, và đây
chính là standard error
```

Mean, SD, SE, 95%CI: tóm lược

- *Mean* – số trung bình
- *Standard deviation* - độ lệch chuẩn phản ảnh độ dao động
- *Coefficient of variation* – hệ số biến thiên phản ảnh độ biến thiên so với giá trị trung bình
- Không có cái gọi là "standard error of the means" (SEM). Chỉ có standard deviation of the means – tức Standard error (SE)
- SE phản ảnh độ dao động của nhiều số trung bình mẫu
- Khoảng tin cậy *95%* - phản ảnh các giá trị khả dĩ với xác suất 95%.