

# XÂY DỰNG KHOẢNG TIN CẬY BẰNG MÔ PHỎNG ĐƠN GIẢN TRONG R THỐNG KÊ MÁY TÍNH & ỨNG DỤNG

(Vũ Quốc Hoàng – 09/12/2018)

## 1. Xây dựng khoảng tin cậy cho tỉ lệ

Giả sử ta muốn khảo sát tỉ lệ giới tính trong một *tổng thể* (population) là tất cả cư dân của một thành phố. Từ tổng thể, ta thu thập được một *mẫu dữ liệu ngẫu nhiên* (random sample) gồm 100 người trong đó có 60 nữ và 40 nam. Dựa vào mẫu dữ liệu này, ta xây dựng *khoảng tin cậy* (confident interval) 95% cho tỉ lệ nữ của thành phố.

Gọi  $p$  là tỉ lệ nữ trong dân cư của thành phố,  $\hat{p}$  là tỉ lệ nữ trong mẫu dữ liệu. Ta có

$$\hat{p} = \frac{60}{100} = 0.6$$

Ta không biết *tham số* (parameter)  $p$  là bao nhiêu, nhưng một cách “hợp lý”, ta có thể dùng *thống kê* (statistic)  $\hat{p}$  để ước lượng  $p$ . Tức là, ta đoán rằng

$$p \approx \hat{p} = 0.6$$

Tuy nhiên, ta không “tin cậy” lắm vào ước lượng này vì tùy theo mẫu dữ liệu thu thập mà ta có kết quả ước lượng khác nhau. Thay vì đưa ra một giá trị ước lượng, ta đưa ra một khoảng ước lượng cho  $p$ , trong đó có tính đến sự biến động của  $\hat{p}$  theo mẫu dữ liệu cỡ  $n = 100$  có thể thu thập từ tổng thể.

*Xây dựng khoảng tin cậy bằng công thức lý thuyết:*

Theo *định lý giới hạn trung tâm* (central limit theorem), khi cỡ mẫu  $n$  đủ lớn, ta có *phân phối mẫu* (sampling distribution) của  $\hat{p}$  “xấp xỉ chuẩn”, tức là

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Từ đó, ta có khoảng ước lượng với *độ tin cậy* (confident level)  $1 - \alpha$  cho  $p$  là

$$\hat{p} \pm z_{1-\alpha/2} se$$

Với  $z_{1-\alpha/2}$  là phân vị mức  $1 - \alpha/2$  của phân phối chuẩn tắc  $Z \sim N(0, 1)$ , tức là

$$P\left(Z \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} \text{ hay } P\left(Z \geq z_{1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

Còn  $se$  là *sai số chuẩn* (standard error) của  $\hat{p}$ , tức là

$$se = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

Vì không biết giá trị của  $p$  nên ta xấp xỉ sai số chuẩn bằng

$$se \approx \widehat{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Với  $n = 100, \hat{p} = 0.6, 1 - \alpha = 0.95$  ta có  $\alpha = 0.05, z_{1-\alpha/2} = z_{0.975} = 1.96, \widehat{se} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.0490$ . Từ đó, khoảng tin cậy 95% cho  $p$  là

$$\hat{p} \pm z_{1-\alpha/2} se = 0.6 \pm 1.96 \times 0.0490 = [0.50, 0.70]$$

Các bước tính toán trên có thể thực hiện trong R như sau:

```
n <- 100; p_hat <- 0.6; (anpha <- 1 - 0.95)
(z <- qnorm(1 - anpha/2))
(se <- sqrt(p_hat*(1 - p_hat)/n))
(conf_int <- p_hat + c(-z*se, z*se))
0.5039818 0.6960182
```

Hoặc có thể thực hiện nhanh bằng hàm `prop.test` như sau:

```
prop.test(60, 100, conf.level = 1 - anpha)
1-sample proportions test with continuity correction

data: 60 out of 100, null probability 0.5
X-squared = 3.61, df = 1, p-value = 0.05743
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.4970036 0.6952199
```

Lưu ý, vì `prop.test` dùng công thức khác (dựa trên phân phối Chi-bình phương) nên kết quả tính có khác chút xíu so với cách dùng công thức dựa trên phân phối chuẩn ở trên.

*Xây dựng khoảng tin cậy bằng mô phỏng trong R:*

Mô tả dữ liệu và thống kê tương ứng

```
n <- 100
data <- c(rep(1, 60), rep(0, 40)) #mẫu dữ liệu gồm 60 nữ, 40 nam
stat <- function(data)           #thống kê cần tính
{
  return(sum(data)/length(data)) #tỉ lệ mẫu
}
```

Một *mẫu bootstrap* (bootstrap sample) là mẫu cùng cỡ được lấy có hoàn lại trên chính dữ liệu. Bằng cách lấy nhiều mẫu bootstrap và tính thống kê tương ứng, ta có tập kết quả thể hiện sự biến động của thống kê theo mẫu, gọi là *phân phối bootstrap* (bootstrap distribution). Kỹ thuật mô phỏng này được gọi là *bootstrap* (bootstrapping).

```
bootstrap <- function(B)
{
  return(replicate(B, stat(sample(data, n, replace = TRUE))))
}
```

Chạy với số lượng mẫu bootstrap  $B$  đủ lớn thì phân phối bootstrap có thể được dùng để “xấp xỉ” phân phối mẫu. Từ đó ta có thể lấy *độ lệch chuẩn* (standard deviation) của phân phối bootstrap để “xấp xỉ” cho sai số chuẩn của thống kê tương ứng và lấy khoảng *phân vị* (percentile) của phân phối bootstrap để “xấp xỉ” cho khoảng tin cậy của tham số tương ứng. Cũng lưu ý là mỗi lần chạy sẽ có kết quả khác nhau do ngẫu nhiên.

Chẳng hạn thử với  $B = 10000$

```
anpha <- 1 - 0.95
boots_dist <- bootstrap(10000)
(se_boots <- sd(boots_dist))
0.04881839
(confint_boots <- quantile(boots_dist, c(anpha/2, 1 - anpha/2)))
2.5% 97.5%
0.5 0.7
```

Vậy sai số chuẩn xấp xỉ bằng bootstrap là 0.0488 và khoảng tin cậy 95% cho  $p$  dựa trên bootstrap là  $[0.50, 0.70]$ . Các kết quả này rất giống với các kết quả tính bằng công thức lý thuyết.

## 2. Xây dựng khoảng tin cậy cho kì vọng

Giả sử ta muốn khảo sát chiều cao trung bình của một tổng thể là tất cả cư dân của một thành phố. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên gồm chiều cao của 40 người như bảng dưới. Dựa vào mẫu dữ liệu này, ta xây dựng khoảng tin cậy 95% cho chiều cao trung bình của tổng thể.

Bảng dữ liệu chiều cao 40 người (đơn vị mét)

|      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|
| 1.56 | 1.47 | 1.59 | 1.65 | 1.62 | 1.78 | 1.69 | 1.49 | 1.92 | 1.55 |
| 1.65 | 1.52 | 1.65 | 1.60 | 1.71 | 1.48 | 1.69 | 1.65 | 1.59 | 1.74 |
| 1.70 | 1.61 | 1.58 | 1.65 | 1.75 | 1.65 | 1.46 | 1.53 | 1.59 | 1.62 |
| 1.60 | 1.55 | 1.57 | 1.46 | 1.57 | 1.63 | 1.46 | 1.68 | 1.53 | 1.48 |

Gọi  $\mu$  là chiều cao trung bình của cư dân thành phố,  $\bar{x}$  là chiều cao trung bình của  $n = 40$  người trong mẫu dữ liệu. Ta có

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 1.61$$

Ta không biết tham số  $\mu$  là bao nhiêu, nhưng một cách “hợp lý”, ta có thể dùng thống kê  $\bar{x}$  để ước lượng  $\mu$ . Tức là, ta đoán rằng

$$\mu \approx \bar{x} = 1.61$$

Tuy nhiên, ta không “tin cậy” lắm vào ước lượng này vì tùy theo mẫu dữ liệu thu thập mà ta có kết quả ước lượng khác nhau. Thay vì đưa ra một giá trị ước lượng, ta đưa ra một khoảng ước lượng cho  $\mu$ , trong đó có tính đến sự biến động của  $\bar{x}$  theo mẫu dữ liệu cỡ  $n = 40$  có thể thu thập từ tổng thể.

*Xây dựng khoảng tin cậy bằng công thức lý thuyết:*

Theo định lý giới hạn trung tâm, khi cỡ mẫu  $n$  đủ lớn, ta có phân phối mẫu của  $\bar{x}$  “xấp xỉ chuẩn”, tức là

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Với  $\sigma^2$  là phương sai của dữ liệu mà ở đây là phương sai của chiều cao của một người trong tổng thể. Từ đó, ta có khoảng ước lượng  $1 - \alpha$  cho  $\mu$  là

$$\bar{x} \pm z_{1-\alpha/2} se$$

Với  $z_{1-\alpha/2}$  là phân vị mức  $1 - \alpha/2$  của phân phối chuẩn tắc  $Z \sim N(0, 1)$  còn  $se$  là sai số chuẩn của  $\bar{x}$ , tức là

$$se = \sqrt{\text{Var}(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

Nếu biết giá trị của  $\sigma$ , ta có thể tính được  $se$  bằng công thức trên. Tuy nhiên, thường thì ta không biết giá trị của  $\sigma$ , khi đó ta có thể xấp xỉ sai số chuẩn bằng

$$se \approx \widehat{se} = \frac{s}{\sqrt{n}}$$

Với  $s$  là độ lệch chuẩn của mẫu dữ liệu, được tính bằng công thức

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = 0.0985$$

Với  $n = 40, \bar{x} = 1.61, s = 0.0985, 1 - \alpha = 0.95$  ta có  $\alpha = 0.05, z_{1-\alpha/2} = z_{0.975} = 1.96, \widehat{se} = \frac{s}{\sqrt{n}} = 0.0156$ . Từ đó, khoảng tin cậy 95% cho  $\mu$  là

$$\bar{x} \pm z_{1-\alpha/2} se = 1.61 \pm 1.96 \times 0.0156 = [1.58, 1.64]$$

Thật ra công thức ước lượng trên chưa tốt lắm vì khi dùng  $s$  thay cho  $\sigma$  thì ta có thể dùng phân phối Student để xấp xỉ tốt hơn phân phối chuẩn, cụ thể ta có

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Từ đó, ta có khoảng ước lượng  $1 - \alpha$  cho  $\mu$  là

$$\bar{x} \pm t_{n-1}^{1-\alpha/2} se$$

Với  $t_{n-1}^{1-\alpha/2}$  là phân vị mức  $1 - \alpha/2$  của phân phối Student với bậc tự do là  $n - 1$ , tức là

$$P(t_{n-1} \leq t_{n-1}^{1-\alpha/2}) = 1 - \frac{\alpha}{2} \text{ hay } P(t_{n-1} \geq t_{n-1}^{1-\alpha/2}) = \frac{\alpha}{2}$$

Với  $n = 40, \bar{x} = 1.61, s = 0.0985, 1 - \alpha = 0.95$  ta có  $\alpha = 0.05, t_{n-1}^{1-\alpha/2} = t_{39}^{0.975} = 2.02, \widehat{se} = \frac{s}{\sqrt{n}} = 0.0156$ . Từ đó, khoảng tin cậy 95% cho  $\mu$  tính dựa trên phân phối Student là

$$\bar{x} \pm t_{n-1}^{1-\alpha/2} se = 1.61 \pm 2.02 \times 0.0156 = [1.58, 1.64]$$

Khi cỡ mẫu  $n$  lớn thì khoảng tin cậy xây dựng dựa trên phân phối chuẩn giống như khoảng tin cậy xây dựng dựa trên phân phối Student.

Các bước tính toán trên có thể thực hiện trong R như sau:

```
data <- c(1.56, 1.47, 1.59, 1.65, 1.62, 1.78, 1.69, 1.49, 1.92, 1.55,
          1.65, 1.52, 1.65, 1.60, 1.71, 1.48, 1.69, 1.65, 1.59, 1.74,
          1.70, 1.61, 1.58, 1.65, 1.75, 1.65, 1.46, 1.53, 1.59, 1.62,
          1.60, 1.55, 1.57, 1.46, 1.57, 1.63, 1.46, 1.68, 1.53, 1.48)
(n <- length(data)); (x_bar <- mean(data)); (anpha <- 1 - 0.95)
(se <- sd(data)/sqrt(n))
(z <- qnorm(1 - anpha/2))
(zconf_int <- x_bar + c(-z*se, z*se))
1.576221 1.637279
(t <- qt(1 - anpha/2, df = n - 1))
(tconf_int <- x_bar + c(-t*se, t*se))
1.575244 1.638256
```

Hoặc có thể thực hiện nhanh bằng hàm `t.test` như sau:

```
t.test(data, conf.level = 1 - anpha)
One Sample t-test

data: data
t = 103.16, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

1.575244 1.638256

### Xây dựng khoảng tin cậy bằng mô phỏng trong R:

Với dữ liệu đã được nhập như trên, ta mô tả thêm thống kê tương ứng (ở đây là trung bình mẫu) và hàm thực hiện bootstrap như phần 1

```
stat <- function(data)      #thống kê cần tính
{
  return(mean(data))        #tỉ lệ mẫu
}
bootstrap <- function(B)
{
  return(replicate(B, stat(sample(data, n, replace = TRUE))))
}
```

Chạy thử với B = 10000

```
anpha <- 1 - 0.95
boots_dist <- bootstrap(10000)
(se_boots <- sd(boots_dist))
0.01533063
(confint_boots <- quantile(boots_dist, c(anpha/2, 1 - anpha/2)))
2.5% 97.5%
1.57725 1.63725
```

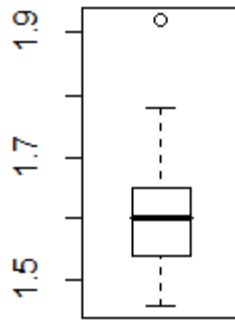
Vậy sai số chuẩn xấp xỉ bằng bootstrap là 0.0153 và khoảng tin cậy 95% cho  $\mu$  dựa trên bootstrap là [1.58, 1.64]. Các kết quả này rất giống với các kết quả tính bằng công thức lý thuyết.

Lưu ý là code chạy mô phỏng trong phần 2 này hoàn toàn giống code chạy mô phỏng trong phần 1, chỉ khác hàm tính thống kê (ở đây là trung bình mẫu so với phần 1 là tỉ lệ mẫu).

### **3. Xây dựng khoảng tin cậy cho trung vị**

Giả sử cùng tổng thể và mẫu dữ liệu ở phần 2 nhưng ta muốn xây dựng khoảng tin cậy 95% cho *trung vị* (median) thay vì trung bình của chiều cao. Mặc dù trung bình thường được sử dụng như là con số mô tả trọng tâm của phân phối, nó lại rất nhạy cảm với *ngoại lệ* (outlier). Nghĩa là, trung bình thay đổi rất nhiều khi mẫu dữ liệu có ngoại lệ. Chẳng hạn trong mẫu dữ liệu ở phần 2 ta có một dữ liệu là 1.92, lớn hơn rất nhiều so với các dữ liệu còn lại. (Trong trường hợp này, đây có thể là chiều cao của một người nước ngoài, đã bị lấy mẫu nhầm 😊). Ta có thể thấy ngoại lệ này qua đồ thị hộp như sau:

```
boxplot(data)
```



Trong những trường hợp như thế này ta có thể dùng trung vị là một thống kê ít bị ảnh hưởng bởi ngoại lệ. Việc xây dựng khoảng tin cậy cho trung vị phức tạp hơn cho trung bình. Tuy nhiên, với phương pháp tổng quát và đơn giản như bootstrap, ta dễ dàng làm được điều này.

Với dữ liệu và đoạn mã bootstrap như phần 2, ta chỉ cần sửa lại thống kê tương ứng để tính trung vị của mẫu thay vì trung bình

```
stat <- function(data)      #thống kê cần tính
{
  return(median(data))      #trung vị mẫu
}
```

Chạy thử với B = 10000

```
anpha <- 1 - 0.95
boots_dist <- bootstrap(10000)
(se_boots <- sd(boots_dist))
0.01854741
(confint_boots <- quantile(boots_dist, c(anpha/2, 1 - anpha/2)))
2.5% 97.5%
1.57 1.65
```

Vậy sai số chuẩn và khoảng tin cậy 95% cho trung vị xấp xỉ bằng bootstrap là 0.0185 và [1.57, 1.65]. Lưu ý là code chạy mô phỏng trong phần này hoàn toàn giống code chạy mô phỏng trong 2 phần trên, chỉ khác hàm tính thống kê (ở đây là trung vị mẫu).