Bài 7.

Ước lượng tham số thống kê

Trong nội dung bài thực hành, chúng ta chỉ xét bài toán ước lượng khoảng (tìm khoảng tin cậy) cho tỷ lệ và kỳ vọng.

I. Tóm tắt lý thuyết

I.1 Khoảng tin cậy cho trung bình của tổng thể:

- Trung bình tổng thể: μ

- Độ lệch chuẩn tổng thể: σ

- Trung bình mẫu: $\overline{x}$ = `mean(x)`

- Độ lệch chuẩn mẫu: $s = \sqrt{s^2}$ = `sd(x)`

- Dung sai: ε

- **Công thức tính dung sai**: phụ thuộc vào 3 trường hợp

TH1: σ đã biết, tổng thể có phân phối chuẩn (*nếu* $n \geq 30$ *thì giả thiết tổng thể có phân phối chuẩn là không còn cần thiết*)

TH2: σ không biết và $n \geq 30$.

TH3: σ không biết và $n < 30$ kèm theo tổng thể có phân phối chuẩn (khi cỡ mẫu nhỏ hơn 30 thì cần tới giả thiết tổng thể có phân phối chuẩn)

$$\varepsilon = \begin{cases} z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}, & \text{TH1} \\ z_{1-\alpha/2}\dfrac{s}{\sqrt{n}}, & \text{TH2} \\ t^{n-1}_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}, & \text{TH3} \end{cases}$$

Trong đó, $z_{1-\alpha/2}$ = `qnorm(1-alpha/2)`; $t^{n-1}_{1-\alpha/2}$ = `qt(1-alpha/2, df = n-1)`

- **Công thức tính khoảng tin cậy**:

$$\text{Khoảng tin cậy} = \text{ trung bình mẫu } \pm \text{ dung sai}$$

I.2 Khoảng tin cậy cho tỷ lệ tổng thể

- Tỷ lệ tổng thể: p

- Tỷ lệ mẫu: $\hat{p}$

- Dung sai: $\varepsilon = z_{1-\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

- Công thức tính khoảng tin cậy:

Khoảng tin cậy = tỷ lệ mẫu $\pm$ dung sai

I.3 Ý nghĩa của khoảng tin cậy 95% cho tham số cần ước lượng (trung bình / tỷ lệ)

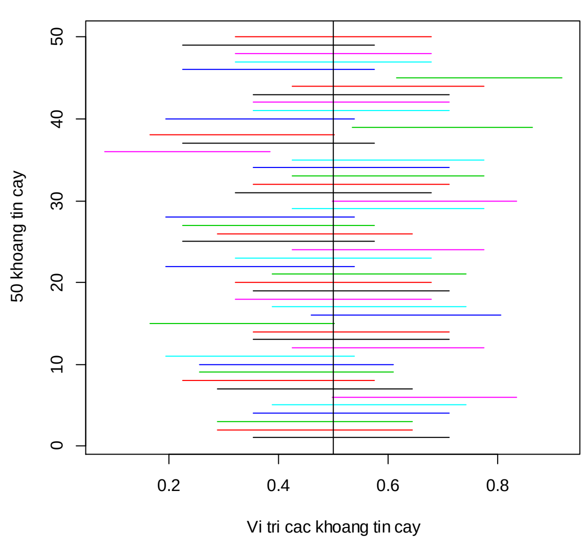
Nếu chúng ta lấy mẫu nhiều lần, mỗi mẫu chúng ta tính 1 khoảng tin cậy thì sẽ có xấp xỉ 95% khoảng tin cậy chứa tham số cần ước lượng.

Ví dụ: đoạn chương trình sau mô phỏng 50 khoảng tin cậy cho tỷ lệ xuất hiện mặt sấp khi tung một đồng xu cân đối

```
> m = 50; n=30; p = .5; alpha = 0.05   # Tung 30 đồng xu cân đối 50 lần
> p.hat = rbinom(m,n,p)/n              # Tính tỷ lệ mẫu
> epsilon = qnorm(1-alpha/2)*sqrt(p.hat*(1-p.hat)/n)
                                       # Dung sai
>  matplot(rbind(p.hat - epsilon, p.hat + epsilon),
           rbind(1:m,1:m),type="l",lty=1,
           xlab = "Vi tri cac khoang tin cay",
           ylab = "50 khoang tin cay")
                                       # Ve 50 khoang tin cay
> abline(v=p)                          # Ve duong thang p = 0.5
```

Chú thích: hàm `matplot(x,y,…)` ở trên cho phép ta vẽ m đoạn thẳng, đoạn thẳng thứ i có 2 hoành độ (của điểm đầu và điểm cuối) nằm trong cột i của ma trận x; và có 2 tung độ bằng nhau là i nằm trong cột i của ma trận y.

Kết quả nhận được:



II. Bài tập

1. Tạo ngẫu nhiên 35 giá trị của biến ngẫu nhiên có phân phối chuẩn với trung bình bằng 10 và độ lệch chuẩn 5. Tìm khoảng tin cậy 95% cho kỳ vọng của biến ngẫu nhiên chuẩn dựa vào số liệu vừa tạo.

2. Số liệu thống kê về doanh số bán hàng của một siêu thị cho ở file data31.xls:

a) Đọc dữ liệu từ file data31.xls vào R.

b) Viết hàm **ci.mean(x, alpha)** xuất ra khoảng tin cậy cho kỳ vọng, với x là vec-tơ dữ liệu, (1-alpha) là độ tin cậy. Áp dụng để tìm khoảng tin cậy 95% và 99% cho doanh số bán hàng trung bình ở siêu thị.

3. File data32.xls chứa số liệu về thời gian tự học của 120 sinh viên trường ĐH Khoa học Tự nhiên.

a. Hãy ước lượng thời gian học nhóm trung bình của sinh viên trường ĐH KHTN, độ tin cậy là 95%. (Dùng hàm **ci.mean(x, alpha)**)

b. Viết hàm hàm **ci.prop(f, n, alpha)** xuất ra khoảng tin cậy cho tỷ lệ, với n là cỡ mẫu; f: số các phần tử thỏa yêu cầu (với tỷ lệ p cần tìm); (1-alpha) là độ tin cậy. Áp dụng để tìm khoảng tin cậy 90%; 95% và 99% cho tỷ lệ sinh viên có thời gian tự học trên 5 giờ mỗi ngày.

4. Bảng sau thống kê chiều cao (Đv: m) của 125 thanh niên 18 tuổi trong một khu vực:

Chiều cao	[1.2,1.4)	[1.4,1.6)	[1.6,1.8)	[1.8,2.0)	[2.0,2.2)
Số thanh niên	6	34	31	42	12

a. Chuyển bảng tần số dạng khoảng ở trên thành dữ liệu dạng véc-tơ cột. Áp dụng hàm **ci.mean** đã ở bài **2** để tìm khoảng tin cậy 95% cho chiều cao trung bình của thanh niên trong khu vực.

b. Những người có chiều cao từ 1.7 m trở lên được xếp vào sức khỏa loại A. Sử dụng hàm ci.prop ở bài **3** để tìm khoảng tin cậy 95% cho tỷ lệ thanh niên đạt sức khoẻ loại A.

5. Viết hàm *ktc.tb()* để tìm khoảng tin cậy cho trung bình biết:

- Input: là trung bình mẫu $\bar{x}$, độ lệch chuẩn của tổng thể σ (có thể biết trước hoặc không), trường hợp không biết σ thì phải nhập độ lệch chuẩn của mẫu s, kích thước mẫu n, và mức ý nghĩa α.

- Output: khoảng tin cậy cho trung bình.

6. Từ hàm được viết trong câu 5) hãy viết hàm *ktc.tb.mau()* để tìm khoảng tin cậy cho trung bình biết:

- Input: vecto dữ liệu mẫu x, độ lệch chuẩn của tổng thể σ (có thể biết trước hoặc không), và mức ý nghĩa α.
- Output: khoảng tin cậy cho trung bình.

7. Đo đường kính của một chi tiết máy do một máy tiện tự động sản xuất, ta ghi nhận được số liệu như sau:

X	12.00	12.05	12.10	12.15	12.20	12.25	12.30	12.35	12.40
n	2	3	7	9	10	8	6	5	3

Bằng cách sử dụng hàm *ktc.tb.mau()* trong câu 6), hãy ước lượng khoảng tin cậy 95% cho đường kính trung bình.