Chương 5. Lý thuyết mẫu

Nguyễn Minh Trí

Trường Đại học Công nghệ Thông tin
Đại học Quốc gia Thành phố Hồ Chí Minh

Mục tiêu

- Biết các khái niệm cơ bản của thống kê.
- Biết biểu diễn đồ họa của dữ liệu và tính các thống kê từ mẫu.
- Biết phân phối lấy mẫu.

5.1 Các khái niệm cơ bản của thống kê

5.1 Các khái niệm cơ bản của thống kê

Định nghĩa 5.1 Thống kê (statistics) là khoa học về thu thập, tổ chức, tóm tắt và phân tích thông tin để rút ra kết luận hoặc trả lời câu hỏi. Ngoài ra, thống kê còn là việc cung cấp thước đo độ tin cậy trong bất kỳ kết luận nào.

- **Thống kê mô tả** (descriptive statistics): Các phương pháp bao gồm chủ yếu là tổ chức, tóm tắt và trình bày dữ liệu dưới dạng bảng, đồ thị và biểu đồ.
- **Thống kê suy luận** (inferential statistics): Các phương pháp rút ra kết luận và đưa ra quyết định về tổng thể từ các mẫu.

Định nghĩa 5.2 **Tổng thể** (population) là tập hợp tất cả các đối tượng hoặc phép đo mà người thu thập quan tâm.

Ví dụ 5.3 Giả sử chúng ta muốn nghiên cứu chiều cao của tất cả nam sinh viên tại một trường đại học nào đó.

- **Tổng thể** là tập hợp các chiều cao đo được của tất cả sinh viên nam trong trường đại học.
- Tổng thể **không** phải là tập hợp tất cả sinh viên nam trong trường đại học.

Trong thực tế, ta khó thể có được thông tin của toàn bộ tổng thể. Mục tiêu chính của thống kê là thu thập và nghiên cứu một tập hợp con của tổng thể, được gọi là **mẫu**, để đưa ra thông tin về một số đặc điểm của tổng thể.

Định nghĩa 5.4 **Mẫu** (sample) là một tập hợp con của dữ liệu được chọn từ tổng thể. **Kích thước mẫu** là số phần tử của tập hợp con đó.

Ví dụ 5.5 Chúng ta muốn ước tính tỷ lệ phần trăm sản phẩm lỗi được sản xuất tại một nhà máy trong một ngày. Trong trường hợp này, “tất cả các sản phẩm được sản xuất trong ngày” là **tổng thể**. Chọn ngẫu nhiên 100 sản phẩm để kiểm tra số sản phẩm lỗi. Khi đó ta có một mẫu là "100 sản phẩm" và kích thước mẫu là 100.

Định nghĩa 5.6 Một mẫu được chọn sao cho mọi phần tử của tổng thể đều có cơ hội được chọn như nhau và độc lập với nhau được gọi là **mẫu ngẫu nhiên đơn giản** (simple random sample).

Ví dụ 5.7 Có 100 quả bóng trong một cái hộp. Lấy ra 10 quả bóng mà không cần nhìn vào trong hộp. Khi đó 10 quả bóng được lấy ra là một mẫu ngẫu nhiên.

Chú ý:

- Việc chọn mẫu luôn được xem là chọn không hoàn lại, tức là một phần tử nào đó không được chọn nhiều hơn 1 lần trong một mẫu.
- Mẫu phải được chọn ngẫu nhiên và kích thước mẫu đủ lớn.

5.2 Biểu diễn đồ họa của dữ liệu

5.2 Biểu diễn đồ họa của dữ liệu

Định nghĩa 5.8

1. Cho một mẫu có kích thước n và các giá trị của dấu hiệu X mà ta muốn nghiên cứu là $x_1 < x_2 < \ldots < x_m$. Số lần lặp lại k_i của x_i được gọi là tần số của x_i.
   Bảng phân bố tần số

X	x_1	x_2	$\cdots$	x_m
Tần số	k_1	k_2	$\cdots$	k_m

2. Tần suất f_i của giá trị x_i :

$$f_i = \frac{k_i}{n}$$

   Bảng phân bố tần suất

X	x_1	x_2	$\cdots$	x_m
Tần suất	f_1	f_2	$\cdots$	f_m

Ví dụ 5.9 Kiểm tra 80 hộp (mỗi hộp chứa 100 chip bán dẫn) để tìm số lượng chip bị lỗi trong mỗi hộp. Số chip bị lỗi trong mỗi hộp như sau

1	3	4	7	2	7	5	5
2	2	4	2	5	4	3	2
2	7	1	3	3	2	5	0
0	1	2	5	5	4	1	3
3	2	6	3	8	2	2	3
1	6	3	4	1	2	5	3
1	3	3	3	2	1	2	5
5	4	1	4	3	1	0	3
2	1	2	4	4	5	3	3
4	0	5	2	5	6	2	1

Số chip bị lỗi	Tần số	Tần suất
0	4	0,05
1	12	0,15
2	18	0,225
3	17	0,2125
4	10	0,125
5	12	0,15
6	3	0,0375
7	3	0,0375
8	1	0,0125
≥ 9	0	0
Tổng	80	1

Người ta thường xác định một số khoảng $C_1, C_2, \ldots, C_m$ sao cho mỗi giá trị mà X nhận được chỉ thuộc một khoảng nào đó. Các khoảng này được gọi là các lớp ghép của X.

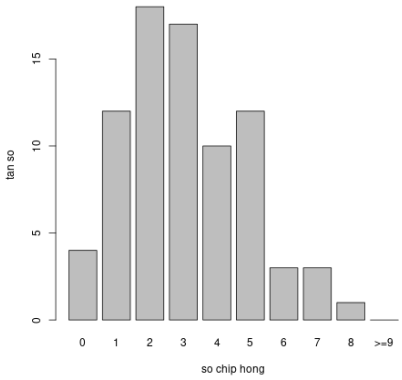
Ví dụ 5.10 Một mẫu về chiều cao của 40 sinh viên được trình bày trong bảng phân bố lớp ghép sau:

Khoảng	Tần số	Tần suất
$(146; 151]$	4	0,1
$(151; 156]$	2	0,05
$(156; 161]$	6	0,15
$(161; 166]$	10	0,25
$(166; 171]$	12	0,3
$(171; 176]$	6	0,15

Định nghĩa 5.11 Biểu đồ gồm các cột có chiều cao biểu thị tần số (tần suất) tương ứng của các loại đối tượng được gọi là biểu đồ cột (bar chart).

Ví dụ 5.12 Số chip bị lỗi trong mỗi hộp như sau

Số chip bị lỗi	Tần số	Tần suất
0	4	0,05
1	12	0,15
2	18	0,225
3	17	0,2125
4	10	0,125
5	12	0,15
6	3	0,0375
7	3	0,0375
8	1	0,0125
≥ 9	0	0
Tổng	80	1



Vẽ biểu đồ cột bằng phần mềm R

Vào trang web

https://www.tutorialspoint.com/execute_r_online.php

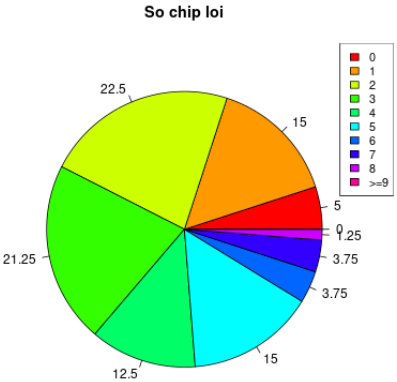
Trong phần Source Code

```
# Tan so cac cot
H <- c(4,12,18,17,10,12,3,3,1,0)
# Ten cac cot
M <- c("0","1","2","3","4","5","6","7","8",">=9")
# Ten file
png(file = "barchart.png")
# Ve bieu do cot
barplot(H,names.arg=M,xlab="so chip hong",ylab="tan so")
# Luu file
dev.off()
```

Định nghĩa 5.13 Một hình tròn được chia thành các phần biểu thị tỷ lệ phần trăm tổng thể hoặc một mẫu thuộc các danh mục khác nhau được gọi là biểu đồ tròn (pie chart).

Ví dụ 5.14 Số chip bị lỗi

Số chip bị lỗi	Tần số	Tỉ lệ %
0	4	5%
1	12	15%
2	18	22,5%
3	17	21,25%
4	10	12,5%
5	12	15%
6	3	3,75%
7	3	3,75%
8	1	1,25%
≥ 9	0	0%
Tổng	80	100%



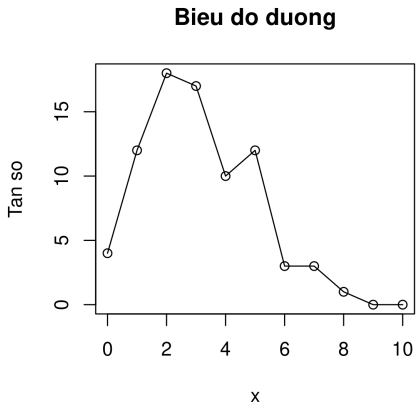
Vẽ biểu đồ tròn bằng phần mềm R

```
# So lieu phan tram
x <- c(5, 15, 22.5, 21.25, 12.5, 15, 3.75, 3.75, 1.25, 0)
# Ten cac danh mục
labels <- c("0", "1", "2", "3", "4", "5", "6", "7", "8", ">=9")
# Dat ten file
png(file = "piechiploi.png")
# Hien ti le phan tram
piepercent<- round(100*x/sum(x), 1)
# Ve bieu do tron
pie(x,labels=piepercent,main="So chip loi",col=rainbow(length(x)))
# Them chu thich
legend("topright",c("0", "1", "2", "3", "4", "5", "6", "7", "8", ">=9"),
cex=0.8, fill=rainbow(length(x)))
# Luu file
dev.off()
```

Định nghĩa 5.15 Biểu đồ đường (line chart) là một loại biểu đồ thống kê trong đó mỗi điểm dữ liệu (giá trị và tần số/tần suất) được biểu diễn bằng một điểm trên đồ thị và những điểm dữ liệu liên tiếp được kết nối bằng một đường thẳng.

Ví dụ 5.16 Số chip bị lỗi

Số chip bị lỗi	Tần số	Tỉ lệ %
0	4	5%
1	12	15%
2	18	22,5%
3	17	21,25%
4	10	12,5%
5	12	15%
6	3	3,75%
7	3	3,75%
8	1	1,25%
≥ 9	0	0%
Tổng	80	100%



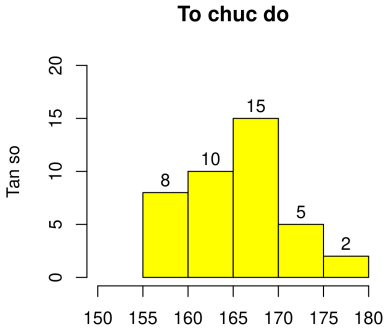
Vẽ biểu đồ đường bằng phần mềm R

```
# Gia tri cua x
x <- 0:10
# Tan so tương ứng
v <- c(4,12,18,17,10,12,3,3,1,0,0)
# Ve line chart
plot(x,v,type='o',main="Bieu do duong",ylab="Tan so")
```

Định nghĩa 5.17 Tổ chức đồ (histogram) là biểu đồ trong đó các lớp ghép được đánh dấu trên trục hoành và tần số/tần suất/tỷ lệ phần trăm được biểu thị bằng độ cao trên trục tung. Trong một tổ chức đồ, các cột được vẽ liền kề nhau mà không có bất kỳ khoảng trống nào.

Ví dụ 5.18 Chiều cao của 40 sinh viên

Khoảng	Tần số
$(155; 160]$	8
$(160; 165]$	10
$(165; 170]$	15
$(170; 175]$	5
$(175; 180]$	2



Vẽ tổ chức đồ bằng phần mềm R

```
# Nhap gia tri
v<- c(175,156,157,158,159,160,161,162,163,164,165,166,167,168,169,170,
171,172,173,174,165,176,167,178,169,170,161,166,167,168,165,166,167,
158,159,160,161,162,167,167)
# Ve Histogram
h<-hist(v,main="To chuc do", xlim=c(150,180),ylim=c(0,20),breaks=5,
ylab="Tan so",xlab=" ",col="yellow")
# Dat so tren cac cot
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
```

5.3 Các số đo mô tả

5.3 Các số đo mô tả

- Một tổng thể có kích thước N : $v_1, v_2, \ldots, v_N$
- Một mẫu có kích thước n nhận các giá trị $x_1, \ldots, x_n$.

Định nghĩa 5.19

1. **Trung bình tổng thể**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} v_i.$$

2. **Phương sai tổng thể**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \mu)^2.$$

3. **Độ lệch chuẩn tổng thể**

$$\sigma = \sqrt{\sigma^2}.$$

4. **Trung bình của một mẫu** (sample mean)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

5. **Phương sai mẫu** (sample variance), ký hiệu s^2,

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

6. **Độ lệch chuẩn mẫu** (sample standard deviation)

$$s = \sqrt{s^2}.$$

Ví dụ 5.20 So sánh giá cà phê tại 4 cửa hàng tạp hóa được chọn ngẫu nhiên ở Thủ Đức cho thấy các mức tăng so với tháng trước là 12, 15, 17 và 20 nghìn đồng cho một túi 1 kg. Tìm trung bình, phương sai của mẫu này.

Giải.

- Trung bình mẫu

$$\overline{x} = \frac{12 + 15 + 17 + 20}{4} = 16 \text{ (nghìn đồng)}$$

- Phương sai mẫu

$$\begin{aligned} s^2 &= \frac{1}{3}\sum_{i=1}^{4}(x_i - \overline{x})^2 \\ &= \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3} = \frac{34}{3} \end{aligned}$$

Tính trung bình, phương sai và độ lệch chuẩn bằng R

```
x <- c(12,15,17,20)
y <- mean(x)
y
[1] 16
z <- var(x)
z
[1] 11.33333
t<- sd(x)
t
[1] 3.366502
```

Định lý 5.21 Nếu s^2 là phương sai của một mẫu có kích thước n thì

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

Định nghĩa 5.22 Giả sử các giá trị của mẫu được sắp xếp từ nhỏ đến lớn. Trung vị mẫu (median) là một số m thỏa mãn

$$m = \begin{cases} x_{(n+1)/2}, & n \text{ lẻ} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & n \text{ chẵn} \end{cases}$$

Ví dụ 5.23 Các số liệu thu được như sau:

a. 100 45 60 130 30

b. 100 45 60 130 30 70

Tìm trung vị mẫu.

Giải. a. Sắp xếp dữ liệu theo thứ tự

30 45 60 100 130

Khi đó trung vị mẫu là 60.

b. Sắp xếp dữ liệu theo thứ tự

30 45 60 70 100 130

Khi đó trung vị mẫu là $\frac{60+70}{2} = 65.$

Định nghĩa 5.24 Cho một phân bố lớp ghép với m khoảng $C_1, C_2, \ldots, C_m$. Giả sử x_i là trung điểm (tâm) của khoảng C_i và k_i là tần số của khoảng C_i với $i = 1, 2, \ldots, m$. Khi đó trung bình mẫu $\overline{x}$ và phương sai mẫu s^2 được xác định bởi

$$\overline{x} = \frac{\sum_{i=1}^{m} k_i x_i}{\sum_{i=1}^{m} k_i}; \quad s^2 = \frac{1}{\sum_{i=1}^{m} k_i - 1} \sum_{i=1}^{m} k_i (x_i - \overline{x})^2$$

Ví dụ 5.25 Chiều cao của 40 sinh viên

Khoảng	Tần số
$(155; 160]$	8
$(160; 165]$	10
$(165; 170]$	15
$(170; 175]$	5
$(175; 180]$	2

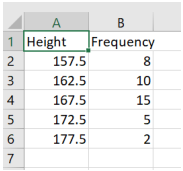
- Trung bình mẫu

$$\overline{x} = \frac{1}{40}(8.157,5 + 10.162,5 + 15.167,5 + 5.172,5 + 2.177,5)$$
$$= 165,375$$

- Phương sai: $s^2 = 30,625$.

Tính trung bình, phương sai và độ lệch chuẩn bằng R

- Tạo file dữ liệu trong Excel (gồm 2 cột Height và Frequency) và lưu dưới dạng .csv (meanvar.csv) chung folder với file .RData



	A	B
1	Height	Frequency
2	157.5	8
3	162.5	10
4	167.5	15
5	172.5	5
6	177.5	2
7		

- Mở chương trình R

```
d <- read.table("meanvar.csv", header=TRUE, sep=",")
d2 <- rep(d$Height, d$Frequency)
multi.fun <- function(x){c(mean=mean(x), var=var(x), median=median(x),
 sd=sd(x))}
multi.fun(d2)
```

5.4 Phân phối lấy mẫu

5.4 Phân phối lấy mẫu

- Khi chọn một mẫu từ một tổng thể, các số đo mô tả tính được từ mẫu đó được gọi là các thống kê (statistic).
- Các thống kê thay đổi theo các mẫu khác nhau mà ta chọn, do đó chúng là các biến ngẫu nhiên.
- Phân phối xác suất của các thống kê được gọi là các phân phối lấy mẫu (sampling distribution)

Định nghĩa 5.26 Cho các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ nhận giá trị từ một tổng thể. Tập hợp các biến ngẫu nhiên $\{X_1, X_2, \ldots, X_n\}$ tạo thành một mẫu ngẫu nhiên có kích thước n nếu

1. Các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ có phân phối giống nhau.
2. Các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập với nhau.

Ví dụ 5.27 Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$.

- Mẫu ngẫu nhiên gồm 3 phần tử $\{X_1, X_2, X_3\}$
- Mẫu ngẫu nhiên này có thể nhận các giá trị $\{2, 4, 6\}, \{2, 6, 7\}, \{4, 6, 8\}$.

Định nghĩa 5.28 Một hàm số được tính từ mẫu ngẫu nhiên $\{X_1, X_2, \ldots, X_n\}$ được gọi là một thống kê (statistic).

Ta có một số thống kê

- $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (trung bình mẫu ngẫu nhiên)
- $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (phương sai mẫu ngẫu nhiên)

Định nghĩa 5.29 Phân phối lấy mẫu (sampling distribution) của một thống kê là phân phối xác suất của một thống kê.

Ví dụ 5.30 Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (không chọn lại). Tìm phân phối lấy mẫu của trung bình mẫu.

Giải. Có tất cả 15 mẫu ngẫu nhiên có kích thước bằng 2

Mẫu	Trung bình mẫu		Mẫu	Trung bình mẫu		Mẫu	Trung bình mẫu
2,4	3		4,6	5		6,7	6,5
2,6	4		4,6	5		6,8	7
2,6	4		4,7	5,5		6,7	6,5
2,7	4,5		4,8	6		6,8	7
2,8	5		6,6	6		7,8	7,5

Phân phối lấy mẫu của trung bình mẫu (kích thước mẫu bằng 2)

$\overline{X}$	3	4	$4,5$	5	$5,5$	6	$6,5$	7	$7,5$
$P(\overline{X} = \overline{x}_i)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

$\overline{X}$	3	4	4,5	5	5,5	6	6,5	7	7,5
$P(\overline{X} = \overline{x}_i)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

- Trung bình của $\overline{X}$

$$E(\overline{X}) = 3.\frac{1}{15} + 4.\frac{2}{15} + \cdots + 7,5.\frac{1}{15} = 5,5 = \mu \text{ (trung bình tổng thể)}$$

- Phương sai của $\overline{X}$

$$\text{Var}(\overline{X}) = \frac{1}{15}(3 - 5,5)^2 + \frac{2}{15}(4 - 5,5)^2 + \cdots + \frac{1}{15}(7,5 - 5,5)^2 = \frac{47}{30}$$

5.4.1 Phân phối lấy mẫu của trung bình mẫu

- Các phân phối lấy mẫu có thể bao gồm vô hạn mẫu có kích thước nào đó.
- Mọi thống kê mẫu đều có phân phối lấy mẫu.

Định lý 5.31 Cho $\{X_1, X_2, \ldots, X_n\}$ là một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể vô hạn có trung bình là μ và phương sai là σ^2. Khi đó

$$E(\overline{X}) = \mu \text{ và } \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Nhận xét: Nếu tổng thể có phân phối chuẩn $N(\mu, \sigma^2)$ thì phân phối lấy mẫu của trung bình có phân phối chuẩn $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Định lý 5.32 Cho $\{X_1, X_2, \ldots, X_n\}$ là một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể có kích thước N với trung bình tổng thể là μ và phương sai tổng thể là σ^2. Khi đó

$$E(\overline{X}) = \mu \text{ và } \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}.$$

Ví dụ 5.33 (Xem **Ví dụ 5.30**) Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (chọn không hoàn lại). Xác định độ lệch chuẩn của trung bình mẫu.

Giải. Ta có $N = 6; n = 2$, trung bình tổng thể $\mu = 5,5$ và phương sai tổng thể

$$\sigma^2 = \frac{1}{6}\left((2-5,5)^2 + (4-5,5)^2 + \cdots + (8-5,5)^2\right) = \frac{47}{12}.$$

Phương sai của trung bình mẫu

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1} = \frac{47/12}{2}.\frac{6-2}{6-1} = \frac{47}{30}.$$

Nhận xét:

- Nếu tổng thể không có phân phối chuẩn thì theo Định lý giới hạn trung tâm, phân phối lấy mẫu của trung bình mẫu $\overline{X}$ sẽ xấp xỉ phân phối chuẩn khi kích thước mẫu n đủ lớn ($n \geq 30$).
- Nếu tổng thể có phân phối chuẩn hay không chuẩn thì phân phối lấy mẫu của trung bình mẫu là xấp xỉ phân phối chuẩn khi kích thước mẫu lớn hơn hoặc bằng 30.

Trường hợp tổng thể có phân phối chuẩn nhưng không biết độ lệch chuẩn.

Định lý 5.34 Nếu $\overline{X}$ là trung bình của mẫu ngẫu nhiên có kích thước n lấy từ một tổng thể có phân phối chuẩn với trung bình tổng thể là μ và phương sai mẫu ngẫu nhiên $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ thì

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

là biến ngẫu nhiên có phân phối Student với bậc tự do $n - 1$.

Ví dụ 5.35 Giả sử lượng chất rắn lơ lửng trong nước thải của một công ty có phân phối chuẩn với trung bình 40 mg/l. Người ta lấy ngẫu nhiên 20 mẫu nước thải và thấy rằng độ lệch chuẩn của mẫu này là $s = 9,4$ mg/l. Xác suất lượng chất thải trung bình của 20 mẫu này nhỏ hơn 46 mg/l là bao nhiêu?

Giải. Theo đề bài, ta có $\mu = 40; n = 20$ và $s = 9,4$. Đặt $\overline{X}$ là trung bình của mẫu ngẫu nhiên. Khi đó

$$\begin{aligned} P(\overline{X} < 46) &= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < \frac{46 - 40}{9,4/\sqrt{20}}\right) \\ &= P(T < 2,855) \\ &= 1 - P(T \geq 2,855) \approx 1 - 0,01 = 0,99. \end{aligned}$$

5.4.2 Phân phối lấy mẫu của tỉ lệ mẫu

Định nghĩa 5.36 Trong một tổng thể có kích thước N, có y phần tử có tính chất $\mathcal{P}$ mà ta quan tâm. Giả sử X là số phần tử có tính chất $\mathcal{P}$ trong một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể. Khi đó tỉ lệ tổng thể

$$p = \frac{y}{N}$$

và tỉ lệ mẫu ngẫu nhiên (sample proportion)

$$\hat{p} = \frac{X}{n}$$

Tỉ lệ mẫu ngẫu nhiên $\hat{p}$ là một thống kê và do đó nó có phân phối lấy mẫu.

Ví dụ 5.37 Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (chọn không hoàn lại). Giả sử ta quan tâm các phần tử nhỏ hơn 5. Tìm phân phối lấy mẫu của tỉ lệ mẫu của mẫu ngẫu nhiên có 2 phần tử.

Giải. Có tất cả 15 mẫu ngẫu nhiên đơn giản có kích thước bằng 2

Mẫu	Tỉ lệ mẫu		Mẫu	Tỉ lệ mẫu		Mẫu	Tỉ lệ mẫu
2,4	1		4,6	1/2		6,7	0
2,6	1/2		4,6	1/2		6,8	0
2,6	1/2		4,7	1/2		6,7	0
2,7	1/2		4,8	1/2		6,8	0
2,8	1/2		6,6	0		7,8	0

Phân phối lấy mẫu của tỉ lệ mẫu ngẫu nhiên (kích thước mẫu bằng 2)

$\hat{p}$	0	$1/2$	1
$P(\hat{p} = \hat{p}_i)$	$\frac{6}{15}$	$\frac{8}{15}$	$\frac{1}{15}$

Phân phối lấy mẫu của tỉ lệ mẫu (kích thước mẫu bằng 2)

$\hat{p}$	0	$1/2$	1
$P(\hat{p} = \hat{p}_i)$	$\frac{6}{15}$	$\frac{8}{15}$	$\frac{1}{15}$

Khi đó

$$E(\hat{p}) = 0.\frac{6}{15} + \frac{1}{2}.\frac{8}{15} + 1.\frac{1}{15} = \frac{1}{3}$$

và

$$\mathrm{Var}(\hat{p}) = \frac{6}{15}(0 - \frac{1}{3})^2 + \frac{8}{15}.(\frac{1}{2} - \frac{1}{3})^2 + \frac{1}{15}.(1 - \frac{1}{3})^2 = \frac{4}{45}.$$

Định lý 5.38 Cho p là tỉ lệ của một tổng thể và $\hat{p}$ là tỉ lệ mẫu ngẫu nhiên có kích thước n. Khi đó

$$E(\hat{p}) = p$$

và

- nếu kích thước tổng thể là N hữu hạn thì

$$\mathrm{Var}(\hat{p}) = \frac{N-n}{N-1}.\frac{p(1-p)}{n}.$$

- nếu kích thước tổng thể là vô hạn thì

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

Ví dụ 5.39 (**Ví dụ 5.37**) Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (chọn không hoàn lại). Giả sử ta quan tâm các phần tử nhỏ hơn 5. Tìm phân phối lấy mẫu của tỉ lệ mẫu ngẫu nhiên có 2 phần tử.

- Trung bình của phân phối lấy mẫu của tỉ lệ mẫu

$$E(\hat{p}) = \frac{1}{3} = \frac{2}{6} = p.$$

- Phương sai phân phối mẫu của tỉ lệ mẫu

$$\mathrm{Var}(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n} = \frac{6-2}{6-1}.\frac{1/3(1-1/3)}{2} = \frac{4}{45}.$$

Định lý 5.40 Nếu kích thước mẫu n đủ lớn thì tỉ lệ mẫu ngẫu nhiên $\hat{p}$ có phân phối chuẩn $N\left(p, \dfrac{p(1-p)}{n}\right)$.

Nhận xét: Định lý 5.40 được áp dụng khi $np, n(1-p) \geq 5$.

Ví dụ 5.41 Chọn ngẫu nhiên 270 người từ một thành phố để ước tính tỉ lệ người không sử dụng mạng xã hội. Người ta thấy rằng tỉ lệ người dân không sử dụng mạng xã hội của thành phố này là 20%. Xác suất tỉ lệ mẫu này từ 16% đến 24% là bao nhiêu?

Giải. Ta có tỉ lệ tổng thể $p = 0,2$ và kích thước mẫu $n = 270$. Vì $np = 54, n(1-p) = 270(1-0,2) = 216 > 5$ nên $\hat{p} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$.

Phương sai của phân phối lấy mẫu của $\hat{p}$ là

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n} = \frac{0,2(1-0,2)}{270} = 0,00059.$$

Khi đó

$$\begin{aligned} P(0,16 \leq \hat{p} \leq 0,24) &= P\left(\frac{0,16-0,2}{\sqrt{0,00059}} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{0,24-0,2}{\sqrt{0,00059}}\right) \\ &= P\left(-1,67 \leq Z \leq 1,67\right) \\ &= 0,905 \end{aligned}$$

Bài tập

Bài 5.1 Cho một tổng thể gồm 5 phần tử $\{1, 2, 3, 4, 5\}$. Xét tất cả các mẫu ngẫu nhiên (không hoàn lại) có kích thước bằng 3 từ tổng thể. Tìm phân phối lấy mẫu của trung bình mẫu. Sau đó, tìm trung bình và phương sai của trung bình mẫu ngẫu nhiên có kích thước bằng 3. (Đáp số: $E(\overline{X}) = 3, \mathrm{Var}(\overline{X}) = 1/3$)

Bài 5.2 Giả sử điểm chỉ số IQ của người có phân phối chuẩn với trung bình 100 và độ lệch chuẩn 15. Chọn một mẫu ngẫu nhiên đơn giản từ 10 người. Tính xác suất điểm chỉ số IQ trung bình của 10 người này lớn hơn 110. (Đáp số: 0,0175)

Bài tập

Bài 5.3 Một nhà sản xuất cầu chì tuyên bố rằng với mức quá tải 20% cầu chì sẽ nổ. Chọn một mẫu ngẫu nhiên gồm 16 cầu chì trong số này đã bị quá tải 20% và thấy rằng thời gian chúng bị nổ có độ lệch chuẩn mẫu là 0,9 phút. Giả sử thời gian cầu chì bị nổ khi bị quá tải 20% có phân phối chuẩn với trung bình là 10 phút. Tính xác suất thời gian nổ trung bình của 16 cầu chì được chọn nhiều hơn 10,4 phút. (Đáp số: 0,05)

Bài 5.4 Người ta ước tính rằng 43% số người tốt nghiệp đại học tin thấy rằng giỏi tiếng Anh là một điều quan trọng. Chọn ngẫu nhiên một mẫu gồm 80 người đã tốt nghiệp đại học. Tính xác suất có hơn một nửa của mẫu này có niềm tin như trên. (Đáp số: 0,102)