CHƯƠNG 6. LÝ THUYẾT MẪU

"Trong một tương lai không xa kiến thức thống kê và tư duy thống kê sẽ trở thành một yếu tố không thể thiếu được trong học vấn của mỗi công dân, giống như là khả năng biết đọc, biết viết vậy"

H. G. WELLS (1920)

6.1. Mẫu số liệu, thống kê mô tả
6.2. Các phương pháp trình bày, biểu diễn mẫu
6.3. Các đặc trưng mẫu
6.4. Phân bố xác suất của các đặc trưng mẫu

Bài 6.1. MẪU SỐ LIỆU, THỐNG KÊ MÔ TẢ

1. Một số khái niệm cơ bản:

Trước khi đi đến các khái niệm cơ bản, ta xét ví dụ sau:

Để điều tra chiều cao trung bình của sinh viên Trường Đại học Công nghệ, người ta lập một danh sách bao gồm tất cả các sinh viên của Trường.

a) Tập hợp toàn bộ các sinh viên của Trường được gọi là ***tập hợp chính*** (hay còn gọi là ***tổng thể*** hay ***dân số***).
b) Mỗi sinh viên được gọi là một cá thể của tập chính.

c) Chiều cao của sinh viên được gọi một biến lượng. Giá trị của biến lượng này thay đổi từ cá thể này sang cá thể khác và được biểu diễn bởi 1 số thực.

d) Do số sinh viên của Trường là lớn, hơn nữa, khi số lượng cá thể đạt đến ngưỡng nào đó lượng thông tin tăng không đáng kể, nên ta không điều tra hết, mà chỉ chọn ra 1 tập hợp con để điều tra.

Tập hợp con được lấy ra để điều tra được gọi là một mẫu, số phần tử của một mẫu được gọi là cỡ mẫu.

Định nghĩa 1.

a) Tập hợp chính (hay dân số) S là tập tất cả các đối tượng có chung một tính chất nào đó mà chúng ta đang quan tâm.

b) Mỗi phần tử của tập hợp chính được gọi là một cá thể.

c) Một biến lượng X là một ánh xạ từ S lên R.

d) Việc chọn ra từ tập hợp chính một tập con nào đó gọi là phép lấy mẫu.

Tập hợp con này được gọi là một mẫu.

Số cá thể của mẫu được gọi là cỡ mẫu.

Ví dụ: lấy mẫu cỡ n=10 để xác định chiều cao TB của Lớp MAT 1101_6 năm học 2012-2013:

SV	1	2	3	4	5	6	7	8	9	10
H	175	172	175	170	164	169	167	161	170	165

Thể hiện 2

SV	1	2	3	4	5	6	7	8	9	10
H	184	180	170	170	172	175	172	170	173	170

lấy mẫu cỡ n=10 để xác định chiều cao TB của Lớp MAT 1101_3 năm học 2012-2013:

SV	1	2	3	4	5	6	7	8	9	10
H	162	175	170	169	172	170	167	172	165	167

Thể hiện 2

SV	1	2	3	4	5	6	7	8	9	10
H	172	169	170	173	172	174	170	166	163	167

Thể hiện 3

SV	1	2	3	4	5	6	7	8	9	10
H	172	174	165	165	175	172	170	171	170	171

Lớp MAT 1101_4 năm học 2013-2014

SV	1	2	3	4	5	6	7	8	9	10
H	168	170	177	171	165	156	168	175	173	165

2. Phương pháp chọn mẫu:

a. Nguyên tắc chọn mẫu:

Tuỳ theo từng yêu cầu của bài toán mà ta chọn một phương pháp hoặc kết hợp nhiều phương pháp chọn mẫu thích hợp. Sau đây là một số phương pháp chọn mẫu thường được sử dụng:

- Chọn mẫu ngẫu nhiên: Để chọn được mẫu ngẫu nhiên, người ta yêu cầu mỗi cá thể trong tổng thể đều có khả năng được lựa chọn như nhau.
- Chọn mẫu theo tỷ lệ: Khi tổng thể bao gồm số lượng lớn và phân thành nhiều bộ phận khác nhau, thì mẫu phải đại diện cho tất cả các bộ phận theo tỷ lệ của từng bộ phận.
- Chọn mẫu theo nhóm trội: Chúng ta quan tâm đến những nhóm tập trung cao dấu hiệu mà ta quan tâm để điều tra. Ví dụ, muốn điều tra việc sử dụng Internet để học tập, tra cứu thông tin, ta tập trung thành phần ở trí thức và sinh viên.

Ở trong giáo trình này, chúng ta tập trung vào mẫu ngẫu nhiên.

b. Định nghĩa 2: Mẫu ngẫu nhiên

Dãy các đại lượng ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập, cùng phân phối với đại lượng ngẫu nhiên X được gọi là mẫu ngẫu nhiên cỡ n từ đại lượng ngẫu nhiên X.

Kết quả của mỗi lần lấy mẫu cỡ n, ta được các giá trị cụ thể $x_1, x_2, \ldots, x_n$. Bộ giá trị $x_1, x_2, \ldots, x_n$ được gọi là 1 thể hiện của mẫu ngẫu nhiên cỡ n từ X.

Ví dụ1. Để xác định chiều cao và trọng lượng trung bình của SV lớp MAT 1101 1 (2011-2012), ta lấy mẫu cỡ 20. Kết quả cụ thể của phép lấy mẫu là 1 thể hiện của mẫu ngẫu nhiên (MNN) cỡ 20:

SV	1	2	3	4	5	6	7	8	9	10
H	165	163	170	170	170	168	170	162	163	168
W	52	51	51	52	52	66	67	45	50	58
SV	11	12	13	14	15	16	17	18	19	20
H	170	157	171	170	165	157	160	159	178	176
W	60	44	61	53	54	50	52	46	55	59

Để xác định chiều cao và trọng lượng trung bình của SV lớp MAT 2078 (2011-2012), ta lấy mẫu cỡ 20. Kết quả cụ thể của phép lấy mẫu là 1 thể hiện của mẫu ngẫu nhiên (MNN) cỡ 20:

SV	1	2	3	4	5	6	7	8	9	10
H	172	166	165	170	165	162	168	172	174	170
W	53	54	50	52	61	52	56	63	55	56
SV	11	12	13	14	15	16	17	18	19	20
H	178	162	168	157	174	160	162	165	164	167
W	67	48	47	45	70	50	50	60	59	53

Lớp MAT 1101_4 năm học 2012-2013

SV	1	2	3	4	5	6	7	8	9	10
H	170	164	168	164	168	168	166	170	170	175
W	60	52	55	50	54	48	49	63	53	57

SV	11	12	13	14	15	16	17	18	19	20
H	160	171	170	163	155	157	162	170	169	165
W	65	51	64	48	49	44	51	52	50	50

Chúng ta đã biết rằng, để chọn được mẫu ngẫu nhiên, người ta yêu cầu mỗi cá thể trong tổng thể đều có khả năng được lựa chọn như nhau.

3. Thống kê mô tả:

Thống kê mô tả được dùng để tổng hợp số liệu, mô tả các đặc trưng quan trọng của các biến lượng bằng các bảng, biểu, đồ thị, sơ đồ và các số trị.

Bài 6.2. Các phương pháp trình bày, biểu diễn mẫu

Giả sử ta có dãy các số liệu quan sát $x_1, x_2, \ldots, x_n$ của một ĐLNN X nào đấy. Giả sử X có hàm phân phối F(x). Ta cần biết các thông tin về F(x), chẳng hạn, giá trị trung bình, phương sai, các mô men, dáng điệu của hàm mật độ f(x), hàm phân phối F(x).

SV	1	2	3	4	5	6	7	8	9	10
H	165	163	170	170	170	168	170	162	163	168
W	52	51	51	52	52	66	67	45	50	58
SV	11	12	13	14	15	16	17	18	19	20
H	170	157	171	170	165	157	160	159	178	176
W	60	44	61	53	54	50	52	46	55	59

Ví dụ

Lớp MAT 1101_4 năm học 2012-2013

SV	1	2	3	4	5	6	7	8	9	10
H	170	164	168	164	168	168	166	170	170	175
W	60	52	55	50	54	48	49	63	53	57

SV	11	12	13	14	15	16	17	18	19	20
H	160	171	170	163	155	157	162	170	169	165
W	65	51	64	48	49	44	51	52	50	50

Bước 1. Ta liệt kê ra các giá trị khác nhau và đếm số lần xuất hiện các giá trị này. Tiếp theo, sắp xếp các giá trị này từ bé tới lớn. Giả sử, sau khi sắp xếp lại ta được $x_{(1)}< x_{(2)}<\ldots<x_{(m)}$, và giả sử $x_{(k)}$ xuất hiện r_k lần (k=1, 2, …, n), trong đó, $r_1+r_2+...+r_m=n$.

Giá trị n được gọi là cỡ mẫu. Các số $r_1, r_2, \ldots, r_m$ được gọi là tần số xuất hiện của các biến cố $\{X=x_1\}, \{X=x_2\}, \ldots, \{X=x_m\}$ tương ứng.

Tần suất của các biến cố $\{X=x_1\}, \{X=x_2\}, \ldots, \{X=x_m\}$ được tính tương ứng:

$f_1=r_1/n, f_2=r_2/n,\ldots, f_m=r_m/n$

(được gọi là tần suất xuất hiện biến cố $\{X=x_1\}, \{X=x_2\}, \ldots, \{X=x_m\}$ tương ứng).

i	1	2	3	4	5	6	7	8	9	10
x_i	176	172	162	160	165	168	165	153	155	160
i	11	12	13	14	15	16	17	18	19	20
xi	155	164	170	158	160	155	160	164	168	165

Ví dụ: Thống kê chiều cao của SV Lớp K59CA

165	172	165	160	169	180	170	164	177	165
169	172	165	165	165	172	167	171	174	159
174	163	164	180	172					

Bảng tần số, tần suất

159	160	163	164	165	167	169	170	171	1
1	1	1	2	6	1	2	1	1	
0.04	0.04	0.04	0.08	0.24	0.04	0.08	0.04	0.04	0

Trong thực hành, ta thường phân chia số liệu quan sát thành các khoảng (đều nhau hoặc không đều nhau), rồi tính tần số và tần suất cho mỗi khoảng.
Nếu số liệu này là kết quả đo chiều cao của người Việt, ta cần biết chiều cao trung bình, độ lệch chuẩn về chiều cao, … Việc phân tích như thế rất cần thiết cho thực tế. Chẳng hạn, ta cần biết có bao nhiêu phần trăm người Việt có chiều cao từ 1,65m đến 1,75m.

Bước 2. Vẽ biểu đồ, tổ chức đồ

Đối với số liệu chưa phân khoảng

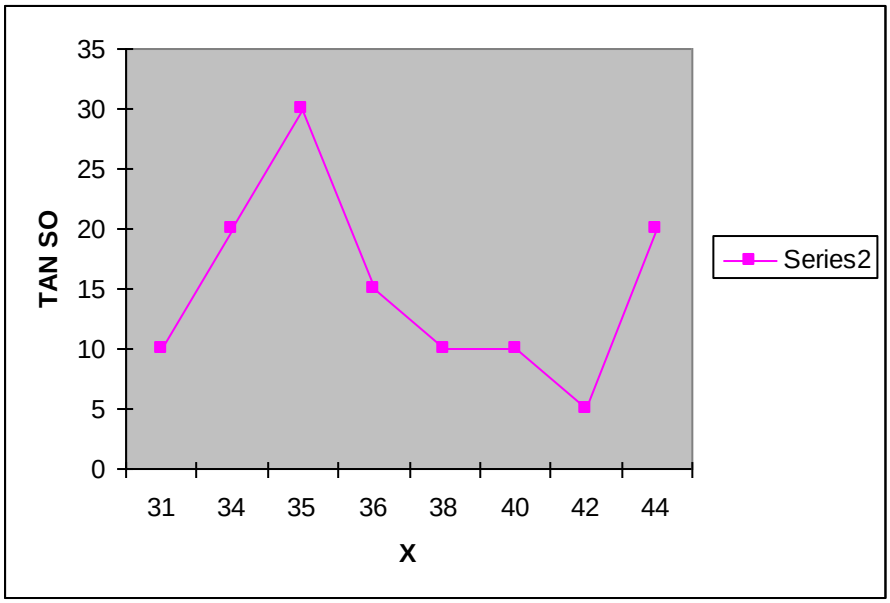
- Chấm trên mặt phẳng các điểm (x_k, r_k), k=1, 2, …, m.
- Nối các điểm $(x_k, 0)$ với các điểm (x_k, r_k), ta được biểu đồ tần số hình gậy.
- Nối liên tiếp điểm (x_k, r_k) với (x_{k+1}, r_{k+1}), ta được biểu đồ đa giác tần số.

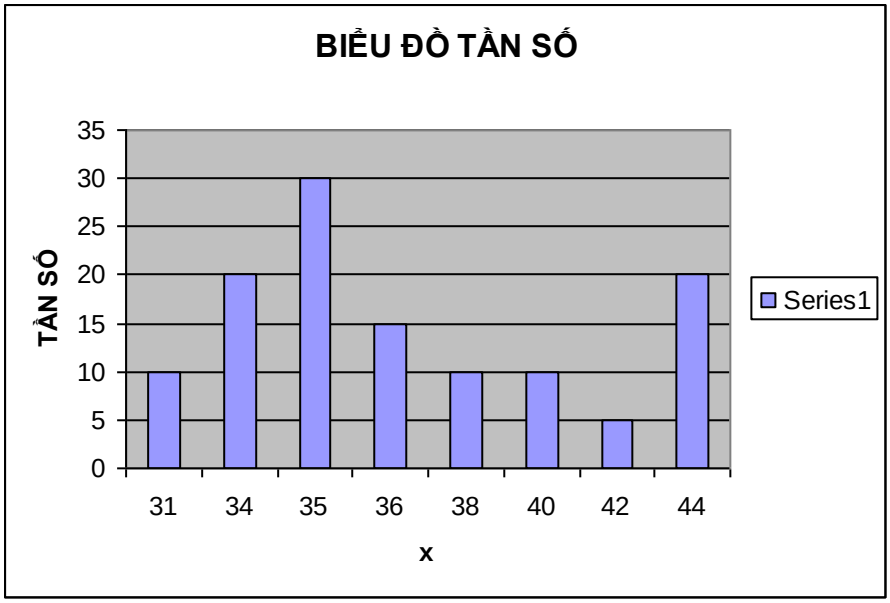
Tương tự,

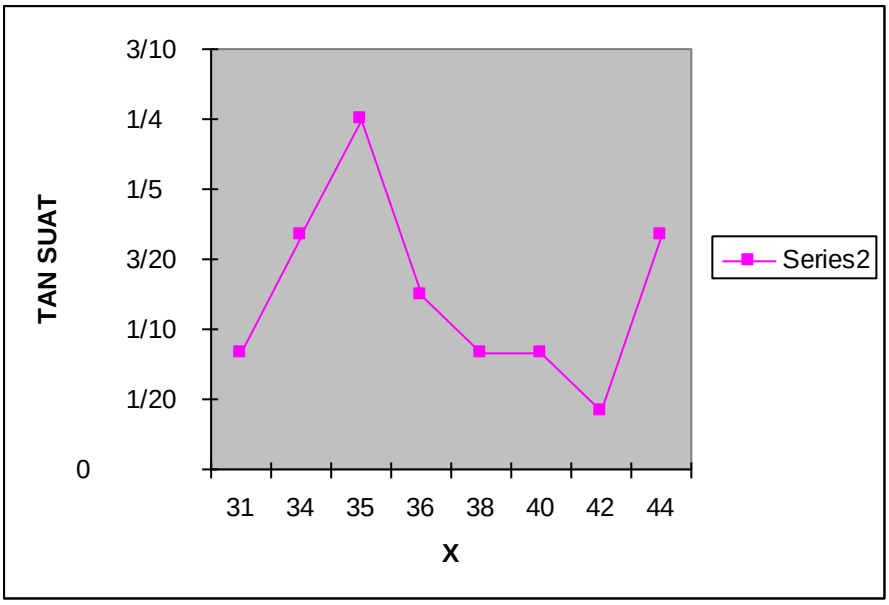
- Chấm trên mặt phẳng các điểm (x_k, f_k), k=1, 2, …, m.
- Nối các điểm $(x_k, 0)$ với các điểm (x_k, f_k), ta được biểu đồ tần suất hình gậy.

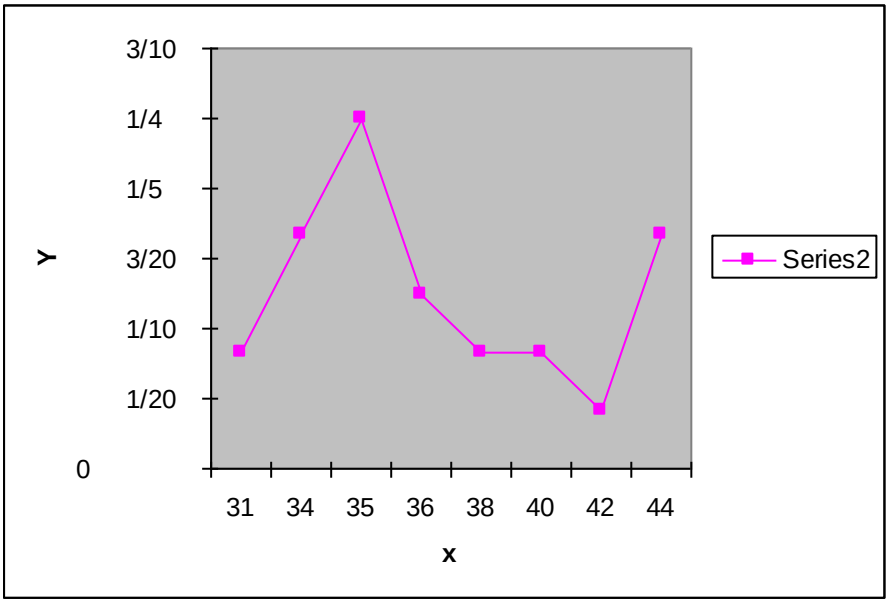
- Nối liên tiếp điểm (x_k, f_k) với (x_{k+1}, f_{k+1}), ta được biểu đồ đa giác tần suất.

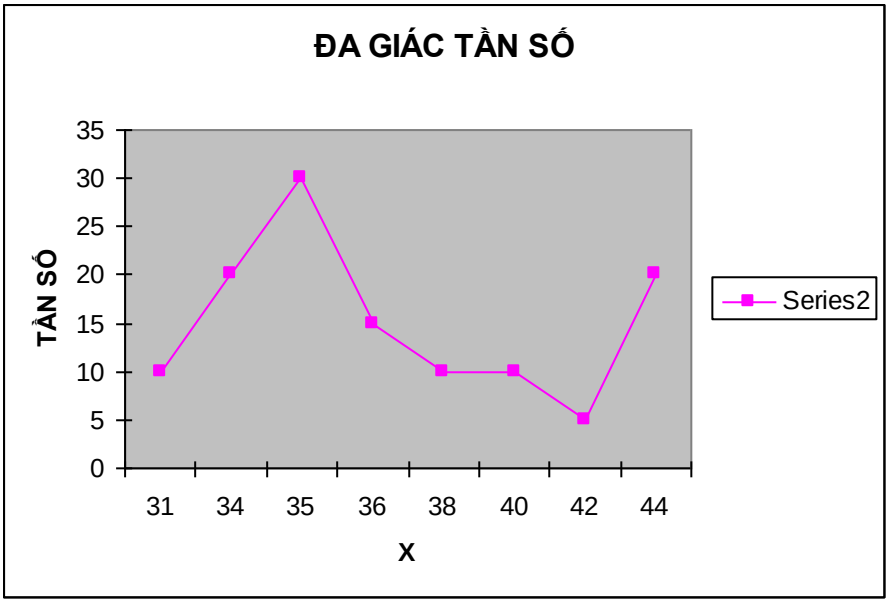
X	31	34	35	36	38	40	42	44
Tần số	10	20	30	15	10	10	5	20
Tần suất	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{2}{12}$

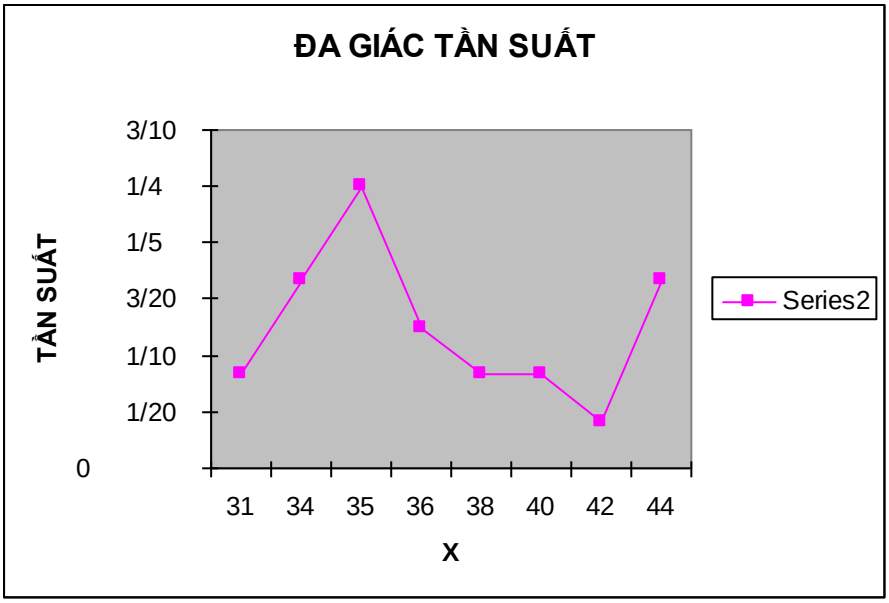





TAN SUAT
3/10
1/4
1/5
3/20
1/10
1/20
0
31
34
35
36
38
40
42
44
X
Series2


Y
3/10
1/4
1/5
3/20
1/10
1/20
0
31
34
35
36
38
40
42
44
x
Series2


ĐA GIÁC TẦN SỐ
TẦN SỐ
35
30
25
20
15
10
5
0
31
34
35
36
38
40
42
44
X
Series2


ĐA GIÁC TẦN SUẤT
TẦN SUẤT
3/10
1/4
1/5
3/20
1/10
1/20
0
31
34
35
36
38
40
42
44
X
Series2

Tổ chức đồ tần số - tổ chức đồ tần suất:

Đối với số liệu đã phân chia thành các khoảng có độ dài bằng nhau:

- Trên mỗi khoảng ta dựng hình chữ nhật có chiều cao bằng tần số (hay tần suất) tương ứng với khoảng đó.
- Tô đậm hoặc kẻ chéo bằng các đường song song các hình chữ nhật này ta thu được tổ chức đồ tần số (hay tổ chức đồ tần suất).

Đối với số liệu đã phân chia thành các khoảng có độ dài không bằng nhau.

- Trên mỗi hình chữ nhật có chiều cao bằng $y_k=\lambda r_k/l_k$ (hay $y_k=\lambda f_k/l_k$).

trong đó l là chiều dài của khoảng, l là số tuỳ chọn, chẳng hạn l=1, sao cho hình vẽ thu được dễ coi.

- Tô đậm hoặc kẻ chéo bằng các đường song song các hình chữ nhật này ta thu được tổ chức đồ tần số (hay tổ chức đồ tần suất).

Ví dụ sau minh hoạ những điều vừa trình bày ở trên:

Khoảng	Tần số	Tần suất
26,5-48,5	2	0,04
48,5-70,5	8	0,16
70,5-92,5	12	0,24
92,5-114,5	12	0,24
114,5-136,5	8	0,16
136,5-158,5	7	0,14
158,5-180,5	1	0,02
180,5-202,5	1	0,02

Tổng	51	1

Bước 3. Tính các đặc trưng mẫu
Trung bình mẫu tính theo công thức:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{k=1}^{m} r_k x_k}{\sum_{k=1}^{m} r_k}$$

Phương sai mẫu tính theo công thức:

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

Độ lệch mẫu tính theo công thức:

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

Bài 6.3. Các đặc trưng mẫu

Trong phần trên ta đã giới thiệu cách tính 3 đặc trưng mẫu là: trung bình mẫu, phương sai mẫu và độ lệch chuẩn mẫu. Sau đây, chúng ta giới thiệu một số đặc trưng quan trọng khác:

1. Trung vị (Median): Ký hiệu là Med(X)

Với một mẫu, trung vị là là giá trị nằm giữa dãy giá trị quan trắc theo thứ tự tăng hay giảm.

Nếu dãy quan trắc có 2n+1 số liệu sắp xếp theo thứ tự tăng dần thì giá trị thứ n+1 là trung vị, nếu dãy quan trắc gồm 2n số liệu thì trung vị là giá trị trung bình của giá trị thứ n và n+1.

Nếu các giá trị x_i có tần số r_i, gọi k là chỉ số bé nhất để $r_1+r_2+...+r_k \geq n/2$. Khi đó ta định nghĩa $Med(X)=x_k$.

Ví dụ: Cho bảng phân bố tần số của đại lượng X như sau:

X	0	1	2	3	4	5	6	7	8	9	10	11
r_i	6	15	43	53	85	72	55	33	18	10	7	3

Kích thước mẫu là 400

Hãy tính trung bình mẫu và trung vị.

Giải

Trung bình mẫu $\bar{x} = 4.645$

Ta thấy số giá trị của mẫu bé hơn hay bằng 3 là:

6+15+43+53=117<200

Số giá trị của mẫu bé hơn hay bằng 4 là:

6+15+43+53+85=202>200

Vậy Med(X)=4.

Trong trường hợp mẫu được cho dưới dạng phân bố ghép lớp ta định nghĩa trung vị như sau:

Giả sử ta có m khoảng với các điểm chia là:

$$a_0<a_1< \ldots <a_m$$

$C_1= [a_0, a_1)$, $C_2= [a_1, a_2)$, …, $C_m= [a_{m-1}, a_m]$. Trong đó khoảng C_i có tần số r_i.

Khoảng C_k được gọi là khoảng trung vị nếu k là chỉ số bé nhất sao cho $r_1+r_2+\ldots+r_k \geq n/2$.

Số trung vị Med(X) là số mà tại đó đường thẳng x=Med(X) chia đôi diện tích của tổ chức đồ tần số (tần suất).

$Med(X)=a_{k-1}+[(n/2)–(r_1+r_2+\ldots+r_{k-1})]/h_k$, h_k – là chiều cao của hình chữ nhật thứ k.

2. Mode: Ký hiệu là Mod(X)

Nếu mẫu được cho dưới dạng bảng phân bố tần số thì mode là giá trị có tần số cực đại.

Trường hợp mẫu được cho dưới dạng bảng phân bố ghép lớp, khoảng mode(X) là khoảng có chiều cao của hình chữ nhật dựng trên khoảng đó là lớn nhất.

Bài 6.4. Phân bố của các đặc trưng mẫu

Giá trị kỳ vọng của trung bình mẫu được cho bởi:

$$E[M_n] = E\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} E[X_j] = \mu \qquad (5.17)$$

do $E[X_j] = E[X] = \mu$ với $\forall$ j. Như vậy trung bình mẫu bằng $E[X] = \mu$ về giá trị trung bình. Vì lý do này, chúng ta nói rằng trung bình mẫu là ước lượng không chệch cho μ.

Hệ thức (5.17) suy ra rằng sai số trung bình bình phương của trung bình mẫu xung quanh μ là bằng phương sai của M_n, nghĩa là,

$E[(M_n - \mu)^2] = E[(M_n - E[M_n])^2]$.

Chú ý rằng $M_n = S_n/n$ trong đó $S_n = X_1 + X_2 + \ldots + X_n$. Từ hệ thức (5.4),

$VAR[S_n] = n\,VAR[X_j] = n\sigma^2$, do X_j là các biến ngẫu nhiên độc lập cùng phân phối. Như vậy,

$$\mathbf{VAR}[M_n] = \frac{1}{n^2}\mathbf{VAR}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Mệnh đề : Giả sử X_j với j=1, 2, ... là các biến ngẫu nhiên Gauss độc lập cùng phân phối, với kỳ vọng μ chưa biết và phương sai σ^2 đã biết. Khi đó :

1) M_n là biến ngẫu nhiên Gauss với kỳ vọng μ và phương sai σ^2/n.

2) $(n-1)S_n^2/\sigma^2$ là biến ngẫu nhiên χ^2 với $n-1$ bậc tự do.

3) $$W = \frac{M_n - \mu}{S_n/\sqrt{n}} = \frac{\sqrt{n}(M_n - \mu)/\sigma}{S_n/\sigma} = \frac{(M_n - \mu)(\sigma/\sqrt{n})}{\left\{\left[(n-1)S_n^2/\sigma^2\right]/(n-1)\right\}^{1/2}}.$$

Có phân phối Student với (n-1) bậc tự do với hàm mật độ:

$$f_{n-1}(y) = \frac{r(n/2)}{r((n-1)/2)\sqrt{\pi(n-1)}}\left(1 + \frac{y^2}{n-1}\right)^{-n/2}$$

Bảng 5.2 Thể hiện các giá trị của $z_{\alpha/2,\, n-1}$ đối với các giá trị đặc thù của $1 - \alpha$ và n.

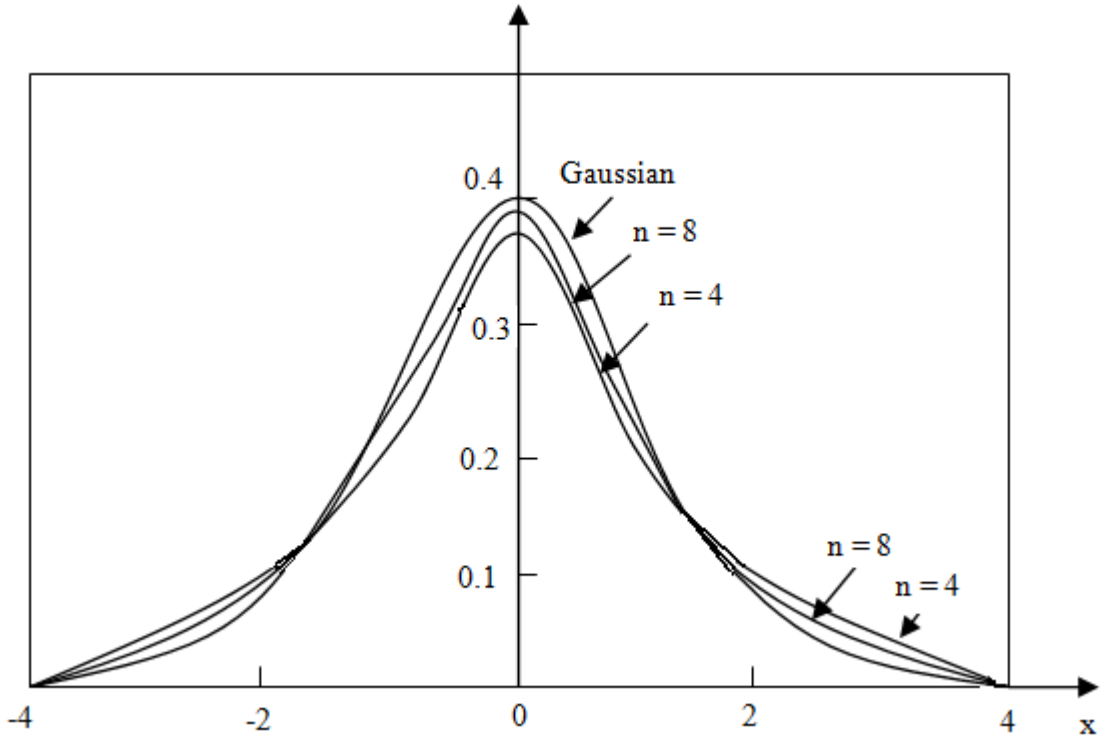
Bảng 5.2

Các giá trị của để tính các khoảng tin cậy trong phương trình (5.43)

	1 – α		
n – 1	**0.90**	**0.95**	**0.99**
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.064
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
15	1.753	2.131	2.947

20	1.725	2.806	2.845
30	1.697	2.042	2.750
40	1.684	2.021	2.704
60	1.671	2.000	2.660
∞	1.645	1.960	2.576

HINH 5.7
Hàm mật độ phân phối
Gauss và Hàm mật độ
phân phối Student với n=4 và 5



(2) : Phân phối được đặt tên bởi W. S. Gosset, người xuất bản dưới cái tên "A. Student".

Phép kiểm nghiệm khi-bình phương bao gồm hai yếu tố trên và tiến hành như sau:

1. Phân hoạch không gian mẫu S_X thành K khoảng không giao nhau.
2. Tính xác suất b_k để kết cục rơi vào khoảng thứ k với giả thiết X có hàm phân phối giả định. Khi đó $m_k = nb_k$ là số kết cục kỳ vọng rơi vào khoảng thứ k trong n lần lặp lại thí nghiệm. (Để nhận thấy điều này chúng ta tưởng tượng thực hiện phép thử Bernoulli mà ở đó "sự thành công" tương ứng với kết cục thuộc vào khoảng thứ k).
3. Thống kê khi-bình phương được xác định theo trọng số sự khác biệt giữa số kết cục quan sát được, N_k, rơi vào khoảng thứ k và giá trị được kỳ vọng m_k:

$$D^2 = \sum_{k=0}^{K} \frac{(N_k - m_k)^2}{m_k}. \qquad (3.75)$$

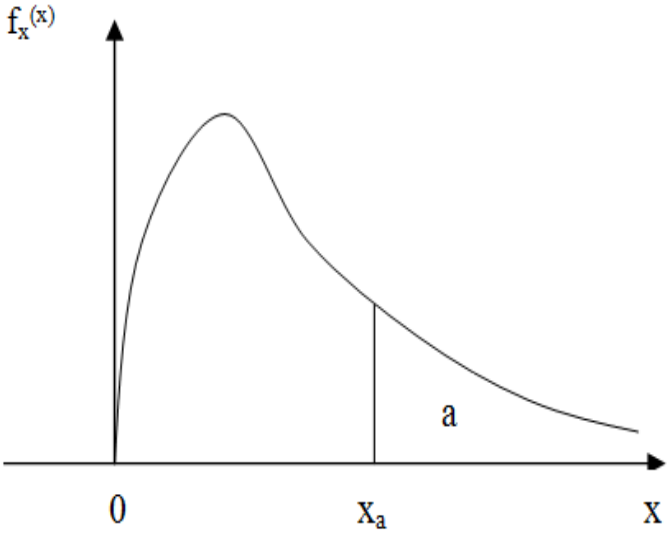
4. Nếu sự phù hợp là tốt khi đó D^2 sẽ nhỏ. Do vậy giả thuyết bị bác bỏ nếu D^2 đủ lớn; nghĩa là, nếu $D^2 \geq t_\alpha$, ở đây t_α là ngưỡng được xác định bởi mức ý nghĩa của tính chất.

Phép kiểm nghiệm khi-bình phương được đặt cơ sở trên thực tế là với n lớn, biến ngẫu nhiên D^2 có hàm mật độ xác suất xấp xỉ hàm mật độ khi-bình phương với $K - 1$ bậc tự do. Như vậy ngưỡng t_α có thể được tính bằng cách tìm điểm mà tại đó :

$$P[X \geq t_\alpha] = \alpha,$$

Ở đây X là biến ngẫu nhiên khi-bình phương với $K - 1$ bậc tự do (xem Hình 3.25). Các ngưỡng với mức ý nghĩa 1%

và 5% và các bậc tự do khác nhau được cho trong Bảng 3.5.



HÌNH 3.25
Ngưỡng trong tiêu chuẩn khi – bình phương được lấy sao cho $P[D^2 > t_\alpha] = \alpha$

BẢNG 3.5
Các giá trị ngưỡng của tiểu chuẩn khi – bình phương

K	**5%**	**1%**
1	3.84	6.63
2	5.99	9.21
3	7.81	11.35
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.76

12	21.03	26.22
13	22.36	27.69
14	23.69	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.51
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
25	37.65	44.31
30	43.77	50.89

VÍ DỤ 3.44

Biểu đồ trên tập {0, 1, 2, …, 9} trong Hình 3.23 nhận được bằng việc lấy số cuối cùng của 114 số điện thoại trong một cột trong danh bạ điện thoại. Số liệu quan trắc có phù hợp với giả thuyết chúng có hàm xác suất rời rạc đều hay không?

Nếu các biến cố có phân phối đều, khi đó mỗi số có xác suất bằng 1/10. Giá trị kỳ vọng của số lần xảy ra mỗi biến cố trong 114 phép thử là 114/10 = 11,4. Khi đó thống kê khi-bình phương là:

$$D^2 = \frac{(17-11.4)^2}{11.4} + \frac{(16-11.4)^2}{11.4} + \ldots + \frac{(7-11.4)^2}{11.4}$$

$$= 9.51.$$

Số bậc tự do là $K - 1 = 10 - 1 = 9$, bởi vậy từ Bảng 3.5 ngưỡng với mức ý nghĩa 1% là 27.1. D^2 không vượt quá ngưỡng, do vậy

chúng ta kết luận rằng số liệu phù hợp với biến ngẫu nhiên phân phối đều.

VÍ DỤ 3.45

Biểu đồ trong Hình 3.24 nhận được bởi việc tạo ra 1000 mẫu từ một chương trình được thiết kế để tạo ra biến ngẫu nhiên có phân phối mũ với tham số 1. Biểu đồ nhận được bởi việc chia nửa dương của đường thẳng thực thành 20 khoảng có cùng độ dài 0.2. Giá trị đúng được cho bởi Bảng 3.6. Biểu đồ thứ hai cũng được xây dựng khi sử dụng 20 khoảng có xác suất bằng nhau. Các số của biểu đồ này được cho bởi Bảng 3.7.

Từ Bảng 3.5 chúng ta tìm được ngưỡng với mức ý nghĩa 5% là 30.1. Các giá trị khi-bình phương cho các biểu đồ tương ứng là 14.2 và 11.6 một cách. Cả hai biểu đồ chuyển tiêu chuẩn phù hợp tốt vào trường hợp này, nhưng có vẻ như phương pháp chọn các khoảng ảnh hưởng đến giá trị của độ đo khi-bình phương.

Ví dụ 3.45 chỉ ra rằng có nhiều cách chọn các khoảng để phân hoạch và điều này có thể dẫn tới những kết quả khác nhau. Những qui tắc quan trọng sau được đề nghị: Thứ nhất, độ rộng có thể của các khoảng nên chọn sao cho chúng đồng xác suất. Thứ hai, các khoảng nên được chọn sao cho giá trị kỳ vọng của các kết cục trong mỗi khoảng lớn hơn hoặc bằng 5. Điều này hiệu chỉnh sự chính xác của xấp xỉ hàm phân phối của D^2 bởi hàm phân phối khi-bình phương.

Chúng ta có được lý luận trên do đã giả thiết rằng phân phối giả định được xác định hoàn toàn. Trong trường hợp điển hình, một hoặc hai tham số của phân phối, nghĩa là giá trị trung bình và phương sai, được ước lượng từ dữ liệu. Thường là nếu có *r* tham số của hàm phân phối được ước lượng từ dữ liệu, thì D^2 được xấp xỉ tốt hơn bởi phân phối khi-bình phương với $K - r - 1$ bậc tự do. Như vậy, mỗi một tham số được ước lượng làm giảm 1 bậc tự do.

BẢNG 3.6

Phép kiểm nghiệm khi-bình phương cho biến ngẫu nhiên mũ, Các khoảng độ dài bằng nhau.

Khoảng	**Giá trị quan trắc O**	**Giá trị kỳ vọng E**	**$(O - E)^2 / E$**
0	190	181.3	0.417484
1	144	148.4	0.130458
2	102	121.5	3.129629
3	96	99.5	0.123115
4	86	81.44	0.255324
5	67	66.7	0.001349
6	59	54.6	0.354578
7	43	44.7	0.064653
8	51	36.6	5.665573
9	28	30	0.133333
10	28	24.5	0.5
11	19	20.1	0.060199
12	15	16.4	0.119512
13	12	13.5	0.166666
14	11	11	0
15	7	9	0.444444
16	9	7.4	0.345945
17	5	6	0.166666
18	8	5	1.8

>19	20	22.4	0.257142
		Giá trị khi-bình phương =	14.13607

BẢNG 3.7
Phép kiểm nghiệm khi-bình phương cho biến ngẫu nhiên mũ. Các khoảng đồng xác suất.

Khoảng	**Quan trắc O**	**Kỳ vọng E**	**$(O - E)^2 / E$**
0	49	50	0.02
1	61	50	2.42
2	50	50	0
3	50	50	0
4	40	50	2
5	52	50	0.08
6	48	50	0.08
7	40	50	2
8	45	50	0.5
9	46	50	0.32
10	50	50	0
11	51	50	0.02
12	55	50	0.5
13	49	50	0.02
14	54	50	0.32
15	52	50	0.08
16	62	50	2.88
17	46	50	0.32
18	49	50	0.02
19	51	50	0.02
		Giá trị khi-bình phương =	11.6

VÍ DỤ 3.46

Biểu đồ trong Bảng 3.8 được thông báo bởi Rutherford, Chadwick, và Ellis trong một bài báo nổi tiếng xuất bản năm 1920. Số các hạt

được phát ra bởi một chất phóng xạ trong chu kỳ thời gian 7.5 giây đã được đếm. Tổng số có 2608 chu kỳ được quan trắc. Giả định rằng số các hạt phát ra trong một chu kỳ thời gian là một biến ngẫu nhiên với phân phối Poisson. Hãy thực hiện phép kiểm nghiệm phù hợp tốt khi-bình phương.

Trong trường hợp này giá trị trung bình của phân phối khi-bình phương chưa biết, mà được ước lượng từ dữ liệu bằng 3.870. D^2 với 12 – 1 –1 = 10 bậc tự do là 12.94. Ngưỡng của mức ý nghĩa 1% là 23.2. D^2 không vượt quá giá trị này, bởi vậy chúng ta có thể kết luận rằng dữ liệu phù hợp tốt với phân phối Poisson.

BẢNG 3.8

Phép kiểm nghiệm khi-bình phương cho biến ngẫu nhiên Poisson

Số	**Quan trắc O**	**Kỳ vọng E**	**$(O – E)^2 /E$**
0	57.00	54.40	0.12
1	203.00	210.50	0.27
2	383.00	407.40	1.46
3	525.00	525.00	.00
4	532.00	508.40	1.10
5	408.00	393.50	0.53
6	273.00	253.80	1.45
7	139.00	140.30	0.01
8	45.00	67.80	7.67
9	27.00	29.20	0.17

10	10.00	11.30	0.15
>11	6.00	5.80	0.01
			12.94

Dựa theo H. Cramer, *Mathematical Methods of Statistics*, Princeton University, Princeton, N. J., 1946, p. 436.