PHẦN 2: THỐNG KÊ

1

CHƯƠNG 5: LÝ THUYẾT MẪU

2

1. MẪU VÀ PHƯƠNG PHÁP MẪU

Giả sử ta cần nghiên cứu một tập hợp có rất nhiều phần tử, vì một số lý do mà ta *không thể khảo sát toàn bộ tập lớn* này (khảo sát *tất cả* các phần tử), nhưng ta lại *muốn có kết quả trên tập lớn*. Ta có thể giải quyết như sau: từ tập hợp lớn lấy ra một tập hợp nhỏ hơn để nghiên cứu, ta thu được kết quả trên tập nhỏ, từ kết quả trên tập nhỏ ta suy ra kết quả cho tập lớn. Phương pháp làm việc như vậy gọi là *phương pháp mẫu*. Tập lớn gọi là *tổng thể* hay *đám đông*, số phần tử của tập lớn gọi là *kích thước tổng thể/đám đông*, ký hiệu là *N*. Tập nhỏ gọi là *mẫu*, số phần tử của mẫu gọi là *kích thước mẫu* hay *cỡ mẫu*, ký hiệu *n*.

3

Một số lý do không thể nghiên cứu toàn bộ tổng thể:

- *Giới hạn về thời gian, tài chính...*

Thí dụ muốn khảo sát xem chiều cao trung bình của thanh niên Việt Nam hiện nay có tăng lên so với trước đây không, ta phải đo chiều cao của toàn bộ thanh niên Việt nam (giả sử xấp xỉ N= 40 triệu người), điều này tuy làm được nhưng rõ ràng tốn nhiều thời gian, tiền bạc, công sức....

Ta có thể khảo sát khoảng 1 triệu thanh niên và từ chiều cao trung bình của n= 1 triệu người này, ta suy ra chiều cao trung bình của toàn bộ thanh niên *VN*.

4

Một số lý do không thể nghiên cứu toàn bộ tổng thể:

- *Phá vỡ tổng thể nghiên cứu.*

Thí dụ ta cất vào kho N= 10000 hộp sản phẩm, muốn biết tỷ lệ hộp hư trong kho sau 1 thời gian bảo quản. Ta phải kiểm tra từng hộp để xác định số hộp hư M= 300, thì tỷ lệ hộp hư trong kho là M/N.

Một sản phẩm sau khi được kiểm tra thì bị mất phẩm chất, khi ta kiểm tra xong cả kho thì cũng "tiêu" luôn cái kho!

Ta có thể lấy ngẫu nhiên n= 100 hộp ra kiểm tra, giả sử có m= 9 hộp hư. Từ tỷ lệ hộp hư 9% ta suy ra tỷ lệ hộp hư của cả kho.

5

Một số lý do không thể nghiên cứu toàn bộ tổng thể:

- *Không xác định được chính xác tổng thể.*

Thí dụ muốn khảo sát xem tỷ lệ những người bị nhiễm *HIV* qua đường tiêm chích ma túy là bao nhiêu phần trăm. Trong tình huống này thì tổng thể chính là những người bị nhiễm *HIV*, nhưng ta không thể xác định chính xác tất cả những người bị nhiễm *HIV* vì chỉ có những người tự nguyện đến trung tâm xét nghiệm, bệnh viện thì mới biết được, còn những người không đi xét nghiệm thì không biết được.

Do đó ta *chỉ biết một phần của tổng thể*, là những người đã đi xét nghiệm. Ngoài ra số người bị nhiễm mới HIV và bị chết do HIV có thể thay đổi từng giây nên *số phần tử của tổng thể thay đổi từng giây.*

6

- Muốn từ kết quả của mẫu suy ra kết quả cho tổng thể tốt thì mẫu phải *đại diện được* cho tổng thể, muốn vậy thì mẫu phải được *lấy một cách ngẫu nhiên*. Trong phạm vi bài giảng này không đề cập đến *kỹ thuật lấy mẫu* (mẫu giản đơn, mẫu hệ thống, mẫu chùm, mẫu phân tổ, mẫu nhiều cấp ...).
- Có 3 *cách lấy mẫu* thông dụng:
  - C1: Lấy ngẫu nhiên n phần tử: *phân phối siêu bội*
  - C2: Lấy lần lượt n phần tử
  - C3: Lấy có hoàn lại n phần tử: *phân phối nhị thức*

* Về mặt xác suất: c1 = c2

* Khi n << N thì c1 xấp xỉ c3

- Ta quy ước là ***mẫu được lấy theo cách có hoàn lại***.
- Mẫu gồm có: *mẫu ngẫu nhiên* và *mẫu cụ thể*. Cần phân biệt rõ mẫu ngẫu nhiên và mẫu cụ thể.

7

- Tổng thể được đặc trưng bởi *dấu hiệu nghiên cứu X*, là một đại lượng ngẫu nhiên. Do đó khi nói về *X* tức là nói về tổng thể.
- *Mẫu ngẫu nhiên* (có cỡ mẫu *n*) được ký hiệu $W_X=(X_1,...,X_n)$ là một véctơ có *n* thành phần, mỗi thành phần X_i là một *ĐLNN*. Các *ĐLNN* này độc lập nhau và có cùng quy luật phân phối giống với *X*.
- *Mẫu cụ thể* (có cỡ mẫu *n*) được ký hiệu $W_x= (x_1,...,x_n)$ là một véctơ có *n* thành phần, mỗi thành phần x_i là một *giá trị (con số) cụ thể*.

- Ứng với *một mẫu ngẫu nhiên* thì có *nhiều mẫu cụ thể* tương ứng với kết quả của các phép thử ngẫu nhiên khác nhau.

8

II. Các đặc trưng số cơ bản của tổng thể và mẫu:

- *Ta xét tổng thể về mặt định lượng:* Tổng thể được đặc trưng bởi dấu hiệu nghiên cứu X, X là ĐLNN.

  Ta có E(X)=μ là *trung bình tổng thể*. Var(X)= σ^2 là *phương sai tổng thể*, và σ là *độ lệch chuẩn của tổng thể*.

- *Ta xét tổng thể về mặt định tính:* tổng thể có kích thước N, trong đó có M phần tử có *tính chất A quan tâm*. Ta có p= M/N gọi là *tỷ lệ tổng thể*.

- Tương tự, ta cũng có trung bình mẫu $\bar{x}$, phương sai mẫu (đã hiệu chỉnh) s^2, tỷ lệ mẫu f.

9

Các đặc trưng số cơ bản của mẫu (dạng cụ thể):

- *Định lượng*:
- Trung bình mẫu: $\bar{x}=\frac{1}{n}\sum x_i$
- Phương sai mẫu (chưa hiệu chỉnh): $\hat{s}^2=\frac{1}{n}\sum(x_i-\bar{x})^2$
- Phương sai mẫu (đã hiệu chỉnh): $s^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2$
- Độ lệch chuẩn mẫu (chưa hiệu chỉnh): $\hat{s}=\sqrt{\hat{s}^2}$
- ***Độ lệch chuẩn*** mẫu (đã hiệu chỉnh): $s=\sqrt{s^2}$
- Ta có: $s=\hat{s}\sqrt{\frac{n}{n-1}}$
- ***Sai số chuẩn*** mẫu (đã hiệu chỉnh): $\frac{s}{\sqrt{n}}$

10

Các đặc trưng số cơ bản của mẫu (dạng cụ thể):

- *Định tính*:

  Trong thực hành ta xác định tỷ lệ mẫu:

  f = m/n

  Với:

  n: cỡ mẫu

  m: số phần tử có tính chất A quan tâm trong mẫu

11

Trong thực hành: Xác định trung bình mẫu, phương sai mẫu (đã hiệu chỉnh) như sau:

x_i	n_i
x_1	n_1
...	...
x_i	n_i
...	...
x_k	n_k
	$n=n_1+...+n_k$

Mẫu dạng điểm

* x_i là giá trị thu thập được

* n_i là số lần xuất hiện của x_i trong mẫu

$$\bar{x}=\frac{1}{n}\sum n_i x_i \quad ; \quad s^2=\frac{1}{n-1}\left(\sum n_i x_i^2-n(\bar{x})^2\right)$$

12

VD2: Điều tra năng suất lúa trên diện tích 100 hecta trồng lúa của một vùng, ta thu được bảng số liệu sau:

Năng suất (tạ / ha)	41	44	45	46	48	52	54
Số ha có năng suất tương ứng	10	20	30	15	10	10	5

1) Tính trung bình mẫu, phương sai mẫu hiệu chỉnh, độ lệch chuẩn mẫu hiệu chỉnh
2) Những thửa ruộng có năng suất từ 48 tạ trở lên là những thửa ruộng có năng suất cao. Tính tỷ lệ thửa ruộng có năng suất cao
3) Tính trung bình mẫu, phương sai mẫu hiệu chỉnh của những thửa ruộng có năng suất cao

13

Giải:

1) Ta lập bảng như sau

x_i	n_i	$n_i x_i$	$n_i x_i^2$
41	10	410	16.810
44	20	880	38.720
45	30	1350	60.750
46	15	690	31.740
48	10	480	23.040
52	10	520	27.040
54	5	270	14.580
Tổng	**n = 100**	**4600**	**212680**

14

Lưu ý: Máy tính Casio fx-570VN Plus có chức năng tính trung bình mẫu, độ lệch chuẩn mẫu (hiệu chỉnh). Xem file hướng dẫn trên trang web của Phạm Trí Cao.

Từ kết quả tính ở bảng trên ta có

Năng suất trung bình $\bar{x} = \frac{4600}{100} = 46$ tạ/ha

Phương sai (đã hiệu chỉnh) của năng suất

$$s^2 = \frac{1}{100-1}\left(212680 - 100 * 46^2\right) = 10,909$$

Độ lệch chuẩn (đã hiệu chỉnh) của năng suất

$$s = \sqrt{s^2} = \sqrt{10,909} = 3,303$$

15

2) Tỷ lệ mẫu là $f = \frac{10+10+5}{100} = 0,25$

3) Lập bảng sau

x_i	n_i	$n_i.x_i$	$n_i.x_i^2$
48	10	480	23040
52	10	520	27040
54	5	270	14580
Tổng	**n = 25**	**1270**	**64660**

$$\bar{x} = \frac{1270}{25} = 50,8$$

$$s^2 = \frac{1}{25-1}[64660 - 25*(50,8)^2] = 6$$

16

VD3: Quan sát tuổi thọ của một số người ta có bảng số liệu sau :

Tuổi (năm)	Số người
20 – 30	5
30 – 40	14
40 – 50	25
50 – 60	6

Mẫu dạng khoảng

1) Tính trung bình mẫu $\bar{x}$, phương sai mẫu s^2.
2) Những người sống dưới 40 tuổi là "chết trẻ". Tìm tỷ lệ người chết trẻ.

17

Giải:

Đưa về dạng điểm, lập bảng tính như VD2.

x_i	n_i
25	5
35	14
45	25
55	6

1) n= 50 ; $\bar{x}$ = 41,40 ; s^2= 68,4082
2) Tỷ lệ mẫu f = (5+14)/ 50 = 0,38

18

- VD4:
- Khảo sát 500.000 người ở một nước, người ta thấy có 75000 người có biểu hiện tâm thần.
- Tìm tỷ lệ mẫu của những người có biểu hiện tâm thần?
- Giải:
- Tỷ lệ mẫu f = 75000 / 500000 = 0,15
- VD5:
- Lô hàng có nhiều sản phẩm, các sản phẩm được đóng vào từng hộp. Mỗi hộp có 10 sản phẩm.
- Lấy 20 hộp từ lô hàng thì thấy có 60 sản phẩm loại A.
- Tìm tỷ lệ mẫu của sản phẩm loại A?
- Giải:
- Tỷ lệ mẫu f= 60/ 20*10 = 60/ 200

19

- VD6:
- Máy tự động sản xuất ra sản phẩm, cứ 10 sản phẩm đóng thành 1 hộp. Lấy ngẫu nhiên 100 hộp để kiểm tra, ta có bảng số liệu sau:

Số sp loại A trong hộp	7	8	9	10
Số hộp	5	25	30	40

- Xác định tỷ lệ mẫu của sản phẩm loại A?
- Giải:
- Tỷ lệ mẫu f = (1/1000).{7(5)+8(25)+9(30+10(40)}
  = 0,905

20

- VD 7: Bảng số liệu về chiều cao của một số người như sau:

Chiều cao (m)	1,3-1,5	1,5-1,7	1,7-1,8	1,8-2,0
Số người	30	70	60	40

- a) Những người có chiều cao trong khoảng từ 1,7m đến 1,8m là những người có chiều cao *mê ly*. Xác định tỷ lệ người *mê ly*?
- b) Những người có chiều cao từ 1,5m trở xuống là những người *mi nhon*. Xác định tỷ lệ người *mi nhon*?
- c) Những người có chiều cao từ 1,5m đến 1,8m là những người *có chiều cao lý tưởng*. Xác định tỷ lệ người *cao lý tưởng?*
- Giải:
- a) Tỷ lệ mẫu f= 60/200
- b) f= 30/200
- c) f= 130/200

21

VD8: Mẫu cụ thể 2 chiều

Ta có bảng số liệu về 2 chỉ tiêu X, Y của 1 loại sản phẩm như sau:

X \ Y	5	10	15	20	25
2	2	1			
4		2	2		
6		4	6	3	1
8			4	3	2

1) Xác định các đặc trưng số của mẫu về chỉ tiêu X, chỉ tiêu Y?

2) Sản phẩm có chỉ tiêu Y<= 15 và X<=6 gọi là sản phẩm loại A. Xác định tỷ lệ sản phẩm loại A của mẫu?

22

Giải:

1) Ta có bảng tần số thực nghiệm của X và Y như sau:

x_i	2	4	6	8
n_i	3	4	14	9

y_i	5	10	15	20	25
n_i	2	7	12	6	3

* Chỉ tiêu X: $n = 30$, $\sum n_x x = 178$

$\sum n_x x^2 = 1156$, $\bar{x} = 178/30 = 5,9333$

$s_x^2 = \frac{1}{n-1}[\sum n_x x^2 - n(\bar{x})^2] = 3,4441$

23

1) Chỉ tiêu Y:

$n = 30$, $\sum n_y y = 455$, $\sum n_y y^2 = 7725$

$\bar{y} = 455/30 = 15,1667$

$s_y^2 = \frac{1}{n-1}[\sum n_y y^2 - n(\bar{y})^2] = 28,4185$

2) Tỷ lệ sản phẩm loại A của mẫu:

f= 17/30 = 0,5667

24

III. PHÂN PHỐI CỦA CÁC ĐẶC TRƯNG MẪU

Định lý:

Tổng thể có quy luật phân phối X với:

$E(X)=\mu$ và $var(X)=\sigma^2$

- Lấy mẫu có hoàn lại:

  $E(\overline{X})=\mu$ và $var(\overline{X})=\sigma^2/n$

- Lấy mẫu không hoàn lại:

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

  $\frac{N-n}{N-1}$ gọi là hệ số hiệu chỉnh

25

Quy luật phân phối xác suất của đặc trưng mẫu NN:

Định lượng:

Ta có $X \sim N(\mu, \sigma^2)$

- $\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Do đó: $P(a<\overline{X}<b)=\phi(\frac{b-\mu}{\sigma/\sqrt{n}})-\phi(\frac{a-\mu}{\sigma/\sqrt{n}})$

$P(|\overline{X}-\mu|<\varepsilon)=2\phi(\frac{\varepsilon}{\sigma/\sqrt{n}})$

26

Quy luật phân phối xác suất của đặc trưng mẫu NN:

Định tính:

$F=\frac{1}{n}\sum X_i$

với X_i có quy luật ppxs 0-1

X	0	1
P	q=1-p	p

$E(F)=p$, $var(F)=\frac{pq}{n}$

27

VD9: Chiều cao thanh niên của vùng M là biến ngẫu nhiên phân phối chuẩn với μ= 165 cm, $\sigma^2= 20^2$ cm^2 .

1) Người ta đo ngẫu nhiên chiều cao của 100 thanh niên vùng đó.

a) Xác suất để chiều cao trung bình của 100 thanh niên đó sẽ sai lệch so với chiều cao trung bình của thanh niên vùng M không vượt quá 1 cm là bao nhiêu?

b) Khả năng chiều cao trung bình của 100 thanh niên trên lớn hơn 168 cm là bao nhiêu?

2) Nếu muốn chiều cao trung bình đo được của *1 số* thanh niên sai lệch so với chiều cao trung bình của tổng thể *(của tất cả thanh niên vùng M)* không vượt quá 3 cm với xác suất là 0,99 thì chúng ta phải tiến hành đo chiều cao của bao nhiêu thanh niên?

28

Giải:

1) $\overline{X}$ là chiều cao tb của 100 thanh niên khảo sát

μ là chiều cao tb của thanh niên toàn vùng M

X~N(165, 20^2) → $\overline{X}$~N(165, 20^2/ 100) = N(165, 2^2)

a) $P(|\overline{X}-\mu|<1)=2\phi(\frac{1}{2})=2(0,1915)=0,3830$

b) $P(\overline{X}>168)=0,5-\phi(\frac{168-165}{2})$

$=0,5-\phi(1,5)=0,5-0,4332=0,0668$

29

Giải:

2) $\overline{X}$ là chiều cao tb của ***n*** thanh niên cần khảo sát

μ là chiều cao tb của thanh niên toàn vùng M

Tìm n sao cho: $P(|\overline{X}-\mu|<3)=0,99$

X~N(165, 20^2) → $\overline{X}$~N(165, 20^2/ n)

$P(|\overline{X}-\mu|<3)=2\phi(\frac{3}{20/\sqrt{n}})=0,99$

$\Rightarrow \phi(\frac{3}{20}\sqrt{n})=0,495=\phi(2,58)$

$\Rightarrow \frac{3}{20}\sqrt{n}=2,58 \Rightarrow$ n= 295,84 ≈ 296 *(làm tròn lên)*

Làm tròn lên của 1 số thập phân là lấy *phần nguyên* của số đó cộng thêm 1

30

Mời ghé thăm trang web:

- https://sites.google.com/a/ueh.edu.vn/phamtricao/
- https://sites.google.com/site/phamtricao/

31