XÁC SUẤT THỐNG KÊ

Tôn Thất Tú

Đà Nẵng, 2019

Chương 4: Thống kê mô tả

1. Tổng thể và mẫu

a. Các khái niệm:

- *Tổng thể*: tập hợp toàn bộ các phần tử thống nhất theo dấu hiệu nghiên cứu. Tổng thể có thể hữu hạn hoặc vô hạn.
- *Mẫu*: một tập con bất kì của tổng thể. Số lượng phần tử của nó được gọi là kích thước hay cỡ mẫu.
- *Phép lấy mẫu*: việc chọn một tập con bất kì của tổng thể.
- *Mẫu ngẫu nhiên*: một mẫu là ngẫu nhiên nếu việc chọn các cá thể được tiến hành độc lập và có xác suất chọn như nhau.

Về mặt toán học, một mẫu ngẫu nhiên kích thước n từ tổng thể có phân phối theo biến ngẫu nhiên X có thể xem như một bộ n biến ngẫu nhiên $\{\mathrm{X}_1, X_2, ..., X_n\}$ độc lập và có cùng phân phối với X.

- Cho mẫu ngẫu nhiên $\{\mathrm{X}_1, X_2, ..., X_n\}$. Khi đó:

+ Bộ n giá trị $\{\mathrm{x}_1, x_2, ..., x_n\}$ cụ thể quan thu được được gọi là *mẫu thực nghiệm*.

+ Một hàm (đo được) $\mathrm{T} = \mathrm{T}(\mathrm{X}_1, X_2, ..., X_n)$ được gọi là một *thống kê* trên mẫu ngẫu nhiên $\{\mathrm{X}_1, X_2, ..., X_n\}$.

b. Phép lấy mẫu đơn giản:

- Lấy có hoàn lại: Chọn ngẫu nhiên một cá thể từ tổng thể, ghi lại các dấu hiệu cần quan tâm và hoàn trả lại vào tổng thể trước khi chọn tiếp lần sau.
- Lấy không hoàn lại: tương tự trên, nhưng phần tử được lấy ra không trả lại vào tổng thể trước khi chọn tiếp.

2. Bảng tần số, tần suất

a. Mẫu không ghép lớp

- Bảng tần số:

Giá trị	x_1	x_2	...	x_m
Tần số	n_1	n_2	...	n_m

trong đó $x_1, ..., x_m$ là các giá trị khác nhau với số lần xuất hiện là $n_1, ..., n_m$.

- Bảng tần suất:

Giá trị	x_1	x_2	...	x_m
Tần suất	f_1	f_2	...	f_m

trong đó $f_i = n_i/n, n = \sum_{i=1}^{m} n_i$. Giá trị f_i được gọi là *tần suất* xuất hiện của x_i trong mẫu.

Ví dụ 1

Khảo sát tuổi của một nhóm học viên học anh văn tại một trung tâm ngoại ngữ, ta thu được bảng số liệu sau:

Tuổi	*15*	*16*	*17*	*18*	*19*
Số lượng	*4*	*6*	*3*	*5*	*7*

Ví dụ 2

Khảo sát ngẫu nhiên một nhóm người về việc họ sử dụng phương tiện gì thường xuyên nhất để đọc báo: báo giấy, máy tính, điện thoại hay máy tính bảng. Kết quả được thể hiện như sau:

Phương tiện	*Báo giấy*	*Máy tính*	*Điện thoại*	*Máy tính bảng*
Số lượng	*5*	*20*	*50*	*15*

b. Mẫu ghép lớp

Khi ta thu được mẫu dữ liệu với nhiều giá trị khác nhau thì người ta tiến hành chia miền giá trị thành nhiều khoảng $[a_{i-1}, a_i)$ không giao nhau.

- Bảng tần số:

Khoảng giá trị	$[a_0, a_1)$	$[a_1, a_2)$	...	$[a_{m-1}, a_m)$
Tần số	n_1	n_2	...	n_m

- Bảng tần suất:

Khoảng giá trị	$[a_0, a_1)$	$[a_1, a_2)$	...	$[a_{m-1}, a_m)$
Tần suất	f_1	f_2	...	f_m

Nhận xét: Thông thường các khoảng chia có độ dài bằng nhau. Tuy nhiên, tùy thuộc vào mục đích nghiên cứu mà ta có thể có những cách chia khoảng khác nhau.

Ví dụ 3

Cân thử 100 trái táo vừa thu hoạch, ta được bảng số liệu sau:

Khối lượng (g)	*[100;120)*	*[120;140)*	*[140;160)*	*[160;180)*	*[180;200)*
Số trái	*12*	*19*	*31*	*23*	*15*

Ví dụ 4

Khảo sát thời gian trung bình (tính bằng giờ) mà một người từ độ tuổi 15 trở lên dành để đọc tin tức thời sự online trong một ngày ở một thành phố, số liệu được thể hiện ở bảng sau:

Độ tuổi	*[15,20)*	*[20,30)*	*[30,40)*	*[40,50)*	≥ 50
Số người	*20*	*25*	*20*	*30*	*15*

3. Các số đặc trưng mẫu

a. Trung bình và phương sai mẫu

Cho $\{x_1, x_2, ..., x_n\}$ là mẫu số liệu của biến ngẫu nhiên X.

- Trung bình mẫu, kí hiệu là $\overline{x}$, được tính theo công thức:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Phương sai mẫu, kí hiệu là s^2, được tính theo công thức:

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

- Độ lệch chuẩn mẫu, kí hiệu s, được tính:

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

Nhận xét:

i) Khi mẫu được cho ở dạng bảng tần số:

X	x_1	x_2	...	x_m
n_i	n_1	n_2	...	n_m

- Kích thước mẫu: $n = n_1 + n_2 + ... + n_m$.
- Trung bình mẫu: $\overline{x} = \frac{1}{n}\sum_{i=1}^{m} n_i x_i$.
- Phương sai mẫu: $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{m} n_i x_i^2 - n\overline{x}^2\right]$.

ii) Khi mẫu ở dạng bảng ghép lớp:

X	$a_0 - a_1$	$a_1 - a_2$	...	$a_{m-1} - a_m$
n_i	n_1	n_2	...	n_m

trong đó $a_{k-1} - a_k = [a_{k-1}; a_k)$.
Đặt $x_k = \dfrac{a_{k-1} + a_k}{2}$ ta được mẫu dạng thu gọn:

X	x_1	x_2	...	x_m
n_i	n_1	n_2	...	n_m

iii) Tính $\overline{x}$ và s bằng máy tính CASIO FX570VN PLUS.
- Mode $\to$ 3 $\to$ 1
- Bật/tắt tần số: Shift $\to$ SETUP $\to$ REPLAY ($\downarrow$) $\to$ 4(Stat)
- Nhập số liệu, kết thúc nhập: bấm AC
- Lấy $\overline{x}$: Shift $\to$ 1 $\to$ 4 $\to$ 2 $\to$ =
- Lấy s: Shift $\to$ 1 $\to$ 4 $\to$ 4 $\to$ =

Ví dụ 5

Chiều cao (mét) của 10 sinh viên đại học:

$$1.75, 1.69, 1.70, 1.82, 1.68, 1.72, 1.70, 1.67, 1.71, 1.68$$

Tính trung bình mẫu, phương sai mẫu và độ lệch chuẩn mẫu.

Đáp số.

$$\bar{x} = 1,712; \quad s^2 = 0.00197; \quad s = 0.0444$$

Ví dụ 6

Doanh thu X (triệu đồng) trong 100 ngày được chọn ngẫu nhiên của 1 cửa hàng:

X	*19,0 - 19,4*	*19,4 - 19,8*	*19,8 - 20,2*	*20,2 - 20,6*	*20,6 - 21,0*
n_i	15	25	30	20	10

Tìm trung bình mẫu và độ lệch chuẩn mẫu.

Giải. Dạng thu gọn:

X	19,2	19,6	20	20,4	20,8
n_i	15	25	30	20	10

Các số đặc trưng mẫu:

$$\bar{x} = 19,94; \quad s = 0,48$$

b. Trung vị mẫu

Sắp xếp mẫu số liệu theo thứ tự tăng dần, giả sử $x_1 \leq x_2 \leq ... \leq x_n$. *Trung vị mẫu*, kí hiệu x_{med}, xác định bởi:

$$x_{med} = \begin{cases} x_{\frac{n+1}{2}}, & \text{nếu } n \text{ lẻ,} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{nếu } n \text{ chẵn.} \end{cases}$$

c. Hệ số tương quan mẫu

Cho $\{(x_1, y_1); (x_2, y_2); ...; (x_n, y_n)\}$ là mẫu hai chiều của vectơ ngẫu nhiên (X, Y). *Hệ số tương quan* mẫu được xác định bởi:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

3. Biểu đồ

a. Biểu đồ cho dữ liệu rời rạc:

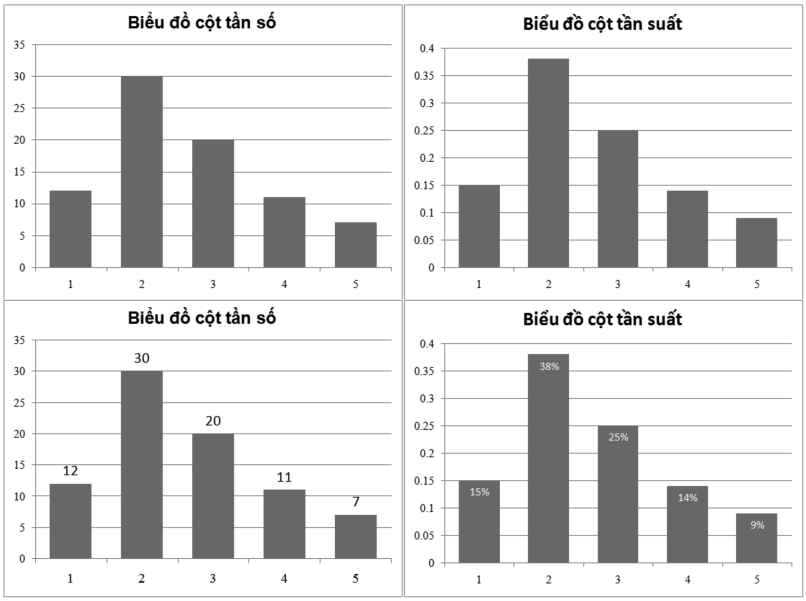
- *Biểu đồ cột*: đồ thị gồm các cột hình chữ nhật có chiều cao bằng tần số (tần suất) tương ứng.
- *Đa giác tần số* : đường gấp khúc nối các điểm $(x_1, n_1), ..., (x_k, n_k)$ trên mặt phẳng.
- *Đa giác tần suất*: đường gấp khúc nối các điểm $(x_1, f_1), ..., (x_k, f_k)$ trên mặt phẳng.
- *Biểu đồ hình tròn*: hình tròn được chia ra các phần có diện tích tỉ lệ với tần suất (tần số) tương ứng.

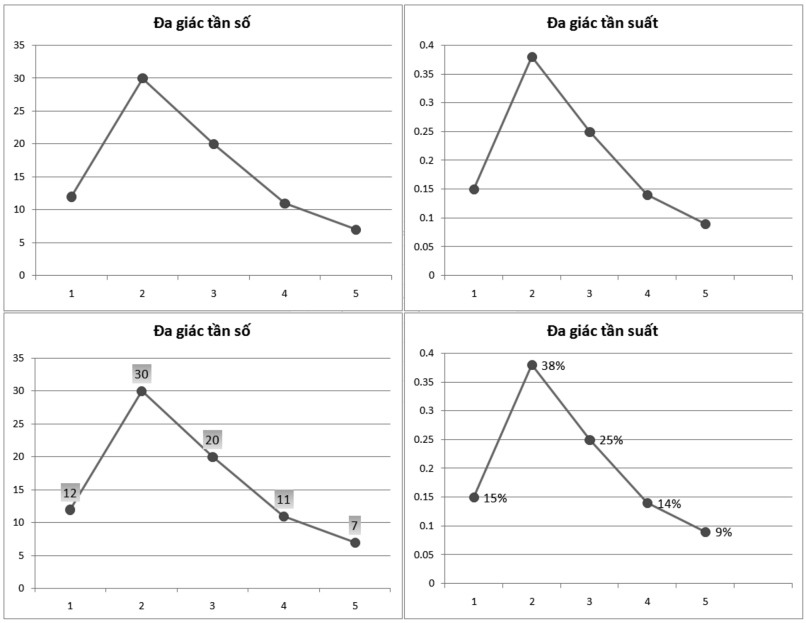
Ví dụ 7

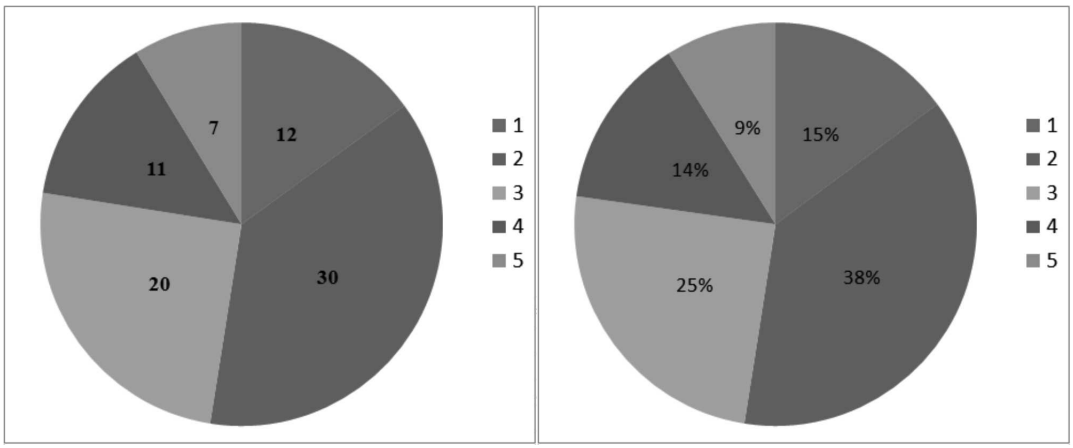
Cho dữ liệu:

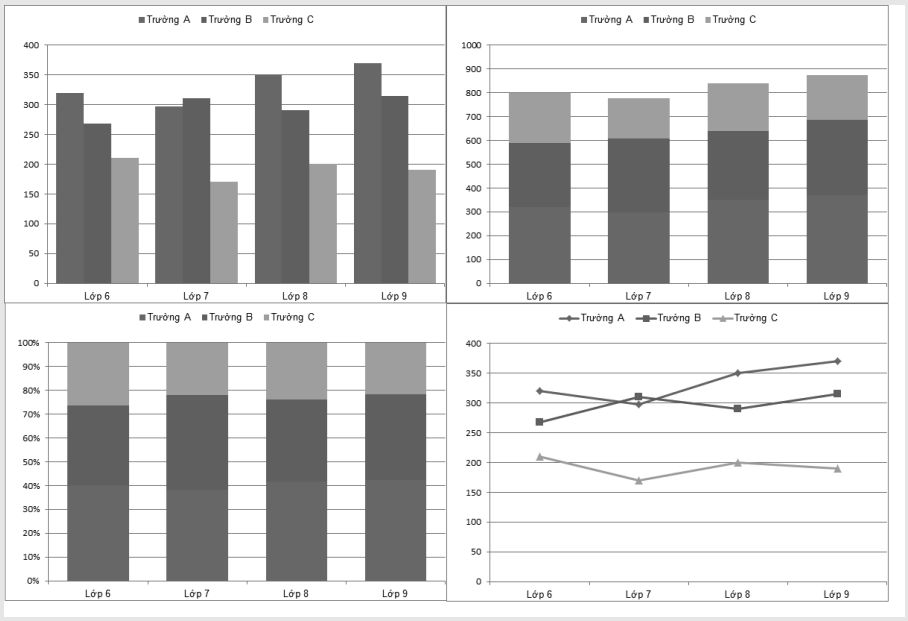
X	1	2	3	4	5
Tần số	12	30	20	11	7
Tần suất	0.15	0.4	0.3	0.14	0.1

Hãy vẽ các biểu đồ:
- *biểu đồ cột (tần số, tần suất)*
- *đa giác tần số, tần suất*
- *biểu đồ hình tròn*


Biểu đồ cột tần số
Biểu đồ cột tần suất
Biểu đồ cột tần số
12
30
20
11
7
Biểu đồ cột tần suất
15%
38%
25%
14%
9%


Đa giác tần số
Đa giác tần suất
Đa giác tần số
12
30
20
11
7
Đa giác tần suất
15%
38%
25%
14%
9%


12
30
20
11
7
1
2
3
4
5
15%
38%
25%
14%
9%



b. Biểu đồ cho dữ liệu liên tục:

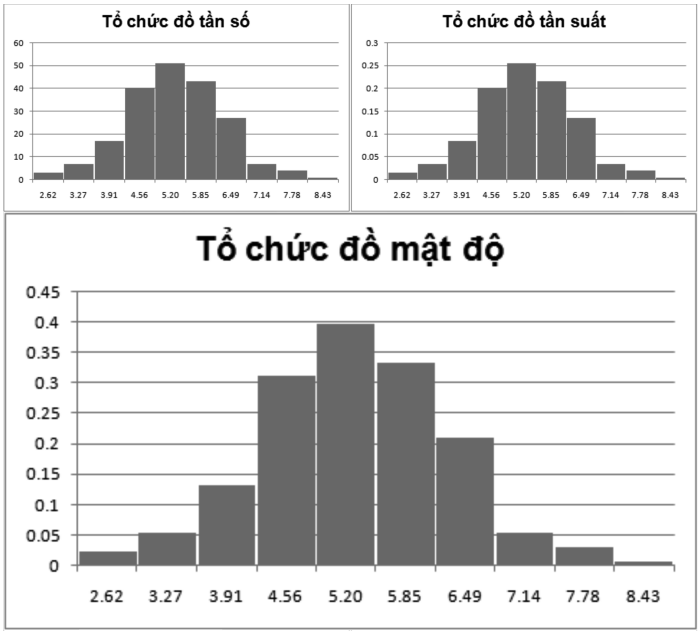
- *Tổ chức đồ (histogram)*: Giả sử mẫu dữ liệu được chia làm m với độ dài các khoảng lần lượt là h_i (thường ta chọn $h_i = h$). Khi đó tổ chức đồ của mẫu dữ liệu này gồm m hình chữ nhật có đáy trùng với trục hoành và:

+ Độ dài cạnh đáy hình thứ i là chiều dài h_i của khoảng thứ i.
+ Chiều cao của hình thứ i bằng d_i, trong đó:
  - ▷ $d_i = n_i$: tổ chức đồ tần số
  - ▷ $d_i = \dfrac{n_i}{n}$: tổ chức đồ tần suất
  - ▷ $d_i = \dfrac{n_i}{n.h_i}$: tổ chức đồ mật độ

với n_i là số lượng các giá trị nằm trong khoảng thứ i và $n = n_1 + n_2 + ... + n_m$.

Ví dụ 8

Dữ liệu được khảo sát từ biến ngẫu nhiên liên tục. Miền dữ liệu được chia thành 10 khoảng đều nhau. Các tổ chức đồ được xây dựng như ở hình sau đây.



- *Biểu đồ xác suất chuẩn*: Giả sử mẫu số liệu của biến ngẫu nhiên liên tục X đã sắp thứ tự tăng dần:

$$x_1 \leq x_2 \leq x_3 \leq ... \leq x_n.$$

Với $i = 1, 2, ..., n,$ đặt

$$z_i = \Phi^{-1}\left(\frac{i - 0.5}{n}\right) \text{ hay } \Phi(z_i) = (i - 0,5)/n.$$

Biểu đồ xác suất chuẩn là tập hợp các điểm có tọa độ $(z_i; x_i), i = 1, 2, ..., n$ trên hệ trục tọa độ Descartes vuông góc Ozx.
Nếu $(z_i; x_i), i = 1, 2, ..., n$ nằm xấp xỉ trên 1 đường thẳng thì có thể xem biến ngẫu nhiên X có phân phối chuẩn. Đường thẳng này có phương trình:

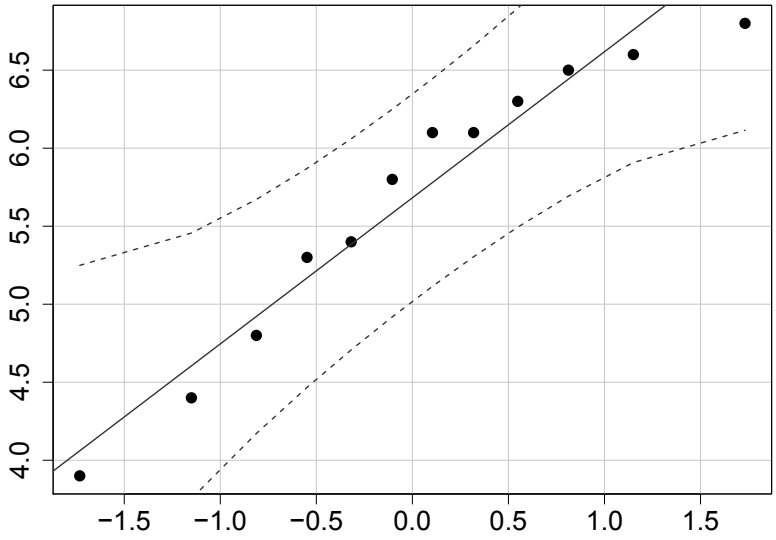
$$x = \hat{\sigma} z + \hat{\mu},$$

với $\hat{\mu} = \bar{x}, \hat{\sigma} = s$ - trung bình và độ lệch chuẩn mẫu được tính dựa trên mẫu đã cho.

Ví dụ 9

Quan sát biểu đồ xác suất chuẩn sau đây.

BIỂU ĐỒ XÁC SUẤT CHUẨN



4. Phân phối mẫu

Định lý 1: Nếu $\{X_1, X_2, ..., X_n\}$ là mẫu ngẫu nhiên của biến ngẫu nhiên X có phân phối chuẩn $N(\mu; \sigma^2)$ thì

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ hay } \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0, 1).$$

Định lý 2: Nếu $\{X_1, X_2, ..., X_n\}$ là mẫu ngẫu nhiên của biến ngẫu nhiên X có phân phối chuẩn $N(\mu; \sigma^2)$ thì:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \sqrt{n}(\overline{X} - \mu)/S \sim T_{n-1}$$

Nhận xét: Trường hợp khi X không có phân phối chuẩn. Nếu kích thước mẫu lớn $(n > 30)$ thì theo định lý giới hạn trung tâm:

$$\overline{X} \text{ có phân phối xấp xỉ } N\left(\mu, \frac{\sigma^2}{n}\right) \text{ và } \frac{\sqrt{n}(\overline{X} - \mu)}{S} \text{ có phân phối xấp xỉ } N(0, 1)$$

Ví dụ 10

Chiều cao của thanh niên tuân theo luật phân phối chuẩn $N(\mu;\sigma^2)$, với $\mu = 165$(cm) và $\sigma = 5$(cm). Tính xác suất chiều cao trung bình của 16 thanh niên được chọn ngẫu nhiên lớn hơn 167 (cm).

Giải. Gọi X_i là chiều cao của thanh niên thứ $i, i = \overline{1,16}$.
Đặt $\bar{X} = (X_1 + ... + X_{16})/16$. Khi đó, $\bar{X}$ cũng có phân phối chuẩn với trung bình $\mu = 165$ và phương sai $\sigma^2/n = 25/16$.
Xác suất cần tìm:

$$P(\bar{X} > 167) = 1 - P(\bar{X} \leq 167) = 1 - \Phi\left(\frac{167 - 165}{5/4}\right) = 0.0548$$

Ví dụ 11

Để nghiên cứu về thâm niên công tác (tính tròn năm) của nhân viên ở một công ty lớn, người ta khảo sát thâm niên của 100 nhân viên được chọn ngẫu nhiên trong công ty. Kết quả như sau:

Thâm niên	*5-7*	*8-10*	*11-13*	*14-16*	*17-19*
Số nhân viên	*8*	*21*	*36*	*25*	*10*

a. Hãy tính giá trị trung bình mẫu và giá trị độ lệch chuẩn mẫu.
b. Giả sử thâm niên công tác của nhân viên ở công ty trên là biến ngẫu nhiên X có kỳ vọng là 12 năm và độ lệch chuẩn là 3 năm. Tính xác suất để trung bình mẫu nhận giá trị lớn hơn 12,5 năm.

Giải. Dạng thu gọn:

Thâm niên	6	9	12	15	18
Số nhân viên	8	21	36	25	10

a. Các số đặc trưng mẫu: $\bar{x} = 12,24; s = 3,27$.
b. Theo định lý giới hạn trung tâm $\bar{X}$ có phân phối xấp xỉ chuẩn $N(12; 3^2/100) = N(12; 0,09)$. Do đó:

$$P(\bar{X} > 12,5) = 1 - \Phi\left(\frac{12,5 - 12}{\sqrt{0,09}}\right) = 0.0478$$