# XỬ LÝ THỐNG KÊ SỐ LIỆU THỰC NGHIỆM TRONG PHÒNG THÍ NGHIỆM

*Tác giả: Nguyễn Văn Lân, PGS/TS*

## CHƯƠNG 3

# TÍNH CÁC SỐ THỐNG KÊ CỦA MẪU

Từ số cá thể của một *tổng thể* hay *lô*, người ta chọn ra một số *n* cá thể để kiểm tra. Tập hợp số cá thể này gọi là *mẫu*. Các *số thống kê* sẽ đặc trưng một cách khái quát chất lượng của mẫu.

## 1. CÁC SỐ THỐNG KÊ THỂ HIỆN VỊ TRÍ

Đó là các giá trị thể hiện mức độ lớn hay bé, nhiều hay ít, cao hay thấp của chất lượng gọi là số trung bình khái quát hóa : số trung bình mũ p.

Giả sử từ *n* giá trị quan trắc $x_i$, *số trung bình mũ p* được tính như sau:

$$M_p = \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p} \quad \text{với } p \neq 0.$$

Nếu p = 0 :

$$M_0 = \lim_{p \to 0} \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p}$$

Nếu p = +∞

$$M_{+\infty} = \lim_{p \to +\infty} \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p} = \max\{x_i\}$$

Nếu p = −∞

$$M_{-\infty} = \lim_{p \to -\infty} \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p} = \min\{x_i\}$$

Từ đó suy ra:

**$M_1 = \bar{\bar{x}}$ là số trung bình cộng**

**Số trung bình cộng**: $\bar{x} = \frac{x_1+x_2+...+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

**$M_2$ = Q là số trung bình toàn phương**

**Số trung bình toàn phương**: $Q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$

**$M_{-1}$ = H là số trung bình điều hòa**

**Số trung bình điều hòa**: $H = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n}\right)^{-1} = \frac{n}{\sum_{i=1}^{n} x_i^{-1}}$

**$M_0$ = G là số trung bình nhân**

**Số trung bình nhân**: $G = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Nếu xem các đại lượng trên là trường hợp đặc biệt của trung bình mũ thì về độ lớn, chúng có mối quan hệ sau :

$$M_{-\infty} \le ... \le M_{-1} \le M_0 \le M_1 \le M_2 \le ... \le M_{+\infty}$$

hay là $\min\{x_i\} \le ... \le H \le G \le \bar{x} \le Q \le ... \le \max\{x_i\}$

Ngoài ra còn có hai dạng số trung bình cũng được sử dụng trong thực tế thống kê là số trung tâm và số mốt.

**Số trung tâm** (còn gọi là **trung vị**) ký hiệu $\tilde{x}$. Để tìm trung vị, phải sắp xếp lại các $x_i$ thành một dãy thứ tự các giá trị từ bé đến lớn : $x_1 \le x_2 \le .... \le x_n$

Khi *n lẻ* , đặt $k = \frac{n+1}{2}$ sẽ được $\tilde{x} = x_k$ (tức $x_i$ ở vị trí thứ i = k của dãy)

Khi *n chẵn*, đặt $k = \frac{n}{2}$ sẽ được $\tilde{x} = \frac{x_k + x_{k+1}}{2}$

*Ví dụ 1* : Có một mẫu với n = 10 kết quả đo sau đây đã được sắp xếp thứ tự tăng

$\{x_i\}$ = 3,2 – 3,2 – 3,4 – 4,6 – 4,8 – 5,2 – 5,6 – 6,4 – 6,8 – 7,6}

Vì n chẵn nên $\tilde{x} = \frac{4,8+5,2}{2} = 5,0$

*Ví dụ 2* : Một mẫu lớn n = 120 đã được xếp thành lớp với tần số $n_i$ tương ứng sau:

| Lớp | 0,5-1,5 | 1,5-2,5 | 2,5-3,5 | 3,5-4,5 | **4,5**-5,5 | 5,5-6,5 | 6,5-7,5 | 7,5-8,5 | 8,5-9,5 |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 8 | 10 | 11 | 16 | **20** | 25 | 15 | 9 | 6 |
| $\Sigma n_i$ | 8 | 18 | 29 | **45** | **65** | 90 | 105 | 114 | 120 |

Áp dụng công thức sau để tìm trung vị:

$$\tilde{x} = L + \left(\frac{0{,}5.n - F_a}{F_w}\right).\Delta$$

trong đó: L – biên dưới của lớp chứa trung vị, ở ví dụ trên là 4,5

$F_a$ – tần số lũy tích của lớp trước lớp chứa trung vị, ở ví dụ trên là 45

$F_w$ – tần số của lớp chứa trung vị, ở ví dụ trên là 20

$\Delta$ – khoảng cách giữa hai lớp kế tiếp, ở ví dụ trên là 1

Vậy theo ví dụ trên: $\tilde{x} = 4{,}5 + \left(\frac{0{,}5.120 - 45}{20}\right).1 = 5{,}25$

**Số mốt** (còn gọi là **yếu vị**) $x_{mod}$ là giá trị $x_i$ xuất hiện nhiều nhất trong tập hợp mẫu{$x_i$}.

$$x_{mod} \approx 3\tilde{x} - 2\bar{x}$$

*Ví dụ 1: Có một mẫu với kết quả đo sau đây đã được sắp xếp thứ tự tăng*

{$x_i$} = 3 4 4 5 5 5 6 6 6 6 7 7 8 9 10

Giá trị xuất hiện nhiều nhất là 6 có 4 lần nên $x_{mod} = 6$

*Ví dụ 2 : Một mẫu lớn n = 100 đã được xếp thành lớp với tần số $n_i$ tương ứng sau:*

| Lớp | 0 –10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 – 50 | 50 – 60 | 60 – 70 | 70 – 80 |
|---|---|---|---|---|---|---|---|---|
| $n_i$ | 5 | 8 | 7 | 12 | **28** | 20 | 10 | 10 |

Số mốt được xác định theo công thức sau :

$$x_{mod} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}\Delta$$

trong đó: L – biên dưới của lớp chứa mode, ở ví dụ trên là 40

$f_1$ – tần số của lớp chứa mode, ở ví dụ trên là 28

$f_0$ – tần số của lớp trước lớp chứa mode, ở ví dụ trên là 12

$f_2$ – tần số của lớp sau lớp chứa mode, ở ví dụ trên là 20

$\Delta$ – khoảng cách giữa hai lớp kế tiếp, ở ví dụ trên là 10

Vậy ở ví dụ trên: $x_{mod} = 40 + \frac{28-12}{2\times 28 - 12 - 20}\times 10 = 40 + \frac{20}{3} = 40{,}66$

Tùy trường hợp mà nên nghiên cứu áp dụng các loại số trung bình trên cho có hiệu quả. Thông thường trong thực tế thống kê, người ta hay sử dụng số trung bình cộng ($\bar{x}$). Khi cần nhấn mạnh đến ảnh hưởng của số đông thì dùng *trung bình cộng có trọng số $m_i$* :

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i x_i}{\sum_{i=1}^{k} m_i}$$

Một số thí nghiệm được quy định thay vì trung bình cộng ($\bar{x}$) là trung vị ($\tilde{x}$) như xác định chất lượng bê tông, độ bền nén thủng của condom, ... Ở những trường hợp tính toán chất lượng mà không có điều kiện loại trừ số lạc (như đánh giá của hội đồng chấm thi người đẹp, người giỏi, đánh giá thực chất trình độ trung bình mà không kể đến sở trường sở đoản, v.v..) thì nên áp dụng *trung vị*.

Trong sản xuất mặt hàng phục vụ cho đối lượng đa dạng về loại, cỡ (như may mặc), số lượng sản phẩm các loại cỡ nếu áp dụng *số mốt* thì việc kinh doanh nhanh chóng mang lại hiệu quả hơn (như ít bị tồn kho).

Khi tính toán năng suất, tỷ suất dịch vụ,... trung bình trong cả thời gian dài thì nên áp dụng *trung bình nhân*. Ví dụ giá một mặt hàng tiêu dùng từ năm 1985 đến 1986 tăng 5 %, từ 1986 đến 1987 tăng 8 %, từ 1987 đến 1988 tăng 77 % vậy từ 1985 đến 1988 giá của nó tăng trung bình bao nhiêu? Không phải tính (105 + 108 + 177)/3 = 130 tức 30 % mà phải tính $(105\times108\times177)^{1/3}$ = 126,14 tức 26,1 %

Trong trường hợp thử nghiệm chi số $N_i$ của n đoạn sợi có chiều dài không đổi L, thì việc tính chi số trung bình đúng đắn là phải sử dụng số trung bình điều hòa.

Bởi vì, chi số trung bình theo định nghĩa là bằng :

$$\overline{N} = \frac{\Sigma L}{\Sigma G_i} = \frac{n.L}{\Sigma G_i} = \frac{n}{\Sigma \frac{G_i}{L}} = \frac{n}{\Sigma N_i^{-1}}$$

Kết quả của phép tính chính là số trung bình điều hòa.

*Ví dụ khác: Một chiếc ô tô chạy đi 100 km từ đồng bằng lên đồi với vận tốc trung bình 30 km/h, chạy về trên đoạn đường đó với vận tốc trung bình 20 km/h. Hỏi vận tốc trung bình cho cả hai lần đi và về là bao nhiêu ?*

Nếu dùng trung bình cộng, ta có : $\bar{x} = \frac{30+20}{2}$ = 25 km/h

Nếu dùng trung bình điều hòa : $H = \frac{2}{\frac{1}{30}+\frac{1}{20}}$ = 24 km/h

Kết quả nào đúng? Nếu chuyến đi 100 km với $v_{tb}$ = 30 km/h sẽ mất 3 h 20 min, chuyến về 100 km với $v_{tb}$ = 20 km/h sẽ mất 5 h. Tổng cộng đi 200 km mất 8 h 20 min , tính ra:

$$v_{tb} = \frac{200}{8h20} = 24 \text{ km/h},$$

vậy vận tốc trung bình tính theo trung bình điều hòa là đúng!

**Số phần tư** $x_{1/4}$ (gồm số phần tư dưới và số phần tư trên).

*Số phần tư dưới* ký hiệu $x_{1/4d}$ nằm ở vị trí dãy số của kết quả đo được sắp xếp theo thứ tự từ bé đến lớn mà 1/4 của *n* giá trị không vượt quá nó và 3/4 của *n* giá trị còn lại bằng và vượt quá nó.

*Số phần tư trên* ký hiệu $x_{1/4t}$ nằm ở vị trí dãy số của kết quả đo được sắp xếp theo thứ tự từ bé đến lớn mà 3/4 của *n* giá trị không vượt quá nó và 1/4 của *n* giá trị còn lại bằng và vượt quá nó.

Cách tính hai số phần tư này như sau.

- Sắp xếp dãy số theo thứ tự tăng dần: $x_1 \le x_2 \le .... \le x_n$
- Tính k = (n+1)/4 và làm tròn đến số nguyên gần nhất $k_1$. Tại vị trí $k_1$, $x_{1/4d} = x_{k1}$. Nếu k nằm chính giữa hai số nguyên thì làm tròn tăng.
- Tính k = 3(n+1)/4 và làm tròn đến số nguyên gần nhất $k_3$. Tại vị trí $k_3$, $x_{1/4t} = x_{k3}$.

Nếu k nằm chính giữa hai số nguyên thì làm tròn giảm hoặc $k_3 = 3(n - 1)/4 + 1$. Đó là những trường hợp mà n = 4m + 1, trong đó m là số nguyên dương.

*Ví dụ 1 : Mẫu có* n = 13 *kết quả đo đã được sắp xếp tăng dần:*

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 122 | 134 | 136 | 140 | 142 | 146 | 156 | 158 | 160 | 168 | 172 | 176 | 180 |

$k_1 = \frac{13+1}{4} = 3,5$ lấy tròn tăng bằng 4 ; $k_3 = \frac{13+1}{4}3 = 10,5$ lấy tròn giảm bằng 10.

Vậy $x_{1/4d} = 140$ và $x_{1/4t} = 168$

*Ví dụ 2 : Một mẫu lớn* n = 100 *đã được xếp thành lớp với tần số* $n_i$ *tương ứng sau:*

| Lớp | 0 –10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 – 50 | 50 – 60 | 60 – 70 | 70 – 80 |
|---|---|---|---|---|---|---|---|---|
| $n_i$ | 5 | 8 | 7 | 12 | 28 | 20 | 10 | 10 |
| Tích lũy | 5 | 13 | 20 | 32 | 60 | 80 | 90 | 100 |

Xem biến ở trường hợp này là liên tục, $k_1 = \frac{n}{4}$ và $k_3 = \frac{3n}{4}$. Công thức tìm số phần tư:

$$x_{1/4} = L + \left(\frac{k_i - F_a}{F_w}\right)\Delta$$

ở đó: L – biên dưới của lớp chứa số phần tư , ở ví dụ trên là 30 và 50

$F_a$ – tần số lũy tích của lớp trước lớp chứa số phần tư, ở ví dụ trên là 20 và 60

$F_w$ – tần số của lớp chứa số phần tư, ở ví dụ trên là 12 và 20

$\Delta$ – khoảng cách giữa hai lớp kế tiếp, ở ví dụ trên là 10

Vậy $x_{1/4d} = 30 + \frac{\frac{100}{4} - 20}{12}10 = 34,2$ và $x_{1/4t} = 50 + \frac{\frac{3\times100}{4} - 60}{20}10 = 57,5$

## 2. CÁC SỐ THỐNG KÊ THỂ HIỆN MỨC ĐỘ PHÂN TÁN

Mức độ phân tán của các giá trị $x_i$ trong tập hợp thể hiện tính không đồng nhất nhiều hay ít của chất lượng qua các số thống kê sau đây :

**Độ rộng** ký hiệu *w*, là phạm vi biến động của các giá trị $x_i$ từ giá trị nhỏ nhất $x_{min}$ cho đến giá trị lớn nhất $x_{max}$. tức là : $w = x_{max} - x_{min}$.

**Độ rộng phần tư** ký hiệu $w_{1/4}$ là hiệu số của số phần tư trên và số phần tư dưới

$$w_{1/4} = x_{1/4t} - x_{1/4d}$$

**Độ rộng phần tư chuẩn hóa** ký hiệu $w_{1/4ch}$ tính theo :

$$w_{1/4ch} = 0{,}7413.w_{1/4}$$

Hệ số 0,7413 lấy từ phân bố chuẩn (standard normal distribution) có $\mu = 0$ và $\sigma = 1$. Độ rộng phần tư $w_{1/4}$ của phân bố này bằng 1,34898 và 1/1,34898 = 0,7413.

**Phương sai** ký hiệu $s^2$, đôi khi ký hiệu bằng *v* nói lên mức độ phân tán của các giá trị $x_i$ so với $\bar{x}$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Trường hợp giá trị quan trắc $x_i$ có tần số $n_i$ thì có thể tính độ lệch chuẩn *s* theo

$$s^2 = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + ... + n_k(x_n - \bar{x})^2}{(n_1 + n_2 + ... n_k) - 1} = \frac{\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}{\sum_{i=1}^{k} n_i - 1}$$

Phương sai được tính với giá trị cỡ mẫu trừ đi 1 được gọi là *phương sai không chệch*, còn mẫu số đúng bằng cỡ mẫu được gọi là *phương sai chệch, ký hiệu* $s_c^2$. Giá trị *s* gọi là *độ lệch chuẩn (không chệch)*.

Độ rộng phần tư $w_{1/4}$ xấp xỉ 1,3.s nên $w_{1/4ch}$ xấp xỉ bằng s.

**Hệ số biến động** ký hiệu *cv*% cũng thể hiện mức độ phân tán của các giá trị $x_i$ so với $\bar{x}$ nhưng ở dạng tương đối :

$$cv\ \% = \frac{s}{\bar{x}} \cdot 100$$

**Hệ số biến động thô** ký hiệu *rcv* % cũng thể hiện mức độ phân tán nhưng tính theo tỷ số giữa độ rộng phần tư chuẩn hóa $w_{1/4ch}$ và trung vị $\tilde{x}$ :

$$rcv\ \% = \frac{w_{1/4ch}}{\tilde{x}} 100$$

***Bài tập 3.1****: Cho dãy kết quả đo gồm* 1,22 1,45 1,28 1,20 1,42 1,38 1,34 1,25 1,30 1,40. *Tìm số trung bình, trung vị, số phần tư dưới, số phần tư trên, độ rộng, độ rộng phần tư, độ lệch chuẩn, hệ số biến động và hệ số biến động thô.*

Trong thống kê, tham số cơ bản của mẫu thường áp dụng là số trung bình cộng và hệ số biến động. Đôi khi, người ta còn hay áp dụng các tham số cơ bản « thô » như trung vị, số phần tư, độ rộng phần tư và hệ số biến động thô. Khi tìm số lạc của một tập hợp mẫu mà không chắc rằng mẫu này có thuộc phân bố chuẩn hay không, nên áp dụng trắc nghiệm số phần tư và độ rộng phần tư. Trong chương trình xử lý các kết quả đo của trắc nghiệm thành thạo, NATA sử dụng các trung vị và độ rộng phần tư chuẩn hóa để tìm những kết quả đo của những phòng thí nghiệm nào có độ lặp lại và độ tái lập đủ lớn để trở thành số lạc.

## 3. SỐ THỐNG KÊ CỦA NHIỀU MẪU CÙNG THỰC HIỆN

Giả sử khi thử nghiệm, người ta đã thực hiện *k* mẫu . Mỗi mẫu có số quan trắc lặp khác nhau $n_i$. Kết quả đo biểu diễn dưới dạng $x_{ij}$ (với i = 1,..., *k* và j = 1,..., $n_i$).

**Số trung bình của mỗi mẫu** tính theo: $$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

**Số trung bình của k mẫu** tính theo: $$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

Trường hợp $n_i$ bằng nhau và bằng $n_o$: $$\bar{x} = \frac{n_o \sum_{i=1}^{k} \bar{x}_i}{n_o k} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$$

**Phương sai (chệch) của mỗi mẫu**: $$s_{ci}^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i}$$

và phương sai trung bình của k mẫu: $$\overline{s_{ci}^2} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i s_{ci}^2$$

Trường hợp $n_i$ bằng nhau và bằng $n_o$: $$\overline{s_{ci}^2} = \frac{1}{k n_o} n_o \sum_{i=1}^{k} s_{ci}^2 = \frac{1}{k} \sum_{i=1}^{k} s_{ci}^2$$

**Phương sai (chệch) giữa các mẫu**: $$s_{ck}^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

Trường hợp $n_i$ bằng nhau và bằng $n_o$:

$$s_{ck}^2 = \frac{1}{k n_o} n_o \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 = \frac{1}{k} \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2$$

**Phương sai (chệch) của mẫu chung**:

$$s_c^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}) \right]^2$$

$$s_c^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ (x_{ij} - \bar{x}_i)^2 + (\bar{x}_i - \bar{x})^2 + 2(x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) \right]$$

Vì $\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)$ kể cả $\sum (\bar{x}_i - \bar{x})$ đều bằng không, nên:

$$s_c^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ (x_{ij} - \bar{x}_i)^2 + (\bar{x}_i - \bar{x})^2 \right] = \frac{1}{\sum_{i=1}^{k} n_i} \left[ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 \right]$$

$$s_c^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 = \overline{s_{ci}^2} + s_{ck}^2$$

Qua công thức trên, ta thấy phương sai chung của các mẫu không những bị ảnh hưởng bởi phương sai bên trong $s_i^2$ của từng mẫu mà sự chênh lệch lớn giữa số trung bình $\bar{x}_i$ của các mẫu cũng có tác động đáng kể.

***Bài tập 3.2*** *: Sau khi thử nghiệm k = 4 mẫu, mỗi mẫu đo* $n_i = 5$ *lần và được kết quả:*

| Mẫu | Lần 1 | Lần 2 | Lần 3 | Lần 4 | Lần 5 |
|---|---|---|---|---|---|
| 1 | 1,28 | 1,24 | 1,25 | 1,22 | 1,26 |
| 2 | 1,18 | 1,20 | 1,20 | 1,22 | 1,20 |
| 3 | 1,20 | 1,16 | 1,25 | 1,23 | 1,22 |
| 4 | 1,17 | 1,24 | 1,28 | 1,20 | 1,22 |

*Hãy tính số trung bình, trung vị, số phần tư dưới, số phần tư trên, phương sai, hệ số biến động và hệ số biến động thô của từng mẫu và chung 4 mẫu.*

## 4. ĐỘ LẶP LẠI VÀ ĐỘ TÁI LẬP

Trong tính toán xử lý kết quả thử nghiệm, có hai số thống kê cũng thường được áp dụng là độ lặp lại và độ tái lập.

**Độ lặp lại** ký hiệu **r** , là giá trị thể hiện sai lệch tuyệt đối giữa hai kết quả thử nghiệm được thực hiện trên cùng một mẫu trong các điều kiện giống nhau như cùng trang thiết bị đo, cùng một phương pháp, cùng một người thao tác, cùng điều kiện môi trường trong quãng thời gian ngắn. Khi được xét với mức chắc chắn 95% thì *r* được tính theo $r = 2\sqrt{2}.s_r$ ; trong đó $s_r$ gọi là độ lệch chuẩn của độ lặp lại, tính theo công thức:

$$s_r^2 = \frac{\Sigma(n_i - 1)s_i^2}{\Sigma n_i - k}$$

Độ lặp lại *r* được sử dụng để so sánh hai kết quả thử nghiệm do một người thực hiện trong cùng điều kiện thí nghiệm đã nêu trên. Nếu hai kết quả sai nhau không quá *r* thì kết quả cuối cùng là trung bình cộng của hai kết quả đó. Còn nếu điều kiện này không đạt, cần xem xét lại phương pháp và làm lại thử nghiệm từ đầu.

**Độ tái lập** ký hiệu **R** là giá trị thể hiện sai lệch tuyệt đối giữa hai kết quả thử nghiệm được thực hiện trên cùng một mẫu và cùng phương pháp nhưng trong các điều kiện khác nhau về trang thiết bị đo, người thao tác, phòng thí nghiệm và thời gian thực hiện. Khi được xét với mức chắc chắn 95% thì *R* được tính theo :

$$R = 2\sqrt{2}.s_R$$

trong đó $s_R$ gọi là độ lệch chuẩn của độ tái lập, tính theo công thức :

$$s_R^2 = s_L^2 + s_r^2$$

Đến lượt $s_L$ rút ra từ công thức :

$$s_L^2 = \frac{1}{n}\left[\frac{\Sigma n_i(\bar{x}_i - \bar{x})^2}{k-1} - s_r^2\right] \quad \text{với } \bar{x} = \frac{\Sigma n_i \bar{x}_i}{\Sigma n_i} \text{ và } n = \frac{1}{k-1}\left[\Sigma n_i - \frac{\Sigma n_i^2}{\Sigma n_i}\right]$$

Độ tái lập *R* được sử dụng để so sánh hai kết quả thử nghiệm do hai người thực hiện trong cùng điều kiện thí nghiệm hoặc khác điều kiện thí nghiệm như đã nêu trên. Nếu hai kết quả sai nhau không quá *R* thì kết quả cuối cùng là trung bình cộng của hai kết quả đó. Còn nếu điều kiện này không đạt, cần xem xét lại phương pháp thao tác của cả hai người, trang thiết bị của hai phòng thí nghiệm và làm lại thí nghiệm .

Khi các $n_i$ giống nhau và bằng $n_o$ :

$$s_r^2 = \frac{\Sigma s_i^2}{k} \text{ và } s_L^2 = \frac{\Sigma(\bar{x}_i - \bar{x})^2}{k-1} - \frac{s_r^2}{n_o} \text{ trong đó } \bar{x} = \frac{\Sigma \bar{x}_i}{k}$$

***Bài tập 3.3*** *: Dựa vào bảng số liệu trong bài tập 3.2, hãy tính độ lặp lại r và độ tái lập R từ các kết quả đo $x_{ij}$.*

Đặc biệt khi $n_i = 2$ , từ $w_i = |x_{i1} - x_{i2}|$ tính ra

$$s_r^2 = \frac{\Sigma w_i^2}{2k} \text{ và } s_L^2 = \frac{\Sigma(\bar{x}_i - \bar{x})^2}{k-1} - \frac{s_r^2}{2} \text{ với } \bar{x} = \frac{\Sigma \bar{x}_i}{k}$$

*Chú ý* : a. Khi tính, nếu thấy $s_L^2 < 0$ thì cho $s_L^2 = 0$.

b. Nếu $x_1$ là trung bình của $n_1$ lần đo, $x_2$ là trung bình của $n_2$ lần đo (với $n_1, n_2 > 1$) thì độ tái lập R' được tính theo:

$$R' = \sqrt{R^2 - r^2\left(1 - \frac{1}{2n_1} - \frac{1}{2n_2}\right)}$$

c. Nếu có 3 kết quả thử nghiệm tham gia so sánh thì hiệu của kết quả lớn nhất với kết quả bé nhất được xét với R' = 1,2.R

# 5. SỐ LẠC TRONG THỬ NGHIỆM

***Số lạc*** được coi như là những giá trị hoặc quá lớn hoặc quá bé so với các giá trị còn lại của tập hợp các kết quả đo, có xác suất xuất hiện rất thấp. Trong một chừng mực nào đó, có thể xem chúng không đại diện cho chất lượng mẫu, nếu được loại ra khỏi phép tính thống kê thì kết quả thử nghiệm sẽ gần với giá trị thực hơn.

Các tính toán nhằm phát hiện số lạc hầu hết đều dựa trên cơ sở giả thiết các đại lượng đo thuộc phân bố chuẩn hoặc gần với phân bố chuẩn và xác suất rủi ro do việc loại bỏ sai lầm thường lấy bằng 5 %.

Số lạc có thể là một giá trị cá thể trong tập hợp các giá trị của mẫu và cũng có thể là kết quả thử nghiệm của một mẫu cá thể trong tập hợp nhiều mẫu được lấy ra từ cùng một tổng thể.

Đối với tập hợp các giá trị của mẫu, số lạc có thể được phát hiện khi:

## 1. *Phương pháp dùng hệ số z* (z-score)

Hệ số $z_i$ của một giá trị $x_i$ nào đó được tính theo :

$$z_i = \frac{|x_i - \bar{x}|}{s}$$

trong đó $\bar{x}$ là số trung bình và s là độ lệch chuẩn của mẫu . Khi $z_i \geq 3$, tức là $x_i \leq \bar{x} - 3.s$ hoặc $x_i \geq \bar{x} + 3.s$ sẽ bị coi là số lạc với mức tin cậy 99,73% nếu đại lượng đo thuộc phân bố chuẩn.

Shiffler (1988) đã chứng minh $z_i$ phụ thuộc cỡ mẫu n, giá trị tối đa của nó bằng $(n-1)/\sqrt{n}$ nên việc **áp dụng hệ số z không được tốt để phát hiện số lạc đối với những mẫu nhỏ cỡ n từ 3 đến 10**.

Để khắc phục nhược điểm này, *Iglewicz* và *Hoaglin* (1993) đề nghị hệ số Z cải tiến ký hiệu $M_i$ như sau :

$$M_i = \frac{0{,}6745.|x_i - \tilde{x}|}{MAD}$$

trong đó MAD (median absolute deviation) = median($|x_i - \tilde{x}|$) với E(MAD) = 0,6745.$\sigma$. Khi $M_i > 3{,}5$, có thể khẳng định được $x_i$ tương ứng là số lạc.

*Ví dụ : Một mẫu n = 14 với các giá trị:*

{$x_i$} = {3,2 3,4 3,7 3,7 3,8 3,9 4,0 4,0 4,1 4,2 4,7 4,8 14 15}

Nếu tính $Z_i$ và $M_i$, ta có:

| $x_i$ | 3,2 | 3,4 | 3,7 | 3,7 | 3,8 | 3,9 | 4,0 | 4,0 | 4,1 | 4,2 | 4,7 | 4,8 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Z_i$ | -0,59 | -0,54 | -0,46 | -0,46 | -0,43 | -0,41 | -0,38 | -0,38 | -0,35 | -0,33 | -0,20 | -0,17 | 2,21 | 2,47 |
| $M_i$ | 1,80 | 1,35 | 0,67 | 0,67 | 0,45 | 0,22 | 0 | 0 | 0,22 | 0,45 | 1,57 | 1,80 | 22,48 | 24,73 |

Các giá trị 14 và 15 có $M_i > 3{,}5$ nên là những số lạc.

### 2. *Phương pháp Carling* (1998)

Hệ số $z_i$ dùng thích hợp cho những đại lượng đo thuộc phân bố chuẩn hoặc gần với phân bố chuẩn. Với phân bố bất kỳ nên dùng trung vị và số phần tư. Từ tập hợp kết quả đo $\{x_i\}$, hãy xác định trung vị $\tilde{x}$ và độ rộng phần tư $x_{1/4}$ (bằng hiệu số giữa số phần tư trên $x_{14t}$ và số phần tư dưới $x_{1/4d}$). Sẽ là số lạc nếu :

$$x_i \le \tilde{x} - 2{,}3.w_{1/4} \text{ hoặc } x_i \ge \tilde{x} + 2{,}3.w_{1/4}$$

### 3. *Phương pháp Tukey* (1977)

Nếu $x_i \le x_{1/4d} - k.w_{1/4}$ hoặc $x_i \ge x_{1/4t} + k.w_{1/4}$

trong đó k = 1,5 ứng với vòng trong và k = 3 ứng với vòng ngoài. Khi $x_i$ vượt vòng trong, có thể được xem là số lạc, còn khi $x_i$ vượt vòng ngoài thì chắc chắn nó là số lạc. Trong hai bất đẳng thức trên, $x_{1/4d}$ là số phần tư dưới, $x_{1/4t}$ là số phần tư trên và $w_{1/4}$ là độ rộng phần tư của mẫu.

### 4. *Phương sơ đồ hộp hiệu chỉnh của Vanderviere và Huber* (2004)

Hai tác giả này đã đưa vào khái niệm giá trị thô của độ bất đối xứng MC (medcouple) để tìm số lạc. Giả sử $\{x_i\}$ đã được sắp xếp tăng dần $\{x_1 \le x_2 \le x_3 \le ... \le x_n\}$. Tìm trung vị của mẫu $\tilde{x}$, sau đó những giá trị lớn hơn trung vị ký hiệu $x_j$ và nhỏ hơn trung vị ký hiệu $x_i$. Tiếp tục tính:

$$MC\{x_i\} = \text{med}\frac{(x_j - \tilde{x}) - (\tilde{x} - x_i)}{(x_j - x_i)}$$

Khoảng chứa giới hạn dưới L và giới hạn trên U để phát hiện số lạc là như sau:

Khi $MC \ge 0 \rightarrow [L,U] = [x_{1/4d} - 1{,}5.\exp(-3{,}4MC).w_{1/4}\ ,\ x_{1/4t} + 1{,}5.\exp(4MC).w_{1/4}]$

Khi $MC \le 0 \rightarrow [L,U] = [x_{1/4d} - 1{,}5.\exp(-4MC).w_{1/4}\ ,\ x_{1/4t} + 1{,}5.\exp(3{,}5MC).w_{1/4}]$

Với ví dụ trên đây cỡ mẫu n = 14 với các giá trị :

$$\{x_i\} = \{3{,}2\ \ 3{,}4\ \ 3{,}7\ \ 3{,}7\ \ 3{,}8\ \ 3{,}9\ \ 4{,}0\ \ 4{,}0\ \ 4{,}1\ \ 4{,}2\ \ 4{,}7\ \ 4{,}8\ \ 14\ \ 15\}$$

Các giá trị tính được là : $x_{1/4d} = 3{,}725$; $x_{1/4t} = 4{,}575$ ; $w_{1/4} = 0{,}85$; MC = 0,427, từ đó L = 3,44 và U = 11,62. Vậy các số lạc là 3,2; 3,4; 14 và 15.

***Bài tập 3.4***

*a. 10 kết quả đo sau đây đã được sắp xếp theo thứ tự tăng. Hãy tìm xem có số lạc không?*

568 570 570 570 572 572 572 578 584 596

*b. Cũng xét phát hiện số lạc như trên với dãy kết quả đo sau:*

114,0 114,2 114,6 116,8 115,0 119,2 115,6 113,6 114,0 114,4