Chương 1

ÁP DỤNG MS-EXCEL TRONG THỐNG KÊ MÔ TẢ

- **Tính các giá trị thông kê mô tả**
- **Xác định và độ chính xác**

A- TÍNH TOÁN GIÁ TRỊ THỐNG KÊ MÔ TẢ

3.1 Khái niệm thống kê

3.1.1 Giá trị trung bình *(Mean, Average)*

Một người tợ săn bắn một con vịt bằng hai loạt đạn, mỗi loạt gồm 5 viên. Loạt thứ nhất cách con vịt một mét về phía trước, loạt thứ hai cách con vịt một mét về phía sau. Trên thực tế con vịt chưa chết nhưng theo *kết quả trung bình* thì con vịt đã chết. Cái mà người thợ săn cần bắn một phát trúng con vịt để được nuôi mồi, cái mà nhà khoa học cần thông thường là sự ước tính để có *giá trị trung bình.*

Giá trị trung bình là giá trị thống kê mô tả hay được dùng nhất để mô tả đặc tính của một mẫu từ dân số. Giả sử bạn có một mẫu gồm N giá trị quan sát được sắp xếp thành một chuỗi thống kê: $X_1, X_2, X_3, \ldots, X_N$

Giá trị trung bình của mẫu được tính bởi biểu thức:

Công thức: $\overline{X} = \dfrac{\sum_{i=1}^{N} X_i}{N}$

Giá trị trung bình của mẫu, $\overline{X}$, là trị số ước tính của giá trị trung bình thực sự của dân số μ .

Thí dụ 3: Khối lượng trung bình (mg) của 9 viên nén thuộc lô A

201, 203, 209, 204, 202, 206, 200, 207, 207

$$\overline{X} = \frac{1839}{9} = 204{,}33mg$$

Thí dụ 4: Khối lượng trung bình *(mg)* của chính viên nén thuộc lô B:

151, 153, 259, 154, 202, 256, 150, 257, 257

$$\overline{X} = \frac{18390}{9} = 204{,}33mg$$

3.1.2 Giá trị trung vị *(Median)*

Giá trị trung vị diễn tả khái niệm trung tam của chuỗi dữ liệu. Nếu một chuỗi dữ liệu có N giá trị quan sát được sắp xếp từ nhỏ đến lớn thì giá trị trung vị số thứ tự (N + 1). Trong thí dụ thứ 3, giá trị trung vị là số thứ 5:

1	2	3	4	5	6	7	8	9
200,	201,	202,	203,	204,	256,	207,	207,	209

$$\text{Số thứ tự thứ } 5 = \frac{(9+1)}{2}$$

3.1.3 Khoảng khảo sát

Là sự khác biệt giữa hai giá trị quan sát: lớn nhất và nhỏ nhất: r = Max – Min

Trong thí dụ 3: r = 209 – 200 = 9; trong thí dụ 4: r = 259 – 150 = 109. Vậy các chuỗi dữ liệu có thể có khoảng quan sát khác nhau cho dù chúng có cùng giá trị trung bình. Nếu khoảng quan sát lớn độ phân tán sẽ cao.

3.1.4 Độ lệch chuẩn *(Standarl Deviation)*

Độ phân tán dữ liệu thường được diễn tả bởi phương sai (Variance) hai độ lệch chuẩn *(căn số bậc 2 có phương sai)*:

$$SD = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{(N-1)}$$

Trong thí dụ 3: S = 3,08; trong thí dụ 4: S = 52,65. Độ phân tán của dữ liệu trong thí dụ 4 cao hơn độ phân tán của dữ liệu trong thí dụ 3:

3.1.5 Sai số chuẩn giá trị trung bình *(Std. error of the mean)*

Giá trị trung bình của mẫu gần bằng giá trị trung bình của dân số hơn là các giá trị quan sát riêng biệt.

$$SEM = SD(\overline{X}) = S_{\overline{X}} = \frac{S}{\sqrt{N}}$$

Trong thí dụ 3: SEM = 1,03; trong thí dụ 4: SEM = 17,55

Khi cỡ màu càng lớn *(N tăng)* thì giá trị trung bình càng gần μ

3.1.6 Giới hạn và khoảng tin cậy

Với một mức tin cậy *(Confidence level)* nhất định là α giới hạn tin cậy *(confidence limits)* của một giá trị trung bình được cho bởi tích số:

$t_\alpha S_{\overline{X}}$: *(N < 30: phân phối Student)*

$\overline{X} - t_\alpha S_{\overline{X}}$: Giới hạn dưới *(LCL: lower control limit)*

$\overline{X} + t_\alpha S_{\overline{X}}$: Giới hạn trên *(LCL: upper control limit)*

và khoảng tin cậy *(Confidence interval)* của giá trị trung bình là:

$$\overline{X} \pm t_\alpha S_{\overline{X}} = (\overline{X} - t_\alpha S_{\overline{X}}, \overline{X} + t_\alpha S_{\overline{X}}, = \overline{X} - t_\alpha S_{\overline{X}} \text{ đến } \overline{X} - t_\alpha S_{\overline{X}}$$

Trong thí dụ 3: $t_\alpha S_{\overline{X}} = 2,37$; Trong thí dụ 4: $t_\alpha S_{\overline{X}} = 40,47$

Giá trị thống kê t_α *(phân phối Student)* cần được thay đổi bởi giá trị thống kê, z_α *(phân phối chuẩn)* trong trường hợp mẫu lớn (N > 30)

Lưu ý: Công cụ phân tích "Deseriptive Statistics" trong chương trình MS-EXCEL chỉ áp dụng cho mẫu nhỏ. trong trường hợp mẫu lớn, bạn phải tính lại tích số $t_\alpha S_{\overline{X}}$, với tα được tra trong bảng tích phân Laplace.

3.1.7 Hệ số phân tán *(Coefficient of variatio)*

Hệ số phân tán còn được gọi là sai số tương đối *(relative deviation)*: $CV \frac{S}{X} = 100$

Trong thí dụ 3: CV = 1,50%; trong thí dụ 4: CV = 25,77%

Hệ số phân tán có liên quan đến độ chuẩn *(cũng như độ chính xác của phương pháp đo lường)* và giá trị trung bình của các kết quả.

3.1.8 Giá trị yếu vị *(Mode)*

Giá trị yếu vị là giá trị có tần số cao nhất trong một chuỗi dữ liệu. trong thí dụ 3: giá trị yếu vị là 207; trong thí dụ 4: giá trị yếu vị là 257.

3.1.9 Giá trị KURT *(Kurtosis)*

Giá trị KURT diễn tả đặc điểm thuộc về đỉnh của dạng phân phối dữ liệu. Giá trị số dương dữ liệu phân phối tương đối có đỉnh. Ngược lai, có giá trị âm khi dữ liệu phân phối tương đối phẳng.

$$KURT = \left\{ \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum \left(\frac{X_1 - \overline{X}}{S} \right) \right\} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

3.1.10 Giá trị SKEW *(Skewness)*

Giá trị SKEW phản ánh mức độ bất đối xứng của dạng phân phối dữ liệu xung quanh giá trị trung bình. Giá trị SKEW có giá trị dương khi dữ liệu phân phối bất đối xứng với đuôi nằm lệch về phía giá trị dương. Ngược lại, nó có giá trị âm dữ liệu phân phối bất đối xứng đuôi nằm lệch về phía các giá trị âm.

$$SKEM = \frac{N}{(N-1)(N-2)} \sum \left(\frac{X_1 - \overline{X}}{S} \right)$$

3.2 Áp dụng MS-EXCEL

Hãy tính khoảng tin cậy với mức α = 0,01, độ lệch chuẩn và hệ phân tán của hai chuỗi dữ liệu trong thí dụ 3 và thí dụ 4:

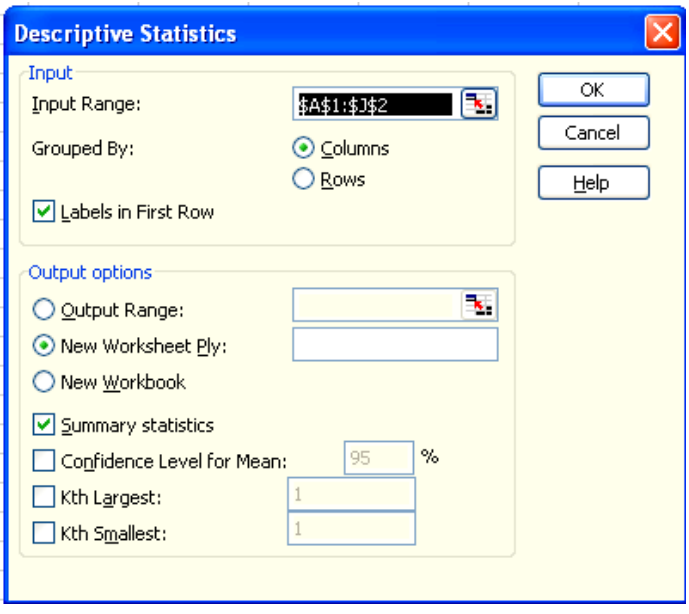
3.2.1 Nhập dữ liệu vào bảng tính

	A	B	C	D	E	F	G	H	I	J
1	TD3	201	203	209	204	202	206	200	207	207
2	TD4	151	153	259	154	202	256	150	257	257

3.2.2 Áp dụng "Descriptive statistics"

a- Nhấp lần lượt đơn lệch Tools và lệch Data Analysis

b- Chọn chương trình Descriptive statistics trong hộp thoại Data Analysis rồi nhấp nút OK



Hình hợp thư thoại data Analysis

c- Trong hộp thư thoại Data Analysis, ấn định lần lượt các chi tiết:

- Phạm vi đầu vào *(Input Range)*,

- Cách sắp xếp theo hàng hay cột *(Group By)*,

- Nhãn dữ liệu *(Labels in Fisrt Row/Column)*,

- Mức tin cậy *(Confidence Level)*,

- Phạm vi đầu ra *(Output Range)*,

- Kết quả tóm tắc *(Summary Satt.istics*).

3.2.3 Tính các giá trị thống kê

Từ đầu ra của MS-EXCEL, bạn phải tính giới hạn tin cậy $t_\alpha S_{\overline{X}}$ đồng thời tính thêm hệ số phân tán $CV = \frac{S}{X} = 100$ bằng cách:

- Chọn B18 trong bảng tính chứa đầu ra của MS-EXCEL, nhập biểu thức = 2.306*B4 dùng con trỏ kéo nút tự điều đến ô D18 *(2,306: giá trị của t với α = 0,05; B4: tọa độ của giá trị $S_{\overline{X}}$).*

- Chọn B19 trong bảng tính chứa đầu ra của MS-EXCEL, nhập biểu thức $= \left(\frac{B7}{B3}\right) x100$ rồi dùng con trỏ kéo nút tự điều đến ô D19 (B7: tọa độ của giá trị S; B3: tọa độ của giá trị $\overline{X}$).

	A	B	C	D
1	***TD3***		***TD4***	
2				
3	Mean	204.3333333	Mean	204.3333333
4	Standard Error	1.027402334	Standard Error	17.54992877
5	Median	204	Median	202
6	Mode	207	Mode	257
7	Standard Deviation	3.082207001	Standard Deviation	52.64978632
8	Sample variance	9.5	Sample variance	2772
9	Kurtosis	-1.330431342	Kurtosis	-2.423671859
10	Skewness	0.058546077	Skewness	0.020438933
11	Range	9	Range	109
12	Minimum	200	Minimum	150
13	Maximum	209	Maximum	259
14	Sum	1839	Sum	1839
15	Count	9	count	9
16	Confidence Level (95.0	2.369189782	Confidence Level (95.0	4047020829
17				
18				
19				

Kết quả:

Giá trị thống kê	*Thí dụ 3*	*Thí dụ 4*
Giới hạn tin cậy 95% ($\overline{X} \pm t_\alpha S_{\overline{X}}$)	204,33 ± 2,37	204,33 ± 40,47
Độ lệch chuẩn (S)	3,08	52,65
Hệ số phân tán	1,50%	25,77%

Lưu ý: Trong thực tế, dữ liệu có khi không đồng nhất nên độ lệch chuẩn S và hệ số phân tan CV có giá trị lớn. Dữ liệu có thể được chuyển dạng một cách phù hợp để phân phối được đồng nhất hơn.

Sự chuyển dạng Logrit (Log transformation)

Dạng logarit *(log, cơ số 10 hay ln, cơ số e)* của X hay 1/X thường được áp dụng khi dữ liệu có giá trị SKEW dương và giá trị trung bình $\overline{X}$ tỉ lệ thuận với độ lệch chuẩn S. Thí dụ:

0,3 12 7 0,41 21 12 4 10 17 9 0,7 6 23 ($\overline{X} = 9{,}42; S = 7{,}5$

Dữ liệu	$\overline{X}$	S
Dạng log (X)	0,71	-0,71
Dạng log (1/X)	0,64	0,64

Sự chuyển dạng căn số (Square-root transformation)

Dạng căn số bậc hai cảu X phù hợp hơn dạng logarit tương ưng khi chuỗi dữ liệu có nhiều giá trị nhỏ. Nếu chuỗi dữ liệu có các số < 10 hay o thì nên dùng dạng căn số của (X + 1). Thí dụ:

0 11 7 4 0 12 4 2 7 9 3 0 12 5 8 10 ($\overline{X} = 9{,}42; S = 7{,}5$

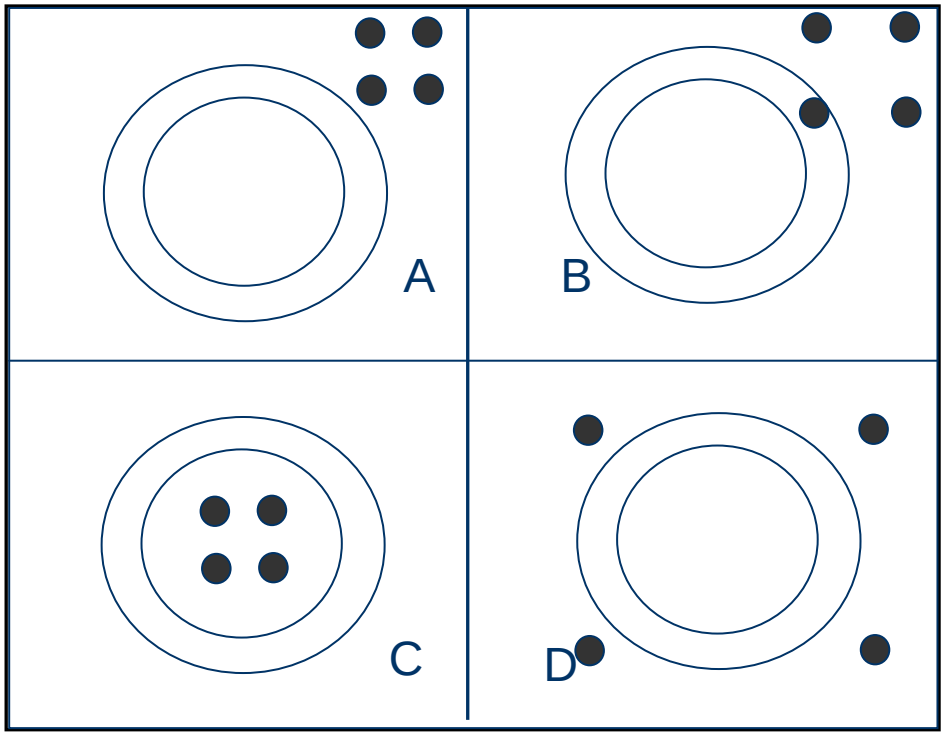
Dữ liệu	$\overline{X}$	S
Dạng $\sqrt{X}$	2,12	1,21
Dạng $\sqrt{X+1}$	2,14	0,92

Các giá trị như xác xuất hay tỉ số hoặc tỉ lệ *(dưới hình thức số lẻ thập phân*) có thể được ***chuyển Dạng arcsin*** *(Arcsine transformation),* $\sqrt{P}$

Các giá thống kê mô tả của dữ liệu được chuyển dạng sẽ khác với dữ liệu nguyên thủy. Bạn nên thận trọng khi chuyển dạng vì điều này có thể ảnh hưởng đến kết quả của các trắc nghiệm thống kê suy lí *(so sánh hai giá trị trung bình, so sánh hai tỉ số, só sánh hai phương sai).*

B- XÁC ĐỊNH ĐỘ ĐÚNG VÀ ĐỘ CHÍNH XÁC

3.3 Khái niệm thống kê



A: chính xác, không đúng B: Không chính xác, không đúng

C: *Chính xác, đúng* D: Không chính xác, không đúng

Độ chính xác *(Precision)*: độ lập lại của các giá trị quan sát

$$P = 100 - CV = \left(1 - \frac{S}{X}\right) \times 100$$

Độ đúng *(Accuracy)*: độ trùng hợp giữa các giá trị quan sát *(hay thực nghiệm)* với giá trị lý thuyết.

$$A = \frac{Giá\ tri\ \ trung\ bình}{Giá\ tri\ lí\ \ \text{thuy h}} \times 100$$

Độ chính xác và độ đúng phản ánh chất lượng của một phương pháp đo lường. một phương pháp tốt vừa chính xác vừa đúng.

Áp dụng MS-EXCEL

Thí dụ 5: người ta xác định hàm lượng hoạt chất trong một mẫu thuốc bằng cách tiến hành song song 10 lần trong hai điều kiện; mẫu không được thêm và được thêm một lượng hoạt chất biết trướclà 10 *mg*

X	9,8	9,7	9,9	10,0	10,1	9,8	10,2	10,0	9,8	9,8
X’	19,5	19,6	19,6	20,0	19,9	19,7	20,0	19,7	19,6	19,5

Tính độ chính xác và độ đúng của phương pháp: (X’ = X + 10*mg*)

Nhập dữ liệu vào bảng tính:

	B3		$\downarrow \lvert X \rVert \nu \rVert fx \rvert = B1 - B2$								
	A	B	C	D	E	F	G	H	I	J	K
1	X’	19,5	19,6	19,6	20	19,9	19,7	20	19,7	19,6	19,5
2	X	9,8	9,7	9,9	10	10,1	9,8	10,2	10	9,9	9,8
3	Xo	B2									
4											

Làm một số phép tính

- Chọn B3 rồi nhấp biểu thức = B1 – B2

- Dùng con trỏ để kéo nút điền tự động từ ô B3 đến ô K3.

	A	B	C	D	E	F	G	H	I	J	K
1	X’	19,5	19,6	19,6	20	19,9	19,7	20	19,7	19,6	19,5
2	X	9,8	9,7	9,9	10	10,1	9,8	10,2	10	9,9	9,8
3	Xo	9,7	9,9	9,7	10	9,8	9,9	9,8	9,7	9,7	9,7

Áp dụng “Descriptive Statistics”

Tương tự phần “Tính các giá trị thống kê mô tả”

Kết quả:

$$P = \left(1 - \frac{0,15}{9,92}\right) \times 100 = 98,5\%$$

$$A = \left(\frac{9,79}{10}\right) \times 100 = 97,9\%$$