Chương 2

ÁP DỤNG MS-EXCEL TRONG THỐNG KÊ SUY LÍ

- So sánh giá trị trung bình
  - *Phương sai biết trước*
  - *Dữ liệu tương ứng từng cặp*
  - *Phương sai bằng nhau*
  - *Phương sai khác nhau*
- So sánh tỉ số
- So sánh phương sai

A- SO SÁNH GIÁ TRỊ TRUNG BÌNH VỚI PHƯƠNG SAI BIẾT TRƯỚC

4.1 Khái niệm thống kê

Nếu mẫu lớn (N > 30) thì phương sai của mẫu , S_i^2 , có thể được xem là phương sai của dân số , σ_i^2 , khi ấy bạn có thể áp dụng trắc nghiệm z để so sánh giá trị trung bình của hai mẫu với phương sai biết trước.

Giả thuyết:

Trắc nghiệm bên phải

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Trắc nghiệm bên trái

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

Trắc nghiệm hai bên

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Giá trị thống kê:

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$ Phân phối chuẩn

Biện luận

Nếu $z < z_\alpha$ (hai bên) hay $z_{\alpha/2}$ (một bên) $\Rightarrow$ Chấp nhận giả thuyết H_0

4.2 Áp dụng Ms-EXCEL

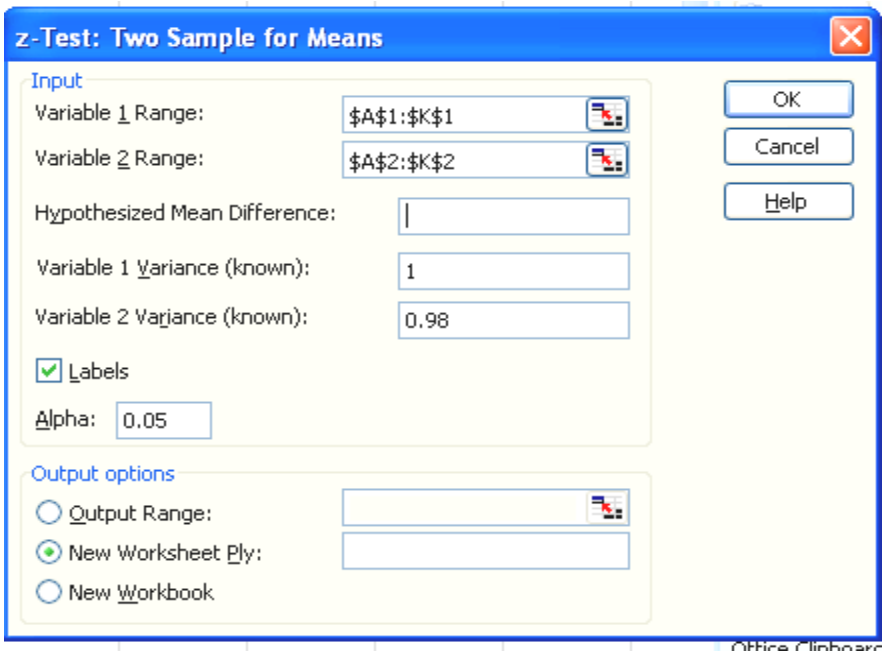
Thí dụ 6: Người ta chọn hai mẫu, mỗi mẫu có 10 máy, từ hai lô (I và II được sản xuất với phương sai biết trước tương ứng là 1 và 0,98) để khảo sát thời gian hoàn thành công việc (phút) của chúng:

I	6	8	9	10	6	15	9	7	13	11
II	5	5	4	3	9	9	6	13	17	12

Hỏi khả năng hoàn thành công việc của hai máy có khác nhau?

Nhập dữ liệu vào bảng tính

	Book1										
	A	B	C	D	E	F	G	H	I	J	K
1	I	6	8	9	10	6	15	9	7	13	11
2	II	5	5	4	3	9	9	6	13	17	12



Hình 4.1:* Hộp thoại z-Test: *Two Sample for Means

Áp dụng "z-test: Two Sample for Means"

a. Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysis.

b. Chọn chương trình z-Test: Two Sample for Means trong hộp thoại DataAnalysis rồi nhấn nút OK.

c. Trong hộp thoại z-Test: Two Sample for Means, ấn lần lượt các chi tiết cũng giống như " hai mẫu có dữ liệu tương ứng", song có thêm các chi tiết:

- Phương sai của dữ liệu 1 *(Variance 1 Variance),*
- Phương sai của dữ liệu 2 *(Variance 2 Variance),*

z-Test: Two Sample for Means		
	I	II
Mean	9.4	8.3
Known Variance	1	0.98
Observations	10	10
Hypothesized Mean Difference	0	
z	2.472066162	
P(Z<=z) one-tail	0.006716733	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	0.013433465	
z Critical two-tail	1.959963985	

Kết quả

H_0: $\mu_1 = \mu_2$ "Khả năng hoàn thành của hai máy như nhau".

H_1: $\mu_1 \neq \mu_2$ "Khả năng hoàn thành của hai máy khác nhau".

$z = 2{,}472 > z_{0,05} = 1{,}960 \Rightarrow$ Bác bỏ giả thuyết H_0.

Vậy khả năng hoàn thành của hai máy khác nhau.

B- SO SÁNH GIÁ TRỊ TRUNG BÌNH DỮ LIỆU TƯƠNG ỨNG TỪNG CẶP

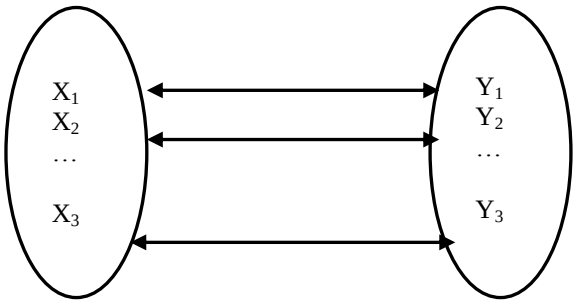
4.3 Khái niệm thống kê

Trong trường hợp hai mẫu nhỏ (N < 30) phụ thuộc (thí dụ: kết quả của một nhóm chuột được xét nghiệm máu hai lần - trước và sau khi uống thuốc - hay một nhóm bệnh nhân trải qua hai thí nghiệm - được thử thuốc trên tay này và giả được trên kia) và không giả định rằng phương sai của hai mẫu bằng nhau, bạn có thể áp dụng trắc nghiệm t để so sánh giá trị trung bình của hai mẫu dữ liệu tương ứng từng cặp.

Giả thuyết

Tương tự trường hợp " hai mẫu với phương sai biết trước".

Giá trị thống kê



$$\mathbf{D_i = X_i - Y_i\ (I = 1, 2, \ldots N)}$$

$$\overline{D} = \frac{\sum_{i=1}^{N} D_i}{N}$$

$$S_D = \sqrt{\frac{\sum_{i=1}^{N} (D_i - \overline{D})^2}{(N-1)}}$$

$$t = \frac{\overline{D} - \mu_D}{S_D / \sqrt{N}} = \frac{\overline{D}}{S_D / \sqrt{N}}$$ **Phân phối Student với γ = N – 1**

Biện luận

Nếu $t < t_\alpha$ hay $t_{\alpha/2}$ ($\gamma = N - 1$) $\Rightarrow$ Chấp nhận giả thuyết H_0

4.4 Áp dụng MS-EXCEL

Thí dụ 7: Hàm lượng *(mg)* của một chế phẩm được xác định trước và sau khi được lão hoá cấp tốc như sau:

Trước	7,5	6,8	7,1	7,5	7,2	6,8	6,9	6,7	6,8	6,8
Sau	6,1	6,3	6,5	6,4	6,8	6,3	6,1	6,4	6,5	6,3

Hãy cho biết hàm lượng hoạt chất có giảm sau thí nghiệm?

Nhập dữ liệu vào bảng tính

	A	B	C	D	E	F	G	H	I	J	K
1	Trước	7.5	6.8	7.1	7.5	7.2	6.8	6.9	6.7	6.8	6.8
2	Sau	6.1	6.3	6.5	6.4	6.8	6.3	6.1	6.4	6.5	6.3

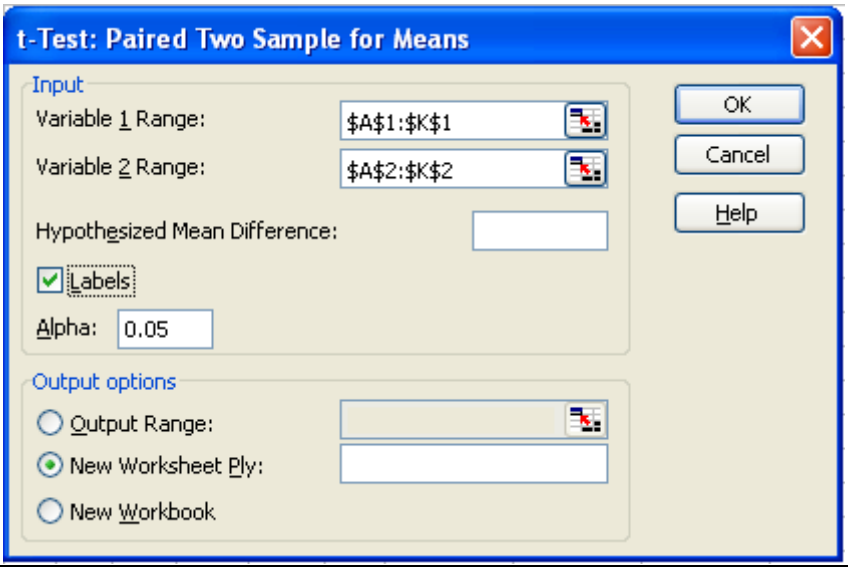
4.5 Áp dụng "t-Test: Paired Two Sample for Means"

a. Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysis.

b. Chọn chương trình t-Test: Paired Two Sample for Means trong hộp thoại Data Analysis rồi nhấp nút OK.

c. Trong hộp thoại t-Test: Paired Two Sample for Means, lần lượt ấn định các chi tiết:

- Phạm vi của dữ liệu 1 *(Variable 1 Range),*
- Phạm vi của dữ liệu 2 *(Variable 2 Range),*
- Nhãn dữ liệu *(Labels),*
- Ngưỡng tin cậy *(Alpha),*
- Sai biệt giữa hai giá trị trung bình ước tính *(Hypothesized Mean Difference),*
- Phạm vi đầu ra *(Output Range).*



Hình 4.2: **Hộp thoại t-Test:** ***Paired Two Sample for Means***

t-Test: Paired Two Sample for Means		
	Trước	*Sau*
Mean	7.01	6.37
Variance	0.089888889	0.042333333
Observations	10	10
Pearson Correlation	0.023415671	
Hypothesized Mean Difference	0	
df	9	
t Stat	5.627619665	
P(T<=t) one-tail	0.000161339	
t Critical one-tail	1.833112923	
P(T<=t) two-tail	0.000322677	
t Critical two-tail	2.262157158	

Kết quả và biện luận:

H_0 : $\mu_1 = \mu_2$ "Hàm lượng thuốc không giảm sau thí nghiệm".

H_1 : $\mu_1 > \mu_2$ "Hàm lượng thuốc giảm sau thí nghiệm".

$t = 5{,}628 > t_{0,05} = 1{,}833 \Rightarrow$ Bác bỏ giả thuyết H_0.

Vậy hàm lượng thuốc giảm sau thí nghiệm**.**

C- SO SÁNH GIÁ TRỊ TRUNG BÌNH VỚI PHƯƠNG SAI BẰNG NHAU

4.6 Khái niệm thống kê

Trong trường hợp hai mẫu nhỏ (N < 30) độc lập và có phương sai bằng nhau*, bạn có thể áp dụng trắc nghiệm t đồng phương sai (homoscedastic t-test) để so sánh giá trị trung bình của hai mẫu ấy.

Giả thuyết

Như trường hợp "hai mẫu có dữ liệu tương ứng từng cặp"

Giá trị thống kê

$$= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$= \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{S_p^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Phân phối Student

$$\gamma = N_1 + N_2 - 2$$

$$S_p^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^1}{N_1 + N_2 - 2}$$

Biện luận

Nếu $t < t_\alpha$ hay $t_{\alpha/2}$ ($\gamma = N_1 + N_2 - 2$) $\Rightarrow$ Chấp nhận giả thuyết H_0.

4.7 Áp dụng MS-EXCEL

Thí dụ 8: Người ta cho 10 bệnh nhân uốngthuốc hạ cholesterol đồng thời cho bệnh nhân khác uống giả được *(placebo)* rồi xét nghiệm về nồng độ cholesterol trong máu *(g/L)* của cả hai nhóm:

Thuốc	1,10	0,99	1,05	1,01	1,02	1,07	1,10	0,98	1,03	1,12
Giả được	1,25	1,31	1,28	1,20	1,18	1,22	1,22	1,17	1,19	1,21

Theo bảng kết quả trên, thuốc có tác dụng hạ cholesterol trong máu?

Nhập dữ liệu vào bảng tính

	A	B	C	D	E	F	G	H	I	J	K
1	Thuốc	1,10	0,99	1,05	1,01	1,02	1,07	1,10	0,98	1,03	1,12
2	Giả được	1,25	1,31	1,28	1,20	1,18	1,22	1,22	1,17	1,19	1,21

4.8 Áp dụng *"t-Test: Two-Sample Assuming Equal Variances"*

a. Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysis.

b. Chọn chương trình t-Test: Two-Sample Assuming Equa! Variances trong hộp thoại Data Analysis rồi nhấp nút OK.

c. Trong hộp thoại t-Test: Two-Sample Assuming Equa! Variances, ấn định lần lượt các chi tiết như "hai mẫu có dữ liệu tương ứng".

t-Test: Two-Sample Assuming Equal Variances		
	Thuốc	*Giả được*
Mean	1.047	1.223
Variance	0.002401111	0.002001111
Observations	10	10
Pooled Variance	0.002201111	
Hypothesized Mean Difference	0	
df	18	
t Stat	-8.388352782	
P(T<=t) one-tail	6.19807E-08	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	1.23961E-07	
t Critical two-tail	2.100922037	

Kết quả và biện luận

$H_0 : \mu_1 = \mu_2$ "Thuốc và giả được tác dụng như nhau".

$H_1 : \mu_1 < \mu_2$ "Thuốc có tác dụng hạ cholesterol".

$t = 8{,}388 > t_{0,05} = 1{,}734 \Rightarrow$ Bác bỏ giả thuyết H_0.

Vậy thuốc tác dụng hạ cholesterol

D- SO SANH GIÁ TRỊ TRUNG BÌNH VỚI PHƯƠNG SAI KHÁC NHAU

4..9 Khái niệm thống kê

Với hai mẫu nhỏ (N < 30) độc lập và có phương sai khác nhau (hai mẫu phân biệt), bạn có thể áp dụng trắc nghiệm t dị phương sai (betero – scedastic – test) để so sánh giá trị trung bình của hai mẫu ấy.

Giả thuyết

Tương tự như trường hợp "hai mẫu với phương sai bằng nhau".

Giá trị thống kê

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} \quad t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Phân phối Student

$$\gamma = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{\left(S_1^2/N_1\right)^2}{N_1 - 1} + \frac{\left(S_2^2/N_2\right)^2}{N_2 - 1}}$$

(Smith - Satterthwaite)

Biện luận

Nếu $t < t_\alpha$ hay $t_{\alpha/2}$ (γ ước tính) $\Rightarrow$ Chấp nhận giả thuyết H_0.

4.10 Áp dụng MS-EXCEL

Thí dụ 9: Thời gian tan rã *(phút)* của một loại viên bao từ hai xí nghiệp dược phẩm *(XNDP)* khác nhau được kiểm nghiệm như sau:

XNDP I	61	71	68	73	71	70	69	74
XNDP II	62	69	65	65	70	71	68	73

Thời gian tan rã của viên bao thuộc hai XNDP có giống nhau?

Nhập dữ liệu vào bảng tính

	A	B	C	D	E	E	G	H	I	J
1	**XNDP I**	61	71	68	73	71	70	69	74	
2	**XNDP II**	62	69	65	65	70	71	68	73	

4.11 Áp dụng *"t-Test: Two-Sample Assuming Unequal Variances"*

a. Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysis.

b. Chọn chương trình t-Test: Two-Sample Assuming Unequal Variances trong hộp thoại Data Analysis rồi nhấp nút OK.

c. Trong hộp thoại t-Test: Two-Sample Assuming Unequal Variances, ấn định lần lượt các chi tiết như " hai mẫu có dữ liệu tương ứng".

t-Test: Two-Sample Assuming Unequal Variances		
	XNDP I	*XNDP II*
Mean	69.625	67.875
Variance	15.98214286	13.26785714
Observations	8	8
Hypothesized Mean Difference	0	
df	14	
t Stat	0.915208631	
P(T<=t) one-tail	0.187788433	
t Critical one-tail	1.761310115	
P(T<=t) two-tail	0.375576865	
t Critical two-tail	2.144786681	

Kết quả và biện luận

$H_0 : \mu_1 = \mu_2$ "Thời gian tan rã của viên bao thuộc hai XNDP giống nhau".

$H_1 : \mu_1 \neq \mu_2$ "Thời gian tan rã của viên bao thuộc hai XNDP khác nhau".

$t = 0{,}915 < t_{0,05} = 2{,}145 \Rightarrow$ Chấp nhận giả thuyết H_0.

Vậy thời gian tan rã của viên bao thuộc hai XNDP giống nhau.

E- SO SÁNH TỈ SỐ

4.12 Khái niệm thống kê

Đối với một thí nghiệm có hai kết quả *(binomial experiment)* – thí dụ, đối với một thuốc được kê đơn: có hay không - bạn thường so sánh hai tỉ số với nhau *(thực nghiệm với lí thuyết hay thực nghiệm với thực nghiệm)*. Song đối với một thí nghiệm có nhiều kết quả (multinomial experiment)-thí dụ, bác sĩ đánh giá tình trạng của các bệnh nhân được điều trị bởi thuốc trong một khoảng thời gian - bạn cần so sánh nhiều tỉ số. Trắc nghiệm "khi" bình phương (X^2) cho phép bạn so sánh không những hai mà còn nhiều tỉ số *(hay tỉ lệ hoặc xác suất*) một cách tiện lợi. X^2 là phân phối về xác suất, không có tính đối xứng và chỉ có giá trị ≥ 0. Giả sử bạn có một công trình nghiên cứu với N thử nghiệm độc lập, mỗi thử nghiệm có k kết quả và mỗi kết quả mang một các xác suất thực nghiệm là P_i(i = 1, 2, …k). Nếu gọi $P_{i,0}$ là các giá trị lí thuyết tương ứng với P_i thì các tần số lí thuyết sẽ là $E_i = NP_{i,0}$. Điều kiện để áp dụng trắc nghiệm X^2 một cách thành công là các tần số lí thuyết E_i phải ≥ 5.

Giả thuyết

$H_0 : P_1 = P_{1,0}, P_2 = P_{2,0},..., P_{k,0} \Leftrightarrow$ "Các cặp P_i và $P_{i,0}$ giống nhau".

H_1: " Ít nhất có một cặp P_i và $P_{i,0}$ khác nhau".

Giá trị thống kê

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right];$$

O_i: các tần số thực nghiệm *(observed frequency);*E_i: các tần số lí thuyết *(expected frequency)*

Biện luận

Nếu $\chi^2 > \chi_a^2 \Rightarrow$ Bác bỏ giả thuyết H_0 *(DF = k – 1)*

Trong chương trình MS-EXCEL có hàm số CHITEST có thể tính:

- Giá trị χ^2 theo biểu thức: $\chi^2 = \sum_{j=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

O_{ij}: tần số thực nghiệm của ô thuộc hàng i và cột j;

E_{ij}: tần số lí thuyết của ô thuộc hàng I với cột j, r: số hàng; và c: số cột.

- Xác suất $P(X > \chi^2)$ với bậc tự do DF = (r – 1)(c – 1); trong đó, r là số hàng và c là số cột trong bảng ngẫu nhiên (contingency table).

Nếu $P(X > \chi^2) > \alpha \Rightarrow$ Chấp nhận giả thuyết H_0, và ngược lại.

4.13 Áp dụng MS-EXCEL

Thí dụ 10: Kết quả điều trị trên hai nhóm bệnh nhân: ruột nhóm dùng thuốc và một nhóm giả dược được tóm tắt như sau:

Biện pháp điều trị	Số bệnh nhân khỏi bệnh	Số bệnh nhân không khỏi
Thuốc	24	15
Giả dược	20	23

Tỉ lệ khỏi bệnh do thuốc (24/39 = 61%) và giả dược (20/43 = 46%) có khác nhau về mặt thống kê?

Nhập dữ liệu vào bảng tính

B9	⇓	**=CHIEST (B3:C4,B7:C8)**		
	A	**B**	**C**	**D**
1	**THỰC NGHIỆM**	**Khỏi bệnh**	**Không khỏi**	**Tổng hàng**
2	Điều trị	24	15	39
3	Thuốc	20	23	43
4	Giả dược	44	38	82
5	Tổng cột			
6	LÝ THUYẾT			
7	Thuốc	20.92682927	18.07317073	
8	Giả dược	23.07317073	19.92682927	
9	GIÁ TRỊ "P"	0.172954847		

Sắp xếp dữ liệu theo bảng trắc nghiệm hai mẫu độc lập.

Tính các tổng số

Tổng hàng *(Row totals):* chọn ô D_3 và nhập biểu thức = SUM(B3:C3).

Dùng con trỏ để kéo nút tự điền từ ô D_3 đến ô D_4.

Tổng cột *(Column totals):* chọn ô B5 và nhập biểu thức = SUM(B3:B4).

Dùng con trỏ để kéo nút tự điền từ ô B5 đến ô C5.

Tổng cộng *(Grand total):* chọn ô D5 và nhập biểu thức = SUM(D3:D4)

Tính các tần số lí thuyết

Tần số lí thuyết = (tổng hàng × tổng cột)/tổng cộng

Khỏi bệnh do thuốc: chọn ô B7 và nhập biểu thức D3*B5/D5

Không khỏi bệnh do thuốc: chọn ô C7 và nhập biểu thức = D3*C5/D5

Khỏi bệnh do giả dược: chọn ô B8 và nhập biểu thức = D4*B5/D5

Không khỏi bệnh do giả dược: chọn ô C8 và nhập biểu thức = D4*C5/D5

Hình

4.13 Áp dụng hàm số "CHITEST"

Tính xác suất $P(X>\chi^2)$ bằng cách chọn ô B9 và nhập biểu thức như trên hay sử dụng hộp thoại của CHITEST.

Kết quả: $P(X > \chi^2) = 0,17 > \alpha = 0,05$ … nhận giả thuyết H_0.

Vậy tỉ lệ khỏi bệnh do thuốc và do giả dược không khác nhau.

F- SO SÁNH PHƯƠNG SAI

4.14 Khái niệm thống kê

Trắc nghiệm so sánh hai phương sai thường được áp dụng để so sánh độ chính xác của hai phương pháp định lượng khác nhau.

Giả thuyết:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

Giá trị thống kê: $$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} = \frac{S_1^2}{S_2^2} = \frac{S_1^2}{S_2^2}$$

Phân phối Fischer: $\gamma_1 = N_1 - 1; \quad \gamma_2 = N_2 - 2$

Biện luận

Nếu $F < F_\alpha(\gamma_1, \gamma_2) \Rightarrow$ Chấp nhận giả thuyết H_0 với xác suất $(1 - \alpha)100\%$

4.15 Áp dụng MS-EXCEL

Thí dụ 11: Một mẫu được phân tích bởi hai phương pháp A và B với kết quả được tóm tắt trong bảng sau:

A	6,4	5,2	4,8	5,2	4,3	4,4	5,1	5,8
B	2,6	3,5	3,4	3,2	3,4	2,8	2,9	2,8

Cho biết phương pháp nào chính xác hơn?

Nhập dữ liệu vào bảng tính

	A	B	C	D	E	F	G	H	I
1		6,4	5,2	4,8	5,2	4,3	4,4	5,1	5,8
2		2,6	3,5	3,4	3,2	3,4	2,8	2,9	2,8

Áp dụng "F-Test Two-Sample for Variances"

a. Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysis.

b. Chọn chương trình F-Test Two-Sample for Variances trong hộp thoại Data Analysis rồi nhấp nút OK.

c. Trong hộp thoại F-Test Two-Sample for Variances, lần lượt ấn định các chi tiết:

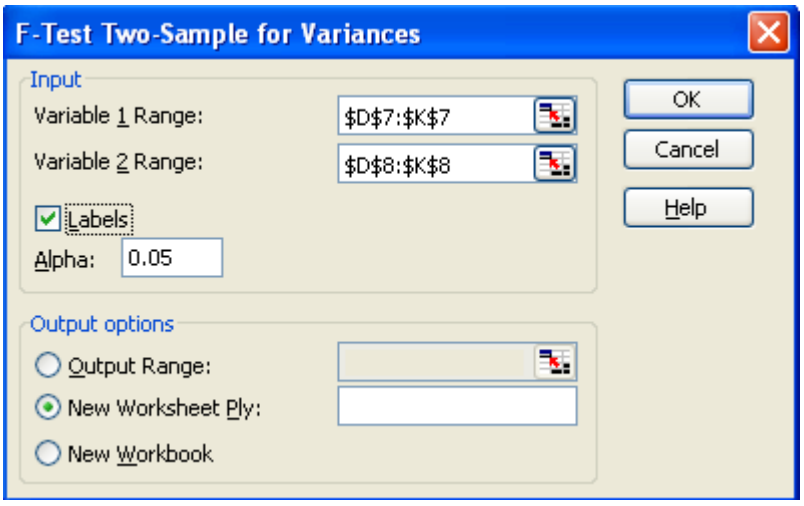
- Tọa độ của dữ liệu 1 *(Variable 1 Range),*

- Tọa độ của dữ liệu 2 *(Variable 2 Range),*

- Nhãn dữ liệu *(Labels),*

- Ngưỡng tin cậy *(Alpha),*

- Tọa độ đầu ra *(Output Range)*



Hình 4.3: Hộp thoại F-Test Two-Sample for Variaces

F-Test Two-Sample for Variances		
	6.4	*2.6*
Mean	4.971428571	3.142857143
Variance	0.269047619	0.092857143
Observations	7	7
df	6	6
F	2.897435897	
P(F<=f) one-tail	0.110575721	
F Critical one-tail	4.283865714	

Kết quả và biện luận

$H_0 : \sigma_A^2 = \sigma_B^2$ "Hai phương pháp có độ chính xác như nhau".

$H_t : \sigma_A^2 > \sigma_B^2$ "Độ chính xác của phương pháp B cao hơn A".

$F = 4{,}171 > F_{0,05} = 3{,}787 \Rightarrow$ Bác bỏ giả thuyết H_0.

Vậy độ chính xác của phương pháp B cao hơn phương pháp A.