BÀI 6 KIÊM DỊNH MỘT PHÂN PHỐI VÀ BẢNG TƯƠNG LIÊN

I- NỘI DUNG

Biến ngẫu nhiên liên tục bằng tổng bình phương của nhiều biến ngẫu nhiên độc lập, phân phối chuẩn tắc được gọi là biến Khi bình phương χ^2.

Biến này được khảo sát tỷ mỷ và lập bảng phân phối χ^2.

Biến χ^2 có nhiều ứng dụng khác nhau, ở đây chúng ta chỉ đề cập đến hai ứng dụng đối với các biến định tính.

a- KIỂM ĐỊNH MỘT PHÂN PHỐI

Để khảo sát một biến định tính X chúng ta lấy một mẫu quan sát gồm N cá thể và căn cứ vào trạng thái của biến X để phân chia thành k lớp (loại) :

(L_i là lớp thứ i, m_i là số lần quan sát thấy X thuộc lớp i).

Biến X	L_1	L_2	. . .	L_k	Tổng
Tần số mi	m_1	m_2	. . .	m_k	$N=\Sigma m_i$

Từ một lý thuyết nào đó, có thể là một lý thuyết đã được xây dựng chặt chẽ, có giải thích cơ chế, cũng có thể chỉ là một lý thuyết mang tính kinh nghiệm, đúc kết từ những quan sát trước đây về biến X, người ta đưa ra một giả thiết H_o thể hiện ở dãy các tần suất lý thuyết $f_1, f_2, \ldots, f_k$ của biến X (có nghĩa là dãy tần suất này được tính từ lý thuyết đã nêu trên). Căn cứ vào tần suất lý thuyết f_i và tần số thực tế m_i chúng ta phải đưa ra một trong hai kết luận:

a) Chấp nhận H_o: tần số thực tế phù hợp với lý thuyết đã nêu (tức là dãy tần số thực tế m_i phù hợp với dãy tần suất f_i).

b) Bác bỏ H_o tức là dãy tần số thực tế m_i không phù hợp với dãy lý thuyết fi đã nêu.

Phù hợp ở đây được hiểu là tỷ lệ giữa các tần số m_i giống như tỷ lệ giữa các tần suất f_i , nói cách khác diễn biến của dãy m_i tương tự như diễn biến của dãy f_i.

Việc kiểm định được thực hiện với mức ý nghĩa α, tức là nếu giả thiết H_0 đúng thì xác suất để bác bỏ một cách sai lầm H_0 bằng α.

a1- Kiểm định χ^2 (còn gọi là Pearson chi square) Kiểm định này dựa trên việc tính gần đúng phân phối nhị thức bằng phân phối chuẩn.

Các bước cần làm gồm:

a/ Tính các tần số lý thuyết theo công thức: $t_i = N. f_i$ (1)

b/ Tính khoảng cách giữa hai số m_i và t_i theo cách tính khoảng cách χ^2

$$\frac{(m_i - t_i)^2}{t_i}$$

c/ Tính khoảng cách giữa hai dãy tần số thực tế m_i và tần số lý thuyết t_i theo công thức :

$$\chi^2_{tn} = \sum_{i=1}^{p} \frac{(m_i - t_i)^2}{t_i} \quad (2)$$

d/ Tìm giá trị tới hạn trong bảng χ^2

(mức ý nghĩa α,bậc tự do k-1, ký hiệu là $\chi^2(\alpha,k-1)$).

e/ Nếu $\chi^2_{tn} \le \chi^2(\alpha, k-1)$ thì chấp nhận H_o:" Dãy tần số thực tế m_i phù hợp với lý thuyết đã nêu".

Nếu $\chi^2_{tn} > \chi^2(\alpha, k-1)$ thì bác bỏ H_o, tức là "Dãy tần số thực tế m_i không phù hợp với lý thuyết đã nêu".

Nếu trong giả thiết H_0 có r tham số cần ước lượng từ mẫu quan sát thì việc tính χ^2 vẫn như cũ nhưng với mỗi tham số cần ước lượng phải bớt đi một bậc tự do tức là phải so χ^2_{tn} với $\chi^2(\alpha, p-1-r)$.

a2- Kiểm định G (còn gọi là Likelihood chi square)

Một kiểm định khác cho kết quả tương tự như kiểm đinh χ^2 thường dùng trong các chương trình máy tính là kiểm định G dựa trên tỷ số hợp lý cực đại.

Các bước cần làm:

a/ Tính lôgarit của tỷ số m_i / t_i tức là lấy $\ln(m_i/t_i)$

b/ Tính $G = 2\sum_{i=1}^{p} m_i \ln(\frac{m_i}{t_i})$

c/ Tính $\chi^2(\alpha, p-1-r)$ rồi so với G để kết luận

Nếu $G \le \chi^2(\alpha, p-1-r)$ thì chấp nhận H0, nếu ngược lại thì bác bỏ H0.

b- BẢNG TƯƠNG LIÊN

Có 2 biến định tính, biến X chia thành k lớp, biến Y chia thành l lớp, qua khảo sát thấy số cá thể có $X = X_i$, $Y = Y_j$ là m_{ij}. Bảng hai chiều chứa m_{ij} gọi là bảng tương liên R_{kxl}

Bảng các tần số m_{ij}

Y / X	Y_1	Y_2	. . .	Y_l	TH_i
X_1	m_{11}	m_{12}	. . .	m_{1l}	TH_1
X_2	m_{21}	m_{22}	. . .	m_{2l}	TH_2
. . .	. . .	. . .	. . .	. . .	. . .
X_k	m_{k1}	m_{k2}	. . .	m_{kl}	TH_k
TC_j	TC_1	TC_2	. . .	TC_l	N

Bài toán đặt ra ở đây là biến X (hàng) và biến Y (cột) có quan hệ hay không?

Giả thiết Ho:" Hàng và cột không quan hệ".

b1-Kiểm định χ^2

Để kiểm tra giả thiết này theo kiểm định χ^2 phải thực hiện các bước sau:

a- Từ giả thiết hàng và cột không quan hệ suy ra các số ở trong ô về lý thuyết phải bằng tổng hàng(TH_i) nhân với tổng cột (TC_j) chia cho tổng số quan sát N (trong thí dụ 7.4 chúng ta sẽ lý giải vấn đề này). Gọi tần số lý thuyết là t_{ij}

$$t_{ij} = \frac{TH_i \times TC_j}{N} \qquad (3)$$

b- Tính khoảng cách giữa 2 tần số m_{ij} và t_{ij} theo khoảng cách χ^2

$$\frac{(m_{ij} - t_{ij})^2}{t_{ij}}$$

c- Tính khoảng cách giữa 2 bảng m_{ij} và t_{ij} bằng χ^2_{tn}:

$$\chi^2_{tn} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(m_{ij} - t_{ij})^2}{t_{ij}} \qquad (4)$$

d- Chọn mức ý nghĩa α và tìm giá trị tới hạn trong bảng 4 $\chi^2(\alpha,(k-1)(l-1))$

e- Kết luận: Ở mức ý nghĩa α nếu $\chi^2_{tn} \le \chi^2(\alpha,(k-1)(l-1))$ thì chấp nhận Ho, ngược lại thì bác bỏ H_o

f - Có thể tính χ^2_{tn} theo công thức tương đương với (4)

$$\chi^2_{tn} = N(\sum_i \sum_j \frac{m_{ij}^2}{TH_i \times TC_j} - 1) \qquad (5)$$

Bài toán về bảng tương liên thường thể hiện dưới hai dạng:

1- X và Y là hai tính trạng, giả thiết H_0:"Hai biến X, Y không quan hệ" (đôi khi còn nói là "X và Y độc lập").

Thường gọi bài toán này là bài toán ***kiểm định tính độc lập*** của hai biến định tính, hay kiểm định tính độc lập của hai tính trạng.

2- Hàng X là các đám đông, cột Y là các nhóm, việc phân chia mỗi đám đông thành các nhóm căn cứ vào một tiêu chuẩn nào đó. Bài toán này thường gọi là ***bài toán kiểm định tính thuần nhất của các đám đông*** (tức là các đám đông có cùng tỷ lệ phân chia), ***hay còn gọi là bài toán kiểm định các tỷ lệ.***

b2- Kiểm định G

Kiểm định G theo các bước sau:

a- Tính $T_1 = \sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij} \ln m_{ij}$ b- Tính $T_2 = \sum_{i=1}^{k} TH_i \times \ln(TH_i)$

c- Tính $T_3 = \sum_{j=1}^{l} TC_j \times \ln(TC_j)$ d- Tính $T_4 = N \times \ln(N)$

e- Tính $G = 2[T_1 - T_2 - T_3 + T_4]$ f- So với $\chi^2(\alpha,(k-1)(l-1))$.

Nếu $G \le \chi^2(\alpha,(k-1)(l-1))$ thì chấp nhận H_0, nếu G lớn hơn thì bác bỏ H_0.

c- BẢNG 4 Ô

Trường hợp đặc biệt của bảng tương liên là bảng chỉ có 2 hàng, 2 cột tạo ra 4 ô, gọi tắt là bảng 4 ô như trong thí dụ 3.

Y / X	Y1	Y2	Tổng hàng
X1	a	b	a +b
X2	c	d	c + d
Tổng cột	a +c	b+d	n = a+b+c+d

Có thể kiểm định giả thiết X độc lập với Y theo cách tính χ^2_{tn} như thí dụ 3, nhưng trong trường hợp bảng 4 ô có thể tính nhanh hơn theo công thức sau (suy ra từ cách tính trên)

$$\chi^2_{tn} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (6)$$

trong trường hợp bảng 4 ô các nhà thống kê thường đưa thêm hiệu chỉnh Yates để tăng độ chính xác của kiểm định

$$\chi^2_{tn} = \frac{n(|\ ad-bc\ |-0{,}5n)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (7)$$

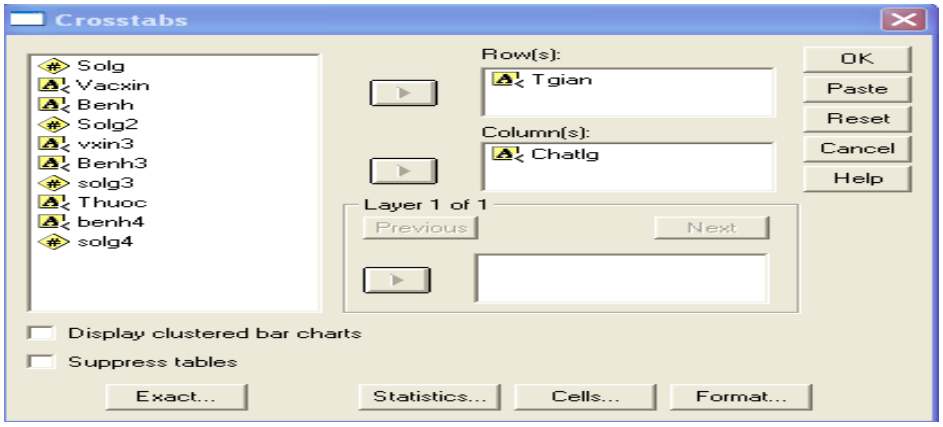
II XỬ LÝ TRONG SPSS

Mở tệp Baitap5.

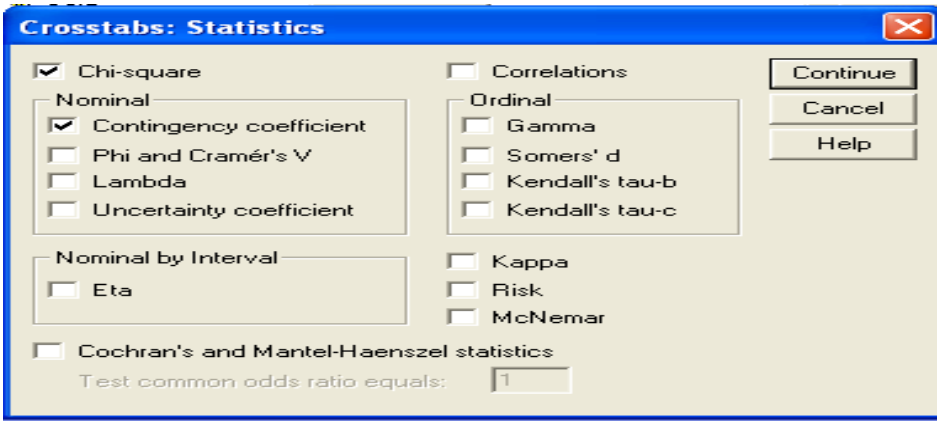
Vào Data Weight cases. Chọn Weight case by Solg

Sau đó vào Analyse Descriptive Statistics Crosstab

Đưa Tgian vào Rows Chatlg vào Columns. Giả thiết H_0:Thời gian thu hoạch không ảnh hưởng đến chất lượng cà chua.



Trong Statistics chọn Chi square và Contigency Coeficient



Kết quả được bảng tương liên

Tgian * Chatlg Crosstabulation

Count

		Chatlg		Total
		Tot	xau	
Tgian	Bthuong	33	12	45
	Som	27	3	30
Total		60	15	75

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.125[b]	1	.077		
Continuity Correction[a]	2.170	1	.141		
Likelihood Ratio	3.363	1	.067		
Fisher's Exact Test				.139	.067
N of Valid Cases	75				

a. Computed only for a 2x2 table

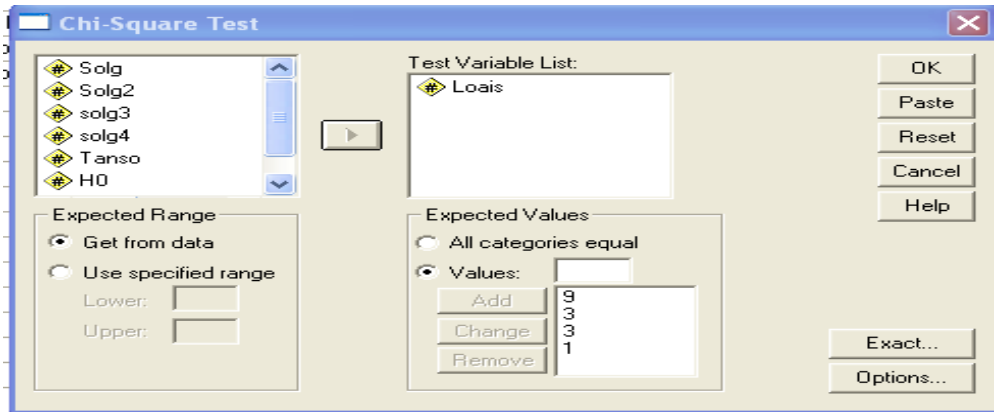
b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6. 00.

Kết luận: Chấp nhận H_0 (vì các mức ý nghĩa sig đều lớn hơn 0,05)

Kiểm định một phân phối: Đậu với 2 tính trội gồm 4 nhóm

Loại	Tần số	Giả thiết H0
AB	59	9
Ab	18	3
aB	26	3
ab	12	1
Tổng số	115	16

Vào Data Weight case chọn weight case by tanso. Vào Analyse Nonparametric Tests Chisquare, chọn Loais vào test variable List. Chọn Values sau đó lần lượt đưa 9, 3, 3, 1 vào (Nhập số 9, Add, nhập số 3, Add v. v. . .)



Kết quả như sau:

Chi-Square Test

Frequencies

Loais

	Observed N	Expected N	Residual
1	59	64.7	-5.7
2	18	21.6	-3.6
3	26	21.6	4.4
4	12	7.2	4.8
Total	115		

Test Statistics

	Loais
Chi-Square[a]	5.224
df	3
Asymp. Sig.	.156
Exact Sig.	.158
Point Probability	.005

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 7.2.

Kết luận: Chấp nhận giả thiết H0:Các kiểu hình phân phối theo tỷ lệ 9:3:3:1