Chương 7

Kiểm định một phân phối và bảng tương liên

Biến ngẫu nhiên liên tục bằng tổng bình phương của nhiều biến ngẫu nhiên độc lập, phân phối chuẩn tắc là biến Khi bình phương χ^2. Biến này được khảo sát tỷ mỷ và lập bảng phân phối (bảng 4). 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.

7.1. Kiểm định một phân phối

Để khảo sát một biến định tính X ta lấy mẫu quan sát gồm N cá thể và căn cứ vào sự thể hiện của biến X để phân chia thành k lớp như bảng sau:

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

Biến X	L_1	L_2	. . .	L_k	Tổng
Tần số Oi	O_1	O_2	. . .	O_k	$N=\Sigma O_i$

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_0 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:

1) Chấp nhận H_0 tức là coi tần số thực tế m_i phù hợp với lý thuyết đã nêu thể hiện ở tần suất f_i.

2) Bác bỏ H_0 tức là dãy tần số thực tế m_i không phù hợp với lý thuyết đã nêu.

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.

Các bước thực hiện:

1) Tính các tần số lý thuyết theo công thức: $E_i = N. f_i$ (7.1)

2) Tính khoảng cách giữa hai số O_i và E_i theo cách tính khoảng cách

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$$

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

$$\chi^2_{TN} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (7.2)$$

4) Tìm giá trị tới hạn trong bảng 4 (cột α, dòng k-1, ký hiệu là $\chi^2(\alpha,k-1)$)

5) Nếu $\chi^2_{tn} \le \chi^2(\alpha,k-1)$ thì chấp nhận H_0: "Tần số thực tế O_i phù hợp với lý thuyết đã nêu". Nếu $\chi^2_{tn} > \chi^2(\alpha,k-1)$ thì bác bỏ H_0, tức là "Tần số thực tế O_i không phù hợp với lý thuyết đã nêu".

Để sử dụng phép thử χ^2, cần thoả mãn các điều kiện sau:

1) Các Oi là các quan sát độc lập

2) Tất cả các Ei đều phải lớn hơn hoặc bằng 5

3) Các Oi và Ei không phải là các tỷ lệ phần trăm.

Ví dụ 7.1: Số liệu thống kê năm 1995 cho thấy, tỷ lệ màu lông (fi) *trắng*, *nâu* và *đen trắng* của thỏ trong một quần thể tương ứng là 0,36; 0,48 và 0,16. Năm 2005, từ 400 con thỏ rút một cách ngẫu nhiên từ quần thể nêu trên có 140 con màu lông *trắng*, 240 con màu *nâu* và 20 con màu *đen trắng*. Câu hỏi đặt ra: Sau 10 năm (từ 1995 đến 2005) tỷ lệ màu lông của thỏ trong quần thể có thay đổi hay không?

Giả thiết H_0: Tỷ lệ màu lông của thỏ trong quần thể sau 10 năm không thay đổi

Ta có thể tóm tắt số liệu quan sát thu được năm 2005 như sau:

Màu lông	Trắng	Nâu	Đen trắng	Tổng số
Tần số (O_i)	140	240	20	400

Dựa vào tỷ lệ ban đầu (năm 1995) ta có các tần suất lý thuyết (ti)

Màu lông	Trắng	Nâu	Đen trắng	Tổng số
fi	0,36	0,48	0,16	1
Ei	400×0,36= 144	400×0,48= 192	400×0,16= 64	400

$$\chi^2_{TN} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(140-144)^2}{144} + \frac{(240-192)^2}{192} + \frac{(20-64)^2}{64} = 42{,}361$$

Bậc tự do df = (3 - 1) = 2; giá trị tới hạn $\chi^2(0{,}05;\ 2) = 5{,}991$

Kết luận: $\chi^2_{TN} < \chi^2(0{,}05,\ 2)$ nên bác bỏ giả thiết H_0. Chứng tỏ tỷ lệ màu lông thỏ trong quần thể sau 10 năm có sự thay đổi.

Ví dụ 7.2: Giả sử chúng ta điều tra giới tính của một quần thể cho trước. Trong một mùa nhất định trong năm người ta thấy tỷ lệ giới tính lúc sinh ra có xu hướng con cái cao hơn. Để giải đáp câu hỏi trên tiến hành chọn ngẫu nhiên 297 con chim mới sinh thì thấy có 167 con cái. Liệu yếu tố mùa có làm ảnh hưởng đến tỷ lệ giới tính hay không?

Đối với trường hợp giới tính, ta luôn thừa nhận tỷ lệ đực cái là 1:1 hay 0,5:0,5. Nếu mùa không làm ảnh hưởng đến tỷ lệ giới tính thì theo ước tính lý thuyết từ 297 con chim quan sát ta sẽ có số chim đực và số chim cái bằng nhau và bằng $297 \times 0,5 = 148,5$.

Ta có bảng tổng hợp sau:

	Đực	Cái	Tổng số
Tần số quan sát (Oi)	130	167	297
Tần số lý thuyết (Ei)	148,5	148,5	297

$$\chi^2_{TN} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2_{TN} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(130-148,5)^2}{148,5} + \frac{(167-148,5)^2}{148,5} = 4,61$$

Bậc tự do df = (2 - 1) = 1; giá trị tới hạn $\chi^2(0,05; 1) = 3,84$

Kết luận: $\chi^2_{TN} < \chi^2(0,05, 1)$ nên bác bỏ giả thiết H_0. Chứng tỏ tỷ lệ giới tính không tuân theo tỷ lệ đực cái 1:1. Điều kiện khí hậu đã làm thay đổi tỷ lệ này.

Hiệu chỉnh Yate

$$\chi^2 = \sum_{i=1}^{k} \frac{(|O_i - E_i| - 0,5)^2}{E_i}$$

Hệ số 0,5 trong công thức nêu trên gọi là hệ số hiệu chỉnh Yate hay còn gọi là hiệu chỉnh tính liên tục để loại bỏ sự thiên lệch. Hiệu chỉnh Yate sẽ được trình bày chi tiết ở phần tiếp theo

Theo ví dụ trên ta có giá trị χ^2 hiệu chỉnh là:

$$\chi^2_{TN} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(|130-148,5| - 0,5)^2}{148,5} + \frac{(|167-148,5| - 0,5)^2}{148,5} = 4,36$$

Giá trị χ^2 hiệu chỉnh (4,36) bé hơn giá trị χ^2 trước khi hiệu chỉnh (4,61), tuy nhiên giá trị hiệu chỉnh vẫn lớn hơn giá trị tới hạn (3,84) cho nên ta vẫn có kết luận tương tự như trên.

7.2. Bảng tương liên l × k

Có 2 biến định tính, biến X chia ra k lớp, biến Y chia ra l lớp, qua khảo sát thu được bảng hai chiều chứa các số quan sát được của các ô O_{ij} (gọi là bảng tương liên):

Bảng các tần số Oij

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

Các số O_{ij} thường được gọi là các tần số thực tế. Bài toán đặt ra ở đây là biến X(hàng) và biến Y(cột) có quan hệ hay không?

Giả thiết H_0 : "hàng và cột không quan hệ" với đối thuyết H_1: "hàng và cột có quan hệ".

Để kiểm tra giả thiết này phải thực hiện các bước sau:

1) 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 E_{ij} ta có :

$$E_{ij} = \frac{TH_i \times TC_j}{N} \qquad (7.3)$$

2) Tính khoảng cách giữa 2 tần số O_{ij} và E_{ij} theo cách tính khoảng cách χ^2

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3) Tính khoảng cách giữa 2 dãy m_{ij} và t_{ij} bằng χ^2_{tn}:

$$\chi^2_{tn} = \sum_1^k \sum_1^l \frac{(m_{ij} - t_{ij})^2}{t_{ij}} \qquad (7.4)$$

4) Chọn mức ý nghĩa α và tìm giá trị tới hạn trong bảng 4 $\chi^2(\alpha,(k-1)(l-1))$ tương ứng với cột α và bậc tự do (k-1)(l-1)

5) Kết luận:
Ở mức ý nghĩa α nếu $\chi^2_{tn} \leq \chi^2(\alpha,(k-1)(l-1))$ chấp nhận H_0, ngược lại thì bác bỏ H_0

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 có quan hệ" hay còn phát biểu một cách khác 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 đám đông thành các nhóm căn cứ vào một tiêu chuẩn nào đó. Bài toán này thường được 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 kiểm định các tỷ lệ.

Ví dụ 7.3: Từ một đàn trước khi cho tiếp xúc với nguồn bệnh, chọn ra 295 động vât thí nghiệm (tiêm vắc xin) và 55 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiếp xúc với nguồn bệnh ta thu được kết quả như trong bảng sau. Liệu vắc xin có làm giảm tỷ lệ chết hay không?

	Kết quả		
Thuốc	Sống	Chết	Tổng hàng
Vắc xin	120	175	295
Đối chứng	30	25	55
Tổng cột	150	200	350

Ở đây có thể coi hàng là các lớp của biến thuốc X (có 2 lớp A, B), cột là là các lớp của biến kết quả Y (có 2 lớp: sống và chết). Cũng có thể coi hàng là các đám đông: "những động vật tiêm vắc xin" và "những động vật không tiêm vắc xin". Cột là sự phân chia mỗi đám đông thành 2 nhóm sống và chết.

Bảng tần số lý thuyết:

	Kết quả		
Thuốc	Sống	Chết	Tổng hàng
Vắc xin	$\frac{295\times150}{350}=126{,}4$	$\frac{295\times200}{350}=168{,}6$	295
Đối chứng	$\frac{55\times150}{350}=23{,}6$	$\frac{55\times200}{350}=31{,}4$	55
Tổng cột	150	200	350

$$\chi^2_{TN} = \frac{(120-126{,}4)^2}{126{,}4}+\frac{(175-168{,}6)^2}{168{,}6}+\frac{(30-23{,}6)^2}{23{,}6}+\frac{(25-31{,}4)^2}{31{,}4}=3{,}64$$

Bậc tự do df = (2-1)(2-1) = 1. Giá trị tới hạn $\chi^2(0{,}05,1) = 3{,}84$

Kết luận: Vì "$\chi^2_{TN} = 3{,}64 < \chi^2(0{,}05,1) = 3{,}84$, ta chưa có đủ bằng chứng để bác bỏ H_0. Hay nói một cách khác vắc xin đã không làm giảm được tỷ lệ chết.

Ví dụ 7.4: Nghiên cứu ảnh hưởng của việc thiến đến sự xuất hiện bệnh tiểu đường ở chuột. Từ 100 chuột thí nghiệm, chia ngẫu nhiên về 1 trong 2 cách xử lý thiến và không thiến. Số chuột ở 2 lô thí nghiệm được theo dõi cho đến 140 ngày tuổi và tiến hành lấy mẫu nghiên cứu từ 42 ngày tuổi. Bệnh tiểu đường được xác định đối với chuột có hàm lượng đường trong máu lớn hơn 200 mg/ dl. Kết quả thí nghiệm được ghi lại ở bảng sau:

Cách xử lý	Kết quả		
	Mắc bệnh	Không mắc bệnh	Tổng
Thiến	26	24	50
Không thiến	12	38	50
Tổng số	38	62	100

Tần suất lý thuyết

Cách xử lý	Kết quả		
	Mắc bệnh	Không mắc bệnh	Tổng
Thiến	$\frac{50\times 38}{100}=19$	$\frac{50\times 62}{100}=31$	50
Không thiến	$\frac{50\times 38}{100}=19$	$\frac{50\times 62}{100}=31$	50
Tổng số	38	62	100

$$\chi^2_{TN}=\frac{(26-19)^2}{19}+\frac{(12-19)^2}{19}+\frac{(24-31)^2}{31}+\frac{(38-31)^2}{31}=8{,}32$$

Đối với trường hợp bảng tương liên 4 ô

a	b
c	d

Có thể tính χ^2_{TN} theo công thức

$$\chi^2_{TN}=n\times\frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}=100\times\frac{(26\times 38-12\times 24)^2}{50\times 50\times 38\times 62}=8{,}32$$

Bậc tự do df = (2-1)(2-1) = 1. Giá trị tới hạn $\chi^2(0{,}05;1) = 3{,}84$

Kết luận: Vì $\chi^2_{TN} = 8{,}32 > \chi^2(0{,}05;1) = 3{,}84$ nên giả thiết H_0 bị bác bỏ. Chứng tỏ, tỷ lệ chuột sau khi thiến mắc bệnh đái đường cao hơn so với chuột không bị thiến.

Hiệu chỉnh Yates

$$\chi^2=\frac{\left(|ad-bc|-\frac{n}{2}\right)^2 n}{(a+b)(a+c)(b+d)(c+d)}$$

Với ví dụ trên ta có giá trị χ^2 hiệu chỉnh là:

$$\chi^2 = \frac{\left(|26\times 38 - 24\times 12| - \frac{100}{2}\right)^2 \times 100}{(26+24)(26+12)(24+38)(12+38)} = 7{,}17$$

Kết luận: Với hiệu chỉnh Yate, giá trị χ^2 thực nghiệm bé hơn ($\chi^2 = 7{,}17$) so với trước khi hiệu chỉnh ($\chi^2 = 8{,}32$). Tuy nhiên giá trị c^2 thực nghiệm vẫn lớn hơn giá trị tới hạn, nên ta có kết luận tương tự về bệnh tiểu đường của chuột như đã nêu ở phần trên.

Lưu ý:
Hệ số điều chỉnh của Yate trong kiểm định một phân phối có 2 lớp và trong bảng tương liên 2×2.

a) Kiểm định một phân phối có 2 lớp

Tính trạng nghiên cứu	Loại 1	Loại 2	Tổng
Tần số thực tế	m_1	m_2	N
Tần số lý thuyết	$t_1=N\times p_1/(p_1+p_2)$	$t_2=N\times p_2/(p_1+p_2)$	N

Để kiểm định giả thiết H_0: "Hai lớp nói trên phân phối theo tỷ lệ $p_1:p_2$ "có thể sử dụng phương pháp χ^2 với nội dung:

Tính $$\chi^2_{tn} = \frac{(m_1 - t_1)^2}{t_1} + \frac{(m_2 - t_2)^2}{t_2}$$

So χ^2_{TN} với giá trị tới hạn χ^2 với mức ý nghĩa α và bậc tự do bằng 1. Nếu $\chi^2_{TN} \le \chi^2_{(\alpha,1)}$ thì chấp nhận H_0, nếu $\chi^2_{tn} > \chi^2(\alpha,1)$ thì bác bỏ H_0.

Bài toán kiểm định này tương đương với bài toán kiểm định một xác suất, việc tính toán dựa trên cách tính xấp xỉ phân phối nhị thức bằng phân phối chuẩn, từ đó suy ra χ^2_{TN} xấp xỉ phân phối χ^2 (là một phân phối liên tục suy ra từ phân phối chuẩn). Trường hợp $N < 100$ phép xấp xỉ không thật tốt, thường cho χ^2_{TN} hơi to do đó Yate đề nghi điều chỉnh lại χ^2_{TN} theo hướng làm nhỏ bớt χ^2_{TN}, điều chỉnh này thường gọi là điều chỉnh do tính liên tục.

Công thức tính χ^2_{TN} điều chỉnh như sau:

$$\chi^2_{tn} = \frac{(|m_1 - t_1| - 0{,}5)^2}{t_1} + \frac{(|m_2 - t_2| - 0{,}5)^2}{t_2}$$

b) Bảng tương liên 4 ô (2 x 2)

Tính trạng A	Tính trạng B		
	Lớp B1	Lớp B2	Tổng hàng
Loại A1	a	b	a+b
Loại A2	c	d	c+d
Tổng cột	a+c	b+d	N=a+b+c+d

Để kiểm định giả thiết H_0: "Hai tính trạng A và B độc lập" có thể dùng phương pháp χ^2 với các nội dung sau:

+ Tính các số lý thuyết

$$\hat{a} = \frac{(a+b)(a+c)}{N} \quad \hat{b} = \frac{(a+b)(b+d)}{N} \quad \hat{c} = \frac{(c+d)(a+c)}{N} \quad \hat{d} = \frac{(c+d)(b+d)}{N}$$

+ Tính $\chi^2_{TN} = \frac{(a-\hat{a})^2}{\hat{a}} + \frac{(b-\hat{b})^2}{\hat{b}} + \frac{(c-\hat{c})^2}{\hat{c}} + \frac{(d-\hat{d})^2}{\hat{d}}$

Có thể tính χ^2_{TN} bằng công thức sau:

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

+ So với giá trị tới hạn χ^2 với mức ý nghĩa α và bậc tự do bằng 1. Nếu $\chi^2_{TN} \le \chi^2(\alpha,1)$ thì chấp nhận H_0, nếu $\chi^2_{TN} > \chi^2(\alpha,1)$ thì bác bỏ H_0.

Bài toán này tương đương với bài toán so sánh hai xác suất, việc tính toán dựa trên cách tính xấp xỉ phân phối nhị thức bằng phân phối chuẩn, từ đó suy ra χ^2_{TN} xấp xỉ phân phối χ^2.

Khi N nhỏ việc xấp xỉ không tốt do đó có một số hướng dẫn như sau:

+ Nếu $N \le 20$ thì không nên dùng phương pháp χ^2_{TN}

+ Nếu $20 < N \le 40$ và có ô có số lý thuyết bé < 5 thì cũng không nên dùng phương pháp χ^2_{TN}

Cả hai trường hợp này nên dùng phương pháp chính xác Fisher (xem phần 7.3)

Nếu $N \ge 100$ thì có thể dùng phương pháp χ^2.

Nếu $N < 100$ và không rơi vào 2 trường hợp đầu thì nên đưa thêm điều chỉnh do tính liên tục Yate nhằm làm nhỏ bớt χ^2_{TN} như sau:

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

7.3. Kiểm định chính xác của Fisher đối với bảng tương liên 2×2

Khi các giá trị ước tính (Ei) trong bảng tương liên 2×2 rất bé (Ei < 5) thì việc sử dụng phép kiểm định χ^2 không còn đảm bảo được độ chính xác. Trường hợp này hay gặp trong nghiên cứu dịch tễ học và phép kiểm định chính xác của Fisher được sử dụng. Phép kiểm định này cho ta một xác suất trực tiếp và chính xác thay vì đi tìm giá trị xác suất từ bảng.

Nếu ta có bảng tương liên 2×2

a	b	a + b
c	d	c + d
a + c	b + d	n

Fisher dựa trên phân phối siêu hình học (hypergeometric distribution) để tính xác suất của phép thử theo công thức.

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!}$$

Các bước thực hiện:

1) Tính p_1 với bảng số liệu đã cho

2) Tính ad – bc.

+ Nếu ad – bc > 0 thì tăng a và d, giảm b và c bằng 1 đơn vị rồi tính xác suất p_2; làm tương tự cho đến khi a bằng min của (a+b) hoặc (a+c)

+ Nếu ad – bc < 0 thì giảm a và d, tăng b và c rồi tính xác suất p_2; làm tương tự cho đến khi a bằng 0

3) Tính P = 2×(p1 + p2 +....+ pn)

4) Nếu xác suất P < 0,05 thì kết luận bác bỏ H_0.

Ví dụ 7.5: Từ một đàn trước khi cho tiếp xúc với nguồn bệnh, chọn ra 10 động vât thí nghiệm (tiêm vắc xin) và 10 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiếp xúc với nguồn bệnh ta thu được kết quả như trong bảng sau. Liệu vắc xin có làm giảm tỷ lệ chết hay không?

	Kết quả		
Thuốc	Sống	Chết	Tổng hàng
Vắc xin	9	1	10
Đối chứng	2	8	10
Tổng cột	11	9	20

1) $p_1 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = \frac{10!10!11!9!}{9!1!2!8!20!} = 0{,}002679$

2) ad – bc = 9×8 - 1×2 = 70 > 0

Tăng a, d và giảm b, c bằng 1 đơn vị ta có

9 +1	2 -1	11	➔	10	1	11
1 - 1	8 + 1			0	9	
10				10		

$$p_2 = \frac{10!10!11!9!}{10!0!1!9!20!} = 0{,}000059537985$$

3) P = 2×(p1 + p2 +....+ pn) = 2×(0,002679 + 0,000059537985) = 0,005477076

4) Với xác suất này, giả thiết H_0 bị bác bỏ. Điều này chứng tỏ vắc xin đã làm giảm tỷ lệ chết.

Ví dụ 7.6: Tương tự như ví dụ 7.5 từ 15 động vât thí nghiệm (tiêm vắc xin) có 2 động vật mắc bệnh và từ 13 động vật đối chứng (không tiêm vắc xin) có 10 động vật mắc bệnh. Liệu vắc xin có làm giảm tỷ lệ mắc bệnh hay không?

	Kết quả		
Thuốc	Mắc bệnh	Không	Tổng hàng
Vắc xin	2	13	15
Đối chứng	10	3	13
Tổng cột	12	16	28

1) $p_1 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = \frac{15!13!12!16!}{2!13!10!3!28!} = 0{,}00098712$

2) ad – bc = 2×3 - 13×10 = -124 < 0

Giảm a, d và tăng b, c bằng 1 đơn vị ta có

2 -1	13 + 1	15	➔	1	14	15
10 + 1	3 - 1	13		11	2	13
12	16	28		12	16	28

$p_2 = \frac{15!13!12!16!}{1!14!11!2!28!} = 0{,}00003846$

Giảm a, d và tăng b, c bằng 1 đơn vị ta có

1 - 1	14 + 1	15	➔	0	15	15
11 + 1	2 - 1	13		12	1	13
12	16	28		12	16	28

$p_3 = \frac{15!13!12!16!}{0!15!12!1!28!} = 0{,}0000004273$

3) P = 2×(p1 + p2 +....+ pn) = 2×(0,00098712 + 0,00003846 + 0,0000004273) = 0,00205202

4) Với xác suất này, giả thiết H_0 bị bác bỏ. Điều này chứng tỏ vắc xin đã làm giảm tỷ lệ mắc bệnh.

Cochran khuyến cáo nên sử dụng phép thử chính xác của Fisher nếu trong thí nghiệm n < 20 hoặc 20 < n <40 và dự đoán bé nhất nhỏ hơn 5.

7.4. Xác định mức liên kết trong dịch tễ học bằng kiểm định χ^2

Trong dịch tễ học, tầm quan trọng của sự liên kết giữa hàng và cột trong bảng tương liên còn được xem xét bởi: 1) nguy cơ tương đối (RR) và 2) tỷ suất chênh (OR).

Nếu ta có bảng tương liên 2×2 như sau:

	Bệnh		Tổng số
Nhân tố	+	-	
+	a	b	a + b
-	c	d	c + d
Tổng số	a + c	b + d	n

Ta có:

$$\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc} \qquad \text{RR} = \frac{a/a+b}{c/c+d}$$

7.4.1. Nghiên cứu cắt ngang (cross sectional studies)

Mục đích của nghiên cứu cắt ngang là tìm ra mối liên hệ giữa yếu tố nguy cơ và bệnh; tức là so sánh tần suất mắc bệnh của nhóm có tiếp xúc và không tiếp xúc. Trong nghiên cứu này toàn bộ các phép đo phải thực hiện trong thời điểm nhất định.

Ví dụ 7.7: Tỷ lệ bò mắc bệnh viêm vú giữa 2 trại (A và B) có sự sai khác có ý nghĩa hay không? Biết rằng sau khi kiểm tra 96 bò ở trại A và 72 bò ở trại B trong 1 ngày thấy số lượng bò măc bệnh viêm vú tương ứng là 36 và 10.

Giả thiết H_0: Tỷ lệ bò mắc bệnh viêm vú ở hai trại là như nhau với đối thiết H_1: Tỷ lệ bò mắc bệnh viêm vú ở 2 trại là khác nhau.

Nếu sử dụng phép thử χ^2 ta được giá trị $\chi^2_{TN} = 11{,}535$; giá trị $\chi^2_{(0{,}05;\ 1)} = 3{,}841$.

Kết luận:

Vì $\chi^2_{TN} > \chi^2$ tới hạn nên có thể kết luận rằng tỷ lệ bò mắc bệnh viêm vú ở hai trại là khác nhau. Mặt khác ta có tỷ suất chênh OR = (36×62)/(60×10) = 3,72; tức là số bò mắc bệnh viêm vú ở trại A cao gấp 3,72 lần so với số bò mắc bệnh ở trại B.

7.4.2. Tiến cứu (cohort studies)

Trong nghiên cứu này động vật được chia thành 2 nhóm; một trong hai nhóm sẽ tiếp xúc với yếu tố nguy cơ của bệnh, nhóm còn lại là đối chứng. Theo dõi trong một thời gian để xác định sự xuất hiện bệnh ở hai nhóm. Căn cứ vào kết quả thu được để kết luận giữa yếu tố nguy cơ và tỷ lệ mắc bệnh. Chính vì vậy nghiên cứu này được gọi là tiến cứu (cohort studies).

Ví dụ 7.8: Xem xét ví dụ 7.5, từ một đàn trước khi cho tiếp xúc với nguồn bệnh, chọn ra 10 động vât thí nghiệm (tiêm vắc xin) và 10 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiếp xúc với nguồn bệnh ta thu được kết quả như trong bảng sau. Liệu vắc xin có làm giảm tỷ lệ chết hay không?

	Kết quả		
Thuốc	Sống	Chết	Tổng hàng
Vắc xin	9	1	10
Đối chứng	2	8	10
Tổng cột	11	9	20

Nếu sử dụng phép thử chính xác của Fisher ta có xác suất P = 0,005477076

Kết luận: Với xác suất này, giả thiết H_0 bị bác bỏ. Điều này chứng tỏ vắc xin đã làm giảm tỷ lệ chết. Bên cạnh đó, nguy cơ tương đối RR = (9/10)/(2/10) = 4,5. Hay nói một cách khác động vật sử dụng vắc xin mức độ sống sót gấp 4,5 lần so với động vật không dùng vắc xin.

7.4.3. Nghiên cứu - bệnh chứng hay hồi cứu (case-control studies)

Trong nghiên cứu bệnh - chứng hay hồi cứu, các nhóm động vật nhiễm bệnh và không nhiễm bệnh được chọn ra, sau đó ta đánh giá trong quá khứ động vật đã tiếp xúc với yếu tố nguy cơ như thế nào. Vì vậy nghiên cứu bệnh - chứng mang ý nghĩa của một hồi cứu.

Ví dụ 7.9: Trong một nghiên cứu, có 62 bò sữa được chẩn đoán ung thư biểu mô mắt và 124 không mắc được chọn ngẫu nhiên từ quần thể. Có mối liên hệ nào giữa giống bò và tỷ lệ mắc bệnh ung thư biểu mô mắt hay không? Nếu số liệu thu thập được như sau:

Giống	Mắc bệnh	Không mắc bệnh	Tổng số
Hereford	44	63	107
Giống khác	18	61	79
Tổng số	62	124	186

Giả thiết H_0: Không có mối liên hệ giữa giống và tỷ lệ mắc bệnh với đối thiết H_1: Có mối liên hệ giữa bệnh và giống..

Sử dụng phép thử χ^2, ta có $\chi^2_{TN} = 6,876$ và $\chi^2 (0,05;1) = 3,841$.

Kết luận:

Vì $\chi^2_{TN} > \chi^2$ tới hạn nên ta bác bỏ H_0 chấp nhận H_1; chứng tỏ có mối liên hệ giữa giống và bệnh. Tỷ suất chênh OR = (44×61)/(18×63) = 2,37. Hay nói cách khác giống Hereford mắc bệnh ung thư biểu mô mắt cao hơn 2,37 lần so với các giống khác.

7.5. Bài tập

7.5.1

Một trung tâm thu tinh nhân tạo tiến hành thử nghiệm 3 phương pháp thụ tinh nhân tạo khác nhau. Tỷ lệ phối có chửa ở 3 phương pháp thu được như sau: ở phương pháp I, có 275 bò có chửa từ 353 bò tham gia thí nghiệm; tương tự ở phương pháp II, các con số này lần lượt là 192 và 256 con, phương pháp III là 261 và 384 con. Tỷ lệ thụ tinh thành công ở 3 phương pháp này có khác nhau hay không?

7.5.2

Chọn mẫu ngẫu nhiên thế hệ con của bò lang Shorthorn thu được kết quả sau đây: 82 con màu lông đỏ, 209 con lang và 89 con trắng. Phân bố màu lông của bò có tuân theo giả thiết rằng màu lông được xác định bởi một cặp allen trội không hoàn toàn? Biết rằng trội không hoàn toàn là trường hợp có một allen trội và dị hợp tử thể hiện sự ảnh hưởng của đồng thời cả 2 allen.

7.5.3

Một thí nghiệm được tiến hành nhằm đánh giá sự liên hệ giữa tỷ lệ viêm nội mạc tử cung và giống. Trong tổng số 700 bò sữa trong nghiên cứu thuần tập (cohort studies), có 500 con giống Holstein Friesian và 200 con giống Jersey. Kết quả nghiên cứu thu được như sau:

		Viêm nội mạc tử cung		Tổng số
		Có	Không	
Giống	Holstein	100	400	500
	Jersey	10	190	200
	Tổng số	110	590	700

Có sự liên hệ giữa tỷ viêm nội mạc tử cung và các giống hay không?