ĐẠI HỌC THĂNG LONG
THANG LONG UNIVERSITY

THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ XÃ HỘI

Hướng dẫn thực hành trên phần mềm R

Phan Thanh Hồng - Nguyễn Thị Nhung

NXB Thống kê, 2015

CHƯƠNG 11

KIỂM ĐỊNH KHI-BÌNH PHƯƠNG

Yêu cầu đối với sinh viên

- Biết cách kiểm chứng tính độc lập trong R.
- Biết cách kiểm chứng mức phù hợp của một phân phối trong R.
- Biết cách kiểm chứng phân phối chuẩn trong R.

Trong phần này, ta sẽ giới thiệu cách dùng hàm `chisq.test` để thực hiện kiểm định khi-bình phương, cụ thể là làm bài toán kiểm định tính độc lập và kiểm định mức phù hợp của một phân phối.

11.1 Kiểm chứng tính độc lập

Giả sử, ta có hai biến định tính trong tổng thể. Kiểm chứng tính độc lập là bài toán đi kiểm định xem hai biến định tính đã cho độc lập hay phụ thuộc vào nhau dựa trên việc kiểm định cặp giả thuyết:
H_0 : Hai biến định tính độc lập (không có mối liên hệ) với nhau;
H_1 : Hai biến định tính không độc lập (có mối liên hệ) với nhau.
Để đưa ra kết luận chấp nhận hay bác bỏ giả thuyết H_0 trong cặp giả thuyết trên, ta tiến hành làm các bước sau:

- Bước 1: Thiết lập ma trận dữ liệu sau:

Biến thứ nhất	Biến thứ hai					Tổng
	1	2	3	…	c	
1	O_{11}	O_{12}	O_{13}	…	O_{1c}	R_1
2	O_{21}	O_{22}	O_{23}	…	O_{2c}	R_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	O_{r1}	O_{r2}	O_{r3}	…	O_{rc}	R_r
Tổng	C_1	C_2	C_3	…	C_c	n

  trong đó

  - r là số thuộc tính của biến định tính thứ nhất, c là số thuộc tính của biến định tính thứ hai, n là cỡ mẫu;
  - $O_{ij}, i = 1, \ldots, r; j = 1, \ldots, c$ là số quan sát có thuộc tính thứ i của biến thứ nhất và thuộc tính thứ j của biến thứ hai;
  - $R_i, i = 1, \ldots, r$ là tổng số quan sát ở dòng i của biến định tính thứ nhất, $C_j, j = 1, \ldots, c$ là tổng số quan sát ở cột j của biến định tính thứ hai.

- Bước 2: Tính các giá trị tần số lí thuyết E_{ij} của ô ở địa chỉ ij theo công thức:

$$E_{ij} = \frac{R_i \times C_j}{n}.$$

- Bước 3: Tính giá trị thống kê

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- Bước 4: Đưa ra kết luận: Bác bỏ H_0 nếu

  - giá trị thống kê lớn hơn giá trị tới hạn tương ứng

$$\chi^2 > \chi^2_{(r-1)(c-1),\alpha},$$

– hoặc $p-$giá trị của bài toán nhỏ hơn mức ý nghĩa α tương ứng.

Hàm `chisq.test` dùng để kiểm chứng tính độc lập được cho như sau:

chisq.test(A)

trong đó

`A` ma trận dữ liệu.

Ví dụ sau minh họa cho ta cách dùng của hàm `chisq.test` trong trường hợp này.

Ví dụ 11.1.1 Một đài truyền hình muốn khảo sát xem có mối liên hệ giữa giai cấp và ý kiến đối với thời lượng phát sóng phim truyện trong tuần hay không. Giai cấp gồm có công nhân, nông dân và trí thức và ý kiến đối với thời lượng phát sóng phim truyện được chia làm bốn mức là tăng thời lượng, giữ thời lượng như cũ, giảm thời lượng và không có ý kiến gì. Bảng sau cho ta số liệu về cuộc thăm dò:

	Tầng lớp		
Ý kiến	Công nhân	Nông dân	Trí thức
Tăng	100	300	20
Như cũ	200	400	30
Giảm	50	80	5
Không ý kiến	30	70	5

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem giai cấp và ý kiến đối với thời lượng phát sóng phim truyện có liên hệ với nhau không.

Lời giải: Ví dụ 11.1.1 là một trường hợp về kiểm chứng tính độc lập (liên hệ) giữa các thuộc tính trong tổng thể, cụ thể là xét xem có hay không có sự liên hệ (độc lập) giữa giai cấp và ý kiến đối với thời lượng phát sóng. Để thực hiện điều này, ta dùng kiểm định khi-bình phương và sử dụng những kết quả cho qua hàm `chisq.test` để đưa ra kết luận. Qui trình thực hiện bài toán được tiến hành qua các bước sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Giai cấp và ý kiến đối với thời lượng phát sóng độc lập với nhau.
  H_1 : Giai cấp và ý kiến đối với thời lượng phát sóng không độc lập với nhau.

- Bước 2: Ta dùng hàm `chisq.test` với các tham số cần thiết được cho như sau:

  – `A` là ma trận dữ liệu.

- Bước 3: Thực hiện kiểm định trên R:

```
> x = scan()
1: 100 200 50 30
5: 300 400 80 70
9: 20 30 5 5
13:
Read 12 items
> A = matrix(x,nrow=4)
> A
     [,1] [,2] [,3]
```

```
[1,]  100  300   20
[2,]  200  400   30
[3,]   50   80    5
[4,]   30   70    5

> chisq.test(A)


        Pearson's Chi-squared test

data:  A
X-squared = 12.0357, df = 6, p-value = 0.06118
```

- Bước 4: Phân tích kết quả trên R: Nhận thấy $p-$giá trị của bài toán bằng 0.06118 lớn hơn mức ý nghĩa $\alpha = 0.05$ nên chấp nhận H_0, hay là có đủ bằng chứng thống kê để cho rằng không có mối liên hệ giữa giai cấp và ý kiến đối với thời lượng phát sóng phim truyện trên đài truyền hình này.

11.2 Kiểm chứng mức phù hợp của một phân phối

Bài toán kiểm chứng mức phù hợp của một phân phối là bài toán kiểm định xem tổng thể có tuân theo một qui luật A nào đó hay không. Để thực hiện điều này, ta đi kiểm định cặp giả thuyết:
H_0 : Tổng thể tuân theo qui luật xác suất A;
H_1 : Tổng thể không tuân theo qui luật xác suất A.
Ta tiến hành làm một số bước sau để đưa ra kết luận chấp nhận hay bác bỏ H_0:

- Bước 1: Chọn một mẫu ngẫu nhiên gồm n phần tử và xếp mỗi phần tử vào đúng một trong k lớp:

Lớp	1	2	...	k	Tổng
Số phần tử quan sát	O_1	O_2	...	O_k	n

  trong đó O_i là số phần tử trong mẫu rơi vào lớp thứ i.

- Bước 2: Tính số phần tử kì vọng theo từng lớp: $E_i = np_i, i = 1, 2, \ldots, k$, trong đó $p_i, i = 1, 2, \ldots, k$ là xác suất kì vọng theo H_0 để một phần tử rơi vào lớp thứ i.

- Bước 3: Tính giá trị thống kê

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

- Bước 4: Đưa ra kết luận: Bác bỏ H_0 nếu:

  – giá trị thống kê lớn hơn giá trị tới hạn tương ứng

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-m-1,\alpha},$$

  với m là số tham số tổng thể ước lượng từ dữ liệu mẫu.

– hoặc $p-$giá trị của bài toán nhỏ hơn mức ý nghĩa α tương ứng.

Hàm `chisq.test` trong trường hợp kiểm chứng mức phù hợp của một phân phối gồm các tham số như sau:

chisq.test(x, p)

trong đó

`x`	véc tơ dữ liệu.
`p`	véc tơ có cùng chiều dài với x chỉ xác suất của phân phối cần kiểm chứng.

Ta xét một số ví dụ minh họa cho hàm `chisq.test` trong trường hợp này.

Ví dụ 11.2.1 Tại một địa phương, số liệu bảo hiểm cho thấy 82% số lái xe không gây ra vụ tai nạn nào trong năm, 15% gây ra đúng một vụ tai nạn và 3% gây ra trên hai vụ tai nạn. Một mẫu ngẫu nhiên gồm 440 kĩ sư được chọn ra thấy có 366 người không gây ra vụ tai nạn nào trong năm, 68 người gây ra đúng một vụ tai nạn và có 6 người gây ra trên hai vụ tai nạn. Tại mức ý nghĩa $\alpha = 5\%$, có thể khẳng định tỉ lệ gây tai nạn của các kĩ sư giống với tỉ lệ gây tai nạn của các lái xe nói chung không?

Lời giải: Ví dụ 11.2.1 là một ví dụ về kiểm chứng mức phù hợp của một phân phối, cụ thể là kiểm định xem các tỉ lệ gây tai nạn của các kĩ sư có giống tỉ lệ gây tai nạn của các lái xe nói chung không. Bài toán kiểm định này được dựa trên kiểm định khi-bình phương và trong R ta sử dụng kết quả trong hàm `chisq.test` để đưa ra kết luận. Qui trình thực hiện bài toán kiểm định được tiến hành qua các bước sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Tỉ lệ gây tai nạn của các kĩ sư giống tỉ lệ gây tai nạn của các lái xe nói chung.
  H_1 : Tỉ lệ gây tai nạn của các kĩ sư không giống tỉ lệ gây tai nạn của các lái xe nói chung

- Bước 2: Ta dùng hàm `chisq.test` với các tham số cần thiết được cho như sau:

  – `x = c(366, 68, 6)` là véc tơ dữ liệu;

  – `p = c(0.82, 0.15, 0.03)` là véc tơ xác suất cần so sánh.

- Bước 3: Thực hiện kiểm định trên R:

```
> x = c(366, 68, 6)
> p0 = c(0.82, 0.15, 0.03)
> chisq.test(x, p=p0)


        Chi-squared test for given probabilities

data:  x
X-squared = 4.0628, df = 2, p-value = 0.1312
```

- Bước 4: Phân tích kết quả trên R: Nhận thấy $p-$giá trị của bài toán bằng 0.1312 lớn hơn mức ý nghĩa $\alpha = 0.05$ nên chấp nhận H_0, hay là có đủ bằng chứng thống kê để cho rằng tỉ lệ gây tai nạn của các kĩ sư giống tỉ lệ gây tai nạn của các lái xe nói chung.

Ví dụ 11.2.2 Một nghiên cứu được tiến hành để xem các vụ động đất ít nhất ở mức trung bình (4.4 độ richter) ở phía nam California có khả năng xảy ra một ngày nào đó cao hơn những ngày khác hay khả năng là như nhau đối với mọi ngày trong tuần. Bảng dữ liệu sau cho ta thông tin về 1100 vụ động đất:

Thứ	Chủ nhật	Hai	Ba	Tư	Năm	Sáu	Bảy
Số vụ động đất	156	144	170	158	172	148	152

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem khả năng xảy ra động đất có như nhau đối với mọi ngày trong tuần không.

Lời giải: Ví dụ 11.2.2 cũng là một ví dụ về kiểm chứng mức phù hợp của một phân phối, ở đây ta muốn kiểm định khả năng xảy ra động đất các ngày trong tuần như nhau không hay xét xem số vụ động đất của các ngày trong tuần có tuân theo phân phối đều hay không. Qui trình kiểm định được tiến hành như sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Khả năng xảy ra động đất các ngày trong tuần như nhau.
  H_1 : Khả năng xảy ra động đất các ngày trong tuần không như nhau.
- Bước 2: Ta dùng hàm `chisq.test` với các tham số cần thiết được cho như sau:
  - `x = c(156, 144, 170, 158, 172, 148, 152)` là véc tơ dữ liệu;
  - `p = c(1/7,1/7, 1/7,1/7,1/7,1/7,1/7)` là véc tơ xác suất cần so sánh.
- Bước 3: Thực hiện kiểm định trên R:

```
> x = c(156, 144, 170, 158, 172, 148, 152)
> p0 = rep(1/7,7)
> chisq.test(x, p=p0)


        Chi-squared test for given probabilities

data:  x
X-squared = 4.2691, df = 6, p-value = 0.6403
```

- Bước 4: Phân tích kết quả trên R: Nhận thấy $p-$giá trị của bài toán bằng 0.6403 lớn hơn mức ý nghĩa $\alpha = 0.05$ nên chấp nhận H_0, hay là có đủ bằng chứng thống kê để cho rằng khả năng xảy ra động đất của các ngày trong tuần là như nhau, không ngày nào có khả năng xảy ra động đất cao hơn những ngày còn lại.

Ví dụ 11.2.3 Người ta cho rằng số vụ cắt điện trong một ngày ở một thành phố tuân theo phân phối Poisson. Bảng dữ liệu sau cho ta thông tin về số vụ cắt điện trong 150 ngày:

Số vụ cắt điện	0	1	2	3	4	5	6	7	8	$\geqslant 9$
Số ngày	0	5	22	23	32	22	19	13	6	8

a. Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem số vụ cắt điện trong một ngày ở thành phố này có tuân theo phân phối Poisson với $\lambda = 4.2$ không.

b. Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem số vụ cắt điện trong một ngày ở thành phố này có tuân theo phân phối Poisson không.

Lời giải: Ví dụ 11.2.3 yêu cầu ta kiểm định xem số vụ cắt điện trong một ngày có tuân theo phân phối Poisson hay không trong cả hai trường hợp biết tham số λ và không biết tham số λ.

a. Trong trường hợp biết $\lambda = 4.2$, ta tiến hành kiểm định theo các bước sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Số vụ cắt điện trong một ngày ở thành phố tuân theo phân phối Poisson với $\lambda = 4.2$.
  H_1 : Số vụ cắt điện trong một ngày ở thành phố không tuân theo phân phối Poisson với $\lambda = 4.2$.
- Bước 2: Ta dùng hàm `chisq.test` với các tham số cần thiết được cho như sau:
  - `x = c(0, 5, 22, 23, 32, 22, 19, 13, 6, 8)` là véc tơ dữ liệu;
  - `p = c(dpois(0:8, lambda=4.2),1-ppois(8,lambda=4.2))` là véc tơ xác suất cần so sánh.
- Bước 3: Thực hiện kiểm định trên R:

```
> x = c(0, 5, 22, 23, 32, 22, 19, 13, 6, 8)
> p0 = c(dpois(0:8, lambda=4.2),1-ppois(8,lambda=4.2))
> chisq.test(x, p=p0)

        Chi-squared test for given probabilities

data:  x
X-squared = 10.375, df = 9, p-value = 0.321
```

- Bước 4: Phân tích kết quả trên R: Nhận thấy $p-$giá trị của bài toán bằng 0.321 lớn hơn mức ý nghĩa $\alpha = 0.05$ nên chấp nhận H_0, hay là không có đủ bằng chứng thống kê để cho rằng số vụ cắt điện trong ngày không tuân theo phân phối Poisson với $\lambda = 4.2$.

b. Khi chưa biết λ, ta sẽ ước lượng λ dựa vào mẫu và tiến hành kiểm định như sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Số vụ cắt điện trong một ngày ở thành phố tuân theo phân phối Poisson.
  H_1 : Số vụ cắt điện trong một ngày ở thành phố không tuân theo phân phối Poisson.
- Bước 2: Ước lượng tham số λ và thực hiện kiểm định trên R:

```
> k = 0:9
> x = c(0, 5, 22, 23, 32, 22, 19, 13, 6, 8)
> lambda = sum(k*x)/sum(x)
> lambda
[1] 4.54
> p0 = c(dpois(0:8, lambda=4.54),1-ppois(8,lambda=4.54))
> chisq.test(x, p=p0)

        Chi-squared test for given probabilities

data:  x
X-squared = 5.9704, df = 9, p-value = 0.7429

> qchisq(0.95,8)
[1] 15.50731
```

- Bước 3: Trong trường hợp chưa biết λ, ta sẽ dùng giá trị thống kê $\chi^2 = 5.9704$ so sánh với giá trị tới hạn $\chi^2_{k-m-1,\alpha} = \chi^2_{10-1-1,0.05} = 15.5$. Do $5.9 < 15.5$ nên ta chấp nhận H_0 hay ta không có đủ bằng chứng thống kê để cho rằng số vụ cắt điện trong một ngày ở thành phố này là không tuân theo phân phối Poisson.

Ví dụ 11.2.4 Đo chiều cao (cm) của 500 người dân Việt Nam ở tuổi trưởng thành ta thu được bảng dữ liệu sau:

Khoảng chiều cao	<150	[150,155)	[155,160)	[160,165)	[165,170)	$\geqslant$ 170
Số người	10	77	164	170	65	14

a. Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem chiều cao của người dân Việt Nam có tuân theo phân phối chuẩn với trung bình là 160 và độ lệch chuẩn là 5 hay không?

b. Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem chiều cao của người dân Việt Nam có tuân theo phân phối chuẩn không.

Lời giải: Ví dụ 11.2.4 yêu cầu ta kiểm định xem chiều cao của người dân Việt Nam dựa trên bộ số liệu đã cho có tuân theo phân phối chuẩn hay không. Ta thực hiện bài toán trong cả hai trường hợp biết và không biết tham số trung bình, phương sai.

a. Trong trường hợp biết $\mu = 160, \sigma^2 = 5^2$, ta tiến hành kiểm định theo các bước sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Chiều cao của người dân tuân theo phân phối chuẩn với $\mu = 160, \sigma^2 = 5^2$.
  H_1 : Chiều cao của người dân không tuân theo phân phối chuẩn với $\mu = 160$, $\sigma^2 = 5^2$.
- Bước 2: Tính các xác suất và thực hiện kiểm định trên R:

```
> x = c(10, 77, 164, 170, 65, 14)
> #%%%%%%%%%%
> p1 = pnorm(150,160,5)
> p2 = pnorm(155,160,5)-pnorm(150,160,5)
> p3 = pnorm(160,160,5)-pnorm(155,160,5)
> p4 = pnorm(165,160,5)-pnorm(160,160,5)
> p5 = pnorm(170,160,5)-pnorm(165,160,5)
> p6 = 1 - pnorm(170,160,5)
> p0 = c(p1,p2,p3,p4,p5,p6)
> #%%%%%%%%%%%
> #Hoặc ta có thể làm như sau:
> p1 = pnorm(150,160,5)
> a = c(150, 155, 160, 165, 170)
> p2 = diff(pnorm(a,160,5))
> p3 = 1 - pnorm(170,160,5)
> p0 = c(p1, p2, p3)
> #%%%%%%%%%%%%
> chisq.test(x, p=p0)

        Chi-squared test for given probabilities

data:  x
X-squared = 2.3684, df = 5, p-value = 0.7962
```

- Bước 3: p-giá trị của phép kiểm định là 0.7962 > 0.05 dẫn đến ta chấp nhận H_0. Vậy tại mức ý nghĩa $\alpha = 5\%$, ta chưa có đủ cơ sở thống kê để cho rằng chiều cao của người dân Việt Nam không tuân theo phân phối chuẩn với trung bình là 160 cm và độ lệch chuẩn là 5 cm.

b. Khi chưa biết μ, σ^2, ta sẽ ước lượng μ và σ^2 dựa vào mẫu và tiến hành kiểm định như sau:

- Bước 1: Đặt cặp giả thuyết:
  H_0 : Chiều cao của người dân tuân theo phân phối chuẩn.
  H_1 : Chiều cao của người dân không tuân theo phân phối chuẩn.
- Bước 2: Ước lượng các tham số μ, σ^2 và thực hiện kiểm định trên R:

```
> k = c(150,152.5,157.5,162.5,167.5,170)
> x = c(10,77,164,170,65,14)
> tb = sum(k*x)/sum(x)
> tb
[1] 159.93
> dlc =sqrt(sum((k-tb)^2*x)/(sum(x)-1))
> dlc
[1] 4.999509
> p1 = pnorm(150,mean=tb,sd=dlc)
> a = c(150, 155, 160, 165, 170)
> p2 = diff(pnorm(a,mean=tb,sd=dlc))
> p3 = 1 - pnorm(170,mean=tb,sd=dlc)
> p0 = c(p1, p2, p3)

> chisq.test(x,p=p0)

        Chi-squared test for given probabilities

data:  x
X-squared = 2.3372, df = 5, p-value = 0.8008

> qchisq(0.95,3)
[1] 7.814728
```

- Bước 3: Trong trường hợp chưa biết μ, σ^2, ta dùng giá trị thống kê $\chi^2 = 2.3372$ so sánh với giá trị tới hạn $\chi^2_{k-m-1,\alpha} = \chi^2_{6-2-1,0.05} = 7.82$. Do $2.3372 < 7.82$ nên ta chấp nhận H_0. Vậy ta chưa có đủ bằng chứng thống kê để cho rằng chiều cao của người dân Việt Nam không tuân theo phân phối chuẩn.

Trong thống kê, để kiểm tra một tổng thể có tuân theo phân phối chuẩn hay không, ngoài cách dùng kiểm định khi-bình phương như trên, ta còn có thể dùng nhiều phương pháp khác. Trong R giới thiệu hàm `shapiro.test` dùng kiểm định một tập dữ liệu có tuân theo phân phối chuẩn hay không dựa trên kiểm định Shapiro–Wilk được đưa ra bởi Samuel Shapiro và Martin Wilk vào năm 1965, chi tiết có thể tham khảo tại:

(`http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test`)
và hàm `jarque.bera.test` trong gói `tseries` theo kiểm định Jarque–Bera được giới thiệu bởi Carlos Jarque và Anil K. Bera. Kiểm định Jarque–Bera kiểm định phân phối chuẩn dựa trên hệ số bất đối xứng skewness và hệ số độ nhọn kurtosis, chi tiết có thể tham khảo tại:

(`http://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test`)

Minh họa sau cho ta cách sử dụng hai hàm này trên R:

```
> DuLieu = rnorm(1000,100,15)
> shapiro.test(DuLieu)
```

```
        Shapiro-Wilk normality test

data:  DuLieu
W = 0.9986, p-value = 0.6191

> library(tseries)
> jarque.bera.test(DuLieu)


        Jarque Bera Test

data:  DuLieu
X-squared = 0.0413, df = 2, p-value = 0.9795
```

Các p-giá trị từ hai hàm này cho ta thấy **DuLieu** tuân theo phân phối chuẩn tại mức ý nghĩa $\alpha = 5\%$.

BÀI TẬP

XI.120. Một mẫu ngẫu nhiên gồm 100 phụ nữ ở thành phố Hồ Chí Minh được chọn ra để hỏi mức độ ưa thích về 5 loại xà phòng A, B, C, D và E thì thu được kết quả như sau:

Loại xà phòng	A	B	C	D	E	Tổng
Số phụ nữ chọn	18	16	23	20	23	100

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm tra xem 5 loại xà phòng này có được ưa thích như nhau đối với phụ nữ ở thành phố HCM hay không.

XI.121. Theo hồ sơ lưu trữ thì nếu qui trình sản xuất là bình thường thì có 93% số sản phẩm không bị sai sót, 5% có một sai sót và 2% có hơn một sai sót. Một mẫu ngẫu nhiên gồm 500 sản phẩm được chọn ra từ đợt sản xuất tuần qua thì thấy 458 sản phẩm không sai sót, 30 sản phẩm có một sai sót và 12 sản phẩm có hơn một sai sót. Hãy kiểm chứng ở mức ý nghĩa 5% xem chất lượng sản phẩm tuần qua có bình thường không.

XI.122. Theo số liệu các năm trước, giải thưởng trúng được khi mua một loại xà phòng trong dịp tết Nguyên đán phân về ba miền Bắc, Trung, Nam theo tỉ lệ $3 : 2 : 2$. Phỏng vấn 400 người trúng giải nhân lễ trao giải thưởng năm nay, thấy có 180 người miền Bắc, 100 người miền Trung và 120 người miền Nam. Tại mức ý nghĩa $\alpha = 5\%$, có thể kết luận tỉ lệ giải thưởng phân về ba miền năm nay còn như mọi năm không.

XI.123. Một nhân viên mỗi ngày tiếp xúc với 5 khách hàng. Giả thiết cho rằng số lần bán được hàng trong một ngày tuân theo phân phối nhị thức với $p = 0.4$. Kết quả trong nhiều ngày như sau:

Số lần bán trong ngày	0	1	2	3	4	5
Số ngày	10	41	60	20	6	3

Ở mức ý nghĩa 5%, ta có thể kết luận rằng phân phối của số lần bán được hàng thực sự tuân theo phân phối đã giả thiết trên không?

XI.124. Người ta muốn kiểm tra xem số người đến chi cục thuế ở một huyện trong một giờ có tuân theo phân phối Poisson không. Quan sát ngẫu nhiên nhiều khoảng thời gian một giờ tại chi cục thuế này, ta thu được bảng số liệu sau:

Số người	0	1	2	3	4	5	trên 5
Số khoảng 1h	8	12	20	25	15	15	5

Tại mức ý nghĩa $\alpha = 5\%$, hãy xét xem số người đến chi cục thuế trong một giờ có tuân theo phân phối Poisson với $\lambda = 3$ không.

XI.125. Một trong những cách để quyết định ai là tác giả của một tác phẩm văn học là so sánh tần số xuất hiện của một từ nào đó. Nghiên cứu số lần xuất hiện của từ "có thể" trong nhiều đoạn văn dài xấp xỉ 200 từ được chọn ngẫu nhiên, người ta thu được kết quả sau:

Số lần xuất hiện	0	1	2	$\geqslant 3$
Số đoạn văn	156	63	29	14

Tại mức ý nghĩa $\alpha = 5\%$, hãy xét xem phân phối tần số xuất hiện của từ "có thể" có tuân theo phân phối Poisson với $\lambda = 0.6$ không.

XI.126. Các phương tiện giải trí ở TPHCM sẽ được đánh giá là tốt, vừa hoặc xấu. Một mẫu ngẫu nhiên gồm nhiều người có mức thu nhập trên trung bình, trung bình và dưới trung bình được chọn ra để hỏi ý kiến. Kết quả xếp lớp chéo như sau:

	Ý kiến		
Mức thu nhập	Tốt	Vừa	Xấu
Trên trung bình	175	124	92
Trung bình	118	110	126
Dưới trung bình	127	82	147

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem có mối liên hệ giữa mức thu nhập với thái độ đối với phương tiện giải trí không.

XI.127. Người ta nghiên cứu xem có mối liên hệ giữa thời gian tìm hiểu trước hôn nhân và tình trạng hiện tại của cuộc hôn nhân hay không. Thời gian tìm hiểu trước hôn nhân được chia làm ba mức là ngắn, trung bình và dài và tình trạng hôn nhân được chia làm ba mức là hạnh phúc, không hạnh phúc và ly dị/ly thân. Bảng sau cho ta số liệu điều tra về 200 cặp vợ chồng có thời gian kết hôn trên 5 năm:

	Thời gian tìm hiểu		
Tình trạng hôn nhân	Ngắn	Trung bình	Dài
Hạnh phúc	38	58	54
Không hạnh phúc	12	14	4
Ly dị/Ly thân	10	8	2

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem thời gian tìm hiểu trước hôn nhân có liên hệ với tình trạng hôn nhân không.

XI.128. Để tìm hiểu xem thời gian nằm bệnh viện của bệnh nhân có phụ thuộc vào các loại bảo hiểm (phần trăm chi phí được bảo hiểm chi trả) không, người ta thu thập một mẫu gồm 660 thời gian nằm viện và được xếp chéo với loại bảo hiểm như sau:

	Số ngày nằm viện		
Loại bảo hiểm	< 5	$5 - 10$	> 10
$< 25\%$	40	75	65
$25\% - 50\%$	30	45	75
$> 50\%$	40	100	190