Bài giảng Xác suất Thống kê và ứng dụng

Nguyễn Thị Nhung

Bộ môn Toán - Đại học THĂNG LONG

Ngày 24 tháng 8 năm 2013

Chương XI

Kiểm định Chi-bình phương

Kiểm chứng tính độc lập

Trong thực tế nhiều khi ta gặp một số tình huống đòi hỏi phải tìm hiểu mối liên hệ giữa các biến định tính, chẳng hạn

- Có hay không có mối liên hệ giữa thời gian nghe nhạc và kết quả học tập?
- Có người yêu có ảnh hưởng đến kết quả học tập hay không?
- Có mối liên hệ giữa độ bền của một cuộc hôn nhân với thời gian yêu nhau trước khi kết hôn hay không?
- Giới tính có ảnh hưởng đến việc thuận tay trái nào hay không?

Nội dung trình bày

Câu hỏi dẫn nhập

Bài toán

Ta xét hai biến định tính và muốn kiểm tra xem mối quan hệ giữa chúng là độc lập hay phụ thuộc. Để thực hiện việc này ta kiểm định cặp giả thuyết sau:

H_0 : *Hai biến định tính độc lập (hay không có mối liên hệ giữa hai biến này);*

H_1 : *Hai biến định tính không độc lập (có mối liên hệ giữa hai biến này).*

Qui trình thực hiện

- Giả sử biến định tính thứ nhất gồm r loại, biến định tính thứ hai gồm c loại. Chọn từ tổng thể ra mẫu gồm n phần tử xếp chéo thành $r \times c$ giá trị $Oij, i = 1, \ldots, r; j = 1, \ldots, c$, trong đó O_{ij} là số quan sát có thuộc tính thứ i của biến thứ nhất và thuộc tính thứ j của biến thứ hai. Khi đó ta có bảng 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

Qui trình thực hiện

- Giả sử H_0 đúng, khi đó tần số lí thuyết E_{ij} của ô ở địa chỉ ij được tính theo công thức:

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

Ta có giá trị kiểm định:

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

- Nếu giả thuyết H_0 đúng và $E_{ij} \geqslant 5, \forall i, j$ thì χ^2 tuân theo phân phối chi- bình phương với $(r-1)(c-1)$ bậc tự do.
- So sánh giá trị kiểm định với $\chi^2_{(r-1)(c-1),\alpha}$, tại mức ý nghĩa α ta đưa ra quyết định bác bỏ H_0 nếu

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

Ví dụ

Bài toán

Một nhà nghiên cứu cho rằng điểm của các học sinh phụ thuộc vào số lượng thời gian chúng nghe nhạc. Một mẫu ngẫu nhiên gồm 400 học sinh được chọn và được xếp lớp chéo giữa điểm trung bình cuối năm với thời gian nghe nhạc hàng tuần như sau:

Thời gian nghe nhạc	Điểm trung bình					Tổng
	Xuất sắc	Giỏi	Khá	Trung bình	Kém	
$<$ 5h	13	10	11	16	5	55
5h - 10h	20	27	27	19	2	95
11h - 20h	9	27	71	16	32	155
$>$ 20h	8	11	41	24	11	95
Tổng	50	75	150	75	50	400

Ở mức ý nghĩa $\alpha = 5\%$ hãy kiểm tra xem điểm trung bình có phụ thuộc vào thời gian nghe nhạc hay không.

Ví dụ

- Đặt cặp giả thuyết cho bài toán:

- Các giá trị tần số lí thuyết E_{ij} được cho tương ứng trong bảng sau:

Thời gian nghe nhạc	Điểm trung bình				
	Xuất sắc	Giỏi	Khá	Trung bình	Kém
$<$ 5h	6.875	10.3125	20.625	10.3125	6.875
5h - 10h	11.875	17.8125	35.625	17.8125	11.875
11h - 20h	19.375	29.0625	58.125	29.0625	19.375
$>$ 20h	11.875	17.8125	35.625	17.8125	11.875

Ví dụ

- Đặt cặp giả thuyết cho bài toán:

- Các giá trị tần số lí thuyết E_{ij} được cho tương ứng trong bảng sau:

Thời gian nghe nhạc	Điểm trung bình				
	Xuất sắc	Giỏi	Khá	Trung bình	Kém
$<$ 5h	6.875	10.3125	20.625	10.3125	6.875
5h - 10h	11.875	17.8125	35.625	17.8125	11.875
11h - 20h	19.375	29.0625	58.125	29.0625	19.375
$>$ 20h	11.875	17.8125	35.625	17.8125	11.875

Ví dụ

- Tính toán giá trị kiểm định

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 63.8296.$$

- Giá trị tới hạn $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{12,0.05} = 21.02607.$
- Kết luận:

Ví dụ

- Tính toán giá trị kiểm định

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 63.8296.$$

- Giá trị tới hạn $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{12,0.05} = 21.02607.$
- Kết luận:

Ví dụ kiểm định tính độc lập trong R

```
> SoHocSinh = c(13, 20, 9, 8, 10, 27, 27, 11, 11, 27,
71, 41, 16, 19, 16, 24, 5, 2, 32, 11)
> A = matrix(SoHocSinh, nrow = 4)
> chisq.test(A)

        Pearson's Chi-squared test

data:  A
X-squared = 63.8296, df = 12, p-value = 4.483e-09
```

Nội dung trình bày

Câu hỏi dẫn nhập

Khi tiến hành thực hiện các bài toán xác suất hoặc thống kê, ta thường giả sử rất nhiều giả thiết, chẳng hạn:

- Giả sử quân xúc xắc là cân đối và đồng chất;
- Giả sử chỉ số IQ của dân chúng tuân theo phân phối chuẩn với trung bình $\mu = 100$ và phương sai $\sigma = 15$;
- Giả sử số vụ án mạng trong ngày tuân theo phân phối Poisson với $\lambda = 10$;
- Giả sử số dặm mà một chiếc ô tô đi được cho đến khi không sử dụng được nữa tuân theo phân phối mũ với $\lambda = \dfrac{1}{20}$.

Làm thế nào để ta có thể đưa ra được những giả sử như trên để thực hiện các bài toán?

Kiểm định về qui luật phân phối xác suất của tổng thể

Bài toán

Giả sử ta chưa biết qui luật phân phối xác suất của tổng thể, ta cần kiểm định xem phân phối của tổng thể có tuân theo một qui luật xác suất A nào đó không bằng cách kiểm định cặp giả thuyết sau:

H_0 : *Tổng thể tuân theo qui luật xác suất A;*

H_1 : *Tổng thể không tuân theo qui luật xác suất A.*

Qui trình thực hiện

- Chọn một mẫu ngẫu nhiên gồm n phần tử mà mỗi phần tử được xếp vào đúng một trong k lớp. Gọi $O_1, O_2, \ldots, O_k$ lần lượt là số phần tử rơi vào k lớp trên.
- Nếu H_0 đúng, tức là tổng thể tuân theo qui luật xác suất A, thì xác suất để một phần tử rơi vào lớp $1, 2, \ldots, k$ lần lượt là $p_1, p_2, \ldots, p_k$ với $p_1 + p_2 + \ldots + p_k = 1$. Khi đó số phần tử kì vọng theo k lớp đó sẽ là $E_i = np_i, i = 1, 2, \ldots, k$.

Lớp	1	2	...	k	Tổng
Số phần tử quan sát	O_1	O_2	...	O_k	n
Xác suất theo H_0	p_1	p_2	...	p_k	1
Số phần tử theo H_0	$E_1 = np_1$	$E_2 = np_2$	...	$E_k = np_k$	n

Qui trình thực hiện

- Chọn một mẫu ngẫu nhiên gồm n phần tử mà mỗi phần tử được xếp vào đúng một trong k lớp. Gọi $O_1, O_2, \ldots, O_k$ lần lượt là số phần tử rơi vào k lớp trên.
- Nếu H_0 đúng, tức là tổng thể tuân theo qui luật xác suất A, thì xác suất để một phần tử rơi vào lớp $1, 2, \ldots, k$ lần lượt là $p_1, p_2, \ldots, p_k$ với $p_1 + p_2 + \ldots + p_k = 1$. Khi đó số phần tử kì vọng theo k lớp đó sẽ là $E_i = np_i, i = 1, 2, \ldots, k$.

Lớp	1	2	...	k	Tổng
Số phần tử quan sát	O_1	O_2	...	O_k	n
Xác suất theo H_0	p_1	p_2	...	p_k	1
Số phần tử theo H_0	$E_1 = np_1$	$E_2 = np_2$	...	$E_k = np_k$	n

Qui trình thực hiện

- Nếu H_0 đúng và cỡ mẫu lớn sao cho $E_i = np_i \geqslant 5, \forall i = \overline{1,k}$ thì biến ngẫu nhiên

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

tuân theo phân phối xấp xỉ phân phối chi- bình phương với $k - m - 1$ bậc tự do, trong đó m là số tham số tổng thể ước lượng từ dữ liệu mẫu.

- Tại mức ý nghĩa α, giả thuyết H_0 bị bác bỏ nếu

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

Qui trình thực hiện

- Nếu H_0 đúng và cỡ mẫu lớn sao cho $E_i = np_i \geqslant 5, \forall i = \overline{1,k}$ thì biến ngẫu nhiên

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

tuân theo phân phối xấp xỉ phân phối chi- bình phương với $k - m - 1$ bậc tự do, trong đó m là số tham số tổng thể ước lượng từ dữ liệu mẫu.

- Tại mức ý nghĩa α, giả thuyết H_0 bị bác bỏ nếu

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

Ví dụ

Bài toán

Để kiểm định xem quân xúc xắc có cân đối và đồng chất hay không, người ta tiến hành tung con xúc xắc 120 lần và nhận được kết quả như sau:

Số chấm	*1*	*2*	*3*	*4*	*5*	*6*	*Tổng*
Số lần tung	*28*	*14*	*26*	*18*	*15*	*19*	*120*

Tại mức ý nghĩa $\alpha = 5\%$ *có thể kết luận con xúc xắc là cân đối và đồng chất hay không?*

Ví dụ

- Bảng mô tả mối quan hệ giữa số lần tung quan sát và số lần tung kì vọng:

Số chấm	1	2	3	4	5	6
Số lần tung quan sát (O_i)	28	14	26	18	15	19
Xác suất kì vọng (p_i)	1/6	1/6	1/6	1/6	1/6	1/6
Số lần tung kì vọng E_i	20	20	20	20	20	20

- Giá trị thống kê:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 8.3.$$

- Giá trị tới hạn $\chi^2_{5,0.05} = 11.07$.

Lời giải

Bài tập tự học

Bài toán

Một công ty thương mại dựa vào kinh nghiệm quá khứ đã xác định rằng vào cuối năm thì 80% số hóa đơn đã được thanh toán đầy đủ, 10% khất lại một tháng, 6% khất lại hai tháng và 4% khất lại trên hai tháng. Vào cuối năm nay công ty kiểm tra một mẫu ngẫu nhiên gồm 400 hóa đơn, ta thấy 287 được thanh toán đầy đủ, 49 khất lại một tháng, 30 khất lại hai tháng và 34 khất lại trên hai tháng. Tại mức ý nghĩa $\alpha = 5\%$, những dữ liệu này gợi ý rằng mô thức của năm nay có còn giống những năm trước nữa không?

Hướng dẫn

- Bảng mô tả mối quan hệ giữa số hóa đơn quan sát và số hóa đơn kì vọng:

Số tháng khất lại	0	1	2	>2	Tổng
Số hóa đơn quan sát (O_i)	287	49	30	34	400
Xác suất kì vọng (p_i)	0.8	0.1	0.06	0.04	1
Số hóa đơn kì vọng (E_i)	320	40	24	16	400

- $\chi^2 = 27.178$; $\chi^2_{3,0.05} = 7.81$.

Ví dụ

Bài toán

Để điều tra xem số vụ án mạng trong ngày ở London có tuân theo phân phối Poisson hay không, người ta điều tra số vụ án mạng xảy ra từ 04/2004 đến 03/2007 và thu được bảng số liệu sau:

Số vụ án mạng	0	1	2	$\geqslant 3$	Tổng
Số ngày	713	299	66	17	1095

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem số vụ án mạng hàng ngày ở London có tuân theo phân phối Poisson hay không?

Lời giải

- Trung bình mẫu là:

$$\overline{x} = \frac{0 \times 713 + 1 \times 299 + 2 \times 66 + 3 \times 17}{1095} = 0.44.$$

- Bảng mô tả mối quan hệ giữa số tần số quan sát và kì vọng:

ố vụ án mạng	0	1	2	$\geqslant 3$	Tổng
Số ngày	713	299	66	17	1095
Xác suất kì vọng (p_i)	0.6440	0.2834	0.0623	0.0103	1
Số từ kì vọng (E_i)	705.1800	310.3230	68.2185	11.2785	1095

- Giá trị thống kê:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 3.475.$$

- $\chi^2_{2,0.05} = 5.99.$

Lời giải

Kiểm chứng mức phù hợp trong R bằng hàm `chisq.test`

- Thực hiện kiểm chứng mức phù hợp trong *R* bằng hàm `chisq.test(x, p =` p_0`)`
- trong đó,
  - `x` là véc tơ chỉ các quan sát trong mẫu,
  - p_0 là véc tơ xác suất chỉ qui luật phân phối của tổng thể.

Ví dụ kiểm chứng mức phù hợp trong R

```
> x = c(287, 49, 30, 34)
> p0 = c(0.8, 0.1, 0.06, 0.04)
> chisq.test(x, p = p0)

        Chi-squared test for given probabilities

data:  x
X-squared = 27.1781, df = 3, p-value = 5.402e-06
```

Thực hiện kiểm chứng mức phù hợp trong R

Bài toán

Để điều tra xem số vụ án mạng trong ngày ở London có tuân theo phân phối Poisson hay không, người ta điều tra số vụ án mạng xảy ra từ 04/2004 đến 03/2007 và thu được bảng số liệu sau:

Số vụ án mạng	0	1	2	$\geqslant 3$	Tổng
Số ngày	713	299	66	17	1095

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem số vụ án mạng hàng ngày ở London:

a. *có tuân theo phân phối Poisson với $\lambda = 0.44$ không?*

b. *có tuân theo phân phối Poisson không?*

Ví dụ kiểm chứng mức phù hợp trong R

- Kiểm định câu **(a)**:

```
> x = c(713, 299, 66, 17)
> p0 = c(dpois(0:2, lambda = 0.44), 1-ppois(0:2, lambda = 0.44))
> chisq.test(x, p = p0)

        Chi-squared test for given probabilities

data:  x
X-squared = 3.5523, df = 3, p-value = 0.314
```

- Kiểm định câu **(b)**:
  - Giá trị thống kê $\chi^2 = 3.55$.
  - Giá trị tới hạn $\chi^2_{2,0.05} = 5.99$.
  - Kết luận: Do $3.474492 < 5.991456$ nên ta chấp nhận H_0.

Nội dung trình bày

Kiểm chứng phân phối chuẩn

Giả sử ta có một mẫu ngẫu nhiên gồm n phần tử, ta cần kiểm tra xem tập dữ liệu này có phải chọn từ một tổng thể tuân theo phân phối chuẩn hay không. Để thực hiện việc này ta kiểm định cặp giả thuyết:
H_0 : Tổng thể tuân theo phân phối chuẩn;
H_1 : Tổng thể không tuân theo phân phối chuẩn.
Ta có thể kiểm định theo một trong hai cách sau:

Kiểm chứng phân phối chuẩn

- Cách 1: Dựa trên qui luật phân phối xác suất
  - Xếp các phần tử của mẫu vào các khoảng phù hợp, giả sử khoảng thứ i có O_i phần tử của mẫu;
  - Ước tính các tham số μ và σ theo mẫu (nếu chưa biết);
  - Tính các giá trị xác suất p_i theo H_0 và tần số E_i theo H_0;
  - Tính giá trị thống kê:

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

  - Tại mức ý nghĩa α ta quyết định bác bỏ H_0 nếu $\chi^2 > \chi^2_{k-2-1,\alpha}$ (ước lượng μ, σ) hoặc $\chi^2 > \chi^2_{k-1,\alpha}$ (không cần ước lượng μ, σ).

Kiểm chứng phân phối chuẩn (Bowman-Shelton)

- Cách 2: Kiểm chứng Bowman-Shelton
  Kiểm chứng này dựa trên hai đặc trưng của phân phối chuẩn thông qua hai hệ số độ nghiêng (Skewness) và hệ số đo độ nhọn (Kurtosis).
  - Đặc trưng thứ nhất của phân phối chuẩn là tính đối xứng qua trung bình. Đặc trưng này thể hiện qua hệ số đo độ nghiêng bằng 0:

$$Skewness = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3/n}{s^3},$$

    trong đó $\overline{x}, s$ là trung bình và độ lệch chuẩn của mẫu.
  - Đặc trưng thứ hai của phân phối chuẩn là hệ thức giữa độ phẳng của các phần đuôi của phân phối so với phần trung tâm thể hiện qua hệ số Kurtosis bằng 3:

$$Kurtosis = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4/n}{s^4},$$

    trong đó $\overline{x}, s$ là trung bình và độ lệch chuẩn của mẫu.

Kiểm chứng phân phối chuẩn (Bowman-Shelton)

- Giá trị thống kê của phép kiểm định là

$$B = n\left(\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24}\right).$$

- Nếu H_0 đúng và cỡ mẫu lớn thì B sẽ có phân phối chi-bình phương với 2 bậc tự do. Nếu cỡ mẫu nhỏ thì ta sẽ so sánh B với giá trị tương ứng trong bảng giá trị Bowman-Shelton.
- Trong trường hợp cỡ mẫu rất lớn, ta sẽ bác bỏ H_0 nếu $B > \chi^2_{2,\alpha}$, nếu cỡ mẫu nhỏ ta sẽ bác bỏ H_0 nếu B lớn hơn giá trị tương ứng trong bảng giá trị Bowman-Shelton.

Thực hiện kiểm định phân phối chuẩn trong R

- Để kiểm chứng phân phối chuẩn trong R ta dùng hàm `shapiro.test(x)`, trong đó x là véc tơ dữ liệu.
- Kiểm định phân phối dữ liệu điểm thi Toán khối A năm 2008 phù hợp với phân phối chuẩn:

```
> shapiro.test(ToanAmoi)

        Shapiro-Wilk normality test

data:  ToanAmoi
W = 0.953, p-value < 2.2e-16
```