THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 17 tháng 12 năm 2018

Nội dung môn học

1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. Kiểm định giả thuyết về tham số một tổng thể
8. Kiểm định giả thuyết về tham số hai tổng thể
9. Phân tích phương sai
10. Kiểm định phi tham số
11. **Kiểm định chi - bình phương**
12. Hồi quy đơn biến
13. Hồi quy đa biến

Nội dung chính

Trong phần này chúng ta sẽ tìm hiểu:

1. kiểm định xem hai yếu tố định tính nào đó có mối quan hệ với nhau hay không?

Nội dung chính

Trong phần này chúng ta sẽ tìm hiểu:

1. kiểm định xem hai yếu tố định tính nào đó có mối quan hệ với nhau hay không? Ví dụ
   - Có hay không mối liên hệ giữa thời gian nghe nhạc với kết quả học tập của sinh viên?

Nội dung chính

Trong phần này chúng ta sẽ tìm hiểu:

1. kiểm định xem hai yếu tố định tính nào đó có mối quan hệ với nhau hay không? Ví dụ
   - Có hay không mối liên hệ giữa thời gian nghe nhạc với kết quả học tập của sinh viên?
   - Có người yêu có ảnh hưởng đến kết quả học tập không?

Nội dung chính

Trong phần này chúng ta sẽ tìm hiểu:

1. kiểm định xem hai yếu tố định tính nào đó có mối quan hệ với nhau hay không? Ví dụ
   - Có hay không mối liên hệ giữa thời gian nghe nhạc với kết quả học tập của sinh viên?
   - Có người yêu có ảnh hưởng đến kết quả học tập không?
   - Giới tính có ảnh hưởng đến việc thuận tay trái không?...

Nội dung chính

Trong phần này chúng ta sẽ tìm hiểu:

1. kiểm định xem hai yếu tố định tính nào đó có mối quan hệ với nhau hay không? Ví dụ
   - Có hay không mối liên hệ giữa thời gian nghe nhạc với kết quả học tập của sinh viên?
   - Có người yêu có ảnh hưởng đến kết quả học tập không?
   - Giới tính có ảnh hưởng đến việc thuận tay trái không?...
2. kiểm định so sánh tỉ lệ hai tổng thể và kiểm định một dãy tỉ lệ cho trước có còn đúng với số liệu thu thập được hay không.

Nội dung chính

Trong phần này chúng ta sẽ tìm hiểu:

1. kiểm định xem hai yếu tố định tính nào đó có mối quan hệ với nhau hay không? Ví dụ
   - Có hay không mối liên hệ giữa thời gian nghe nhạc với kết quả học tập của sinh viên?
   - Có người yêu có ảnh hưởng đến kết quả học tập không?
   - Giới tính có ảnh hưởng đến việc thuận tay trái không?...
2. kiểm định so sánh tỉ lệ hai tổng thể và kiểm định một dãy tỉ lệ cho trước có còn đúng với số liệu thu thập được hay không.

$\longrightarrow$ Thủ tục kiểm định các bài toán này dựa trên phân phối Chi bình phương.

Phần XI

Kiểm định chi bình phương

Mục lục

Example

Một điều tra giới tính và quan điểm nên lập gia đình muộn hay sớm cho thấy trong số 200 nam có 120 người cho rằng nên lập gia đình muộn, trong khi đó có 85 trong số 160 nữ cho rằng nên lập gia đình muộn. Qua số liệu trên có thể khẳng định rằng quan điểm về kết hôn sớm hay muộn có phụ thuộc vào giới tính không?

Example

Một điều tra giới tính và quan điểm nên lập gia đình muộn hay sớm cho thấy trong số 200 nam có 120 người cho rằng nên lập gia đình muộn, trong khi đó có 85 trong số 160 nữ cho rằng nên lập gia đình muộn. Qua số liệu trên có thể khẳng định rằng quan điểm về kết hôn sớm hay muộn có phụ thuộc vào giới tính không?

Để dễ hình dung ta có thể lập thành một bảng sau:

	Sớm	Muộn	Tổng dòng
Nam	80	120	200
Nu	75	85	160
Tổng cột	155	205	360

Bảng: Tần số thực tế

Phân tích tình huống

Cặp giả thuyết:
H_0: *Quan điểm về thời gian kết hôn và giới tính là độc lập.*
H_1: *Quan điểm về thời gian kết hôn và giới tính là không độc lập.*

Phân tích tình huống

Giả sử H_0 xảy ra, khi đó bảng tần số sẽ thế nào?

	Sớm	*Muộn*	*Tổng dòng*
Nam	$200 - x$	x	*200*
Nu	$160 - y$	y	*160*
Tổng cột	*155*	*205*	*360*

Nếu H_0 xảy ra, thì khi đó tỉ lệ nam cho rằng nên kết hôn muộn và tỉ lệ nữ có quan điểm tương tự phải xấp xỉ nhau:

$$\frac{x}{200} \approx \frac{y}{160} \Rightarrow \frac{x}{200} \approx \frac{y}{160} \approx \frac{x+y}{200+160} \approx \frac{205}{360}$$

Do đó,

$$x \approx \frac{200 \times 205}{360}, y \approx \frac{160 \times 205}{360}$$

Phân tích tình huống

Bởi thế nếu H_0 xảy ra thì bảng thu được xấp xỉ bảng sau:

	Sớm	*Muộn*	*Tổng dòng*
Nam	$\frac{200 \times 155}{360}$	$\frac{200 \times 205}{360}$	*200*
Nữ	$\frac{160 \times 155}{360}$	$\frac{160 \times 205}{360}$	*160*
Tổng cột	*155*	*205*	*360*

Bảng: *Tần số lí thuyết*

Phân tích tình huống

Vậy, nếu H_0 mà đúng thì

- *bảng tần số thực tế phải xấp xỉ với bảng tần số lí thuyết.*
- *Nếu hai bảng này càng khác xa nhau, khả năng H_0 đúng càng thấp.*

$\longrightarrow$ *Cần một quy luật nào đó giúp ta không những đo được độ xa gần đó mà còn giúp được ta vạnh ra ranh giới quyết định vùng nào được coi là gần ở mức chấp nhận được, vùng nào là xa tới mức không chấp nhận được.*

Bài toán

Ta xét hai biến định tính và muốn kiểm tra xem mối quan hệ giữa chúng là độc lập hay phụ thuộc.
Cặp giả thuyết sau:
H_0 : *Hai biến định tính độc lập (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).*

Chọn mẫu và lập bảng tần số

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 và lập bảng tần số chéo gồm r dòng, n cột. Trong bảng, kí hiệu $Oij, i = 1, \ldots, r; j = 1, \ldots, c$, tần 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

Logic của kiểm định

Giả sử H_0 đúng, tức là hai yếu tố là độc lập, khi đó về mặt lí thuyết ô ở vị trí ij phải mang giá trị là:

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

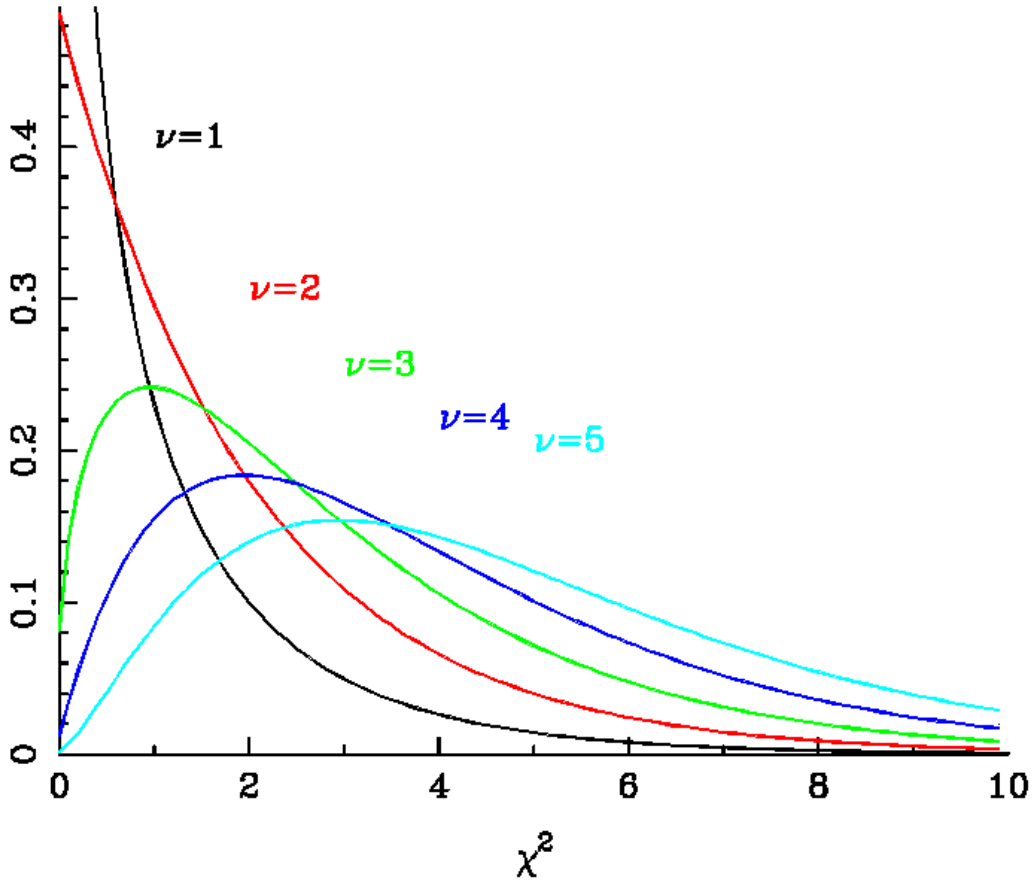
Khi đó bằng cảm nhận ta thấy rằng nếu các tần số quan sát được trong thực tế O_{ij} càng xa so với tần số lí thuyết E_{ij} (là tần số khi giả định H_0 đúng) thì khả năng bác bỏ H_0 càng lớn.

Theorem

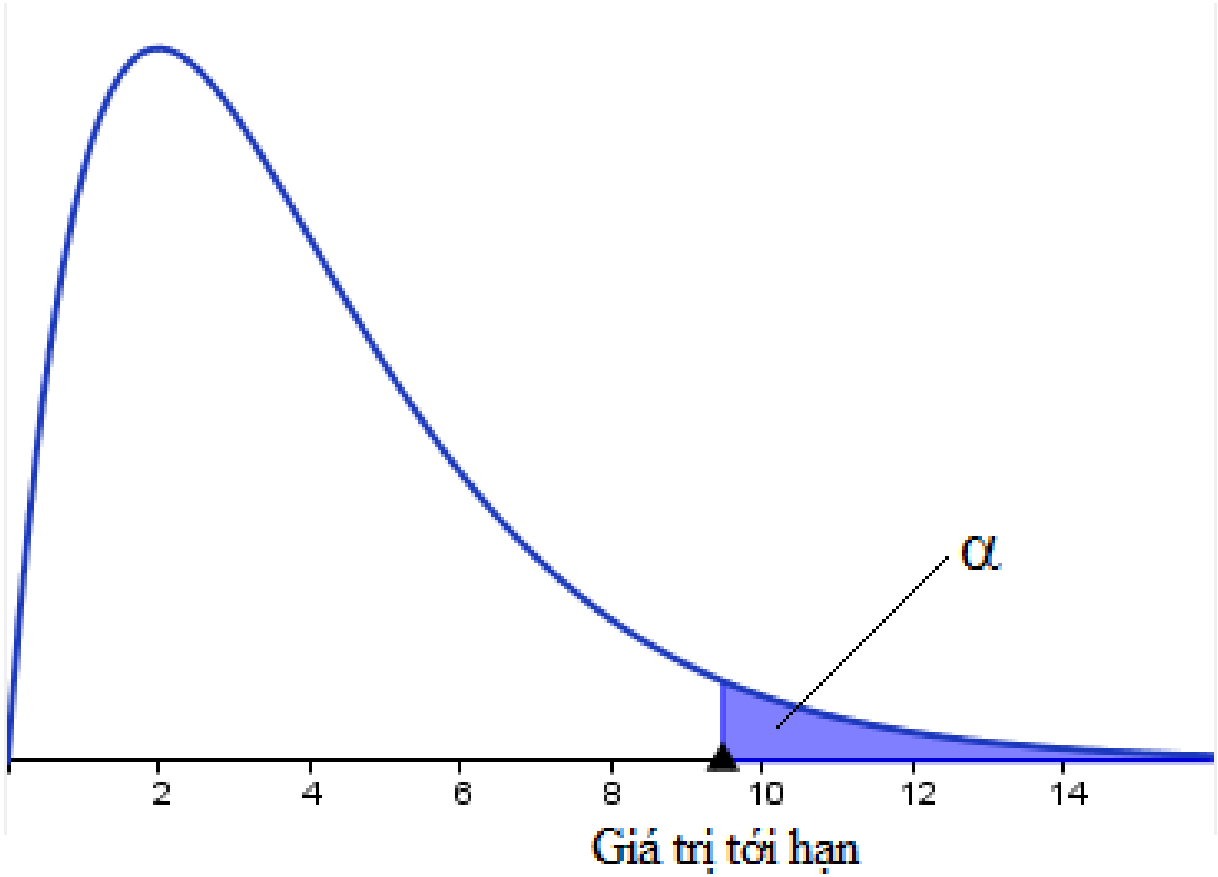
Nếu H_0 đúng và $E_{ij} \geq 5, \forall i, j$ thì đại lượng

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

tuân theo phân phối chi − bình phương với $(r-1)(c-1)$ bậc tự do.



Hình: Đường biểu diễn đồ thị hàm mật độ của phân phối Chi bình phương với các bậc tự do 1, 2, 3, 4, 5



Hình: Giá trị tới hạn $\chi^2_{6,\alpha}$

Quy tắc bác bỏ

Tại mức ý nghĩa α ta đưa ra quyết định bác bỏ H_0 nếu

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

hoặc

$$P-value = P(\chi^2_{(r-1)(c-1)} > \chi^2) = 1-pchisq(\chi^2,(r-1)(c-1)) < \alpha$$

Example

Kết quả sau đây cho mối quan hệ giữa mức độ hài lòng về thu nhập và kết quả công việc đạt được:

	Không hài lòng	Bình thường	Hài lòng	Tổng dòng
Xuất sắc	8	18	30	56
Tốt	10	15	20	45
Trung bình	19	45	35	99
Tổng cột	37	78	85	200

Liệu có mối liên hệ giữa mức độc hoàn thành công việc và sự hài lòng về thu nhập ở mức ý nghĩa 5% hay không?

Solution

H_0: Hai yếu tố mức độ hài lòng về thu nhập và kết quả công việc độc lập nhau.
H_1: Hai yếu tố trên có mối liên hệ với nhau.
Giả sử H_0 đúng, khi đó ta có bảng tần số lí thuyết sau:

	Không hài lòng	*Bình thường*	*Hài lòng*
Xuất sắc	*10.36*	*21.84*	*23.8*
Tốt	*8.325*	*17.55*	*19.125*
Trung bình	*18.315*	*38.610*	*42.075*

Giá trị kiểm định

$$\chi^2 = \frac{(8-10.36)^2}{10.36} + \frac{(21.84-18)^2}{21.84} + \frac{(8.325-10)^2}{8.325} + \frac{(17.55-15)^2}{17.55} + \frac{(19.125-20)^2}{19.125} + \frac{(18.315-19)^2}{18.315} + \frac{(38.61-45)^2}{38.61} + \frac{(42.075-35)^2}{42.075} \approx 5.848303.$$

Solution

Giá trị tới hạn ứng với mức sai lầm 0.05 mà ta chấp nhận khi bác bỏ H_0 *là* $\chi^2_{4,0.05} \approx 9.487729$.

Ta thấy $\chi^2 < 9.487729$ *nên ta chấp nhận* H_0. *Vậy, ở mức ý nghĩa 5% với dữ liệu trên ta cho rằng mức độ hài lòng vào lương bổng độc lập với kết quả công việc.*

Chú ý, *Nếu dùng P - giá trị thì ta có: P - giá trị* $=$
$P(\chi^2_4 > 4.205953) = 1 - pchisq(5.8483, 4) \approx 0.2108 > 0.05$ *nên ta chấp nhận* H_0.

Thực hiện kiểm định tính độc lập trong R

- Để kiểm chứng tính độc lập trong R, ta dùng hàm `chisq.test(A)`, trong đó A là ma trận chỉ bảng gồm các quan sát của hai thuộc tính cần kiểm định tính độc lập.
- Để lập được một ma trận cấp $m \times n$ ta dùng hàm
  `matrix(x, nrow = m, ncol = n, byrow = FALSE, dimnames = NULL)`,
  trong đó
  - `x` là véc tơ chỉ các phần tử của ma trận;
  - `nrow = m` là tham số chỉ số hàng bằng m của ma trận;
  - `ncol = n` là tham số chỉ số cột bằng n của ma trận;
  - `byrow = FALSE (TRUE)` là tham số chỉ việc sắp xếp các phần tử trong véc tơ x theo cột (hàng) trước, mặc định là `FALSE` tức là theo cột trước;
  - `dimnames` là tham số ghi tên cột và hàng của ma trận, mặc định là `NULL`.

Solution

Trong R:

```
> x=c(8,18,30,10,15,20,19,45,35)
> A=matrix(x,nrow = 3,byrow = T)
> A
     [,1] [,2] [,3]
[1,]    8   18   30
[2,]   10   15   20
[3,]   19   45   35
> chisq.test(A)

Pearson's Chi-squared test

data:  A
X-squared = 5.8483, df = 4, p-value = 0.2108
```

Example

Hãy thực hiện kiểm định bài toán ban đầu tại mức ý nghĩa 5%

Example

Từ tập dữ liệu ChiTieu2010.csv, hãy kiểm định xem yếu tố nghèo và khu vực có mối liên hệ với nhau hay không? Sử dụng mức ý nghĩa 5%.

Example

Từ tập dữ liệu ChiTieu2010.csv, hãy kiểm định xem yếu tố nghèo và số người trong hộ phân theo nhóm: ít (≤ 2), bình thường (từ 3 đến 5), nhiều (từ 6 trở lên) có mối liên hệ với nhau hay không? Sử dụng mức ý nghĩa 5%.

Bài toán (Kiểm định nhiều tỉ lệ)

Giả sử một tổng thể có k nhóm, để kiểm định xem tỉ lệ k nhóm này trong tổng thể có tuân theo một dãy tỷ lệ $p_1, p_2, \ldots, p_k$ hay không ta kiểm định cặp giả thuyết:

H_0 : *Số phần tử của tổng thể phân bổ vào k nhóm tuân theo dãy tỷ lệ* $p_1, p_2, \ldots, p_k$;

H_1 : *Số phần tử của tổng thể phân bổ vào k nhóm không tuân theo dãy tỷ lệ* $p_1, p_2, \ldots, p_k$.

Example

Tỉ lệ người trúng xổ số của 3 miền Bắc, Trung, Nam năm trước là $0.2 : 0.3 : 0.5$. Hãy kiểm định xem tỉ lệ này nay nay có thay đổi không?

Example

Một thời báo kinh tế cho rằng tỉ lệ người khách hàng yêu thích mua hàng qua mạng, đến trực tiếp cửa hàng mua, nhờ người mua giúp là 30%, 60%, 10%. Hãy kiểm định xem điều này có đúng không?

Quy trình kiểm định

- Chọn một mẫu ngẫu nhiên gồm n phần tử mà mỗi phần tử được xếp vào đúng một trong k nhóm. Gọi $O_1, O_2, \ldots, O_k$ lần lượt là số phần tử rơi vào k nhóm trên.
- Nếu H_0 đúng thì xác suất để một phần tử rơi vào nhóm $1, 2, \ldots, k$ 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 nhóm đó sẽ là $E_i = np_i, i = 1, 2, \ldots, k$:

Nhóm	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

Quy trình kiểm định

- Nếu H_0 đúng và cỡ mẫu lớn sao cho $E_i = np_i \geq 5, \forall i = \overline{1,k}$ thì đại lượng

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

tuân theo phân phối chi- bình phương với $k-1$ bậc tự do.

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

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

Hoặc

$$P - value = P(\chi^2_{k-1} > \chi^2) < \alpha$$

Example

Theo báo cáo tổng điều tra dân số của hai năm trước đây tại một tỉnh, tỉ lệ những vợ chồng có một con là 15%, có 2 con là 55 %, trên 2 con là 30%.
Sau bốn năm với những chiến dịch tuyên truyền, người ta muốn đánh giá lại hiệu quả của nó. Một cuộc điều tra ngẫu nhiên trên 500 vợ chồng cho thấy có 100 cặp có 1 con, 300 cặp có 2 con mà 100 cặp có trên 2 con. Với dữ liệu đó, có thể kết luận là những chiến dịch tuyên truyền có làm thay đổi tỉ lệ sinh hay không? chọn mức ý nghĩa 5%.

Example

Theo báo cáo tổng điều tra dân số của hai năm trước đây tại một tỉnh, tỉ lệ những vợ chồng có một con là 15%, có 2 con là 55 %, trên 2 con là 30%.
Sau bốn năm với những chiến dịch tuyên truyền, người ta muốn đánh giá lại hiệu quả của nó. Một cuộc điều tra ngẫu nhiên trên 500 vợ chồng cho thấy có 100 cặp có 1 con, 300 cặp có 2 con mà 100 cặp có trên 2 con.
Với dữ liệu đó, có thể kết luận là những chiến dịch tuyên truyền có làm thay đổi tỉ lệ sinh hay không? chọn mức ý nghĩa 5%.

Solution

Ta có cặp giả thuyết như sau:
H_0 : *Hiện tại tỉ lệ số cặp vợ chồng có 1 con là 0.15, có 2 con là 0.55, có trên 2 con là 0.3.*
H_1 : *Tỉ lệ nói trên nay đã khác.*

Solution

Giả sử rằng H_0 đúng, khi đó trong 500 cặp vợ chồng được chọn sẽ có: $0.15 \times 500 = 75$ *cặp có 1 con, 275 cặp có 2 con và 150 cặp có trên 2 con.*

Ta tính được đại lượng

$$\chi^2 = \frac{(100-75)^2}{75} + \frac{(300-275)^2}{275} + \frac{(100-150)^2}{150} = 27.27$$

Trong khi đó nếu H_0 đúng chỉ có 5% các giá trị tính được lớn hơn $\chi^2_{2,0.05} = 5.991$.

Như vậy, với việc chấp nhận sai lầm 5% ta sẽ bác bỏ H_0. Tức là tỉ lệ các hộ sinh 1 con, 2 con và nhiều hơn 2 con đã thay đổi so với trước đây.

1. Kiểm chứng tính độc lập
2. Kiểm định nhiều tỉ lệ
3. **Kiểm định sự phù hợp của một phân phối**
4. Kiểm định chi bình phương với R

Kiểm chứng sự phù hợp của một phân phối

Khi làm một bài toán xác suất hoặc thống kê, ta thường gặp các giả thiết:

- Giả sử quân xúc xắc cân đối và đồng chất;

Kiểm chứng sự phù hợp của một phân phối

Khi làm một bài toán xác suất hoặc thống kê, ta thường gặp các giả thiết:

- Giả sử quân xúc xắc cân đối và đồng chất;
- Giả sử chỉ số IQ của người dân tuân theo phân phối chuẩn với trung bình $\mu = 100$ và phương sai $\sigma = 15$;

Kiểm chứng sự phù hợp của một phân phối

Khi làm một bài toán xác suất hoặc thống kê, ta thường gặp các giả thiết:

- Giả sử quân xúc xắc cân đối và đồng chất;
- Giả sử chỉ số IQ của người dân tuân theo phân phối chuẩn với trung bình $\mu = 100$ và phương sai $\sigma = 15$;
- Giả sử số km mà một chiếc ô tô đi được cho đến khi bỏ đi tuân theo phân phối mũ với tham số $\lambda = \frac{1}{20}$.

Kiểm chứng sự phù hợp của một phân phối

Khi làm một bài toán xác suất hoặc thống kê, ta thường gặp các giả thiết:

- Giả sử quân xúc xắc cân đối và đồng chất;
- Giả sử chỉ số IQ của người dân tuân theo phân phối chuẩn với trung bình $\mu = 100$ và phương sai $\sigma = 15$;
- Giả sử số km mà một chiếc ô tô đi được cho đến khi bỏ đi tuân theo phân phối mũ với tham số $\lambda = \frac{1}{20}$.

Làm thế nào để đưa ra được những giả sử như trên trong các bài toán?

Bài toán (Kiểm định sự phù hợp của một phân phối)

Giả sử ta chưa biết phân phối của một tổng thể. Ta cần kiểm định xem phân phối của tổng thể có tuân theo một phân phối xác suất A cho trước hay không.

Để giải bài toán trên, cặp giả thiết của chúng ta sẽ là:
H_0 : Tổng thể tuân theo phân phối A.
H_1 : Tổng thể không tuân theo phân phối A.

Giải quyết vấn đề

Trước tiên ta chọn mẫu ngẫu nhiên gồm n phần tử, rồi chia vào k nhóm. Giả sử $O_1, O_2, \ldots, O_k$ là số phần tử thuộc vào các nhóm.
Bây giờ, giả sử H_0 là đúng, tức là tổng thể tuân theo phân phối xác suất A. Từ đây ta sẽ tính được là khi đó về mặt lí thuyết thì tỉ lệ phần tử của tổng thể rơi vào k nhóm trên lần lượt là $p_1, p_2, \ldots, p_k$.
Và bây giờ bai toán trở thành kiểm định:
H_0: Tỉ lệ phân bổ các phần tử của tổng thể vào k nhóm theo tỉ lệ: $p_1, p_2, \ldots, p_k$.
H_1: Tỉ lệ phân bổ các phần tử của tổng thể vào k nhóm không tuân theo tỉ lệ: $p_1, p_2, \ldots, p_k$.
Và ta trở về với bài toán so sánh nhiều tỉ lệ.

Example (Ví dụ giả tưởng)

Sau một thời gian nghiên cứu cách đọc và thống kê tự động các thông tin trên internet. Một sinh viên 58TH1 đã lập ra một phần mềm giúp trả lời tự động bài kiểm tra trắc nghiệm của trường. Bạn sinh viên này muốn biết xem liệu phần mềm này có thực sự giúp trả lời các câu hỏi không. Theo đó nếu nó giúp được thì tỉ lệ trả lời đúng của nó ít nhất là phải khác so với việc đánh ngẫu nhiên các câu trả lời. Cho phần mềm này thử trả lời 100 đề mỗi đề có 5 câu hỏi và số câu trả lời đúng mỗi bài được cho dưới đây

Số câu đúng trong một đề	1	2	3	4	5
Tần số	2	23	30	36	9

Hãy kiểm định xem, phân phối số câu đúng có tuân theo phân phối nhị thức B(5,0.25) không? Nếu đúng thì điều này có nghĩa là phần mềm không hề giúp ích gì, vì nó giống trả lời hú họa.
Lựa chọn mức ý nghĩa 5% cho các kết luận.

Solution

Ta có cặp giả thuyết:

H_0 : *số câu trả lời đúng tuân theo B(5,0.25).*

H_1 : *số câu trả lời đúng không tuân theo B(5,0.25)*

Nếu số câu đúng là tuân theo $B(5, 0.25)$ *thì tỉ lệ số câu đúng sẽ là (dùng lệnh dbinom(0:5,5,0.25))*

Số câu đúng	*0*	*1*	*2*	*3*	*4*	*5*
Xác suất xảy ra	*0.2373*	*0.3955*	*0.2637*	*0.0879*	*0.0146*	*0.0010*

Vậy nếu H_0 *đúng thì ta có bảng tần số lí thuyết:*

Số câu đúng trong một đề	*0*	*1*	*2*	*3*	*4*	*5*
Tần số	*24*	*40*	*26*	*9*	*1*	*0*

Solution

Ta phải phân lại nhóm để đảm bảo mỗi ô ít nhất 5. Khi đó bảng tần số lí thuyết có dạng:

Số câu đúng trong một đề	*0*	*1*	*2*	≥ 3
Tần số	*24*	*40*	*26*	*10*

Theo cách chia đó thì thực tế quan sát được tần số số câu trả lời đúng như sau:

Số câu đúng trong một đề	*0*	*1*	*2*	≥ 3
Tần số	*0*	*2*	*23*	*74*

Ta tính được

$$\chi^2 = \frac{(24-0)^2}{24} + \frac{(40-2)^2}{40} + \frac{(26-23)^2}{26} + \frac{(10-74)^2}{10} \approx 470.$$

P - value = pchisq(470,3,lower.tail = FALSE) $= 1.51254 \times 10^{-101} < 0.05$ *nên bác bỏ* H_0.

Vậy, phần mềm này có tác động tới kết quả trả lời.

Example (Ví dụ giả tưởng)

Sau một thời gian nghiên cứu cách đọc và thống kê tự động các thông tin trên internet. Một sinh viên 58HT đã lập ra một phần mềm giúp trả lời tự động bài kiểm tra trắc nghiệm của trường. Bạn sinh viên này muốn biết xem liệu phần mềm này có thực sự giúp trả lời các câu hỏi không. Theo đó nếu nó giúp được thì tỉ lệ trả lời đúng của nó ít nhất là phải khác so với việc đánh ngẫu nhiên các câu trả lời. Cho phần mềm này thử trả lời 100 đề mỗi đề có 5 câu hỏi và số câu trả lời đúng mỗi bài được cho dưới đây

Số câu đúng trong một đề	1	2	3	4	5
Tần số	2	23	30	36	9

Ta thấy rằng, có tất cả 500 câu trong 100 đề trên, và số câu trả lời đúng tổng cộng lên đến 327 câu, tức là chiếm xấp xỉ 65%. Liệu có phải phần mềm này giúp tăng khả năng đúng khi trả lời mỗi câu lên tới 65%? Lựa chọn mức ý nghĩa 5% cho các kết luận.

Solution

Làm tương tự như vậy, ta có cặp giả thuyết:
H_0 : *số câu trả lời đúng tuân theo B(5,0.65).*
H_1 : *số câu trả lời đúng không tuân theo B(5,0.65)*
Nếu số câu đúng là tuân theo $B(5, 0.65)$ *thì tỉ lệ số câu đúng sẽ là (dùng lệnh dbinom(0:5,5,0.65))*

Số câu đúng	*0*	*1*	*2*	*3*	*4*	*5*
Xác suất	*0.0053*	*0.0488*	*0.1811*	*0.3364*	*0.3124*	*0.11603*

Như vậy nếu H_0 *đúng thì ta có bảng tần số lí thuyết:*

Số câu đúng	*0*	*1*	*2*	*3*	*4*	*5*
Tần số	*1*	*5*	*18*	*34*	*31*	*12*

Solution

Tương tự như trên, ta phải dồn 2 cột đầu của bảng xác suất lí thuyết rồi mới nhân 100 vào xác suất để được bảng sau (làm như vậy để tránh tích lũy nhiều sai số):
Nếu H_0 đúng thì ta có

Số câu đúng	≤ 1	2	3	4	5
Tần số	5	18	34	31	12

Thực tế quan sát cho ta ta bảng:

Số câu đúng trong một đề	≤ 1	2	3	4	5
Tần số	2	23	30	36	9

Ta tính được $\chi^2 \approx 5.215929$ *và có p-value* $\approx 0.2658506 > 0.05$ *nên chấp nhận* H_0.
Vậy, tại mức ý nghĩa 5% có thể nói phần mềm trên giúp đạt xác suất trả lời đúng mỗi câu là 65%.

1. Kiểm chứng tính độc lập
2. Kiểm định nhiều tỉ lệ
3. Kiểm định sự phù hợp của một phân phối
4. Kiểm định chi bình phương với R

Kiểm đinh chi bình phương với R

1. Với bài toán kiểm định mối quan hệ độc lập giữa hai biến định tính:
    1. Đầu tiên ta lập bảng tần số chéo, có thể dùng table(biến 1, biến 2).
    2. Lập ma trận A là biểu đồ chéo nói trên. Nếu có dữ liệu sơ cấp, thì đơn giản đặt: A = table(biến 1, biến 2). Kiểm tra điều kiện mọi ô tần số trong bảng tần số lí thuyết đều thỏa mãn ≥ 5.
    3. Kiểm định bởi hàm: chisq.test(A)
2. Với bài toán kiểm định sự phù hợp của một phân phối:
    1. Đầu tiên ta lập véc tơ xác suất lí thuyết, XS. Nhân với cỡ mẫu để kiểm tra điều kiện mọi phần tử đều ≥ 5. Nếu không thỏa mãn, có thể dồn lại một số cột, và khi đó sẽ tạo véc tơ XS mới theo cách dồn đó.
    2. Lập véc tơ tần số ứng với véc tơ XS trên.
    3. Kiểm định bằng hàm: chisq.test(TanSo,p=XS)

Example

Một công ty muốn đánh giá xem hiệu quả của chiến lược quảng cáo đến thị phần của mình. Trước khi thực hiện chiến lược quảng cáo thị phần của công ty này là 46 %, của công ty đối thủ chính là 38%, phần còn lại thuộc về các đối thủ khác. Sau khi thực hiện chiến dịch quảng cáo người ta lấy một mẫu 200 khách hàng ngẫu nhiên có dùng mặt hàng được quảng cáo cho thấy 100 người thích dùng sản phẩm của công ty này, 80 người cho rằng họ thích sản phẩm của đối thủ cạnh tranh nói trên, còn lại dùng sản phẩm của các nhà sản xuất khác.
Tại mức ý nghĩa 5%, thị phần về mặt hàng nói trên có thay đổi so với trước khi chiến dịch quảng cáo được thực hiện không?