Bài 10.

KIỂM ĐỊNH GIẢ THUYẾT VỀ PHÂN PHỐI VÀ SỰ ĐỘC LẬP

Nội dung:

Xét biến ngẫu nhiên *X,* với mức ý nghĩa α ta cần kiểm định

Giả thuyết H_0: *X* có phân phối $F(x)$

Đối thuyết H_1: *X* không có phân phối $F(x)$

Phương pháp kiểm định: Chi – bình phương.

I. Kiểm định phân phối rời rạc:

Tóm tắt lý thuyết: xét biến ngẫu nhiên rời rạc *X* với mẫu quan trắc được trình bày như sau

X	x_1	x_2	...	x_{k-1}	x_k
Tần số	n_1	n_2	...	n_{k-1}	n_k

Cỡ mẫu $N = \sum_{i=1}^{k} n_i$.

1. Đặt giả thuyết H_0: *X* có phân phối $F(x)$
2. Tính giá trị Chi – bình phương thực nghiệm:

$$Q^2 = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$$

với

- n_i: tần số thực nghiệm.
- n'_i: tần số lý thuyết. $n'_i = Np_i$, với $p_i = P(X = x_i)$ là xác suất xảy ra biến cố $(X = x_i)$ được với giả sử H_0 đúng.

3. Xác định mức ý nghĩa α và bậc tự do $df = k - r - 1$, trong đó r là số tham số ước lượng.
4. Tính giá trị $\chi^2_{\alpha,df}$.
5. Bác bỏ H_0 nếu: $Q^2 > \chi^2_{\alpha,df} \Rightarrow$ Kết luận.

Thực hành trong R:

Cách 1: tính toán từng bước như trên sử dụng các hàm trong **R** để kiểm định.

Cách 2: kiểm định trực tiếp dùng các hàm có sẵn

Sử dụng hàm `chisq.test(n, p = prob)`, với

- $n = (n_1, n_2, \dots, n_k)$: véc-tơ chứa các giá trị tần số thực nghiệm.

- $prob = (p_1, p_2, \ldots, p_k)$: với $p_i = P(X = x_i), i = 1, \ldots, k$ là xác suất biến ngẫu nhiên X nhận giá trị x_i tương ứng với giả sử H_0 đúng.

Ngoài ra, có thể sử dụng hàm `goodfit`: (kiểm định cho phân phối Poisson, nhị thức)

```
goodfit(n, type = "poisson", method ="MinChisq", par=NULL)
```

Nếu kiểm định phân phối nhị thức thì thay `type = "binomial"`
`par`: một danh sách chỉ ra số tham số của phân phối tương ứng. Nếu là phân phối poisson, tham số là lambda=?, nếu là nhị thức là prob =? Mặc định sẽ ước lượng tham số từ mẫu.

Chú ý: để sử dụng hàm `goodfit`, phải cài đặt gói (package) **vcd** trước. Cách cài đặt:

- Cài trực tiếp qua internet, chọn server cài đặt, rồi chọn tên gói.
- Download file nén chứa gói vcd về, giải nén vào thư mục "C:\Program Files\R\R-2.6.0\library" (Tên đường dẫn thay đổi tùy theo nơi cài đặt trên máy, tổng quát là chép gói vcd vào thư mục library của R).

Ví dụ: Tại một trang trại, người nông dân ghi lại số lượng bê cái được sinh ra trong lần sinh sản đầu tiên của từng con bò cái trong trang trại của ông ta, kết quả cho bởi bảng sau:

Số bê cái	0	1	2	3	4	5
Số bò cái	4	19	41	52	26	8

Với mức ý nghĩa 5%, gọi X là số lượng bê cái sinh ra tương ứng với mỗi lần sinh của một con bò cái. Hãy xét xem X có tuân theo luật phân phối nhị thức $B(5,p)$ hay không?

```
x <- 0:5;x
#e: vecto chua cac tan so thuc nghiem
e <- c(4,19,41,52,26,8);e
#Gia thuyet: X co phan phoi nhi thuc, X ~ B(5,p)
#Tinh cac xac suat pi = P(X = xi) theo pp nhi thuc
#Tinh gia tri p, EX = np => p = EX/n, EX uoc luong bang trung binh
mau
x.mean <- sum(e*x)/sum(e);x.mean
p = x.mean/5;p
#Tinh vecto xac suat prob
prob = dbinom(0:5,5,p)
#Kiem dinh dung ham chisq.test(n =e, p = prob)
test = chisq.test(e, p = prob)
test
```

Bài tập:

1. Bảng sau thống kê số vụ tai nạn xe máy / ngày ở quận 5 trong 80 ngày:

Số vụ tai nạn	Số ngày
0	34
1	25
2	11
3	7
4	3

Với mức ý nghĩa 5% và dùng phương pháp Chi – bình phương, hãy kiểm tra xem số vụ tai nạn xe máy hàng ngày có tuân theo luật phân phối poisson hay không?

3. Trong một nhà máy sản xuất ôtô chỉ có một dây chuyền sản xuất, nếu dây chuyền bị hư thì nhà máy phải tạm ngưng đến khi dây chuyền được sửa xong. Gọi X là số lần tạm ngưng trong một ngày, ta có bảng thống kê sau trong 1400 ngày:

Số lần tạm ngưng	**Số ngày**
0	728
1	447
2	138
3	48
4	26
5	13
>= 6	0

Với mức ý nghĩa 5% và dùng phương pháp Chi – bình phương, hãy kiểm tra xem mô hình trên có tuân theo luật phân phối Poisson hay không?

3. Ở khâu kiểm tra sản phẩm của một nhà máy sản xuất bóng đèn. Người ta kiểm tra ngẫu nhiên 30 lô hàng. Mỗi lô hàng người ta lấy ra 5 bóng để kiểm tra. Đối với mỗi bóng đèn có hai khả năng có thể xảy ra: sáng hoặc không sáng. Ta có bảng kết quả sau:

Số bóng đèn sáng	0	1	2	3	4	5	Tổng
Số lô hàng (n_i)	1	2	4	11	9	3	30

Với mức ý nghĩa α= 5% và dùng phương pháp Chi - bình phương, hãy kiểm tra xem mô hình dữ liệu cung cấp ở trên có tuân theo luật phân phối nhị thức hay không?

II. Kiểm định phân phối liên tục:

Đối với biến ngẫu nhiên có phân phối liên tục, để kiểm định, biến đổi mẫu khảo sát về bảng tần số dạng khoảng:

X	$(a_1, a_2]$	$(a_2, a_3]$	...	$(a_{k-1}, a_k]$	$(a_k, a_{k+1}]$
Tần số	n_1	n_2	...	n_{k-1}	n_k

Các bước kiểm định tương tự như trong trường hợp rời rạc:

Phát biểu giả thuyết H_0: X có phân phối $F(x)$.

Tính giá trị Chi – bình phương thực nghiệm:

$$Q^2 = \sum_{i=1}^{k} \frac{\left(n_i - n'_i\right)^2}{n'_i}$$

Với $n'_i = Np_i$ và $p_i = P\left(a_i \le X < a_{i+1}\right) = F\left(a_{i+1}\right) - F\left(a_i\right)$; trong đó $F\left(x\right)$ là phân phối được phát biểu trong giả thuyết H_0.

Bác bỏ H_0 khi: $Q^2 > \chi^2_{\alpha,df}$ với $df = k - r - 1$ (r: số tham số ước lượng).

<u>Thực hành trong R:</u>

Biến đổi số liệu dạng véc-tơ về bảng tần số dạng khoảng, dùng hàm cut, table.
Sử dụng các hàm tính xác suất tích luỹ của phân phối liên tục (`pnorm`, `pexp`, …) để tính các p_i.
Cách 1: tính trực tiếp giá trị Chi – bình phương thực nghiệm Q^2 và so sánh với $\chi^2_{\alpha,df}$ (dùng hàm `qchisq(alpha,df, lower.tail = FALSE)`)
Cách 2: dùng hàm `chisq.test(n, p = prob)` với $prob = \left(p_1, p_2, \ldots, p_k\right)$.
<u>Ví dụ:</u> Tải biến X là chiều cao những thanh niên từ 18 tuổi trở lên trong một khu vực từ tập tin **height6.rda**, dùng lệnh load('height6.rda'); hãy kiểm tra xem X có phân phối chuẩn với alpha = 0.05.

```
#Chuyen du lieu X ve dang bang tan so theo khoang
#Co the can cu theo bieu do histogram de chia khoang
#Hoac xac dinh so khoang chia bang cach tim range = max(X) - min(X)
hist(X); max(X); min(X)
X.range = max(X) - min(X)
#Dua theo histogram co the chia X thanh 6 doan
#[130,140), [140,150), ... , [180,190), [190,200)
tab = table(cut(X,breaks = seq(130,200,by=10)))
```

```
#Co mau
N = tab[]
a = seq(130,190,by=10)
b = seq(140,200,by=10)
#Gia su chua biet mu va sigma
#Uoc luong mu va sigma tu mau
mu = mean(X)
sigma = sd(X)
#Tinh vec-to xac suat prob
k = length(N)
prob = c(pnorm(b[1],mu,sigma),pnorm(b[2:(k-1)],mu,sigma) -
   pnorm(a[2:(k-1)],mu,sigma),1-pnorm(a[k],mu,sigma))
sum(prob)
#Kiem dinh dung ham chisq.test
test = chisq.test(N, p = prob)
        Chi-squared test for given probabilities
data:  N
X-squared = 0.6581, df = 6, p-value = 0.9954
```

Bài tập:

4. Bảng sau thống kê chiều cao (Đv: m) của 125 thanh niên 18 tuổi trong một khu vực:

Chiều cao	[1.2,1.4)	[1.4,1.6)	[1.6,1.8)	[1.8,2.0)	[2.0,2.2)
Số thanh niên	6	34	31	42	12

Với mức ý nghĩa 1%, hãy kiểm tra xem chiều cao của các thanh niên trong khu vực này có tuân theo luật phân phối chuẩn hay không?

5. Thời gian sống hay còn gọi là tuổi thọ (Đv: giờ) của 300 linh kiện điện tử trong một hệ thống máy tính được cho bởi bảng thống kê sau:

Tuổi	[0,50)	[50,100)	[100,150)	[150,200)	[200,300)	[300,400)	[400,500)	[500,+∞)
Số linh kiện	63	47	55	34	29	27	24	21

Biết rằng tuổi thọ trung bình của các linh kiện này là 200 giờ. Với mức ý nghĩa 5%, hãy kiểm tra xem tuổi thọ của các linh kiện có tuân theo phân phối mũ hay không?

6. Số liệu trong data61.xls cho biết mức lương trên 1 năm (Đv: 1000 USD) của 44 nhân viên công ty ANZ. Với mức ý nghĩa 1%, hãy kiểm tra xem mức lương trên có tuân theo luật phân phối chuẩn hay không?

III. Kiểm định với phân phối xác suất cho trước:

Dùng hàm:

```
chisq.test(n, p = prob)
```

với

$n = (n_1, n_2, \ldots, n_k)$: véc-tơ tần số thực nghiệm

$prob = (p_1, p_2, \ldots, p_k)$: véc-tơ chứa các xác suất cho trước.

Bài tập:

7. Năm 1986 tỷ lệ bác sĩ theo các chuyên môn như sau:

Chuyên môn	Tổng quát	Nội khoa	Giải phẫu	Còn lại
Tỷ lệ	0,180	0,339	0,270	0,211

Năm 1989 thống kê 500 bác sĩ có số liệu sau:

Chuyên môn	Tổng quát	Nội khoa	Giải phẫu	Còn lại
Tần số	80	162	156	102

Hỏi tỷ lệ chuyên môn hai năm nói trên có thay đổi không (mức ý nghĩa 5%)?

8. Có một lô hàng rất nhiều mà người chào hàng cho biết : tỷ lệ hỏng là 10%, đạt là 60%, tốt là 30%. Người ta kiểm tra một số sản phẩm thấy có 30 sản phẩm hỏng, 80 sản phẩm đạt, 40 sản phẩm tốt. Hỏi người chào hàng nói có đúng không ? (mức ý nghĩa 1%)

IV. Kiểm định sự độc lập:

Xét hai thuộc tính *A* và *B*;

- A có r mức: $a_1, a_2, \ldots, a_r$
- B có c mức: $b_1, b_2, \ldots, b_r$

Bảng số liệu khảo sát cho hai thuộc tính A và B:

Hàng						
Cột		$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	$\cdots$	$\boldsymbol{b_c}$	Tổng hàng
	$\boldsymbol{a_1}$	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\bullet}$
	$\boldsymbol{a_2}$	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$\boldsymbol{a_r}$	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\bullet}$
	Tổng cột	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet c}$	N

Cần kiểm định giả thuyết:

H_0: *A* độc lập với *B*

(Đối thuyết: *A* không độc lập với *B*)

Phương pháp Chi – bình phương:

- Tính giá trị Chi – bình phương thực nghiệm

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

Trong đó:

n_{ij} : tần số thực nghiệm

n'_{ij} : tần số lý thuyết; $n'_{ij} = \dfrac{n_{i\bullet} \times n_{\bullet j}}{N} = \dfrac{\text{Tổng hàng} \times \text{Tổng cột}}{\text{Tổng tất cả}}$

- Bác bỏ H_0 nếu: $Q^2 > \chi^2_{\alpha,df}$ với $df = (r-1)(c-1)$ ⇒ Kết luận.

(Trong R, tìm $\chi^2_{\alpha,df}$ dùng hàm `qchisq(alpha, df, lower.tail = FALSE)`)

Thực hành trong R:

Cách 1: tính toán từng bước như để sử dụng các hàm trong **R** để kiểm định.

Cách 2: sử dụng hàm

`chisq.test(M)`

Trong đó: M là ma trận chứa các tần số thực nghiệm

$$M = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1c} \\ n_{21} & n_{22} & \cdots & n_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1} & n_{r2} & \cdots & n_{rc} \end{bmatrix}$$

Đọc $p-$ giá trị xuất từ hàm và kết luận.

Ví dụ:

Một báo cáo khoa học trong y khoa tuyên bố rằng việc sở hữu một thú cưng trong nhà (chó hoặc mèo) sẽ làm tăng khả năng sống sót của những người chủ mà thường bị lên cơn đau tim. Một mẫu ngẫu nhiên gồm 95 người đã lên cơn đau tim được chọn để khảo sát. Dữ liệu của mỗi người khảo sát được chia làm 2 loại:

- Những người sống sót/tử vong 1 năm sau khi lên cơn đau tim.
- Người sống sót/tử vong có nuôi thú cưng trong nhà hay không.

Bảng kết quả:

	Có nuôi thú cưng	Không nuôi thú cưng
Sống sót	28	44
Tử vong	8	15

Với mức ý nghĩa 5%, liệu dữ liệu trên có ủng hộ kết luận rằng sự sống sót (khi bị lên cơn đau tim) và nuôi thú cưng là độc lập?
Đặt giả thuyết H_0: Sự sống sót khi lên cơn đau tim và nuôi thú cưng là độc lập với nhau.

```
> M = matrix(c(28,44,8,15),nrow=2,byrow=T)
> M
     [,1] [,2]
[1,]   28   44
[2,]    8   15
> chisq.test(M)
        Pearson's Chi-squared test with Yates' continuity
correction
data:  M
X-squared = 0.0114, df = 1, p-value = 0.9152
```

P giá trị = 0.9152 > 0.05 ⇒ kết luận: không bác bỏ H_0. Tức là việc nuôi thú cưng không ảnh hưởng đến khả năng sống sót khi bị lên cơn đau tim.

Bài tập:

9. Kiểm tra chất lượng sản phẩm do 3 nhà máy sản xuất hàng xuất khẩu ta ghi nhận được bảng số liệu sau

	Chất lượng			
Nhà máy	Tốt	Đạt	Phải sửa	Thứ phẩm
Nhà máy A	40	125	18	17
Nhà máy B	29	91	14	16
Nhà máy C	31	84	18	17

Với mức ý nghĩa 1%, hãy xét xem chất lượng sản phẩm của 3 nhà máy có như nhau?

10. Vé máy bay của hãng hàng không Việt Nam Airline được chia làm 3 loại: Hạng thường (C), hạng trung (B) và hạng doanh nhân (A). Hành khách đi máy bay của VN Airlines nằm trong 1 trong 2 dạng sau: bay nội địa hoặc quốc tế. Khảo sát 920 hành khách đã bay của hãng, cho kết quả sau:

	Loại chuyến bay	
Loại vé	Nội địa	Quốc tế
Hạng thường	29	22
Hạng trung	95	121
Hạng doanh nhân	518	135

Có ý kiến cho rằng hành khách mua loại vé nào (A, B, C) sẽ phụ thuộc vào việc người đó bay nội địa hay quốc tế. Với mức ý nghĩa 5%, hãy kiểm tra ý kiến trên.

11. Công ty B được đặt hàng sản xuất một loại yên xe đạp phù hợp cho cả người béo lẫn người gầy. Họ đưa ra 4 mẫu yên khác nhau, và muốn thử chất lượng của cả 4 mẫu. Họ cho 7 người béo gầy khác nhau ngồi thử lên xe trong 5 phút (thứ tự người thử là ngẫu nhiên). Những người ngồi thử sau đó sẽ cho điểm đánh giá mức độ thoải mái của mẫu yên xe, với 1 điểm là tệ nhất và 10 điểm là thoải mái nhất. kết quả được cho trong bảng bên dưới. Từ kết quả này và với mức ý nghĩa 5%, ta có thể bác bỏ giả thuyết gốc là 4 mẫu yên xe đều thoải mái như nhau không? Hãy lập thống kê đầy đủ của kiểm định chi bình phương dẫn tới kết luận này.

Seat 1	Seat 2	Seat 3	Seat 4
3	**9**	**5**	**7**
7	**8**	**6**	**8**
8	**10**	**9**	**7**
4	**8**	**3**	**6**
6	**8**	**5**	**10**
9	**5**	**7**	**6**
8	**10**	**7**	**9**

12. Thực hiện 1 cuộc khảo sát về ảnh hưởng về việc có người yêu và kết quả học tập trong sinh viên. Kết quả khảo sát trên 50 sinh viên của trường ĐH KHTN với 2 câu hỏi là:

1 - Bạn đã có người yêu chưa?

Có 3 mức độ: 0 = "Chưa có"; 1 = "Đã có"; 2 = "Đang tìm hiểu" (Biến "Nguoi_yeu")

2 - Kết quả học tập của bạn?

Có 3 mức độ: 0 = "Giỏi"; 1 = "Khá"; 2 = "Trung bình" (Biến "Hoc_tap")

Số liệu cho bởi file data61.txt.

a. Chuyển các giá trị 0,1,2 trong biến Hoc_tap thành các mô tả tương ứng, Hoc_tap = (G, K, TB)

b. Với mức ý nghĩa 5%, hãy kiểm định xem việc có người yêu có ảnh hưởng đến kết quả học tập hay không?

13. Một công ty cần phát một đoạn quảng cáo trên đài phát thanh. Trước khi thuê người đọc đoạn quảng cáo này, công ty kiểm tra bốn người với 4 giọng đọc khác nhau gọi là A, B, C, D. 10 người ngẫu nhiên được chọn để nghe đoạn quảng cáo này và đánh giá chất lượng của bốn giọng đọc này. Chất lượng giọng được xếp loại từ tốt nhất (1) đến dở nhất (4). Số liệu cho bởi bảng sau:

Giọng đọc			
A	B	C	D
1	3	2	4
3	4	2	1
4	2	1	3
3	1	2	4
1	2	4	3
3	1	2	4
3	2	4	1
2	3	4	1
3	4	1	2
4	2	3	1

Có ý kiến cho rằng chất lượng của 4 giọng đọc này là như nhau, hãy kiểm tra ý kiến trên. Mức ý nghĩa 5%.