Bài giảng 12: So sánh nhiều nhóm bằng kiểm định Ki bình phương

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Nội dung

- Vấn đề thực tế - so sánh nhiều nhóm
- Khái niệm "độc lập"
- Giới thiệu kiểm định Ki bình phương
- Thao tác R

So sánh nhiều nhóm

Trường hợp 1: Bệnh nhân nhập viện

Số bệnh nhân ung thư nhập viện mỗi tháng

	1	2	3	4	5	6	7	8	9	10	11	12
Số ca bệnh	40	34	30	44	39	58	51	55	36	48	33	38

Câu hỏi: phân bố ngẫu nhiên, không có khác biệt giữa các tháng?

Trường hợp 2: tình trạng kinh tế

- Bill Clinton đắc cử tổng thống 1996
- Lí do đắc cử: do kinh tế khá
- Nghiên cứu trên 800 người

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn
Trung học (n=430)	91	104	235
Cao đẳng (n=160)	39	73	48
Đại học (n=210)	18	31	161

Tử vong trong tai nạn tàu Titanic

Hạng	Chết	Sống	Tổng số
I	123	200 (62%)	323
II	158	119 (43%)	277
III	528	181 (26%)	709
Total	809	500 (38%)	1309

http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic3info.txt

Có mối liên quan giữa hạng hành khách và nguy cơ tử vong?

Khái niệm độc lập

Independence – độc lập

- Hai biến độc lập khi hoàn toàn không có liên quan với nhau
- Hệ số tương quan (coefficient of correlation) = 0
- Nếu A và B độc lập thì:

$$P(A \& B) = P(A) \times P(B)$$

Triết lí và mục đích của Chi square

- Khai thác khái niệm độc lập
- Kiểm định sự độc lập giữa hai biến
- Nếu hai biến *không* độc lập => có liên quan (association)

Kiểm định ý nghĩa thống kê (test of significance)

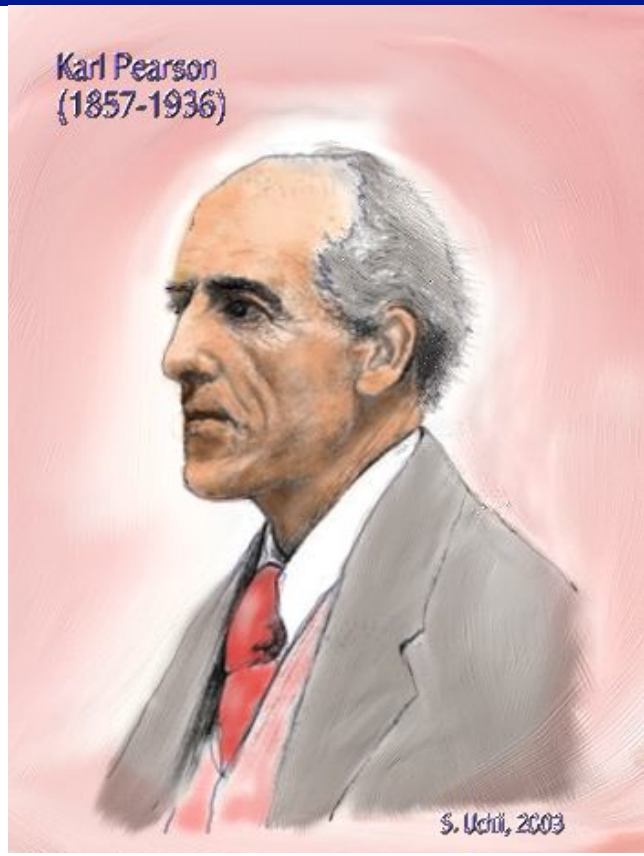
- Triết lí phản nghiệm (falsificationism) của Popper
- Bước 1: phát biểu giả thuyết vô hiệu (null hypothesis)
- Bước 2: thu thập dữ liệu (D)
- Bước 3: tính xác suất D xảy ra nếu giả thuyết vô hiệu đúng

Kiểm định ý nghĩa thống kê (test of significance)

- Bước 1: biến A và B độc lập (không có mối liên quan giữa trình độ học vấn và kinh tế)
- Bước 2: thu thập dữ liệu (D) liên quan đến A và B
- Bước 3: tính xác suất D xảy ra nếu A và B độc lập

Karl Pearson

- Học trò của Francis Galton
- Một trong những "cha đẻ" của mathematical statistics
- Sáng lập bộ môn thống kê học ở University College London (1911)
- Tác giả cuốn *The Grammar of Science*
- Cha đẻ của "Chi square test" (và nhiều phương pháp khác)



Logic của Chi square test

- Nếu hai biến độc lập: ước tính giá trị kì vọng (e**xpected values - E)**
- So sánh giá trị kì vọng với giá trị quan sát (**observed data – O**)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- **Nếu X^2 lớn, bác bỏ giả thuyết vô hiệu**

Kiểm định Ki bình phương

Nghiên cứu về nhập viện

- Chúng ta có 506 bệnh nhân
- Nếu không có khác biệt giữa các tháng, chúng ta kì vọng mỗi tháng có 506 / 12 = 42 ca

	1	2	3	4	5	6	7	8	9	10	11	12
O	40	34	30	44	39	58	51	55	36	48	33	38
E	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**42**

Giá trị kì vọng và quan sát (1)

	1	2	3	4	5	6	7	8	9	10	11	12
O	40	34	30	44	39	58	51	55	36	48	33	38
E	42	42	42	42	42	42	42	42	42	42	42	42
D=O-E	-2	-8	-12	2	-3	16	9	13	-6	6	-9	-4

Giá trị kì vọng và quan sát (2)

	1	2	3	4	5	6	7	8	9	10	11	12
O	40	34	30	44	39	58	51	55	36	48	33	38
E	42	42	42	42	42	42	42	42	42	42	42	42
D=O-E	-2	-8	-12	2	-3	16	9	13	-6	6	-9	-4
D^2	4	64	144	4	9	256	81	169	36	36	81	16
D^2/E	**.11**	**1.58**	**3.51**	**.08**	**.24**	**5.95**	**1.85**	**3.91**	**.90**	**.81**	**2.99**	**.41**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$X^2 = 0.11 + 1.58 + 3.51 + \ldots + 0.41 = 21.3$

Khái niệm degree of freedom (bậc tự do)

- Chính xác là **"degree of freedom for error"**
- Đo lượng huyết áp của 100 bệnh nhân
- Chúng ta có thể ước tính tham số của biến số (mean, median, v.v.)
- Mỗi thông số được ước tính phải "tốn" mất 1 bậc tự do; còn lại ***n – 1*** tự do (degrees of freedom – df)

Bậc tự do (nghiên cứu nhập viện)

- Có 12 số liệu (cho 12 tháng)
- "Mất" 1 thông số để ước tính số trung bình
- Còn lại 11 bậc tự do

Bậc tự do (nghiên cứu nhập viện)

- Còn lại 11 bậc tự do
- $X^2 = 21.3$ phải so sánh với df = 11

- Câu hỏi: **xác suất mà $X^2 = 21.3$ (hay cao hơn) nếu giả thuyết độc lập đúng là bao nhiêu?**

```
1-pchisq(21.3, 11)
```

Tóm lược

- Kiểm định Ki bình thương dựa vào khái niệm "độc lập"
- Tính giá trị kì vọng (E) từ giả thuyết độc lập
- So sánh E với giá trị thực tế: $X^2 = (O – E)^2 / E$
- Tính xác suất X^2 (theo bậc tự do) nếu giả thuyết độc lập là đúng

Khái niệm độc lập

Trường hợp 2: tình trạng kinh tế

- Bill Clinton đắc cử tổng thống 1996
- Lí do đắc cử: do kinh tế?
- Nghiên cứu trên 800 người

Trình độ học vấn	**Tệ hơn**	**Không khác**	**Tốt hơn**
Trung học (n=430)	91	104	235
Cao đẳng (n=160)	39	73	48
Đại học (n=210)	18	31	161
Tổng số	**148**	**208**	**444**

Tử vong trong tai nạn tàu Titanic

Hạng	Chết	Sống	Tổng số
I	123	200 (62%)	323
II	158	119 (43%)	277
III	528	181 (26%)	709
Total	809	500 (38%)	1309

http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic3info.txt

Có mối liên quan giữa hạng hành khách và nguy cơ tử vong?

Independence – độc lập

- Hai biến độc lập khi hoàn toàn không có liên quan với nhau
- Hệ số tương quan (coefficient of correlation) = 0
- Nếu A và B độc lập thì:

$$P(A \& B) = P(A) \times P(B)$$

Giá trị kì vọng: xác suất trình độ học vấn

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn	Xác suất
Trung học (n=430)				0.537
Cao đẳng (n=160)				0.200
Đại học (n=210)				0.263

430 / 800 = 0.537
160 / 800 = 0.200
210 / 800 = 0.263

Giá trị kì vọng: xác suất tình trạng kinh tế

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn	Xác suất
Trung học (n=430)				0.537
Cao đẳng (n=160)				0.200
Đại học (n=210)				0.263
Tổng số	148	208	444	
Xác suất	**0.185**	**0.260**	**0.555**	**1.000**

Giá trị kì vọng *nếu độc lập*

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn	Xác suất
Trung học	0.537*0.185	0.537*0.260	0.537*0.555	**0.537**
Cao đẳng	0.200*0.185	0.200*0.260	0.200*0.555	**0.200**
Đại học	0.263*0.185	0.263*0.260	0.263*0.555	**0.263**
Xác suất	**0.185**	**0.260**	**0.555**	**1.000**

x 800

Giá trị kì vọng

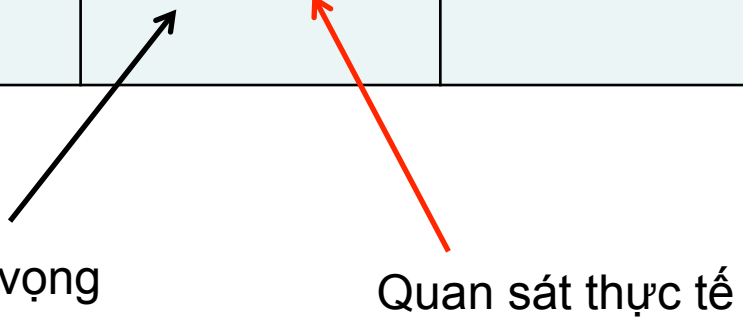
Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn	Xác suất
Trung học (n=430)	79	112	238	**0.537**
Cao đẳng (n=160)	30	42	89	**0.200**
Đại học (n=210)	39	55	117	**0.263**
Xác suất	**0.185**	**0.260**	**0.555**	**1.000**

0.537*0.185*800 = 79

0.537*0.260*800 = 112

Giá trị kì vọng và quan sát

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn
Trung học (n=430)	79 (**91**)	112 (**104**)	238 (**235**)
Cao đẳng (n=160)	30 (**39**)	42 (**73**)	89 (**48**)
Đại học (n=210)	39 (**18**)	55 (**31**)	117 (**161**)



So sánh giá trị kì vọng và quan sát

- E = giá trị kì vọng (expected value)
- O = giá trị quan sát (observed value)

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

Giá trị kì vọng và quan sát (2)

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn
Trung học (n=430)	79 (**91**)	112 (**104**)	238 (**235**)
Cao đẳng (n=160)	30 (**39**)	42 (**73**)	89 (**48**)
Đại học (n=210)	39 (**18**)	55 (**31**)	117 (**161**)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$X^2 = (79-91)^2/91 + (112-104)^2/112 + \ldots + (117-161)^2/117 = 86.0$

Phân tích với R

Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn
Trung học (n=430)	91	104	235
Cao đẳng (n=160)	39	73	48
Đại học (n=210)	18	31	161

```
# nhập dữ liệu

dat = matrix(c(91, 104, 235, 39, 73, 48, 18,
31, 161), nrow=3, byrow=T)

# dùng hàm chisq.test

chisq.test(dat)
```

Phân tích với R

```
> chisq.test(dat)


Pearson's Chi-squared test

data:  dat

X-squared = 86.023, df = 4, p-value < 2.2e-16
```

Phân tích dữ liệu Titanic bằng epitools

Hạng	Chết	Sống	Tổng số
I	123	200 (62%)	323
II	158	119 (43%)	277
III	528	181 (26%)	709
Total	809	500 (38%)	1309

```
library(epitools)
data = matrix(c(123, 200, 158, 119, 528, 181), byrow=T,ncol=2)
riskratio(data)
```

```
> riskratio(data)
          Outcome
Predictor  Disease1 Disease2 Total
  Exposed1      123      200   323
  Exposed2      158      119   277
  Exposed3      528      181   709
  Total         809      500  1309

$measure
          risk ratio with 95% C.I.
Predictor   estimate     lower     upper
  Exposed1 1.0000000        NA        NA
  Exposed2 0.6938087 0.5909899 0.8145156
  Exposed3 0.4122920 0.3541358 0.4799986

$p.value
          two-sided
Predictor    midp.exact fisher.exact  chi.square
  Exposed1           NA           NA          NA
  Exposed2 3.564252e-06 4.126370e-06 3.48938e-06
  Exposed3 0.000000e+00 8.842906e-29 2.80403e-29
```

Hạng	Tỉ số nguy cơ	Trị số P
Phổ thông	1.00	
Thương gia	0.69 (0.59 – 0.81)	<0.001
Hạng nhất	0.41 (0.35 – 0.48)	<0.001

Tóm lược

- Tính giá trị kì vọng (E) từ giả thuyết độc lập
- So sánh E với giá trị thực tế: $X^2 = (O - E)^2 / E$
- Hàm **`chisq.test(data)`**
- Có thể dùng **`epitools`**