

# Phân tích mô tả biến định tính (phân nhóm)

**Tuan V. Nguyen**

Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam



# Biến định danh (nominal variable)

- Sống / chết
- Nam / Nữ
- Nhóm máu: O, A, B, AB
- Cư dân ở Hà Nội, Sài Gòn, Đà Nẵng, Kiên Giang

# Ordinal scale: có thứ tự, nhưng không có khoảng cách

- Độ hài lòng:
  - Rất không hài lòng, không hài lòng, không có ý kiến, hài lòng, rất hài lòng
- Trình độ học vấn: tiểu học, trung học, đại học

# Đếm số ca

- Đơn giản nhất trong đo lường: ĐẾM
- Số đếm là “nền tảng” để ước tính nguy cơ (risk), tỉ lệ, tỉ số, v.v.

# Tỉ lệ

- Proportion, percent
- Tỷ số là một phần của mẫu số
- Thường mô tả bằng phần trăm
- Ví dụ: nghiên cứu cắt ngang trên 2392 đối tượng, trong số đó có 111 người nghèo
- $Prelavelce = 111 / 2392 = 0.046$  (hay 4.6%)

# KTC95% của một tỉ lệ (prevalence)

- $X = 111$  người mắc bệnh,  $N = 2392$  người
- $P = 111 / 2392 = 0.046$
- Standard deviation (độ lệch chuẩn):

$$s = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.046 \times 0.954}{2392}} = 0.0043$$

- KTC95% :  $p \pm 1.96s = 0.046 \pm 1.96 \times 0.0043 = 0.038$  to  $0.055$   
**(3.8 đến 5.5%)**

# KTC95% của một tỉ lệ: chính xác

```
library(rms)  
binconf(x=111, n=2392, method="all")  
  
> binconf(x=111, n=2392, method="all")  
          PointEst Lower Upper  
Exact            0.046 0.038 0.056  
Wilson          0.046 0.039 0.056  
Asymptotic      0.046 0.038 0.055
```

# **So sánh hai tỉ lệ**

# Tình huống chung

- Nghiên cứu có 2 nhóm, với hai tỉ lệ  $p_1$  và  $p_2$  (và cỡ mẫu  $n_1$  và  $n_2$ )
- Chúng ta muốn biết hai tỉ lệ thật sự khác nhau có ý nghĩa thống kê?
- Có 3 cách so sánh:
  - **Hiệu số**  $d = p_1 - p_2$
  - Tỉ số  $RR = p_1 / p_2$
  - Tỉ số odds  $OR = [p_1 / (1 - p_1)] / [p_2 / (1 - p_2)]$

# Binomial distribution – phân bố nhị phân

Population (true) proportion:  
 $\pi$



Sample proportion:  
 $p$

Lí thuyết:

- $p = (\text{cases} / \text{tổng số}) = x / N$   
 $p$  là ước số khách quan của  $\pi$
- Độ lệch chuẩn (standard deviation) của  $p$ :  $S = \sqrt{\frac{p(1-p)}{N}}$
- KTC 95% của  $\pi$  is:  $p \pm 1.96 \times S$

# So sánh 2 nhóm: Sample và population

	Sample (mẫu)		Population (quần thể)	
	Group 1	Group 2	Group 1	Group 2
N	$n_1$	$n_2$	Infinite	Infinite
Xác suất outcome	$p_1$	$p_2$	$\pi_1 = ?$	$\pi_2 = ?$
Hiệu số	$d = p_1 - p_2$		$\delta = \pi_1 - \pi_2$	
Tình trạng	Biết		Không biết	

Mục tiêu: dùng dữ liệu của mẫu để suy luận cho quần thể

# Phân tích so sánh 2 nhóm

	Sample (mẫu)	
	Nhóm 1	Nhóm 1
N	$n_1$	$n_2$
Xác suất outcome	$p_1$	$p_2$
Độ lệch chuẩn	$s_1$	$s_2$

Hiệu số ảnh hưởng

$$d = p_1 - p_2$$

Độ lệch chuẩn của  $d$

$$s = \sqrt{s_1^2 + s_2^2}$$

$$z = d / s$$

$$\text{KTC95\% of } d = d \mp 1.96s$$

# Hiệu quả chống gãy xương của zoledronic acid

	Placebo	Zoledronic acid
Số bệnh nhân	1062	1065
Số ca gãy xương	139	92
Không gãy xương	923	973
Tỉ lệ gãy xương	0.131	0.086
Độ lệch chuẩn	0.0103	0.0086

Hiệu số ảnh hưởng  $d = 0.131 - 0.086 = \textcolor{red}{0.045}$

Độ lệch chuẩn của  $d$   $s = \sqrt{0.0103^2 + 0.0086^2} = \textcolor{red}{0.013}$

KTC95%  $0.045 \pm 1.96 * s = \textcolor{red}{0.018 \text{ đến } 0.071}$

Z test  $z = 0.045 / 0.013 = 3.30$

P-value  $2 * (1 - \text{pnorm}(3.30)) = 0.0009$

# Diễn giải d và KTC95%

- Nếu zoledronic acid không có hiệu quả
  - $d = 0$
  - KTC95% của d dao động từ âm đến dương
- Nhưng kết quả cho thấy
  - $d \neq 0$
  - KTC95% của d đều dương
- Do đó, zoledronic acid có hiệu quả giảm nguy cơ gãy xương

# Dùng hàm prop.test

```
prop.test(x=c(139, 92), n=c(1062, 1065), correct=FALSE)
```

2-sample test for equality of proportions without continuity correction

```
data: c(139, 92) out of c(1062, 1065)
X-squared = 10.877, df = 1, p-value = 0.0009736
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01811559 0.07088471
sample estimates:
 prop 1      prop 2
0.13088512 0.08638498
```

# Ki bình phương (Chi squared test)

# Khái niệm "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:

$$P(A \text{ & } B) = P(A) \times P(B)$$

# 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)

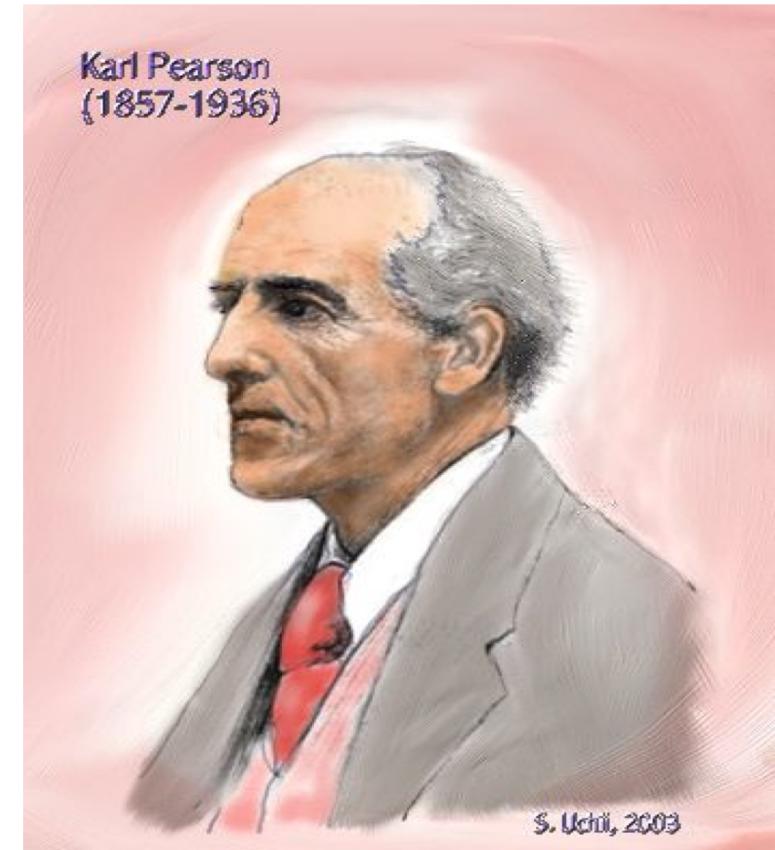
- Triết lí phản nghiệm (falsificationism) của Popper
- Bước 1: phát biểu giả thuyết vô hiệu (null hypothesis)
- Bước 2: thu thập dữ liệu (D)
- Bước 3: tính xác suất D xảy ra nếu giả thuyết vô hiệu đúng

# 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 (**expected 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  $\chi^2$  lớn, bác bỏ giả thuyết vô hiệu

# 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

# 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 <sup>2</sup>	4	64	144	4	9	256	81	169	36	36	81	16
D <sup>2</sup> /E	.11	1.58	3.51	.08	.24	5.95	1.85	3.91	.90	.81	2.99	.41

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

$$\chi^2 = 0.11 + 1.58 + 3.51 + \dots + 0.41 = 21.3$$

# 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
- $\chi^2 = 21.3$  phải so sánh với  $df = 11$
- Câu hỏi: xác suất mà  $\chi^2 = 21.3$  (hay cao hơn) nếu giả thuyết độc lập đúng là bao nhiêu?

1-pchisq(21.3, 11)

# Tóm lược

- Kiểm định Ki bình thường 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

# 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
<b>Tổng số</b>	<b>148</b>	<b>208</b>	<b>444</b>

# 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:

$$P(A \text{ & } B) = P(A) \times P(B)$$

# 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

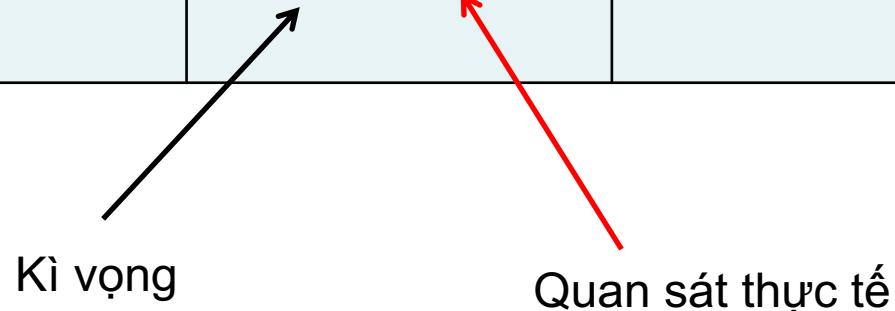
Trình độ học vấn	Tệ hơn	Không khác	Tốt hơn	Xác suất
Trung học (n=430)	79	112	238	0.537
Cao đẳng (n=160)	30	42	89	0.200
Đại học (n=210)	39	55	117	0.263
Xác suất	0.185	0.260	0.555	1.000

$$0.537 * 0.185 * 800 = 79$$

$$0.537 * 0.260 * 800 = 112$$

# Giá trị kì vọng và quan sát

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



# So sánh giá trị kì vọng và quan sát

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

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

# Giá trị kì vọng và quan sát (2)

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

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

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

# 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

# Tóm lược

- Mô tả một tỉ lệ
- So sánh 2 tỉ lệ: phương pháp z test (kiểm định z)
- So sánh nhiều tỉ lệ: kiểm định Ki bình phương