

Bài giảng 9: Phân tích mô tả số liệu phân nhóm: Ước tính tỉ lệ

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia
Đại học Tôn Đức Thắng, Việt Nam

Chúng ta sẽ học

- Cách ước tính một tỉ lệ, tỉ suất
- Cách tính khoảng tin cậy 95% cho một tỉ lệ và tỉ suất

Có bốn loại số liệu chính

Qualitative
(định tính)



Nominal (danh)

Ordinal (thứ tự)

Quantitative
(định lượng)



Interval (khoảng)

Ratio (tỉ số)

Đo lường định lượng

Interval level

- Classification + Ordering + Standard distance
- Tập hợp đối tượng có thể mô tả bằng đơn vị chỉ ra sự khác biệt ca này với ca khác
- *Ex: nhiệt độ*

Ratio level

- Classification + Ordering + Standard distance + Natural zero
- Biến định lượng và có "natural zero"
- *Ex: thu nhập, độ tuổi, huyết áp, v.v.*

Đo lường định tính

Nominal level

- Classification
- Tập hợp đối tượng có thể phân theo nhóm không trùng hợp nhau (mutually exclusive)
- *Ex: tôn giáo, giới tính, địa điểm*

Ordinal level

- Classification + Ordering
- Tập hợp số có ý nghĩa thứ bậc
- *Ex: trình độ học vấn, mức độ hài lòng, giai tầng xã hội, v.v.*

Biến định danh (nominal variable)

- Alive vs. dead
- Male vs. female
- Rabies vs. no rabies
- Blood group O, A, B, AB
- Resident of Michigan, Ohio, Indiana...

**Phân tích mô tả:
Ước tính một tỉ lệ**

Số liệu liên quan đến tỉ lệ

- *Proportion, percentage, probability (xác suất)*
- Estimate (ước số)
- Khoảng tin cậy 95% (95% confidence interval)

Ví dụ: Ước tính tỉ lệ hiện hành (prevalence)

- Điều tra ngẫu nhiên 1000 người trong độ tuổi 60-69 ($n = 1000$)
- Kết quả: 100 bị ung thư ($x = 100$)
- Chúng ta muốn ước tính tỉ lệ hiện hành và suy luận tỉ lệ trong quần thể (population), tức π

Ước tính tỉ lệ hiện hành (prevalence)

- Ước số (còn gọi là *point estimate*) chỉ đơn giản là:

$$P = \frac{\text{no. cases}}{\text{no. of people}} = \frac{100 \text{ people}}{1000 \text{ people}} = 0.10$$

Chúng ta kết luận: Tỉ lệ mắc bệnh ung thư là 10%

Nhưng còn KTC95%?

Khoảng tin cậy 95% của tỉ lệ

- Cần ước tính độ lệch chuẩn:

$$s = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.10 \times 0.90}{1000}} = 0.009$$

- KTC95%

$$p \pm 1.96s = 0.10 \pm 1.96 \times 0.009$$
$$= \mathbf{0.08 \text{ to } 0.12}$$

Kết luận: Nếu lấy mẫu nhiều lần, 95% tỉ lệ sẽ dao động trong khoảng 8 đến 12%.

Suy luận: Xác suất 95% là π nằm trong khoảng 8 – 12%.

Phương pháp chính xác (exact method)

- Cách tính vừa mô tả là "xấp xỉ" (approximation)
- Xấp xỉ dựa vào giả định rằng tỉ lệ (P) ước tính từ nhiều mẫu tuân theo luật phân bố chuẩn (normal distribution)
- Phương pháp chính xác hơn ... dùng package **epitools**
- Trường hợp n nhỏ, nên dùng phương pháp chính xác

```
library(epitools)
```

```
binom.exact(x=100, n=1000, methods="all")
```

```
> binom.exact(100, 1000)
```

	x	n	proportion	lower	upper	conf.level
1	100	1000	0.1	0.08210533	0.1202879	0.95

```
library(binom)
```

```
binom.confint(x=100, n=1000, methods="all")
```

```
> binom.confint(x=100, n=1000, methods="all")
```

	method	x	n	mean	lower	upper
1	agresti-coull	100	1000	0.1000000	0.08284688	0.1202145
2	asymptotic	100	1000	0.1000000	0.08140615	0.1185939
3	bayes	100	1000	0.1003996	0.08206073	0.1191877
4	cloglog	100	1000	0.1000000	0.08239444	0.1195577
5	exact	100	1000	0.1000000	0.08210533	0.1202879
6	logit	100	1000	0.1000000	0.08288164	0.1201906
7	probit	100	1000	0.1000000	0.08264461	0.1198768
8	profile	100	1000	0.1000000	0.08243331	0.1196133
9	lrt	100	1000	0.1000000	0.08243172	0.1196130
10	prop.test	100	1000	0.1000000	0.08245237	0.1206909
11	wilson	100	1000	0.1000000	0.08290944	0.1201520

```
library(binom)
```

```
binom.bayes(x=100, n=1000)
```

```
> binom.bayes(x=100, n=1000)
```

	method	x	n	shape1	shape2	mean	lower	upper	sig
1	bayes	100	1000	100.5	900.5	0.1003996	0.08206073	0.1191877	0.05

**Phân tích mô tả:
Ước tính một tỉ suất (rate)**

Tỉ suất (rate)

- Có yếu tố thời gian
- Tử số: Số ca với biến cố quan tâm
- Mẫu số: **Số đối tượng * thời gian** (còn gọi là person-years, person-months, v.v.)
- Phân bố **Poisson**

Ví dụ: *incidence rate*

- 45 cá nhân được theo dõi trong thời gian 4 năm
- Trong thời gian theo dõi có 3 người mắc bệnh
- Ước tính
 - tỉ lệ phát sinh (incidence proportion)
 - **tỉ suất phát sinh (incidence rate)**

Tỉ lệ phát sinh

- Dùng phương pháp prevalence
- $P = 3/45 = 0.067$ (6.7%)

```
> binom.confint(x=3, n=45, methods="all")
```

	method	x	n	mean	lower	upper
1	agresti-coull	3	45	0.06666667	0.016334990	0.1851631
2	asymptotic	3	45	0.06666667	-0.006214379	0.1395477
3	bayes	3	45	0.07608696	0.011993564	0.1522221
4	cloglog	3	45	0.06666667	0.017337640	0.1638879
5	exact	3	45	0.06666667	0.013965097	0.1826845
6	logit	3	45	0.06666667	0.021660639	0.1872841
7	probit	3	45	0.06666667	0.019474979	0.1742618
8	profile	3	45	0.06666667	0.017060470	0.1638663
9	lrt	3	45	0.06666667	0.017034597	0.1638635
10	prop.test	3	45	0.06666667	0.017376865	0.1931102
11	wilson	3	45	0.06666667	0.022932033	0.1785660

Tỉ suất phát sinh

- **45 cá nhân** được theo dõi trong thời gian **4 năm**; **3 người mắc bệnh**
- Incidence rate (tỉ lệ phát sinh) :

$$I = 3 / (45 * 4) = 0.017$$

- Sai số chuẩn của I là:

$$SE = \text{sqrt}(3) / (45 * 4) = 0.0096$$

- KTC95% là: **0.017 \pm 1.96x0.0096**

-0.0018 đến 0.036

Dùng **epitools**

- KTC95% âm – không thể!

```
library(epitools)
```

```
> library(epitools)
> pois.exact(x=3, pt=45*4)
```

	x	pt	rate	lower	upper	conf.level
1	3	180	0.01666667	0.003437062	0.04870709	0.95

Tóm lược

- Ước tính một tỉ lệ, tỉ suất: đơn giản
- Tính khoảng tin cậy 95%: **nên dùng phương pháp chính xác**