Phân tích mô tả
biến liên tục

Tuan V. Nguyen

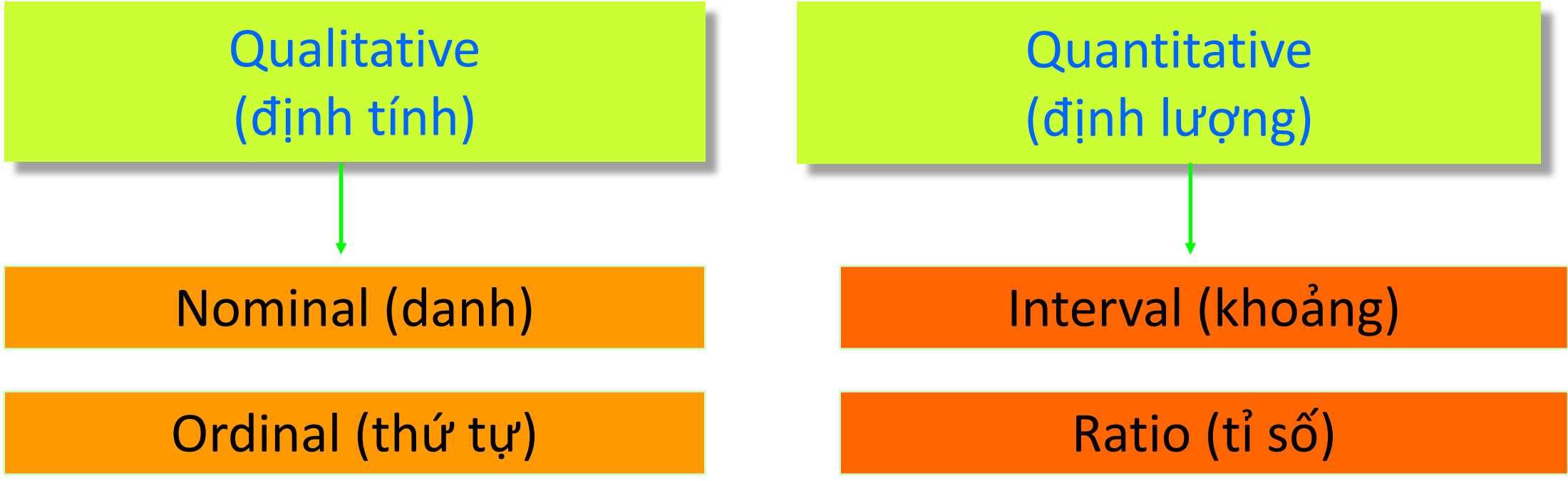
Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam

Phân loại biến đo lường



Phân loại biến đo lường

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, điểm thi*

Biến 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.*

Mô tả biến liên tục

Biến liên tục

- Trung bình (mean, average)
- Trung vị (median)
- Độ lệch chuẩn (standard deviation)
- Bách phân vị (percentile)

Trung vị (median)

- Một biến với giá trị như sau:

1, 9, 6, 7, 8, 4, 5, 3, 25

- Sắp xếp lại từ thấp đến cao. Trung vị là số chính giữa

1, 3, 4, 5, 6, 7, 8, 9, 25

Trung vị (median): Dùng R

```
x = c(1, 9, 6, 7, 8, 4, 5, 3, 25)
median(x)
```

```
> median(x)
[1] 6
```

Phương sai (variance)

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

Nghiên cứu 1: 6, 7, 8, 4, 5, và 6

$$s^2 = \frac{(6-6)^2 + (7-6)^2 + (8-6)^2 + (5-6)^2 + (6-6)^2}{6-1} = \frac{10}{5} = 2$$

Nghiên cứu 2: 10, 2, 3, 9

$$s^2 = \frac{(10-6)^2 + (2-6)^2 + (3-6)^2 + (9-6)^2}{4-1} = \frac{50}{3} = 16.7$$

Độ lệch chuẩn (standard deviation)

Căn số bậc 2 của phương sai $s = \sqrt{s^2}$

- Nghiên cứu 1: 6, 7, 8, 4, 5, và 6, phương sai là:

$$s^2 = \frac{(6-6)^2 + (7-6)^2 + (8-6)^2 + (5-6)^2 + (6-6)^2}{6-1} = \frac{10}{5} = 2$$

- Nghiên cứu 2: 10, 2, 3, 9, phương sai là:

$$s^2 = \frac{(10-6)^2 + (2-6)^2 + (3-6)^2 + (9-6)^2}{4-1} = \frac{50}{3} = 16.7$$

- Nghiên cứu 1, s = sqrt(2) = 1.41

  Nghiên cứu 2, s = sqrt(16.7) = 4.1

Khoảng tin cậy 95% (95% confidence interval)

- Nếu một biến (tuân theo luật phân bố chuẩn) có trung bình là *m,* và độ lệch chuẩn *s*
- Chúng ta có thể **suy luận** rằng : 95% các giá trị của biến dao động trong khoảng

$$m - 1.96*s$$

đến

$$m + 1.96*s$$

Khoảng tin cậy 95% (95% confidence interval)

- Chiều cao trung bình nữ giới (16-65 tuổi):
  - Trung bình: m = 163.5 cm
  - Độ lệch chuẩn: s = 6.1 cm
- Chúng ta có thể **suy luận** rằng : 95% phụ nữ có chiều cao dao động trong khoảng

163.5 - 1.96*6.1 đến 163.5 + 1.96*6.1

151.5 đến 175.5 cm

Phân tích mô tả với R

Ba package chính cho phân tích mô tả

- `table1`
- `compareGroups`
- `DescTools`

Package "table1"

```
library(table1)

table1(~ var1 + var2 + var3 ... | group, data=xxx)

table1(~ var1 + var2 + var3 ... | group, overall=F,
data=xxx)

table1(~ var1 + var2 + var3 ... | group, overall=F,
transpose=T, data=xxx)
```

Dùng package "table1"

```
library(table1)

table1(~Group + Region + Age + Height
+ Bust + Waist + Hip, data=hh)
```

	Overall (n=46)
Group	
A Hau	30 (65.2%)
Hoa Hau	16 (34.8%)
Region	
Central	2 (4.3%)
North	28 (60.9%)
South	16 (34.8%)
Age	
Mean (SD)	19.4 (2.08)
Median [Min, Max]	19.0 [16.0, 24.0]
Missing	1 (2.2%)
Height	
Mean (SD)	170 (4.49)
Median [Min, Max]	171 [158, 180]
Bust	
Mean (SD)	84.1 (2.78)
Median [Min, Max]	84.5 [78.0, 87.0]
Missing	30 (65.2%)
Waist	
Mean (SD)	61.7 (1.99)
Median [Min, Max]	61.0 [58.0, 65.0]
Missing	30 (65.2%)
Hip	
Mean (SD)	90.2 (2.97)
Median [Min, Max]	90.5 [84.0, 95.0]
Missing	30 (65.2%)

```
library(table1)

table1(~ Region + Age + Height | Group, data=hh)
```

	A Hau (n=30)	Hoa Hau (n=16)	Overall (n=46)
Region			
Central	0 (0%)	2 (12.5%)	2 (4.3%)
North	18 (60.0%)	10 (62.5%)	28 (60.9%)
South	12 (40.0%)	4 (25.0%)	16 (34.8%)
Age			
Mean (SD)	19.6 (2.21)	19.3 (1.88)	19.4 (2.08)
Median [Min, Max]	19.0 [16.0, 24.0]	19.2 [16.1, 22.2]	19.0 [16.0, 24.0]
Missing	1 (3.3%)	0 (0%)	1 (2.2%)
Height			
Mean (SD)	170 (3.72)	170 (5.78)	170 (4.49)
Median [Min, Max]	170 [160, 180]	171 [158, 180]	171 [158, 180]

Package `DescTools`

- Package "`DescTools`" : bảng số liệu & biểu đồ
- Hàm chính: **`Desc`**

```
Desc(var ~ group, options)
```

Mô tả biến biến liên tục

```
phd = read.csv("~/Dropbox/_Conferences and Workshops/Banking
University/Data/Income and PhDs.csv")
```

```
> library(DescTools)
> options("scipen"=100, "digits"=6)
> Desc(phd$Salary)
-------------------------------------------------------------------------
phd$Salary (integer)

     length           n         NAs      unique          0s        mean      meanCI
         62          62           0         = n           0  53'941.89  51'030.44
                100.0%        0.0%                    0.0%              56'853.33

        .05         .10         .25      median         .75         .90         .95
  39'141.85   41'956.80   47'492.00   53'537.50   59'415.75   68'766.70   74'050.00

      range          sd       vcoef         mad         IQR        skew        kurt
  77'176.00   11'464.52        0.21    9'060.91   11'923.75       -0.63        3.61

lowest : 6'327, 37'939, 38'340, 39'115, 39'652
highest: 71'219, 74'199, 74'343, 75'822, 83'503
```

Biến liên tục:
So sánh 2 nhóm

Lí giải: so sánh hai nhóm

- Gọi số trung bình của nhóm 1 và 2 là μ_1 và μ_2
- Gọi khác biệt giữa hai nhóm là $\delta = \mu_1 - \mu_2$
- Nếu hai nhóm có số trung bình giống nhau, thì $\delta = 0$
- Chúng ta không biết μ_1, μ_2 và δ
- Nhưng chúng ta có dữ liệu từ mẫu: m_1, m_2, và d.

Kiểm định t

	Mẫu		Quần thể	
	A	B	A	B
N	n_1	n_2	Infinite	Infinite
Mean	x_1	x_2	$\mu_1 = ?$	$\mu_2 = ?$
SD (standard deviation)	s_1	s_2	$\sigma_1 = ?$	$\sigma_2 = ?$
Difference	$d = x_1 - x_1$		$\delta = \mu_1 - \mu_2$	
Status	Biết		Không biết	

Muốn dùng dữ liệu của mẫu để suy luận cho quần thể. Giả thuyết là $\delta = 0$

Kiểm định t

$$t = \frac{m_1 - m_2}{SE_{m_1 - m_2}} = \frac{signal}{noise}$$

$$SE_{m_1 - m_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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

Chỉ số *t* gọi là "statistic"

t = tín hiệu / nhiễu; giá trị t cao có nghĩa là sự khác biệt có thể không phải ngẫu nhiên

t tuân theo luật phân bố t

khi *n > 30* và t > 2, chúng ta có thể nói rằng sự khác biệt có ý nghĩa thống kê (statistically significant)

Lí giải

- Nếu hai nhóm không khác nhau, thì $d = 0$ (d là ước số của δ)

  (Nhưng nếu hai nhóm khác nhau thì $d \neq 0$)

Kiểm định thống kê: **nếu δ = 0 thì xác suất chúng ta quan sát t hay cao hơn t là bao nhiêu?** → Trị số P

t-test được triển khai trong R: hàm "t.test"

```
> phd = read.csv("~/Dropbox/_Conferences and Workshops/Banking
University/Data/Income and PhDs.csv")
> head(phd)
  id TimeSincePhD NPubs Sex Citations Salary
1  1            3    18   1        50  51876
2  2            6     3   1        26  54511
3  3            3     2   1        50  53425
4  4            8    17   0        34  61683
5  5            9    11   1        41  52926
6  6            6     6   0        37  47034
```

t-test được triển khai trong R: hàm "t.test"

```
t.test(phd$Salary ~ phd$Sex)
```

```
> t.test(phd$Salary ~ phd$Sex)

data:  phd$Salary by phd$Sex
t = 2.0396, df = 53.705, p-value = 0.04632
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    99.7255 11694.1772
sample estimates:
mean in group 0 mean in group 1
       56509.91        50612.96
```

Diễn giải kết quả t test

```
> t.test(phd$Salary ~ phd$Sex)

data:  phd$Salary by phd$Sex
t = 2.0396, df = 53.705, p-value = 0.04632
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    99.7255 11694.1772
sample estimates:
mean in group 0 mean in group 1
       56509.91        50612.96
```

Tính trung bình, nữ có lương thấp hơn nam $5897 (50613 - 56510), và khoảng tin cậy 95% dao động từ $100 đến $50613. Sự khác biệt có ý nghĩa thống kê (P = 0.046).

Mô tả biến liên tục

- Nếu biến số tuân theo luật phân bố chuẩn
  - Trung bình, độ lệch chuẩn, n
- Nếu biến số lệch khỏi luật phân bố chuẩn
  - Trung vị, bách phân vị 25 – 75%
- Triển khai trong R: `table1, compareGroups, DescTools`