Bài giảng 9: Phân tích mô tả với package "tables"

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Ton Duc Thang University, Vietnam

Chúng ta muốn biến từ dữ liệu thô ...

Garvanid	age	gender	actn3	weight	height	fnbmd	lsbmd	trbmd	wtbmd	lean	fat	quadstrength
8	67.4	Female	RX	72	166	0.97	1.33	0.85	0.85	.	.	18
9	68.5	Male	XX	87	184	1.01	1.49	0.87	0.84	.	.	36
10	62.1	Female	RR	72	173	0.84	1.21	0.62	0.71	.	.	14
27	64.6	Female	RX	85	167	0.86	1.07	0.78	0.68	.	.	27
28	76.2	Female	RX	48	153	0.65	0.87	0.5	0.46	.	.	20
33	74.1	Female	RX	52	156	0.83	0.85	0.71	0.61	32.61	22.35	22
34	75.3	Female	RR	70	160	0.79	1.19	0.7	0.51	.	.	8
36	62.2	Male	XX	97	171	1.16	1.44	1.16	0.94	.	.	31
37	59.8	Female	XX	60	161	0.79	0.91	0.61	0.63	.	.	35
38	66.4	Male	RR	64	170	0.85	1.12	0.83	0.71	.	.	30
40	67.7	Male	RX	82	179	0.75	1.07	0.96	0.44	.	.	41
41	65.1	Male	RR	101	174	0.94	1.22	0.88	0.65	.	.	45
42	62.1	Female	XX	82	151	0.99	1.42	0.96	0.94	.	.	22
47	72.5	Female	RX	56	150	0.62	0.75	0.56	0.42	.	.	12
49	72.2	Female	RR	46	154	0.58	0.79	0.49	0.38	.	.	19
51	68.7	Female	RX	65	163	0.79	1.28	0.8	0.66	.	.	.
53	80.4	Male	RR	76	178	0.99	1.23	0.99	0.78	.	.	18
56	69.3	Female	RX	55	162	0.75	0.73	0.52	0.58	.	.	26
57	63	Male	RR	86	160	.	1.49	.	.	.	.	30
58	60.9	Female	RX	58	161	0.59	0.76	0.5	0.5	.	.	32
60	58.6	Female	RX	64	151	0.94	0.94	0.73	0.68	.	.	21
61	59.8	Male	RX	86	184	1.17	1.46	1.1	1.01	.	.	46
62	59	Female	RR	75	169	0.95	1.05	0.64	0.78	.	.	9
64	71.5	Female	XX	55	155	0.63	0.94	0.57	0.46	.	.	26
65	69.3	Male	XX	114	177	1.19	1.38	1.25	0.96	.	.	46

thành các bảng số liệu (information)

Table 1. Explanatory Variables for Patients with Oropharyngeal Cancer and Control Patients.*

Explanatory Variable	Patients with Oropharyngeal Cancer (N=100)	Control Patients (N=200)	Unadjusted Odds Ratio (95% CI)†
	number (percent)		
Demographic characteristics			
Sex			
Female	14 (14)	28 (14)	1.0
Male	86 (86)	172 (86)	1.0 (0.5–2.0)
Age			
<50 yr	34 (34)	68 (34)	1.0
50–64 yr	51 (51)	102 (51)	1.0 (0.6–1.7)
≥65 yr	15 (15)	30 (15)	1.0 (0.5–2.0)
Highest educational level			
Some high school	11 (11)	15 (8)	1.0
High-school graduate or some college	41 (41)	71 (36)	0.8 (0.3–1.9)
College graduate	48 (48)	114 (57)	0.6 (0.3–1.4)‡

D'Souza G, et al. Case–Control Study of Human Papillomavirus and Oropharyngeal Cancer. NEJM 2007;356:1944-56

2. CLINICAL MANIFESTATIONS

Parenteral Nutrition

Table 3.10. Time on PN

Type	Group	Recruiment (n)	Mean	Min	Max
Parenteral Nutrition	All	51	41.37 ± 46.3	1	281
	Dischargeable	21	55.86 ± 64.87	6	281
	Died	29	31.28 ± 23.58	1	82
Total PN	Dischargeable	21	11.14 ±9.66	4	47
Adequate requirement	All	23	20.13 ± 23.82	3	122
	Dischargeable	10	30.90± 33.34	4	22
≥ RMEE	All	40	25.70 ± 22.14	2	101
	Dischargeable	19	26 ± 26.73	2	101
< RMEE	Dischargeable	21	12.14 ± 25.77	1	180

RMEE: Resting Metabolic Energy Expenditure

Nội dung

- Phân tích mô tả
- Giới thiệu package "tables"
- Các hàm phổ biến và cách sử dụng

Phân tích mô tả

- Một bước rất quan trọng để "cảm nhận" dữ liệu
- Mô tả biến liên tục: trung vị, trung bình, độ lệch chuẩn, bách phân vị, v.v.
- Khi mô tả có phân nhóm: Vấn đề trở nên khá phức tạp

Dữ liệu

```
setwd("~/Dropbox/World Bank 2014/Data for 2015
   workshop")

pisa = read.csv("~/Dropbox/World Bank 2014/Data for 2015
   workshop/PISA DATA.csv", header=T)

attach(pisa)

# t=tabular(REGION ~ AREA*(n=1+Percent("col")))

# html(t, "test.html")
```

Hàm summary trong R

- R có hàm summary() có thể dùng để tóm tắt dữ liệu
- Ví dụ:

```
attach(pisa)

summary(pisa)
```

Dùng package psych

- **psych**: Package rất có ích cho phân tích mô tả và phân tích đa biến, đa phần ứng dụng trong khoa học xã hội
- Hàm **describe** và **describeBy** có thể sử dụng cho phân tích mô tả

Dùng psych

```
setwd("~/Dropbox/World Bank 2014/Data for 2015
  workshop")

pisa = read.csv("~/Dropbox/World Bank 2014/Data for
  2015 workshop/PISA DATA.csv", header=T)

attach(pisa)

library(psych)

describe(pisa)

describe(cbind(PV1MATH, PV1SCIE, PV1READ), skew=F)

describeBy(cbind(PV1MATH, PV1SCIE, PV1READ),
  group=AREA, skew=F, range=F)
```

Một phần output của psych

```
> describeBy(cbind(PV1MATH, PV1SCIE, PV1READ), group=AREA, skew=F,
   range=F)

INDICES: REMOTE
         var   n   mean    sd   se
PV1MATH    1 324 461.10 71.34 3.96
PV1SCIE    2 324 498.77 70.42 3.91
PV1READ    3 324 471.80 75.09 4.17
------------------------------------------------------------
INDICES: RURAL
         var    n   mean    sd   se
PV1MATH    1 2278 501.29 77.23 1.62
PV1SCIE    2 2278 518.91 71.10 1.49
PV1READ    3 2278 502.51 67.99 1.42
------------------------------------------------------------
INDICES: URBAN
         var    n   mean    sd   se
PV1MATH    1 2357 526.38 86.88 1.79
PV1SCIE    2 2357 541.13 77.92 1.61
PV1READ    3 2357 518.82 73.03 1.50
```

Phân tích mô tả với psych

- Khá tốt!
- Nhưng output theo nhóm được trình bày theo dòng – có khi "khó" đọc
- Thông thường, chúng ta đọc theo cột

	Nhóm 1	**Nhóm 2**	**Nhóm 3**
Biến số 1	mean, sd	mean, sd	mean, sd
Biến số 2	mean, sd	mean, sd	mean, sd
Biến số 3	mean, sd	mean, sd	mean, sd
....			

Giới thiệu package "tables"

Package "tables"

- Mô phỏng theo PROC TABULATE của SAS
- Hàm chủ yếu: **`tabular()`**
- Rất có ích cho phân tích mô tả đơn giản, mô tả theo nhóm
- Có thể dùng cho biến phân loại và biến liên tục
- Cách trình bày có thể theo ý của người sử dụng

Cách sử dụng hàm tabular()

- **Ví dụ**: Chúng ta muốn có một bảng số liệu về số học sinh cho từng vùng (REGION)

REGION	Số học sinh
CENTRAL	xx
NORTH	xx
SOUTH	xx

```
attach(pisa)
library(tables)
tabular(REGION~1)
```

Cách sử dụng hàm tabular(): thêm %

- **Ví dụ**: Chúng ta muốn có một bảng số liệu về số học sinh + tính phần trăm phân bố cho từng vùng (REGION)

REGION	Số học sinh	Phần trăm
CENTRAL	xx	yy
NORTH	xx	yy
SOUTH	xx	yy

```
attach(pisa)
library(tables)
tabular(REGION~1*(n=1 + Percent("col")))
```

```
> tabular(REGION~1)

 REGION  All
 CENTRAL 1591
 NORTH   1751
 SOUTH   1617

> tabular(REGION~1*(n=1 + Percent("col")))

         All
         n
 REGION  All  Percent
 CENTRAL 1591 32.08
 NORTH   1751 35.31
 SOUTH   1617 32.61
```

Phân tích 2 biến phân loại

- Mục tiêu: Muốn biết số học sinh từng miền (REGION) và vùng (AREA)

REGION	URBAN	RURAL	REMOTE
CENTRAL	xx	yy	zz
NORTH	xx	yy	zz
SOUTH	xx	yy	zz

```
library(tables)
tabular(REGION ~ AREA)
```

Phân tích 2 biến phân loại

- Mục tiêu: Muốn biết số học sinh và phần trăm từng miền (REGION) và vùng (AREA)

REGION	URBAN	RURAL	REMOTE
CENTRAL	xx	yy	zz
NORTH	xx	yy	zz
SOUTH	xx	yy	zz

```
library(tables)
tabular(REGION ~ AREA*(n=1 + Percent("row")))
tabular(REGION ~ AREA*(n=1 + Percent("col")))
```

```
tabular(REGION ~ AREA*(n=1 + Percent("row")))
```

	AREA					
	REMOTE		RURAL		URBAN	
	n		n		n	
REGION	All	Percent	All	Percent	All	Percent
CENTRAL	121	7.605	771	48.46	699	43.93
NORTH	141	8.053	865	49.40	745	42.55
SOUTH	62	3.834	642	39.70	913	56.46

```
> tabular(REGION ~ AREA*(n=1 + Percent("col")))
```

	AREA					
	REMOTE		RURAL		URBAN	
	n		n		n	
REGION	All	Percent	All	Percent	All	Percent
CENTRAL	121	37.35	771	33.85	699	29.66
NORTH	141	43.52	865	37.97	745	31.61
SOUTH	62	19.14	642	28.18	913	38.74

Phân tích 3 biến phân loại

- Mục tiêu: Muốn biết số học sinh và phần trăm từng miền (REGION), vùng (AREA) và loại trường (TYPE)

REGION	AREA	PUBLIC	PRIVATE
CENTRAL	REMOTE	yy	zz
	RURAL	yy	zz
	URBAN	yy	zz
NORTH	REMOTE	yy	zz
	RURAL	yy	zz
	URBAN	yy	zz
SOUTH	REMOTE	yy	zz
	RURAL	yy	zz
	URBAN	yy	zz

Phân tích 3 biến phân loại

REGION	AREA	PUBLIC	PRIVATE
CENTRAL	REMOTE	yy	zz
	RURAL	yy	zz
	URBAN	yy	zz
NORTH	REMOTE	yy	zz
	RURAL	yy	zz
	URBAN	yy	zz
SOUTH	REMOTE	yy	zz
	RURAL	yy	zz
	URBAN	yy	zz

```
library(tables)
tabular(REGION*AREA ~ TYPE)
tabular(REGION*AREA ~ TYPE*(n=1 + Percent("row")))
```

```
> tabular(REGION*AREA ~ TYPE*(n=1 + Percent("row")))
```

		TYPE			
		PRIVATE		PUBLIC	
		n		n	
REGION	AREA	All	Percent	All	Percent
CENTRAL	REMOTE	0	0.000	121	100.00
	RURAL	68	8.820	703	91.18
	URBAN	69	9.871	630	90.13
NORTH	REMOTE	12	8.511	129	91.49
	RURAL	68	7.861	797	92.14
	URBAN	102	13.691	643	86.31
SOUTH	REMOTE	0	0.000	62	100.00
	RURAL	44	6.854	598	93.15
	URBAN	64	7.010	849	92.99

Phân tích biến liên tục với tabular()

- **Mục tiêu**: Muốn biết điểm trung bình + SD của học sinh từng miền (REGION)

REGION	Điểm trung bình	Độ lệch chuẩn
CENTRAL	xx	yy
NORTH	xx	yy
SOUTH	xx	yy

```
library(tables)
tabular(REGION ~ PV1MATH*(n=1+mean+sd))
```

```
> tabular(REGION ~ MATH*(n=1+mean+sd))

          MATH
          n
 REGION   All  mean  sd
 CENTRAL 1591 513.1 75.99
 NORTH   1751 519.1 90.54
 SOUTH   1617 498.9 81.38
```

Dùng tabular() khi biến phân nhóm là số

- Mục tiêu: Muốn biết điểm toán cho từng trường (SCHOOLID)
- SCHOOLID được mã hoá bằng số
- Do đó, phải chuyển sang "factor"

```
library(tables)
School = as.factor(SCHOOLID)
tabular(School ~ PV1MATH*(n=1+mean+sd))
```

Phân tích biến liên tục với tabular()

- **Mục tiêu**: Muốn biết điểm toán và đọc (trung bình + SD) của học sinh từng miền (REGION)

REGION	Điểm môn toán	Điểm môn văn/đọc
CENTRAL	xx	yy
NORTH	xx	yy
SOUTH	xx	yy

```
library(tables)
tabular(REGION ~ (PV1MATH + PV1READ)*(n=1+mean
+sd))
```

```
> tabular(REGION ~ (PV1MATH + PV1READ)*(n=1+mean
+sd))
```

	PV1MATH			PV1READ		
	n			n		
REGION	All	mean	sd	All	mean	sd
CENTRAL	1591	513.1	75.99	1591	511.8	68.43
NORTH	1751	519.1	90.54	1751	510.4	75.62
SOUTH	1617	498.9	81.38	1617	502.4	70.99

Phân tích biến liên tục với tabular()

- **Mục tiêu**: Muốn biết điểm toán (trung bình + SD) của học sinh từng miền (REGION) và loại trường (TYPE)

```
library(tables)
tabular(REGION*TYPE ~ PV1MATH*(n=1+mean+sd))
```

```
> tabular(REGION*TYPE ~ MATH*(n=1+mean+sd))

                  MATH
                  n
 REGION  TYPE     All  mean  sd
 CENTRAL PRIVATE  137 467.9 60.12
         PUBLIC  1454 517.4 75.96
 NORTH   PRIVATE  182 471.4 63.36
         PUBLIC  1569 524.7 91.60
 SOUTH   PRIVATE  108 486.3 52.51
         PUBLIC  1509 499.8 83.00
```

Export sang dạng html

```
setwd("~/Dropbox/World Bank 2014/Data for 2015
  workshop")

t = tabular(REGION*TYPE ~ PV1MATH*(n=1+mean+sd))

html(t, "test.html")
```

Từ html có thể chuyển sang Word và biên tập lại cho "thẩm mĩ" hơn

Package "table"

- Rất có ích cho các phân tích mô tả (biến liên tục và phân nhóm)
- Có thể dùng để xây dựng các bảng số liệu đa biến và phức tạp
- Có thể "exported" sang latex hoặc html