OpenIntro Statistics

Rodrigo Dias Martins

11 de mar??o de 2015

CHAPTER 1 - INTRODUCTION TO DATA

Exercise 1.1

Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year.

```
setwd("/Volumes/E-Books and articles/e-Books & articles/R/OpenIntro Statistics/openintroData")
load("stent.rda")
s30 <- as.data.frame(rbind(table((split(stent, stent$period)[[1]])[, 1:2]),
    colSums(table((split(stent, stent$period)[[1]])[, 1:2]))))
rownames(s30) <- c("control", "treatment", "TOTAL")
s365 <- as.data.frame(rbind(table((split(stent, stent$period)[[2]])[, 1:2]),
    colSums(table((split(stent, stent$period)[[2]])[, 1:2]))))
rownames(s365) <- c("control", "treatment", "TOTAL")
library(pander)
pander(s30, "Events - 0 to 30 days period", emphasize.strong.rows = 3)
```

Table 1: Events - 0 to 30 days period

	no event	stroke
control	214	13
treatment	191	33
TOTAL	**405**	**46**

```
pander(s365, "Eventsp - 0 to 365 days period", emphasize.strong.rows = 3)
```

Table 2: Eventsp - 0 to 365 days period

	no event	stroke
control	199	28
treatment	179	45
TOTAL	**378**	**73**

ANSWER :

- Proportion of patientes who had a stroke in treatment group :

```
round(45/224,3) ; paste((round(45/224,3)*100),'%',sep="")
```

```
## [1] 0.201
```

```
## [1] "20.1%"
```

- Proportion of patientes who had no stroke in treatment group :

```
round(179/224,3) ; paste((round(179/224,3)*100),'%',sep="")
```

```
## [1] 0.799
```

```
## [1] "79.9%"
```

- Proportion of patientes who had stroke in control group :

```
round(28/227,2) ; paste((round(28/227,3)*100),'%',sep="")
```

```
## [1] 0.12
```

```
## [1] "12.3%"
```

- Proportion of patientes who had no stroke in control group :

```
round(199/227,2) ; paste((round(199/227,3)*100),'%',sep="")
```

```
## [1] 0.88
```

```
## [1] "87.7%"
```

Exercise 1.2

We consider a publicly available data set that summarizes information about the 3,143 counties in the United States, and we call this the `county` data set. This data set includes information about each county: its name, the state where it resides, its population in 2000 and 2010, per capita federal spending, poverty rate, and five additional characteristics. How might these data be organized in a data matrix?

ANSWER :

```
setwd("/Volumes/E-Books and articles/e-Books & articles/R/OpenIntro Statistics/openintroData")
load("county.rda")
library(pander)
panderOptions("table.split.table", 110)
pander(summary(county), "county summary")
```

Table 3: county summary (continued below)

name	state	pop2000	pop2010	fed_spend
Length:3143	Length:3143	Min. : 67	Min. : 82	Min. : 0.000
Class :character	Class :character	1st Qu.: 11210	1st Qu.: 11104	1st Qu.: 6.964
Mode :character	Mode :character	Median : 24608	Median : 25857	Median : 8.669
NA	NA	Mean : 89623	Mean : 98233	Mean : 9.991
NA	NA	3rd Qu.: 61766	3rd Qu.: 66699	3rd Qu.: 10.857
NA	NA	Max. :9519338	Max. :9818605	Max. :204.616
NA	NA	NA's :3	NA	NA's :4

poverty	homeownership	multiunit	income	med_income	smoking_ban
Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. : 7772	Min. : 19351	Length:3143
1st Qu.:11.0	1st Qu.:69.50	1st Qu.: 6.10	1st Qu.:19030	1st Qu.: 36952	Class :character

poverty	homeownership	multiunit	income	med_income	smoking_ban
Median :14.7	Median :74.60	Median : 9.70	Median :21773	Median : 42445	Mode :character
Mean :15.5	Mean :73.26	Mean :12.33	Mean :22505	Mean : 44270	NA
3rd Qu.:19.0	3rd Qu.:78.40	3rd Qu.:15.90	3rd Qu.:24814	3rd Qu.: 49142	NA
Max. :53.5	Max. :91.30	Max. :98.50	Max. :64381	Max. :115574	NA
NA	NA	NA	NA	NA	NA

```
panderOptions("table.split.table", 100)
panderOptions("round", 4)
panderOptions("keep.trailing.zeros", TRUE)
pander(head(cbind(entry = 1:nrow(county), county), 5), "How to organize it in a data matrix")
```

Table 5: How to organize it in a data matrix (continued below)

entry	name	state	pop2000	pop2010	fed_spend	poverty
1	Autauga County	Alabama	43671	54571	6.068	10.6
2	Baldwin County	Alabama	140415	182265	6.140	12.2
3	Barbour County	Alabama	29038	27457	8.752	25.0
4	Bibb County	Alabama	20826	22915	7.122	12.6
5	Blount County	Alabama	51024	57322	5.131	13.4

homeownership	multiunit	income	med_income	smoking_ban
77.5	7.2	24568	53255	none
76.7	22.6	26469	50147	none
68.0	11.1	15875	33219	none
82.9	6.6	19918	41770	none
82.0	3.7	21070	45549	none

- Types of variables in `county`

```
str(county)
```

```
## 'data.frame':    3143 obs. of  11 variables:
##  $ name         : chr  "Autauga County" "Baldwin County" "Barbour County" "Bibb County" ...
##  $ state        : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ pop2000      : int  43671 140415 29038 20826 51024 11714 21399 112249 36583 23988 ...
##  $ pop2010      : int  54571 182265 27457 22915 57322 10914 20947 118572 34215 25989 ...
##  $ fed_spend    : num  6.07 6.14 8.75 7.12 5.13 ...
##  $ poverty      : num  10.6 12.2 25 12.6 13.4 25.3 25 19.5 20.3 17.6 ...
##  $ homeownership: num  77.5 76.7 68 82.9 82 76.9 69 70.7 71.4 77.5 ...
##  $ multiunit    : num  7.2 22.6 11.1 6.6 3.7 9.9 13.7 14.3 8.7 4.3 ...
##  $ income       : num  24568 26469 15875 19918 21070 ...
##  $ med_income   : num  53255 50147 33219 41770 45549 ...
##  $ smoking_ban  : chr  "none" "none" "none" "none" ...
```

Exercise 1.3

Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

ANSWER :

- **Number of siblings** = mumerical, discrete.
- **Student height** = numerical, continuous.
- **Statistics course (Y/N)** = categorical.

Exercise 1.4

Consider the variables group and outcome (at 30 days) from the stent study in Section 1.1. Are these numerical or categorical variables?

ANSWER : Categorical.

```
setwd("/Volumes/E-Books and articles/e-Books & articles/R/OpenIntro Statistics/openintroData")
load("stent.rda")
s30 <- as.data.frame(rbind(table((split(stent, stent$period)[[1]])[, 1:2]),
    colSums(table((split(stent, stent$period)[[1]])[, 1:2]))))
rownames(s30) <- c("control", "treatment", "TOTAL")
library(pander)
pander(s30, "Events - 0 to 30 days period", emphasize.strong.rows = 3)
```

Table 7: Events - 0 to 30 days period

	no event	stroke
control	214	13
treatment	191	33
TOTAL	**405**	**46**

```
pander(head(cbind(entry = 1:nrow(stent), stent), 5), "Head of stent data")
```

Table 8: Head of stent data

entry	group	outcome	period
1	treatment	stroke	0-30 days
2	treatment	stroke	0-30 days
3	treatment	stroke	0-30 days
4	treatment	stroke	0-30 days
5	treatment	stroke	0-30 days