OpenIntro Statistics

Rodrigo Martins

March 29, 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.

```r
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**

```r
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 :

```r
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
Washington County: 30	Texas : 254	Min. : 67	Min. : 82	Min. : 0.000
Jefferson County : 25	Georgia : 159	1st Qu.: 11210	1st Qu.: 11104	1st Qu.: 6.964
Franklin County : 24	Virginia: 134	Median : 24608	Median : 25857	Median : 8.669
Jackson County : 23	Kentucky: 120	Mean : 89623	Mean : 98233	Mean : 9.991
Lincoln County : 23	Missouri: 115	3rd Qu.: 61766	3rd Qu.: 66699	3rd Qu.: 10.857
Madison County : 19	Kansas : 105	Max. :9519338	Max. :9818605	Max. :204.616
(Other) :2999	(Other) :2256	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	comprehensive: 524
1st Qu.:11.0	1st Qu.:69.50	1st Qu.: 6.10	1st Qu.:19030	1st Qu.: 36952	none :1911

poverty	homeownership	multiunit	income	med_income	smoking_ban
Median :14.7	Median :74.60	Median : 9.70	Median :21773	Median : 42445	partial : 681
Mean :15.5	Mean :73.26	Mean :12.33	Mean :22505	Mean : 44270	NA's : 27
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

```r
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`

```r
str(county)
```

```
## 'data.frame':    3143 obs. of  11 variables:
##  $ name         : Factor w/ 1877 levels "Abbeville County",..: 83 90 101 151 166 227 237 250 298 320 ...
##  $ state        : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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  : Factor w/ 3 levels "comprehensive",..: 2 2 2 2 2 2 2 3 2 1 ...
```

```r
plot(fed_spend ~ poverty, county, xlim = c(0, 55), ylim = c(0, 32), xlab = "Poverty Rate (Percent)",
    ylab = "Federal Spending Per Capita", las = 1, yaxt = "n", pch = 19, col = "lightblue",
    cex = 0.8)
axis(2, at = c(0, 10, 20, 30), las = 2)
text(45, 33, "32 countries with higher\nfederal spending are not shown", pos = 1)
points(fed_spend ~ poverty, county, type = "p", pch = 15, cex = 0.3)
points(fed_spend[county$name == "Owsley County"] ~ poverty[county$name == "Owsley County"],
    county, type = "p", pch = 1, cex = 1.8, col = "red", lwd = 2)
segments(x0 = county$poverty[county$name == "Owsley County"], y0 = -1, x1 = county$poverty[county$name ==
    "Owsley County"], y1 = county$fed_spend[county$name == "Owsley County"] -
    0.8, lty = 2, col = "red")
segments(x0 = -2, y0 = county$fed_spend[county$name == "Owsley County"], x1 = county$poverty[county$name ==
    "Owsley County"] - 0.8, y1 = county$fed_spend[county$name == "Owsley County"],
    lty = 2, col = "red")
```

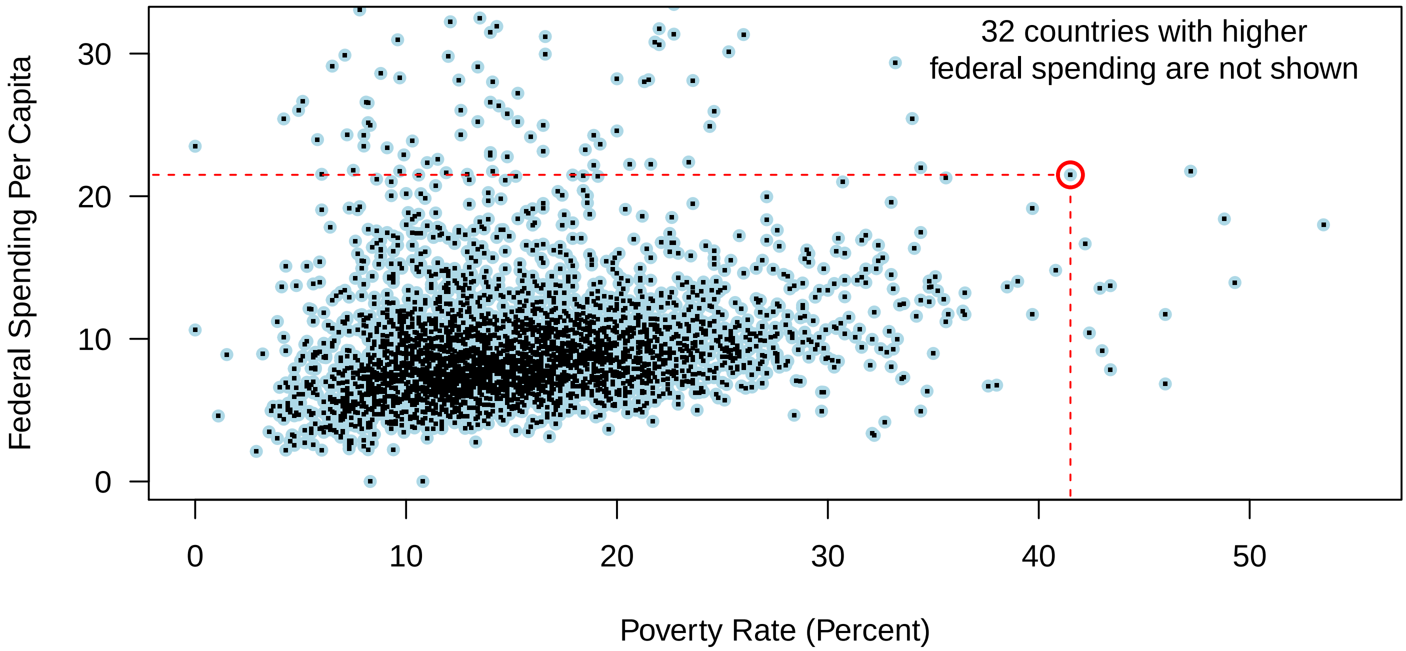

Figure 1: A scatterplot showing fed spend against poverty. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

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[, 1:2]), 5), "Head of stent data")
```

Table 8: Head of stent data

entry	group	outcome
1	treatment	stroke
2	treatment	stroke
3	treatment	stroke
4	treatment	stroke
5	treatment	stroke

```
plot(homeownership ~ multiunit, county, ylim = c(0, 90), xlab = "Percent of Units in Multi-Unit Structures",
    ylab = "Percent of Homeownership", las = 1, xaxt = "n", yaxt = "n", pch = 19,
    col = "lightblue", cex = 0.8)
axis(1, at = c(0, 20, 40, 60, 80, 100), lab = paste0(seq(0, 100, by = 20), "%"),
    las = 1)
axis(2, at = c(0, 20, 40, 60, 80), lab = paste0(seq(0, 80, by = 20), "%"), las = 2)
points(homeownership ~ multiunit, county, type = "p", pch = 15, cex = 0.3)
```

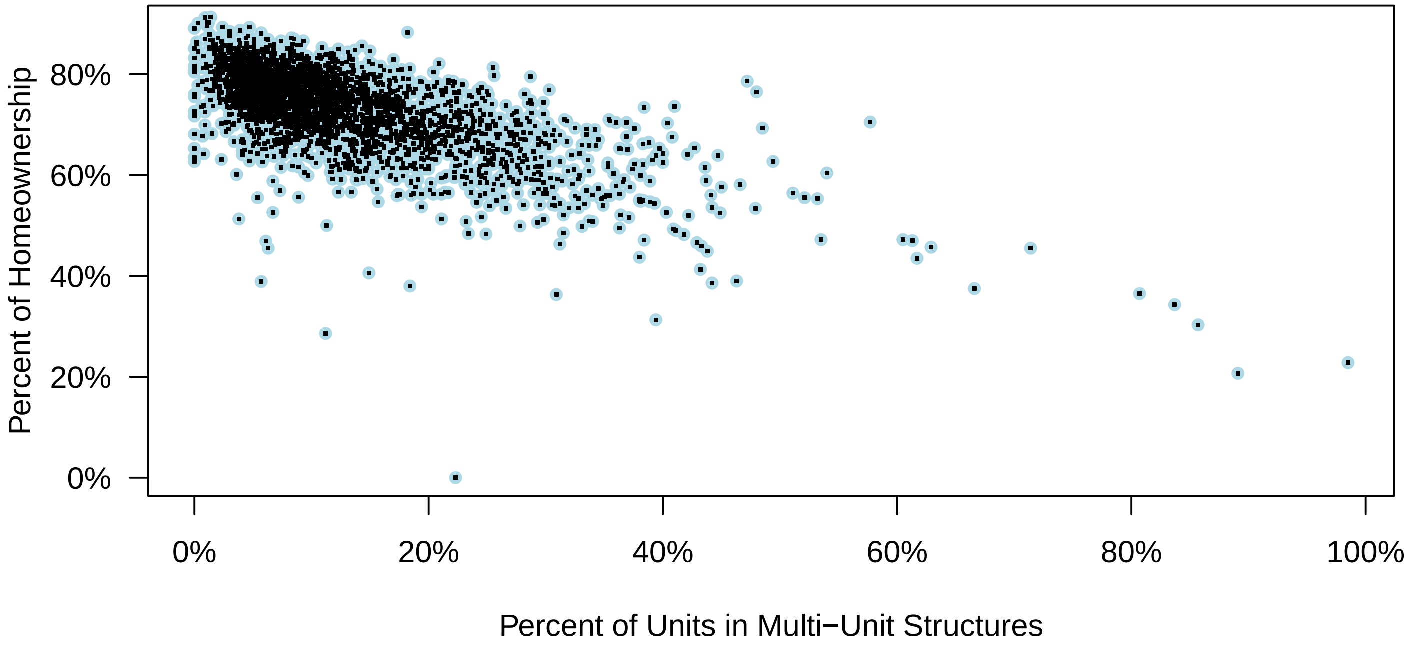

Figure 2: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties.

Exercise 1.5

Examine the variables in the `email50` data set, which are described in Table 1.4 on page 4. Create two questions about the relationships between these variables that are of interest to you.

- Is the number of destinataries (*to_multiple*) related to *spam* messages ?
- Is the number of dollar signs (*dollar*) related to *spam* messages ?

```
setwd("/Volumes/E-Books and articles/e-Books & articles/R/OpenIntro Statistics/openintroData")
load('email50.rda')
library(pander)
panderOptions("table.split.table", 120)
panderOptions('keep.trailing.zeros', TRUE)
pander(head(email50,5),"Head - email50 data set")
```

Table 9: Head - email50 data set (continued below)

spam	to_multiple	from	cc	sent_email	time	image	attach	dollar
0	0	1	0	1	2012-01-04 05:19:16	0	0	0
0	0	1	0	0	2012-02-16 12:10:06	0	0	0
1	0	1	4	0	2012-01-04 07:36:23	0	2	0
0	0	1	0	0	2012-01-04 09:49:52	0	0	0
0	0	1	0	0	2012-01-27 01:34:45	0	0	9

Table 10: Table continues below

winner	inherit	viagra	password	num_char	line_breaks	format	re_subj
no	0	0	0	21705	551	1	1
no	0	0	0	7011	183	1	0
no	0	0	0	631	28	0	0
no	0	0	0	2454	61	0	0
no	0	0	1	41623	1088	1	0

exclaim_subj	urgent_subj	exclaim_mess	number
0	0	8	small
0	0	1	big
0	0	2	none
0	0	1	small
0	0	43	small