Chapter 2

Displaying and Describing Categorical Data

Data from categorical variable are described in both in graphical and tabular form. Data for *categorical* variable organized into one of several groups (categories) and can only be counted. Bar graphs (Pareto charts), pie charts and line graphs are discussed in this chapter. *Contingency* tables and *Simpson's Paradox* are discussed.

2.1 Summarizing a Categorical Variable

Exercise 2.1 (Summarizing a Categorical Variable)

1. *Company stocks.* Consider types of stocks (A, B or C) for small and large companies purchased in years 2010, 2011, 2012, 2013 or 2014.

company	stock	year	company	stock	year
small	A	2010	large	C	2011
small	B	2010	small	C	2010
small	C	2010	large	B	2010
large	B	2014	small	A	2013
small	B	2010	small	A	2013
small	B	2012	small	B	2013
large	B	2010	small	B	2010
large	A	2012	large	C	2010
large	C	2012	large	B	2014
large	C	2010	large	A	2010

Import chapter2.company.stock.size text file into R. Use R Studio Environment panel, click on Import Dataset, then Find local file. Then type following R script in the Console:

```
> data <- chapter2.company.stock.size  # shortens up file name to "data"
> attach(data) # makes data current working dataframe
> head(data) # if first five cases correct, indication of correct data
```

Fill in the blanks.

company	counts	proportions
large	______	______
small	______	______
total	______	______

```
> table(company) # counts for large and small companies
> sum(as.vector(table(company))) # convert table to vector, then sum for total
> prop.table(as.vector(table(company))) # convert table to vector, then find proportion
```

stock	Frequency	Relative Frequency
A	______	______
B	______	______
C	______	______
total	______	______

```
> table(stock); sum(as.vector(table(stock))); prop.table(as.vector(table(stock)))
```

year →	2010	2011	2012	2013	2014	total
count	______	______	______	______	______	______
percentage	______	______	______	______	______	______

```
> table(year); sum(as.vector(table(year))); 100*prop.table(as.vector(table(year)))
```

2. *Age distribution comparison.* Age distribution of a random sample of 463 people living in Uppsala, a city in Sweden, is compared to age distribution to *all* of Sweden, where, notice percentages, not counts, are given for Sweden population.

age	Uppsala	Sweden
under 5	47	6.7%
5 to 16	75	14.1%
16 to 65	296	69.5%
over 65	45	9.7%
total	463	100%

Import chapter2.age.distribution text file into R. Use R Studio Environment panel, click on Import Dataset, then Find local file. Then type following R script in the Console:

```
> data <- chapter2.age.distribution; attach(data); head(data)
```

Fill in the blanks.

age →	under 5	5 to 16	16 to 65	over 65	total
Uppsala count	______	______	______	______	______
percentage	______	______	______	______	______

```
> options(digits = 2) # restricts output to 2 digits of accuracy
> 100*prop.table(Uppsala)
```

What would the age distribution be for Uppsala if the age distribution in this town matched the age distribution of all of Sweden?

age →	under 5	5 to 16	16 to 65	over 65	total
Uppsala (using Sweden %) count	______	______	______	______	______
percentage	______	______	______	______	______

```
> 463*prop.table(Sweden) # count if Uppsala matches Sweden age distribution
```

2.2 Displaying a Categorical Variable

Bar, Pareto and pie charts are discussed in this section.

Exercise 2.2 (Displaying a Categorical Variable)

1. *Patient health costs.* Sample of twenty patients costs, where "great" means small annual health costs and "bad" means higher average annual health costs, are listed below. Distribution table, bar graph, Pareto chart and pie charts for data given below.

costs	number of patients	proportion of patients
bad	2	$\frac{2}{20} = 0.10$
poor	4	$\frac{4}{20} = 0.20$
fair	5	$\frac{5}{20} = 0.25$
good	8	$\frac{8}{20} = 0.40$
great	1	$\frac{1}{20} = 0.05$
total	20	1.0

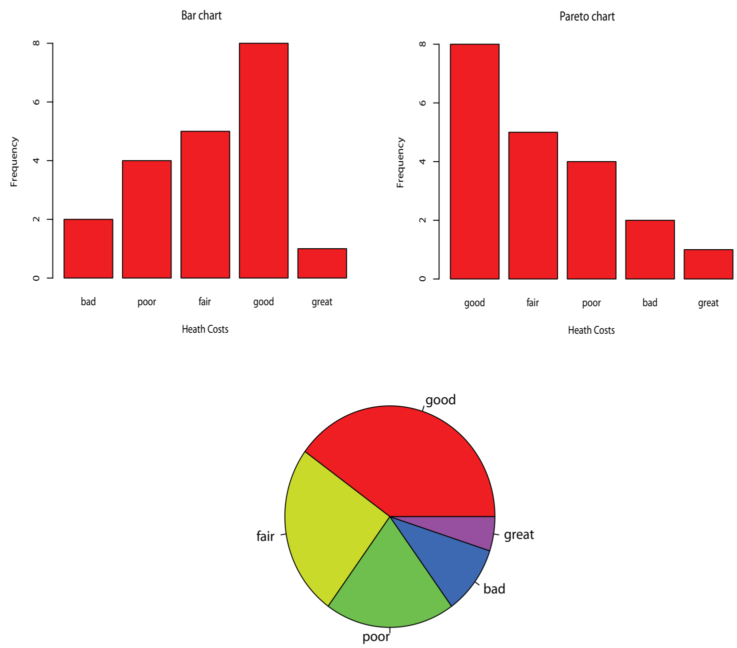

Figure 2.1 (Bar, Pareto and pie charts for health costs)

```
> data <- chapter2.health.costs; attach(data); head(data)
> barplot(frequency, main = "Bar chart", xlab="Heath Costs", ylab="Frequency", col="red", names.arg=costs)
> pareto.data <- data[rev(order(frequency)),]; attach(pareto.data) # Pareto order
> barplot(frequency, main="Pareto chart", xlab="Heath Costs", ylab="Frequency", col="red", names.arg=costs)
> pie(frequency,col=rainbow(5),labels=as.character(costs))
```

(a) This data is **categorical** / **quantitative** because data grouped into five categories: bad, poor, fair, good and great.

(b) Of 20 patients, **2** / **4** / **5** / **8** are in fair health or a proportion of $\frac{5}{20} = 0.25$.

(c) Height of each vertical bar in bar graph corresponds to frequency for each category. For example, vertical bar for "good" category has a height (or proportion) of (choose one) **5** / **8** / **9**.

(d) Adding heights of all vertical bars in five categories together, we get (choose one) **8** / **15** / **20**.

(e) Pareto chart is a bar graph where bars are arranged left to right in **decreasing** / **increasing** order.

(f) **True** / **False**. Another possible variation of a bar graph has *proportion* rather than frequency along y-axis. Heights of this version of bar graph *do* necessarily add to one.

(g) **True / False** *Width* of each vertical bar has *no* meaning.

(h) Angle spanned by each wedge in pie chart is **smaller than / in proportion to / larger than** size of category. Wedges *must* add to a "whole" in pie chart since wedge angles add to 360^o; *all* data must be included.

2. *Graphical misrepresentations: unequal widths.*

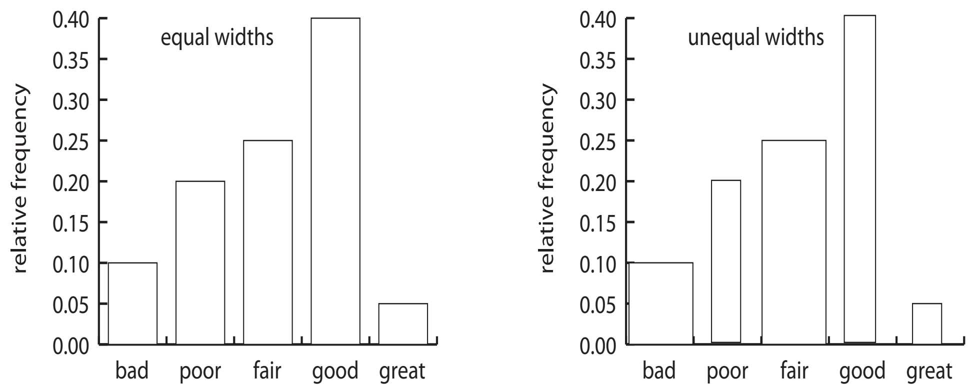

Figure 2.2 (Graphical misrepresentations: unequal widths)

Bar graph on right possibly misleading because it seems "bad" and "fair" health occur **less frequently than / as frequently as / more frequently than** other categories.

3. *Graphical misrepresentations: truncated and adjusted scale.*

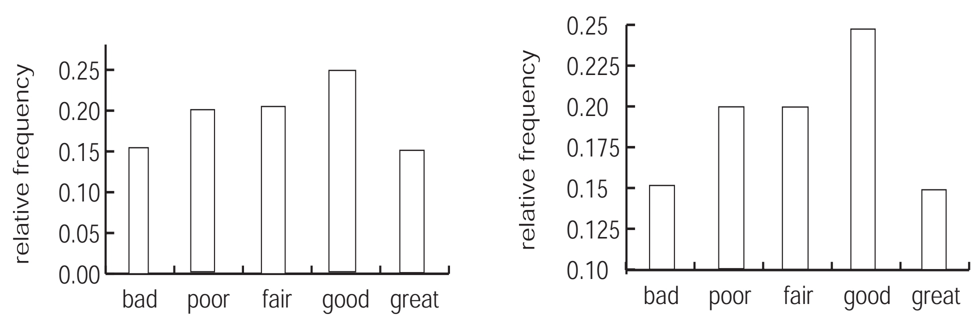

Figure 2.3 (Graphical misrepresentations: truncated and adjusted scale)

Bar graph on right possibly misleading because it seems **greater / same / lesser** difference between categories.

2.3 Exploring Two Categorical Variables: Contingency Tables

We look at contingency tables to determine the association of paired qualitative data. We look at marginal distributions, conditional distributions and bar graphs. We also discuss Simpson's Paradox, analogous to lurking variables in paired quantitative data.

Exercise 2.3 (Exploring Two Categorical Variables: Contingency Tables)

1. *Fathers, sons and college.* Data from a sample of 80 families in a midwestern city gives record of college attendance by fathers and their oldest sons.

	college	no college
father	18	7
son	22	33

(a) *Marginal distributions.* Fill in the blanks.

	college	no college	
father	18	7	______
son	22	33	______

(Marginal) distribution of father-son is $\mathbf{(25, 55)}$ / $\mathbf{(40, 40)}$.

```
> data <- chapter2.father.son.table; attach(data); head(data)
> data.matrix <- as.matrix(data[,2:3]) # convert data frame to useable matrix
> dimnames(data.matrix) <- list(data$X,c("college","no.college")); data.matrix
> margin.table(data.matrix, 1) # row totals
```

	college	no college
father	18	7
son	22	33
	______	______

(Marginal) distribution of college attendance is $\mathbf{(25, 55)}$ / $\mathbf{(40, 40)}$.

```
> margin.table(data.matrix, 2) # column totals
```

(b) *Conditional distributions.*
Complete proportion of row totals table: condition on father or son.

	college	no college	row totals
father	$\frac{18}{25}$ = ______	$\frac{7}{25} = 0.28$	25 $(\frac{25}{25} = 1)$
son	$\frac{22}{55}$ = ______	$\frac{33}{55} = 0.6$	55 $(\frac{55}{55} = 1)$

Percent of fathers that attend college is **72%** / **28%**.
Conditional distribution of college attendance or not for father in this study is $\mathbf{(0.72, 0.28)}$ / $\mathbf{(0.4, 0.6)}$.

```
> prop1 <- prop.table(data.matrix, 1); prop1 # proportion of the row totals
```

Complete proportion of column totals table: condition on college attendance.

	college	no college
father	$\frac{18}{40} =$ ______	$\frac{7}{40} = 0.175$
son	$\frac{22}{40} =$ ______	$\frac{33}{40} = 0.875$
column totals	$40\ (\frac{40}{40} = 1)$	$40\ (\frac{40}{40} = 1)$

Percent of college students who are fathers is **45%** / **55%**.
Conditional distribution of father or son for college attendance is
$\mathbf{(0.45, 0.55)}$ / $\mathbf{(0.175, 0.875)}$.

```
> prop2 <- prop.table(data.matrix, 2); prop2 # proportion of the column totals
```

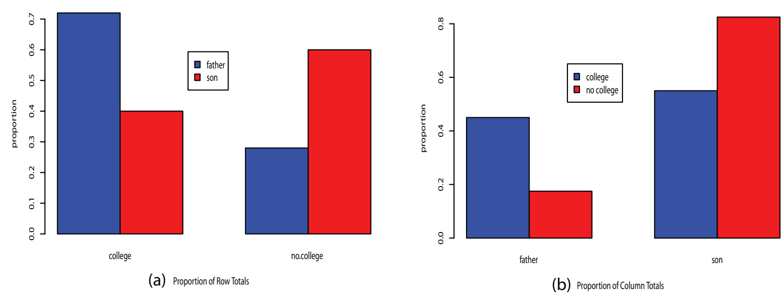

Figure 2.4 (Split bar plots for conditional distributions.)

```
> barplot(prop1,col=c("blue","red"), beside=T, ylab="proportion") # proportion of row totals matrix
> legend("topleft",c("father","son"),fill=c("blue","red")) # click in plot to locate son/father legend
> barplot(t(prop2),col=c("blue","red"), beside=T, ylab="proportion") # transpose prop of column totals matrix
> legend("topleft",c("college","no college"),fill=c("blue","red")) # click in plot to locate college/not legend
```

According to proportion of row totals split plot (a),
a **greater** / **lesser** proportion of fathers than sons attend college
indicating there appears to be **an** / **no** association:
sons attend college if fathers do not attend college.

From plot (b),
a **greater** / **lesser** proportion of college students are fathers.

Proportion of grand totals table:

	college	no college	row totals
father	$\frac{18}{80} =$ ______	$\frac{7}{80} =$ ______	$25\ (\frac{25}{80} = 1)$
son	$\frac{22}{80} =$ ______	$\frac{33}{80} =$ ______	$55\ (\frac{55}{80} = 1)$
column totals	$40\ (\frac{40}{80} = 0.5)$	$40\ (\frac{40}{80} = 0.5)$	$80\ (\frac{80}{80} = 1)$

Percent of all people who are fathers attending college is **22.5%** / **55%**.

```
> prop2 <- prop.table(data.matrix) # proportion of the grand total
```

2. *Contingency table: association between drug, flu symptoms and gender lurking variable.* Are flu symptoms influenced by drug?

flu symptoms →	reduced	not reduced	totals
drug	100	50	150
no drug	200	100	300
totals	300	150	450

```
> data <- chapter2.flu.drug; attach(data); head(data)
> data.matrix <- as.matrix(data[,2:3]) # convert data frame to useable matrix
> dimnames(data.matrix) <- list(data$X,c("flu better","flu worse")); data.matrix
```

(a) *Flu symptoms conditional on drug distribution.*
Complete conditional table.

flu symptoms →	reduced	not reduced	
drug	$\frac{100}{150}$ = ______	$\frac{50}{150} = 0.33$	$\frac{150}{150}$ = ______
no drug	$\frac{200}{300}$ = ______	$\frac{100}{300} = 0.33$	$\frac{300}{300} = 1$
	$\frac{300}{450}$ = ______	$\frac{150}{450} = 0.33$	$\frac{450}{450} = 1$

```
> prop1 <- prop.table(data.matrix, 1); prop1 # proportion of the row totals
```

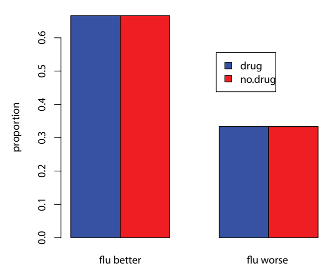

Figure 2.5 (Bar graph: flu symptoms conditional on drug.)

```
> barplot(prop1,col=c("blue","red"), beside=T, ylab="proportion") # proportion of row totals matrix
> legend(locator(1),c("drug","no drug"),fill=c("blue","red")) # click in plot to locate drug/no.drug legend
```

There is (choose one) **an** / **no** association:
flu symptoms same whether drug given or not.

(b) *Lurking variable: gender.* Doctors suspect gender is confounding results. Consequently, *to control for gender*, they tabulate effect of drug on males and, separate from this, tabulate effect of drug on females.

male	reduced	not reduced	subtotals
drug	80	40	120
no drug	100	80	180
subtotals	180	120	300

female	reduced	not reduced	subtotals
drug	20	10	30
no drug	100	20	120
subtotals	120	30	150

Complete conditional table for both males and females.

males	reduced	not reduced	subtotals
drug	$\frac{80}{120}=$ ______	$\frac{40}{120}=$ ______	$\frac{120}{120}=$ ______
no drug	$\frac{100}{180}=0.55$	$\frac{80}{180}=0.44$	$\frac{180}{180}=$ ______
subtotals	$\frac{180}{300}=0.6$	$\frac{120}{300}=0.4$	300 $\frac{300}{300}=1$

females	reduced	not reduced	subtotals
drug	$\frac{20}{30}=$ ______	$\frac{10}{30}=$ ______	$\frac{30}{30}=$ ______
no drug	$\frac{100}{120}=0.83$	$\frac{20}{120}=0.17$	$\frac{120}{120}=$ ______
subtotals	$\frac{120}{150}=0.8$	$\frac{30}{150}=0.2$	$\frac{150}{150}=1$

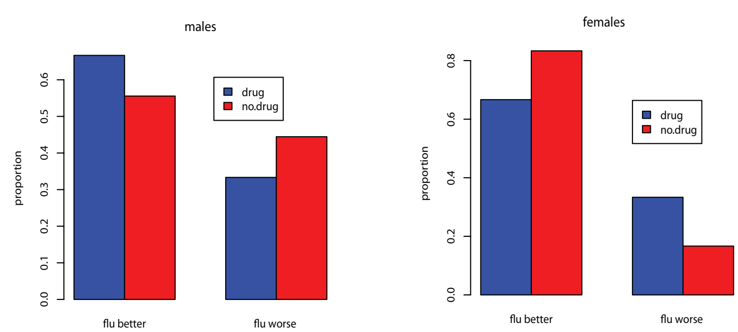

Figure 2.6 (Bar graph: flu conditional on drug, males/females.)

```
> data <- chapter2.flu.drug.male; attach(data); head(data)
> data.matrix <- as.matrix(data[,2:3]) # convert data frame to useable matrix
> dimnames(data.matrix) <- list(data$X,c("flu better","flu worse")); data.matrix
> prop1 <- prop.table(data.matrix, 1); prop1 # proportion of the row totals
> barplot(prop1,col=c("blue","red"), beside=T, ylab="proportion", main="male") # proportion of row totals matrix
> legend(locator(1),c("drug","no drug"),fill=c("blue","red")) # click in plot to locate drug/no.drug legend
>
> data <- chapter2.flu.drug.female; attach(data); head(data)
> data.matrix <- as.matrix(data[,2:3]) # convert data frame to useable matrix
> dimnames(data.matrix) <- list(data$X,c("flu better","flu worse")); data.matrix
> prop1 <- prop.table(data.matrix, 1); prop1 # proportion of the row totals
> barplot(prop1,col=c("blue","red"), beside=T, ylab="proportion", main="female") # proportion of row totals matrix
> legend(locator(1),c("drug","no drug"),fill=c("blue","red")) # click in plot to locate drug/no.drug legend
```

There is (choose one) **an** / **no** association for *males*:
more likely flu symptoms reduced when taking drug than not taking drug.
There is (choose one) **an** / **no** association for *females*:
less likely flu symptoms reduced when taking drug than not taking drug.

(c) **True / False** Although combined study demonstrates *no* association between drug and reduced flu symptoms, a positive association between drug and reduced flu symptoms occurs for males, whereas a negative association between drug and reduced flu symptoms occurs for females. This is an example of *Simpson's Paradox* where association changes with introduction of third (lurking) variable.

3. *More contingency tables: company stocks.*
Consider types of stocks (A, B or C) for small and large companies and for different years.

company	stock	year	company	stock	year
small	A	2010	large	C	2011
small	B	2010	small	C	2010
small	C	2010	large	B	2010
large	B	2014	small	A	2013
small	B	2010	small	A	2013
small	B	2012	small	B	2013
large	B	2010	small	B	2010
large	A	2012	large	C	2010
large	C	2012	large	B	2014
large	C	2010	large	A	2010

```
> data <- chapter2.company.stock.size; attach(data); head(data)
```

Fill in blanks: number of stock type for both large and small companies.

O_i	stock type →	A	B	C	row totals
company	large	2	_____	_____	10
	small	_____	_____	_____	10
	column totals	5	9	6	20

```
> data.table <-table(company,stock); data.table
```

Fill in blanks: calculate contingency table of stock type versus company size (divide by company (row) totals).

O_i	stock type →	A	B	C	row totals
company	large	0.2	_____	_____	1
	small	_____	_____	_____	1
	column totals	0.5	0.9	0.6	20

```
> prop1 <- prop.table(data.table, 1); prop1 # proportion of the row totals
```

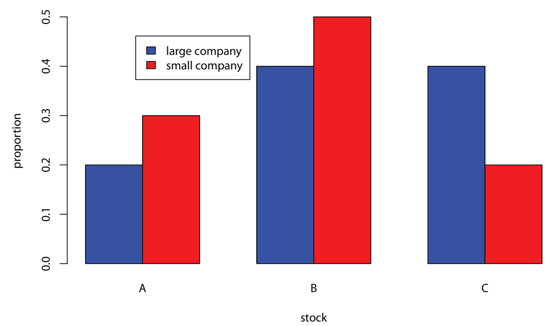

Figure 2.7 (Company size given stock type.)

```
> barplot(prop1,col=c("blue","red"), beside=T, ylab="proportion", xlab="stock") # proportion of row totals matrix
> legend(locator(1),c("large company","small company"),fill=c("blue","red")) # click in plot to locate small or large com
```

There is (choose one) **an** / **no** association:
stock types different for different size companies.

Percent of large companies who buy stock B
20% / **30%** / **40%**.

```
> prop1 <- prop.table(data.table, 1); prop1 # proportion of the row totals
```

Percent of stock B bought by large companies
44% / **56%**.

```
> prop2 <- prop.table(data.table, 2); prop2 # proportion of the column totals
```

Percent of all transactions which were stock B bought by large companies
20% / **30%** / **40%**.

```
> prop <- prop.table(data.table); prop # proportion of the grand total
```

A segmented bar chart (or spine plot) could also be used here.

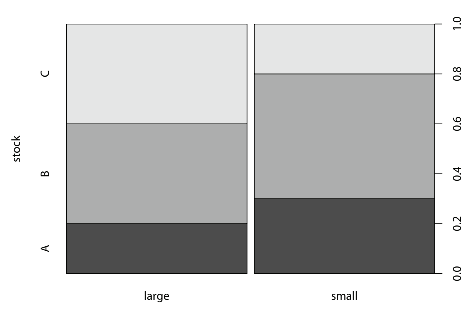

Figure 2.8 (Segmented bar chart.)

```
> spineplot(company,stock)
```

A mosiac plot could also be used here.

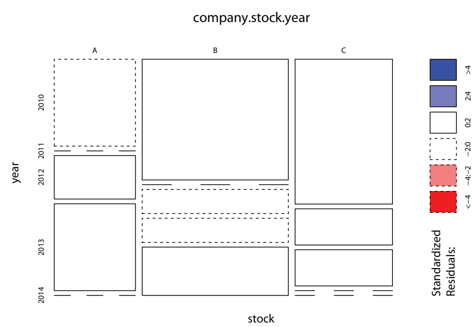

Figure 2.9 (Mosaic plot.)

```
> company.stock.year <-table(stock,year); company.stock.year
> mosaicplot(company.stock.year, shade=TRUE)
```

All white rectangles, no red or blue rectangles, in the mosaic plot indicates there are no outlying cell counts in this contingency table, that all counts are relatively the same as one another. Also notice the mosaic plot acts like a segmented bar charts but with the additional feature of proportional in *both* x and y direction; in this case, in both year and stock type.

2.4 Segmented Bar Graphs and Mosaic Plots

Covered in previous sections.

2.5 Simpson's Paradox

Covered in previous sections.