

Elementary Statistics

Brian E. Blank

January 17, 2018

FIRST PRESIDENT UNIVERSITY PRESS

Preface

These notes are amplifications of lectures for an elementary statistics course, M2200, that the author gave at First President University in the Fall 2014 and Spring 2015 semesters. Additional changes were incorporated during the Spring semesters of 2016 and 2017. Graphics, tables, and certain other material not presented in the lectures are included, but, by-and-large, the notes are the actual lecture notes in typeset form.

The topics that are covered here are those that are found in standard college-level statistics textbooks that are not calculus-based. In point of fact, the course on which these notes are based *does* list Calculus I (M131 at First President University) as a prerequisite. The material of Calculus I is mainly devoted to differential calculus, and we have no essential need for any of the topics in differential calculus. However, at First President University the basic ideas of integration up to and including the Fundamental Theorem of Calculus are usually also included in Calculus I. In a few places in these notes, there are references to integration. Some familiarity with integration will probably be helpful for obtaining a deeper understanding of the theory of some of the topics in elementary statistics. However, none of the suggested exercises in these lecture notes and none of the questions found in the exams for the course on which these lecture notes are based require calculus.

As these notes will emphasize, the subject of statistics often does not have a standard, unique definition for an important measurement or concept. Readers of these notes should be aware that the conventions adopted in these notes do not (indeed, cannot) always coincide with common conventions (because if they agree with some common conventions, then they will disagree with other common conventions). For example, the author adopts the definitions of *quartiles* used in the popular statistics program R. Different definitions are wired into the TI-83 calculator. In the course on which these notes are based, exam questions require the definitions used in these notes. Asked one student who got a wrong answer supplied by her TI-83, “Why don’t you use the definition that the TI-83 uses?” The answer is that the author uses R and does not use the TI-83 (or any other calculator). It is hoped that all students in the class will become experts in statistics who write their own sets of lecture notes. Such students can then adopt their own house rules.

As the reader may infer from the preceding paragraph, the author is not a devotee of calculators. That said, the practice of statistics is saturated with mundane calculations that are in no way edifying. The author believes that it is beneficial to perform each type of calculation a few times with low level technology (such as a basic scientific calculator that is not statistics-enabled). Higher level technology should be used only after the experience of hand-to-hand combat with the calculations.

Professional statisticians do not use calculators. Why should you? A number of software packages are available—*Minitab*, *R*, *SAS*, *STATA*, and *SPSS* all have their aficionados. Of these, *R* has a substantial advantage: it is *free*. Throughout these notes there are subsections that illustrate basic *R* usage. Although *R* has a well-deserved reputation of being difficult to master, it is actually quite easy to learn how to do the basic chores of statistics by using *R*’s built-in functions interactively. If you perceive an enduring need for statistical analysis in your future, then you should not let this opportunity to simultaneously learn elementary statistics and basic *R* functionality slip by.

Several subsections in these lecture notes are marked “Optional”. You may skip over all of these optional subsections. Doing so will not compromise either your understanding of any material that follows or your

ability to answer any of the exam or homework questions put to you.

Each chapter concludes with a section of exercises. Many of these exercises are taken from past exams at First President University. Every exercise is fully solved.

These notes are under constant revision. Check frequently for updates in which errors have been corrected and minor improvements and clarifications have been made. The author would appreciate feedback alerting him to errors (even minor typos) and discussions that are less than clear. They will be corrected or improved in subsequent postings.

The URL for the original posting of these notes is <http://math.wustl.edu/~brian/stats/2200-01.pdf>. The author retains copyright. Permission is ***not*** granted for posting this file at any other URL. Uploading this file to Course Hero or any similar site is expressly ***forbidden***.

Chapter 1. Data—Categorical

1.1 Tables—Variables and Cases

Statistics is the science of drawing conclusions from data. The practice of Statistics includes not only the analysis of data that have already been obtained, but also the design of studies and experiments that yield the data that will be analyzed. A large part of Statistics consists of distinguishing events that were unlikely to have occurred “by chance” from those that might reasonably have occurred by chance.

In statistics, the term *population* refers to all members of a set that will be studied. A population might consist of people (as in, for example, a drug trial), but it may not. Events, such as all accidents on the New Jersey Turnpike in 2013, or places, such as all towns within 50 km of a nuclear reactor, might constitute the population of a particular study. A *sample* is a subset of a population. In general, data is collected from a sample in order to draw inferences about the entire population.

The data that we will analyze in this course will typically be organized into two-dimensional tables composed of vertical *columns* and horizontal *rows*. Each column corresponds to a *variable*. The name of the variable will usually appear above the column. Each row corresponds to a *case*. One case of a company’s customer records is shown here:

Name	Customer Number	Telephone	Zip	Age Group	Sex	Last Purchase	Purchases Yr to date
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Vasquez, José	723605	310-555-0100	90210	26-40	M	46.95	82.94
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1.1.1: Customer Database

The data “Vasquez, José”, 310-555-0100, 90210, and “M” (for *male*)—to cite just four of the row entries—are *values* of the variables *Name*, *Telephone*, *Zip*, and *Sex* respectively. Notice that data is not necessarily numerical. The variables *Name* and *Sex* provide obvious examples of variables with non-numerical values. The first of these, *Name*, is, for the most part, an *identifier variable*. Its values are of no real significance to the company. Had José’s parents named him Jerome instead, the line of the table would have had a different first entry but would have had the same information of interest to the company. In other words, the *Name* variable serves as a label for the line of data in which it is found. In this case, the variable *Name* may actually play more of a role than identifier. For example, the company might generate a promotional email addressed to Mr. Vasquez by using the part of the value of *Name* that precedes the comma.

Further to the observation that data need not be numerical, notice that even the *numbers* that appear in a table are not always numerical data. Customer numbers, telephone numbers, and zip codes are examples

of such data. They are numbers, but they are not used as numbers. Would you ever calculate the sum of two telephone numbers or subtract one zip code from another? If you compare two forwards in hockey, do you derive any useful information knowing that the player number of one is greater¹ than the number of the other? Although the values of the *Customer Number* variable are numbers, *Customer Number* is actually an identifier variable (and does that job better than the *Name* variable because customers with the same name can be assigned unique customer numbers.) A numerical variable is sometimes called a **quantitative variable**. As we have remarked, variables with values that are numbers are *not* necessarily quantitative variables.

The variables *Zip*, *Age Group*, and *Sex* are called **categorical variables** because their values are categories. Other terms for such variables are **qualitative variables** and **nominal variables**. A categorical variable might have numbers as values. For example, each position in baseball is assigned a counting number from 1 to 9 (1 for the pitcher, 2 for the catcher, 3 for the first baseman, and so on). In a table of baseball statistics, we might choose to use these numbers as values of the *Position variable* rather than the names of the positions.

It should be noted that the terms “case” and “value” have different meanings. The distinction is clear in Table 1.1.1 because eight variables are tabulated: as a result, there are eight values for each case. However, even when only one variable *Y* is under consideration, there is a distinction between a case and a value. In particular, there is no connection between the number of cases and the number of (different) values of a variable. For example, no matter how many cases in the data set tabulated in Table 1.1.1, the variable *Sex* has only two possible values. In the other direction, the number of cases in the Social Security Administration’s database is, by design, less than the number of possible values of the *Social Security Number* variable. It should also be noted that there is a distinction between the possible values of a variable and the values actually attained. For example, in a study of the success of a cancer treatment, the variable *Five Year Survival* has two possible values: *Yes* and *No*. If the treatment is particularly effective, then only the value *Yes* might be observed among the cases of the study.

Descriptive, Inferential, and Summary Statistics

A *descriptive statistic* is a number that is used to describe an aspect of a collection of observations. For example, the U.S. Fish and Wildlife Service set traps in a program to control a troublesome coyote population. The traps snared other species as well. The kill rates for various species were recorded. These numbers are descriptive statistics.

Often *certain* descriptive statistics tell us pretty much everything of interest about a collection of information. By focusing on these statistics, we can avoid the noise generated by numbers of lesser importance. Such descriptive statistics are called *summary statistics* because they summarize.² In the coyote control program mentioned in the preceding paragraph, three summary statistics come to mind: 44,982 (the total number of animals that were killed in the program), 25,026 (the number of coyotes that were killed in the program), and 19,956 (the number of unintended animal casualties). You might have an interest in knowing that 2,698 opossums, 1,367 porcupines, and 3,345 skunks died in program to reduce the coyote population. But, if you are interested only in the big picture, then these individual species kill rates can be distracting details.

Descriptive statistics are can be informative, but very often they are only the first step in an application of statistics. Statistics that are used in conjunction with probability theory to draw an inference are called **inferential statistics**.

For an example of inferential statistics, consider the 2004 U.S. presidential campaign. One month before the election, both the Rasmussen and Gallup polls showed that John Kerry was trailing George W. Bush in the popular vote. In the Gallup poll, 1016 voters were surveyed and 498 expressed a preference for Kerry.

¹In addition to their arithmetic properties, numbers also have the important property of *order*: if two numbers are unequal, then one is greater than the other. In Chapter 4, we will encounter numerical variables whose values are used only for order and not for arithmetic. They are numerical variables nonetheless.

²Sometimes coming up with jargon is as easy as falling off a log: descriptive statistics describe and summary statistics summarize. Duh.

The sample proportion in favor of Kerry, namely $498/1016$, or 0.49 is certainly a descriptive statistic. But we can also use it to draw an inference.

According to the Rasmussen poll, only 45.9% of voters favored John Kerry. Could this sample proportion be consistent with Gallup's, or did one of the pollsters make an error?

The important point about a survey is that, if it is properly designed, then we gain correct information from it, but we cannot expect it to provide us with exactness. If a different group of 1016 voters had been selected, we could not have expected that *exactly* 498 voters in the poll would favor Kerry. Sample variability is to be expected. To settle the question of consistency, two other numbers can be calculated: $0.49 - (0.01568)(2.5758)$, or 0.4496 , and $0.49 + (0.01568)(2.5758)$, or 0.5304 . At this point it is not important to understand how these numbers are determined—we will revisit this example in Chapter 9. Their relevance is that with some understanding of probability theory (which we will gain in Chapters 7 and 8), we can be 99% sure that the true proportion of voters favoring Kerry was between 0.4496 and 0.5304 . Because the proportion obtained by Rasmussen was 0.459 , which is indeed between 0.4496 and 0.5304 , we infer that the two polls were consistent.

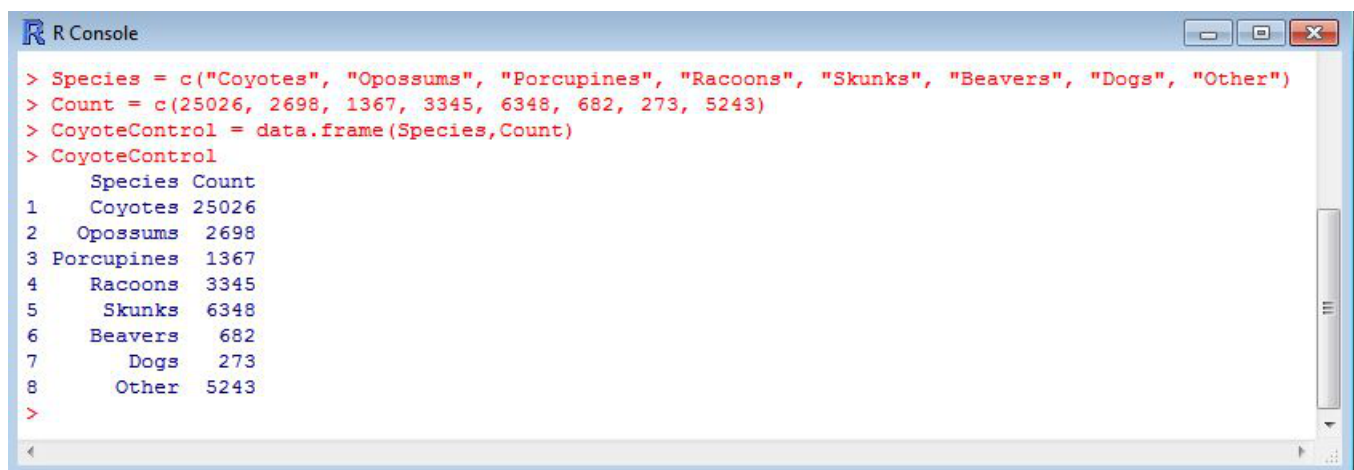
Ordered Data Structures in R

If you anticipate frequent use of statistics, then you should put aside your TI-83 and learn a statistical package. The leading freeware package is called R. Basic instruction in R will be provided in optional subsections of these notes.

A vector $\langle x_1, x_2, \dots, x_N \rangle$ in R is created by concatenating the entries using R's `c()` command. In the code that follows, vectors are created and assigned to the user-defined names `Species` and `Count`.

```
> Species = c("Coyotes", "Opossums", "Porcupines",
              "Racoons", "Skunks", "Beavers", "Dogs", "Other")
> Count = c(25026, 2698, 1367, 3345, 6348, 682, 273, 5243)
```

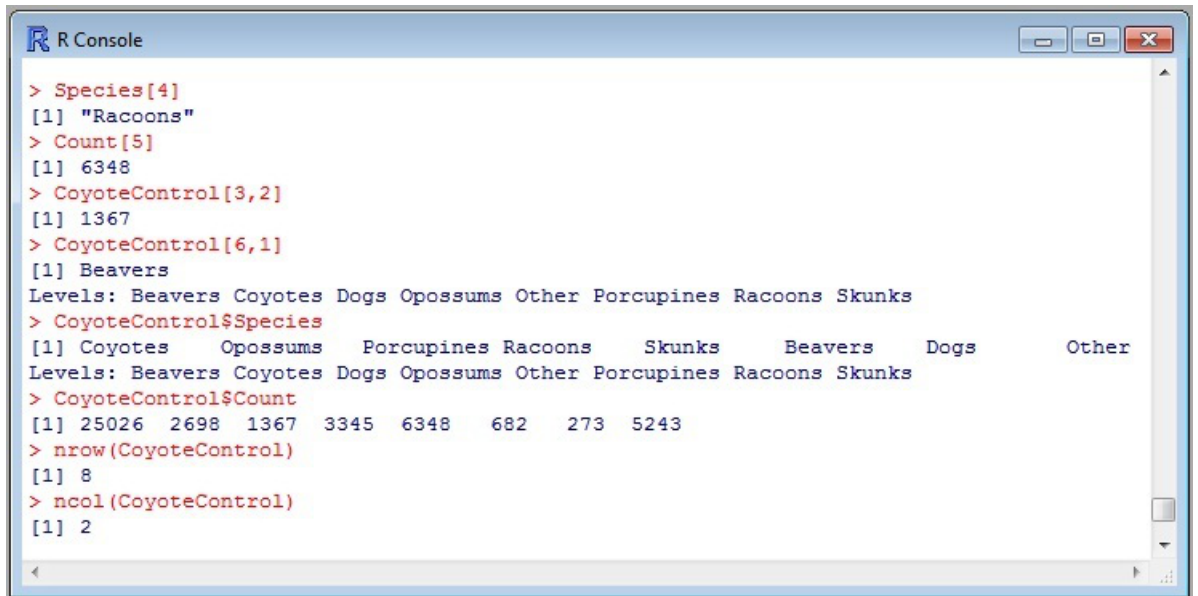
The first line creates a vector of character strings. The second line creates a numerical vector. In Figure 1.1.1, a screen capture shows how these vectors can be used to form a table, specifically, a *data frame* in R, using R's `data.frame` command. The red printing in the R session is user input; the blue is R's output. In this session, the first three lines are assignments, which are effected behind the scenes and do not elicit a response from R.



```
R Console
> Species = c("Coyotes", "Opossums", "Porcupines", "Racoons", "Skunks", "Beavers", "Dogs", "Other")
> Count = c(25026, 2698, 1367, 3345, 6348, 682, 273, 5243)
> CoyoteControl = data.frame(Species, Count)
> CoyoteControl
  Species Count
1  Coyotes 25026
2 Opossums  2698
3 Porcupines 1367
4  Racoons  3345
5   Skunks  6348
6  Beavers   682
7    Dogs   273
8   Other  5243
>
```


Figure 1.1.1: Animals Killed in a Coyote Control Program³

The screen capture that follows shows a continuation of the R session. It illustrates the method of extracting entries from vectors and data frames. Also shown are the commands `nrow` and `ncol` for obtaining the number of rows and the number of columns of a data frame.



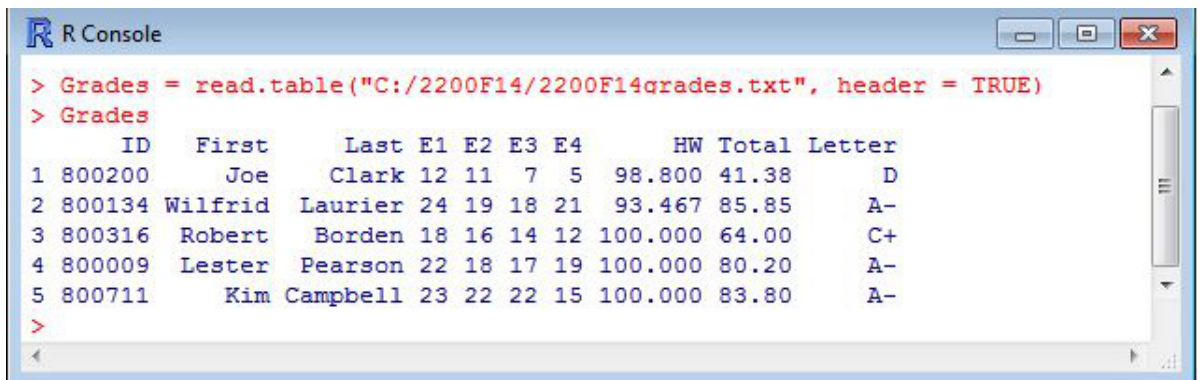
```

> Species[4]
[1] "Raccoons"
> Count[5]
[1] 6348
> CoyoteControl[3,2]
[1] 1367
> CoyoteControl[6,1]
[1] Beavers
Levels: Beavers Coyotes Dogs Opossums Other Porcupines Raccoons Skunks
> CoyoteControl$Species
[1] Coyotes Opossums Porcupines Raccoons Skunks Beavers Dogs Other
Levels: Beavers Coyotes Dogs Opossums Other Porcupines Raccoons Skunks
> CoyoteControl$Count
[1] 25026 2698 1367 3345 6348 682 273 5243
> nrow(CoyoteControl)
[1] 8
> ncol(CoyoteControl)
[1] 2

```

Figure 1.1.2: Extracting Entries of vectors and data Frames in R

Often an existing table (from a spreadsheet or other database) is read into an R session. Figure 1.1.3 is a screen capture that illustrates how this is done. The file consists of five lines of grades from a statistics course given at First President University. The names and IDs have been fictionalized.



```

> Grades = read.table("C:/2200F14/2200F14grades.txt", header = TRUE)
> Grades

```

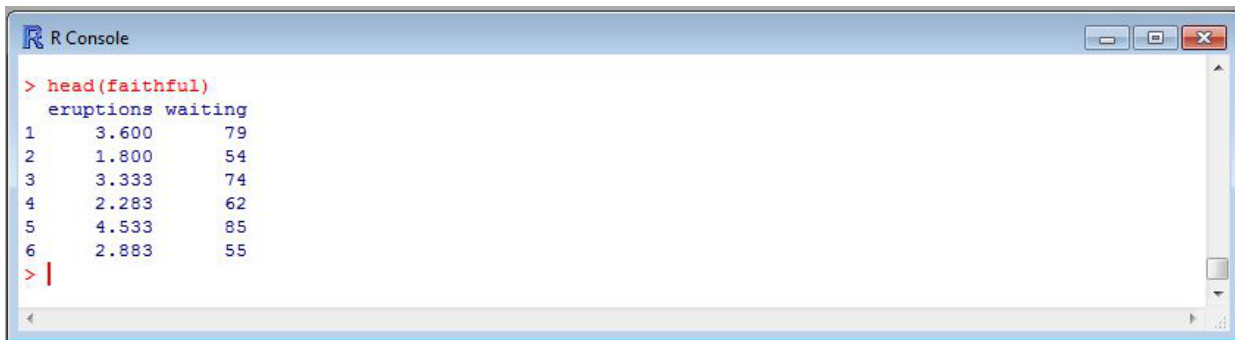
	ID	First	Last	E1	E2	E3	E4	HW	Total	Letter
1	800200	Joe	Clark	12	11	7	5	98.800	41.38	D
2	800134	Wilfrid	Laurier	24	19	18	21	93.467	85.85	A-
3	800316	Robert	Borden	18	16	14	12	100.000	64.00	C+
4	800009	Lester	Pearson	22	18	17	19	100.000	80.20	A-
5	800711	Kim	Campbell	23	22	22	15	100.000	83.80	A-

Figure 1.1.3: M2200 Grades, Fall 2014

It is useful to know that R comes with several built-in data sets. One favorite is `faithful`, which contains 272 rows and two columns of eruption data for Old Faithful, a geyser in Wyoming. The first column is the

³In the 1990s, the U.S. Fish and Wildlife Service set steel-jaw leghold traps in order to control the coyote population. Of the 44,982 animals that were killed in the program, only 25,026 were the intended species.

duration in minutes of an observed eruption and the second column is the time in minutes until the next eruption. R's command `head(dataFrameName)` prints the headers and the first few lines of data.



```

R Console
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
>

```

Figure 1.1.4: Built-in Data Frame, Faithful

To see a list of all datasets that are available, enter

```
> data(package = .packages(all.available = TRUE))
```

After a brief delay, a window will open displaying an eclectic list of over 300 datasets. A brief selection is shown in the screen capture of Figure 1.1.5.



```

R data sets
Seatbelts           Road Casualties in Great Britain 1969-84
Theoph              Pharmacokinetics of Theophylline
Titanic             Survival of passengers on the Titanic
ToothGrowth         The Effect of Vitamin C on Tooth Growth in Guinea Pigs
UCBAdmissions       Student Admissions at UC Berkeley
UKDriverDeaths      Road Casualties in Great Britain 1969-84
UKgas               UK Quarterly Gas Consumption
USAccDeaths         Accidental Deaths in the US 1973-1978
USArrests           Violent Crime Rates by US State
USJudgeRatings      Lawyers' Ratings of State Judges in the US Superior Court
USPersonalExpenditure Personal Expenditure Data
VADeaths            Death Rates in Virginia (1940)
WWWusage            Internet Usage per Minute
WorldPhones         The World's Telephones
ability.cov         Ability and Intelligence Tests
airmiles            Passenger Miles on Commercial US Airlines, 1937-1960
airquality          New York Air Quality Measurements
anscombe            Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu              The Joyner-Boore Attenuation Data
attitude            The Chatterjee-Price Attitude Data
austres             Quarterly Time Series of the Number of Australian Residents
beaver1 (beavers)   Body Temperature Series of Two Beavers
beaver2 (beavers)   Body Temperature Series of Two Beavers
cars                Speed and Stopping Distances of Cars
chickwts            Chicken Weights by Feed Type
co2                 Mauna Loa Atmospheric CO2 Concentration
crimtab             Student's 3000 Criminals Data
discoveries         Yearly Numbers of Important Discoveries

```

Figure 1.1.5: Some Data Sets Available in R

1.2 Frequency Tables

Consider an up-coming election with three candidates: Tweedle-Dee, Tweedle-Dum, and Twitter-Dot. By fixing a moment of time for each of the registered voters, we obtain a *Voter Preference* variable. *Voter Preference* is a categorical variable with, in this case, four possible values (categories): voters in favor of Dee, voters in favor of Dum, voters in favor of Dot, and undecided voters: each registered voter belongs to one of these four categories.

Suppose that a survey of 1000 registered voters is commissioned to determine voter preferences. This subpopulation of registered voters will constitute the cases of the identifier variable *Surveyee* and of the variable of interest, *Voter Preference*. Using the order in which the responses are obtained to provide the values of the identifier *Surveyee*, we could tabulate the results of the survey as follows:

Surveyee	Voter Preference
1	Dot
2	Dee
3	Dee
4	Dum
5	Undecided
6	Dot
⋮	⋮
1000	Dum

Table 1.2.1: Voter Preference Survey—Raw Data

What would be the point of such a table, however? The first variable, *Surveyee*, is completely immaterial: there would be no difference in the results of the survey if, for example, the order of the third and fourth surveyees had been reversed. What matters is a variable that we will introduce and call *Count*. The four possible values of the *Voter Preference* variable, namely Dee, Dum, Dot, and Undecided, will be the cases of the *Count* variable. Supposing that of the 1000 surveyees, 345 preferred Dee, 388 preferred Dum, 189 preferred Dot, and 78 were undecided, here is a more meaningful way of tabulating the results of the survey:

Voter Preference	Count
Dee	345
Dum	388
Dot	189
Undecided	78
Total	1000

Table 1.2.2: Voter Preference

Such a table is called a ***frequency table***. It contains all the information about the variable *Voter Preference*: the collection of all possible cases of the variable as well as the number of times each case occurred. This information is called the ***distribution*** of the variable *Voter Preference*. In general, a frequency table of a categorical variable X is a tabulation of the distribution of the variable X .

Although the frequency table of a categorical variable X can be very useful, the distribution of X depends on a somewhat arbitrary parameter: the number of cases. In the example of *Voter Preference* that we have been considering, the distribution of *Voter Preference* shown in Table 1.2.2 depends on the number, 1000, of surveyees. Had 1500 registered voters been surveyed, the numbers in Table 1.2.2 would have been quite different. For this reason, *relative counts* are often presented instead of raw counts. To obtain the relative counts of a variable X , the value for each case of *Count* is divided by the total number of cases of X , and these fractions are presented in decimal form. In this example, the total is 1000, so the relative counts are $345/1000$, $388/1000$, $189/1000$, and $78/1000$ (or 0.345, 0.388, 0.189, and 0.078). In tabular form, we have

Voter Preference	Relative Count
Dee	0.345
Dum	0.388
Dot	0.189
Undecided	0.078
Total	1.000

Table 1.2.3: Voter Preference—Relative Counts

Such a table is called a *relative frequency table*. Often percentages are displayed rather than decimal fractions. Each relative count value is multiplied by 100 and presented as a percentage. The resulting table is also called a relative frequency table.

Voter Preference	Count (%)
Dee	34
Dum	39
Dot	19
Undecided	8

Table 1.2.4: Voter Preference—Percentages

For a more user-friendly table, we have rounded to whole numbers. In this example, every value of *Relative Count* $\times 100$ would ordinarily have been rounded up. The danger of rounding percentages is that the total will not necessarily come to 100. Here we decided to risk the protests of Candidate Dee and round 34.5 down to 34 in order to avoid the inconvenience of having a percentage total that does not equal 100. Had the author resisted the temptation to fiddle with statistics so early in the manuscript, he might have rounded 34.5 up to 35 and added the standard disclaimer, “Figures may not add to 100% due to rounding.”⁴

Relative frequency tables 1.2.3 and 1.2.4 tell us a good deal about the distribution of the categorical variable *Voter Preference*: the collection of all possible values (categories) of *Voter Preference* as well as how the collected data is distributed among these categories. They do not, however, tell us the number of cases obtained for *Voter Preference*. That information can be useful. Toss a coin twice and get two heads, there is no suspicion that there is anything hinky about the coin. Toss a coin 1000 times and get 1000 heads, there is a strong suspicion that each side of the coin is a head. For that reason, when relative frequency tables are presented, the size of the study is also disclosed.

⁴One common rounding convention rarely taught in schools is to round $n.5$ to the nearest *even* integer, n or $n + 1$.

Calculating Totals and Relative Frequencies in R

Many calculations involving distributions are especially easy in R thanks to the builtin shortcuts. Consider, for example, the distribution of votes in the 2016 U.S. Presidential election (in the order Donald Trump, Hillary Clinton, Gary Johnson, Jill Stein, Other):

```
> popular.vote = c(62980160, 65845063, 4488931, 1457050, 931810)
```

To obtain the total vote, we simply use the `sum` function:

```
> total.vote = sum(popular.vote)
[1] 135703014
```

In R, we can multiply (or divide) every number in a vector by multiplying (or dividing) the vector. Thus, to obtain relative frequencies and percentages for the popular vote, we use the following code

```
> popular.vote.proportions = popular.vote/sum(popular.vote)
> popular.vote.proportions
[1] 0.464102883 0.485214448 0.033079081 0.010737050 0.006866539
> popular.vote.percentages = 100*popular.vote.proportions
> popular.vote.percentages
[1] 46.4102883 48.5214448 3.3079081 1.0737050 0.6866539
```

Now let us be a bit fancier using the hypothetical election poll involving Dee, Dum, and Dot. In input Line 1, the vector of counts is created. Line 2 is a bit tricky because the variable `counts` appears on both sides of the equals sign. The first point to observe is that the assignment in Line 2 would not be valid had `counts` not already received a value in the previous line: Every variable on the right side of an equals sign must already have a value. That existing value, namely the vector `c(345, 388, 189, 78)`, is used in every occurrence of `counts` on the right side of Line 2. Thus, `sum(counts)` results in $345 + 388 + 189 + 78$, or 1000. The vector `c(counts, sum(counts))` is therefore `345, 388, 189, 78, 1000`. This new vector, which has length 5, is then assigned to the variable on the left side of the equals sign. Because `counts` already had a value, a vector of length 4, that value is overwritten by the new, longer vector. Line 4 uses `counts` to fill the entries of a matrix with 1 column. The filling procedure is row-by-row, although for this column matrix filling the matrix by column would have resulted in the same matrix. Finally, names are added given for the rows and the column, the matrix is used to create a table, and the table is displayed in Line 8.

```
> counts = c(345, 388, 189, 78)      # Line 1
> counts = c(counts, sum(counts))    # Line 2
> counts # Line 3
[1] 345 388 189 78 1000
> poll.matrix = matrix(counts, ncol = 1, byrow = TRUE) # Line 4
> colnames(poll.matrix) = c("Count") # Line 5
> rownames(poll.matrix) = c("Dee", "Dum", "Dot", "Undecided", "Total") # Line 6
> poll.results = as.table(poll.matrix) # Line 7
> poll.results # Line 8
```

	Count
Dee	345
Dum	388
Dot	189
Undecided	78
Total	1000

Let us go back over this example from a more sophisticated point of view. This time we will rely on R's builtin functions and get a more detailed look as a result. Until we get to the new R functions, the only difference will be that we omit Line 2 (and remove the row name Total since there will be no such row in the table). The new functions are `margin.table` and `prop.table`. For now, each of these functions is called with two arguments separated by a comma: the name of a table and the number 2:

```
> counts = c(345, 388, 189, 78)
> poll.matrix = matrix(counts, ncol = 1, byrow = TRUE)
> colnames(poll.matrix) = c("Count")
> rownames(poll.matrix) = c("Dee", "Dum", "Dot", "Undecided")
> poll.results = as.table(poll.matrix)
> poll.results
```

	Count
Dee	345
Dum	388
Dot	189
Undecided	78

```
> margin.table(poll.results,2)
Count
1000
> prop.table(poll.results,2)
```

	Count
Dee	0.345
Dum	0.388
Dot	0.189
Undecided	0.078

The call `margin.table(poll.results,2)` sums the counts in the column of the table `poll.results` and displays the calculated sum. If `poll.results` had had more than one column, then, for each column, the column sum would have been calculated. The number 2 in the argument list tells R that column actions are required. Replace the 2 with a 1 in the argument list of `margin.table` and the functionality is the same but applied to rows. We will see this more general functionality in Section 1.4.

The call `prop.table(poll.results,2)` sums the counts in the column of the table `poll.results` and divides every entry in the column by the column sum. The result is a vector of relative frequencies. If the table `poll.results` had had more than one column, then each entry of the table would have been divided by the column sum of the column in which it was located. The number 2 in the argument list tells R that column actions are required. Replace the 2 with a 1 in the argument list of `prop.table` and the functionality is the same but applied to rows. We will see this more general functionality in Section 1.4.

1.3 Graphical representations

Homer: Here's good news! According to this eye-catching article [in spoof newspaper *US of A Today*], SAT scores are declining at a slower rate!

Lisa: Dad, I think this paper is a flimsy hodgepodge of pie graphs, factoids, and Larry King^{5,6}

⁵According to Wikipedia, King "wrote a regular newspaper column in [spoofable newspaper] *USA Today* for almost 20 years, from shortly after that newspaper's origin in 1982 until September 2001. The column consisted of short 'plugs, superlatives and dropped names'." Source: http://en.wikipedia.org/wiki/Larry_King Retrieved May 22, 2014

⁶Source: *The Simpsons*, Season 3 Episode 5, "Homer Defined", first aired on Sunday 8:00 PM EST Oct 17, 1991 on FOX.

As the editors of *US of A Today* are aware, hungry humans instinctively discern any inequities in the slicing of a pizza. **Pie graphs** are therefore a favorite device for presenting the data of relative frequency tables. They are often made colorful by shading each sector with its own hue. Here is a pie graph⁷ of the data presented in Table 1.2.4:

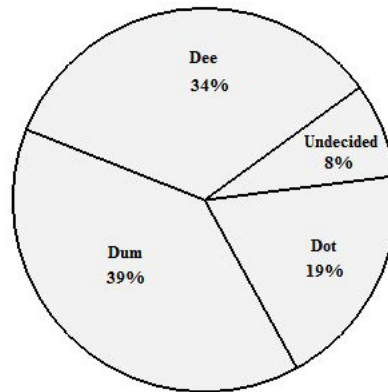


Figure 1.3.1: Voter Preference—Percentages

The angles of the sectors are proportional to the values found in the relative frequency tables for *Voter Preference*: in degree measure, $(34/100) \cdot 360$, $(39/100) \cdot 360$, $(19/100) \cdot 360$, and $(8/100) \cdot 360$. Notice that the sum of these four angles is 360° .

In Figure 1.3.2, we have presented the same data in a different graphical manner by using squares. In doing so, we have violated a standard graphical principle, thereby risking a misperception. The human eye is drawn more to the areas of the squares than to their side lengths. Whereas the side lengths are correct representations of the voter preferences, the areas are *not*.

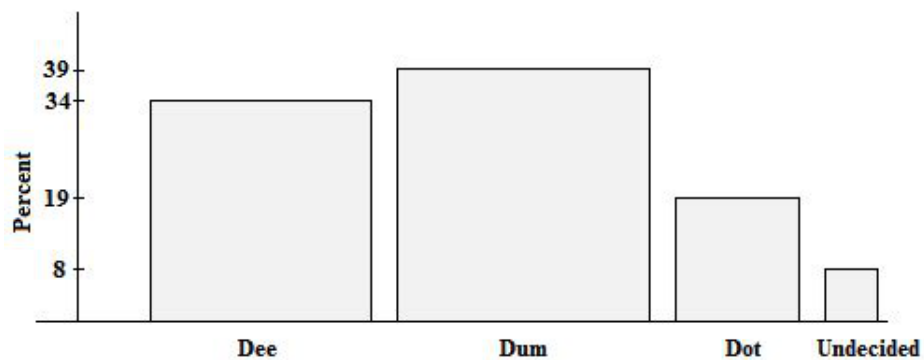


Figure 1.3.2: Voter Preference—Percentages

The percentage of voters favoring Dum, namely 39%, is about twice the percentage, 19%, favoring Dot. We can see that relationship by resolutely focusing on the heights of the squares. That is hard to do, however:

⁷Alas, *My world is monochrome and welcome to it*, says the author.

like it or not, our eyes focus not on the heights but on the areas. And the area of Dum’s square is about four times that of Dot. Figure 1.3.2 is misleading.

We can fix the problem of Figure 1.3.2 by changing the squares to rectangles with equal length bases. In that way the areas will be proportional to the values of *Voter Preference*, just as the sectors of the pie graph were. The result is the **bar graph** shown below:

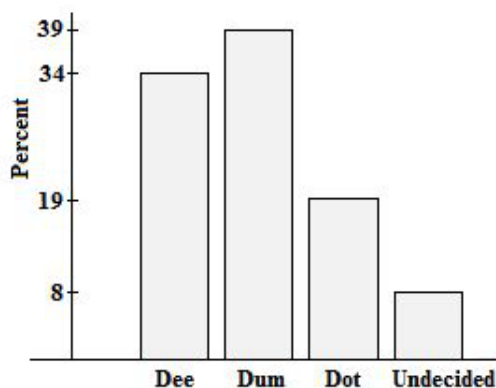


Figure 1.3.3: Voter Preference—Percentages

The observation we have made is sometimes stated formally as the **Area Principle**:

In a graphical display of the values of a categorical variable X , for every value x of X the area of the region representing x should be proportional to the count of x .

The preceding discussion inevitably prompts the observation that there are two types of people: those who refer to graphs, and those who refer to charts.⁸ Which should it be: pie graphs or pie charts? Bar graphs or bar charts? There is no need for a dogmatic answer: both “graph” and “chart” are correct and common. Nevertheless, Google searches reveal that “pie chart” is nearly four times as popular as “pie graph” and that, likewise, “bar chart” outnumbers “bar graph” by a factor close to four.

Type of Display	Graph	Chart
Pie display	About 6,620,000	About 24,400,000
Bar display	About 32,700,000	About 121,000,000

Table 1.3.1: Google Hits for Pie Graph, Pie Chart, Bar Graph, and Bar Chart (Data retrieved: May 22, 2014)

Artists often produce artistic alternatives to pie and bar charts. Figure 1.3.4, for example, is equivalent to a pie chart, and Figure 1.3.5 is an artistically rendered bar chart. Both appeared in *USA Today* and both conform to the Area Principle⁹.

⁸It has also been observed that there are three types of people: those who can count to two, and those who cannot. And, as everybody knows, there are two types of people: those who divide people into two types, and those who do not.

⁹However, Figure 1.3.4 does not follow an important principle of journalism: Anticipate and answer all questions that might arise from a story. On seeing this graphic, readers, such as the author, who suffer from mathematical obsessive compulsive disorder, will be compelled to sum all the volumes in the chart. The total, 44.73 gallons, does not equal 42 gallons, which is equivalent to one barrel of oil. That discrepancy will cause anxiety in readers with mathematical OCD. It turns out that in the refining of crude oil, processing and chemical changes increase the volume.

Products Made from a Barrel of Crude Oil (Gallons)

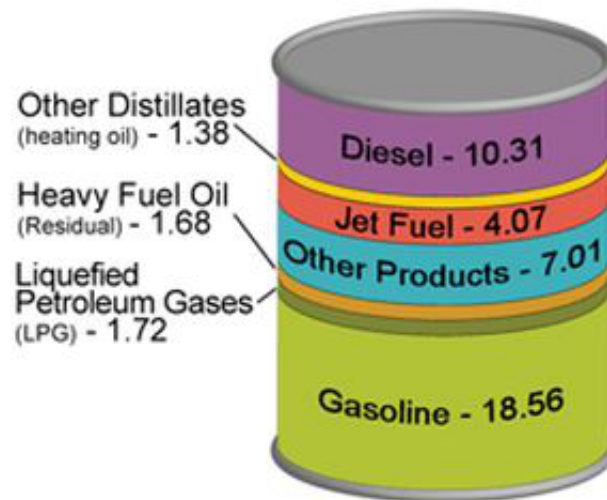


Figure 1.3.4: Products Refined from One Barrel of Oil

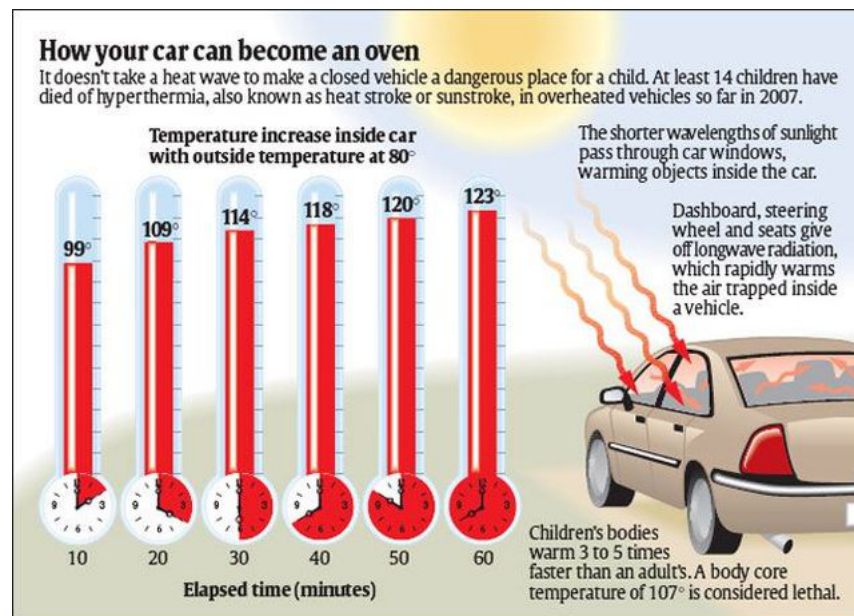


Figure 1.3.5: Temperature Inside Car as a Function of Time (Outdoor Temperature: 80°F)

Sometimes the bars of bar charts are aligned horizontally instead of vertically.

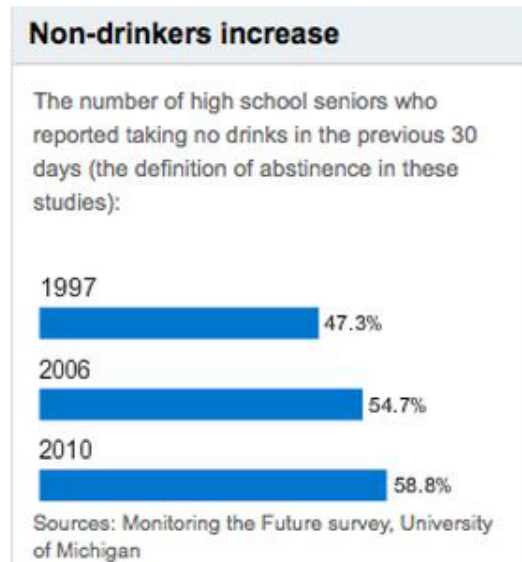
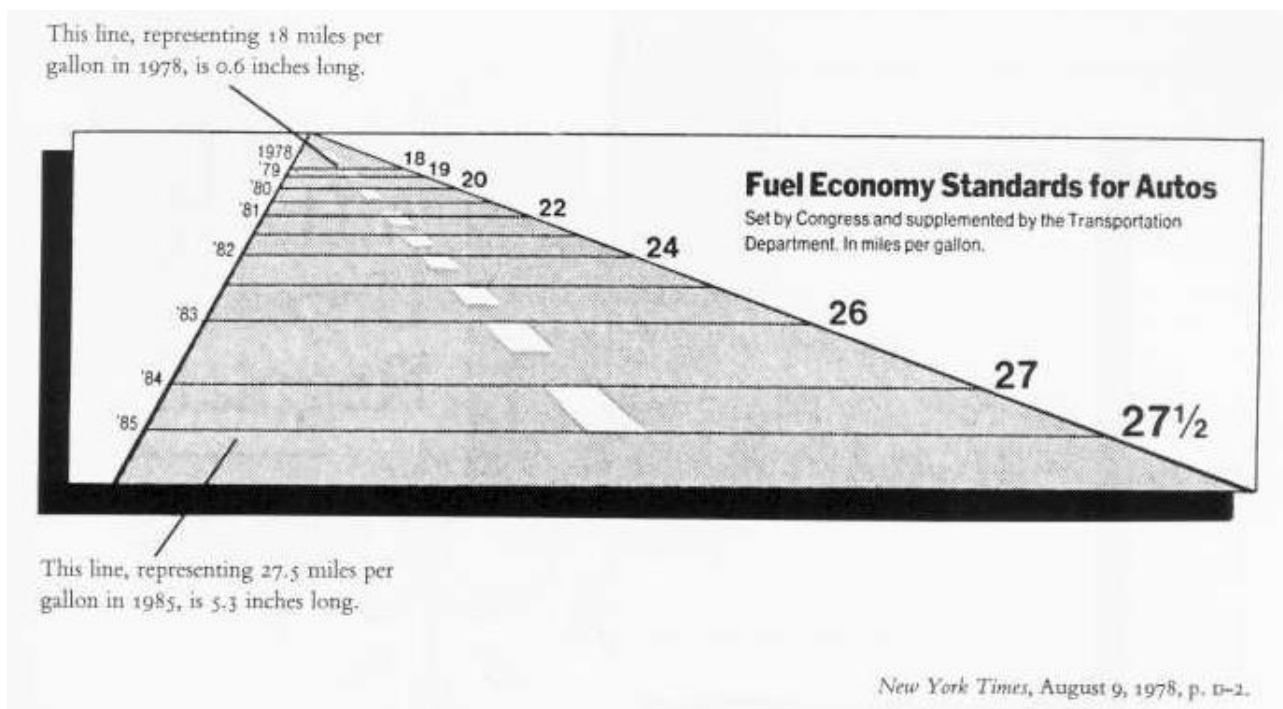


Figure 1.3.6: Percentage of High School Seniors Not Consuming Alcoholic Beverages

Sometimes the bars of bar charts are reduced in width to lines. In this case, the lengths of the lines should be proportional to the quantities they represent. The following figure, which originally appeared in the New York Times, was cited by E.R. Tufte as an egregious violation of the Area Principle.¹⁰



¹⁰ *The Visual Display of Quantitative Information*, E.R. Tufte, Graphics Press, 1983, p.57

Figure 1.3.7: Fuel Economy Standards for Autos, 1978–1985

The intention of Figure 1.3.7 was to illustrate the mandated 53% increase in fleet fuel economy, from 18 mpg to 27.5 mpg, that was to be implemented between 1978 and 1985. The magnitude of increase shown by the bars of the graph is actually 783%. In Figure 1.3.8, the right edge of the road has been repositioned so that the horizontal lines are proportional to the quantities they represent. The result is an honest but less dramatic graph.

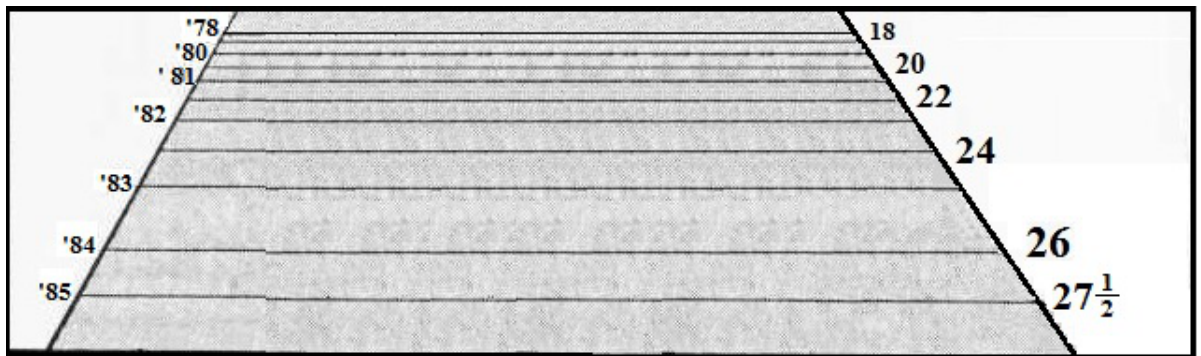
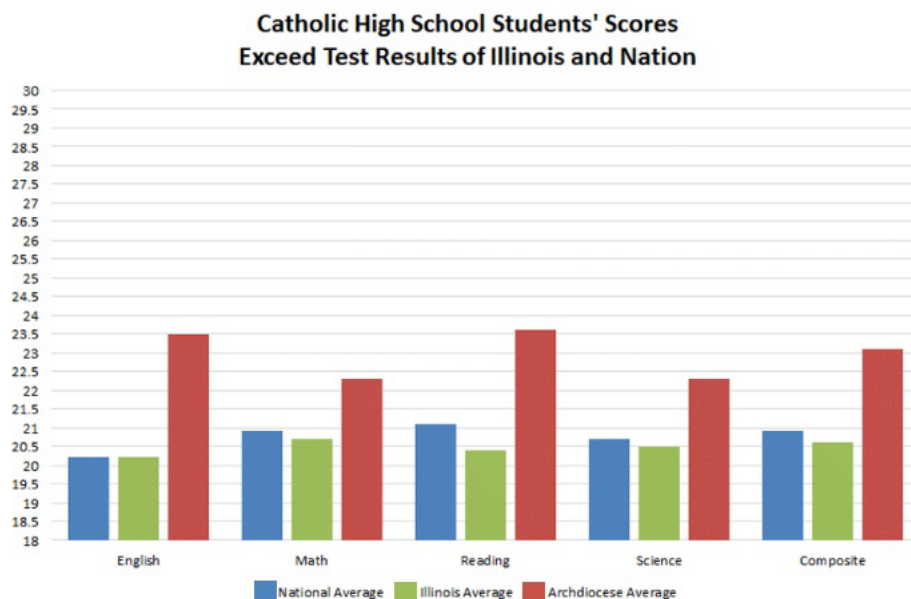


Figure 1.3.8: Fuel Economy Standards for Autos, 1978–1985 (Reproportioned)

Figure 1.3.9¹¹, which shows national, Illinois, and Chicago Archdiocese average ACT scores for 2013, is an example of a somewhat subtle violation of the Area Principle. If the author lived in Chicago, and if the author had a child approaching school age, he would, on glimpsing Figure 1.3.9, say to himself, *I may be an atheist, and everything I know about the Chicago Archdiocese I learned from Jake and Elwood Blues*¹², but, *all that notwithstanding, my child is going to be educated by the expert teachers of the Chicago Archdiocese*. On closer examination of the graphic, the author would say to himself, *They did not obey the Area Principle*.



¹¹<http://schools.archchicago.org/Academics/Charts/ACT.aspx> Retrieved May 7 27, 2014. The author became aware of this graphic from Brian Fitzpatrick's blog, <http://fitz.blogspot.com>

¹²As documented in *The Blues Brothers* <http://www.imdb.com/title/tt0080455/>

Figure 1.3.9: National, Illinois, and Chicago Archdiocese Average ACT Scores, 2013

Do you see what is wrong? For a hint, if needed, consider the following graphic.

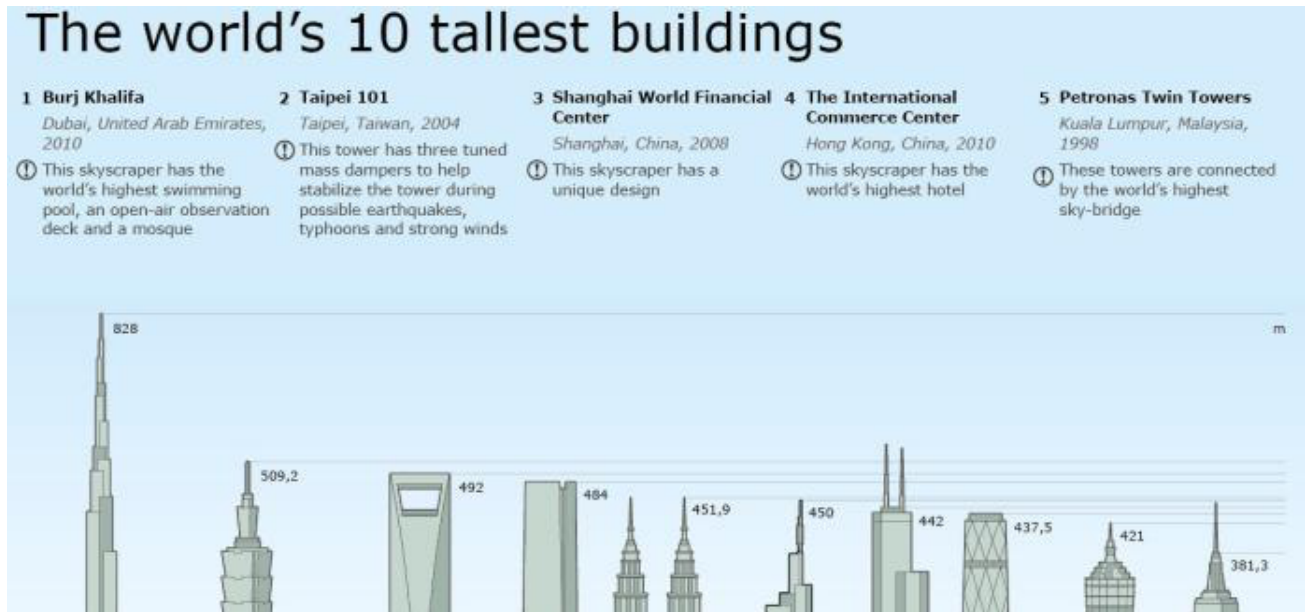


Figure 1.3.10: Top Stories of World's Tallest Buildings (Heights in meters)

This graphic seems to show that the Burj Khalifa is twice as tall as the other leading skyscrapers. In this case, the graphic has obviously been altered by truncating the first few hundred meters of the buildings. This deception can be avoided by displaying the full graphic.

The world's 10 tallest buildings

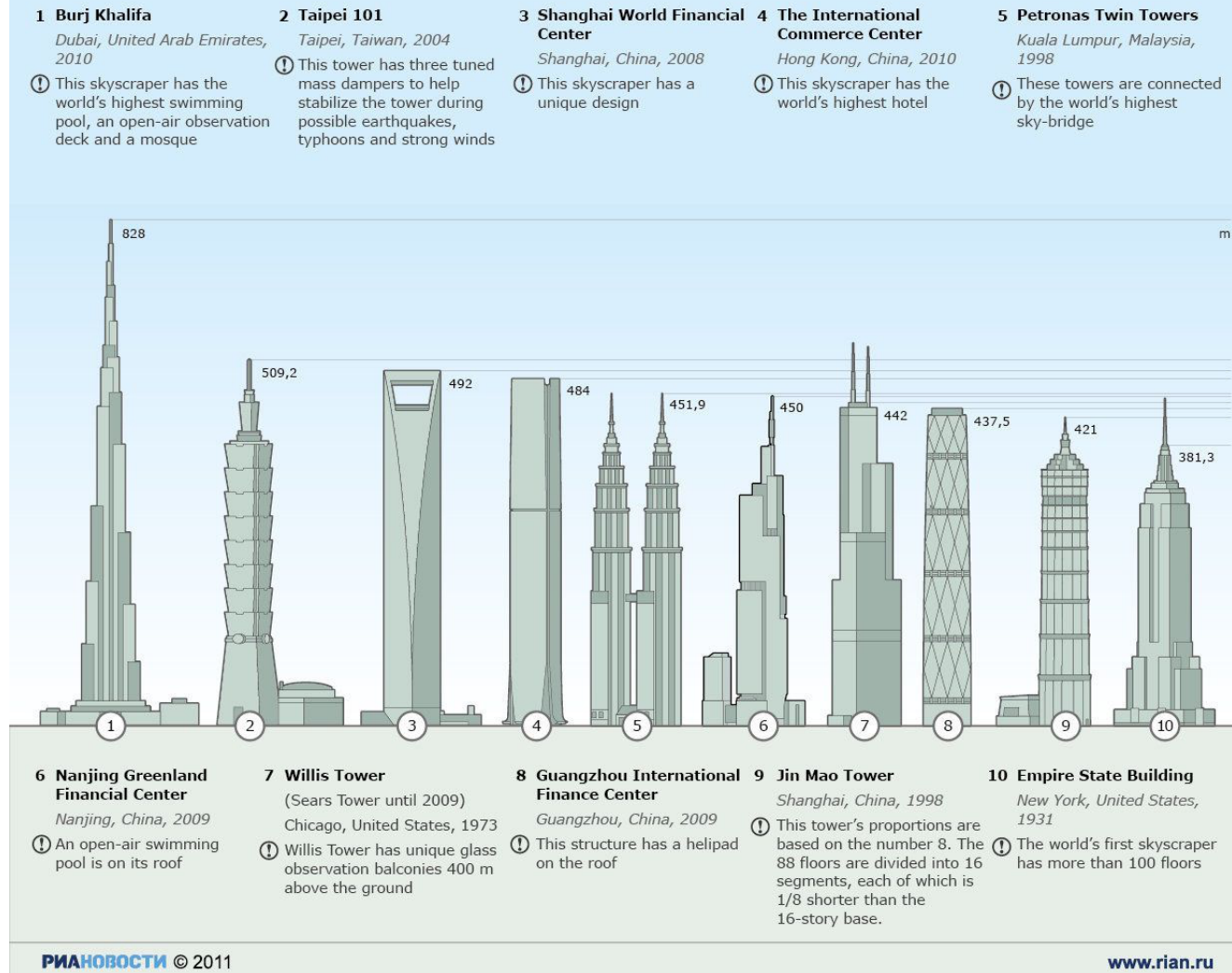


Figure 1.3.11: The World's Tallest Buildings (Heights in meters)

In the same way, the Chicago Archdiocese bar chart can be corrected by restoring the first 18 units of the bars. Refer to Figure 1.3.12. The issue that the chart compares kumquats to pomegranates remains.

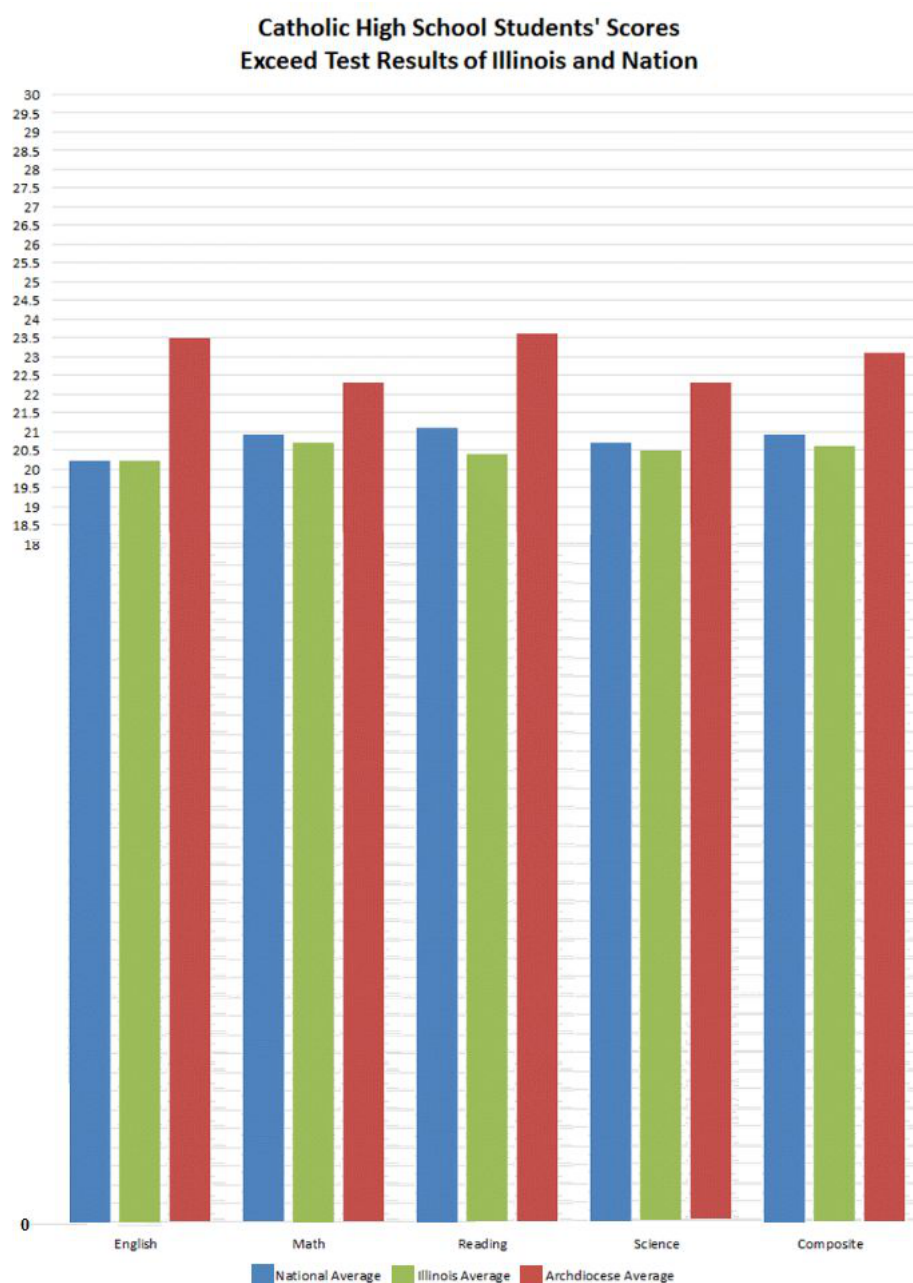


Figure 1.3.12: National, Illinois, and Chicago Archdiocese Average ACT Scores, 2013

Pie Charts in R

Creating a basic pie chart in R is as easy as pie. If `list` is a list of positive numbers, then simply enter `pie(list)` at a prompt. For example,

```
> distribution = c(9, 22, 16, 5, 12)
> pie(distribution)
```

produces the plain pie chart in Figure 1.3.13.

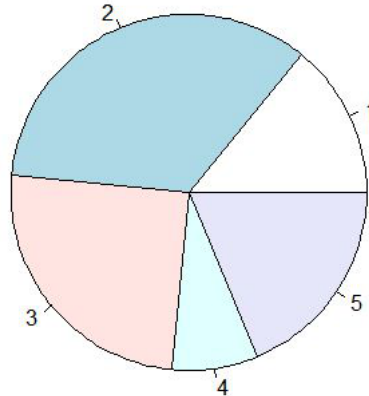


Figure 1.3.13: Plain Vanilla Pie Chart

In the code given, `distribution` is a user-defined name that has no *a priori* significance in R. It is the name assigned to the vector $\langle 9, 22, 16, 5, 12 \rangle$, which is coded as `c(9, 22, 16, 5, 12)` in R. The code could have been given more compactly in one line as `pie(c(9, 22, 16, 5, 12))`, but such a shortcut makes the code more difficult to decipher. Furthermore, assigning a list to a name allows for expedited future references if needed.

Notice that the numbers in the list that is to be pie-charted do not have to be percentages; that is, the numbers in the list do not have to sum to 100. For instance, the listed numbers of `distribution` sum to 64. The third number in the list, 16, is 25% of the total. As you can see, the sector labelled “3” has 25% of the area of the disk.

Fancier pie charts are not very much more difficult. Consider the voter preference poll of Section 1.3. We will pie-chart the percentages, label the sectors of the pie chart using the optional parameter `labels = ...`, assign a title to the pie chart using the optional parameter `main = ...`, make our own choice of colors—shades of grey are currently very popular—for the sectors using the optional parameter `col = ...`, and insert a legend using the `legend` command. The optional parameter `cex = ...` has been included to shrink the legend so that it does not overlap the pie chart. Thus, the code

```
> VoterPref = c(34,39,19,8)
> sliceLabels = c("Dee","Dum","Dot","Undecided")
> shadesOfGrey = c("grey55","grey67","grey79","grey91")
> pie(VoterPref, main = "Voter Preferences", col = shadesOfGrey, labels = sliceLabels)
> legendLabels = c("Dee 34%", "Dum 39%", "Dot 19%", "Undecided 8%")
> legend("bottomleft", legendLabels, fill = shadesOfGrey, cex = 0.75)
```

produces the snazzy pie chart in Figure 1.3.14.

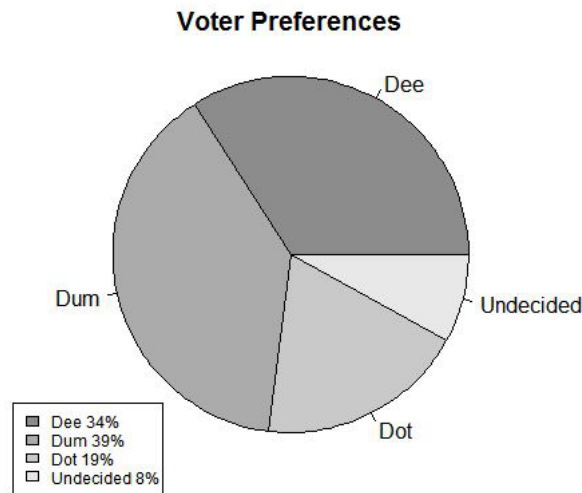


Figure 1.3.14: Work of Art Pie Chart

That was a piece of cake! In the code given, `VoterPref`, `sliceLabels`, `shadesOfGrey`, and `legendLabels` are all user-defined names to which lists have been assigned. All other terms that appear in the code have a *a priori* meaning in R.

Additional pie-charting functionality can be obtained by installing and loading the `plotrix` package. For example, here is a 3-dimensional pie chart. The parameter `explode = ...` controls the separation (explosion) of the wedges. Thus, the code

```
> library(plotrix)
> VoterPref = c(34,39,19,8)
> sliceLabels = c("Dee","Dum","Dot","Undecided")
> shadesOfGrey = c("grey55","grey67","grey79","grey91")
> pie3D(VoterPref, labels = sliceLabels, col = shadesOfGrey, explode = .1, radius = pi/2)
```

results in the 3D pie chart in Figure 1.3.15.

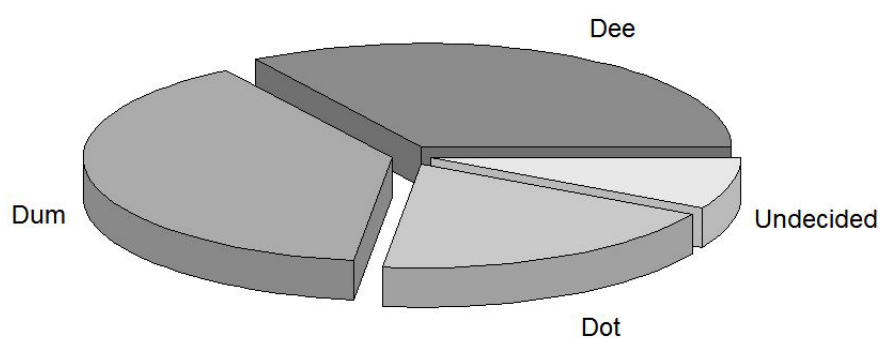


Figure 1.3.15: Three Dimensional Exploded Pie Chart

Bar Charts in R

There are no barriers to creating a basic bar chart in R. If `list` is a list of positive numbers, then simply enter `barplot(list)` at a prompt. The example we show is slightly less barebones. The optional parameters `col = ...` and `names.arg = ...` are employed to specify shades of grey colorings and bar labels. The optional argument `axis.lty = 1` results in a visible horizontal baseline, which is otherwise not drawn. The optional parameters `xlab = ...` and `ylab = ...` permit the labelling of the horizontal and vertical axes.

```
> VoterPref = c(34,39,19,8)
> barLabels = c("Dee","Dum","Dot","Undecided")
> shadesOfGrey = c("grey55","grey67","grey79","grey91")
> barplot(VoterPref, col = shadesOfGrey, names.arg = barLabels,
          axis.lty = 1, xlab = "Candidates", ylab = "Percentage Support")
```

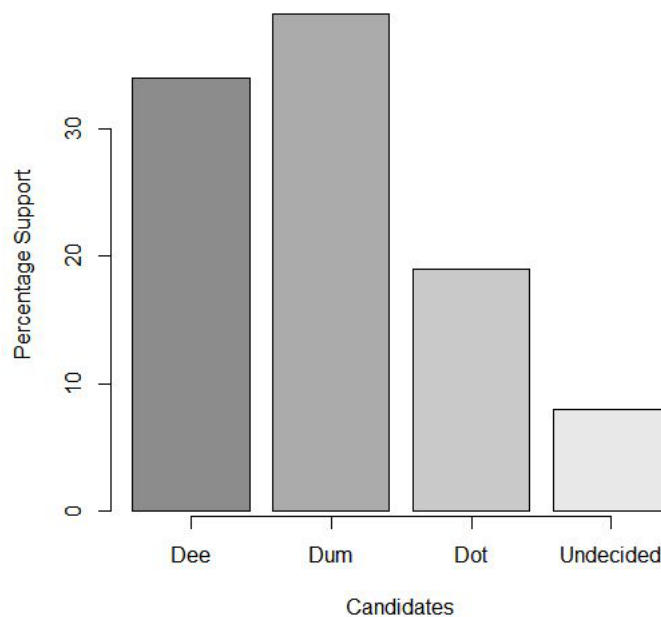


Figure 1.3.16: A Bar Chart Produced in R

In many instances a particular type of bar chart is particularly useful for understanding taking in at a glance the largest counts. In this type of bar chart we arrange the bars so that they decrease in height from left to right. We then add an increasing piecewise linear graph above to show the consecutive cumulative sums of the bars from left to right. Such a chart is called a *Pareto chart*.

To illustrate, we will continue with the R session with the popular vote percentages of the U.S. 2016 Presidential Election. The command we will use is `pareto.chart` and it is contained in the `qcc` package. Chances are, the first time you use `pareto.chart`, you will have to install the `qcc` package:

```
> install.packages("qcc")
```

Running the install command finds the package in CRAN and installs it on your computer. You will still need to load the package when you want to use it:

```
> library(qcc)
```

We will give names to the percentages in `popular.vote.percentages` and then call on `pareto.chart` to produce the desired plot:

```
> names(popular.vote.percentages) = c("Donald Trump", "Hillary Clinton", "Gary Johnson",  
+ "Jill Stein", "Other")  
> pareto.chart(popular.vote.percentages)
```

Pareto chart analysis for popular.vote.percentages

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Hillary Clinton	48.5214448	48.52144	48.5214448	48.52144
Donald Trump	46.4102883	94.93173	46.4102883	94.93173
Gary Johnson	3.3079081	98.23964	3.3079081	98.23964
Jill Stein	1.0737050	99.31335	1.0737050	99.31335
Other	0.6866539	100.00000	0.6866539	100.00000

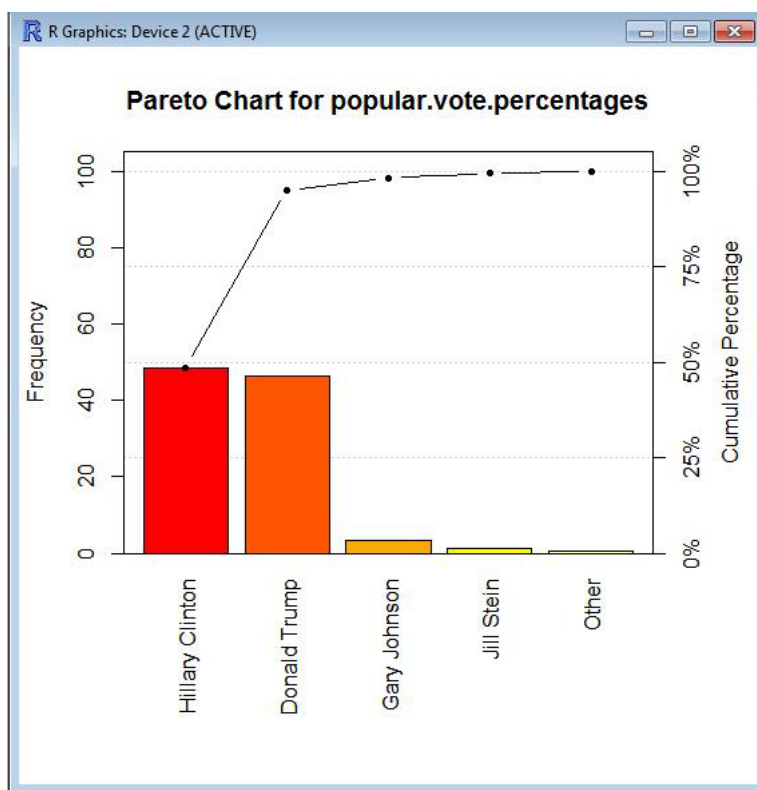


Figure 1.3.16: A Pareto Chart in R

1.4 Contingency Tables

We will continue with the voter survey discussed in Section 1.2, but now we will suppose that additional information was obtained in the survey. For example, categorical data such as the gender, race, and age group of each respondent is often of interest in political surveys. Let us focus on *two* of the categorical

variables: *Gender* and *Voter Preference*. In the discussion that follows, it is significant that the number of categorical variables that we are simultaneously considering is two. The numbers of values (categories) that these variables have—in this case, 2 for *Gender* and 4 for *Voter preference*—are immaterial.

The values of the categorical variable *Gender* appear as column identifiers in Table 1.4.1 below. The four values of the categorical variable *Voter Preference* appear as row identifiers. As you can see, the table illustrates the breakdown of each candidate's *Count* by gender. The total value of the two numerical values per row is equal to the value of *Count* for the candidate. This value is presented in a column that may be regarded as being outside and to the right of the table; this column is generally considered to be a margin of the table. The header *Total* of this margin is short for *Total Count per Candidate (Regardless of Gender)*. Why do we make this distinction? The identifier of the marginal column is not a value of the categorical variable *Gender*, as are the entries to the left of it in its row.

	Female	Male	Total
Dee	171	174	345
Dum	193	195	388
Dot	96	93	189
Undecided	46	32	78
Total	506	494	1000

Table 1.4.1: Voter Preference by Gender

The column sums have also been included, and they are also considered to be *outside* of the table. The values of these sums are given as a row below the table. This marginal row has been labelled *Total*, which is short for *Total Count per Gender (Regardless of Voter preference)*. The values in this row represent, from left to right, the number of female voters surveyed, the number of male voters surveyed, and the total number, 1000, of voters surveyed. Of course, this last number must also equal the sum of the values above it in the right margin. Why is there a distinction between this marginal row and the rows above it? Its identifier is not a value of the categorical variable *Voter Preference*, as are the entries above it in its column. Note that all the marginal entries are obtained by adding entries of the table. The entries of the table are, by contrast, obtained by counting the survey responses.

Table 1.4.1 is called a **contingency table**. A synonym for “contingency table” is **cross-tabulation**. The term **two-way table** is also in use. The key feature of a contingency table is that there is an underlying variable X , each value of which is assigned to a category of a first categorical variable C and to a category of a second categorical variable R . The identifiers of the columns of the table are the values of C . The identifiers of the rows of the table are the values of R . If c is a value of C and if r is a value of R , then the entry n in the column for c and the row for r is the number of cases of X that belong to both the category c and the category r . If categorical variable R has p values and categorical variable C has q values, then the contingency table has p rows and q columns (not including the top row and left column of identifiers and not including the marginal row and column). We say that the contingency table is a $p \times q$ contingency table. It contains $p \cdot q$ **cells**. The numbers in the cells are said to be **joint frequencies**. In the example we have been considering, X is the variable *Surveyee*, C is the variable *Gender*, and R is the variable *Voter Preference*. Table 1.4.1 is a 4×2 contingency table. The joint frequency of being female and preferring Dee is 171. The joint frequency of being male and undecided is 32, and so on: there are 6 other joint frequencies.

Look at Table 1.4.1 and disregard the two columns that correspond to the two cases, Female and Male, of the variable *Gender*. Together, the column of cases of *Voter Preference* and the margin at the right constitute the frequency table of *Voter Preference*. Compare with Table 1.2.2. In the context of Table 1.2.2, the data of the Count column was said to be the distribution of *Voter Preference*. In the present context, that data is said to be the **marginal distribution** of *Voter Preference* because the data is located in the margin of the

contingency table. For the same reason, the data in the margin below Table 1.4.1 is the marginal distribution of *Gender*.

The reason for the name *contingency table* is that such a table gives the distribution of one categorical variable contingent upon the distribution of the other categorical variable. Table 1.4.1 tells us the distribution of *Voter Preference* contingent on the gender of the surveyee. Symmetrically, Table 1.4.1 also gives us the distribution of *Gender* contingent on the voter preference of the surveyee.

Often several contingency tables can be constructed from a comprehensive data set. Let us examine two contingency tables associated with the sinking of the *RMS Empress of Ireland* in the early hours of 29 May 1914.¹³ The underlying variable playing the role of *X* in the generic discussion above is *Persons on Board*. The first categorical variable we will consider, the one playing the role of *C* in the generic discussion above, is *Mortality Outcome*. It has two values: *Survived* and *Perished*. The second categorical variable we will consider, the one playing the role of *R* in the generic discussion above, is *Type of Voyager*. It has four values: 1st Class Passenger, 2nd Class Passenger, 3rd Class Passenger, and Crew Member.

	Survived	Perished	Total
1 st Class	36	51	87
2 nd Class	48	205	253
3 rd Class	133	584	717
Crew	248	172	420
Total	465	1012	1477

Table 1.4.2: Mortality Outcome by Type of Voyager

Did the crew put their own lives at risk by selflessly doing everything possible to preserve the lives of their passengers? Did first-class passengers have greater chances of survival than the riff-raff? These are the sort of questions on which contingency tables are intended to shed light. From Table 1.4.2 we see that the survival rate of the crew, 248 out of 420, or 59%, was substantially higher than the survival rate of first-class passengers (41%).¹⁴ This rate was, in turn, significantly higher than the survival rates for second-class passengers (19%) and third-class passengers (18.5%).¹⁵

Women and children first! So goes the well-known¹⁶ code of conduct for emergencies at sea. Let us introduce another categorical variable that separates the passengers into three categories: Men, Women, and Children. Table 1.4.3 is the resulting contingency table.

	Survived	Perished	Total
Men	172	437	609
Women	41	269	310
Children	4	134	138
Total	217	840	1057

Table 1.4.3: Passenger Mortality Outcome for Men, Women, and Children

¹³As it happens, this discussion is being written 100 years to the day after this maritime disaster occurred. Despite its name, the ocean liner *Empress of Ireland* was built in Scotland to serve a trans-Atlantic route between Canada and England. It went down in the Saint Lawrence River near Rimouski, Québec, after a collision with a Norwegian cargo ship. The author, his birth and upbringing on an island in the Saint Lawrence River notwithstanding, learned of the incident (while landlocked in the Midwest) thanks to *Lost Liners*, an excellent PBS documentary, which was released on VHS after its broadcast, but which has never been silvered. See <http://www.pbs.org/lostliners/> if interested. The data presented in these notes is from the Wikipedia page http://en.wikipedia.org/wiki/RMS_Empress_of_Ireland, retrieved 28 May 2014, and does not agree exactly with the numbers found on the PBS site.

¹⁴There was little time for rescue. The ship was submerged within fourteen minutes of the collision. The situation must have quickly become one of every man for himself.

¹⁵Lower class passengers had berths deeper in the ship.

¹⁶It is even the subject of an interesting Wikipedia page http://en.wikipedia.org/wiki/Women_and_children_first.

The extremely low survival rate of children, less than 3%, may be contrasted with the survival rate of 51% for children on board the Titanic.¹⁷

Contingency tables sometimes provide percentages instead of counts. However, percentages, if not properly presented can lead to confusion. The issue is, Percentage of *what*? Stated in terms of the potentially ambiguous arithmetic operation, In the division that is carried out to obtain the percentage, what is the divisor?

In Table 1.4.2, consider, for example, the 248 surviving crew members. What do we mean when we speak of the percentage of surviving crew members? Arithmetically, we can divide 248 by 420 and multiply by 100 to obtain 59.05%. This is the percentage of survivors in the group of crew members. Because the data for crew members is set out in a row, we can, for clarity, refer to 59.05% as a **row percentage**. We can also divide 248 by 465 and multiply by 100 to obtain 53.33%. This is the percentage of crew members in the group of survivors. Because the data for survivors is set out in a column, we can, for clarity, refer to 53.33% as a **column percentage**. Finally, we can divide 248 by 1477 and multiply by 100 to obtain 16.79%. This is the percentage of surviving crew members in the group of all voyagers. Because the data for all voyagers occupies the entire table, we can, for clarity, refer to 16.79% as a **table percentage**.

Let us return to Table 1.4.2 and focus on only one column. For the purpose of having a concrete discussion, we will select the column containing the counts per voyager type of those who survived (with the understanding that a similar discussion pertains to the column with the counts per voyager type of those who died). In Table 1.4.4 below, we have rewritten the counts in the first column of Table 1.4.4 as column percentages: 36 of 465 has become 7.74%, 48 of 465 has become 10.32%, 133 of 465 has become 28.60%, and 248 of 465 has become 53.33%. The column entries comprise what we call a **conditional distribution**: they provide the relative frequencies of the four categories of *Type of Voyager* for the “condition” *Survived* (i.e., for the value *Survived* of the other categorical variable, *Mortality Outcome*). Note: frequencies are often more informative when given as relative frequencies. In that form they seem to be offered as percentages more often than as fractions.

	Percentage of Survivors
1 st Class	7.74%
2 nd Class	10.32%
3 rd Class	28.60%
Crew	53.33%
Total	100%

Table 1.4.4: Percentage of Survivors by Type of Voyager
Conditional Distribution of *Type of Voyager* for the “Condition” *Survived*

Table 1.4.4 provides us with a good example of the familiar maxim, A little knowledge is a dangerous thing.¹⁸ A natural but incorrect inference Table 1.4.4 might lead to is that third class passengers has a greater survival rate than first class passengers. In fact, as we observed in the paragraph following Table 1.4.2, the survival rate for first class passengers was 41% whereas the survival rate for third class passengers was 18.5%. What Table 1.4.4 does not show is that the total number of third class passengers greatly exceeded the number of first class passengers. Thanks to the greater number of third class passengers, there were more third class survivors. But, contrary to what Table 1.4.4 seems to suggest, third class passengers were less likely to survive. There were over 8 times as many third class passengers as first class passengers, but only four times as many third class survivors as first class survivors.

It should be noted that the row data determine conditional distributions as well. For example, if we focus on the second row of Table 1.4.2, then we obtain conditional distribution

Survived: 48	Perished: 205
--------------	---------------

, which is the conditional distribution of third class passengers by condition of surviving or perishing.

¹⁷The Titanic remained afloat for more than 160 minutes after it collision—more than eleven times as long as the Empress of Ireland.

¹⁸The key word is “little,” having the sense of “incomplete.” If interested, see <http://www.phrases.org.uk/meanings/a-little-knowledge-is-a-dangerous-thing.html> for the origins of the phrase.

Let us briefly return to contingency table 1.4.2. From it, we easily see that the two categorical variables, *Mortality Outcome* with its values *Survived* and *Perished*, and *Type of Voyager* with its values *1st Class Passenger*, *2nd Class Passenger*, *3rd Class Passenger*, and *Crew Member*, have a relationship that can be described in ordinary, nontechnical English as *dependent*. For a voyager on that fateful trip, survival depended strongly on the voyager's type: so-so for a crew member but miserable for a third class passenger. In fact, in the technical jargon of Statistics, we use the same adjective ***dependent*** for the categorical variables *Mortality Outcome* and *Type of Voyager*. The term ***associated*** is also used.

For an example of variables that are ***independent***, consider Table 1.4.5, which uses data from the General Social Survey (GSS) undertaken in 2002.

	Not Too Happy	Pretty Happy	Very Happy	Total
Male	67	391	216	674
Female	69	399	220	688
Total	136	790	436	1362

Table 1.4.5: Happiness by Gender

To get a better idea of how gender influences happiness, let us calculate the row conditional distributions, which we will express using percentages in Table 1.4.6.

	Not Too Happy	Pretty Happy	Very Happy	Total
Male	10%	58%	32%	100%
Female	10%	58%	32%	100%

Table 1.4.6: Happiness: Percentages by Gender

From Table 1.4.6, we see immediately that entries (expressed as percentages) for the two conditional distributions of the categorical variable *Gender* are the same for all three values of the categorical variable *Happiness*. In ordinary English, we would say that *Happiness* is *Gender-independent*. We say the same thing in the technical jargon of Statistics. More generally, we say that two categorical variables are ***independent*** if the conditional distributions of one of the variables (expressed by relative frequencies or percentages) have the same value at each category of the other variable. We will return to this important concept later in these notes.

If It Looks Like a Duck but does not Quack Like a Duck, Then Maybe It is Not a Duck

We conclude this section with the warning that not every table that looks like a contingency table is a contingency table. It must also quack like a contingency table to be a contingency table. Consider the following table, Table 1.4.7, which displays life expectancy data.¹⁹

	Female	Male
Australia	84.6	80.5
Italy	85.0	80.2
Japan	87.0	80.0
Luxembourg	84.1	79.7
Singapore	85.1	80.2
Switzerland	85.1	80.7
USA	81.0	76.0

¹⁹The six countries other than the USA in the table were the only countries that made the top ten for both genders. USA was not in the top ten for either gender.

Table 1.4.7: Life Expectancy, in years (Source: World Health Organization, report issued 15 May 2014)

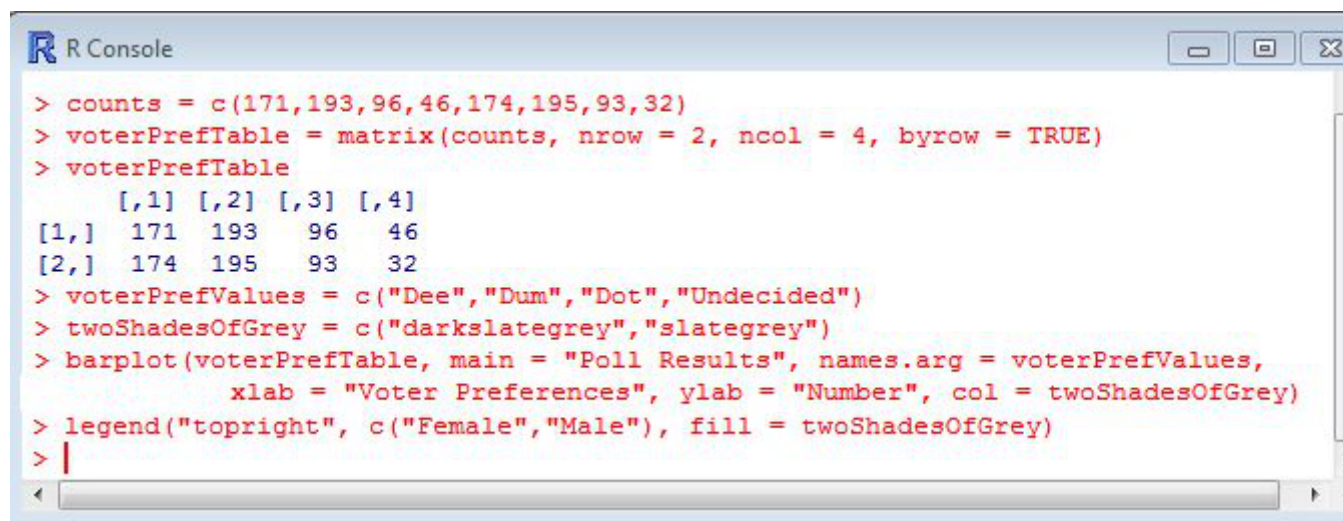
Table 1.4.7 is interesting. Who knew that the citizens of Luxembourg are so long-lived? Clearly the two categorical variables are *Gender* for the columns and *Country* for the rows. At first glance Table 1.4.7 might appear to be a contingency table.²⁰ However, Table 1.4.7 is *not* a contingency table: the cell entries are not frequencies (i.e., counts within categories). In Galliformesical terms, Table 1.4.7 does not quack like a contingency table.

Segmented and Side-by-Side Bar Charts in R

There are two common ways of using bar charts to display the information of a contingency table for categorical variables *X* and *Y*. In the first method, the distribution of one of the categorical variables, say *X*, is represented by bars, which, let us assume, are vertical. The bar representing the total count for a value *x* of *X* is composed of smaller bars, stacked vertically. These bars have heights that are proportional to the conditional distribution of *Y* conditioned on the value *x* of *X*.

To illustrate, let us reconsider contingency table 1.4.1. We will bar chart the distribution of the categorical variable *Voter Preference*, the marginal distribution in the right margin of Table 1.4.1. There will be a bar for each category, Dee, Dum, Dot, and Undecided. In outline, the bar chart will look like Figure 1.3.3. But each bar will be composed of two components: one for each category, *Male* and *Female*, of the other categorical variable, *Gender*. Thus, the bar representing the count of 345 for Dee will consist of a bar representing the 174 male surveyeys supporting Dee stacked on a bar representing the 171 female surveyeys supporting Dee. We will use different colors to help distinguish the division of the bars. Shades of grey again, but new shades of grey this time: R has deep reserves of shades of grey upon which we can draw. More than fifty.

Creating the stacked bar chart in R is not difficult, but it does require care. Although Table 1.4.1 is a 4×2 contingency table, we use the cell data to create a 2×4 matrix in R. Figure 1.4.1, a screen capture of an R session, shows how this is done.

**Figure 1.4.1: An R Session with Code for a Stacked (or Segmented) Bar Chart**

In the first line of Figure 1.4.1, a screen capture of an R session, we have created a vector by entering the first column of Table 1.4.1 followed by the second column. We then used the resulting vector to fill in the

²⁰And would resemble a contingency table even more closely were all the entries rounded to the nearest whole number.

cells of a matrix *by row*, thereby creating a 2×4 matrix, which we have called `voterPrefTable`. In the third line, we call on `voterPrefTable`. Doing so has no effect other than to have R display the matrix so we can see that it has size 2×4 . The stacking of the numbers in the columns of this matrix is similar to the stacking of the sub-bars in the bar chart, except that the stacking in the bar chart will be in the opposite order (174 over 171, 195 over 193, 93 over 96, and 32 over 46). The final two lines of code in Figure 1.4.1 create the stacked bar chart with legend.

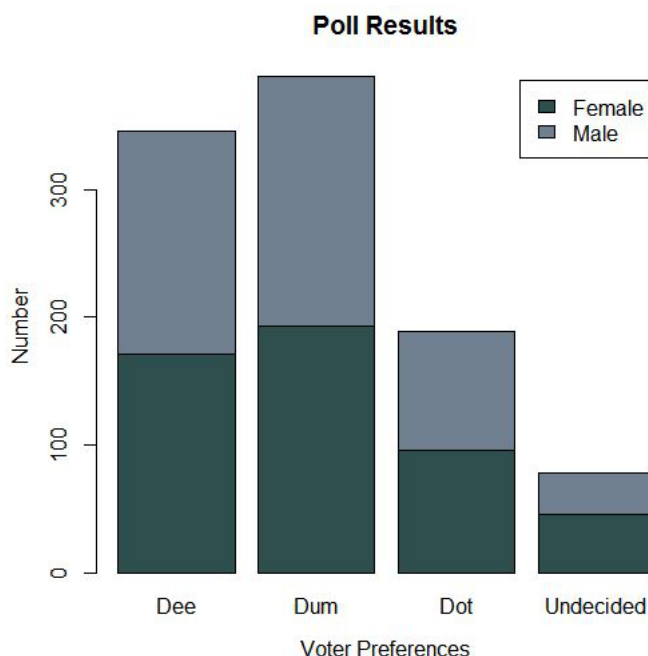


Figure 1.4.2: Stacked Bar Chart of Voter Preference by Gender

An alternative approach is to plot the component sub-bars side-by-side instead of stacked. In this approach, there are no gaps between the side-by-side sub-bars but gaps between the groups. The only essential difference in the R code is the insertion of the argument `beside = TRUE` in `barplot`. (However, having tired of shades of grey, we turned to shades of honeydew.) Thus, the following code,

```
> counts = c(171,193,96,46,174,195,93,32)
> voterPrefTable = matrix(counts, nrow = 2, ncol = 4, byrow = TRUE)
> voterPrefValues = c("Dee","Dum","Dot","Undecided")
> melonShades = c("honeydew1","honeydew3")
> barplot(voterPrefTable, main = "Poll Results", beside = TRUE, names.arg = voterPrefValues,
  xlab = "Voter Preferences", ylab = "Number", col = melonShades)
> legend("topright", c("Female","Male"), fill = melonShades)
```

results in the side-by-side bar chart of Figure 1.4.3:

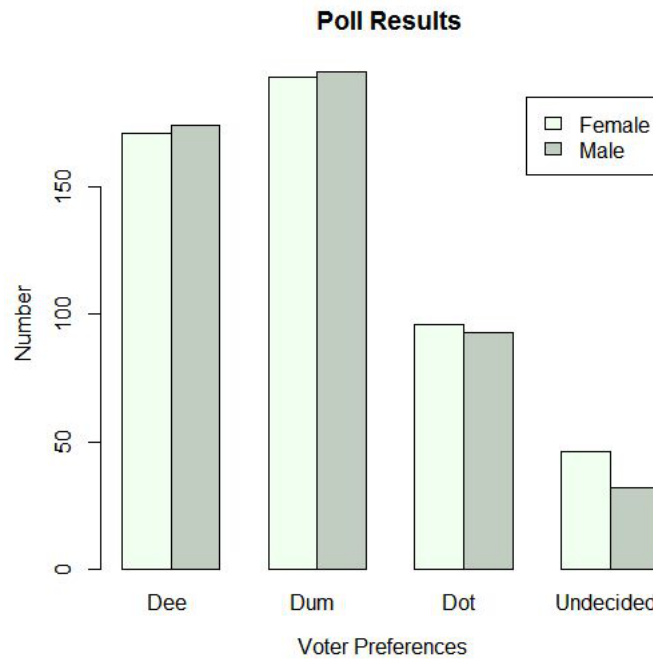


Figure 1.4.3: Side-by-Side Bar Chart of Voter Preference by Gender

Contingency Tables in R

In 1995, the General Social Survey sought to answer the question, Do Americans really like country & western music (CWM), and, if so, where can such Americans be found? The two categorical variables are *Opinion of CWM* and *Region*. The five *Opinion* values are: Strongly Likes, Likes, Mixed, Dislikes, Strongly Dislikes. The nine values of *Region* are: New England, Middle Atlantic, E. North Central, W. North Central, South Atlantic, E. South Central, W. South Central, Mountain, Pacific. In this subsection we will create a contingency table and do contingency things with it. Each of the five values of *Opinion* gives rise to a column of the table. Each of the nine values of *Region* gives rise to a row of the table. The joint frequencies were not available in a copy-and-paste format, so the first step of creating the table is data entry: The 45 joint frequencies are manually entered.

```
> NewEngland = c(5, 13, 8, 3, 0)
> MiddleAtlantic = c(21, 30, 39, 9, 5)
> ENorthCentral = c(41, 60, 40, 17, 9)
> WNorthCentral = c(8, 23, 11, 4, 2)
> SouthAtlantic = c(36, 48, 22, 13, 1)
> ESouthCentral = c(26, 15, 5, 5, 2)
> WSouthCentral = c(27, 24, 10, 7, 1)
> Mountain = c(8, 16, 6, 3, 2)
> Pacific = c(28, 40, 32, 9, 5)
```

Next we concatenate these nine vectors to get one vector, `matrix.entries`, which contains the 45 entries of our desired contingency table. we have chosen the name `matrix.entries` for obvious reasons, but we could have selected any other valid name, such as Bob. In the creation of the table, we call on the R function

`matrix` with three comma-separated arguments. The first is the vector of matrix entries, which we have named `matrix.entries`. The second is `ncol = 5`, which tells R that our matrix will have 5 columns. We might just as well have chosen `nrow = 9` instead of `ncol = 5`. We could even have included both the number of columns and the number of rows in the argument list but there is no need for that: because the vector that is the first argument has 45, or 9×5 , entries, if we specify only one of `ncol = 5` and `nrow = 9`, then R is smart enough to figure out the omitted one. We also need to tell R to fill the matrix by entering the cell data by row (`byrow = TRUE`) or by column (`byrow = FALSE`). We name the matrix `cw.matrix` and ask R to display the matrix so we can verify its entries.

```
> matrix.entries = c(NewEngland,MiddleAtlantic,ENorthCentral,WNorthCentral,
+ SouthAtlantic,ESouthCentral,WSouthCentral,Mountain,Pacific)
> cw.matrix = matrix(matrix.entries, ncol=5, byrow = TRUE)
> cw.matrix
      [,1] [,2] [,3] [,4] [,5]
[1,]    5   13    8    3    0
[2,]   21   30   39    9    5
[3,]   41   60   40   17    9
[4,]    8   23   11    4    2
[5,]   36   48   22   13    1
[6,]   26   15    5    5    2
[7,]   27   24   10    7    1
[8,]    8   16    6    3    2
[9,]   28   40   32    9    5
```

Observe how R refers to the rows and columns of `cw.matrix`. If we wanted to extract the i^{th} row, then we call on `cw.matrix[i,]`. For example, the call `cw.matrix[8,]` would result in `c(8, 16, 6, 3, 2)`. Similarly, we can extract the j^{th} column by means of the call `cw.matrix[,j]`. Of course, if we want just one entry, then we specify its row and column. For example, `cw.matrix[5,2]` extracts the number 48.

Our next step is to name the rows and columns so we do not have to rely on memory. After taking that step, we tell R to create a contingency table called `country.western.crosstab` by treating the matrix `cw.matrix` as a table

```
> colnames(cw.matrix) = c("Strongly Likes", "Likes", "Mixed", "Dislikes", "Strongly Dislikes")
> rownames(cw.matrix) = c("New England", "Middle Atlantic", "E. North Central", "W. North Central",
+ "South Atlantic", "E. South Central", "W. South Central", "Mountain", "Pacific")
> cw.matrix
      Strongly Likes Likes Mixed Dislikes Strongly Dislikes
New England           5   13    8      3           0
Middle Atlantic       21   30   39      9           5
E. North Central      41   60   40     17           9
W. North Central       8   23   11      4           2
South Atlantic        36   48   22     13           1
E. South Central      26   15    5      5           2
W. South Central      27   24   10      7           1
Mountain              8   16    6      3           2
Pacific              28   40   32      9           5
> country.western.crosstab = as.table(cw.matrix)
```

Now for the coup de grâce. We use each of the functions `margin.table` and `prop.table` in three ways. The one-argument call `margin.table(country.western.crosstab)` calculates the sum of *all* joint frequencies in

the table and returns the table total. The two-argument call `margin.table(country.western.crosstab,1)` sums the entries by row. It returns the marginal distribution of the categorical variable whose values correspond to rows. This is the marginal distribution that is often adjoined to the table as a right margin. The two-argument call `margin.table(country.western.crosstab,2)` sums the entries by column. It returns the marginal distribution of the categorical variable whose values correspond to columns. This is the marginal distribution that is often adjoined to the table as a margin at the bottom.

The one-argument call `margin.table(country.western.crosstab)` calculates table relative frequencies. Every cell entry is divided by the table total. Multiply any of the resulting numbers by 100 to obtain its table percentage. The two-argument call `prop.table(country.western.crosstab,1)` calculates the row relative frequencies. Each row is a relative frequency conditional distribution. Multiply by 100 and you obtain the row percentages. The two-argument call `prop.table(country.western.crosstab,2)` calculates the column relative frequencies. Each column is a relative frequency conditional distribution. Multiply by 100 and you obtain the column percentages.

```
> margin.table(country.western.crosstab)
[1] 739
> margin.table(country.western.crosstab,1)
      New England  Middle Atlantic E. North Central W. North Central
              29             104             167             48
      South Atlantic E. South Central W. South Central      Mountain
              120             53             69             35
      Pacific
              114
> margin.table(country.western.crosstab,2)
      Strongly Likes      Likes      Mixed      Dislikes
              200             269             173             70
      Strongly Dislikes
              27
> prop.table(country.western.crosstab)
      Strongly Likes      Likes      Mixed      Dislikes      Strongly Dislikes
New England      0.0067659 0.0175913 0.0108254 0.0040595      0.0000000
Middle Atlantic      0.0284168 0.0405954 0.0527740 0.0121786      0.0067659
E. North Central      0.0554804 0.0811908 0.0541272 0.0230041      0.0121786
W. North Central      0.0108254 0.0311231 0.0148850 0.0054127      0.0027064
South Atlantic      0.0487145 0.0649526 0.0297700 0.0175913      0.0013532
E. South Central      0.0351827 0.0202977 0.0067659 0.0067659      0.0027064
W. South Central      0.0365359 0.0324763 0.0135318 0.0094723      0.0013532
Mountain      0.0108254 0.0216509 0.0081191 0.0040595      0.0027064
Pacific      0.0378890 0.0541272 0.0433018 0.0121786      0.0067659
> prop.table(country.western.crosstab,1)
      Strongly Likes      Likes      Mixed      Dislikes      Strongly Dislikes
New England      0.1724138 0.4482759 0.2758621 0.1034483      0.0000000
Middle Atlantic      0.2019231 0.2884615 0.3750000 0.0865385      0.0480769
E. North Central      0.2455090 0.3592814 0.2395210 0.1017964      0.0538922
W. North Central      0.1666667 0.4791667 0.2291667 0.0833333      0.0416667
South Atlantic      0.3000000 0.4000000 0.1833333 0.1083333      0.0083333
E. South Central      0.4905660 0.2830189 0.0943396 0.0943396      0.0377358
W. South Central      0.3913043 0.3478261 0.1449275 0.1014493      0.0144928
Mountain      0.2285714 0.4571429 0.1714286 0.0857143      0.0571429
Pacific      0.2456140 0.3508772 0.2807018 0.0789474      0.0438596
```

```
> prop.table(country.western.crosstab,2)
```

	Strongly Likes	Likes	Mixed	Dislikes	Strongly Dislikes
New England	0.025000	0.048327	0.046243	0.042857	0.000000
Middle Atlantic	0.105000	0.111524	0.225434	0.128571	0.185185
E. North Central	0.205000	0.223048	0.231214	0.242857	0.333333
W. North Central	0.040000	0.085502	0.063584	0.057143	0.074074
South Atlantic	0.180000	0.178439	0.127168	0.185714	0.037037
E. South Central	0.130000	0.055762	0.028902	0.071429	0.074074
W. South Central	0.135000	0.089219	0.057803	0.100000	0.037037
Mountain	0.040000	0.059480	0.034682	0.042857	0.074074
Pacific	0.140000	0.148699	0.184971	0.128571	0.185185

1.5 Simpson's Paradox

Stan Musial, aka “Stan the Man,” was an accomplished baseball player for the St. Louis Cardinals. Many years ago, when the author was relatively young, he came across some of Musial's batting statistics in an article. In Musial's first two years in the National League, 1941 and 1942, he had 20 hits in 47 at-bats (1941) and 147 hits in 467 at-bats (1942). Batting average is calculated by dividing the number of hits a batter obtains by the number of his at-bats. Musial's batting averages those two years were therefore $20/47$, or 0.426, and $147/467$, or 0.315. The author thought it might be interesting to compare those stats with those of one of the stars of the American League, Joe DiMaggio, aka “The Yankee Clipper” and “Joltin' Joe”, who also played in that era.

Player	Year	At Bats	Hits	Batting Average
Musial	1941	47	20	0.426
Musial	1942	467	147	0.315
DiMaggio	1941	541	193	0.357
DiMaggio	1942	610	186	0.305

Table 1.5.1: Batting Statistics of Musial and Joe DiMaggio

In each of the two years Musial had the higher batting average: 0.426 versus 0.357 in 1941, and 0.315 versus 0.305 in 1942. It stands to reason (famous last words) that Musial must have had the higher batting average over the two years combined. Of course, the average of Musial's yearly averages, namely $(0.426 + 0.315)/2$, or 0.371, is indeed higher than the comparable average, $(0.357 + 0.305)/2$, or 0.331, for DiMaggio. However, these averages of averages do not represent the players' aggregate batting averages for the two years. In 1941 and 1942, Musial had $20 + 147$ hits in $47 + 467$ at bats, for an aggregate batting average of $(20 + 147)/(47 + 467)$, or 0.325. DiMaggio's aggregate batting average for the two years was $(193 + 186)/(541 + 610)$, or 0.329, which is actually *greater* than Musial's 0.325. Eh?

According to our calculation, DiMaggio had a higher batting average than Musial over the two years, even though in each of those two years he had a lower average. It does not seem to make sense! The author's first reaction was to double-check the arithmetic and root out the error. The arithmetic, however, checked out. Here's to you, Joe DiMaggio—you did have a higher batting average than Stan the Man. Faced with an unexpected but inarguable fact, the author then thought to himself, *That's weird*. Then he turned his attention to other nerdy pursuits. Had the author had better instincts, he would have thought to himself, *Publish this interesting observation; someday, somebody might name the phenomenon Blank's Paradox*.

Actually, the paradox was already well-known to statisticians when the author rediscovered it. Edward Simpson described this paradox in 1951, and it is now usually called *Simpson's Paradox*. In fact, pioneering statistician George Udny Yule had observed Simpson's Paradox some fifty years before Simpson. He was not

even the first GUY to do so: Karl Pearson, one of the founding fathers of mathematical statistics, noticed Simpson's Paradox in 1899. The paradox that eventually would be called Simpson's Paradox was even described in a textbook seventeen years before Simpson wrote his paper. In *An Introduction to Logic and Scientific Method*, published by Harcourt, Brace, and World in 1934, authors Morris Cohen and Ernest Nagel cited mortality rates in 1910 from tuberculosis in Richmond, Virginia and New York, New York. As they observed: The death rate for African-Americans was lower in Richmond than in New York, the death rate for Caucasians was lower in Richmond than in New York, but the death rate for the total combined population of African-Americans and Caucasians was *higher* in Richmond than in New York.

To understand what causes Simpson's Paradox, let us use a simplifying model. Consider the hypothetical Shlabotnik brothers, Dominic and Joe, each of whom played two years in the minor leagues with the following hypothetical records:

Player	Year	At Bats	Hits	Batting Average
Dominic	2012	1	1	1.000
Dominic	2013	600	192	0.320
Joe	2012	600	222	0.370
Joe	2013	600	186	0.310

Table 1.5.2: Hypothetical Batting Statistics of the Shlabotnik Brothers

As you can see, Dominic had a batting higher average than Joe in each year. Joe's aggregate average for the two years is $(222 + 186)/(600 + 600)$, or 0.340. This figure is actually the average of his averages because he had the same number of at-bats in the two years under consideration.²¹ Notice that this aggregate average, 0.340, is quite a bit better than the 0.320 average of Dominic in the second of his two years. If Dominic's aggregate two year average was close to his 2013 average of 0.320, then Joe would have a higher two year aggregate average than Dominic despite trailing him in each individual year. Now, Dominic's 2013 average was based on 600 at-bats. We do not expect his average in the previous year, based as it was on only one at-bat, to substantially impact his aggregate average. Indeed, that aggregate average is $(1 + 192)/(1 + 600)$, or 0.321, which, as expected, is about the same as Dominic's single year 2013 average, and therefore considerably less than Joe's two year aggregate average. Here's to you, Joe Shlabotnik.

Additional examples of Simpson's Paradox will be mentioned, but a general description²² of the phenomenon is in order before we get to that.

Simpson's Paradox is a paradox in which a trend that appears in different groups of data disappears when these groups are combined, and the reverse trend appears for the aggregate data.

The Stanford Encyclopedia of Philosophy describes Simpson's Paradox this way²³:

An association between a pair of variables can consistently be inverted in each subpopulation of a population when the population is partitioned.

Simpson's Paradox received some public attention in the 1970s as the result of an allegation of sex bias in admissions to the Graduate School of the University of California–Berkeley. In the Fall of 1973, admissions data from Berkeley's Graduate School suggested that male applicants were more likely to be admitted than were female applicants. The reason for such a suspicion can be seen from the following table.

²¹Suppose that in group A a property holds a times out of x cases and that in group B the same property holds b times out of y cases. The frequencies for groups A and B are $f_A = a/x$ and $f_B = b/y$. The aggregate frequency for C , the merger of the two groups, is $f_C = (a + b)/(x + y)$. A bit of algebra shows that the average $(f_A + f_B)/2$ of the individual group averages is equal to the aggregate average f_C if and only if either the denominators x and y are the same or the individual averages f_A and f_B are the same.

²²As found in the Wikipedia page http://en.wikipedia.org/wiki/Simpson's_paradox Date retrieved: 25 May 2014

²³<http://plato.stanford.edu/entries/paradox-simpson/>

Sex	Applicants	Applicants Admitted	Percentage Admitted
Male	8442	3738	44.3%
Female	4321	1494	34.6%
Total	12763	5232	41.0%

Table 1.5.3: Percentage by Gender of Applicants Admitted to the Berkeley Graduate School, Fall 1973

Because of this data, Associate Dean E.A. Hammel of the Berkeley Graduate School initiated an investigation. Berkeley statistician, Peter Bickel, and computer programmer J.W. O’Connell were enjoined. The results of their inquiry received wider dissemination than a typical statistical study for a number of reasons: Women’s Studies was then a recently created academic discipline, the bias investigation concerned an important societal issue, Simpson’s Paradox was little known and therefore very surprising, and the conclusions of the investigation were published not in a statistical journal with a circulation limited to experts, but in the preeminent cross-discipline scientific journal, *Science*²⁴.

Assuming that male and female applicants were equally qualified and that Berkeley wished to admit 5232 applicants, which is to say 41.0% of the 12,763 applicants (as it in fact did), if the acceptance decisions were made without sex bias, then we would expect about 0.41×8442 , or 3461, male applicants accepted, and about 0.41×4321 , or 1771, female applicants accepted.

Sex	Actual Number	Expected Number	Surplus (+) or Shortfall (-)
Male	3738	3461	+277
Female	1494	1771	-277

Table 1.5.4: Applicants Admitted, Actual and Expected, to the Berkeley Graduate School, Fall 1973

The meaning of “expected” in this context should be clarified. If we were to flip a fair coin 100 times, then we would “expect” 50 heads. But we would also be aware that, by chance, we might not toss *exactly* 50 heads. If we observed only 47 heads, for example, then we would not suspect that we had been handed an unfair coin. Similarly, the Berkeley researchers were not surprised that the observed numbers of acceptances, 3738 and 1771, did not exactly match the expected numbers 3461 and 1771. Nevertheless, the difference of 277 between the numbers that were observed and the numbers that were expected was sufficiently large that the researchers ruled out “chance” as the explanation—as we shall see, Statistics provides us with methods for such determinations.

Having determined that there was something about the data that required explanation, the researchers turned their attention to the 101 individual admitting departments, which the researchers described as ‘independent, decision-making units.’ Because the mechanics of admissions precluded bias at the institutional level, the reasoning was that sex bias would be found in the admissions of at least one of the departments. Of the 101 departments²⁵, sixteen were immediately eliminated because they either had no women applicants or denied admission to no applicants of either sex. Of the 85 remaining departments, only 4 seemed to exhibit a bias that would contribute to the overall deficit of 277 women from the admitted list. However, these 4

²⁴See PJ Bickel, EA Hammel, and JW O’Connell, Sex Bias in Graduate Admissions: Data from Berkeley, *Science*, New Series, Vol. 187, No. 4175 (Feb. 7, 1975), pp. 398-404. Stable URL: <http://www.jstor.org/stable/1739581>. You have access to this article by first going to <http://www.jstor.org>, logging in with your university credentials, and then typing the name of the article in the search box.

²⁵Data for the six largest departments may be found in *Statistics*, D. A. Freedman, Robert Pisani, Roger Purves, W W Norton & Company Incorporated, 1978, pp. 12-15. The page numbers may differ in more recent editions. The authors, Berkeley statisticians who were prohibited by university policy from identifying the departments, referred to them using the letters A, B, C, D, E, F. It is that data set that has been cited most frequently in the subsequent literature.

departments together accounted for a deficit of only 26 women. Even more bewilderingly, six departments appeared to be biased in the opposite direction; together they accounted for a deficit of 64 men.

Having tried to account for a deficit of 277 women, the researchers found instead that there was a deficit of $64 - 26$, or 38, men. As they stated,

These results are confusing. After all, if the campus had a shortfall of 277 women in graduate admissions, and we look to see who is responsible, we ought to find somebody. So large a deficit ought not simply to disappear. There is even a suggestion of a surplus of women.

The researchers went on to say that they had “stumbled” onto Simpson’s Paradox. Remember that Simpson’s Paradox, like many paradoxes, is a paradox only when the examination of the data has been superficial. Just as we were able to resolve Simpson’s Paradox in the context of Stan Musial and Joe DiMaggio, so were the Berkeley researchers able to resolve their paradox. Looking more carefully, the researchers found two missing pieces to the puzzle: “The tendency of men and women to seek entry to different departments is marked” and “Not all departments are equally easy to enter.” The researchers were able to account for the deficit of 277 women without the assumption of sex bias. As they explained

The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into. Moreover this phenomenon is more pronounced in departments with large numbers of applicants.

It is interesting to note that, although Simpson’s Paradox drew scant attention before 1972, it received two high-profile observations in the 1970s. The Berkeley sex bias inquiry was one. The other concerned US taxation. Between 1974 and 1978, the tax rate *decreased* in each income category²⁶, yet the overall tax rate *increased* from 14.1% to 15.2%. The cause of this occurrence of Simpson’s Paradox was so well-understood that even politicians were able to discuss the matter. The decade of the 1970s was marked by high inflation in the USA (by historical American standards)²⁷. Wages increased as prices increased, with the result that taxpayers were pushed into higher income brackets.²⁸

The resolution of Simpson’s Paradox in each of the examples discussed involves a *confounding* or *lurking* variable. A final example of Simpson’s Paradox will illustrate this type of variable. Consider Table 1.5.5. The underlying variable is *Average SAT Score*. The row and column categories are *Year* and *Ethnic Group*.

Year	White	Black	Asian	Mexican	Puerto Rican	American Indian	Total
1981	519	412	474	438	437	471	504
2002	527	431	501	446	455	504	504

Table 1.5.5: Average SAT Scores (Source: *Math on Trial: How Numbers Get Used and Abused in the Courtroom*, Leila Schneps and Coralie Colmez, Basic Books, 2013)

²⁶There were five income categories: under \$5,000, \$5,000 to \$9,999, \$10,000 to \$14,999, \$15,000 to \$99,999, and \$100,000 or more.

²⁷For example, in 1970, the USPS domestic letter rate was \$0.06 for the first ounce. By 1978 that had risen by 250% to \$0.15. A further 20% increase to \$0.18 was implemented in 1981. In other words, during the eleven year period 1970–1981 the postal rate went up 300%. That amounted to an annual 9.99% inflation rate for postage (taking compounding into account) over the eleven years. At the time of this writing in 2014, the postal rate is \$0.49. That amounts to a 3.29% annual inflation rate in postage between 1978 and 2014. It may be of interest to contrast that with one example of textbook inflation. The author received his copy of the first (1978) edition of the statistics text by Freedman, Pisani, and Purves as an instructor’s freebie, but a postcard still tucked inside the book lists the price as \$13.95. The current (2014) price on Amazon is \$128.46. That amounts to a 6.17% annual inflation rate between 1978 and 2014. Publishing hematophagia, eh.

²⁸See Simpson’s Paradox in Real Life, Clifford H. Wagner, *The American Statistician*, Vol. 36, No. 1 (Feb., 1982), pp. 46–48, <http://www.jstor.org/stable/2684093>

Do you see the paradox? From 1981 to 2002 the average SAT score of every population group improved. Nevertheless, the overall average SAT score remained the same. The lurking variable, so-called in this context because its presence cannot be detected in Table 1.5.5, is the distribution of the total population among the ethnic groups. It is also known in this context as a confounding variable, because it results in a variable taking on a value that is contrary to our expectations. Refer to Table 1.5.6, and notice that the population group with the highest average score in 1981 was a much smaller segment of the total population in 2002.

Year	White	Black	Asian	Mexican	Puerto Rican	American Indian	Total
1981	85%	9%	3%	2%	1%	0%	100%
2002	65%	11%	10%	4%	1%	1%	92%

Table 1.5.6: Percentage Population by Ethnic Group

Exercises

Exercises 1–6 are concerned with the table that follows.

Age Group	Male	Female
< 20	5,078,695	4,853,746
20–29	18,024,284	17,899,947
30–39	18,346,087	18,291,443
40–49	20,306,348	20,248,986
50–59	19,101,742	19,382,381
60–69	12,982,415	13,255,192
70–79	6,881,694	7,271,261
> 79	3,540,548	4,153,617

Licensed Drivers (USA, 2009) by Gender and Age Group Source: Federal Highway Administration <https://www.fhwa.dot.gov/policyinformation/statistics/2009/dl20.cfm>, Retrieved 26 August 2014.

1. What population underlies the Licensed Drivers table? What two categorical variables are the basis of the table?
2. Is the Licensed Drivers table above a contingency table? Justify your answer.
3. What are the marginal distributions that are associated with the Licensed Drivers table? (They exist even though they have not been explicitly given.) How many licensed drivers were there in the USA in 2009?
4. What are the three percentages associated with the number 18,024,284? (Describe the meaning of these percentages unambiguously and give their values.)
5. Associated with the Licensed Drivers table are two conditional distributions that involve the number 7,271,261. What are these conditional distributions? (A complete answer includes more than numbers: it explains the variable that underlies the distribution as well as the “conditioning” values.)
6. Are the two categorical values featured in the table independent?

Exercises 7–10 are concerned with the results of a Pew Research Center telephone survey released in November, 2010. In the poll, 2625 adult Americans were read the statement, “There is only one true love for each person.” Surveyees were asked to respond with “Agree,” “Disagree”, or “Do not know.” The gender of surveyees was also recorded. As a result of the survey, cells of the following table were filled in.

	Male	Female	Total
Agree			
Disagree			
Do not know			
Total			

One True Love Survey, Pew Foundation, 18 November 2010.

7. Of the surveyees, 78 responded “Do not know.” There were 1,077 more surveyees who disagreed than agreed. Fill the table as a frequency table, so far as is possible from the given information.
8. In the survey, 363 females agreed, 1005 females disagreed, and 34 males did not know. Use this information to complete the table started in Exercise 7.
9. What proportion of females agree? What proportion of surveyees who agree are female? What proportion of surveyees are female and agree?
10. One cell entry in the frequency table is 807. What are the row, column, and table percentages for that entry? In words, what do these three percentages signify?
11. Eighth grade students in Nebraska (NE) and New Jersey (NJ) were divided into three groups, A, B, and C. An examination of mean math scores on the National Assessment of Educational Progress resulted in the following table:

	A	B	C	All
NE	281	236	259	277
NJ	283	242	260	271

Mean Eighth Grade Math Scores²⁹

If $NE(A)$, $NE(B)$, $NE(C)$, (respectively $NJ(A)$, $NJ(B)$, $NJ(C)$) are the number of eighth grade students in Nebraska (respectively New Jersey) in groups A, B, and C, what inequality involving these numbers can be inferred? (We know that NE is less populated than NJ, so we expect that $NE(A) < NJ(A)$, $NE(B) < NJ(B)$, and $NE(C) < NJ(C)$. These inequalities are neither here nor there.)

²⁹H. Wainer and L. Brown, *American Statistician*, **58** (2004), p.119.

Solutions to the Exercises

Chapter 1

1. The population under study is composed of all licensed drivers in the United States in the year 2009. The two categorical variables that give rise to the table are *Age Group* and *Gender*.
2. The table is indeed a contingency table. That is because all the categories of one of the variables are represented by the rows, all the categories of the other variable are represented by the columns, and every cell is filled with a joint frequency.
3. If, for each row, we add all the entries of that row, then we obtain the right margin, which is the (marginal in this context) distribution of *Age Group*:

< 20	20–29	30–39	40–49	50–59	60–69	70–79	> 79
9,932,441	35,924,231	36,637,530	40,555,334	38,484,123	26,237,607	14,152,955	7,694,165

If, for each column, we add all the entries of that column, then we obtain the bottom margin, which is the (marginal in this context) distribution of *Gender*:

Male	Female
104,261,813	105,356,573

The number of licensed drivers in the USA in 2009 can be obtained by adding the entries of either marginal distribution. The sum comes to 209,618,386 whichever way it is calculated.

4. The three percentages associated with the number 18,024,284 are the row percentage $(18,024,284/35,924,231) \times 100\%$, or 50.17 %, the column percentage $(18,024,284/104,261,813) \times 100\%$, or 17.29 %, and the table percentage $(18,024,284/209,618,386) \times 100\%$, or 8.60 %. (The divisors are respectively, the row sum of the row that contains 18,024,284, the column sum of the column that contains 18,024,284), and the sum of the entries of each of the two marginal distributions. The values of all three divisors were calculated in the preceding exercise.) The row percentage is the percentage of males among all licensed drivers aged 20–29. The column percentage is the percentage of male licensed drivers aged 20–29 among all licensed male drives. The table percentage is the percentage of licensed male drivers aged 20–29 among all licensed drivers.
5. The conditional distributions that involve the number 7,271,261 are the row and the column containing that number. That is

Male	Female
6,881,694	7,271,261

and

< 20	20–29	30–39	40–49	50–59	60–69	70–79	> 79
4,853,746	17,899,947	18,291,443	20,248,986	19,382,381	13,255,192	7,271,261	4,153,617

These conditional distributions are, respectively, the gender distribution of licensed drivers in the 70–79 age group, and the age group distribution of licensed female drivers.

6. Let us express in terms of percentages the conditional distributions of gender for each age group category. In other words, we will present the table of row percentages.

Age Group	Male	Female
< 20	51.13%	48.87%
20–29	50.17%	49.83%
30–39	50.07%	49.93%
40–49	50.07%	49.93%
50–59	49.64%	50.36%
60–69	49.48%	50.52%
70–79	48.62%	51.38%
> 79	46.02%	53.98%

If the population under consideration consisted only of licensed drivers between 20 and 69, then I'd declare Gender and Age Group to be independent. But the three pairs of numbers, 51.13 and 48.87, 48.62 and 51.38, and especially 46.02 and 53.98 show two gaps that are large enough that it is likely that Gender and Age Group *are* dependent in the youngest and oldest sectors of the population. That is enough to deduce dependence for the population under study.

7. Let a be the number of surveyees that agreed. Then the number that disagreed is $a + 1077$. The total number of surveyees, namely 2625, is therefore $a + (a + 1077) + 78$. We solve the equation $a + (a + 1077) + 78 = 2625$ to find $a = 735$ and $a + 1077 = 1812$. We can fill in the right margin of the table:

	Male	Female	Total
Agree			735
Disagree			1812
Do not know			78
Total			2625

8. In total, 78 surveyees did not know. Therefore, $78 - 34$, or 44 females did not know. Similarly, $735 - 363$, or 372 males agreed, and $1812 - 1005$, or 807 males disagreed. The frequency table is

	Male	Female	Total
Agree	372	363	735
Disagree	807	1005	1812
Do not know	34	44	78
Total	1213	1412	2625

9. Proportion of females who agree: $363/1412$, or 0.2571.
 Proportion of surveyees who agree that are female: $363/735$, or 0.4939.
 Proportion of surveyees who are female and agree: $373/2625$, or 0.1421.
10. Row percentage: $100 \cdot 807/1812$, or 44.54—the percentage of disagreeing surveyees who are male.
 Column percentage: $100 \cdot 807/1213$, or 66.53—the percentage of male surveyees who disagree.
 Table percentage: $100 \cdot 807/2625$, or 30.74—the percentage of surveyees who are male and disagree.
11. Simpson's Paradox! The explanation for the reversal of the combined group means is that the proportion of group A in NE is greater than the proportion of group A in NJ. That is,

$$\frac{NJ(A)}{NJ(A) + NJ(B) + NJ(C)} < \frac{NE(A)}{NE(A) + NE(B) + NE(C)}.$$

Solutions to the Chapter 1 Exercises

1. The population under study is composed of all licensed drivers in the United States in the year 2009. The two categorical variables that give rise to the table are *Age Group* and *Gender*.
2. The table is indeed a contingency table. That is because all the categories of one of the variables are represented by the rows, all the categories of the other variable are represented by the columns, and every cell is filled with a joint frequency.
3. If, for each row, we add all the entries of that row, then we obtain the right margin, which is the (marginal in this context) distribution of *Age Group*:

< 20	20–29	30–39	40–49	50–59	60–69	70–79	> 79
9,932,441	35,924,231	36,637,530	40,555,334	38,484,123	26,237,607	14,152,955	7,694,165

If, for each column, we add all the entries of that column, then we obtain the bottom margin, which is the (marginal in this context) distribution of *Gender*:

Male	Female
104,261,813	105,356,573

The number of licensed drivers in the USA in 2009 can be obtained by adding the entries of either marginal distribution. The sum comes to 209,618,386 whichever way it is calculated.

4. The three percentages associated with the number 18,024,284 are the row percentage $(18,024,284/35,924,231) \times 100\%$, or 50.17 %, the column percentage $(18,024,284/104,261,813) \times 100\%$, or 17.29 %, and the table percentage $(18,024,284/209,618,386) \times 100\%$, or 8.60 %. (The divisors are respectively, the row sum of the row that contains 18,024,284, the column sum of the column that contains 18,024,284), and the sum of the entries of each of the two marginal distributions. The values of all three divisors were calculated in the preceding exercise.) The row percentage is the percentage of males among all licensed drivers aged 20–29. The column percentage is the percentage of male licensed drivers aged 20–29 among all licensed male drivers. The table percentage is the percentage of licensed male drivers aged 20–29 among all licensed drivers.
5. The conditional distributions that involve the number 7,271,261 are the row and the column containing that number. That is

Male	Female
6,881,694	7,271,261

and

< 20	20–29	30–39	40–49	50–59	60–69	70–79	> 79
4,853,746	17,899,947	18,291,443	20,248,986	19,382,381	13,255,192	7,271,261	4,153,617

These conditional distributions are, respectively, the gender distribution of licensed drivers in the 70–79 age group, and the age group distribution of licensed female drivers.

6. Let us express in terms of percentages the conditional distributions of gender for each age group category. In other words, we will present the table of row percentages.

Age Group	Male	Female
< 20	51.13%	48.87%
20–29	50.17%	49.83%
30–39	50.07%	49.93%
40–49	50.07%	49.93%
50–59	49.64%	50.36%
60–69	49.48%	50.52%
70–79	48.62%	51.38%
> 79	46.02%	53.98%

If the population under consideration consisted only of licensed drivers between 20 and 69, then I'd declare Gender and Age Group to be independent. But the three pairs of numbers, 51.13 and 48.87, 48.62 and 51.38, and especially 46.02 and 53.98 show two gaps that are large enough that it is likely that Gender and Age Group *are* dependent in the youngest and oldest sectors of the population. That is enough to deduce dependence for the population under study.

7. Let a be the number of surveyees that agreed. Then the number that disagreed is $a + 1077$. The total number of surveyees, namely 2625, is therefore $a + (a + 1077) + 78$. We solve the equation $a + (a + 1077) + 78 = 2625$ to find $a = 735$ and $a + 1077 = 1812$. We can fill in the right margin of the table:

	Male	Female	Total
Agree			735
Disagree			1812
Do not know			78
Total			2625

8. In total, 78 surveyees did not know. Therefore, $78 - 34$, or 44 females did not know. Similarly, $735 - 363$, or 372 males agreed, and $1812 - 1005$, or 807 males disagreed. The frequency table is

	Male	Female	Total
Agree	372	363	735
Disagree	807	1005	1812
Do not know	34	44	78
Total	1213	1412	2625

9. Proportion of females who agree: $363/1412$, or 0.2571.
 Proportion of surveyees who agree that are female: $363/735$, or 0.4939.
 Proportion of surveyees who are female and agree: $373/2625$, or 0.1421.
10. Row percentage: $100 \cdot 807/1812$, or 44.54—the percentage of disagreeing surveyees who are male.
 Column percentage: $100 \cdot 807/1213$, or 66.53—the percentage of male surveyees who disagree.
 Table percentage: $100 \cdot 807/2625$, or 30.74—the percentage of surveyees who are male and disagree.
11. Simpson's Paradox! The explanation for the reversal of the combined group means is that the proportion of group A in NE is greater than the proportion of group A in NJ. That is,

$$\frac{NJ(A)}{NJ(A) + NJ(B) + NJ(C)} < \frac{NE(A)}{NE(A) + NE(B) + NE(C)}.$$