So you decided to take A.P. Statistics, Congratulations!

Attached is a copy of first two chapters of the textbook to use for your

required summer assignment.

This assignment is due your first day of class!

Here is your summer assignment:

1> **Read** chapter 1 – Introduction to Stats
Throughout the year we will follow the textbook very closely from this chapter up until chapter 26. You will be required to read each chapter. This chapter introduces you to how each problem or situation is approached, and the different types of data.

2> Page 7 JUST CHECKING – When reading each chapter make sure you can answer these questions. They tend to ask the most important questions that you are responsible for learning. **Answer these 2 questions on the answer sheet.**

3> At the end of each chapter make sure you **read** and understand these 3 parts:

i. What can go Wrong?
ii. What have we learned?
iii. Terms

It is a great chapter summary.

4> **Complete Exercises** pg. 10 #1, 3, 9, 13, 14, 17, 21, 25

6> **Read** chapter 2 – This chapter discusses qualitative data and the different types of graphs used to display the data.

7> Page 24 JUST CHECKING – **answer these questions**

8> **Complete the worksheet** on the back of this page

A Texas Instrument graphing calculator is required for this course. You should look into getting one before school starts. (Model number TI 84+)

If you have any questions you can email Mrs. Fisher at
jtribble1014@yahoo.com

Enjoy Your Summer and see you in September!

Please put all your answers on the answer sheets at the end of the packet

Welcome to A.P. Statistics with Mrs. Fisher

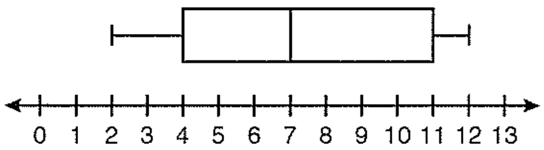
Here is some math you should remember from your previous math classes.
Place your answers and work on the answer sheet provided.

1> Find y when x = 4, $y = 135.798 - 2.642x$

2> In the equation of the line, $y = 135.798 - 2.642x$, what is the slope and y-intercept

3> If 98 students out of 149 students prefer Pepsi over coke, what percent of the students prefer Pepsi?

4> Based on the box-plot below, which statement is *false*?



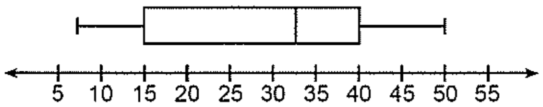
A> The median is 7

B> The range is 12

C> The first quartile is 4

D> The third quartile is 11

5> The box-and-whisker plot below represents the ages of 12 people.



What percentage of these people are age 15 or older?

A> 25 B> 35 C> 75 D> 85

6> Mr. Letona doesn't believe Mrs. Fisher's die is fair. He rolls the 6 sided die 20 times and comes up with following data. Create a dot-plot of the following results.

1	4	6	1	5	3	2	5	4	6
3	5	4	2	1	6	3	2	1	1

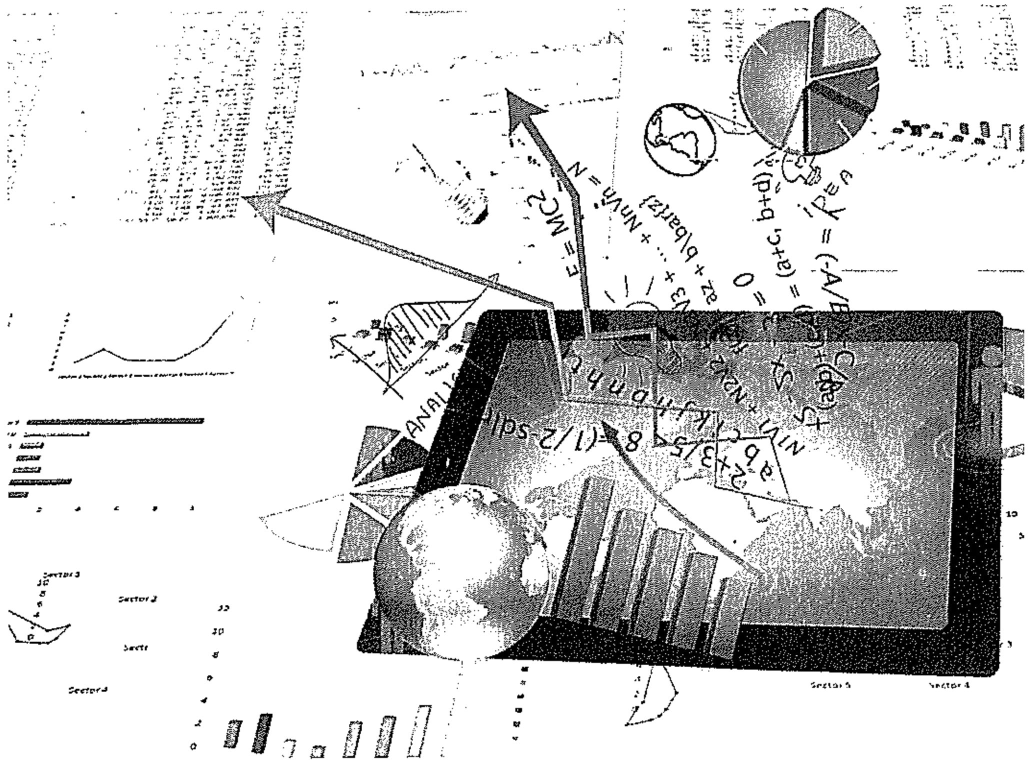
7> At a new dig site 30 artifacts were found by Archaeologists. The depth, in inches, at which the artifacts were found was recorded and given below.

44	51	11	42	76	36	64	37	43	41
62	36	74	51	72	37	28	38	61	47
36	41	22	37	51	46	85	29	53	41

Analyze the Data: what is the sample size, average, sample standard deviation, range, interquartile range, 5 number summary, are there any outliers?

chapter 1

Stats Starts Here[1]



"But where shall I begin?" asked Alice. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."

—*Lewis Carroll, Alice's Adventures in Wonderland*

Statistics gets no respect. People say things like "You can prove anything with Statistics." People will write off a claim based on data as "just a statistical trick." And a Statistics course may not be your friends' first choice for a fun elective.

But Statistics *is* fun. That's probably not what you heard on the street, but it's true. Statistics is about how to think clearly with data. We'll talk about data in more detail soon, but for now, think of **data** as any collection of numbers, characters, images, or other items that provide information about something. Whenever there are data and a need for understanding the world, you'll find Statistics. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

So, What Is (Are?) Statistics?

Q: What is Statistics?
A: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.
Q: What are statistics?
A: Statistics (plural) are particular calculations made from data.
Q: So what is data?
A: You mean, "what *are* data?" Data is the plural form. The singular is datum.
Q: OK, OK, so what are data?
A: Data are values along with their context.

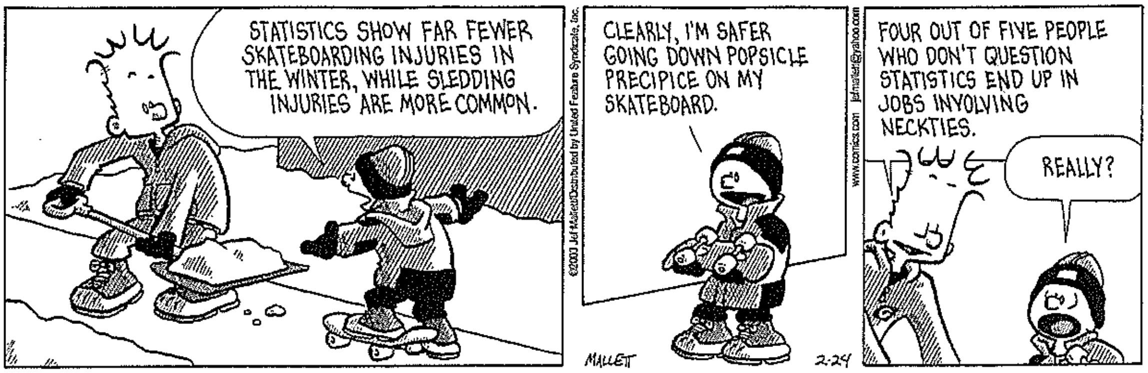
It seems every time we turn around, someone is collecting data on us, from every purchase we make in the grocery store, to every click of our mouse as we surf the Web.

Consider the following:

- If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to the *Wall Street Journal* (10/18/2010),[2] much of your personal information has probably been sold to marketing or tracking companies. Why would Facebook give you a free account and let you upload as much as you want to its site? Because your data are valuable! Using your Facebook profile, a company might build a profile of your

[1] We could have called this chapter "Introduction," but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this here, in the footnote, because nobody reads footnotes either.

[2] blogs.wsj.com/digits/2010/10/18/referers-how-facebook-apps-leak-user-ids/


STATISTICS SHOW FAR FEWER SKATEBOARDING INJURIES IN THE WINTER, WHILE SLEDDING INJURIES ARE MORE COMMON.
©2003 Jef Mallett/Distributed by United Feature Syndicate, Inc.
CLEARLY, I'M SAFER GOING DOWN POPSICLE PRECIPICE ON MY SKATEBOARD.
MALLETT
2-24
www.comics.com jefmallett@yahoo.com
FOUR OUT OF FIVE PEOPLE WHO DON'T QUESTION STATISTICS END UP IN JOBS INVOLVING NECKTIES.
REALLY?

Frazz © 2013 Jef Mallett. Distributed by Universal Uclick. Reprinted with permission. All rights reserved.

interests and activities: what movies and sports you like; your age, sex, education level, and hobbies; where you live; and, of course, who your friends are and what *they* like. From Facebook's point of view, your data are a potential gold mine. Gold ore in the ground is neither very useful nor pretty. But with skill, it can be turned into something both beautiful and valuable. What we're going to talk about in this book is how you can mine your own data and learn valuable insights about the world.

- Like many other retailers, Target stores create customer profiles by collecting data about purchases using credit cards. Patterns the company discovers across similar customer profiles enable it to send you advertising and coupons that promote items you might be particularly interested in purchasing. As valuable to the company as these marketing insights can be, some may prove startling to individuals. Recently coupons Target sent to a Minneapolis girl's home revealed she was pregnant before her father knew![3]
- How dangerous is texting while driving? Researchers at the University of Utah tested drivers on simulators that could present emergency situations. They compared reaction times of sober drivers, drunk drivers, and texting drivers.[4] The results were striking. The texting drivers actually responded more slowly and were more dangerous than those who were above the legal limit for alcohol.

Are You a Statistic?
The ads say, "Don't drink and drive; you don't want to be a statistic." But you can't be a statistic.

We say: "Don't be a datum."

In this book, you'll learn how to design and analyze experiments like this. You'll learn how to interpret data and to communicate the message you see to others. You'll also learn how to spot deficiencies and weaknesses in conclusions drawn by others that you see in newspapers and on the Internet every day. Statistics can help you become a more informed citizen by giving you the tools to understand, question, and interpret data.

Statistics in a Word

Statistics Is about Variation
Data vary because we don't see everything and because even what we do see and measure, we measure imperfectly.

So, in a very basic way, Essential Statistics is about the real, imperfect world in which we live.

It can be fun, and sometimes useful, to summarize a discipline in only a few words. So,

Economics is about . . . *Money (and why it is good).*

Psychology: *Why we think what we think (we think).*

Biology: *Life.*

Anthropology: *Who?*

History: *What, where, and when?*

Philosophy: *Why?*

Engineering: *How?*

Accounting: *How much?*

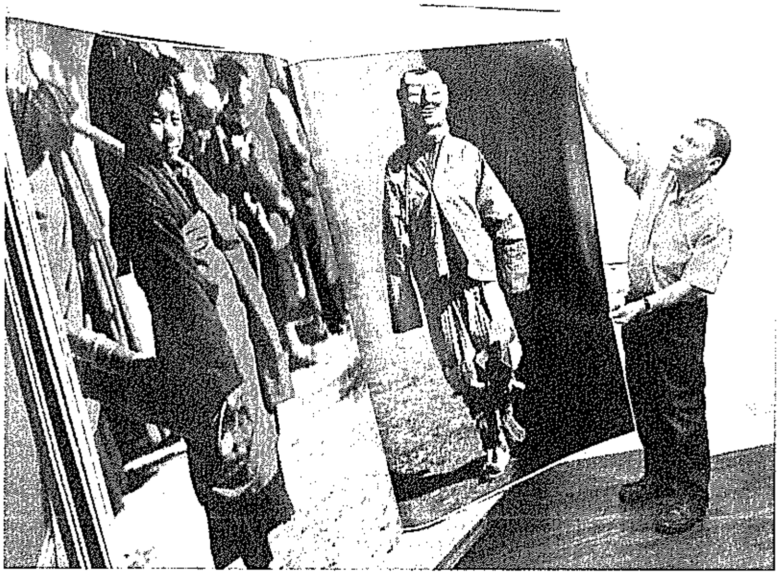
In such a caricature, Statistics is about . . . ***Variation.***

[3] http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

[4] "Text Messaging During Simulated Driving," Drews, F. A. et al. Human Factors: hfs.sagepub.com/content/51/5/762

Data vary. Ask different people the same question and you'll get a variety of answers. Statistics helps us to make sense of the world described by our data by seeing past the underlying variation to find patterns and relationships. This book will teach you skills to help with this task and ways of thinking about variation that are the foundation of sound reasoning about data.

But What *Are* Data?



Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2010, the company's sales reached $34.2 billion (a nearly 40% increase from the previous year). Amazon has sold a wide variety of merchandise, including a $400,000 necklace, yak cheese from Tibet, and the largest book in the world. How did Amazon become so successful and how can it keep track of so many customers and such a wide variety of products? The answer to both questions is *data.*

But what are data? Think about it for a minute. What exactly *do* we mean by "data"? Do data have to be numbers? The amount of your last purchase in dollars is numerical data. But your name and address in Amazon's database are also data even though they are not numerical. What about your ZIP code? That's a number, but would Amazon care about, say, the *average* ZIP code of its customers?

Let's look at some hypothetical values that Amazon might collect:

105-2686834-3759466	Ohio	Nashville	Kansas	10.99	440	N	B00000I5Y6	Katherine H.
105-9318443-4200264	Illinois	Orange County	Boston	16.99	312	Y	B000002BK9	Samuel P.
105-1872500-0198646	Massachusetts	Bad Blood	Chicago	15.98	413	N	B000068ZVQ	Chris G.
103-2628345-9238664	Canada	Let Go	Mammals	11.99	902	N	B000001OAA	Monique D.
002-1663369-6638649	Ohio	Best of Kansas	Kansas	10.99	440	N	B002MXA7Q0	Katherine H.

A S *Activity*: What Is (Are) Data? Do you really know what are data and what are just numbers?

Try to guess what they represent. Why is that hard? Because there is no *context.* If we don't know what values are measured and what is measured about them, the values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

Order Number	Name	State/Country	Price	Area Code	Previous Album Download	Gift?	ASIN	New Purchase Artist
105-2686834-3759466	Katherine H.	Ohio	10.99	440	Nashville	N	B0000015Y6	Kansas
105-9318443-4200264	Samuel R	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
105-1372500-0198646	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals
002-1663369-6638649	Katherine H.	Ohio	10.99	440	Best of Kansas	N	B002MXA7Q0	Kansas

The W's:
Who
What
and in what units
When
Where
Why
How

Now we can see that these are purchase records for album download orders from Amazon. The column titles tell what has been recorded. Each row is about a particular purchase.

What information would provide a **context**? Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who, what, when, where,* and (if possible) *why*. Often, we add *how* to the list as well. The answers to the first two questions are essential. If we don't know *what* values are measured and *who* those values are measured on, the values are meaningless.

Who and *What*

In general, the rows of a data table correspond to individual **cases** about *Who*m (or about which—if they're not people) we record some characteristics. Cases go by different names, depending on the situation.

- Individuals who answer a survey are called **respondents**.
- People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**.
- Animals, plants, websites, and other inanimate subjects are often called **experimental units**.
- Often we simply call cases what they are: for example, *customers, economic quarters*, or *companies.*
- In a database, rows are called **records**—in this example, purchase records. Perhaps the most generic term is *cases*, but in any event the rows represent the *who* of the data.

The characteristics recorded about each individual are called **variables.** These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. *Name*, *Price*, *Area Code*, and whether the purchase was a *Gift* are some of the variables Amazon collected data for. Variables may seem simple, but we'll need to take a closer look soon.

We must know *who* and *what* to analyze data. Without knowing these two, we don't have enough information to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world. If possible, we'd like to know the *when* and *where* of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico. And knowing *why* the data were collected can tell us much about its reliability and quality.

A S **Activity: Consider the context** . . .Can you tell who's *Who* and what's *What*? And *Why*? This activity offers real-world examples to help you practice identifying the context.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

For Example IDENTIFYING THE "WHO"

In December 2011, *Consumer Reports* published an evaluation of 25 tablets from a variety of manufacturers.

QUESTION: Describe the population of interest, the sample, and the *Who* of the study.

ANSWER: The magazine is interested in the performance of tablets currently offered for sale. It tested a sample of 25 tablets, which are the "Who" for these data. Each tablet selected represents all tablets of that model offered by that manufacturer.

How the Data Are Collected

How the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics is the design of sound methods for collecting data.[5] Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. Identifying the W's is a habit we recommend.

The first step of any data analysis is to know what you are trying to accomplish and what you want to know. To help you use Statistics to understand the world and make decisions, we'll lead you through the entire process of *thinking* about the problem, *showing* what you've found, and *telling* others what you've learned. Every guided example in this book is broken into these three steps: *Think*, *Show*, and *Tell*. Identifying the problem and the *who* and *what* of the data is a key part of the *Think* step of any analysis. Make sure you know these before you proceed to *Show* or *Tell* anything about the data.

A S *Activity:* **Collect data in an experiment on yourself.** With the computer, you can experiment on yourself and then save the data. Go on to the subsequent related activities to check your understanding.

More About Variables (*What?*)

The Amazon data table displays information about several variables: *Order Number*, *Name*, *State/Country*, *Price*, and so on. These identify *what* we know about each individual. Variables such as these can play different roles, depending on how we plan to use them. While some are merely identifiers, others may be categorical or quantitative. Making that distinction is an important step in our analysis.

Privacy and the Internet You have many Identifiers: a social security number, a student ID number, possibly a passport number, a health insurance number, and probably a Facebook account name. Privacy experts are worried that Internet thieves may match your identity in these different areas of your life, allowing, for example, your health, education, and financial records to be merged. Even online companies such as Facebook and Google are able to link your online behavior to some of these identifiers, which carries with it both advantages and dangers. The National Strategy for Trusted Identities in Cyberspace (www.wired.com/images_blogs/threatlevel/2011/04/NSTIC strategy_041511.pdf) proposes ways that we may address this challenge in the near future.

Identifiers

For some variables, such as a *student ID*, each individual receives a unique value. We call a variable like this, an **identifier variable**. Identifiers are useful, but not typically for analysis.

Amazon wants to know who you are when you sign in again and doesn't want to confuse you with some other customer. So it assigns you a unique identifier. Amazon also wants to send you the right product, so it assigns a unique Amazon Standard Identification Number (ASIN) to each item it carries. Identifier variables themselves don't tell us anything useful about their categories because we know there is exactly one individual in each. You'll want to recognize when a categorical variable is playing the role of an identifier so you aren't tempted to analyze it.

Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? What color are your eyes? We call variables like these **categorical variables.**[6] Some variables are clearly categorical, like the variable *State/Country*. Its values are text and those values tell us what category the particular case falls into. Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" or "What is your marital status?" yield categorical values. But numerals are often used to label categories, so categorical variable values can also be numerals. For example, Amazon collects telephone area codes that *categorize* each phone number into a geographical region. So area code is considered a categorical variable even though it has numeric values.

[5] Coming attractions: to be discussed in Part III. We sense your excitement.

[6] You may also see them called *qualitative* variables.

Quantitative Variables

When a variable contains measured numerical values with measurement *units*, we call it a **quantitative variable.** Quantitative variables typically record an amount or degree of something. For a quantitative variable, its measurement **units** provide a meaning for the numbers. Even more important, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in Euros, dollars, pennies, yen, or Estonian krooni.

Either/Or?

Some variables with numeric values can be treated as either categorical or quantitative depending on what we want to know. Amazon could record your *Age* in years. That seems quantitative, and it would be if the company wanted to know the average age of those customers who visit their site after 3 A.M. But suppose Amazon wants to decide which album to feature on its site when you visit. Then thinking of your age in one of the categories Child, Teen, Adult, or Senior might be more useful. So, sometimes whether a variable is treated as categorical or quantitative is more about the question we want to ask rather than an intrinsic property of the variable itself.

Suppose a course evaluation survey asks, "How valuable do you think this course will be to you?" 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. Or if she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative.

But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but the teacher will have to imagine that it has "educational value units," whatever they are. Because there are no natural units, she should be cautious. Variables that report order without natural units are often called *ordinal variables*. But saying "that's an ordinal variable" doesn't get you off the hook. You must still look to the *why* of your study and understand what you want to learn from the variable to decide whether to treat it as categorical or quantitative.

A S *Activity*: **Recognize variables measured in a variety of ways.** This activity shows examples of the many ways to measure data.

A S *Activities*: **Variables.** Several activities show you how to begin working with data in your statistics package.

For Example IDENTIFYING "WHAT" AND "WHY" OF TABLETS

RECAP: A *Consumer Reports* article about 25 tablet computers lists each tablet's manufacturer, cost, battery life (hrs.), operating system (iOS/Android/RIM), and overall performance score (0–100).

QUESTION: Are these variables categorical or quantitative? Include units where appropriate, and describe the "Why" of this investigation.

ANSWER: The variables are

- manufacturer (categorical)
- cost (quantitative, $)
- battery life (quantitative, hrs.)
- operating system (categorical)
- performance score (quantitative, no units)

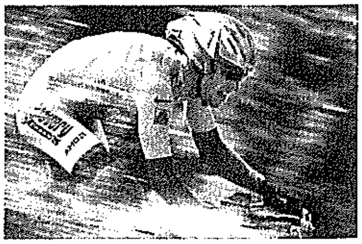
The magazine hopes to provide consumers with the information to choose a good tablet.

Just Checking

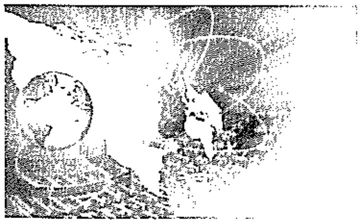
In the 2004 Tour de France, Lance Armstrong made history by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and set a new record for the fastest average speed—41.65 kilometers per hour. A cancer survivor, Armstrong became an international celebrity. But it was all too good to be true. In 2012, following revelations of doping, the International Cycling Union stripped Armstrong of all of his titles and records and banned him from professional cycling for life.

You can find data on all the Tour de France races on the DVD. Keep in mind that the entire data set has over 100 entries.

1. List as many of the W's as you can for this data set.
2. Classify each variable as categorical or quantitative; if quantitative, identify the units.



Year	Winner	Country of Origin	Total Time (h/min/s)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	94.33.00	25.3	6	2428	60	21
1904	Henri Cornet	France	96.05.00	24.3	6	2388	88	23
1905	Louis Trousseller	France	112.18.09	27.3	11	2975	60	24
⋮								
1999	Lance Armstrong (DQ)	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong (DQ)	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong (DQ)	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong (DQ)	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong (DQ)	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong (DQ)	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong (DQ)	USA	86.15.02	41.65	21	3608	189	155
⋮								
2011	Cadel Evans	Australia	86.12.22	39.788	21	3430	198	167
2012	Bradley Wiggins	Great Britain	87.34.47	39.928	20	3497	219	153
2013	Chris Froome	Great Britain	83.56.40	40.551	21	3404	219	170



A S *Self-Test*: **Review concepts about data.**

Like the Just Checking sections of this textbook, but interactive. (Usually, we won't reference the *ActivStats* self-tests here, but look for one whenever you'd like to check your understanding or review material.

There's a World of Data on the Internet These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the data sets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a website. We'll sometimes suggest search terms and offer other guidance.

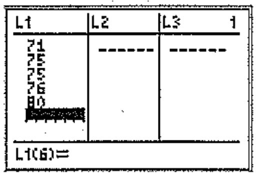
Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and extra symbols such as money indicators ($, ¥, £); few statistics packages can handle these.

WHAT CAN GO WRONG?

- **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- **Just because your variable's values are numbers, don't assume that it's quantitative.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. Think about *how* the data were collected. People who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan website. The question that respondents answered may be posed in a way that influences responses.

TI Tips WORKING WITH DATA

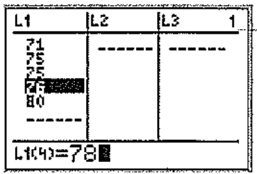
You'll need to be able to enter and edit data in your calculator. Here's how:



```
L1     L2     L3    1
71     ------ ------
75
75
76
80
------
L1(6)=
```

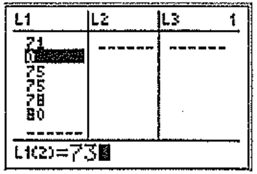
TO ENTER DATA: Hit the STAT button, and choose EDIT from the menu. You'll see a set of columns labeled L1, L2, and so on. Here is where you can enter, change, or delete a set of data.

Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under L1, type in 71, and hit ENTER (or the down arrow). There's the first player. Now enter the data for the rest of the team.



```
L1     L2     L3    1
71     ------ ------
75
75
76
80
------
L1(4)=78
```

TO CHANGE A DATUM: Suppose the 76" player grew since last season; his height should be listed as 78". Use the arrow keys to move the cursor onto the 76, then change the value and ENTER the correction.

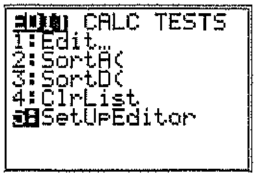


```
L1     L2     L3    1
71     ------ ------
0
75
75
78
80
------
L1(2)=73
```

TO ADD MORE DATA: We want to include the sixth man, 73" tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit 2ND INS, then ENTER the 73 in the new space.

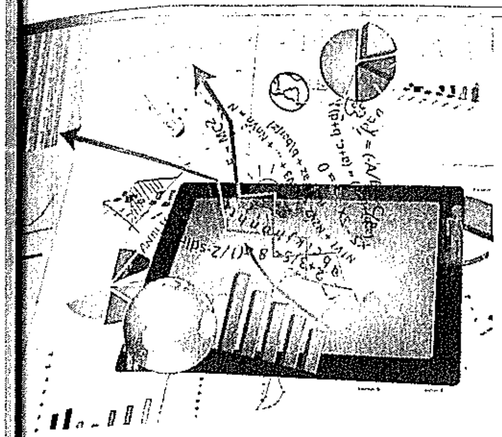
TO DELETE A DATUM: The 78" player just quit the team. Move the cursor there. Hit DEL. Bye.

TO CLEAR THE DATALIST: Finished playing basketball? Move the cursor atop the L1. Hit CLEAR, then ENTER (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.



```
EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```

LOST A DATALIST? Oops! Is L1 now missing entirely? Did you delete L1 by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the STAT EDIT menu, and run SetUpEditor to recreate all the lists.



What Have We Learned?

We've learned that data are information in a context.

- The W's help nail down the context *Who, What, When, Why, Where,* and *hoW.*
- We must know at least the *Who, What,* and *hoW* to be able to say anything useful based on the data. The *Who* are the cases. The *What* are the *variables.* A variable gives information about each of the cases. The *hoW* helps us decide whether we can trust the data.

We treat variables in two basic ways: as *categorical* or *quantitative.*

- Categorical variables identify a category for each case. Usually, we think about the counts of cases that fall into each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something; they must have *units.*
- Sometimes we treat a variable as categorical or quantitative depending on what we want to learn from it, which means that some variables can't be pigeonholed as one type or the other. That's an early hint that in Statistics we can't always pin things down precisely.

Terms

Data Systematically recorded information, whether numbers or labels, together with its context. (p. 1)

Data table An arrangement of data in which each row represents a case and each column represents a variable. (p. 3)

Context The context ideally tells *Who* was measured, *What* was measured, *How* the data were collected, *Where* the data were collected, and *When* and *Why* the study was performed. (p. 4)

Case A case is an individual about whom or which we have data. (*Who*). (p. 4)

Respondent Someone who answers, or responds to, a survey. (p. 4)

Subject A human experimental unit. Also called a participant. (p. 4)

Participant A human experimental unit. Also called a subject. (p. 4)

Experimental unit An individual in a study for which or for whom data values are recorded. Human experimental units are usually called subjects or participants. (p. 4)

Record Information about an individual in a database. (p. 4)

Variable A variable holds information about the same characteristic for many cases. (*What*). (p. 4)

Sample The cases we actually examine in seeking to understand the much larger population. (p. 4)

Population All the cases we wish we knew about. (p. 4)

Identifier variable A categorical variable that records a unique value for each case, used to name or identify it. (p. 5)

Categorical variable A variable that names categories (whether with words or numerals) is called categorical. (p. 5)

Quantitative variable A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units. (p. 6)

Units A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams. (p. 6)

On the Computer DATA

Activity: **Examine the Data.** Take a look at your own data from your experiment (p. 8) and get comfortable with your statistics package as you find out about the experiment test results.

> "Computers are useless; they can only give you answers."
>
> —*Pablo Picasso*

Most often we find statistics on a computer using a program, or *package*, designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a *tab* character and the delimiter that marks the end of a case to be a *return* character.
- Where to put the data. (Usually this is handled automatically.)
- What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

Exercises

1. **Voters** A February 2010 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?

2. **Job hunting** A June 2011 Gallup Poll asked Americans, "Thinking about the job situation in America today, would you say that it is now a good time or a bad time to find a quality jobs?" The choices were "Good time" or "Bad time". What kind of variable is the response?

3. **Medicine** A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?

4. **Stress** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

(Exercises 5–12) For each description of data, identify Who and What were investigated and the population of interest.

5. **The news** Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, answer the questions above. Include a copy of the article with your report.

6. **The Internet** Find an Internet source that reports on a study and describes the data. Print out the description and answer the questions above.

7. **Bicycle safety** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. [*NY Times*, Dec. 10, 2006]

8. **Investments** Some companies offer 401(k) retirement plans to employees, permitting them to shift part of their before-tax salaries into investments such as mutual funds. Employers typically match 50% of the employees'

contribution up to about 6% of salary. One company, concerned with what it believed was a low employee participation rate in its 401(k) plan, sampled 30 other companies with similar plans and asked for their 401(k) participation rates.

9. **Honesty** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University alternately taped two posters over the coffee station. During one week, it was a picture of flowers; during the other, it was a pair of staring eyes. They found that the average contribution was significantly higher when the eyes poster was up than when the flowers were there. Apparently, the mere feeling of being watched—even by eyes that were not real—was enough to encourage people to behave more honestly. [*NY Times*, Dec. 10, 2006]

10. **Biological Instinct** A study published by a team of American and Canadian psychologists found that during ovulation, a woman can identify a man's sexual orientation simply by looking at his face. To explore the subject, the authors conducted several investigations, the first of which involved 40 undergraduate women who were asked to guess the sexual orientation of 80 men based on photos of their faces. Half of the men were gay, and the other half were straight. All held similar expressions in the photos and were deemed to be equally attractive. The results of the study revealed that the closer a woman was to her peak ovulation, the more accurate her assessment.

    (*Source:* http://news.yahoo.com/ovulating-women-better-gaydar-183800184.html)

11. **Blindness:** A study begun in 2011 examines the use of stem cells in treating two forms of blindness, Stargardt's disease, and dry age-related macular degeneration. Each of the 24 patients entered one of two separate trials in which embryonic stem cells were to be used to treat the condition.

12. **Molten iron** The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. The company is interested in keeping the pouring temperature of the molten iron (in degrees Fahrenheit) close to the specified value of 2550 degrees. Cleveland Casting measured the pouring temperature for 10 randomly selected crankshafts.

(Exercises 13–26) For each description of data, identify the W's, name the variables, specify for each variable whether its use indicates that it should be treated as categorical or quantitative, and, for any quantitative variable, identify the units in which it was measured (or note that they were not provided).

13. **Weighing bears** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.

14. **Schools** The State Education Department requires local school districts to keep these records on all students: age, race or ethnicity, days absent, current grade level, standardized test scores in reading and mathematics, and any disabilities or special educational needs.

15. **Arby's menu** A listing posted by the Arby's restaurant chain gives, for each of the sandwiches it sells, the type of meat in the sandwich, the number of calories, and the serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.

16. **Age and party** Gallup conducted a series of telephone polls involving 20,392 American adults during 2011. Among the reported results were the voters' gender, age, race, party affiliation, whether they were of Hispanic ethnicity, education, region, adults in the household, and phone status (cell phone only/landline only/both, cell phone mostly, and having an unlisted landline number).

17. **Babies** Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998–2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

18. **Flowers** In a study appearing in the journal *Science*, a research team reports that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years show that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.

19. **Herbal medicine** Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of the benefits of the compound.

20. **Vineyards** Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

21. **Streams** In performing research for an ecology class, students at a college in upstate New York collect data on local streams each year. They record a number of biological,

chemical, and physical variables, including the stream name, the substrate of the stream (limestone, shale, or mixed), the acidity of the water (pH), the temperature (°C), and the BCI (a numerical measure of biological diversity).

22. **Fuel economy** The Environmental Protection Agency (EPA) tracks fuel economy of automobiles based on information from the manufacturers (Ford, Toyota, etc.). Among the data the agency collects are the manufacturer, vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.

23. **Refrigerators** In 2012, *Consumer Reports* rated bottom-freezer refrigerators. It listed 102 models, giving the brand, cost, size (cu ft), temperature performance, noise (poor, fair, etc.), ease of use, energy efficiency, estimated annual energy cost, an overall rating (good, excellent, etc.), and the exterior dimensions.

24. **Walking in circles** People who get lost in the desert, mountains, or woods often seem to wander in circles rather than walk in straight lines. To see whether people naturally walk in circles in the absence of visual clues, researcher Andrea Axtell tested 32 people on a football field. One at a time, they stood at the center of one goal line, were blindfolded, and then tried to walk to the other goal line. She recorded each individual's sex, height, handedness, the number of yards each was able to walk before going out of bounds, and whether each wandered off course to the left or the right. No one made it all the way to the far end of the field without crossing one of the sidelines. [*STATS* No. 39, Winter 2004]

25. **Kentucky Derby 2012** The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs, Louisville, Kentucky. The race started as a 1.5-mile race, but in 1896, it was shortened to 1.25 miles because experts felt that 3-year-old horses shouldn't run such a long race that early in the season. (It has been run in May every year but one—1901—when it took place on April 29). Here are the data for the first four and several recent races.

Year	Winner	Jockey	Trainer	Owner	Time
2013	Orb	J. Rosario	C. McGaughey III	Phipps/Janney	2:02.89
2012	I'll Have Another	M. Gutierrez	D. O'Neill	Reddam Racing	2:01.83
2011	Animal Kingdom	J. Velazquez	H. G. Motion	Team Valor	2:02.04
2010	Super Saver	C. Borel	T. Pletcher	WinStar Farm	2:04.45
2009	Mine That Bird	C. Borel	B. Woolley	Double Eagle Ranch	2:02.66
. . .					
1878	Day Star	J. Carter	L. Paul	T.J. Nichols	2:37.25
1877	Baden Baden	W. Walker	E. Brown	Daniel Swigert	2:38
1876	Vagrant	R. Swim	J. Williams	William Astor	2:38.25
1875	Aristides	O. Lewis	A. Williams	H.P. McGrath	2:37.75

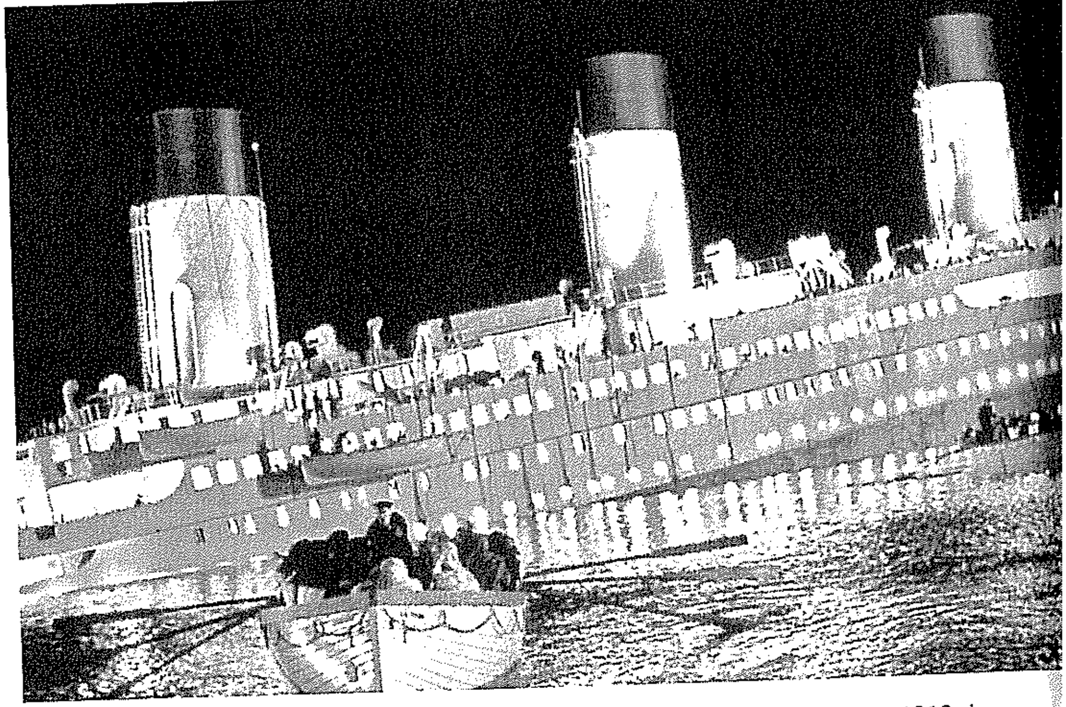
26. **Indy 2013** The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day weekend nearly every year since 1911. Even during the first race, there were controversies. Ralph Mulford was given the checkered flag first but took three extra laps just to make sure he'd completed 500 miles. When he finished, another driver, Ray Harroun, was being presented with the winner's trophy, and Mulford's protests were ignored. Harroun averaged 74.6 mph for the 500 miles. In 2013, the winner, Tony Kanaan, averaged 187.433 mph.

Here are the data for the first five races and five recent Indianapolis 500 races.

Year	Winner	Time	Average Speed (mph)
1911	Ray Harroun	6:42:08.039	74.602
1912	Joe Dawson	6:21:06.144	78.719
1913	Jules Goux	6:35:05.108	75.933
1914	René Thomas	6:03:45.060	82.474
1915	Ralph DePalma	5:33:55.619	89.840
. . .			
2009	Hélio Castroneves	3:19:34.6427	150.318
2010	Dario Franchitti	3:05:37.0131	161.623
2011	Dan Wheldon	2:56:11.7267	170.265
2012	Dario Franchitti	2:58:51	167.734
2013	Tony Kanaan	2:40:03.4181	187.433

chapter 2

Displaying and Describing Categorical Data



Who	People on the *Titanic*
What	Survival status, age, sex, ticket class
When	April 14, 1912
Where	North Atlantic
How	A variety of sources and Internet sites
Why	Historical interest

What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet's cry of "Iceberg, right ahead" and the three accompanying pulls of the crow's nest bell signaled the beginning of a nightmare that has become legend. By 2:15 A.M., the *Titanic*, thought by many to be unsinkable, had sunk, leaving more than 1500 passengers and crew members on board to meet their icy fate.

Here are some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person's *Survival* status (Dead or Alive), *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

Table 2.1

Part of a data table showing four variables for nine people aboard the *Titanic*

Survival	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

The problem with a data table like this—and in fact with all data tables—is that you can't *see* what's going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

A S *Video*: **The Incident** tells the story of the *Titanic*, and includes rare film footage.

The Three Rules of Data Analysis

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *Think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *Tell* others about your data is with a well-chosen picture.

These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.

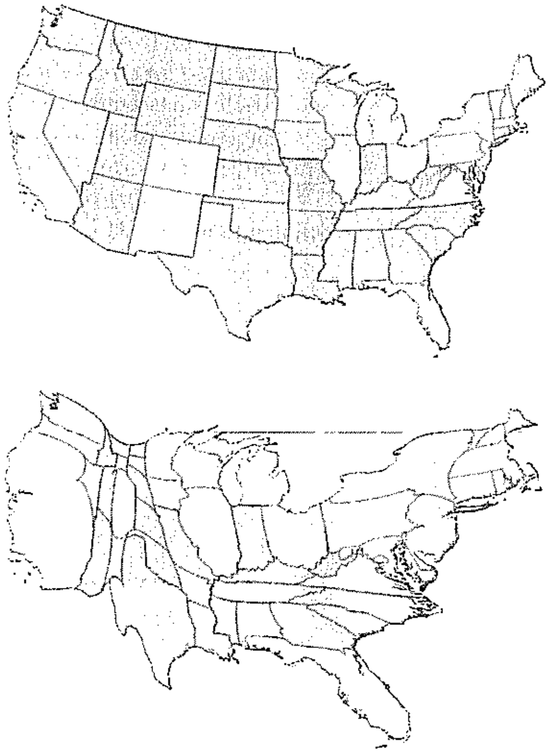


Figure 2.1

A picture to tell a story In November 2012, Barack Obama was re-elected as president of the United States. News reports commonly showed the election results with maps like the one on top, coloring states won by Obama blue and those won by his opponent Mitt Romney red. Even though Romney lost, doesn't it look like there's more red than blue? That's because some of the larger states like Montana and Wyoming have far fewer voters than some of the smaller states like Maryland and Connecticut. The strange-looking map on the bottom cleverly distorts the states to resize them proportional to their populations. By sacrificing an accurate display of the land areas, we get a better impression of the votes cast, giving us a clear picture of Obama's victory.
(*Source:* www-personal.umich.edu/~mejn/election/2012/)

The Area Principle

The best data displays, like the distorted electoral map above, observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. But a bad picture can distort our understanding rather than help it. On the next page is a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn't seem right. What's wrong? The lengths of the ships *do* match the totals in the table. However, experience and psychological tests

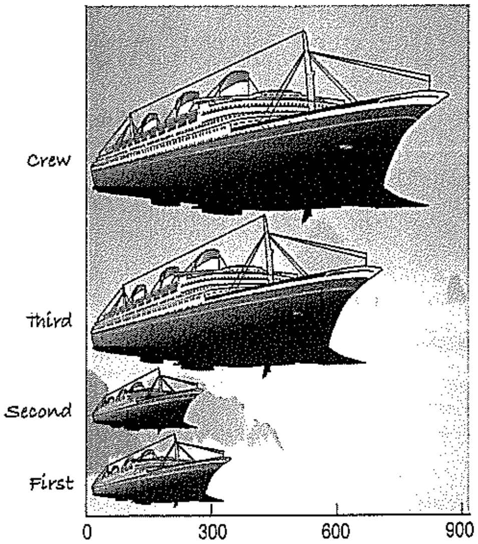


Figure 2.2

How many people were in each class on the *Titanic*? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

show that our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it's the associated *area* in the image that we notice. There were about 3 times as many crew as second-class passengers, and the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. That just isn't a correct impression.

Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say err) with Statistics.

A S *Activity*: **Make and Examine a Table of Counts.** Even data on something as simple as hair color can reveal surprises when you organize it in a data table.

Frequency Tables: Making Piles

To make an accurate picture of data, the first thing we have to do is to make piles. We pile together things that seem to go together, so we can see how the cases distribute across different categories. For categorical data, piling is easy. We just count the number of cases corresponding to each category and put them in a table.

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table,** which records the totals and the category names. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."

Class	Count
First	325
Second	285
Third	706
Crew	885

Table 2.2

A frequency table of the *Titanic* passengers

Even when we have thousands of cases, a variable like ticket *Class,* with only a few categories, has a frequency table that's easy to read.

For a variable with dozens or hundreds of categories, a frequency table will be much harder to read. You might want to combine categories into larger headings. For example, instead of counting the number of students from each state, you might group the states into regions like "Northeast," "South," "Midwest," "Mountain States," and "West." If the number of cases in several categories is relatively small, you can put them together into one category labeled "Other."

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Table 2.3

A relative frequency table for the same data

Counts are useful, but sometimes we want to know the fraction or **proportion** of the data in each category, so we divide the counts by the total number of cases. Usually we multiply by 100 to express these proportions as **percentages. A relative frequency table** displays the *percentages,* rather than the counts, of the values in each category. Both types of tables show how the cases are distributed across the categories. In this way, they describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

Bar Charts

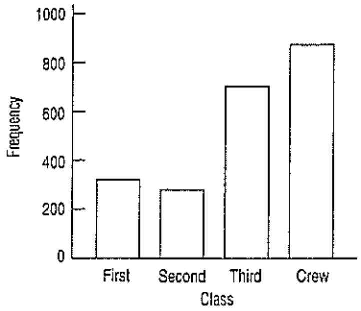
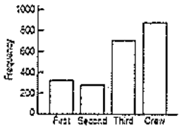


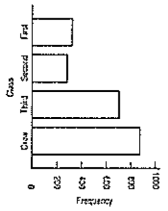
Figure 2.3

People on the *Titanic* by Ticket *Class* With the area principle satisfied, we can see the true distribution more clearly.

Here's a chart that obeys the area principle. It's not as visually entertaining as the ships, but it does give an *accurate* visual impression of the distribution. The height of each bar shows the count for its category. The bars are the same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it's easy to see that the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers, something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

A **bar chart** displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. Bar charts should have small spaces between the bars to indicate that these are freestanding bars that could be rearranged into any order. The bars are lined up along a common base.

Usually they stick up like this 

but sometimes they run sideways like this 

If we really want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages and use a **relative frequency bar chart.**

A S *Activity*: Bar Charts. Watch bar charts grow from data; then use your statistics package to create some bar charts for yourself.

What a Bar Chart Is, and Isn't For some reason, some computer programs give the name "bar chart" to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don't be misled. "Bar chart" is the term for a *display of counts of a categorical variable* with bars.

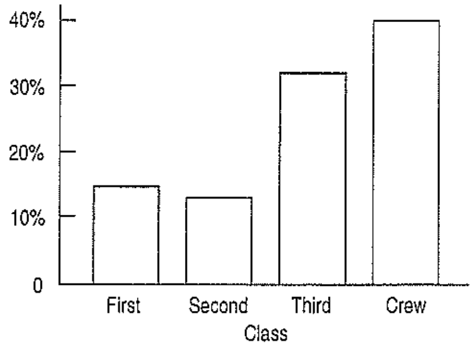


Figure 2.4

The relative frequency bar chart looks the same as the bar chart (Figure 2.3) but shows the proportion of people in each category rather than the counts.

Pie Charts

Another common display that shows how a whole group breaks into several categories is a pie chart. **Pie charts** show the whole group of cases as a circle. They slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category.

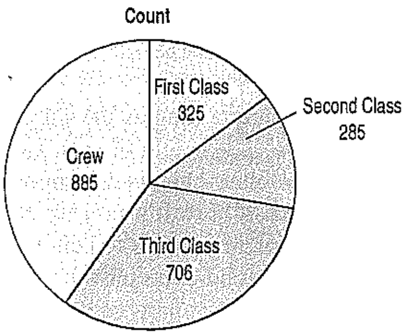
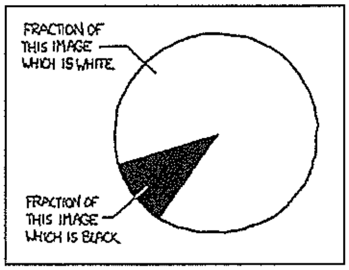


Figure 2.5
Number of *Titanic* passengers in each class


FRACTION OF THIS IMAGE WHICH IS WHITE
FRACTION OF THIS IMAGE WHICH IS BLACK

© 2013 Randall Munroe. Reprinted with permission. All rights reserved.

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near 1/2, 1/4, or 1/8. For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to 1/8 of the total. It's harder to see that there were about twice as many third-class as first-class passengers. Which category had the most passengers? Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

Think Before You Draw

Our first rule of data analysis is *Make a picture*. But what kind of picture? We don't have a lot of options—yet. There's more to Statistics than pie charts and bar charts, and knowing when to use each type of graph is a critical first step in data analysis. That decision depends in part on what type of data we have.

It's important to check that the data are appropriate for whatever method of analysis you choose. Before you make a bar chart or a pie chart, always check the **Categorical Data Condition:** The data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or a pie chart, you'll need to also make sure that the categories don't overlap so that no individual is counted twice. If the categories do overlap, you can still make a bar chart, but the percentages won't add up to 100%. For the *Titanic* data, either kind of display is appropriate because the categories don't overlap.

Throughout this course, you'll see that doing Statistics right means selecting the proper methods. That means you have to *Think* about the situation at hand. An important first step, then, is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

Contingency Tables: Children and First-Class Ticket Holders First?

A S *Activity*: Children at Risk. This activity looks at the fates of children aboard the *Titanic*; the subsequent activity shows how to make such tables on a computer.

Only 32% of those aboard the *Titanic* survived. Was that survival rate the same for men and women? For children and adults? For all ticket classes? It's often more interesting to ask if one variable relates to another. For example, was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into a lifeboat?

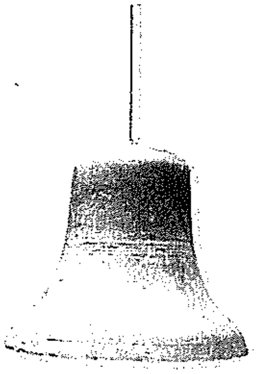
To answer that question we can arrange the counts for the two categorical variables, *Survival* and ticket *Class*, in a table. Table 2.4 shows each person aboard the *Titanic* classified according to both their ticket *Class* and their *Survival*. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a **contingency table**.

Table 2.4
Contingency table of ticket *Class* and *Survival* The bottom line of "Totals" is the same as the previous frequency table.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Each **cell** of the table gives the count for a combination of values of the two variables. The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class*. The right column of the table is the frequency distribution of the variable *Survival*. When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its **marginal distribution.** The marginal distribution can be expressed either as counts or percentages.

If you look down the column for second-class passengers to the first row, you'll find the cell containing the 118 second-class passengers who survived. Looking at the cell to its right we see that more third-class passengers (178) survived. But, does that mean that third-class passengers were more *likely* to survive? It's true that *more* third-class passengers survived, but there were many more third-class passengers on board the ship. To compare the two numbers fairly, we need to express them as percentages—but as a percentage of what?



A bell-shaped artifact from the *Titanic*

For any cell, there are three choices of percentage. We could express the 118 second-class survivors as 5.4% of all the passengers on the *Titanic* (the *overall percent*), as 16.6% of all the survivors (the *row percent*), or as 41.4% of all second-class passengers (the *column percent*). Each of these percentages is potentially interesting.

Statistics programs offer all three. Unfortunately, they often put them all together in each cell of the table. The resulting table holds lots of information, but it can be hard to understand:

Table 2.5

Another contingency table of ticket ***Class*** This time we see not only the counts for each combination of *Class* and *Survival* (in bold) but the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful.

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	**203**	**118**	**178**	**212**	**711**
		% of Row	28.6%	16.6%	25.0%	29.8%	100%
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
		% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	Count	**122**	**167**	**528**	**673**	**1490**
		% of Row	8.2%	11.2%	35.4%	45.2%	100%
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
		% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	Count	**325**	**285**	**706**	**885**	**2201**
		%of Row	14.8%	12.9%	32.1%	40.2%	100%
		% of Column	100%	100%	100%	100%	100%
		% of Table	14.8%	12.9%	32.1%	40.2%	100%

To simplify the table, let's first pull out the percent of table values:

Table 2.6

A contingency table of *Class* by *Survival* with only the table percentages

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

These percentages tell us what percent of *all* passengers belong to each combination of column and row category. For example, we see that although 8.1% of the people aboard the *Titanic* were surviving third-class ticket holders, only 5.4% were surviving second-class ticket holders. Is this fact useful? Comparing these percentages, you might think that the chances of surviving were better in third class than in second. But be careful. There were many more third-class than second-class passengers on the *Titanic,* so there were more third-class survivors. That group is a larger percentage of the passengers, but that's not really what we want to know. Overall percentages don't answer questions like this.

Percent of What? The English language can be tricky when we talk about percentages. If you're asked "What percent *of the survivors* were in second class?" it's pretty clear that we're interested only in survivors. It's as if we're restricting the *Who* in the question to the survivors, so we should look at the number of second-class passengers among all the survivors—in other words, the row percent.

But if you're asked "What percent were second-class passengers who survived?" you have a different question. Be careful; here, the *Who* is everyone on board, so 2201 should be the denominator, and the answer is the table percent.

And if you're asked "What percent of the second-class passengers survived?" you have a third question. Now the *Who* is the second-class passengers, so the denominator is the 285 second-class passengers, and the answer is the column percent.

Always be sure to ask "percent of what?" That will help you to know the *Who* and whether we want *row, column,* or *table* percentages.

For Example FINDING MARGINAL DISTRIBUTIONS

A recent Gallup poll asked 1008 Americans age 18 and over whether they planned to watch the upcoming Super Bowl. The pollster also asked those who planned to watch whether they were looking forward more to seeing the football game or the commercials. The results are summarized in the table:

		Sex		
		Male	**Female**	**Total**
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
	Total	**492**	**516**	**1008**

QUESTION: What's the marginal distribution of the responses?

ANSWER: *To determine the percentages for the three responses, divide the count for each response by the total number of people polled:*

$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

According to the poll, 47.5% of American adults were looking forward to watching the Super Bowl game, 23.5% were looking forward to watching the commercials, and 29% didn't plan to watch at all.

Conditional Distributions

Rather than look at the overall percentages, it's more interesting to ask whether the chance of surviving the *Titanic* sinking *depended* on ticket class. We can look at this question in two ways. First, we could ask how the distribution of ticket *Class* changes between survivors and nonsurvivors. To do that, we look at the *row percentages:*

Table 2.7
The conditional distribution of ticket *Class* conditioned on each value of *Survival: Alive* and *Dead*

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
		28.6%	16.6%	25.0%	29.8%	100%
	Dead	122	167	528	673	1490
		8.2%	11.2%	35.4%	45.2%	100%

By focusing on each row separately, we see the distribution of class under the *condition* of surviving or not. The sum of the percentages in each row is 100%, and we divide that up by ticket class. In effect, we temporarily restrict the *Who* first to survivors and make a pie chart for them. Then we refocus the *Who* on the nonsurvivors and make

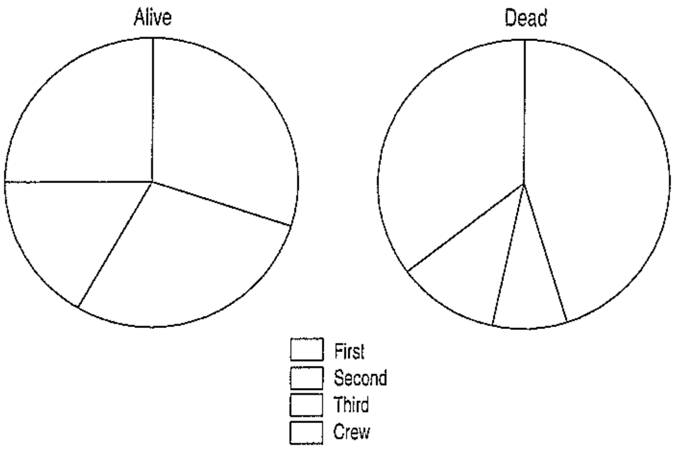


Figure 2.6

Pie charts of the conditional distributions of ticket *Class* for the survivors and nonsurvivors, separately Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a reasonable choice.

their pie chart. These pie charts show the distribution of ticket classes *for each row* of the table: survivors and nonsurvivors. The distributions we create this way are called **conditional distributions,** because they show the distribution of one variable for just those cases that satisfy a condition on another variable.

For Example FINDING CONDITIONAL DISTRIBUTIONS

RECAP: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

QUESTION: How do the conditional distributions of interest in the commercials differ for men and women?

		Sex		
		Male	Female	Total
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
	Total	492	516	1008

ANSWER: Look at the group of people who responded "Commercials" and determine what percent of them were male and female:

$$\frac{81}{237} = 34.2\% \quad \frac{156}{237} = 65.8\%$$

Women make up a sizable majority of the adult Americans who look forward to seeing Super Bowl commercials more than the game itself. Nearly 66% of people who voiced a preference for the commercials were women, and only 34% were men.

But we can also turn the question around. We can look at the distribution of *Survival* for each category of ticket *Class*. To do this, we look at the *column percentages*. Those show us whether the chance of surviving was roughly the same *for each of the four classes*. Now the percentages in each column add to 100%, because we've restricted the *Who*, in turn, to each of the four ticket classes:

Table 2.8

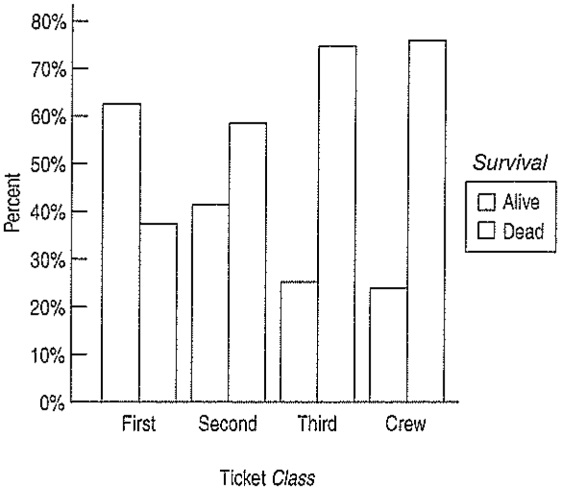
A contingency table of *Class* by *Survival* with only counts and column percentages Each column represents the conditional distribution of *Survival* for a given category of ticket *Class*.

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325 100%	285 100%	706 100%	885 100%	2201 100%

Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could display the percentages surviving and not for each *Class* in a side-by-side bar chart:

Figure 2.7

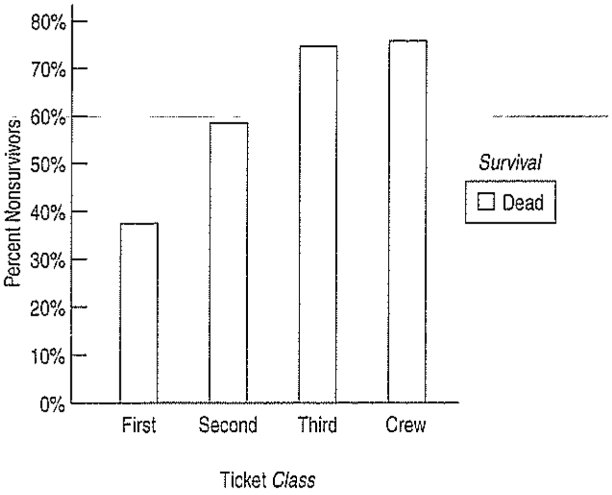
Side-by-side bar chart showing the conditional distribution of *Survival* for each category of ticket *Class* The corresponding pie charts would have only two categories in each of four pies, so bar charts seem the better alternative.



These bar charts are simple because, for the variable *Survival*, we have only two alternatives: Alive and Dead. When we have only two categories, we need to know only the percentage of one of them. We can simplify the display even more by dropping one category. Here are the percentages of dying *across the classes* displayed in one chart:

Figure 2.8

Bar chart showing just nonsurvivor percentages for each value of ticket *Class* Because we have only two values, the second bar doesn't add any information. Compare this chart to the side-by-side bar chart shown earlier.



TI-*nspire*

Conditional distributions and association. Explore the *Titanic* data to see which passengers were most likely to survive.

Now it's easy to compare the risks. Among first-class passengers, 37.5% perished, compared to 58.6% for second-class ticket holders, 74.8% for those in third class, and 76.0% for crew members.

If the risk had been about the same across the ticket classes, we would have said that survival was *independent* of class. But it's not. The differences we see among these conditional distributions suggest that survival may have depended on ticket class. You may find it useful to consider conditioning on each variable in a contingency table in order to explore the dependence between them.

It is interesting to know that *Class* and *Survival* are associated. That's an important part of the *Titanic* story. And we know how important this is because the margins show us the actual numbers of people involved.

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are *not.*[1] In a contingency table, when the distribution of *one* variable is the same for all categories of another, we say that the variables are **independent**. That tells us there's no association between these variables. We'll see a way to check for independence formally later in the book. For now, we'll just compare the distributions.

For Example LOOKING FOR ASSOCIATIONS BETWEEN VARIABLES

RECAP: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

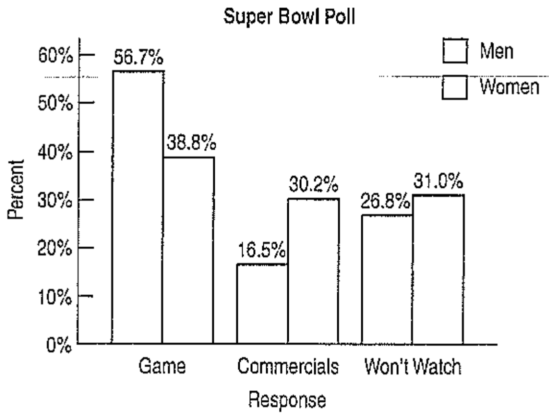
		Sex		
		Male	**Female**	**Total**
Response	**Game**	279	200	**479**
	Commercials	81	156	**237**
	Won't watch	132	160	**292**
	Total	**492**	**516**	**1008**

QUESTION: Does it seem that there's an association between interest in Super Bowl TV coverage and a person's sex?

ANSWER: First find the distribution of the three responses for the men (the column percentages):

$$\frac{279}{492} = 56.7\% \quad \frac{81}{492} = 16.5\% \quad \frac{132}{492} = 26.8\%$$

Then do the same for the women who were polled, and display the two distributions with a side-by-side bar chart:



Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl telecast: 31% of the women said they didn't plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer's sex and what the viewer is most looking forward to. While more women are interested in the game (39%) than the commercials (30%), the margin among men is much wider: 57% of men said they were looking forward to seeing the game, compared to only 16.5% who cited the commercials.

[1] This kind of "backwards" reasoning shows up surprisingly often in science—and in Statistics. We'll see it again.

Just Checking

A Statistics class reports the following data on *Sex* and *Eye Color* for students in the class:

		Eye Color			
		Blue	Brown	Green/Hazel /Other	Total
Sex	Males	6	20	6	32
	Females	4	16	12	32
	Total	10	36	18	64

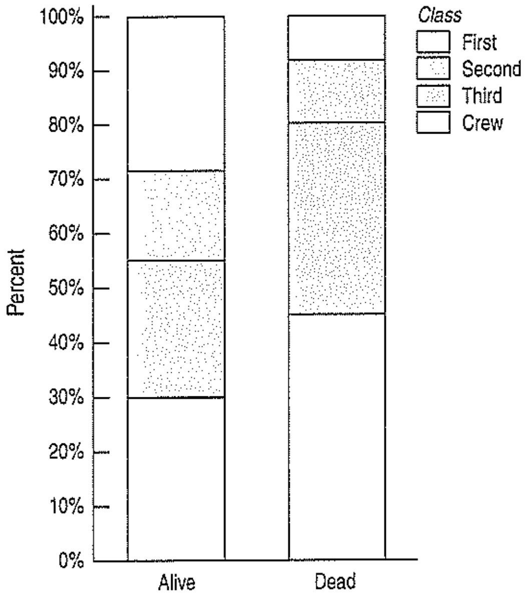
1. What percent of females are brown-eyed?
2. What percent of brown-eyed students are female?
3. What percent of students are brown-eyed females?
4. What's the distribution of *Eye Color*?
5. What's the conditional distribution of *Eye Color* for the males?
6. Compare the percent who are female among the blue-eyed students to the percent of all students who are female.
7. Does it seem that *Eye Color* and *Sex* are independent? Explain.

Segmented Bar Charts

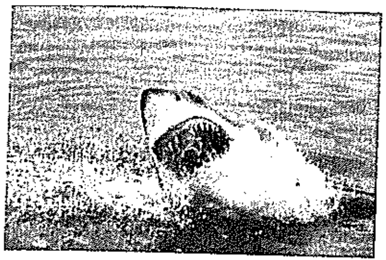
We could display the *Titanic* information by dividing up bars rather than circles. The resulting **segmented bar chart** treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket *Class* are different, indicating again that *Survival* was not independent of ticket *Class*.

Figure 2.9

A segmented bar chart for *Class* by *Survival* Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to *percentages.* Compare this display with the side-by-side pie charts of the same data in Figure 2.6.



Step-by-Step Example EXAMINING CONTINGENCY TABLES



We asked for a picture of a man eating fish. This is what we got.

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001). Their results are summarized in this table:

		Prostate Cancer	
		No	Yes
Fish Consumption	Never/seldom	110	14
	Small part of diet	2420	201
	Moderate part	2769	209
	Large part	507	42

Question: Is there an association between fish consumption and prostate cancer?

THINK ⇨ **Plan** Be sure to state what the problem is about.

I want to know if there is an association between fish consumption and prostate cancer.

Variables Identify the variables and report the W's.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer.

Be sure to check the appropriate condition.

✓ ***Categorical Data Condition:*** *I have counts for both fish consumption and cancer diagnosis. The categories of diet do not overlap, and the diagnoses do not overlap. It's okay to draw pie charts or bar charts.*

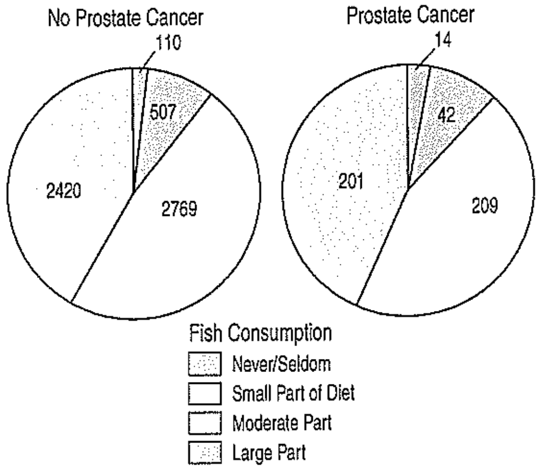
SHOW ⇨ **Mechanics** It's a good idea to check the marginal distributions first before looking at the two variables together.

		Prostate Cancer		
		No	Yes	Total
Fish Consumption	Never/seldom	110	14	124 (2.0%)
	Small part of diet	2420	201	2621 (41.8%)
	Moderate part	2769	209	2978 (47.5%)
	Large part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Two categories of the diet are quite small, with only 2.0% Never/Seldom eating fish and 8.8% in the "Large part" category. Overall, 7.4% of the men in this study had prostate cancer.

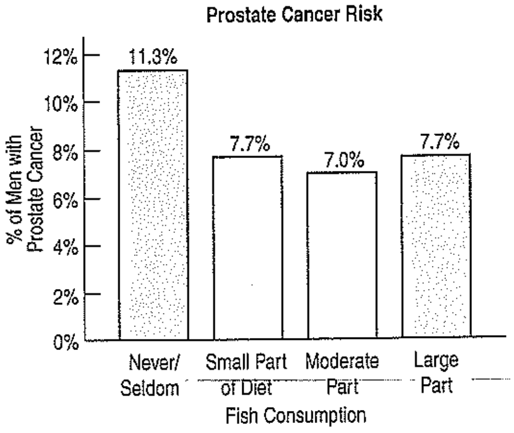
(continued)

Then, make appropriate displays to see whether there is a difference in the relative proportions. These pie charts compare fish consumption for men who have prostate cancer to fish consumption for men who don't.



It's hard to see much difference in the pie charts. So, I made a display of the row percentages. Because there are only two alternatives, I chose to display the risk of prostate cancer for each group:

Both pie charts and bar charts can be used to compare conditional distributions. Here we compare prostate cancer rates based on differences in fish consumption.



TELL ➡ **Conclusion** Interpret the patterns in the table and displays in context. If you can, discuss possible real-world consequences. Be careful not to overstate what you see. The results may not generalize to other situations.

Overall, there is a 7.4% rate of prostate cancer among men in this study. Most of the men (89.3%) ate fish either as a moderate or small part of their diet. From the pie charts, it's hard to see a difference in cancer rates among the groups. But in the bar chart, it looks like the cancer rate for those who never/seldom ate fish may be somewhat higher.

However, only 124 of the 6272 men in the study fell into this category, and only 14 of them developed prostate cancer. More study would probably be needed before we would recommend that men change their diets.[2]

[2] The original study actually used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish actually *was* substantially greater. Using pairs is a special way of gathering data. We'll discuss such study design issues and how to analyze the data in the later chapters.

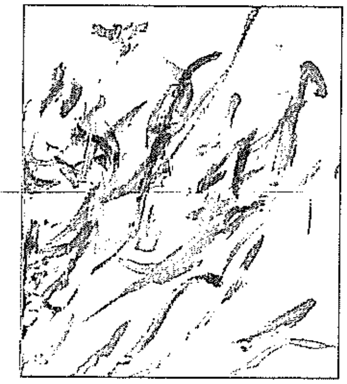
This study is an example of looking at a sample of data to learn something about a larger population, one of the main goals of this book. We care about more than these particular 6272 Swedish men. We hope that learning about their experiences will tell us something about the value of eating fish in general. That raises several questions. What population do we think this sample might represent? Do we hope to learn about all Swedish men? About all men? How do we know that other factors besides that amount of fish they ate weren't associated with prostate cancer? Perhaps men who eat fish often have other habits that distinguish them from the others and maybe those other habits are what actually kept their cancer rates lower.

Observational studies, like this one, often lead to contradictory results because we can't control all the other factors. In fact, a later paper, published in 2011, based on data from a cancer prevention trial on 3400 men from 1994 to 2003, showed that some fatty acids may actually increase the risk of prostate cancer.[3] We'll discuss the pros and cons of observational studies and experiments where we can control the factors in Chapter 11.

[3] "Serum phospholipid fatty acids and prostate cancer risk: Results from the Prostate Cancer Prevention Trial," *American Journal of Epidemiology*, 2011.

WHAT IF ○○○ the variables really ARE independent?

We told you that Statistics is about variation, remember? That presents a problem when we must decide whether we think two variables are independent. If prostate cancer is independent of fish consumption, we expect to see *exactly* the same cancer rate regardless of the amount of fish in the Swedish men's diets. But the real world never cooperates "exactly"; there's always a bit of variation in such percentages. When the percentages are glaringly different (as for passengers' chances of surviving the *Titanic* disaster for various ticket classes) we conclude that there is an association. But when they're "close enough" (as for fish consumption and prostate cancer) we conclude that the variables seem independent. This raises an important question: How close is "close enough"?



Some people think a 5-year-old (or even a chimpanzee) could create modern art like this painting. To see whether there truly are discernable artistic qualities in modern art, researchers[4] paired paintings by real artists with similar works by children or animals. Then they showed those pairs to a group of art students and another group of students without any art expertise, asking which of each pair they preferred.

Here's a table that summarizes the results with counts and row percentages.

STUDY		Preferred Painting Done by	
		Modern Artist	Other
Observer	Art Student	250 (62.5%)	150 (37.5%)
	Non-Art Student	180 (56.25%)	140 (43.75%)

While we can see that both groups did prefer works by artists, did expertise play a role? True, the art students were more likely to choose the "real" art, but could this apparent difference be the result of random chance in this particular sample of people, or does it indicate that experts really can see artistic qualities the rest of us may miss? In other words, do these results provide evidence that the ability to distinguish modern art from that of children or animals is associated with the observer's art expertise?

We investigate by asking, "What if…" What if expertise and preference actually are independent, and these 720 observations just fell into the table randomly? How often would a difference in distributions

[4] Hawley-Dolan, Angelina, and Winner, Ellen, "*Seeing the Mind Behind the Art,*" © 2011 Psychological Science. http://pss.sagepub.com/content/22/4/435

this large (or larger) happen just by chance? To find out, we ran a computer simulation[5] that mimics this study. Our simulation replicates the same number of decisions by each group while maintaining the same overall level of preference for real art, but the simulation makes each decision randomly. The simulation's first trial produced the top table. Compare the results for the observers.

TRIAL #1		Preferred Painting Done by	
		Modern Artist	Other
Observer	Art Student	241 (60.25%)	159 (39.75%)
	Non-Art Student	189 (59.06%)	131 (40.94%)

These percentages barely differ at all. If the study had come out like this, it certainly wouldn't have suggested an association.

But when we did it again, our simulation produced this table:

TRIAL #2		Preferred Painting Done by	
		Modern Artist	Other
Observer	Art Student	266 (66.5%)	134 (33.5%)
	Non-Art Student	164 (51.25%)	156 (48.75%)

This time the association looks even stronger than what appeared in the real study, yet we know this is just random chance at work. Just how likely is something like this to happen?

We simulated again. And again. And… Well, we ran 10,000 trials! Tables like #2, where the apparent association was at least as strong as what actually showed up in the study, appeared 1,382 times. This suggests that we wouldn't be surprised to see an association this strong in our sample even if art and non-art students are equally likely to pick works by a modern artist. There's almost a 14% chance that the researchers' results could just have been random variability in their sample rather than meaningful evidence of any actual association.

So what does this mean? While 14% may seem somewhat low, it's not really all that unusual—about 1 chance in 7. We'll see later on that statisticians commonly consider observed results to be "statistically significant" only if there's less than a 5% chance (a 1-in-20 shot) they could have arisen by accident. They'd conclude that this study doesn't provide convincing evidence that there's an association between expertise and the ability to identify "real" modern art.

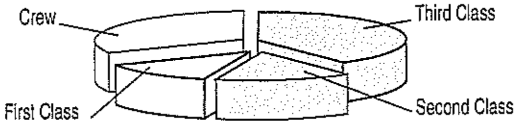
By the way, if you like the painting, you might be able to have one for peanuts. Check with the artist.[6]

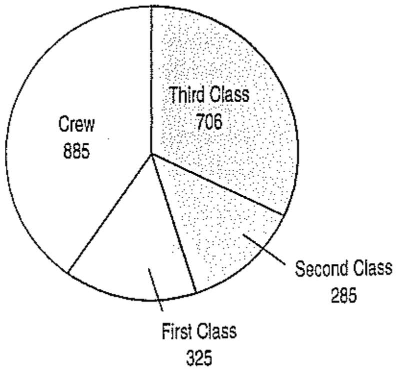
[5] Simulations are really cool. You'll learn to do them in Chapter 10, so don't drop the course yet.

[6] Jojo, the elephant. http://www.elephantartgallery.com/paintings/1129.php

WHAT CAN GO WRONG?

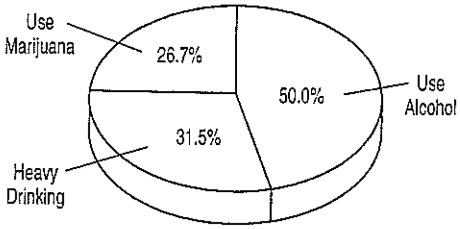
- **Don't violate the area principle.** This is probably the most common mistake in a graphical display. It is often made in the cause of artistic presentation. Here, for example, are two displays of the pie chart of the *Titanic* passengers by class:





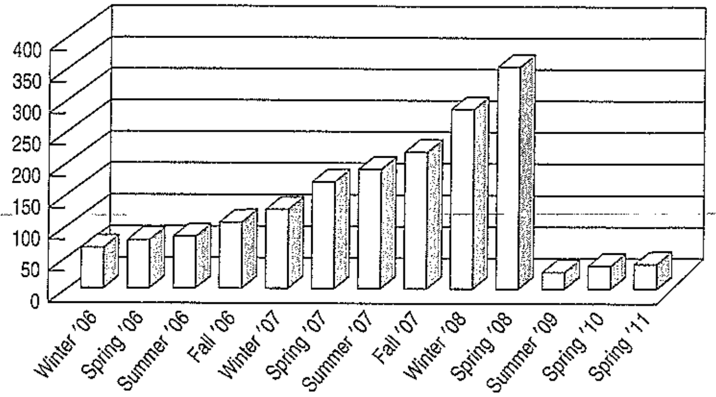
The one on the left looks pretty, doesn't it? But showing the pie on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each class—the principal feature that a pie chart ought to show.

□ **Keep it honest.** Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviors as reported by the Centers for Disease Control. What's wrong with this plot?



Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category. Of course, showing the pie on a slant makes it even harder to detect the error.

The following chart shows the average number of texts in various time periods by American cell phone customers in the period 2006 to 2011.



It may look as though text messaging decreased suddenly some time around 2010, which probably doesn't jibe with your experience. In fact, this chart has several problems. First, it's not a bar chart. Bar charts display counts of categories. This bar chart is a plot of a quantitative variable (average number of texts) against time—although to make it worse, some of the time periods are missing. Even though these flaws are already fatal, the worst mistake is one that can't be seen from the plot. In 2010, the company reporting the data switched from reporting the average number of texts per year (reported each quarter) to average number of texts per month. So, the numbers in the last three quarters should be multiplied by 12 to make them comparable to the rest.

□ **Don't confuse similar-sounding percentages.** These percentages sound similar but are different:

- The percentage of the passengers who were both in first class and survived: This would be 203/2201, or 9.2%.
- The percentage of the first-class passengers who survived: This is 203/325, or 62.5%.
- The percentage of the survivors who were in first class: This is 203/711, or 28.6%.

In each instance, pay attention to the *Who* implicitly defined by the phrase. Often there is a restriction to a smaller group (all aboard the *Titanic,* those in first class, and those who survived, respectively) before a percentage is found. Your discussion of results must make these differences clear.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	**1490**
	Total	**325**	**285**	**706**	**885**	**2201**

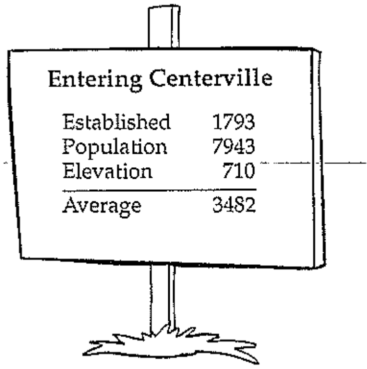
□ **Don't forget to look at the variables separately, too.** When you make a contingency table or display a conditional distribution, be sure you also examine the marginal distributions. It's important to know how many cases are in each category.

□ **Be sure to use enough individuals.** When you consider percentages, take care that they are based on a large enough number of individuals. Take care not to make a report such as this one:

We found that 66.67% of the rats improved their performance with training. The other rat died.

□ **Don't overstate your case.** Independence is an important concept, but it is rare for two variables to be *entirely* independent. We can't conclude that one variable has no effect whatsoever on another. Usually, all we know is that little effect was observed in our study. Other studies of other groups under other circumstances could find different results.

Simpson's Paradox



□ **Don't use unfair or silly averages.** Sometimes averages can be misleading. Sometimes they just don't make sense at all. Be careful when averaging different variables that the quantities you're averaging are comparable. The Centerville sign says it all.

When using averages of proportions across several different groups, it's important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work: Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

Table 2.9

On-time flights by ***Time of Day*** **and** ***Pilot*** Look at the percentages within each *Time of Day* category. Who has a better on-time record during the day? At night? Who is better overall?

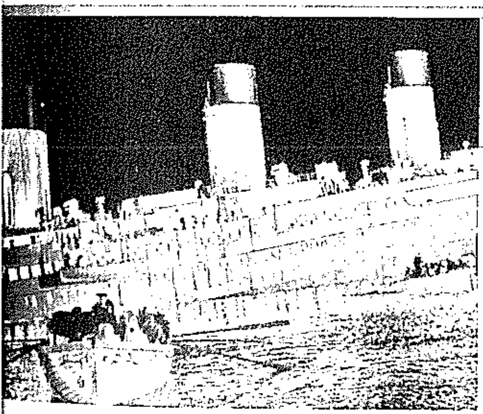
		Time of Day		
		Day	Night	Overall
Pilot	Moe	90 out of 100 90%	10 out of 20 50%	**100 out of 120** **83%**
	Jill	19 out of 20 95%	75 out of 100 75%	**94 out of 120** **78%**

Look at the daytime and nighttime flights separately. For day flights, Jill had a 95% on-time rate and Moe only a 90% rate. At night, Jill was on time 75% of the time and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

What's going on here is a problem known as **Simpson's paradox,** named for the statistician who discovered it in the 1950s. It comes up rarely in real life, but there have been several well-publicized cases. As we can see from the pilot example, the problem is *unfair averaging* over different groups. Jill has mostly night flights, which are more difficult, so her *overall average* is heavily influenced by her nighttime average. Moe, on the other hand, benefits from flying mostly during the day, with its higher on-time percentage. With their very different patterns of flying conditions, taking an overall average is misleading. It's not a fair comparison.

The moral of Simpson's paradox is to be careful when you average across different levels of a second variable. It's always better to compare percentages or other averages *within* each level of the other variable. The overall average may be misleading.

Simpson's Paradox One famous example of Simpson's paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% of female applicants got in. It looked like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc.), It turned out that, within each school, the women were admitted at nearly the same or, in some cases, much *higher* rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates (Law and Medicine, for example, admitted fewer than 10%). Men tended to apply to Engineering and Science. Those schools have admission rates above 50%. When the *average* was taken, the women had a much lower *overall* rate, but the average didn't really make sense.



What Have We Learned?

We've learned to analyze categorical variables.

- The methods in this chapter apply to categorical variables only. We always check the Categorical Variable Condition before proceeding.
- We summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percents.
- We display the distributions in a pie chart or bar chart.

When we want to see how two categorical variables are related, we put the counts (and/or percentages) in a contingency table.

- We look at the marginal distribution of each variable.
- We also look at the conditional distribution of a variable within each category of the other variable.
- We compare these marginal and conditional distributions by using pie charts, bar charts, or segmented bar charts.
- We examine the association between categorical variables by comparing conditional and marginal distributions. If the conditional distributions of one variable are roughly the same for each category of the other, we say the variables are independent.

Terms

Area principle In a statistical display, each data value should be represented by the same amount of area. (p. 15)

Frequency table (Relative frequency table) A frequency table lists the categories in a categorical variable and gives the count (or percentage of observations for each category. (p. 16)

Distribution The distribution of a variable gives

- the possible values of the variable and
- the relative frequency of each value. (p. 16)

Bar chart (Relative frequency bar chart) Bar charts show a bar whose area represents the count (or percentage) of observations for each category of a categorical variable. (p. 17)

Pie chart Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category. (p. 17)

Categorical Data Condition The methods in this chapter are appropriate for displaying and describing categorical data. Be careful not to use them with quantitative data. (p. 18)

Contingency table A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once to reveal possible patterns in one variable that may be contingent on the category of the other. (p. 19)

Marginal distribution In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table. (p. 19)

Conditional distribution The distribution of a variable restricting the *Who* to consider only a smaller group of individuals is called a conditional distribution. (p. 21)

Independence Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other. We'll show how to check for independence in a later chapter. (p. 23)

Segmented bar chart A segmented bar chart displays the conditional distribution of a categorical variable within each category of another variable. (p. 24)

Simpson's paradox When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox." (p. 31)

Name: ______________________________ A.P. Stats Summer Work

Pg 7 – Just Checking

1> Who? ______________________________

What? ______________________________

Where? ______________________________

When? ______________________________

How? ______________________________

2> Categorical Variables (qualitative)? ______________________________

Quantitative Variables (with units)? ______________________________

Pg 10 - #1, 3, 9, 13, 17, 21, 25

1> ______________________ 3> ______________________

9> Who? ______________________________

What? ______________________________

Population of interest? ______________________________

13> Who? ______________________________

Where? ______________________________

Why? ______________________________

How? ______________________________

What? Categorical Variables (qualitative)? Quantitative Variables (with units)?

______________________ ______________________

______________________ ______________________

______________________ ______________________

14> Who? ______________________________

Where? ______________________________

Why? ______________________________

How? ______________________________

What? Categorical Variables (qualitative)? Quantitative Variables (with units)?

______________________ ______________________

______________________ ______________________

______________________ ______________________

______________________ ______________________

17> Who? ______

Where? ______

Why? ______

How? ______

What? Categorical variables (qualitative)? ______ Quantitative variables (with units)? ______

21> Who? ______

Where? ______

Why? ______

How? ______

What? Categorical variables (qualitative)? ______ Quantitative variables (with units)? ______

25> Who? ______

Where? ______

Why? ______

How? ______

What? Categorical variables (qualitative)? ______ Quantitative variables (with units)? ______

Pg 24 – Just Checking

1> ______ 2> ______ 3> ______

4> Blue = ______ Brown = ______ Other = ______

5> Blue = ______ Brown = ______ Other = ______

6> ______

7> ______

Name: ____________________

1> y = ____________________

2> slope = ______________ y-intercept = ______________

3> ____________________

4> ____________________

5> ____________________

6> Dotplot:

7> Sample size = ____________________

Average = ____________________

Sample standard deviation = ____________________

Range = ____________________

Interquartile range = ____________________

5 number summary minimum = ______________

Q1 = ______________

Median = ______________

Q3 = ______________

Maximum = ______________

Outliers?