

Name: _____

AP Statistics Summer Packet

Welcome to AP Statistics. This course will be unlike any other math class you have ever taken before! Before taking this course you will need to be competent in basic algebra, be familiar with basic statistical plots (box-and-whiskers plots, scatter plots, bar graphs, histograms, circle (pie) graphs, and stem-and-leaf plots), and you must be willing to explain your answers, not just simply get the correct answer. In this class you will learn to describe and analyze sets of data and use that analysis to draw conclusions about the situations that gave that data.

For this class, it is not required, but recommended, that students purchase a graphing calculator that you can have at home and in this class on a daily basis. A Texas Instruments TI-84 is the calculator that will be used in this course. You may already own or choose to buy a different brand (Casio, HP, etc), but I will only be teaching how to use a TI calculator. Whatever calculator you buy, if you choose to buy one, make sure it is capable of performing the following: Statistical plots - box and whisker, modified box and whisker, histogram, scatter plot Regression equations and correlation statistics Distribution & probability density functions - normal, binomial, and geometric Statistical tests - t, z, χ^2 , and confidence intervals. These are functions that can be easily googled.

Every student in this class will be required to complete a summer assignment, which will be due the first day of school. The purpose of this assignment is to ensure that you enter this course with the required background knowledge in order to be successful. All answers must be on your own paper and legible. Be careful with spelling, sentence structure, and grammar. A large part of this class is clear communication of your answers. This is five - part assignment. Please make sure to bind your work in a pocket folder or report cover, with your name clearly marked on every page. This is worth a test grade and any missing part will result in a major deduction of your grade. If you have any questions, please feel free to email at mmitchell@gnbvt.edu.

Part I: Introduction to Statistics

Sta-tis-tics

Etymology: German *Statistik*: study of political facts and figures, from New Latin *statisticus*: of politics, from Latin *status*: state. Date: 1770

1 : a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data [note: this is for Statistics with a uppercaseS]

2 : a collection of quantitative data [note: this is for statistics with a lowercaseS]

Source: <http://www.merriam-webster.com/dictionary/statistics>

Answer the following in complete, well written sentences, to the best of your ability:

- 1) Before you saw this definition, how would you have defined Statistics? Has your definition changed after reading this?

- 2) How one collects the data is extremely important. Explain how you would conduct a survey to determine the percentage of GNB Voc-Tech students who are satisfied with the quality of education that they are receiving. Due to resource constraints, however, you will only be able to ask 100 students.

- 3) You have worked with data before in your science classes, if nowhere else. Provide one example from your life in which you have worked with data. How did you collect it? How did you analyze it? How did you present your findings? What conclusions did you come to?

- 4) Tell me what you have heard from other people about this class.

- 5) This class is an elective. So, why did you sign up for it?

- 6) Find a newspaper or magazine article involving statistics. Bring your article to class on the first day. Write a summary of your article on a separate sheet of paper.

Read the following to learn more about early uses of statistics...

Censuses throughout history

Early population counts generally were not concerned with determining the total size of the population or including detailed information about people. Their main goal was to discover who was available for military duty and who held taxable property. These counts usually did not give an accurate number or picture of the population. They often left out large segments of society, such as women and children, men attempting to avoid military service or taxation, and native inhabitants of an area.

The earliest known population counts were made thousands of years ago by the ancient Babylonians, Chinese, and Egyptians. Around 2500 B.C., the Babylonians recorded on clay tablets information about the taxpaying part of the population. These tablets included such data as the number of farm animals, farm products, and households for districts within the kingdom. Tax returns from around 2300 B.C. for parts of ancient China indicate some kind of population count. About 1300 B.C., Egypt was divided into administrative districts. The government registered and counted heads of households and members of the households within these districts.

The fourth book of Bible, the Book of Numbers, describes the census, or numbering, of the tribes in ancient Israel to determine the number of men of fighting age (Numbers 1: 1-46; Numbers 26: 1-51). In 594 B.C., the Greek lawmaker Solon introduced a form of enumeration and registration to reform tax laws in Greece.

The Romans employed census takers known as censors to determine the number of people who were eligible for taxation and military duty. The Roman censor was responsible for officially registering all citizens in a particular area, evaluating their property, collecting revenue, and guarding public morals. Perhaps the best-known Roman census is described in the New Testament story of the birth of Jesus Christ (Luke 2:1-7). This census took place about 5 B.C., when Joseph and Mary traveled to Bethlehem to record their names in a census ordered by the Roman emperor Augustus.

The practice of taking censuses declined in Europe after the fall of the West Roman Empire in A.D. 476. One of the few attempts to count people during the Middle Ages occurred in England in 1086. That year, commissioners sent by William the Conqueror traveled the kingdom and recorded, for tax purposes, the names of all English landowners and the value of their lands and houses, tenants, and servants. The resulting document, known as the Domesday Book, provides historians with a censuslike description of England at that time.

Through the years, with the rise in trade, the growth of towns, and the development of nations, rulers and government officials increasingly recognized the importance of counting people and goods. In 1665, King Louis XIV of France ordered a census in New France, in what is now Quebec, Canada. This census recorded the name of each person, along with such information as age, marital status, occupation, and relationship to the head of the household. The main purpose of this census was to collect information about the colony's progress, rather than to assess how much military service or tax revenue the colonists might provide. Because of this purpose, census historians generally consider the New France enumeration to be the model for modern censuses.

Likewise, in 1703, there was a house-to-house census in Iceland for reasons other than taxation and military service. This census inquired into the effects of economic conditions and natural disasters. The government then used the information to develop programs for economic and social improvement.

A number of European countries undertook censuses of individual cities and provinces in the early 1700's. However, none of these enumerations counted the total population of a nation until 1749. That year, the Swedish government conducted the first national census.

The first modern census—one that was complete, direct, and scheduled to be repeated at regular intervals—was the United States census of 1790. In the 1800's, a number of other countries began taking regular censuses. In 1853, an International Statistical Congress was held in Brussels, Belgium. This conference represented the first attempt to adopt international recommendations and requirements to help in comparing population census data among various countries.

After World War II ended in 1945, censuses became especially important as an aid in planning for the economic reconstruction of countries that had been heavily damaged in the war. In 1946, the United Nations established a separate Population and Statistical Commission, which recognized the need for census statistics. Since then, the United Nations has published a number of principles and recommendations for population and housing censuses to assist countries in the planning of censuses. Following these recommended standards allows for international comparison of collected data. In addition, the United Nations Fund for Population Activities provides many countries with financial and expert assistance for the planning of censuses.

Today, most censuses are proclaimed by a government decree or law and planned and executed by a statistical agency, a permanent or semipermanent census bureau, or both. These census acts or laws require every person to answer the questions to the best of his or her knowledge. Refusal to cooperate can result in a fine or even imprisonment.

Draaijer, Gera. 'Census.' *World Book Advanced*. World Book, 2011. Web. 5 June 2011.

7) After reading this, provide a conjecture for why the word "Statistics" is rooted in the Latin for "state".

Part 2: Completing a Survey

Complete a survey on the Census at School Website (follow the steps below):

- 1.) Log on to <http://ww2.amstat.org/censusatschool/>
- 2.) Click on Student Section
- 3.) Follow the steps on this page
- 4.) When you click on **Online Survey**, you will be asked for a class ID and a password.

Class ID: 305703

Password: bears

- 5.) Complete the survey

Part 3: Conducting a Survey

Create a survey of 5 questions that you can ask 30 people throughout the summer. The questions should contain some quantitative data and some categorical data (see Part 5). The questions can deal with anything that you think would be interesting to study. Write the questions below and create some appropriate method for displaying the answers to the questions asked.

Question 1:

Question 2:

Question 3:

Question 4:

Question 5:

Part 4: Observing Data

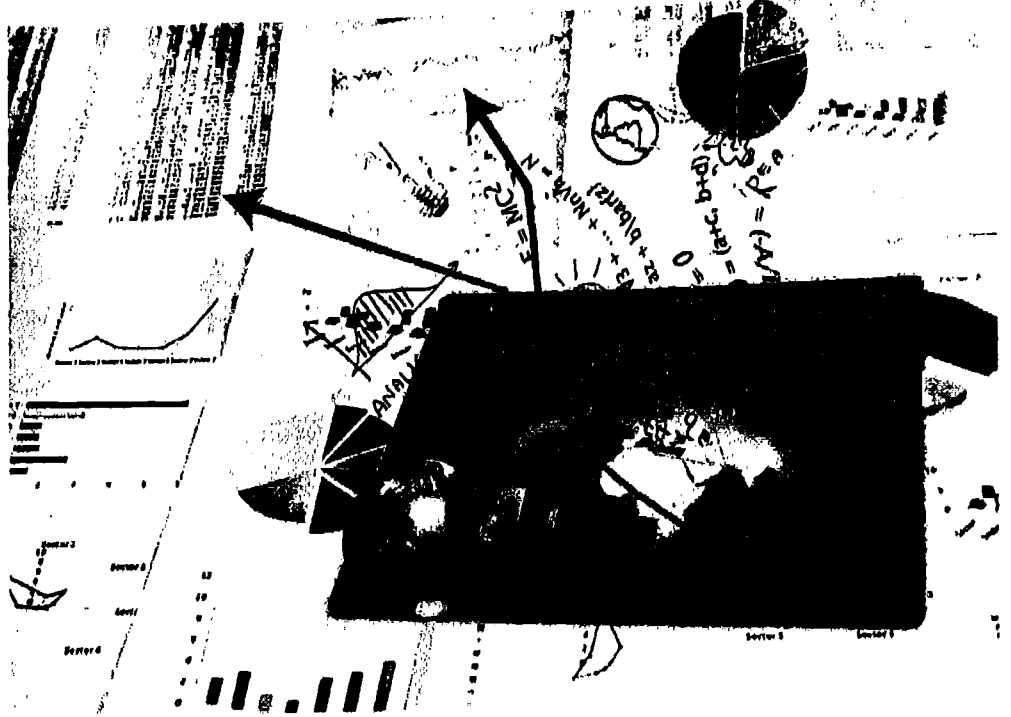
Pick something quantitative that you was to study for a month (high temperature, wind speed, number of home runs hit in Major League Baseball, a certain stock's closing price, etc.) and track that variable for 30 days throughout the summer. Right down the values in the table below.

Variable Measured : _____

Day	Date	Measurement
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		

Part 5: Chapter 1 & 2

Read the attached chapter 1 of your textbook. Answer all even questions at the end of the section. Attach answers to this packet.



"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

¹We 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.

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



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.

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.

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."

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:

105-2688834-3759466	Ohio	Nashville	Kansas	10.99	440	N	B0000015Y6	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	B0000010AA	Monique D.
002-1663369-6638649	Ohio	Best of Kansas	Kansas	10.99	440	N	B002MXA7Q0	Katherine H.

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
105-2688834-3759466	Katherine H.	Ohio	10.99	440	Nashville	N	B0000015Y6	Kansas
105-9318443-4200264	Samuel R.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
105-1872500-0198646	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	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

Who and What

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.

In general, the rows of a data table correspond to individual cases about *Whom* (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.

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.

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.

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.

A *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.

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.

More About Variables (What?)

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.

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.

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.

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

⁶You 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.

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

 **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	3463	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



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.

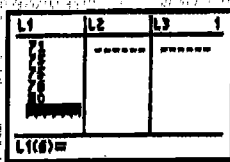
Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics 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.

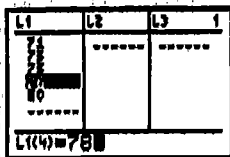
TI Tips WORKING WITH DATA

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

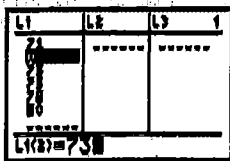


TO ENTER DATA: Hit the **STAT** button, and choose **EDIT** from the menu. You'll see a set of columns labeled **L1**, **L2**, and so on. Here is where you can enter, change, or delete a set of data.

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.



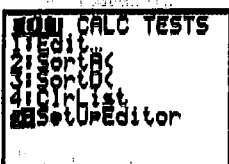
TO CHANGE A DATUM: Suppose the 76" player grew since last season; his height should be listed as 78". Use the arrow keys to move the cursor onto the 76, then change the value and **ENTER** the correction.



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.



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.

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?
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'

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

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. **Gaydar** A study conducted by a team of American and Canadian researchers found that during ovulation, a woman can tell whether a man is gay or straight by looking at his face. To explore the subject, the authors conducted three 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 face. Half of the men were gay, and the other half were straight. All held similar expressions in the photos or were deemed to be equally attractive. None of the women were using any contraceptive drugs at the time of the test. The result: the closer a woman was to her peak ovulation the more accurate her guess.
(Source: news.yahoo.com/does-ovulation-boost-womans-gaydar-210405621.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	Phlips/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

Data



Many years ago, most stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought 6 weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

The answer is data. Collecting data on their customers, transactions, and sales lets companies track their inventory and helps them predict what their customers prefer. These data can help them predict what their customers may buy in the future so they know how much of each item to stock. The store can use the data and what it learns from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

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 2006, the company's revenue reached \$10.7 billion. Amazon has expanded into selling a wide selection of merchandise, from \$400,000 necklaces¹ to yak cheese from Tibet to the largest book in the world.

Amazon is constantly monitoring and evolving its Web site to serve its customers better and maximize sales performance. To decide which changes to make to the site, the company experiments, collecting data and analyzing what works best. When you visit the Amazon Web site, you may encounter a different look or different suggestions and offers. Amazon statisticians want to know whether you'll follow the links offered, purchase the items suggested, or even spend a

"Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."

—Ronny Kohavi,
Director of Data Mining
and Personalization,
Amazon.com



¹ Please get credit card approval before purchasing online.

longer time browsing the site. As Ronny Kohavi, director of Data Mining and Personalization, said, "Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them."

But What Are Data?

We bet you thought you knew this instinctively. 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 some data record names or other labels. The names in Amazon.com's database are data, but not numerical.

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book, like 0321570448, may have a numerical value, but it's really just another name for *Stats: Modeling the World*.

Data values, no matter what kind, are useless without their 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. Answering these questions can provide the **context** for data values. The answers to the first two questions are essential. If you can't answer *Who* and *What*, you don't have **data**, and you don't have any useful information.

THE W'S:

WHO

WHAT

and in what units

WHEN

WHERE

WHY

HOW

Data Tables

Here are some data Amazon might collect:

B000001OAA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B00000I5Y6	440	B000002BK9	Let Go	Y

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

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

Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals

Now we can see that these are four purchase records, relating to CD orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who*. But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases).

A common place to find the *Who* of the table is the leftmost column. The other *W*'s might have to come from the company's database administrator.²

Who

In general, the rows of a data table correspond to individual cases about *Whom* (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. Individuals who answer a survey are referred to as *respondents*. People on whom we experiment are *subjects* or (in an attempt to acknowledge the importance of their role in the experiment) *participants*, but animals, plants, Web sites, and other inanimate subjects are often just called *experimental units*. In a database, rows are called *records*—in this example, purchase records. Perhaps the most generic term is *cases*. In the Amazon table, the cases are the individual CD orders.

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.

Sometimes people just refer to data values as *observations*, without being clear about the *Who*. Be sure you know the *Who* of the data, or you may not know what the data say.

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 March 2007, *Consumer Reports* published an evaluation of large-screen, high-definition television sets (HDTVs). The magazine purchased and tested 98 different models from a variety of manufacturers.

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

The magazine is interested in the performance of all HDTVs currently being offered for sale. It tested a sample of 98 sets, the "Who" for these data. Each HDTV set represents all similar sets offered by that manufacturer.

What and Why

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. Variables may seem simple, but to really understand your variables, you must *Think* about what you want to know.

Although area codes are numbers, do we use them that way? Is 610 twice 305? Of course it is, but is that the question? Why would we want to know whether Allentown, PA (area code 610), is twice Key West, FL (305)? Variables play different roles, and you can't tell a variable's role just by looking at it.

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? . . . What kinds of things can we learn about variables like these? A natural start is to *count* how many cases belong in each category. (Are you listening to music while reading this? We could count

²In database management, this kind of information is called "metadata."

It is wise to be careful. The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.



To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Now that phones have push-buttons, area codes have finally become just categories.

By international agreement, the International System of Units links together all systems of weights and measures. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	°Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

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.

the number of students in the class who were and the number who weren't.) We'll look for ways to compare and contrast the sizes of such categories.

Some variables have measurement **units**. Units tell how each value has been measured. But, more importantly, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement. The units tell us how much of something we have or 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, yen, or Estonian krooni.

What kinds of things can we learn about measured variables? We can do a lot more than just counting categories. We can look for patterns and trends. (How much did you pay for your last movie ticket? What is the range of ticket prices available in your town? How has the price of a ticket changed over the past 20 years?)

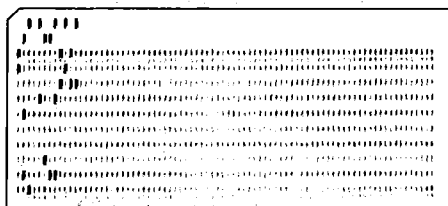
When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical variable**.³ When a measured variable with units answers questions about the quantity of what is measured, we call it a **quantitative variable**. These types can help us decide what to do with a variable, but they are really more about what we hope to learn from a variable than about the variable itself. It's the questions we ask a variable (the *Why* of our analysis) that shape how we think about it and how we treat it.

Some variables can answer questions only about categories. If the values of a variable are words rather than numbers, it's a good bet that it is categorical. But some variables can answer both kinds of questions. Amazon could ask for your *Age* in years. That seems quantitative, and 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 CD to offer you in a special deal—one by Raffi, Blink-182, Carly Simon, or Mantovani—and needs to be sure to have adequate supplies on hand to meet the demand. Then thinking of your age in one of the categories—child, teen, adult, or senior—might be more useful. If it isn't clear whether a variable is categorical or quantitative, think about *Why* you are looking at it and what you want it to tell you.

A typical 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? Once again, we'll look to the *Why*. A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When 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 we should be careful about treating *Educational Value* as

³You may also see it called a *qualitative variable*.

One tradition that hangs on in some quarters is to name variables with cryptic abbreviations written in uppercase letters. This can be traced back to the 1960s, when the very first statistics computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and the earliest statistics programs limited variable names to six or eight characters, so variables were called things like PRSRF3. Modern programs do not have such restrictive limits, so there is no reason for variable names that you wouldn't use in an ordinary sentence.



purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, she should be cautious. Variables like this 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 to decide whether to treat it as categorical or quantitative.

FOR EXAMPLE

Identifying "What" and "Why" of HDTVs.

Recap: A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0–100).

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

The "what" of this article includes the following variables:

- manufacturer (categorical);
- cost (in dollars, quantitative);
- screen size (in inches, quantitative);
- type (categorical);
- performance score (quantitative).

The magazine hopes to help consumers pick a good HDTV set.

Counts Count

In Statistics, we often count things. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases are shipped. They'd probably start by counting the number of purchases shipped by ground transportation, by second-day air, and by overnight air. Counting is a natural way to summarize the categorical variable *Shipping Method*. So every time we see counts, does that mean the variable is categorical? Actually, no.

We also use counts to measure the amounts of things. How many songs are on your digital music player? How many classes are you taking this semester? To measure these quantities, we'd naturally count. The variables (*Songs*, *Classes*) would be quantitative, and we'd consider the units to be "number of . . ." or, generically, just "counts" for short.

So we use counts in two different ways. When we count the cases in each category of a categorical variable, the category labels are the *What* and the individuals counted are the *Who* of our data. The counts themselves are not the

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.

data, but are something we summarize about the data. Amazon counts the number of purchases in each category of the categorical variable *Shipping Method*. For this purpose (the *Why*), the *What* is shipping method and the *Who* is purchases.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Other times our focus is on the amount of something, which we measure by counting. Amazon might record the number of teenage customers visiting their site each month to track customer growth and forecast CD sales (the *Why*). Now the *What* is *Teens*, the *Who* is *Months*, and the units are *Number of Teenage Customers*. *Teen* was a category when we looked at the categorical variable *Age*. But now it is a quantitative variable in its own right whose amount is measured by counting the number of customers.

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	...
...	...

Identifying Identifiers

What's your student ID number? It is numerical, but is it a quantitative variable? No, it doesn't have units. Is it categorical? Yes, but it is a special kind. Look at how many categories there are and at how many individuals are in each. There are as many categories as individuals and only one individual in each category. While it's easy to count the totals for each category, it's not very interesting. 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.

Identifier variables themselves don't tell us anything useful about the categories because we know there is exactly one individual in each. However, they are crucial in this age of large data sets. They make it possible to combine data from different sources, to protect confidentiality, and to provide unique labels. The variables *UPS Tracking Number*, *Social Security Number*, and Amazon's *ASIN* are all examples of identifier variables.

You'll want to recognize when a variable is playing the role of an identifier so you won't be tempted to analyze it. There's probably a list of unique ID numbers for students in a class (so they'll each get their own grade confidentially), but you might worry about the professor who keeps track of the average of these numbers from class to class. Even though this year's average ID number happens to be higher than last's, it doesn't mean that the students are better.

Where, When, and How

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.)

We must know *Who*, *What*, and *Why* to analyze data. Without knowing these three, we don't have enough 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.

How the data are collected can make the difference between insight and non-sense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics, to be discussed in Part III, is the design of sound methods for collecting data.

Throughout this book, whenever we introduce data, we'll provide a margin note listing the *W's* (and *H*) of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why*, the *Who*, and the *What*. Identifying them is a key part of the *Think* step of any analysis. Make sure you know all three before you proceed to *Show* or *Tell* anything about the data.



JUST CHECKING

In the 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fastest Tour de France in its 100-year history. In 2004, he made history again by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and once again set a new record for the fastest average speed. You can find data on all the Tour de France races on the DVD. Here are the first three and last ten lines of the data set. Keep in mind that the entire data set has nearly 100 entries.

- List as many of the *W's* as you can for this data set.
- 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 Trousselier	France	112.18.09	27.3	11	2975	60	24
1999	Lance Armstrong	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong	USA	86.15.02	41.65	21	3608	189	155
2006	Óscar Periero	Spain	89.40.27	40.78	20	3657	176	139
2007	Alberto Contador	Spain	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	87.52.52	40.50	21	3559	199	145

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 Web site. 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 such extra symbols 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. The context colors our interpretation of the data, so those 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 Web site. The question that respondents answered may have been posed in a way that influenced their responses.

TI Tips

Working with data

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

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.

L1	L2	L3	1
----	----	----	

L1(4)=

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	----	----	

L1(4)=71

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.

Skills

THINK

- ▶ Be able to identify the *Who, What, When, Where, Why, and How* of data, or recognize when some of this information has not been provided.
- ▶ Be able to identify the cases and variables in any data set.
- ▶ Be able to identify the population from which a sample was chosen.
- ▶ Be able to classify a variable as categorical or quantitative, depending on its use.
- ▶ For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).

TELL

- ▶ Be able to describe a variable in terms of its *Who, What, When, Where, Why, and How* (and be prepared to remark when that information is not provided).

DATA ON THE COMPUTER

A/S

Activity: Examine the Data. Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

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 2007 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. **Mood.** A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". 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. **Movies.** Some motion pictures are profitable and others are not. Understandably, the movie industry would like to know what makes a movie successful. Data from 120 first-run movies released in 2005 suggest that longer movies actually make *less* profit.
11. **Fitness.** Are physically fit people less likely to die of cancer? An article in the May 2002 issue of *Medicine and Science in Sports and Exercise* reported results of a study that followed 25,892 men aged 30 to 87 for 10 years. The most physically fit men had a 55% lower risk of death from cancer than the least fit group.
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.** The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 2007. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person had voted in the 2006 midterm congressional election.
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 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 2006, *Consumer Reports* published an article evaluating refrigerators. It listed 41 models, giving the brand, cost, size (cu ft), type (such as top freezer), estimated annual energy cost, an overall rating (good, excellent, etc.), and the repair history for that brand (percentage requiring repairs over the past 5 years).

- 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. **Horse race 2008.** 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.

Date	Winner	Margin (lengths)	Jockey	Winner's Payoff (\$)	Duration (min:sec)	Track Condition
May 17, 1875	Aristides	2	O. Lewis	2850	2:37.75	Fast
May 15, 1876	Vagrant	2	B. Swim	2950	2:38.25	Fast
May 22, 1877	Baden-Baden	2	W. Walker	3300	2:38.00	Fast
May 21, 1878	Day Star	1	J. Carter	4050	2:37.25	Dusty
May 1, 2004	Smarty Jones	2 3/4	S. Elliott	854800	2:04.06	Sloppy
May 7, 2005	Giacomo	1/2	M. Smith	5854800	2:02.75	Fast
May 6, 2006	Barbaro	6 1/2	E. Prado	1453200	2:01.36	Fast
May 5, 2007	Street Sense	2 1/4	C. Borel	1450000	2:02.17	Fast
May 3, 2008	Big Brown	4 3/4	K. Desormeaux	1451800	2:01.82	Fast

- 26. **Indy 2008.** The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day 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 2008, the winner, Scott Dixon, averaged 143.567 mph.

Here are the data for the first five races and five recent Indianapolis 500 races. Included also are the pole winners (the winners of the trial races, when each driver drives alone to determine the position on race day).

Year	Winner	Pole Position	Average Speed (mph)	Pole Winner	Average Pole Speed (mph)
1911	Ray Harroun	28	74.602	Lewis Strang	
1912	Joe Dawson	7	78.719	Gil Anderson	
1913	Jules Goux	7	75.933	Caleb Bragg	
1914	Rene Thomas	15	82.474	Jean Chassagne	
1915	Ralph DePalma	2	89.840	Howard Wilcox	98.580
2004	Buddy Rice	1	138.518	Buddy Rice	220.024
2005	Dan Wheldon	16	157.603	Tony Kanaan	224.308
2006	Sam Hornish Jr.	1	157.085	Sam Hornish Jr.	228.985
2007	Dario Franchitti	3	151.744	Hélio Castroneves	225.817
2008	Scott Dixon	1	143.567	Scott Dixon	221.514



JUST CHECKING

Answers

1. Who—Tour de France races; What—year, winner, country of origin, total time, average speed, stages, total distance ridden, starting riders, finishing riders; How—official statistics at race; Where—France (for the most part); When—1903 to 2008; Why—not specified (To see progress in speeds of cycling racing?)

2.

Variable	Type	Units
Year	Quantitative or Categorical	Years
Winner	Categorical	
Country of Origin	Categorical	
Total Time	Quantitative	Hours/minutes/seconds
Average Speed	Quantitative	Kilometers per hour
Stages	Quantitative	Counts (stages)
Total Distance	Quantitative	Kilometers
Starting Riders	Quantitative	Counts (riders)
Finishing Riders	Quantitative	Counts (riders)