1

Stats Starts Here[1]

WHERE ARE WE GOING?

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 statistics courses don't have the reputation of being students' 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 the science of learning from data. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

This is a text about understanding the world by using data. So we'd better start by understanding data. There's more to that than you might have thought.

1.1 What Is Statistics?
1.2 Data
1.3 Variables
1.4 Models

"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

1.1 What Is Statistics?

People around the world have one thing in common—they all want to figure out what's going on. You'd think with the amount of information available to everyone today this would be an easy task, but actually, as the amount of information grows, so does our need to understand what it can tell us.

At the base of all this information, on the Internet and all around us, are data. We'll talk about data in more detail in the next section, but for now, think of **data** as any collection of numbers, characters, images, or other items that provide information about something. What sense can we make of all this data? You certainly can't make a coherent picture from random pieces of information. Whenever there are data and a need for understanding the world, you'll find statistics.

This text will help you develop the skills you need to understand and communicate the knowledge that can be learned from data. By thinking clearly about the question you're trying to answer and learning the statistical tools to show what the data are saying, you'll acquire the skills to tell clearly what it all means. Our job is to help you make sense of the concepts and methods of statistics and to turn it into a powerful, effective approach to understanding the world through data.

[1]We were thinking of calling this chapter "Introduction" but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this down here in the footnotes because nobody reads footnotes either.

> "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, former Director of Data Mining and Personalization, Amazon.com*

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.

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."

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 text will teach you skills to help with this task and ways of thinking about variation that are the foundation of sound reasoning about data.

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 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 is how you can mine your own data and learn valuable insights about the world.
- Americans spend an average of 4.9 hours per day on their smartphones. Trillions of text messages are sent each year.[3] Some of these messages are sent or read while the sender or the receiver is driving. How dangerous is texting while driving?

  How can we study the effect of texting while driving? One way is to measure reaction times of drivers faced with an unexpected event while driving and texting. 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 drivers who were above the legal limit for alcohol.

In this text, 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.

1.2 Data

STATISTICS IS ABOUT...

- Variation: Data vary because we don't see everything, and even what we do see, we measure imperfectly.
- Learning from data: We hope to learn about the world as best we can from the limited, imperfect data we have.
- Making intelligent decisions: The better we understand the world, the wiser our decisions will be.

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 2017, the company's sales reached almost $178 billion (more than 30% over 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"? You might think that data have to be numbers, but data can be text, pictures, web pages,

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

[3] informatemi.com/blog/?p=133

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

and even audio and video. If you can sense it, you can measure it. Data are now being collected automatically at such a rate that IBM estimates that "90% of the data in the world today has been created in the last two years alone."[5]

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

B0000010AA	0.99	Chris G.	902	105-2686834-3759466	1.99	0.99	Illinois
Los Angeles	Samuel R.	Ohio	N	B000068ZVQ	Amsterdam	New York, New York	Katherine H.
Katherine H.	002-1663369-6638649	Beverly Hills	N	N	103-2628345-9238664	0.99	Massachusetts
312	Monique D.	105-9318443-4200264	413	B00000I5Y6	440	B000002BK9	0.99
Canada	Detroit	440	105-1372500-0198646	N	B002MXA7Q0	Ohio	Y

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	Download	Gift?	ASIN	Artist
105-2686834-3759466	Katherine H.	Ohio	0.99	440	Amsterdam	N	B0000015Y6	Cold Play
105-9318443-4200264	Samuel R.	Illinois	1.99	312	Detroit	Y	B000002BK9	Red Hot Chili Peppers
105-1372500-0198646	Chris G.	Massachusetts	0.99	413	New York, New York	N	B000068ZVQ	Frank Sinatra
103-2628345-9238664	Monique D.	Canada	0.99	902	Los Angeles	N	B0000010AA	Blink 182
002-1663369-6638649	Katherine H.	Ohio	0.99	440	Beverly Hills	N	B002MXA7Q0	Weezer

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 *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**.

[5] www-01.ibm.com/software/data/bigdata/what-is-big-data.html

DATA BEATS INTUITION

Amazon monitors and updates its website to better serve customers and maximize sales. To decide which changes to make, analysts experiment with new designs, offers, recommendations, and links. Statisticians want to know how long you'll spend browsing the site and whether you'll follow the links or purchase the suggested items. As Ronny Kohavi, former director of Data Mining and Personalization for Amazon, 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."

- 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.

Look at all the columns to see exactly what each row refers to. Here the cases are different purchase records. You might have thought that each customer was a case, but notice that, for example, Katherine H. appears twice, in both the first and the last rows. A common place to find out exactly what each row refers to is the leftmost column. That value often identifies the cases, in this example, it's the order number. If you collect the data yourself, you'll know what the cases are. But, often, you'll be looking at data that someone else collected and you'll have to ask or figure that out yourself.

Often the cases are a **sample** from some larger **population** that we'd like to understand. Amazon doesn't care about just these customers; it wants to understand the buying patterns of *all* its customers, and, generalizing further, it wants to know how to attract other Internet users who may not 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.

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.

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, to be discussed in Part III, is the design of sound methods for collecting data. Throughout this text, 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 text 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.

EXAMPLE 1.1 Identifying the *Who*

In 2015, *Consumer Reports* published an evaluation of 126 computer 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 126 tablets, which are the *Who* for these data. Each tablet selected represents all similar tablets offered by that manufacturer.

1.3 Variables

The characteristics recorded about each individual are called **variables**. They are usually found as the columns of a data table with a name in the header that identifies what has been recorded. In the Amazon data table we find the variables *Order Number, Name, State/Country, Price*, and so on.

Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? We call variables like these **categorical**, or **qualitative**, **variables**. (You may also see them called **nominal variables** because they name categories.) 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. 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. (But see the story in the following box.)

> "Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cookbook, believing the recipes will work without understanding why. A more *cordon bleu* attitude . . . might lead to fewer statistical soufflés failing to rise."
>
> *—The Economist, June 3, 2004, "Sloppy stats shame science"*

AREA CODES—NUMBERS OR CATEGORIES?

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. Since the advent of push-button phones, area codes have finally become just categories.

Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" and "What is your marital status?" yield categorical values. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases have been shipped in the recent past. Amazon might start by counting the number of purchases shipped in each category: ground transportation, second-day air, and next-day air. Counting is a natural way to summarize a categorical variable such as *Shipping Method*. Chapter 2 discusses summaries and displays of categorical variables more fully.

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 quantitative variables, 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 Mauritanian Ouguiya (MRU).[6] We'll see how to display and summarize quantitative variables in Chapter 2.

Some quantitative variables don't have obvious units. The Dow Jones Industrial "Average" has units (points?) but no one talks about them. Percentages are ratios of two quantities and so the units "cancel out." But they are still percentages of something. So although it isn't imperative that a quantitative variable have explicit units, when they are not explicit, be careful to think about whether adding their values, averaging them, or otherwise treating them as numerical makes sense.

Sometimes a variable with numerical values can be treated as either categorical or quantitative depending on what we want to know from it. 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 AM. 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.

Identifiers

For a categorical variable like *Survived*, each individual is assigned one of two possible values, say *Alive* or *Dead*.[7] But for a variable with ID numbers, such as a *student ID*, each individual receives a unique value. We call a variable like this, which has exactly as many values as cases, 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. You'll want to recognize when a variable is playing the role of an identifier so you aren't tempted to analyze it.

Identifier variables themselves don't tell us anything useful about their categories because we know there is exactly one individual in each. Identifiers are part of what's called **metadata**, or data about the data. Metadata are crucial in this era of large datasets because by uniquely identifying the cases, they make it possible to combine data from different sources, protect (or violate) privacy, and provide unique labels.[8] Many large databases are *relational* databases. In a relational database, different data tables link to one another by matching identifiers. In the Amazon example, the *Customer Number*, *ASIN*, and *Transaction Number* are all identifiers. The IP (Internet Protocol) address of your computer is another identifier, needed so that the electronic messages sent to you can find you.

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/NSTICstrategy_041511.pdf) proposes ways that we may address this challenge in the near future.

Ordinal Variables

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? Often the best way to tell is to look to the *Why* of the study. 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 purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construct. Because there are no natural units, she

[6] As of 9/7/2018 $1 = 35.95 MRU.

[7] Well, maybe three values if you include Zombies.

[8] The National Security Agency (NSA) made the term "metadata" famous in 2014 by insisting that they only collected metadata on U.S. citizens' phone calls and text messages, not the calls and messages themselves. They later admitted to the bulk collection of actual data.

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.

EXAMPLE 1.2

Identifying the *What* and *Why* of Tablets

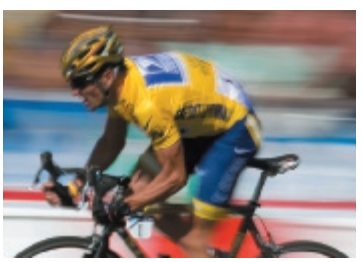
RECAP: A *Consumer Reports* article about 126 tablets lists each tablet's manufacturer, price, battery life (hrs.), the operating system (Android, iOS, or Windows), an overall quality score (0–100), and whether or not it has a memory card reader.

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

ANSWER: The variables are

- manufacturer (categorical)
- price (quantitative, $)
- battery life (quantitative, hrs.)
- operating system (categorical)
- quality score (quantitative, no units)
- memory card reader (categorical)

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. In 2012, he was banned for life for doping offenses, stripped of all of his titles and his records expunged. You can find data on all the Tour de France races in the dataset **Tour de France 2017**. Here are the first three and last seven lines of the dataset. Keep in mind that the entire dataset has over 100 entries.

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

Year	Winner	Country of Origin	Age	Team	Total Time (hours)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	32	La Française	94.55	25.7	6	2428	60	21
1904	Henri Cornet	France	20	Cycles JC	96.10	25.3	6	2428	88	23
1905	Louis Trousseller	France	24	Peugeot	110.45	27.1	11	2994	60	24
...					....					
2011	Cadel Evans	Australia	34	BMC	86.21	39.79	21	3430	198	167
2012	Bradley Wiggins	Great Britain	32	Sky	87.58	39.83	20	3488	198	153
2013	Christopher Froome	Great Britain	28	Sky	94.55	40.55	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.93	40.74	21	3663.5	198	164
2015	Christopher Froome	Great Britain	30	Sky	84.77	39.64	21	3660.3	198	160
2016	Christopher Froome	Great Britain	31	Sky	89.08	39.62	21	3529	198	174
2017	Christopher Froome	Great Britain	32	Sky	86.34	40.997	21	3540	198	167
2018	Geraint Thomas	Great Britain	32	Sky	83.28	40.210	21	3349	176	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 datasets we use in this text 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.

1.4 Models

What is a **model** for data? Models are summaries and simplifications of data that help our understanding in many ways. We'll encounter all sorts of models throughout the text. A model is a simplification of reality that gives us information that we can learn from and use, even though it doesn't represent reality exactly. A model of an airplane in a wind tunnel can give insights about the aerodynamics and flight performance of the plane even though it doesn't show every rivet.[9] In fact, it's precisely because a model is a simplification that we learn from it. Without making models for how data vary, we'd be limited to reporting only what the data we have at hand says. To have an impact on science and society we'll have to generalize those findings to the world at large.

Kepler's laws describing the motion of planets are a great example of a model for data. Using astronomical observations of Tycho Brahe, Kepler saw through the small anomalies in the measurements and came up with three simple "laws"—or models for how the planets move. Here are Brahe's observations on the declination (angle of tilt to the sun) of Mars over a twenty-year period just before 1600:

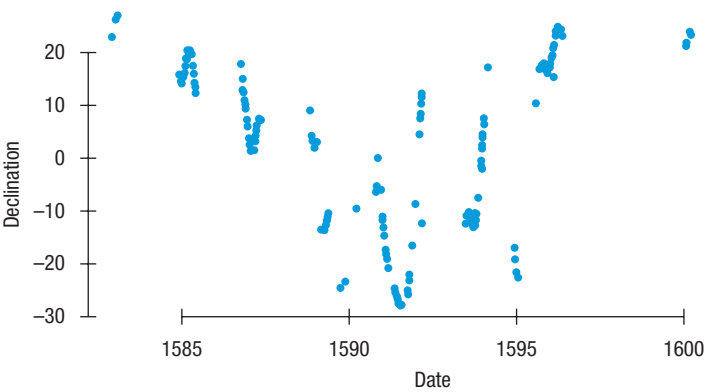


Figure 1.1
A plot of declination against time shows some patterns. There are many missing observations. Can you see the model that Kepler came up with from these data?

[9] Or tell you what movies you might see on the flight.

Here, using modern statistical methods is a plot of the model predictions from the data:

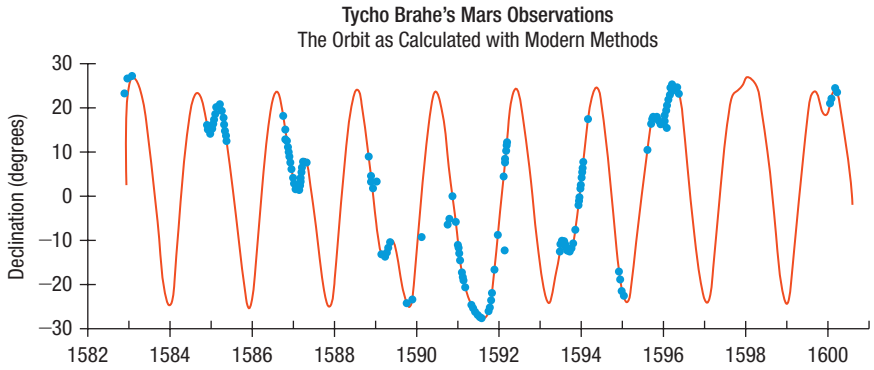


Figure 1.2
The model that Kepler proposed filled in many of the missing points and made the pattern much clearer.

Later, after Newton laid out the physics of gravity, it could be shown that the laws follow from other principles, but Kepler derived the models from data. We may not be able to come up with models as profound as Kepler's, but we'll use models throughout the text. We'll see examples of models as early as Chapter 5 and then put them to use more thoroughly later in the text when we discuss inference.

WHAT CAN GO WRONG?

- **Don't label a variable as categorical or quantitative without thinking about the data and what they represent.** The same variable can sometimes take on different roles.
- **Don't assume that a variable is quantitative just because its values are numbers.** 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 website. The question that respondents answered may be posed in a way that influences responses.

CHAPTER REVIEW

Understand that data are values, whether numerical or labels, together with their context.

- *Who, what, why, where, when* (and *how*)—the W's—help nail down the context of the data.
- We must know *who, what,* and *why* 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 *Why* helps us decide which way to treat the variables.
- Stop and identify the W's whenever you have data, and be sure you can identify the cases and the variables.

Consider the source of your data and the reasons the data were collected. That can help you understand what you might be able to learn from the data.

Identify whether a variable is being used as categorical or quantitative.

- Categorical variables identify a category for each case. Usually we think about the counts of cases that fall in each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something. They typically have units or are ratios of quantities that have units.
- Sometimes we may treat the same variable as categorical or quantitative depending on what we want to learn from it, which means some variables can't be pigeonholed as one type or the other.

REVIEW OF TERMS

The key terms are in chapter order so you can use this list to review the material in the chapter.

Data Recorded values, whether numbers or labels, together with their context (p. 27).

Data table An arrangement of data in which each row represents a case and each column represents a variable (p. 29).

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. 29).

Case An individual about whom or which we have data (p. 29).

Respondent Someone who answers, or responds to, a survey (p. 29).

Subject A human experimental unit. Also called a participant (p. 29).

Participant A human experimental unit. Also called a subject (p. 29).

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. 30).

Record Information about an individual in a database (p. 30).

Sample A subset of a population, examined in hope of learning about the population (p. 30).

Population The entire group of individuals or instances about whom we hope to learn (p. 30).

Variable A variable holds information about the same characteristic for many cases (p. 31).

Categorical (or qualitative) variable A variable that names categories with words or numerals (p. 31).

Nominal variable The term "nominal" can be applied to a variable whose values are used only to name categories (p. 31).

Quantitative variable A variable in which the numbers are values of measured quantities (p. 31).

Unit A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams (p. 31).

Identifier variable A categorical variable that records a unique value for each case, used to name or identify it (p. 32).

Metadata Data about the data. Metadata can provide information to uniquely identify cases, making it possible to combine data from different sources, protect (or violate) privacy, and label cases uniquely (p. 32).

Ordinal variable The term "ordinal" can be applied to a variable whose categorical values possess some kind of order (p. 33).

Model A description or representation, in mathematical and statistical terms, of the behavior of a phenomenon based on data (p. 34).

TECH SUPPORT

Entering Data

These days, nobody does statistics by hand. We use technology: a programmable calculator or a statistics program on a computer. Professionals all use a *statistics package* designed for the purpose. We will provide many examples of results from a statistics package throughout the text. Rather than choosing one in particular, we'll offer generic results that look like those produced by all the major statistics packages but don't exactly match any of them. Then, in the Tech Support section at the end of each chapter, we'll provide hints for getting started on several of the major packages.

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 with cases in the rows and variables in the columns. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a tab character (comma is another common delimiter) and the delimiter that marks the end of a case to be a *return* character. The data used in this text can be found on the text's website at www.pearsonglobaleditions.com.
- 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.
- Excel is often used to help organize, manipulate, and prepare data for other software packages. Many of the other packages take Excel files as inputs. Alternatively, you can copy a data table from Excel and paste it into many packages, or export Excel spreadsheets as tab delimited (.txt) or comma delimited files (.csv), which can be easily shared and imported into other programs. All data files provided with this text are in tab-delimited text (.txt) format.

EXCEL

To open a file containing data in Excel:

- Choose **File** > **Open**.
- Browse to find the file to open. Excel supports many file formats.
- Other programs can import data from a variety of file formats, but all can read both tab delimited (.txt) and comma delimited (.csv) text files.
- You can also copy tables of data from other sources, such as Internet sites, and paste them into an Excel spreadsheet. Excel can recognize the format of many tables copied this way, but this method may not work for some tables.
- Excel may not recognize the format of the data. If data include dates or other special formats ($, €, ¥, etc.), identify the desired format. Select the cells or columns to reformat and choose **Format** > **Cell**. Often, the General format is the best option for data you plan to move to a statistics package.

DATA DESK

To read data into Data Desk:

- Click the **Open File** icon or choose **File** > **Open**. The dialog lets you specify variable names (or take them from the first row of the data), the delimiter, or how to read formatted data.
- **File** > **Import** works the same way, but instead of starting a new data file, it adds the data in the file to the current data file. Data Desk can work with multiple data tables in the same file.
- If the data are already in another program, such as, for example, a spreadsheet, **Copy** the data table (including the column headings). In Data Desk choose **Edit** > **Paste** variables. There is no need to create variables first; Data Desk does that automatically. You'll see the same dialog as for Open and Import.

JMP

To import a text file:

- Choose **File > Open** and select the file from the dialog. At the bottom of the dialog screen you'll see **Open As:**—be sure to change to **Data (Using Preview)**. This will allow you to specify the delimiter and make sure the variable names are correct. (**JMP** also allows various formats to be imported directly, including .xls files.)

You can also paste a dataset in directly (with or without variable names) by selecting:

- **File > New > New Data Table** and then **Edit > Paste** (or **Paste with Column Names** if you copied the names of the variables as well).

Finally, you can import a dataset from a URL directly by selecting:

- **File > Internet Open** and pasting in the address of the website. JMP will attempt to find data on the page. It may take a few tries and some edits to get the dataset in correctly.

MINITAB

To import a text or Excel file:

- Choose **File > Open Worksheet**. From **Files of type**, choose **Text (*.txt)** or **Excel (*.xls; *xlsx)**.
- Browse to find and select the file.
- In the lower right corner of the dialog, choose **Open** to open the data file alone, or **Merge** to add the data to an existing worksheet.
- Click **Open**.

R

R can import many types of files, but text files (tab or comma delimited) are easiest. If the file is tab delimited and contains the variable names in the first row, then:

```
> mydata = read.delim(file.choose())
```

will give a dialog where you can pick the file you want to import. It will then be in a data frame called mydata. If the file is comma delimited, use:

```
> mydata = read.csv(file.choose())
```

COMMENTS

RStudio provides an interactive dialog that may be easier to use. For other options, including the case that the file does not contain variable names, consult **R** help.

SPSS

To import a text file:

- Choose **File > Open > Data**. Under "Files of type," choose **Text (*.txt,*.dat)**. Select the file you want to import. Click **Open**.
- A window will open called **Text Import Wizard**. Follow the steps, depending on the type of file you want to import.

STATCRUNCH

StatCrunch offers several ways to enter data. Click **MyStatCrunch > My Data**. Click a dataset to analyze the data or edit its properties.

Click a dataset link to analyze the data or edit its properties to import a new dataset.

- Choose **Select a file on my computer**,
- Enter the URL of a file,
- Paste data into a form, or
- Type or paste data into a blank data table.

For the "select a file on my computer" option, StatCrunch offers a choice of space, comma, tab, or semicolon delimiters. You may also choose to use the first line as the names of the variables.

After making your choices, select the **Load File** button at the bottom of the screen.

StatCrunch has direct access to the datasets on the text's website.

EXERCISES

SECTION 1.1

1. Grocery shopping Many grocery store chains offer customers a card they can scan when they check out and offer discounts to people who do so. To get the card, customers must give information, including a mailing address and e-mail address. The actual purpose is not to reward loyal customers but to gather data. What data do these cards allow stores to gather, and why would they want that data?

2. Online shopping Online retailers such as Amazon.com keep data on products that customers buy, and even products they look at. What does Amazon hope to gain from such information?

3. Parking lots Sensors in parking lots are able to detect and communicate when spaces are filled in a large covered parking garage next to an urban shopping mall. How might the owners of the parking garage use this information both to attract customers and to help the store owners in the mall make business plans?

4. Satellites and global climate change Satellites send back nearly continuous data on the earth's land masses, oceans, and atmosphere from space. How might researchers use this information in both the short and long terms to help study changes in the earth's climate?

SECTION 1.2

5. Lottery Every year, the administration of a lottery competition performs statistical analysis of previous records. They devise a list of the numbers drawn at each draw, the amount of money won at each draw, the number of times each number has been drawn during that particular year, and the number of draws since each number was last drawn. Identify the *Who* in this list.

6. Nobel laureates The website www.nobelprize.org allows you to look up all the Nobel prizes awarded in any year. The data are not listed in a table. Rather you drag a slider to the year and see a list of the awardees for that year. Describe the *Who* in this scenario.

7. Family growth The National Center for Health Statistics of a country conducts a national survey on family growth, which consists of over 5,000 interviews in each interviewing year. The sample chosen is representative of men and women aged 15–49 living in households. This extensive survey gathers information on "family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health." Describe the sample, the population, the *Who,* and the *What* of this survey.

8. Facebook. Facebook uploads more than 350 million photos every day onto its servers. For this collection, describe the *Who* and the *What*.

SECTION 1.3

9. Grade levels A person's grade in school is generally identified by a number.

a) Give an example of a *Why* in which grade level is treated as categorical.

b) Give an example of a *Why* in which grade level is treated as quantitative.

10. License plate numbers Passenger car license plates in a city are in the format 1ABC123.

a) In what sense are the car license plate numbers categorical?

b) Is there any ordinal sense to the car license plate numbers? In other words, does a license plate tell you anything about the date when the car was registered?

11. Referendum In a questionnaire administered to the members of a particular gym, the members were asked to specify how many hours they spend training at the gym every week. What kind of variable is the response?

12. Tablet dissolution Manufacturers in a pharmaceutical manufacturing company employ the technique of tablet dissolution to ensure that each drug is delivered properly to patients. This technique is used to measure the rate at which a drug releases from a dosage form. What kind of variable is the company studying?

13. Job satisfaction All employees in a company were asked to respond to the following question relating to their job satisfaction: "How satisfied are you in your current job position?" The possible choices were "Very Satisfied," "Satisfied," "Neutral," "Unsatisfied," and "Very Unsatisfied." What kind of variable is the response?

14. Stress A medical researcher measures the increase in heart rate of patients who are taking a stress test. What kind of variable is the researcher studying?

SECTION 1.4

15. Voting and elections Pollsters are interested in predicting the outcome of elections. Give an example of how they might model whether someone is likely to vote.

16. Weather Meteorologists utilize sophisticated models to predict the weather up to ten days in advance. Give an example of how they might assess their models.

17. The news Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, identify as many of the W's as you can. Include a copy of the article with your report.

18. The Internet Find an Internet source that reports on a study and describes the data. Print out the description and identify as many of the W's as you can.

(Exercises 19–26) *For each description of data, identify Who and What were investigated and the population of interest.*

19. **Brain age** Researchers have found that women's brains are nearly four years younger than those of men of the same chronological age. To explore the subject, the researchers used a brain scanning technique called positron emission tomography to measure the flow of oxygen and glucose in the brains of 121 women and 84 men aged 20–82. Through the scans, they could notice how sugar was being turned into energy in the volunteers' brains. The researchers also used a computer algorithm to predict the ages of the volunteers based on brain metabolism as measured by the scans. The findings of the study suggest that "changes in how the brain uses energy over a person's lifetime proceed more gradually in women than they do in men." (www.theguardian.com/science/2019/feb/04/womens-brains-are-four-years-younger-than-mens-study-finds)

20. **Hula-hoops** The hula-hoop, a popular children's toy in the 1950s, has gained popularity as an exercise in recent years. But does it work? To answer this question, the American Council on Exercise conducted a study to evaluate the cardio and calorie-burning benefits of "hooping." Researchers recorded heart rate and oxygen consumption of participants, as well as their individual ratings of perceived exertion, at regular intervals during a 30-minute workout. (www.acefitness.org/certifiednewsarticle/1094/)

21. **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. (Source: *NY Times*, Dec. 10, 2006)

22. **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.

23. **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. (Source: *NY Times*, Dec. 10, 2006)

24. **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. (www.blindness.org/index.php?view=article&id=2514:stem-cell-clinical-trial-for-stargardt-disease-set-to-begin-&option=com_content&Itemid=122)

25. **Not-so-diet soda** A look at 474 participants in the San Antonio Longitudinal Study of Aging found that participants who drank two or more diet sodas a day "experienced waist size increases six times greater than those of people who didn't drink diet soda." (*J Am Geriatr Soc.* 2015 Apr;63(4):708–15. doi: 10.1111/jgs.13376. Epub 2015 Mar 17.)

26. **Brewery** The management of a brewery is adamant on keeping the roasting temperature of barley exactly at the specified value of 232 degrees Celsius. This temperature yields perfectly roasted barley that is full of flavor yet not burnt. The management measured the temperature for 10 randomly selected portions of roasted barley.

(Exercises 27–40) *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).*

27. **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.

28. **Properties** The Property Management Department keeps these records on all properties in a particular district: age, type of property, location, number of residents living in that property, area, number of rooms, and whether the property has a garage.

29. **Cocktail menu** The owner and manager of a cocktail stall serves several different cocktails. On the stall's menu, the components, the number of calories, and the serving size in milliliters of each served cocktail are specified. This information is useful to assess the nutritional value of the different cocktails.

30. **Labor force** A representative telephone survey of 2,000 workers was conducted during the first quarter of 2019 for the purpose of gathering information about the labor force. Among the reported results were the respondent's age, gender, region (Northeast, South, etc.), occupation, and yearly average income.

31. **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).

32. **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.

33. **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

Place	Name	Official Time	Country
1	Kipchoge, Eliud	02:02:37	KEN
2	Geremew, Mosinet	02:02:55	ETH
3	Wasihun, Mule	02:03:16	ETH
4	Kitata, Tola Shura	02:05:01	ETH
5	Farah, Mo	02:05:39	GBR
6	Tola, Tamirat	02:06:57	ETH
7	Abdi, Bashir	02:07:03	BEL
8	Gebresilasie, Leul	02:07:15	ETH
9	Rachik, Yassine	02:08:05	ITA
10	Hawkins, Callum	02:08:14	GBR

Source: Excerpt from https://results.virginmoneylondonmarathon.com/2019/?pid=leaderboard

later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of benefits of the compound.

34. **Insurance claims** For fraud-detection purposes, an automobile insurance company keeps record of the following information about insurance claims: year of claim, claimant's age, claimant's gender, claimant's marital status, claimant's total economic loss (in thousand dollars), and whether the driver of the claimant's vehicle was uninsured.

35. **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).

36. **Vehicle registration** The Department of Motor Vehicles of a country receives daily applications for motor vehicle registrations. Among the data collected in the applications are the vehicle manufacturer, vehicle type (car, SUV, etc.), year of manufacture, color, engine number, chassis number, and seating capacity.

37. **Refrigerators** In 2013, *Consumer Reports* published an article evaluating refrigerators. It listed 353 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).

38. **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. (Source: *STATS* No. 39, Winter 2004)

39. **London Marathon** The London Marathon is an annual marathon held in London, United Kingdom, as part of the World Marathon Majors. It started in 1981 and has been held in the spring of every year since then. The London Marathon is set around the River Thames, beginning around Blackheath and finishing in The Mall alongside St James's Park. This marathon holds the Guinness world record as the largest annual fundraising event in the world. Above are the data for the top ten male finishers aged 18–39 of the 2019 race held on April 28, 2019.

40. **Indy 500 2018** 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 2013, the winner, Tony Kanaan, averaged over 187 mph, beating the previous record by over 17 mph!

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

Year	Driver	Time (hr:min:sec)	Speed (mph)
1911	Ray Harroun	6:42:08	74.602
1912	Joe Dawson	6:21:06	78.719
1913	Jules Goux	6:35:05	75.933
1914	René Thomas	6:03:45	82.474
1915	Ralph DePalma	5:33:55.51	89.840
…			
2013	Tony Kanaan	2:40:03.4181	187.433
2014	Ryan Hunter-Reay	2:40:48.2305	186.563
2015	Juan Pablo Montoya	3:05:56.5286	161.341
2016	Alexander Rossi	3:00:02.0872	166.634
2017	Takuma Sato	3:13:3.3584	155.395
2018	Will Power	2:59:42.6365	166.935

41. **Formula 1 on the computer** Load the Formula 1 data for the ten most recent races in Australia into your preferred statistics package, and answer the following questions:

a) What was the name of the winning driver in 2014?
b) Which country had the highest frequency of winning in the past ten races?
c) What was the winning time in 2011?
d) What is the name of the most recent Finnish winner and in what year did they win?

Year	Driver	Time (hr:min:sec)	Country
2010	Jenson Button	1:33:36.531	UK
2011	Sebastian Vettel	1:29:30.259	Germany
2012	Jenson Button	1:34:09.565	UK
2013	Kimi Räikkönen	1:30:03.225	Finland
2014	Nico Rosberg	1:32:58.710	Germany
2015	Lewis Hamilton	1:31:54.067	UK
2016	Nico Rosberg	1:48:15.565	Germany
2017	Sebastian Vettel	1:24:11.672	Germany
2018	Sebastian Vettel	1:29:33.283	Germany
2019	Valtteri Bottas	1:25:27.325	Finland

Source: Extracted from https://www.formula1.com/en/results.html/2010/races/861/australia/race-result.html

42. **Indy 500 2018 on the computer** Load the **Indy 500 2018** data into your preferred statistics package and answer the following questions:

a) What was the average speed of the winner in 1920?
b) How many times did Bill Vukovich win the race in the 1950s?
c) How many races took place during the 1940s?

JUST CHECKING

Answers

1. *Who*—Tour de France races; *What*—year, winner, country of origin, age, team, 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 2016; *Why*—not specified (To see progress in speeds of cycling racing?)

2.

Variable	Type	Units
Year	Quantitative or Identifier	Years
Winner	Categorical	
Country of Origin	Categorical	
Age	Quantitative	Years
Team	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)