PREFACE

This textbook embeds the usual set of introductory inferential statistical tests into a more pragmatic framework—one that develops a student's capacity to construct arguments and evaluate claims using quantitative evidence. We propose this approach because the principal use of statistics in the social sciences is to support arguments. Toward this end, the text will focus on building student competency in three interrelated areas (1) constructing good comparisons from observational data; (2) using inferential statistics (i.e., completing a statistical test), and (3) executing data management, data visualizations, and statistical tests using statistical software.

While statistical tests are important for evaluating a claim, they are but one step in the scientific process. Indeed, questions regarding the validity of claims rarely focus on whether researchers performed the calculations in the statistical test correctly or even whether the researchers used the correct test. Rather, the concerns typically focus on the methods chosen to measure critical elements of the researcher's broader argument (e.g., measurement validity), whether the comparisons in the analysis are appropriate (e.g., internal validity), and the degree to which we may generalize the results (e.g., external validity).

This text establishes a series of important points that can be regarded as guideposts for effective use of statistics. First, statistics aims to test arguments or claims, and these claims must attend to (1) measurement issues, (2) statistical design, (3) data collection, (4) data transformation, (5) statistical testing, and (6) interpretation. Consequently, these concepts should be taught together. Statistics textbooks generally recognize that it is more effective to teach statistical software together with the assumptions, internal logic, and interpretation of statistical tests. In this text, we take the next step and argue that it is just as important to also include design and measurement issues (e.g., constructing valid comparisons, measurement validity, and pitfalls of causal reasoning). In discussing these design issues, we will incorporate new ideas on statistical designs for observational data.

Second, causal arguments should precede tests of statistical significance. We must begin with a theory or a claim, consider appropriate measures of the key theoretical elements, and choose among a series of available data arrangements before we run a statistical test. More importantly, attention to details of this nature requires no advanced training in statistics. It is instead a process built on logic and context. Thus, a series of factors determine whether an analysis supports a causal claim, only some of which can be supported by statistical testing.

Third, the nature of the claim (association or cause) has important implications for gathering and arranging quantitative evidence. If we are trying to establish an association between the minimum wage and unemployment, we must concern ourselves with the choices of whether we test for an association in a single geographic unit over time (e.g., the entire United States) or by comparing across geographic units at a point in time (e.g., comparing the 50 U.S. states

in 2019). By emphasizing measurement and design issues, we highlight the correspondence between argument, data collection and transformation, and statistical tests. Without an understanding of measurement and design, students have no sense of how they might use statistical tests outside of a classroom experience. Worse yet, exclusive focus on running statistical tests leads students to conclude that a statistical test settles the argument.

Fourth, one cannot learn to use statistics effectively by following a fixed recipe. We concede that one important way to teach and test knowledge of some of the principles supporting statistical tests is to require hand calculations (i.e., computation of test statistics without the use of statistical software). However, using statistics properly requires judgement that varies with the nature of the argument, limitations of the data, and competing arguments on the issue. Consequently, this text focuses less on calculations and more on choices among appropriate statistical procedures and the pitfalls that threaten arguments built on quantitative evidence.

Fifth, all evidence-based arguments are built on comparisons, and strong arguments make meaningful comparisons. Comparisons are most meaningful when they support an argument and limit (or eliminate) the scope of alternative interpretations. If the comparison is flawed, statistical tests will be worthless. That is, arguments typically fail because they are built on weak or flawed comparisons, as opposed to reliance on poor statistical technique.

My experience serving as an advisor for student research projects for more than 20 years suggests that increased attention to arguments, measurement, and statistical designs is a core need. Most junior- and senior-level students can execute statistical tests on prepared data sets, especially with the convenience of commercial software. Problems emerge when these students must construct comparisons.

Common design questions that typically arise in the context of executing an inquiry using observational data include the following: Is there a way to create an empirical measure(s) from available data that neatly aligns with the research question? What is the appropriate unit of analysis? What is the appropriate time frame? Should we compare over time or across entities at a point in time? Should we aggregate the data by geography or disaggregate to the individual level? Is a simple count sufficient or must we divide the count by population or some other measure? Can we identify a natural experiment? What is the best counterfactual? This text will use examples to alert students to these issues and propose a series of steps to produce workable solutions.

If we are generating the data as part of the analysis, we may be able to randomly assign observational units (typically people) across conditions as in a vaccine trial where the experimental group receives the vaccine while the control group receives a placebo. However, we do not typically have that luxury in the social sciences. Instead, researchers in the social sciences (e.g., criminology, economics, political science, and sociology) must use observational data. Because most studies in the social sciences employ observational data and observational data raise a series of design issues that typically do not occur in data generated though a random assignment procedure, we focus on design issues for observational data.

This increased attention to arguments, measurements, and statistical designs also follows from two important shifts in the social sciences: (1) a shift in emphasis in empirical methods toward explicit attention to research designs and (2) a heightened emphasis on student research. Over the past 25 years, empirical researchers in the social sciences have argued for a more explicit focus on better and more clearly articulated research designs. Angrist and Pischke (2010) refer to

this shift as the "credibility revolution."[1] A defining feature of this credibility revolution is using random-assignment procedures as a guidepost in selecting and arranging observational data.

This focus on issues of design leads to a key insight regarding instruction in statistics: Measurement issues and poor research design are pitfalls to causal reasoning. It is difficult for students to appreciate the critical importance of measurement and research design without explicit coverage of the topic. Debates over validity of claims made in quantitative analyses rarely, if ever, focus on the precision of test statistic calculations. Instead, they focus on the series of judgments that users of statistics make when they execute comparisons. These judgments are not always simple since they must account for data constraints and the nature of the theoretical claim.

This text is intended to support a course for social science students with no prior background in statistics. The text presumes only a background in algebra. While some may point out that many statistical designs (e.g., difference-in-differences) are associated with statistical tests that require more advanced training (e.g., multiple regression), we contend that it is possible to usefully employ many of these designs without the benefit of advanced statistical training or tests. In short, design issues precede and are analytically separate from tests of statistical significance.

We also contend that students will benefit if they learn to think more broadly about quantitative evidence as early in their college education as possible. Students will benefit from this broader approach because these skills are useful across the rest of their curriculum as well as in their future careers. The recent rise in society's emphasis on worker numeracy, data science, and knowledge of informatics makes such early learning even more important. Better training on these topics will allow students to become better critical thinkers in their social science and business courses. Instead of memorizing the received wisdom transcribed in textbooks, students may begin to think critically about the claims advanced based on the quality of the evidentiary support. Such thinking is a core analytical skill.

Because many of these concepts are abstract, the text includes excerpts from papers using quantitative evidence to highlight the choices researchers made in constructing comparisons, why the researchers chose a particular comparison (or counterfactual), and how the comparison supports or fails to support the researcher's argument. For our examples, we focus on social and policy issues that are likely familiar to students and have relatively straightforward causal arguments. We avoid arguments with long causal chains that presume a significant amount of training in the social sciences. This allows use of the text early in the students' college careers before they have extensive training in the social sciences. It also allows use of the text across the social science disciplines.

The organization of the text generally follows a logical progression. Early portions of the text emphasize measurement issues, data visualization, statistical fallacies, threats to validity, and design issues. Later portions of the text will focus on statistical tests (e.g., one-sample *t*-tests, two-sample *t*-tests, correlation, and regression). We emphasize the intuition behind the tests and the value of conducting tests on a set of logically connected comparisons. We consider an understanding of sampling distributions and the central limit theorem to be critical. Without such foundational understanding, students may execute statistical tests without understanding their implications or limitations.

The text fosters a logical progression of skills aimed at producing exemplary outcomes in student research projects. We offer instruction in use of statistical software and data visualization

to correspond with the topical coverage in the text and argue that data visualization is not only a method to show known results but also a tool for uncovering the nature of relationships among variables.

The strategy regarding software coverage will vary with chapter type. Some chapters address more theoretical/conceptual issues and therefore do not require software instruction (i.e., chapters 1, 2, 5, 6, and 7). For the chapters that require software support, we integrate the software commands into the text. We integrate the commands into the text because decisions about which data transformations and comparisons to execute cannot be separated from the commands that execute data transformations and comparisons.

Our coverage is selective by necessity. Explaining the full set of data management and display functions for just one software program could consume a volume by itself. Rather than attempting to explain the syntax and features of a large number of software commands, we instead aim to build student understanding of the capacity of the software package and provide a vocabulary of associated concepts. With this understanding, students should be able to execute on-line searches for more detailed information on more esoteric commands.

The remaining chapters include software instruction on creating new variables, univariate measures, and statistical tests. In these chapters, we explain concepts that can then be executed using software. The thinking comes first; students need to understand what the software is doing. Once students achieve a level of understanding, they can progress to using software commands to execute the task (e.g., two-sample t-tests). In these chapters, we attach commands and outputs in an appendix. To ensure the text meets the needs of a wider range of instructors, we will include Stata and Excel commands in these appendices. However, these online appendices will not be software manuals. Instruction on the full capabilities of software like Stata and Excel is easily obtained elsewhere.

We include discussion of Excel because a significant percentage of social science data is stored and accessed as Excel files. Excel also requires less upfront investment for students to learn simple data tasks and statistical analyses. On the other hand, Excel is not equipped to easily complete more advanced data management tasks that are routine in dedicated statistical software packages, and it lacks an accessible programming language for students to write their own code. Many instructors may find it valuable for students to turn in programming scripts with their work as it allows them to evaluate the student's process instead of just the result of their analyses. It is also our judgment that students generally benefit from increased exposure to coding because it helps develop widely applicable problem-solving skills. However, nearly all software tasks in the book may be completed in Excel and we leave the ultimate judgment about software choice to the instructor.

To support software instruction as well as instruction more generally, we include a series of data sets with the book. These data sets vary in length from small (30-40 observations on 3 to 5 variables) to large (500-5,000 observations on 10-30 variables) and include notation on sources and variable definitions. The data sets are available at **sagepub.com** (or contact your Sage representative at **sagepub.com/findmyrep**) and are easily downloadable in Excel and Stata formats. End-of-chapter questions will often refer to these data sets. We will also, where possible, write end-of-chapter questions that do not rely on data sets.

Chapter 1

Making the Right Comparison

Understanding the Rules and Limitations of Quantitative Reasoning

The purpose of Chapter 1 is to introduce the elements of reasoning with quantitative data.

LEARNING OBJECTIVES

1. Explain why "statistics" is not just completing a set of calculations using a prescribed formula.
2. Differentiate between positive and normative statements and between deduction and induction.
3. Examine association, causal claims, and spurious correlation in specific applications.
4. Identify the strengths and weaknesses of induction and deduction in supporting causal claims.
5. Explain the value of conjoint use of deduction and induction.
6. Support that establishing cause requires deliberate attention to theory and design.
7. Demonstrate how measurement efforts can fail because empirical measures are not aligned with theoretical ideas and because the measurement process is flawed.

Two of the most important questions in the English language, or indeed any language, are "So what?" and "How do you know?" More often than not, we appeal to quantitative evidence to answer these two critical questions. Consider debates regarding Covid-19. "So what?" questions aim to establish the relevance of the topic. Establishing relevance requires an answer that is clear and precise. Numbers fit the bill; we should be concerned because (as of this writing) Covid-19 has caused the deaths of more than a million Americans, a surge in the unemployment rate to 14.8%, and disrupted learning for 55 million students in grades K–12.[1] The second question is equally important. Anyone can make up a set of numbers. Convincing skeptics requires that we report a logical process used to generate the numbers. Executing such a task requires judgment and skill. In a world in which we increasingly settle debates by appeals to quantitative

measures, answering "How do you know?" is more important than ever. This text aims to offer instruction on the logical process to produce reliable quantitative evidence and to critique evidence offered by others.

Broadly conceived, statistics is a logical process that aids our efforts to produce reliable knowledge of the world around us. In this chapter, we offer an overview of this process. We focus on offering a set of rules for organizing our thoughts and placing statistical procedures in their proper place within the set of rules. To make the rules clear, we introduce a set of concepts and analyze the relations among these concepts. At the most basic level, all quantitative arguments are built on comparisons. As we shall show, some comparisons are better than others, and the quality of the comparison depends on a series of contextual factors. To properly account for these contextual factors, we need to understand why our analyses must vary with context. Understanding these concepts helps us to draw appropriate conclusions about the best procedure to use in a particular context as well as the proper interpretation of the results of the procedure. We do not want to overlook significant pieces of evidence, nor do we want to make strong claims based on weak evidence.

Positive and Normative Statements

Of course, not all statements or claims about the world are of the sort that can be verified. Some claims are value judgments rather than attempts to describe the world. We refer to these value judgments as **normative statements**. For instance, we may claim that U.S. law should guarantee abortion rights, changes to the tax code should aim to reduce unemployment, or state governments should work to raise the proportion of college degree holders in the population. Each of these statements is prescriptive. That is, rather than describe the world, they make a claim about how the world ought to be. Normative statements are opinions and, consequently, cannot be tested and shown to be false. That is, the statements are not falsifiable. Some of these statements are relatively uncontroversial while others are not. For instance, the claim that abortion rights should be guaranteed through legislation will likely elicit disagreement from a significant percentage of the U.S. population, while a claim that South Carolina should work to raise the proportion of college degree holders will not.

Often normative claims that appear uncontroversial become the subject of controversy when we realize that achieving the normative goal brings us into conflict with other desirable normative goals. For instance, altering the tax code to reduce unemployment can also increase income inequality. That is, we may cut taxes on wages as a method to induce employers to hire more workers, but, in the process, high-earning workers may keep more of their earnings. This, in turn, would increase the income difference between rich and poor.

Similarly, we may increase taxes on gasoline to reduce the level of pollution. However, the poor spend a higher percentage of their income on gasoline, and, as a result, the burden of such a tax falls more heavily on the poor. In addition, the job losses from decreased employment in fossil fuel production fall more heavily on the less educated. Failing to acknowledge that a normative claim conflicts with other widely held normative values or claims will lead skeptics to dismiss the argument because it fails to account for values that they consider important.

In contrast to normative statements, **positive statements** or claims are claims that seek to describe the world. Positive statements tell us what is. However, positive statements are not necessarily true. Indeed, one important characteristic of a positive statement is that it is, at least in theory, falsifiable. That is, we can imagine an observation or a set of observations that would show the statement to be false. A major focus of this book is to specify a set of procedures or best practices that lead to positive statements that are unlikely to be false. In testing these positive claims, we compile evidence. To produce this evidence, we use two basic types of procedures: **deduction** and **induction**. As we shall see, these procedures work best (or most effectively) when they are used in tandem.

Question for Review 1.1: Offer a new example of a seemingly uncontroversial normative statement (or goal) that becomes controversial once we recognize that achieving the goal conflicts with other normative statements (or goals).

Deduction and Induction

Deduction is top-down logic while inductive reasoning is bottom-up logic. Deductive reasoning begins with a set of assumptions (or premises) and employs logic to derive conclusions. If the assumptions are true and we follow the rules of deductive logic, then the conclusion that we reach must be true. This seems clear enough, but note that there are no guarantees that the assumptions are true. Consequently, the typical weaknesses of deduction are logical errors and false (or unsupported) assumptions. Deduction has been with us for some time. Logical arguments (also referred to as syllogisms) that date to the ancient Greeks employ deduction. For instance, we may claim the following:

1. All college students are tall.
2. Jane is a college student.
3. Jane is tall.

Note that the conclusion "Jane is tall" must follow from the assumptions and that questions regarding the validity of the conclusion will focus typically on the assumptions (e.g., "All college students are tall.").

Standard models in the social sciences like the supply-and-demand model or the median-voter model are examples of deduction—once we accept the assumptions, the conclusion must follow. The median voter model contends that when we use a direct majority vote to answer a yes-or-no decision (a binary decision), the preferences of the median voter will determine the winning option. To support this conclusion, the median-voter model relies on several assumptions. For instance, the model assumes that we can arrange votes and policies along a one-dimensional distribution and that the proposal subject to the vote will receive support from all voters to the right (or to the left) of a certain point in the

distribution. The usefulness of the model is that we can apply it to particular cases to organize our inquiry. For instance, we may be able to explain why one referendum on affirmative action passed while another failed. Once again, as in the case of our simple syllogism regarding Jane, we reason from the general to the particular, and the weakness is typically in the validity of the premises.

Frequently, deduction relies on portions of larger theories. For instance, we may employ the theory that price and quantity demanded are inversely related (i.e., consumers purchase less when the price of an item is higher) rather than an entire model of the market. This theory of demand contends that an increase in the price of some good reduces the quantity consumed of the good because the higher price implies that the consumer must forgo more units of all other goods to obtain the good in question and because the now higher price reduces the consumer's buying power. We can then use this theory to support the claim that an increase in a subsidy for health insurance that cuts the out-of-pocket costs of insurance to consumers will reduce the number of uninsured citizens. Once again, as in the case of the median-voter model, we reason from the general to the particular.

Induction, by contrast, generally reasons from the particular to the general and therefore has different strengths and weaknesses. Under induction, we make positive statements based on experiences and observations. These observations may include our own direct experience as well as what we learn from others. We then synthesize or aggregate these observations to make more general statements about the world. Picking up on the previous topics (affirmative action referenda and health insurance), we can offer examples of induction.

First, we may examine a series of referenda on affirmative action that occur at different times and locations, use polling data to determine the median voter's preferences, record whether the referendum passed and whether the result was consistent with the median voter's preferences. In a similar way, we may examine a series of health insurance subsidy programs and record whether payment of the subsidy raises the percentage of the target population that has health insurance. In each case then, we draw a conclusion about whether the median voter's preferences predict referenda outcomes more generally and whether health insurance subsidies raise the number of insured citizens.

Note that induction provides *some* evidence, but not full assurance, for the general conclusions. The validity of the conclusion of an inductive argument is only probable, based on the evidence gathered. It is possible that we have not seen a representative set of cases. For instance, in other cases (not observed), the median voter's preferences may not predict the outcome. Indeed, falsification offers the strongest evidence; it implies that the theory is invalid. When we *fail* to falsify, we find support for the theory. However, the evidence may also support other theories, some of which we may not even consider as possibilities. Thus, we may conclude that the most reliable theories are those that have survived repeated attempts at falsification.

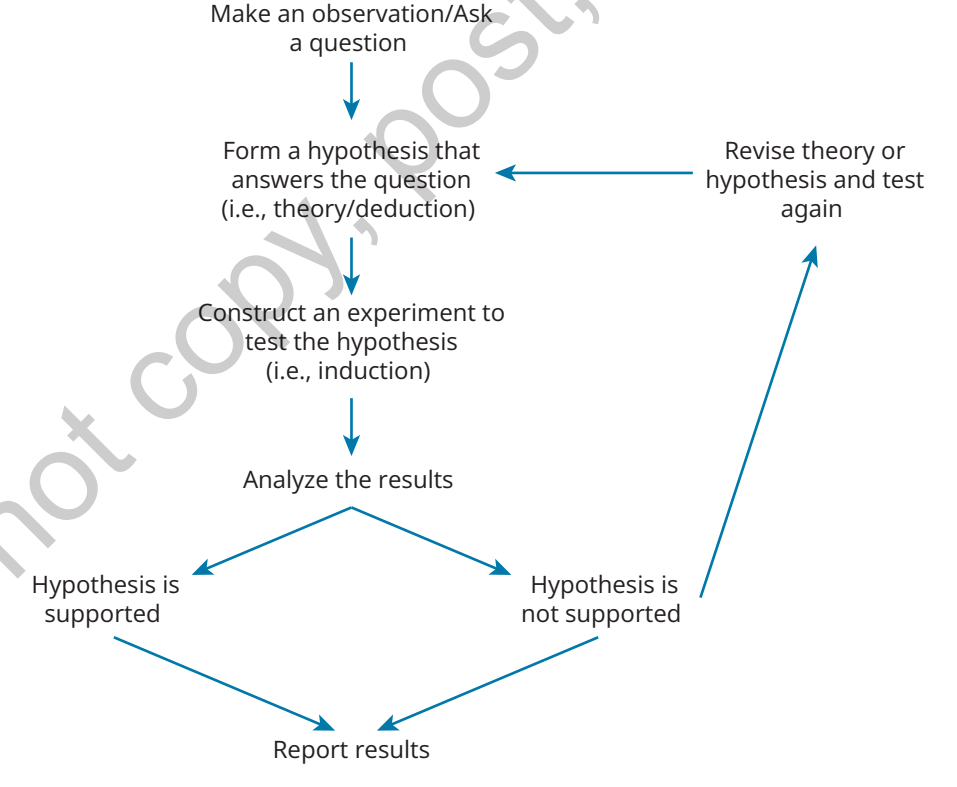
Question for Review 1.2: What are the respective weaknesses of deduction and induction?

Using Deduction and Induction Together

Because the weaknesses (or blind spots) differ between induction and deduction, use of deduction and induction in tandem leads to more reliable knowledge. This use of deduction and induction together is the essence of the scientific method. Figure 1.1 provides a schematic representation of the scientific method. The first seeds of a theory (deduction) are often a set of off-hand observations or anecdotes (induction). The theory then informs more detailed data gathering and hypothesis tests (induction), which in turn informs theoretical refinements. The critical element is a feedback loop that uses the results from the tests to modify the theory and facilitate the design and execution of new experiments. This feedback loop produces more reliable knowledge because induction and deduction have different weaknesses. Thus, one compensates for the other.

The weakness of deduction is that the premises may not accurately capture the process at work. For instance, we may claim that higher prices reduce consumption (deduction). Induction offers a method to test these claims. The weakness of induction is that specific findings may not generalize to other cases. Deduction offers guidance on the extent or limits of our ability to

FIGURE 1.1 ■ The Scientific Method



generalize. For instance, the law of demand applies to publicly traded goods in arms-length transactions (transactions between strangers).

In addition to a reliance on induction or deduction, positive statements answer a question. For instance, opinion data can answer the question, "How many Americans support legal restrictions on abortion?" A display of the monthly unemployment rate for the United States over the past 20 years can answer the question, "How does the current unemployment rate compare with recent unemployment rates?" A table that reports the percentage of the 25- to 44-year-olds that have a four-year college degree for states in the south Atlantic United States can answer the question, "Which south Atlantic states have high (or low) percentages of college-degree holders in their population?" Note also that in each of these cases, the question is built on a comparison. In the first instance, we are simply comparing yesses and noes. In the second instance, we are comparing later values of the unemployment rate with earlier values. In the final instance, we compare among U.S. states.

From Table 1.1, we can infer that there is little support for a ban on abortion among U.S. adults in May 2023. From Figure 1.2, we may explain that the U.S. unemployment rate experienced an enormous and historically unique increase in 2020. Finally, Table 1.2 allows us to explain that residents of Maryland and Virginia in the 25-to-44-year age range show a higher percentage of bachelor's degree completion than residents of South Carolina.

In addition, we can imagine an observation or a set of observations that would show the statement to be false. For instance, subsequent investigation might reveal that 53 percent of Americans support an abortion ban, not 13 percent, and that the original sample was not representative of the U.S. population. In a similar way, we might discover that a coding error distorted our measure of unemployment or bachelor's degree attainment.

Question for Review 1.3: Table 1.2 answers the question "Which south Atlantic states have high (or low) percentages of college-degree holders in their population?" Does the table answer other questions?

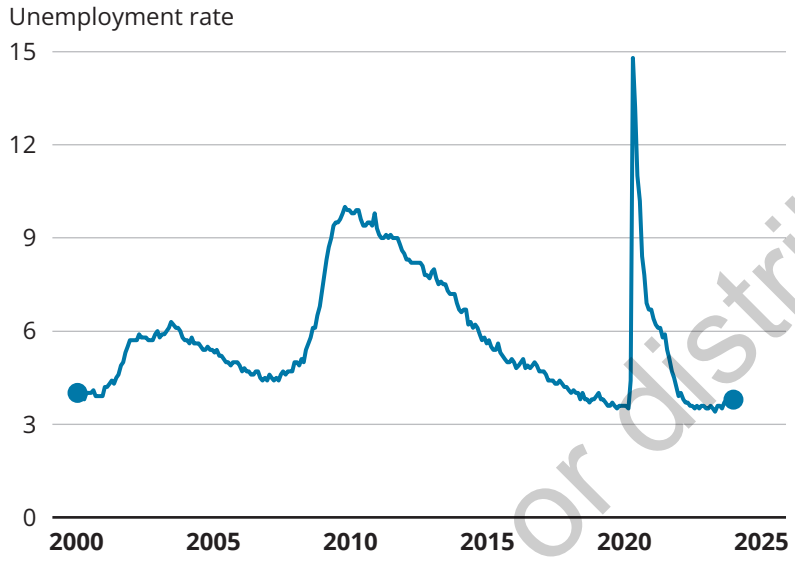
TABLE 1.1 ■ Americans' Preferences Regarding Abortion Regulation, May 2023

Response	Percentage
Legal under any circumstance	34
Legal only under certain circumstances	51
Illegal in all circumstances	13

Source: Saad, L. (2023). Broader Support for Abortion Rights Continues Post-Dobbs, Gallup Poll Report. June 14, 2023. https://news.gallup.com/poll/506759/broader-support-abortion-rights-continues-post-dobbs.aspx

May 2023 Gallup survey of 1,011 U.S. adults. Question: "Do you think abortions should be legal under any circumstances, legal only under certain circumstances, or illegal in all circumstances?"

FIGURE 1.2 ■ Unemployment Rate (U-3) 2000 to 2023



Source: U.S. Bureau of Labor Statistics.

TABLE 1.2 ■ Percentage of Residents 25 to 44 years with a Bachelor's Degree and Median Annual Wages for South Atlantic U.S. States in 2018

State	Percentage with Bachelor's Degree	Median Annual Wage
South Carolina	30.8	$33,750
Florida	32.2	$34,560
Georgia	34	$35,950
North Carolina	35.7	$35,750
Virginia	43.2	$40,820
Maryland	43.9	$44,690

Source: Bachelor's degree: American Community Survey, 2018, 5-year estimates. Wages: Bureau of Labor Statistics

Cause and Association

One important subset of positive statements are statements regarding cause and association. **Association** implies that knowledge of the value of one variable (x) offers information about the value of a second variable (y). By contrast, **cause** implies that a change in one variable (x) will reliably produce a change in the other variable (y). Restated, the change in y would *not* have occurred but for the presence of x (i.e., x implies y *and* not x implies not y). If x causes y,

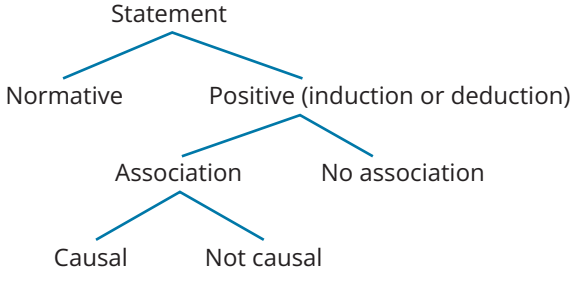
then knowledge of *x* implies information about the value of *y*. Thus, all causal relationships are associations. However, the reverse is not true. The reverse is not true because associations can occur for reasons other than *x* producing a change in *y*. We refer to these noncausal associations as **spurious correlations**.

Because statements of both association and cause are positive, they are also both descriptive and falsifiable. To make a statement regarding cause or association we need a minimum of two variables. For instance, the public opinion data on abortion in Table 1.1 do not support any statements of association or cause. Figure 1.1, by contrast, shows a relationship between time and the unemployment rate. We see for instance that the unemployment rate generally falls with time from 2010 to the first months of 2020. In early 2020, the association reverses, and the unemployment rate rises abruptly.

Table 1.2 shows an association between percentage of 25- to 44-year-olds with a bachelor's degree and median annual wages for South Atlantic states. As the percentage of 25- to 44-year-olds with a bachelor's degree rises, the median annual wage also tends to rise. The association is not perfect as Georgia shows a slightly higher median wage than North Carolina but a slightly lower percentage of bachelor's degrees among 25- to 44-year-olds. Still, knowing the value of one of the variables provides information about the value of the other. That is, knowing the percentage of 25- to 44-year-olds in a state with a bachelor's degree gives us reliable information about the median annual wage in the state. We can therefore claim that bachelor's degrees (measured as a percentage 25- to 44-year-olds at the state level) predict state-level median annual wages. Of course, we prefer to know the actual value of the median annual wage as we rarely see perfect associations in the social sciences. Later, we will discuss methods to measure the strength of association between two variables and these measures will gauge the usefulness of *x* in making predictions of *y*.

To clarify the relationships between the statement types discussed thus far, Figure 1.3 shows a decision tree. Initially, we divide the world into normative (i.e., statements of opinion) and positive (i.e., statements of fact) statements. These positive statements can be supported using either deductive or inductive processes and can show associations among variables. However, not all associations follow from a causal relationship, and much of the debate in social science centers on whether the relationship between two variables is in fact causal (and *not* spurious).

FIGURE 1.3 ■ Relationships Among Statement Types



While the dividing line between causal relationships and spurious correlations is clear in theory, in practice, clear identification of causal relationships is difficult. As we have noted, all causal relationships are associations, while the reverse is not true. To differentiate causal relationships from spurious correlations we need more information beyond the simple fact of an association. Consequently, causal claims require more information (or support) than statements of association. As we shall see, social scientists often spend quite a bit of time and effort trying to separate spurious correlations from causal relationships. Typically, these spurious correlations follow from associations between variables that occur at random or from a third factor that is often not obvious. We refer to these third factors as **confounding variables**.

For instance, we may argue that the surge in Covid-19 cases in March of 2020 caused the U.S. unemployment rate to rise or that a high percentage of 25- to 44-year-olds in the population with a college degree causes higher median annual wages. Such causal statements are an important focus in the social sciences (and in the sciences) because understanding the causal roots of an outcome clarifies the process that generates the outcome and is the first step in working to change that outcome.

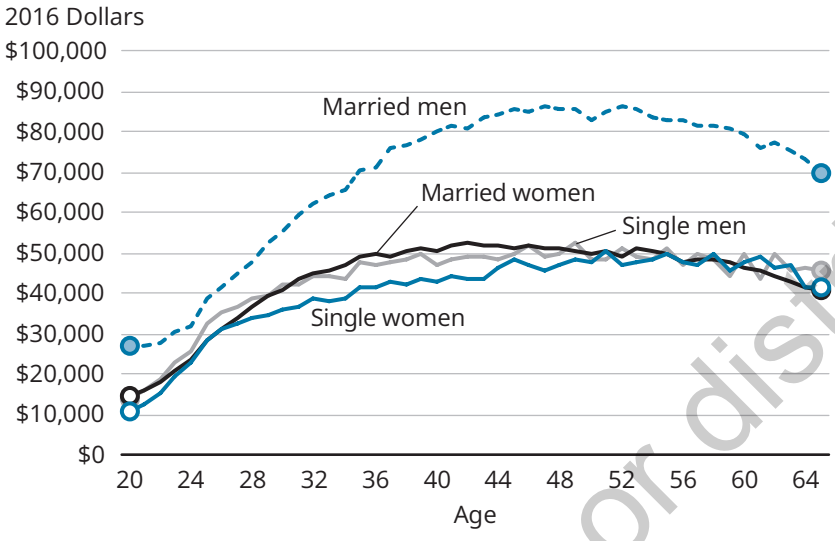
Earlier, we argued in favor of using deduction and induction in tandem as part of a strategy to derive reliable knowledge. This strategy is particularly effective in efforts to separate causal relationships from spurious correlations. Returning to Figure 1.1 and beginning with deduction, specific detailed theories about how one variable affects another are valuable tools to build a case for cause. To demonstrate, let us consider a well-known association: the so-called marriage wage premium. The marriage wage premium refers to the fact that in the United States and other developed countries, married men earn more than unmarried men. Figure 1.4 shows that for the United States, this gap is considerable and that it increases for men in their late 20s through the late 40s.[2] In the late 40s, the gap stabilizes and then decreases slightly for older men.

A series of recent analyses by economists and sociologists have attempted to determine whether any portion of this gap is causal. Restated, does the marriage wage premium appear because (1) marriage leads to changes in behavior and these behavioral changes raise wages (causal), or (2) men who marry have different attributes than men who do not marry and these attributes lead to both higher wages and marriage (spurious correlation)? To distinguish between causal relationships and spurious correlations, we need to think carefully about how and why wages may vary when the relation is either causal or spurious.

Researchers have offered a series of reasons that married men might have higher wages than unmarried men. Some of these explanations imply a causal explanation while others suggest a spurious relationship. One causal explanation is that marriage permits specialization between marriage partners, and this allows men to specialize in labor-market tasks. This specialization will in turn cause faster wage growth after marriage. Other causal arguments include lifestyle changes and employer favoritism. That is, marriage may cause men to make better lifestyle choices, or employers believe married men deserve more to meet the financial needs of their family. In such cases, wages for married men rise relative to unmarried men following marriage.[3]

By contrast, the arguments for a spurious correlation between male earnings and marriage contend that men with higher ability are more likely to marry. We may refer to this higher ability as a potential confounding variable in our effort to assess whether marriage causes higher

FIGURE 1.4 ■ 2016 Wage and Salary Income by Gender, Marital Status, and Age



Source: IPUMS-USA, University of Minnesota; www.ipums.org

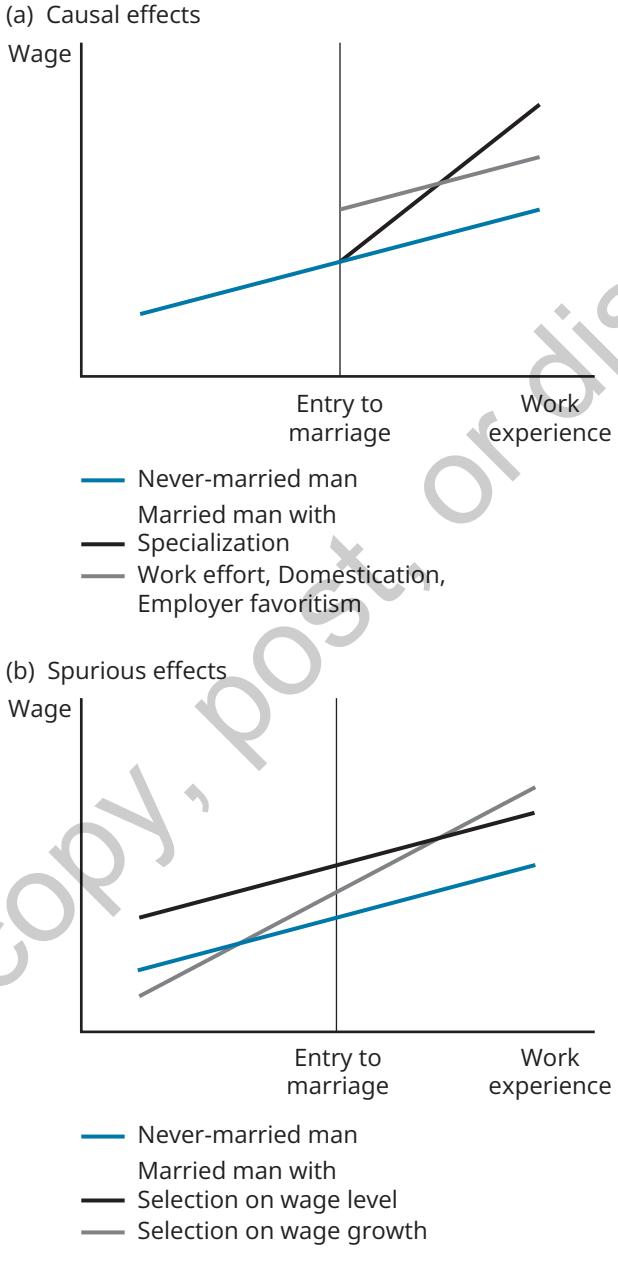
Note: Data cover employed men and women with at least a high school diploma. From Vandenboucke, G. (2018). Married Men Sit Atop the Wage Ladder. St. Louis Fed Economic Synopses. https://fraser.stlouisfed.org/title/economic-synopses-6715/married-men-sit-atop-wage-ladder-624534

wages for men. Higher ability men will in general have higher wages or faster wage growth prior to marriage. Restated, high ability causes both marriage and high wages. Because high ability causes both marriage and high wages, we see a spurious correlation between marriage and high wages. Importantly, the attributes that lead to higher wages or faster wage growth pre-date the marriage. The key then is to observe premarital earnings for men that eventually marry and compare them with similar men who never marry.

Figure 1.5 illustrates the pre- and postmarriage wage trends for each of these possible explanations for the male marriage premium.[4] Panel a of Figure 1.5 depicts the possible patterns in wages for married and unmarried men if marriage causes higher wages. In such cases, men who marry should experience either an increase in wages that coincides with the marriage or an increase in wage growth that coincides with the marriage.

The figure depicts each of these outcomes as lines that rise relative to the solid blue line (wages for unmarried men). If postmarriage wage changes follow from greater specialization, we should see faster postmarriage wage growth for married men as the men discover new and more efficient methods to cooperate with their new spouse to complete household tasks thereby freeing up more time for their labor-market effort. We see this outcome as the black line. If instead, postmarriage wage changes follow from greater changes in work effort or employer favoritism, we should see a one-time increase in wages for married men following marriage. We see this outcome as the gray line.

FIGURE 1.5 ■ Hypotheses on the Effects of Heterosexual Marriage on Men's Wage Trajectory from Ludwig and Brüderl



Source: Ludwig, V. and Brüderl, J. (2018). Is There a Male Marriage Premium? *American Sociological Review.* 83(4). 744-770.

Panel b of Figure 1.5 depicts the possible patterns in wages for married and unmarried men if the relationship between marriage and wages is a spurious correlation. If the relation between marriage and wages is a spurious correlation, unmarried men who later marry should show either higher wages or faster wage growth compared unmarried men who never marry in the premarriage period.[5] It is important that we focus on the premarriage period for those who later marry. During this period, marriage cannot have a causal effect on wages as neither group is married (yet). If we see differences in wages between men who never marry and those who marry later, then we know that those who marry later are different in important respects than those who never marry. Thus, these differences are likely to cause both the marriage and the higher wages, and the relation we see between marriage and wages is spurious.

In general, our strategy to identify causal effects will depend on our theory of how changes in one variable (marriage) lead to changes in a second variable (wages), as well as the available data. The procedure above assumes that researchers cannot directly observe postmarriage changes in behavior for married men. Instead, it assumes that unobserved changes in behavior caused by marriage also cause changes in wages after men marry. If we can devise and collect reliable measures of specialization (e.g., men complete fewer household tasks of a certain type) and compile data on these measures for men who marry, we can test an additional causal link. We can compare time devoted to certain household tasks for men both before and after they marry. If the effect is causal, men who marry should show decreases in time devoted to certain household tasks following marriage, and these changes should predict higher wages following marriage.

While the marriage wage premium is just one example, it offers insights into the typical difficulties of separating spurious correlations from causal relationships as well as the methods that researchers employ to surmount these difficulties. From this example, we see the following:

- Careful thought is needed to specify exactly why (1) the relation may be causal (i.e., changes in the A variable lead to changes in the B variable), and (2) the relationship may be spurious (i.e., possible confounding variables that may lead to a link between the A variable and the B variable).
- These thought processes (or theories) should be linked to specific comparisons in the data.
- Breaking down the causal chain can often lead to additional tests of whether the relation is causal.

Question for Review 1.4: Why is it more difficult to establish that A causes B than it is to establish that A is associated with B?

Linking Deduction With Induction—Measurement Validity

When we analyze data (or employ induction), our efforts are always informed at some level by theory (or deduction). This must be the case as there are literally millions of analyses we can undertake. Choosing among them requires a theory—even if the theory is stunningly

simple. As noted previously, the scientific method provides guideposts for aligning deduction or a theoretical idea with induction or an empirical construct. Nevertheless, we can never avoid making judgments about the best measures of our theoretical constructs to employ in our analysis. We refer to the correspondence between theoretical constructs and empirical measures as **measurement validity**. That is, measurement validity implies that deductive processes must align with our inductive processes.

In some cases, the process is rather straightforward. The federal, state, and local governments keep counts of crimes such as murder, and government officials maintain these counts (i.e., empirical measure) in a way that aligns with commonly understood definitions of "murder" (i.e., theoretical construct). Of course, we are left to decide the time frame over which we will tabulate the count (e.g., monthly or yearly) and whether we should divide the count by population (i.e., a divisor) to facilitate comparisons across states or over time.

In other cases, the theoretical idea may not match neatly with the empirical measure because available measures admit to multiple interpretations. That is, the measure is overly broad. Consider once again Table 1.1. Table 1.1 reports results of a Gallup survey of 1,011 adults aged 18 and over, living in the United States on the question "Do you think abortions should be legal under any circumstances, legal only under certain circumstances, or illegal in all circumstances?" Suppose we wish to use these survey results to assess "opposition to current abortion law" (i.e., theoretical construct) among Americans.

Reasonable critics might be concerned whether these survey responses accurately capture opposition to current abortion law as 51% of respondents answered that abortion should be "legal only under some circumstances." The circumstances under which these respondents oppose abortion are unclear. Indeed, the circumstances can be either more (or less) restrictive than current law (which now varies across U.S. states), and it is therefore not entirely clear whether these respondents are in this sense abortion supporters or abortion opponents. As such, these survey responses have weak measurement validity for purposes of assessing opposition to current abortion law.

Just as an overly broad empirical measure can reduce measurement validity, so can an overly broad theoretical construct. Consider, for instance, the unemployment data reported in Figure 1.2. As you will recall, Figure 1.2 shows the U.S. unemployment rate over time. If the purpose is to show the unemployment rate, the measure is valid. If, however, we make more a general argument about labor-market conditions over time in the United States (i.e., the theoretical construct) using the unemployment rate, the potential for a mismatch between the theoretical construct and the empirical measure increases. The potential for a mismatch rises because the unemployment rate does not perfectly capture labor-market conditions. The unemployment rate does not always accurately capture labor-market conditions because we are forced to make a series of judgments in creating the ratio.

The most often used unemployment rate is the ratio of the number of unemployed workers to the sum of the employed workers plus the unemployed workers, times 100 as seen in Equation 1.1. Economists refer to this statistic as U-3.

$$\text{Unemployment Rate} = \frac{Unemployed}{Unemployed + Employed} \times 100 \tag{1.1}$$

The U.S. Bureau of Labor Statistics defines an unemployed person as someone who reports that they do not have a job and has looked for work in the past 4 weeks. Of course, this calculation would generate a slightly different answer if we chose 3 weeks or 6 weeks or simply dropped the search requirement altogether. Under the government's definition, individuals who are not currently employed and have not looked for work are not included as "unemployed." We include the 4-week requirement to eliminate individuals who are not seeking work (e.g., retired individuals, active-duty military, stay-at-home parents). Counting such people will produce a misleading indicator of the ability of the labor market to absorb willing workers. Similarly, the U-3 definition counts all part-time workers as employed, even those who have been forced to settle for part-time employment when they would in fact prefer full-time employment.

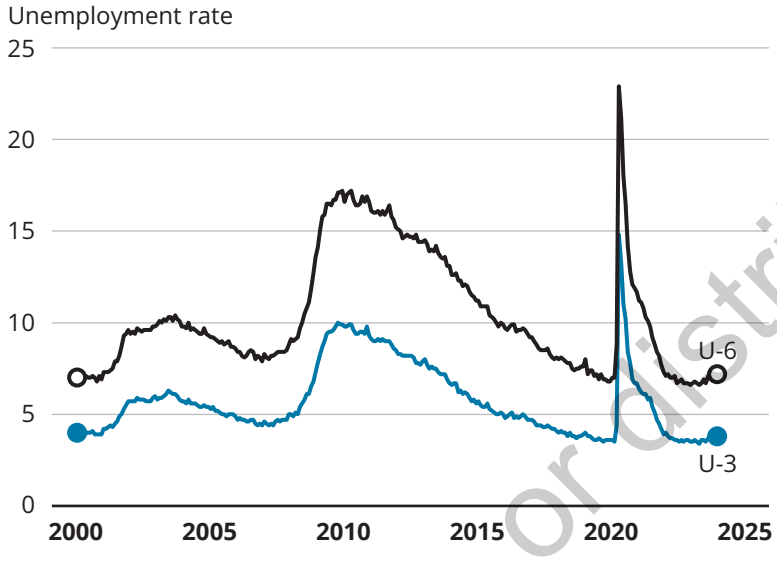
Because the U-3 measure counts unemployment in this way, unemployment rates may *sometimes* rise even when the economy is generally improving. Improvements in hiring activity may induce workers outside of the labor force to initiate a search for work (i.e., enter the labor force). As these workers enter the labor force, the number of unemployed workers rises at a faster rate than the labor force, and the unemployment rate rises. In so doing, U-3 rises with an improvement in labor-market conditions. The general strategy in cases like this is not to discard the measurement but rather to construct additional measurements that avoid the weakness.

Of course, the new measurements have their own weaknesses, but taken together, the set of measurements produces a more accurate picture. In the case of unemployment, one alternative measure of labor-market performance is the U-6 unemployment rate. The U-6 measure counts as unemployed individuals who "currently are neither working nor looking for work but indicate that they want and are available for a job and have looked for work sometime in the past 12 months."[6] It also includes as unemployed part-time workers who report a desire for full-time work. However, statements about the desire for labor-force participation unsupported by concrete action raise concerns that the individual remains jobless for reasons other than labor-market weakness. Figure 1.6 shows the U-3 and U-6 unemployment rates from 2000 to 2023.

Regarding measurement validity, we can draw two general conclusions. First, the measurement validity of a particular calculation (e.g., unemployment rate) cannot be judged separately from the theoretical idea that is motivating the claim (e.g., labor-market conditions). Second, we typically adapt to measurement validity issues by trying to construct measures that better align the theoretical idea (e.g., labor-market conditions) or by using multiple measures. While these added measures may each have their own weaknesses, the weaknesses differ across measures, and taken together, the measures allow us to develop a more accurate picture of the phenomenon of interest. There are 18 instances of the U-3 and U-6 unemployment rates moving in opposite directions in the period from January 2000 through January 2020. In 11 cases, U-3 fell, and U-6 rose, while in seven cases, U-3 rose while U-6 fell. In these cases, we must think carefully about why the two measures would change in opposite directions. For instance, U-3 may fall while U-6 rises because some unemployed workers become discouraged about their job prospects and simply stop looking.

A final important impediment to measurement validity is outcomes that follow from decisions that are nested within other decisions. This nesting effect can cause difficult-to-interpret changes in ratios intended to measure a given theoretical construct. Consider the case of abortion. Suppose we define abortion as the intentional act of ending a pregnancy and we intend to

FIGURE 1.6 ■ U.S. U-3 and U-6 Employment Rates 2000–2023



Source: U.S. Bureau of Labor Statistics.

measure the prevalence of abortion and compare the prevalence over time in the United States. Note that it is not possible to have an abortion unless you are first pregnant, and only a certain segment of population can become pregnant (female and of childbearing age). Suppose also that we are concerned about the number of abortions. While there are a series of possible measures of abortion prevalence, let us consider the following measures of abortion:

1. The total number of abortions in the United States
2. The number of abortions per 100,000 in population
3. The number of abortions per 100,000 women
4. The number of abortions per 100,000 women of childbearing age (15-44)
5. The ratio of abortions to live births

The choice (or choices) that we make regarding the appropriate measure of abortion prevalence will depend on the claim (or theory) that we wish to investigate. If we are concerned that a rise in abortion prevalence will restrict access to abortion at a fixed number of abortion providers, the first measure is the preferred measure.

If, on the other hand, we are concerned that American women are choosing abortion more (or less) frequently, the first measure is potentially misleading. The number of abortions may rise because (1) the population, or more specifically the population of women of childbearing ages, is larger, and (2) more American women take actions that lead to pregnancy. Item 2

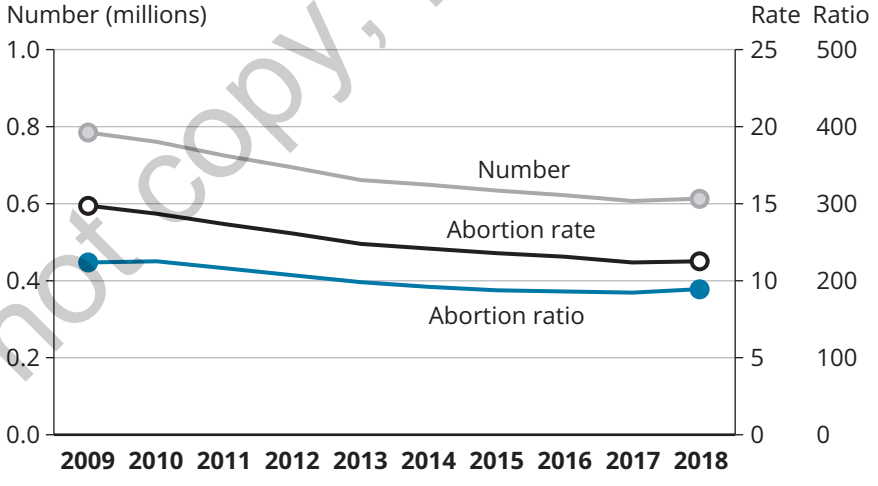
adjusts the abortion level using the total population, but this adjustment is potentially misleading because the number of men, children, and elderly may be rising (or falling) relative to the number of women in childbearing years. Thus, Item 4 is a better adjustment to the number of abortions than either Items 2 or 3.

Nevertheless, even using Item 4, the propensity to choose abortion may be affected by changes in behavior leading to pregnancy (e.g., more unprotected sex). Forming the ratio of live births to abortions (Item 5) is one method to capture behaviors leading to pregnancy. As such, we should prefer Item 5 to Items 1 through 4 in circumstances in which we wish to measure propensity to choose abortion.

Figure 1.7 reports the number (Item 1), rate (Item 4), and ratio (Item 5) for abortions by year for the United States, 2009–2018. The abortion number declines steadily through 2017 as does the number of abortions per 1,000 women 15-44 years. However, the number of abortions per 1,000 live births shows little change from 2015 onward. This suggests that decreases in live births offset changes in abortion in 2016 and 2017, and thus some of the decrease in abortion numbers in 2016 and 2017 is the result of fewer pregnancies.

A final point regarding measurement of outcomes built on nested decisions is the lag that may occur between decision 1 (e.g., pregnancy) and decision 2 (e.g., abortion) may lead to a mismatch between the theoretical concept and the empirical measure. This is not typically a problem with abortion as the data are annual. The pregnancy decision and the abortion decision will be separated by a matter of months. By contrast, the marriage and divorce decision are typically

FIGURE 1.7 ■ Number*, Rate, and Ratio*** of Abortions Performed by Year—Selected Reporting Areas, United States, 2009–2018**



Note: *Number of abortions per 1,000 women aged 15-44 years. **Number of abortions per 1,000 live births. ***Data are for 48 reporting states; excludes California, District of Columbia, Maryland, and New Hampshire.

Source: Kortsmit K, Jatlaoui TC, and Mandel MG, et al. (2020) Abortion Surveillance—United States, 2018. MMWR Surveill Summ 2020;69(No. SS-7): 1-29.

separated by years.[7] As such, measures of divorce may need to account for the lag between the marriage and divorce decisions. That is, a surge in marriage in 2010 can lead to a surge in divorce in 2018 without a change in the propensity of a randomly selected marriage to dissolve. In cases such as this, we may need to further refine the empirical measure to account for this lag.

Question for Review 1.5: Explain why we cannot judge measurement validity for a formula (e.g., Equation 1.1) based only on the equation. Why do we need additional information? What is that information?

A Note of Caution on Measurement

Gathering and presenting quantitative measures to support claims (or arguments) is valuable even in cases in which the theoretical idea may not be perfectly captured by our empirical measure. Linking theoretical claims with empirical measures increases the precision of the argument and decreases the scope for disagreement. Critics of your argument must rebut your evidence, and such a requirement tends to focus the dispute and reduce differences in interpretation. Lord Kelvin offers a particularly eloquent statement of this view:

> I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be.[8]

Nevertheless, there are pitfalls in this effort to link theoretical claims and empirical measures. Links between theoretical claims and empirical measures are often imperfect. While imperfect measures often lead to added insight, we must be aware of their imperfections and consider the impact of these imperfections as we draw conclusions. Where the link between the theoretical claim and the empirical measure is especially tenuous, we must be prepared to abandon the empirical measure.

In many cases, theoretical constructs are difficult to measure. The temptation then is to focus on the easily quantified elements of the general concept and ignore all other information. This tendency to focus on quantifiable aspects of a general concept to the exclusion of other aspects is often referred to as the "McNamara fallacy" after the U.S. Secretary of Defense Robert McNamara. McNamara served as Defense Secretary during the Vietnam War and was responsible for a series of key military decisions. President John F. Kennedy nominated McNamara to the post following a successful stint running Ford Motor Company where he returned the company to profitability using a management strategy that relied on key statistical metrics.

Building on his experience at Ford, McNamara contended that progress in the Vietnam War (the theoretical concept) could be measured by the body count (the empirical measure). That is, if U.S. casualties were below Viet Cong casualties, the U.S. would eventually

prevail. As we know, the U.S. did not prevail in the war and many analysts, often with the benefit of hindsight, blamed McNamara. Yankelovich offers a particularly scathing critique of McNamara's strategy:

> The first step is to measure whatever can be easily measured. This is OK as far as it goes. The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide.[9]

A contemporary example of the McNamara fallacy is college rankings. Perhaps the most popular and influential of all college rankings are U.S. News and World Report (USNWR) college rankings.[10] USNWR ranks institutions based on an index and contends that the index is a "tool to select and compare schools."[11] In 2025, the USNWR national university ranking calculates rank using a weighted average of graduation and retention rates (21%), graduation rates of low-income students (11%), a comparison of predicted versus actual graduation rates (10%), assessments of quality from peer institutions (20%), faculty resources (11%) (e.g., faculty compensation), earnings relative to high school graduates (5%), standardized test scores (5%) (e.g., SAT and ACT scores), financial resources per student (8%), and graduate indebtedness (5%).[12]

To the extent that prospective college students and their families view USNWR rank as a stand in for quality and ignore other information, they commit the McNamara fallacy. For instance, graduation rates receive relatively heavy weight in the index, and the relation between graduation rates and academic quality is less than clear. One method to boost graduation rates is to increase the percentage of A grades and reduce the percentage of failing grades. Of course, none of this amounts to a general argument against quantification. Rather, the lesson calls for attention to (1) the relationship between the theoretical construct and the empirical measure and (2) nonquantifiable aspects of the theoretical construct.

Linking Deduction With Induction—Measurement Reliability

In addition to empirical measures that do not align with theoretical ideas, we must also take care to identify procedures that fail to accurately measure the theoretical idea. If our procedure produces accurate measures of the theoretical idea, we claim it has high **measurement reliability**. In essence, we may fail to link empirical measures with theoretical ideas because there is a mismatch in definitions (measurement validity) or because the procedure used to create the empirical measure is flawed (measurement reliability). As we typically cannot compare the measure produced by our preferred procedure to the "truth" (i.e., an accurate measure of the theoretical idea) to assess accuracy, we must resort to other means to determine measurement validity.

One commonly employed tactic is repeated testing. If we are reasonably certain that the theoretical idea shows little change over time, then we can repeat the procedure for the empirical measure and consider whether the measures are consistent to assess measurement reliability. Consider IQ testing as a measure of the theoretical idea of intelligence. Assuming intelligence

changes little over the course of a week, a test on the same individual that produces an IQ score of 100 on Monday, 130 on Tuesday, and 88 on Wednesday indicates that the test has a measurement reliability problem.

Another important source of measurement unreliability follows from incomplete recall or self-presentation effects on surveys. Suppose that we wish to measure the prevalence of arthritis in the population.[13] One method to measure arthritis prevalence is to ask respondents to report on diagnoses they have received from doctors. For instance, we might provide respondents with a list of medical conditions that includes arthritis and ask, "Which of the following conditions do you have now that a doctor has told you about?"

A competing procedure might ask a series of questions about pain and swelling of specific joints. From the responses, we can construct an alternate measure of arthritis. The second procedure captures people who are unaware that they have arthritis as well as those who avoid reporting because they wish to avoid any social stigma associated with the disease. In addition, the definition may exclude respondents who are using arthritis as a reason for not working. In fact, those who are not working are more likely to report arthritis under the first procedure but not the second.

Researchers attempting to study the prevalence of racism face similar hurdles. Survey items that ask respondents "Are you a racist?" are unlikely to elicit truthful responses. Instead, researchers typically use clever tactics to induce truthful responses and increase measurement reliability. For instance, one well-known study sent resumes from fictitious applicants in response to job listings and counted callbacks from the employers.[14] The resumes sent were identical except for the name of the fictitious applicant. In some cases, the fictitious employee had a name that most Americans associate with African Americans (e.g., Lakisha Washington or Jamal Jones), while in other cases, the fictitious employee had a name most would associate with whites (Emily Walsh or Greg Baker). Because the difference in race was implied and participant firms did not realize they were being observed as part of study, truthful responses and a reliable measure of racism or racial discrimination were more likely.

Measurement reliability is also a concern any time that we use sample data to draw an inference about a population. To draw inferences about a population from a sample, we collect quantitative measures from some randomly selected subset (sample) of a larger group (population) and use the sample as the basis to generate information about the population. In general, larger sample sizes increase measurement reliability. That is, a larger sample size permits a more precise estimate of the population value. We refer to this error that follows from the use of sample to estimate a population value as **sampling error**. Increases in sampling error decrease measurement reliability.[15] Sampling error and its implications will be the focus of the second half of this book. For now, we simply wish to introduce the issue and show that it is an important component of measurement reliability.

All polls (political and otherwise) are subject to sampling error. As sampling error rises, so does the margin of error for the poll. This margin of error measures how close the survey result will likely be to the population value. A smaller margin of error implies that the range of values around the sample proportion that is likely to include the population proportion is also smaller. Suppose that we wish to estimate the proportion of the population that believes dogs are better pets than cats. If our sample proportion is 0.6 (60% of the sample prefers dogs) and our margin of error is 0.02, then our population proportion is likely in the range (or interval) from 0.58 to 0.62.

To better understand this process, it is best to imagine that we conduct repeated polls. Each poll has the same sample size, and we draw (at random) participants from the population and calculate the proportion of dog lovers. Because each individual poll will likely not include the same members of the population, our poll results will vary. Thus, we can only speak about possible population values as a range (or an interval)—there will always be some probability that the actual population value is outside this range. This can occur simply because by random chance we happen to select all dog lovers (or dog haters) into our sample.

Fortunately, we can precisely estimate this interval using the sample data if the sample is random (i.e., each member of the population has an equal chance of selection into the sample). While there is a chance that the population value is outside the interval, our calculation allows us to choose the chance that the population value is outside the interval. In general, larger intervals increase the chance (or the probability) that the interval contains the true population value. For instance, if we want our interval to contain the true population proportion 99% of the time, we will need a larger interval than in the case where we want our interval to contain the true population proportion 95% of the time (all else constant).

Table 1.3 shows margins of error for a series of sample sizes and for confidence levels of 95% and 99%. We denote sample size using the letter *n*. The table also shows the upper and lower bounds for a confidence interval built on the assumption the sample proportion is 0.6.[16] A margin of error of plus or minus 0.025 at the 95% confidence level implies that if we conducted the same survey 100 times (drawing a new random sample each time), we would expect the result to be within 2.5 percentage-points of the true population value in 95 of those 100 instances.

From the table, we can see that increases in the sample size decrease the margin of error and increase measurement reliability. For instance, increasing the sample size (*n*) from 100 to 400 at the 95% confidence level cuts the margin of error from about 10% to 5%. Assuming our sample proportion is 0.6 implies that with a sample size of 100, we can be 95% confident that the population proportion falls between 0.5 and 0.7. If we increase the sample size to 400, we can be 95%

TABLE 1.3 ■ Margins of Error on a Proportion Estimate for a Series of Sample Sizes at 95 and 99 Percent Confidence Levels

Sample Size (n)	Margin of Error 95%	Margin of Error 99%	Lower Bound 95%	Upper Bound 95%	Lower Bound 99%	Upper Bound 99%
50	0.139	0.182	0.461	0.739	0.418	0.782
64	0.123	0.161	0.478	0.723	0.439	0.761
100	0.098	0.129	0.502	0.698	0.471	0.729
400	0.049	0.065	0.551	0.649	0.536	0.665
1000	0.031	0.041	0.569	0.631	0.559	0.641
1600	0.025	0.032	0.576	0.625	0.568	0.632
3200	0.017	0.023	0.583	0.617	0.577	0.623
6400	0.012	0.016	0.588	0.612	0.584	0.616

confident that the population proportion falls between 0.55 and 0.65 (with a sample proportion of 0.6)—a smaller interval and a more precise estimate.

A margin of error of plus or minus 0.025 at the 99% confidence level implies that if we conducted the same survey 100 times (drawing a new random sample each time), we would expect the result to be within 2.5 percentage-points of the true population value in 99 of those 100 instances. To increase the chance that the interval includes the population value holding the margin of error constant, we need a larger sample size.

From the table, we can see that a sample size of nearly 3,200 would be needed to produce a margin of error of plus or minus 0.025 at the 99% confidence level (if n = 3,200, the margin of error is 0.023). By contrast, a sample size of only 1,600 is needed to produce a margin of error of plus or minus 0.025 at the 95% confidence level. If we instead hold the sample size constant and increase the confidence level to 99% from 95%, then the margin of error must increase. For instance, at a sample size of 400 and a 95% confidence level, the margin of error is 0.049. Increasing the confidence level to 99%, increases the margin of error to 0.065.

This type of measurement reliability is a very common concern because sample data are frequently used to estimate population values. Consequently, this use of sample data to estimate population values will assume a central role in the second portion of this book.

Question for Review 1.6: What is sampling error? When does it occur? Is it possible to eliminate sampling error? Explain.

EXERCISES

1. Read "The Evolving Politics of the Common Core" by Ashley Jochim and Leslie Lavery. https://www.brookings.edu/wp-content/uploads/2016/06/common_core.pdf

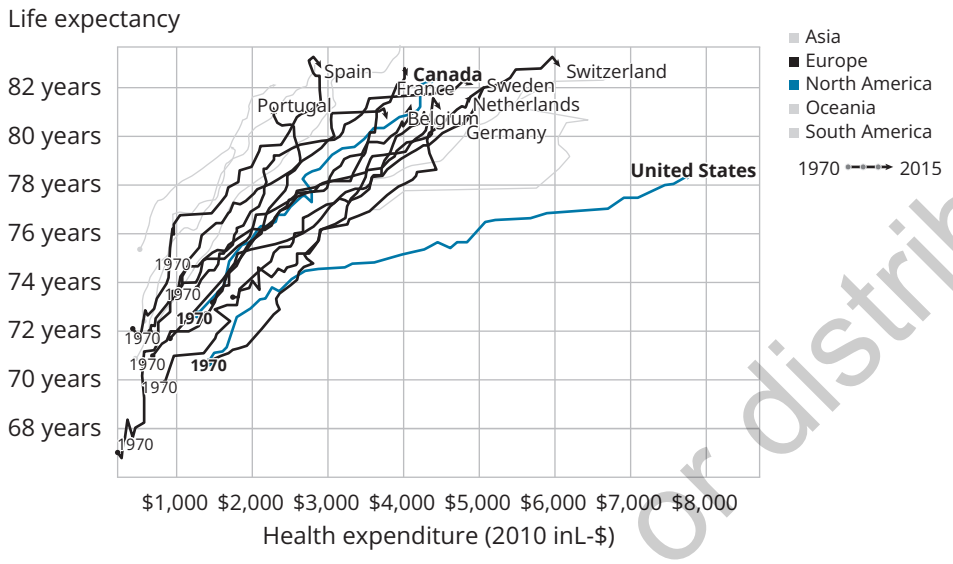
(a) In a paragraph, explain the authors claims regarding the evolving politics of the common core.

(b) How do the authors support their claims? What evidence do they collect, and what does it show? Concentrate on Figures 1 and 2.

(c) Discuss the measurement validity issues related to your answers in parts a and b. Why are they measurement validity issues? How would you address such issues to improve the argument?

2. Figure 1.Q2 (life expectancy vs. health expenditure, 1970-2015) and others like it (https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure) have been used to argue that the U.S. health care system produces some of the worst health outcomes in the developed world. Defenders of the U.S. system often raise measurement validity issues in response. What is measurement validity? Identify some measurement validity issues in the figure. Why are they measurement validity issues?

FIGURE 1.Q2 ■ Life Expectancy vs. Health Expenditure, 1970–2015



Source: https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure

3. Measuring abortion poses a particular set of problems related to the divisor. For instance, measuring abortion is more difficult than measuring race (in a population).

(a) Explain how measuring divorce is similar to measuring abortion. Be specific about the difficulties.

(b) List a set of alternate measures for divorce. Explain their weaknesses and strengths.

4. In a recent *Wall Street Journal* article ("At What Age Do You Meet Your Best Friend?," July, 15, 2019, p. A11), Claire Ansberry reports on a survey of 10,000 people commissioned by Snap Inc. The survey examined friendship in nine countries and reported on the age at which respondents met their best friend. The results appear below. Is this a good comparison? Explain.

Average age people met their best friend	21
Average age Gen Z (13-23) met their best friend	12.9
Average age millenials (24-39) met their best friend	17.9
Average age Gen X (40-54) met their best friend	23.8
Average baby boomers (55-75) met their best friend	29.8

5. In a recent analysis of crime data ("Developmental Estimates of Subnational Crime Rates Based on the National Crime Victimization Survey" https://www.ojp.gov/ncjrs/virtual-library/abstracts/developmental-estimates-subnational-crime-rates-based-national), Robert Fay and Mamadou Diallo consider the two main sources for data on crime in the

United States: the Uniform Crime Reports (UCR) and the National Crime Victimization Survey (NCVS). They note the following:

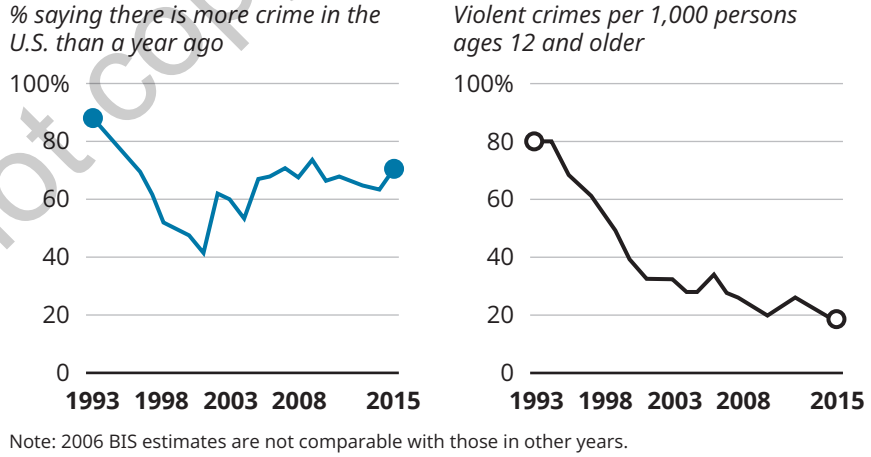
- The UCR is based on police reports, while the NCVS asks survey respondents about crimes in the prior 6 months regardless of whether they reported the crimes to police.
- The UCR covers crimes committed against all persons and businesses, while the NCVS covers persons age 12 and older and excludes crimes against the homeless, institutionalized populations, and businesses.
- The two collections differ in regard to the types of crimes included. In the UCR, violent crime includes murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. In the NCVS, violent crime includes rape and sexual assault, robbery, aggravated assault, and simple assault. Simple assault, which is not included in the UCR, is the largest component of the NCVS estimate of total violent crime.
- In both the UCR and the NCVS, property crime is divided into burglary, motor vehicle theft, and larceny (theft) other than motor vehicle theft. However, the NCVS denominator for property crime rates is the number of U.S. households, while the UCR denominator is the estimated population.

Discuss the implications for measurement validity and measurement reliability when we use these measures to assess the effectiveness of policing in the United States.

6. Consider the relationship between education and wages shown in Table 1.2. Why might the relationship be spurious? Why might the relationship show that more education causes higher wages?

7. Consider the measurement validity and/or measurement reliability issues raised by the data in Figure 1.Q7.

FIGURE 1.Q7 ■ Public Perception of Crime Rates at Odds With Reality



Note: 2006 BIS estimates are not comparable with those in other years.

Source: https://www.pewresearch.org/short-reads/2016/11/16/voters-perceptions-of-crime-continue-to-conflict-with-reality/

8. Review the following claims. Indicate whether the statements are normative or positive. Explain.

(a) Covid-19 cases in New Jersey in January of 2021 (about 5,000 per day on average) were sufficient to require a shutdown of all in-person instruction at all schools in the state.

(b) Because Covid-19 is transmitted through aerosols, transmission is reduced in outdoor venues.

(c) Overweight individuals are more likely to die from a Covid-19 infection than those who are not overweight.

(d) State governors should expend all available resources to drive the Covid-19 infection rate to zero.

(e) African-Americans are more susceptible to Covid-19 infections.

9. Researchers at the University of California are studying longevity. Toward that end, researchers visit study participants every 6 months and perform neurological and neuropsychological tests. They also "obtain information about diet, activities, medical history, medications and numerous other factors" (https://mind.uci.edu/research-studies/90plus-study/). Based on these data, the researchers concluded that "In the 1,700-person survey, people who drank about two glasses of beer or wine per day were 18 percent less likely to experience premature death than those who abstain." Consider whether alcohol causes longer life. Why might alcohol consumption not cause longer life?

10. Review the research report "Americans and Social Trust: Who, Where and Why": https://www.pewresearch.org/social-trends/2007/02/22/americans-and-social-trust-who-where-and-why/

(a) What problems or issues do you see regarding measurement validity? Explain why the problems you have identified are measurement validity issues.

(b) What problems or issues do you see regarding measurement reliability? Explain why the problems you have identified are measurement reliability issues.

11. Suppose that you are a consultant hired to assess company culture at a large financial services firm. Business culture generally refers to the set of behavioral and procedural norms that govern interactions among employees. This includes policies, procedures, ethics, values, employee behaviors, and attitudes. Your colleague proposes a survey of current employees on the following questions. Employees would respond to each item on a Likert scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree).

- Do you enjoy our company's culture?
- Do you feel connected to your coworkers?

- Do you feel your colleagues work as a team?

(a) Assess the measurement validity issues that you should consider in collecting and analyzing your data. Explain why they are measurement validity issues.

(b) Assess the measurement reliability issues that you should consider in collecting and analyzing your data. Explain why they are measurement reliability issues.

Chapter 2

Making the Right Comparison

Observations, Variable Types, Data Displays, and Data Conversions

The purpose of Chapter 2 is to explain variable types with an emphasis on informational content and converting measures.

LEARNING OBJECTIVES

1. Illustrate and define data set types (time-series, cross-sectional, and panel).
2. Illustrate and define variable types.
3. Explain the correspondence between variable types and data display types.
4. Devise and construct ratios that better capture theoretical ideas.
5. Explain the costs and benefits of converting variables.
6. Identify situations that require inflation-adjusted data and execute adjustments to nominal data using a price index.
7. Identify seasonal variation in data and execute adjustments to data by calculating an annualized percentage change.
8. Identify noise in data and execute adjustments to data using a moving average.

Data sets do not typically fall from the sky. Instead, we must create them. Because the process of creating a data set is time consuming, we typically create data sets with a particular goal in mind, usually answering a question of interest. To construct a usable data set, or a data set that produces an answer to our question, we must attend to series of concerns. What is the unit of observation? What is the time frame over which we will collect the data? How do we construct variables and comparisons that will offer answers to our question? What are the units of measure for the variable? How is the variable defined? What is the level of measurement for the variable? This chapter aims to explain the dimensions and the characteristics of data and data sets with the goal of building your capacity to construct data sets that will allow for accurate and useful comparisons.

In this chapter, we aim to carry forward a key theme from Chapter 1: Statistics is not simply a matter of completing a set of calculations using a prescribed formula. Instead, statistics, or more broadly, effective use of quantitative data, requires that we attend to issues of theory and design. Supporting a claim requires that we have an idea or an intent to assess or measure some aspect of the world and we construct measures that plausibly (though perhaps not perfectly) align with those ideas (i.e., measurement validity). Because our measures must align with our ideas, gathering data and creating new measures from those data is an intentional process. In addition, certain types of data displays require certain types of data sets and variable types. The data set and the variable types also must align with our theoretical constructs.

Toward that end, this chapter considers data set and variable types, converting and constructing new variables, and aligning data sets and variable types with a standard set of data displays. The first part of the chapter considers data sets, variable types, and converting data across variable types while the second part introduces several commonly used data displays (or figures) and connects the requirements of the data display to the definitions offered in the first part of the chapter. The final portion of the chapter considers judgments we must make in creating ratio variables and three typical data conversions.

Data Sets and Variable Types

Data sets are arrangements of data values. These sets can include one or more variables and are typically arranged in rows and columns (i.e., a matrix). A **variable** is a measured outcome that follows from a definition. The measured outcome may be numeric or nonnumeric. As the name suggests, a variable varies; it can take two or more values. Each variable can include one or more **observations**. An observation is a measurement of a variable using the variable definition for a specific object. In Table 2.1, we report a data set with 24 observations on each of four variables.

For instance, the first observation for the population variable in Table 2.1 is 282,162,411.

This population variable measures humans living (not deceased) within the geographic boundaries of the United States. In general, we can describe data sets by noting the number of observations and the number of variables. Each of the observations in Table 2.1 is associated with a year. For reasons that we explain later in the text, we typically prefer data sets that include more observations than we see in Table 2.1. We use shorter data sets in this chapter to save space and reduce confusion.

Because the data in Table 2.1 include repeated observations for the same object of analysis (the United States), we refer to the data as **time-series data**. The first variable, year, is ordered from earliest to latest year. The other variables (Population, Murders, and Murder Rate) also include an observation for each of the 20 years in the data set. The first, second, and third variables (Year, Population, and Murders) are **count variables**. A count variable is **discrete** as it can only take on certain values, in this case, integer values. We will not record half a murder or half a person. For fractions of a year, it is better to think of alternative variables like month, week, day, or so on.

These count variables may be measured at a point in time or over an interval of time. For instance, the population count variable is measured as of July 1 of each year. We refer to variables measured at a point in time as **stock variables**. By contrast, the murder count variable

TABLE 2.1 ■ U.S. Murder and Nonnegligent Manslaughter for the United States 2000–2023

year	murders	population	murder rate
2000	13,232	282,162,411	4.69
2001	14,087	284,968,955	4.94
2002	14,279	287,625,193	4.96
2003	14,457	290,107,933	4.98
2004	14,248	292,805,298	4.87
2005	14,990	295,516,599	5.07
2006	15,103	298,379,912	5.06
2007	14,928	301,231,207	4.96
2008	14,305	304,093,966	4.70
2009	13,764	306,771,529	4.49
2010	13,188	309,327,143	4.26
2011	12,810	311,583,481	4.11
2012	12,992	313,877,662	4.14
2013	12,412	316,059,947	3.93
2014	12,281	318,386,329	3.86
2015	13,783	320,738,994	4.30
2016	15,320	323,071,755	4.74
2017	15,304	325,122,128	4.71
2018	14,617	326,838,199	4.47
2019	14,712	328,329,953	4.48
2020	19,041	331,526,933	5.74
2021	16,727	332,048,977	5.04
2022	19,971	333,271,411	5.99
2023	17,722	334,914,895	5.29

Murder count includes nonnegligent manslaughter incidents.

Murder rate is per 100,000 in population. (murder rate = (murders/population) × 100,000). Murders are collected by the Federal Bureau of Investigation and compiled in the Uniform Crime Reports.

murders: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/shr population: https://fred.stlouisfed.org/series/POPTOTUSA647NWDB

is measured over an interval of time, in this case a year. We refer to such variables as **flow variables**. The count for the flow variable begins on January 1 and ends on December 31.

Following December 31, the count is reset to zero and process starts again. In many cases, the variable definition will allow us to infer whether the variable is a stock or flow. In cases where it is not clear, we should specify the point in time at which the count is recorded or the interval of time over which the count is recorded. If we opted to use a two-year period over which to calculate the murder count, the murder count would rise, but the underlying reality regarding murder would be unchanged.

The final variable in Table 2.1, Murder Rate, is a **ratio variable**. We create this ratio by taking the Column 3 variable (Murders) and dividing by the Column 2 variable (Population) and then multiplying by 100,000 (i.e., [Murders/Population] * 100,000). In general, ratio variables are continuous rather than discrete. While discrete data can only take certain values (e.g., 1, 2, or 3), **continuous variables** can take any value (e.g., 2.8916). In most cases, we must make a judgement on how many decimal places to report for a ratio variable. Typically, we should report enough decimal places to highlight the comparisons we wish to make. Too many decimal places may make it difficult to see changes across the observations. In most cases, three significant digits are sufficient to highlight variation. In Table 2.1, we have made the judgment that two decimal places are sufficient for the Murder Rate variable.

We typically construct ratio variables in this way to make better comparisons. In this case, we want to better compare murders in the United States over time. For instance, the first observation in the data set for the murder variable indicates that 13,232 murders occurred in the year 2000 in the United States. The last observation in the data set for the murder variable indicates that 17,722 murders occurred in the year 2023 in the United States. So, we can conclude that the number of murders rose.

However, it is not clear that Americans were more likely to be murdered in 2019 than 2000. The reason for this is that the population of the United States grew by more than 52 million from 2000 to 2023. Similarly, far more murders occur in California than Louisiana, but the population of California is far greater than the population of Louisiana. If we aim to gauge the relative differences in danger, rather than assess the number of criminal prosecutors we should hire, a better comparison of murders between California and Louisiana would adjust for this difference in population size.

One obvious method to adjust for this difference is to simply divide the murders by the population and then multiply by 100,000 as we see in the final column of Table 2.1. Of course, we can derive the same basic result by simply dividing murders by population and finding a value of 0.00005. But such figures are hard to read and compare, and as a result, we often rescale the variable to facilitate comparisons. I hope that you will agree that 5.0 is easier to read than 0.00005. It is difficult to see that 0.00049 is nearly 10 times larger than 0.00005! Thus, we often use judgment to convert raw counts into rates and then rescale the rates in the interest of facilitating comparisons.

The comparisons we wish to make determine the adjustments we make and the ratios we form. Toward that end, we can form ratios of ratios. For instance, we may decide to examine the relative size of the murder rate changes across years. Is the murder rate rising (or falling) quickly or slowly? One method to do this is to calculate the year-over-year percentage change in the murder rate. We can express this idea as: ([murder rate year t +1 – murder rate year t]/ murder rate year t) × 100 where t indicates time (or year).

Question for Review 2.1: Distinguish between an observation and a variable. Can a variable have multiple observations? Can an observation have multiple variables? Explain.

In Table 2.2, we report a data set with 21 observations on each of seven variables. Each of the observations is associated with a U.S. city at a given point in time—in this case, the year 2020.

TABLE 2.2 ■ Murder Rates (per 100,000) and Temperature for Selected U.S. Cities in 2020

City	State	Population	Murders	Murder Rate	Texas Dummy	Murder Rate Rank	Temp
Detroit	Michigan	659,616	322	48.82	0	1	81
Memphis	Tennessee	650,937	289	44.40	0	2	86
Milwaukee	Wisconsin	589,105	177	30.05	0	3	70
Dallas	Texas	1,363,028	236	17.31	1	4	97
Houston	Texas	2,346,155	398	16.96	1	5	91
Nashville	Tennessee	688,013	106	15.41	0	6	84
Columbus	Ohio	911,383	132	14.48	0	7	79
Albuquerque	New Mexico	562,065	70	12.45	0	8	84
Fort Worth	Texas	929,509	110	11.83	1	9	95
Phoenix	Arizona	1,708,960	188	11.00	0	10	106
Oklahoma City	Oklahoma	663,661	62	9.34	0	11	88
Los Angeles	California	4,000,587	351	8.77	0	12	81
Sacramento	California	519,050	42	8.09	0	13	86
Portland	Oregon	662,941	53	7.99	0	14	75
Seattle	Washington	771,517	51	6.61	0	15	73
San Francisco	California	881,514	48	5.45	0	16	75
Austin	Texas	1,000,276	45	4.50	1	17	93
Mesa	Arizona	527,361	21	3.98	0	18	104
El Paso	Texas	685,288	27	3.94	1	19	86
San Jose	California	1,029,542	40	3.89	0	20	79
San Diego	California	1,437,608	55	3.83	0	21	72

Murder count includes nonnegligent manslaughter incidents.

Murder rate is per 100,000 in population. (murder rate = (murders/population) × 100,000) Temperature is the daily high temperature for July 1, 2021: https://www.timeanddate.com/. Data are collected by the Federal Bureau of Investigation and compiled in the Uniform Crime Reports. Population is measured as of July 1, 2020. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home.

Because the data include a single observation for each of a series of objects of analysis (U.S. cities), we refer to the data as **cross-sectional data**. Thus, in contrast to Table 2.1, Table 2.2 compares cities at a point in time. Just as in Table 2.1, it is important that we (or those who compile the counts) use consistent definitions for each of the observations on a particular variable. For instance, the definition of "murder" should be the same in Detroit as in San Diego, otherwise the comparison is flawed. Recording written definitions of each of the variables in your data set at the time you compile your data is an excellent way to avoid confusion regarding definitions.

Another important element of data definition is ensuring that we have a variable (or set of variables) that allows us to uniquely identify each of the observations in the data. In Table 2.2, the first observation for the "City" variable is "Detroit." This tells us that each cell in the first row is an observation associated with Detroit. We refer to City as a **nominal variable** as it is not a number but rather a name. "State"—also a nominal variable—allows us to precisely define each observation. It is important to know that the fourth observation ("Dallas") is Dallas, Texas and not Dallas, Pennsylvania. (Yes, there is a Dallas, Pennsylvania.)

While we can conduct mathematical operations on the count (e.g., population) and ratio (murder rate) variables described previously, we cannot conduct a mathematical operation on a variable like "City" or "State." That is, we cannot take the mathematical average (or mean) of the "City" variable. Nevertheless, we can create one or more **categorical** or **dummy** variables to capture the information in a nominal variable. A categorical variable assigns observations into mutually exclusive categories based on a characteristic of the observation. A dummy variable is a type of categorical variable that classifies information into two mutually exclusive categories and assigns a value of 0 to one category and a value of 1 to the other.

For example, this data set includes observations for five Texas cities. We can, therefore, create a dummy variable that indicates whether the observation is a city in Texas (or any other piece of information that might be extracted from the nominal variable). We can even create a set of dummy variables that classify all observations based on the "State" variable. In such a case, we would show a column for each state that appears in the data set and enter a 1 for all observations for that state. A dummy for California (i.e., California dummy) would show a 1 for Los Angeles, Sacramento, San Francisco, San Jose, and San Diego and a 0 otherwise.

Creating dummy variables from nominal data has two important advantages. First, it allows us to group other variables to facilitate comparisons. We can use the Texas dummy to compare the murder rate for Texas cities with the murder rate for the other cities in the data set. Second, we can conduct mathematical operations directly on the dummy variable. For example, the average (or mean) of the Texas dummy is 0.238. This tells us that 23.8% of the observations in the data are for Texas cities. We can also create dummy variables from other variable types. For instance, we might divide the cities based on murder rate (a ratio variable) and assign a 1 to cities with a murder rate above the average and 0 to cities with a murder rate below the average.

Other conversions are possible. We can also convert a ratio variable (like Murder Rate) into an **ordinal variable**. Ordinal variables record data in categories that are numerically ordered and mutually exclusive—an observation must fit into one and only one category. Thus, an ordinal variable is a variable for which the ordering (or rank) matters but not the difference between the variables. For these data, we order the observations (or cities) based on the murder rate and assign a unique rank to each observation depending on the order. We call this new variable "Murder Rate

Rank." Consistent with the previous definition, the order matters but not the difference between the variables. For instance, the difference between the murder rate for Detroit and Memphis (the cities with the number 1 and 2 highest murder rates) is about 4.4 murders per 100,000 in population (48.82 – 44.40 = 4.42). By contrast, the difference between the murder rate for Memphis and Milwaukee (the cities with the number 2 and 3 highest murder rates) is about 14.4 murders per 100,000 in population (44.40 – 30.05 = 14.35). Thus, we know from Murder Rate Rank that Detroit has a higher murder rate than Memphis but not how much higher.

The average for an ordinal variable constructed in this way is not meaningful. The average will always be $(n + 1)/2$ where n is the number of observations. In this case, $n = 21$, and the average is 11. Nevertheless, the average is meaningful for some other types of ordinal variables. Suppose that we are interested in perceptions of crime in addition to crime counts and crime rates. We might ask a randomly selected individual to respond to the claim "This city is generally safe for tourists and residents" for each of the 21 cities in our data. We can include as possible responses to this survey item: strongly disagree, disagree, neutral, agree, and strongly agree. Suppose also that we assign the value of 1 to "strongly disagree," 2 to "disagree," 3 to "neutral," 4 to "agree," and 5 to "strongly agree." (We refer to this set of response options as a **Likert scale**.)

Because the response options are ordered based on degree of agreement and the difference in the level of agreement between responses is unknown, the response to the item is an ordinal variable. However, one person's perceptions of crime may not be a terribly persuasive measure of "perceptions of crime." Indeed, the sole survey respondent may not have even visited all the cities in the data set. The crime perceptions of a large number of randomly selected residents for each of our 21 cities will likely yield a more persuasive measure of crime perceptions. Suppose then that we gather data from a large-scale survey of randomly selected residents conducted in each of the 21 cities in Table 2.2. Each of the individual responses to the question is an ordinal variable. In this case, we can calculate an average response across survey respondents for each city (i.e., a ratio variable) that is meaningful. If the average response to the question is 3.3 in Detroit and 4.1 in Memphis, we can conclude that we have some evidence that Detroit residents perceive more crime than Memphis residents. Thus, we have created a ratio variable from an ordinal variable. The key difference between the Murder Rate Rank described previously and this survey response is that scaling occurs within the observations for the survey response rather than across the observations as in Murder Rate Rank.

It is also important to note that anytime we convert a ratio (like Murder Rate) or a count variable (like Murders) to an ordinal variable (like Murder Rate Rank) or a categorical variable, we lose information. Ratios and counts have more information than ordinal variables as the difference in each of the observations conveys information about the degree of difference. In our previous example, we know from Murder Rate Rank that Detroit has a higher murder rate than Memphis but not how much higher. This information is clear from the murder rate. Ordinal variables also have more information than categorical variables. We can use Murder Rate Rank to create a categorical variable that identifies high murder rate rank cities with a 1 (and assigns 0 to low rank cities). In so doing, we lose information on the ordering of the high- and low-ranked cities.

Moving from categorical to ordinal to ratio also causes a loss of information, but the mechanics of the loss are different. Returning to our crime perceptions example, we created a

ratio from an ordinal variable, but to do so we had to aggregate more than one response to the survey (ordinal data) and divide through by the number of responses. In so doing, we create a single observation on a ratio variable from a larger number of observations on an ordinal variable. When we take the average of the responses, we lose information on, among other things, the distribution of the individual responses to the survey. If we have only the average survey response (ratio variable), we cannot recover the set of responses to the survey (ordinal variable).

In a similar way, we can create a ratio variable from a categorical variable. Consider our example in Table 2.2. In that table, we created a categorical variable to indicate Texas cities. Calculating the average of the categorical variable gives us a ratio variable (count of Texas cities in the data/count of total cities in the data = 0.238); we create a single observation on a ratio variable from multiple observations on a categorical variable. However, we lose information on which cities are in Texas.

This suggests that we would never want to undertake conversions that decrease informational content. However, this is not the case. Deleting or suppressing certain types of information can aid analysis and facilitate creation of tables and figures. We shall discuss decisions to convert variables and suppress certain information later in the text.

The final variable in Table 2.2, Temp, is an **interval variable**. Temp records the daily high for each city in the data set on July 1, 2021. All interval variables are measured on a scale where each point is placed at an equal distance (or interval), but zero has no meaning. For instance, a temperature reading of 0 does not mean zero heat. Because each point is placed at an equal distance, differences between values are meaningful (like a count or ratio variable and unlike an ordinal variable).

Question for Review 2.2: Explain the process for converting nominal variables into dummy variables. Do we lose information in such a conversion?

Table 2.3 offers a summary of these variable types. Keeping these definitions in mind will be helpful as we consider the relation between variable types and data displays later in this chapter.

To this point, we have considered data sets (or data tables) that include either (1) repeated observations for the same object of analysis (i.e., time-series data) or (2) a single observation for each of a series of objects of analysis (i.e., cross-sectional data). However, researchers often compile data that include both time-series data and cross-sectional data in a single data set. We refer to data sets that have both time-series and cross-sections as **panel data**. Panel data sets offer advantages over data sets that include only a time-series or a cross-section as they permit simultaneous comparison within the time-series for more than one time series and allow for comparisons between time-series. Table 2.4 shows a simple panel data set. As you can see, Table 2.4 shows a randomly selected subset of six cities (the cross section) from the data set shown in Table 2.2. Consistent with the definition offered in Table 2.5, Table 2.4 reports observations for years 2018 to 2020 (the time series) for each of these six cities.

There are several ways that we might arrange such data for analysis. However, we can facilitate most types of analysis by creating an observation for city/year outcome. That is, each observation is a unique combination of the cross-section variable (City) and the time-series variable (Year). If we adopt this city/year format, we create 18 observations (6 × 3 = 18). A competing

TABLE 2.3 ■ Summary of Variable Types

Variable Type	Characteristics
Categorical	Represents an attribute using a limited number of classes or numbers. We can convert a categorical variable into a set of dummy variables.
Count	Assigns value based on the sum of cases that meet a definition. The variable is discrete as it can take on only certain values—usually positive integers. It may be a stock or flow measure. Mean and standard deviation are meaningful.
Dummy	Represents an attribute using zeroes and ones where 1 indicates the presence of the attribute, and 0 indicates absence. The mean value of the variable shows the proportion of observations with the attribute. We can convert a set of dummy variables into a nominal or categorical variable.
Interval	Assigns value based on a continuous scale where each point is placed at an equal distance (or interval), but zero has no meaning. Zero is an arbitrarily chosen point, rather than the absence of value, and each point on the scale is equally distanced. The difference between values on the scale is meaningful as is the mean and standard deviation.
Nominal	Represents an attribute using a name rather than a number. Nominal data can be grouped. We can represent this type of variable in a frequency plot (i.e., histogram). Calculating the mean or standard deviation is not possible.
Ordinal	Assigns value based on rank or other ordering. It requires sequence but does not offer information on the absolute or numerical difference in the measure that serves as the basis for the ranks. Mean and standard deviation are meaningful only when the ordering is external to the data set (e.g., Likert scale).
Ratio	Assigns value based on an expression that includes a numerator and denominator. Frequently, the numerator and denominator are counts, but they can be ratios. Values are continuous. The difference between values on the scale is meaningful as is the mean and standard deviation.

TABLE 2.4 ■ Murder Rates (per 100,000) for Selected U.S. Cities, 2018–2020

City	State	City Number	Year	Murders	Murder Rate
Milwaukee	Wisconsin	1	2018	100	16.79
Milwaukee	Wisconsin	1	2019	97	16.44
Milwaukee	Wisconsin	1	2020	177	30.05
Nashville	Tennessee	2	2018	91	13.26
Nashville	Tennessee	2	2019	85	12.67
Nashville	Tennessee	2	2020	106	15.41
Fort Worth	Texas	3	2018	58	6.49
Fort Worth	Texas	3	2019	69	7.59

(*Continued*)

TABLE 2.4 ■ Murder Rates (per 100,000) for Selected U.S. Cities, 2018–2020 *(Continued)*

City	State	City Number	Year	Murders	Murder Rate
Fort Worth	Texas	3	2020	110	11.83
Los Angeles	California	4	2018	258	6.40
Los Angeles	California	4	2019	258	6.48
Los Angeles	California	4	2020	351	8.77
Seattle	Washington	5	2018	32	4.31
Seattle	Washington	5	2019	28	3.72
Seattle	Washington	5	2020	51	6.61
Mesa	Arizona	6	2018	17	3.37
Mesa	Arizona	6	2019	9	1.74
Mesa	Arizona	6	2020	21	3.98

Murder count includes nonnegligent manslaughter incidents.

Murder rate is per 100,000 in population. (murder rate = (murders/population) × 100,000) Data are collected by the Federal Bureau of Investigation and compiled in the Uniform Crime Reports. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home.

TABLE 2.5 ■ Data Set Types

Data Set Type	Characteristics
Time series	A data set that is composed of repeated observations on the same object of analysis at different points in time (e.g., individual, state, or country)
Cross-Sectional	A data set that is composed of a single observation for each of a series of objects of analysis at a certain point in time or over the same timeframe (e.g., U.S. cities)
Panel	A data set that is composed of repeated observations at different points in time on multiple objects of analysis (i.e., includes a time series on two or more cross-sections).

possibility is to create only 6 observations (one for each city) and show 2018 murders, 2019 murders, and 2020 murders as separate columns (likewise for murder rate). Such an arrangement would decrease the number of observations but increase the number of variables.

Because we have adopted a city/year arrangement, we must include variables in our data that uniquely identify each of the 18 observations in the data set. The nominal variables "City" and "State" identify the cross-section, while "Year" identifies the time series. We can facilitate some types of analysis by substituting a unique number ("City Number") for the nominal variables that identify the cross-section.

Variable Types and Data Displays

Understanding these definitions or distinctions related to data sets and the variables included in data sets has important implications. We make these distinctions to facilitate analysis and presentation of data. In general, the question of interest suggests the type of data gathered and the type of data imply certain types of data displays. These data displays include both figures and tables. A table arranges data in rows and columns. We refer to intersection points between rows and columns in a table as cells. Well-designed tables aim to facilitate comparisons among cells in a specific column (or row). A figure is any type of illustration (other than a table) designed to facilitate comparisons. In this section of the chapter, we will consider a series of commonly used figures (i.e., line, scatter, and bar) and their data requirements. The main goal is to motivate the distinctions we have drawn among data types in the first portion of the chapter rather than offer comprehensive coverage of data displays.

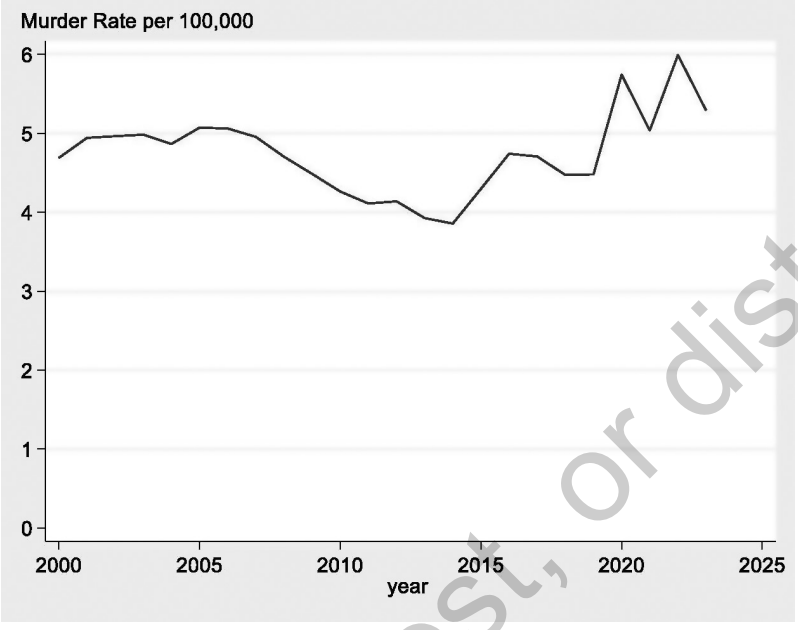
Just as we can speak of grammar and style in the context of an essay or article, figures (and tables) have a grammar. Figures have a grammar because, like an essay, a figure serves as a tool of communication and analysis. That is, we can use figures and tables to convey information to others as well as derive conclusions from data sets. By compressing or arranging the data in certain ways, relationships among variables can become more noticeable. We will consider some elements of this grammar here. First and foremost, figures should aim to maximize the information-to-ink ratio; figures should aim to clearly convey or illuminate the key comparisons the author wishes to make with the data using as little ink or complexity as possible. To the extent possible, the main comparison and its meaning should be discernable simply by viewing the figure without reading any of the associated text. To achieve this goal, figures should include clear data definitions in footnotes and detailed titles that highlight the comparison the table author wishes to make.

For instance, the data set reported in table 2.1 answers the question "What is the trend in murders over the past 20 years in the United States?" The general trend in the U.S. murder rate was stable from 2000 to 2006. After 2006, the murder rate fell until 2014. The murder rate rose in 2015–2016 and again in 2020. Extracting this information from Table 2.1 is difficult. To present these data in a form that enables readers to see the trend more easily, we create Figure 2.1.

Figure 2.1 is a line graph. A line graph requires that we have one variable (typically displayed on the x-axis) that is ordered sequentially (usually lowest to highest) and recorded on a common unit of measure. Each of these x-values must have one (and only one) associated y-value. Most often, the x-value is time (as it is in Figure 2.1). In the case of Figure 2.1, the common unit of measure for the x-value that is ordered sequentially is a year. The y-value, by contrast, can be an ordinal, count, interval, or ratio variable. These y-values either increase or decrease with changes in the x-value. In Figure 2.1, the y-value is the murder rate. The title ("U.S. Murder and Nonnegligent Manslaughter Rates for the United States, 2000–2023") highlights that the figure compares U.S. murder rates over a specific period. Though line graphs typically use time as the x-variable, we could use Murder Rate Rank (ordinal) as the x-variable and population (count variable) as the y-variable to check for a relation between the murder rate and city size.

Figure 2.2 shows a bar graph that summarizes some of the data reported in Table 2.2. In particular, Figure 2.2 shows a relation between a nominal variable (City) reported in the y-axis and a ratio variable reported on the x-axis (Murder Rate). We can reverse the axes here and report the nominal variable on the x-axis and the ratio variable on the y-axis without loss of

FIGURE 2.1 ■ U.S. Murder and Nonnegligent Manslaughter Rates for the United States, 2000–2023



Notes: Murder count includes non-negligent manslaughter incidents. Murder rate is per 100,000 in population. (murder rate = (murders/population) × 100,000).

Source: Uniform Crime Reports. Murders: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/shr Population: https://fred.stlouisfed.org/series/POPTOTUSA647NWDB

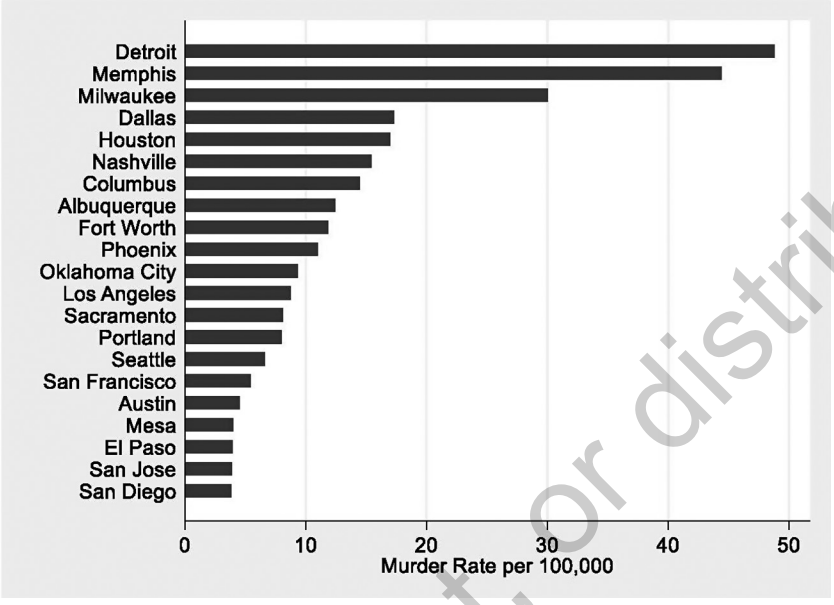
clarity. However, bar graphs generally require that at least one of the variables has a limited number of outcomes. As such, these graphs generally require that one of the variables is either a nominal, categorical, or ordinal variable.

While the nominal variable might be displayed in alphabetical order, displaying the nominal variable (City) based on the level of the ratio variable (Murder Rate) facilitates comparisons among the cities. From Figure 2.2, we can see the scale of differences between high-murder rate cities like Detroit and Memphis and low-murder rate cities like San Jose and San Diego.

Question for Review 2.3: Explain why the data in Table 2.2 are well suited to a bar graph and not a line graph.

Figure 2.3 summarizes the panel data reported in Table 2.4 using a line graph like Figure 2.1. As in Figure 2.1, the data include one variable that is ordered sequentially (usually lowest to highest) and recorded on a common unit of measure. Once again, this variable is time measured in years. Rather than a single y-value associated with time, we have six separate y-values. Each of the y-values is reported on the same time scale and is a ratio variable. From Figure 2.3 we see that in each of the six cities, the murder rate increased in 2020 relative to both 2019 and 2018.

FIGURE 2.2 ■ Murder Rates by U.S. City for 2020



Notes: Murder count includes non-negligent manslaughter incidents. Murder rate is per 100,000 in population. (murder rate = (murders/population) × 100,000).

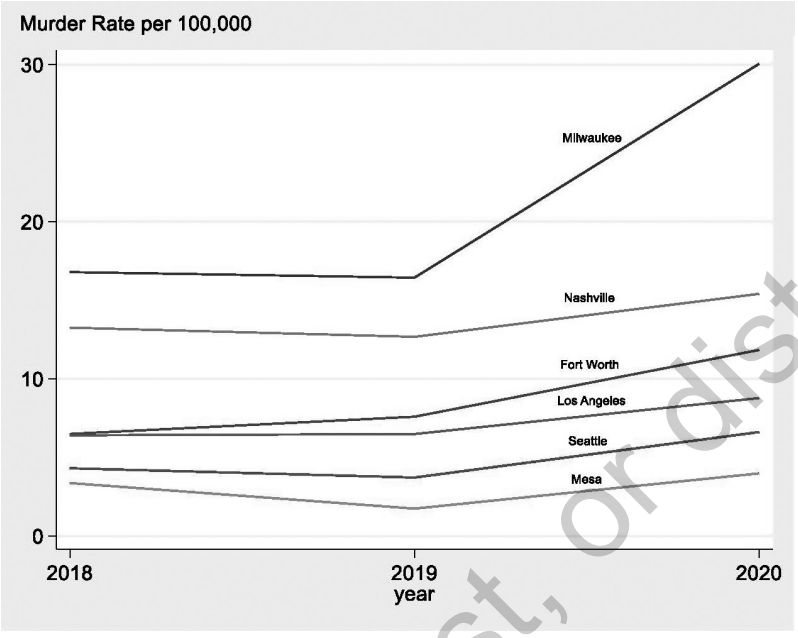
Source: Uniform Crime Reports. Population is measured as of July 1, 2020. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home

To consider additional types of data displays and their correspondence to data types, we introduce Table 2.6. Table 2.6 reports data on changes in a series of crime categories from 2019 to 2020 for the 21 cities reported in Table 2.2. Each of these variables is a ratio, or more precisely, a ratio of a ratio. We first calculate the crime rate per 100,000 for 2019 and the crime rate per 100,000 for 2020 in each of our crime categories (murder, rape, robbery, etc.) by dividing the crime count by the population and then multiplying the ratio by 100,000. Second, we calculate the percentage change in the rate by subtracting the 2019 rate from the 2020 rate, dividing by the 2019 rate, and then multiplying the ratio by 100. The first ratio corrects for population differences and the second ratio corrects for differences in the 2019 crime rate.

Note that we do not have data for the 2019–2020 burglary rate change for Dallas and Albuquerque. In such cases, we must leave the entry blank. The common mistake here is to insert a zero for a missing observation. Inserting zero leads to incorrect analyses because it suggests (in this case) that there was no change from 2019 to 2020 in the burglary rate for Dallas and Albuquerque. This is very different from not knowing the change. The correct mean value for the 2019–2020 burglary rate change is –3.36. Inserting zero for the missing observation will incorrectly reduce the size of the decrease in the average burglary rate change.

These calculations also follow logically from the evidence derived from our earlier figures. Given our results in Figure 2.3 that show a significant increase in murder rates across a series

FIGURE 2.3 ■ Murder Rates for Selected U.S. Cities, 2018–2020



Notes: Murder count includes nonnegligent manslaughter incidents. Murder rate is per 100,000 in population. (murder rate = (murders/population) × 100,000).

Source: Uniform Crime Reports. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home

of cities in 2020, we may consider whether other crime categories showed similar increases for 2020 and whether cities that showed larger increases in their murder rate relative to other cities will also show larger increases in other crime. To consider the relationship between two ratio variables, we arrange the data in a scatter diagram. A scatter diagram requires paired data; each observation has one associated observation. We can think of the pair as a x-y coordinate.

Unlike a line graph, the x-variable does not need to be ordered sequentially. Consequently, the data requirements for a scatter diagram are easier to meet than the requirements for a line graph. While the general goal of a line graph is to detect trends in the y-variable as the x-variable increases, the scatter diagram is typically deployed to assess the degree (or strength) of association between two variables. By strength of association, we aim to measure whether a high value of the x-variable is typically paired with either a low or high value of the y-variable. If a high value of the x-variable is typically paired with a high value of the y-variable, we describe the association as positive. If a high value of the x-variable is typically paired with a low value of the y-variable, we describe the association as negative. If a high value of the x-variable is sometimes paired with a high value of the y-variable and sometimes a low value of the y-variable, we conclude there is no association between the variables.

TABLE 2.6 ■ Percentage Change in the Crime Rates 2019–2020 for Selected U.S. Cities

State	City	Murder Rate Change	Rape Rate Change	Robbery Rate Change	Vehicle Theft Rate Change	Aggravated Assault Rate Change	Burglary Rate Change
Arizona	Mesa	129.2	–14.42	7.94	30.8	14.29	13.5
Wisconsin	Milwaukee	82.8	–14.63	–3.68	27.38	21.56	–7.06
Oregon	Portland	80.5	–32.96	–18.75	–2.37	3.26	–11.94
Washington	Seattle	77.93	–29.77	–8.75	18.96	0.38	27.93
Texas	Fort Worth	56	–14.62	–13.5	1.94	37.1	–15.15
Ohio	Columbus	51.33	–13.29	–2.6	–14.78	24.55	–6.35
Tennessee	Memphis	49	–18.02	–12.39	9.43	34.34	–25.69
Arizona	Phoenix	41.16	–7.77	0.86	3.27	22.19	–23.08
Texas	Houston	40.57	–17.83	–5.68	10.59	29.38	–9.29
California	Los Angeles	35.33	–13.25	–17.42	34.62	7.1	–0.78
Texas	Austin	33.45	–21.1	6.9	30.14	22.9	9.26
California	San Jose	24.06	–16.28	–12.17	14.46	1.89	–2.42
California	Sacramento	22.24	–2.6	–16.28	–8.55	22.34	–6.3
Tennessee	Nashville	21.59	–21.19	–16.51	5.03	11.56	8.83
California	San Francisco	20	–38.89	–21.83	40.42	–14	60.47
Texas	Dallas	16.9	–33.02	–30.04	–6.69	13	
Michigan	Detroit	16.82	–35.11	–19.76	–17.4	23.48	–35.51
California	San Diego	8.95	–14.37	–11.19	–5.15	8.1	–7.08
New Mexico	Albuquerque	–13.82	–12.67	–16.81	–9.36	4.05	
Oklahoma	Oklahoma City	–19.48	–2.54	–10.26	–2.49	1.18	–8.11
Texas	El Paso	–32.85	–23.34	–17.3	–60.44	–11.51	–25.04

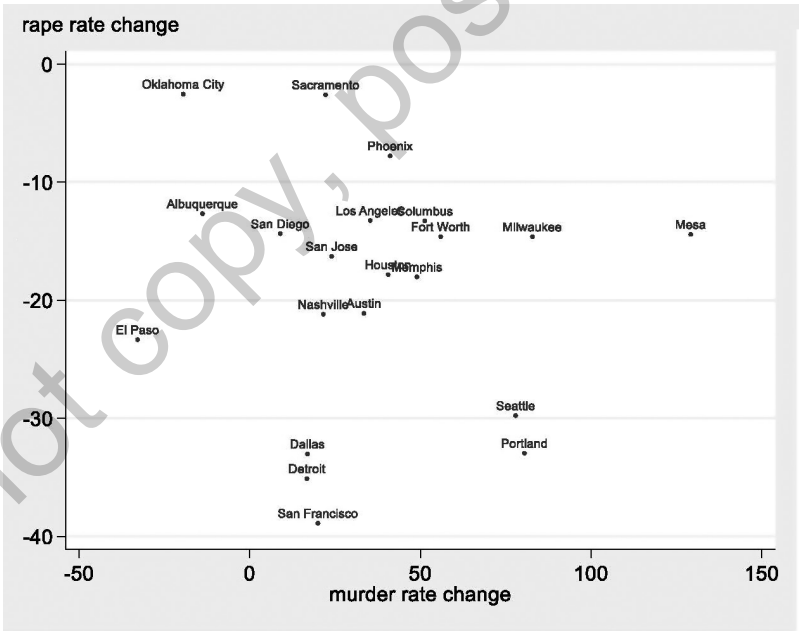
Crime rate change is [(crime rate 2020 – crime rate 2019)/ crime rate 2019] × 100.

Crime rate is per 100,000 in population. (crime rate = (crime count/population) × 100,000) Data are collected by the Federal Bureau of Investigation and compiled in the Uniform Crime Reports. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home.

Using the data in Table 2.6, we examine two associations using our city-level data: (1) the association between the change in the murder rate and the change in the rape rate, and (2) the association between the change in the murder rate and the change in the vehicle theft rate. The average murder rate increase over the period 2019–2020 for the 21 cities in our data is 35.3%, and we wish to assess whether the cities which saw a higher murder rate increase also saw a higher rape rate increase. If there is no association between the variables, the x-y coordinates will be randomly scattered across the figure. If there is a positive association between the variables, the x-y coordinates will generally follow a diagonal line that runs from the lower left of the figure to the upper right (i.e., a high x-value is typically paired with a high y-value). A negative association shows the opposite, a diagonal line that runs from the upper left of the figure to the lower right (i.e., a high x-value is typically paired with a low y-value).

In Figure 2.4, we consider the relationship between changes in murder rates and changes in rape rates. From the figure, we see that the murder rate change 2019–2020 is not associated with the change in the rape rate change as the coordinates are randomly scattered across the figure. Put differently, the change in the murder rate does not predict change in the rape rate. Figure 2.5 reports data on the relationship between changes in

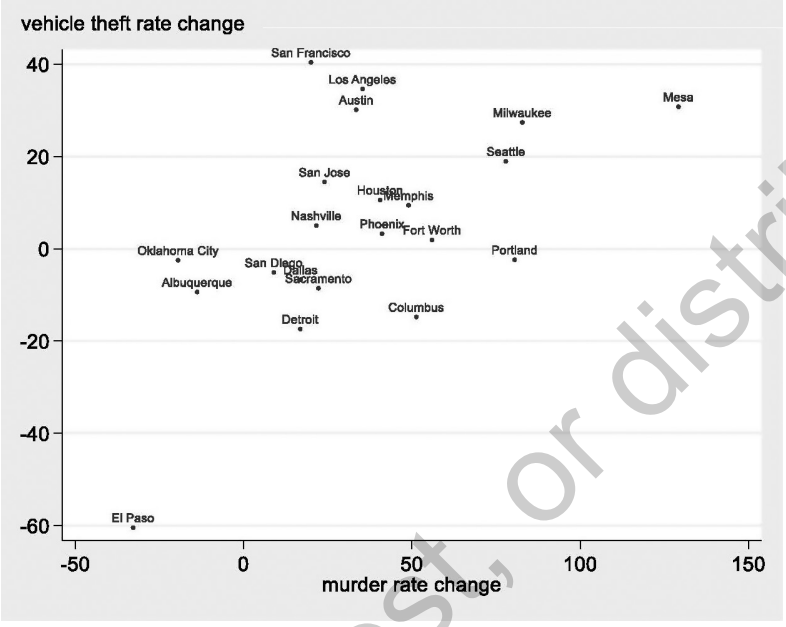
FIGURE 2.4 ■ Association Between Changes in Murder Rates (2019–2020) and Changes in Rape Rates (2019–2020)



Notes: Crime rate change is [(crime rate 2020 – crime rate 2019)/ crime rate 2019] × 100. Crime rate is per 100,000 in population. (crime rate = (crime count/population) × 100,000).

Source: Uniform Crime Reports. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home

FIGURE 2.5 ■ Association Between Changes in Murder Rates (2019–2020) and Changes in Vehicle Theft Rates (2019–2020)



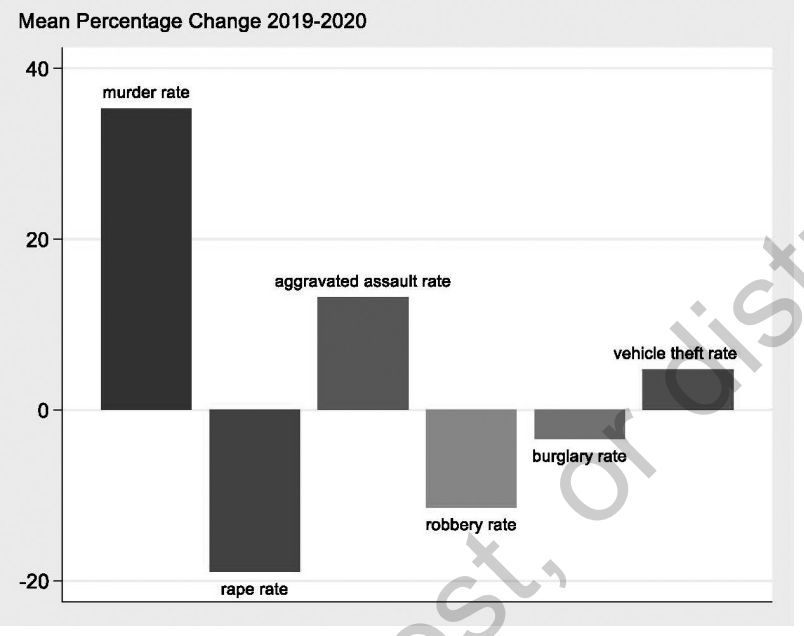
Notes: Crime rate change is [(crime rate 2020 – crime rate 2019)/ crime rate 2019] × 100. Crime rate is per 100,000 in population. (crime rate = (crime count/population) × 100,000).

Source: Uniform Crime Reports. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home

murder rates and changes in vehicle theft rates. The figure shows that the change in the murder rate 2019–2020 is positively associated with the change in the vehicle theft rate 2019–2020 as the coordinates are generally distributed in a diagonal line that runs from the lower left to the upper right. Put differently, a larger change in the murder rate predicts a larger change in the vehicle theft rate. While these measures may seem a bit imprecise, we will discuss methods that allow a more precise measurement of the strength of association later in the text.

Of course, there are other ways to report or summarize data from Table 2.6. We might also ask whether average increases in crime differ across crime categories for our sample of 21 cities. In this case, we have a limited number of nominal variables (crime categories) and one ratio variable for each crime category. In essence, we are averaging the columns in Table 2.6 and comparing the averages. Because we have a limited number of nominal variables and a ratio variable that corresponds to each nominal variable, we can construct a bar graph. Figure 2.6 shows this bar graph. From the figure we can see that the average robbery, rape, and burglary rates fall from 2019 to 2020 while the average murder, vehicle theft, and aggravated assault rates rise.

FIGURE 2.6 ■ Mean Percentage Change 2019–2020 in Crime Rates by Category for Selected U.S. Cities



Notes: Crime rate change is [(crime rate 2020 – crime rate 2019)/ crime rate 2019] × 100. Crime rate is per 100,000 in population. (crime rate = (crime count/population) × 100,000).

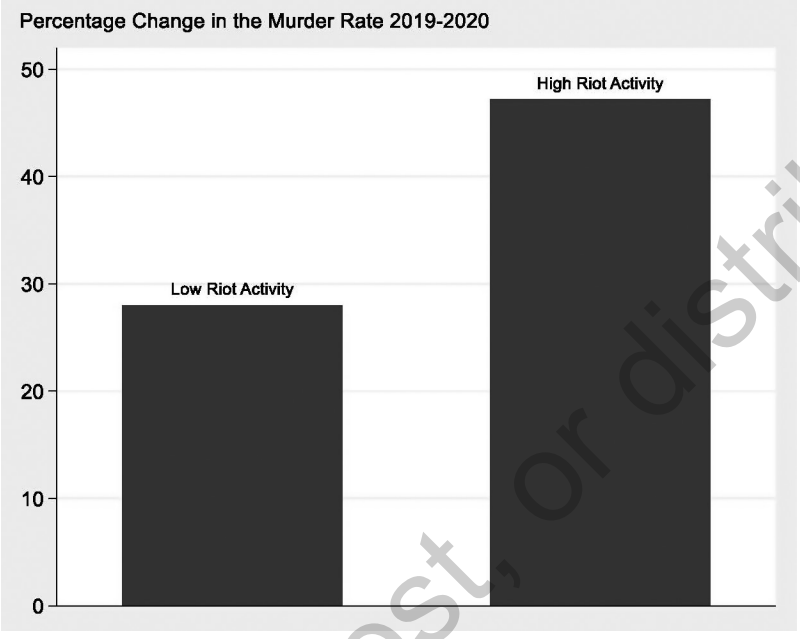
Source: Uniform Crime Reports. See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home

Another possibility is to create a categorical variable to compare outcomes within one of the crime categories. For instance, Figure 2.7 divides the cities in our 21-city sample into two groups based on the level of riot activity in 2020: one with the top half of cities and one with bottom half. (The riot definition and data are from the Armed Conflict Location and Event Data Project (https://acleddata.com/), a nonprofit organization supported by academic researchers.) Thus, we use a categorical variable (high riot activity) to compare murder rates (a ratio variable) across cities.

We see that murder rates rose more from 2019 to 2020 in high-riot activity cities (43%) than they did in low-riot activity cities (28%). Note that this is weak evidence that riots *caused* the increase in the murder rate. High-riot activity cities may differ in a number of important respects from low-riot activity cities. High riot activity may be associated with higher rates of COVID-19 infection, or the murder rate might be causing the riots. We will return to issues of cause and association in later chapters.

Question for Review 2.4: Consider the data reported in Table 2.4 and displayed in Figure 2.3. Would you make the same choices regarding the type of figure if we wished to compare 30 cities rather than 6? Explain.

FIGURE 2.7 ■ Changes in the Murder Rate (2019–2020) for U.S. Cities with High and Low Riot Activity



Sources: Uniform Crime Reports (murder). See https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home. Armed Conflict and Event Data Location Project (riots). See https://acleddata.com/.

Choice of Divisor in Creating Ratios

To this point in the chapter, we have created several ratio variables. In most of these cases, the choice of the divisor was rather straightforward. If we wish to gain insight on which U.S. cities are more dangerous (or the probability that a city resident is murdered), we can divide the annual level murders by the population in that year. Forming this ratio produces the proportion of residents that were murdered in that year and allows us to infer the annual murder risk. However, sometimes the choice of divisor is not straightforward. As we noted in the first chapter, we must attend to measurement validity. Thus, we must show care in considering the alignment between the ratio and the theoretical idea.

Consider the data on driving and roads reported in Table 2.7. The table reports state-level data for U.S. states east of the Mississippi in 2019 including 26 observations on seven count variables. The first count variable, Vehicle Miles Traveled, is the sum of all miles (measured in millions) logged by residents of the state in that year. For instance, the first entry in the table reports that residents of Alabama logged 31.5 billion miles in 2019 (31,565 million). This is far higher than the residents of Connecticut who logged only 19.1 billion. However, Alabama has more population, licensed drivers, and registered vehicles than Connecticut.

TABLE 2.7 ■ State-Level Data on Road Use for 2019

STATE	TOTAL VMT (in millions)	Total miles of Roadway	Registered Vehicles	Motor Fuel Use (thousands of gallons)	Motor Vehicle Fatalities	Licensed Drivers	Population
Alabama	31565	100685	5288208	3729856	930	4026151	4903185
Connecticut	19178	21577	2878548	1803906	249	2608061	3565287
Delaware	5702	6499	1016927	641745	132	812529	973764
Florida	115984	123104	17833720	11245020	3183	15560628	21477737
Georgia	63590	128461	8594567	6362062	1491	7261266	10617423
Illinois	59800	145967	10691947	6280426	1009	8546932	12671821
Indiana	29932	96906	6223460	4398411	809	4589405	6732219
Kentucky	25952	79954	4383223	3079142	732	3030329	4467673
Maine	6221	22819	1130056	838573	157	1046129	1344212
Maryland	38131	32373	4203994	3299060	521	4463862	6045680
Massachusetts	38649	36791	5061260	3256400	334	4950056	6892503
Michigan	54868	122181	8440065	5831913	985	7141494	9986857
Mississippi	20382	77487	2066681	2464662	643	2058036	2976149
New Hampshire	7437	16185	1363379	843669	101	1195211	1359711
New Jersey	47910	38950	6033015	4778073	559	6377413	8882190
New York	69388	113929	11389158	7312463	931	12194360	19453561
North Carolina	57240	107628	8527388	6290787	1373	7620001	10488084
Ohio	57261	123031	10901279	6802637	1153	8032792	11689100
Pennsylvania	57773	120714	10800315	6544855	1059	8987676	12801989
Rhode Island	5245	6004	868942	459836	57	761046	1059361
South Carolina	31139	79234	4516143	3771947	1001	3877968	5148714
Tennessee	45902	96167	5817887	4619097	1135	5422429	6829174
Vermont	2928	14254	620428	381930	47	564894	623989
Virginia	52025	75348	7647692	5180442	831	5888196	8535519
West Virginia	10489	38877	1668113	1355886	260	1130389	1792147
Wisconsin	34260	115673	5666400	3572163	566	4296646	5822434

Source: Data are from "Highway Statistics 2019" compiled by the U.S. Department of Transportation Federal Highway Administration https://www.fhwa.dot.gov/policyinformation/statistics/2019/. All variables compiled on an annual basis.

To account for such differences, we can create a ratio variable by dividing vehicle miles traveled by Population, Licensed Drivers, and Registered Vehicles. In each case, we derive a slightly different result. Dividing by Population includes nondrivers (e.g., children) in the ratio. Dividing by Licensed Drivers avoids counting children but may be a misleading indicator of the intensity of vehicle use as each licensed driver may have more than one vehicle. Dividing by Registered Vehicles is a good indicator of the intensity of vehicle use but a weaker indicator of driving intensity.

In creating such ratios, it is important that we attend to the units of measure. For instance, Vehicle Miles Traveled is in millions of miles. If we forget this and assume that the variable is simply miles, then we understate the amount of driving by a factor of one million—a huge mistake. Dividing Vehicle Miles Traveled (in millions) by Population equals millions of vehicle miles per person while dividing Vehicle Miles Traveled by Fuel Use equals millions of vehicle miles traveled per thousand gallons of motor fuel or thousands of vehicle miles per gallon of fuel. Multiplying this figure by 1,000 equals vehicle miles per gallon of fuel—a more manageable figure that avoids confusion with decimal places.

Measuring highway fatalities creates similar issues. Fatalities will generally be higher in states with more population, vehicle miles traveled, registered vehicles, and licensed drivers. As such, we can produce ratio variables by dividing Fatalities by any of these count variables.

Fatalities may rise because of increases in risk taking, drivers, vehicle occupants, or miles driven. Thus, dividing by Population may not accurately capture changes in vehicle occupants, miles driven, risk taking, or the number of drivers. Dividing by Licensed Drivers accounts for changes in the number of drivers but not vehicle occupants, miles driven, or risk taking. Dividing by Vehicle Miles Traveled (VMT) accounts for miles driven but not changes in the number of drivers, vehicle occupants, or risk taking. This suggests that constructing multiple ratios can help us account for the causes of changes in vehicle fatalities over time and across locations.

Other Types of Data Conversions: Adjusting for Inflation

In these final sections, we consider three very common types of data conversion or correction: corrections for inflation, seasons, and noise. These corrections are often necessary when we analyze time-series data. When inflation occurs, time-series comparisons measured in currency units (e.g., U.S. Dollars, Mexican Pesos, or Japanese Yen) are flawed. The comparison is flawed because the currency unit changes in value over time. Positive rates of inflation reflect increases in the cost of living or viewed alternatively, a decrease in the value of the currency unit. With positive rates of inflation, the buying power of the currency unit falls over time.

To correct for changes in the value of the currency unit over time, we construct a **price index** that aims to measure changes in prices. A price index measures changes in prices by comparing current prices for a set of goods to past prices for that same set of goods (a ratio variable). The index then corrects for the change in prices by multiplying the values measured in currency units by the ratio of price index values. We refer to values that are corrected for inflation as **real values** and values that have not been corrected for inflation as **nominal values**.

One common method government uses to construct a price index is the **market-basket approach**. Under this market-basket approach, the government selects a group of goods,

referred to not surprisingly as the market basket, and tracks the cost of the market basket over time. The items selected for the market basket reflect typical purchases of consumers. The government then tracks the cost of the market basket over time by constructing a ratio of the cost of the market basket in the current year to the cost of the market basket at some arbitrarily selected starting point and multiplying by 100. We refer to this arbitrarily selected starting point as the base year. Equation (2.1) shows the components of the index.

$$\text{Price Index} = \frac{\textit{Cost of the market basket in the current year}}{\textit{Cost of the market basket in the base year}} \times 100 \tag{2.1}$$

Given this, the value of the price index must equal 100 in the base year. It must equal 100 because in the base year, the base year and the current year are the same and the numerator and the denominator are the same in Equation 2.1. If, for instance, the price index in a given year is 105, this indicates that, on average, prices have risen by 5% since the base year. To convert the nominal value to a real value, we simply create the ratio of the base year price index to the current year price index and multiply the ratio by the nominal value as in Equation 2.2:

$$\text{Real Value} = \text{Nominal Value} \times \frac{\textit{Base Year Price Index}}{\textit{Current Year Price Index}} \tag{2.2}$$

In the United States, the **Consumer Price Index** (CPI), compiled by the U.S. Bureau of Labor Statistics, employs this market-basket approach. To illustrate the inflation-correction procedure using the CPI, we have gathered data on average U.S. house prices for the period 2015–2020. These data appear in Table 2.8. The data are compiled quarterly (January, April, July, and October) in each year, and we thus refer to the data as quarterly data. The first column of Table 2.8 reports the quarter for each observation, and the second column shows the average nominal house price for the United States. The third column reports the associated CPI. The final column shows the average real house price for the United States expressed in 2015 dollars. We derive this final column by multiplying the average nominal house price by the ratio of the base year CPI over the current year CPI as shown in Equation 2.2.

For example, the first observation (Jan 2015) on Real House Price (in 2015 dollars) is $350,457. We derive this value using Equation 2.2. That is, we multiply the Nominal House Price for Jan 2015 by the ratio of the base year (2015) CPI (100) to the CPI for the current quarter (99.30) (350,457 = 348,000 × [100/99.3]). The result of this inflation adjustment appears in Figure 2.8. From the figure, we can see the nominal and the real values are nearly the same for quarters near the base year of 2015. In subsequent years, the nominal values rise faster than the real values. Nominal values rise faster than real values because the difference between the base year value of the price index and the current value of the price index increases. We conclude that average house prices changed little over the period once we account for changes in the buying power of the dollar.

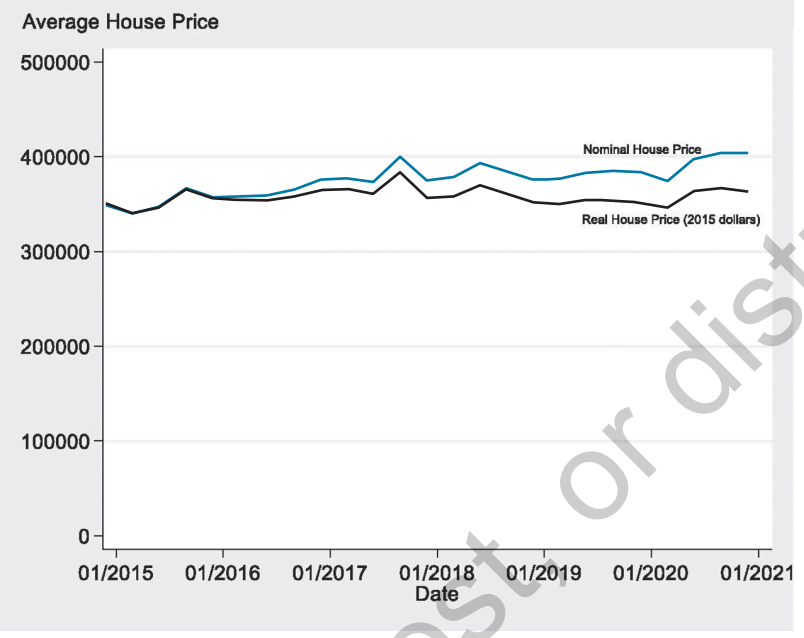
Question for Review 2.5: We can understand a price index constructed using the market-basket approach as a type of weighted average of prices. Explain.

TABLE 2.8 ■ Nominal and Real Average House Prices for the United States, 2015–2020

Date	Nominal House Price (in Current Dollars)	CPI (2015 = 100)	Real House Price (in 2015 Dollars)
Jan 2015	348,000	99.30	350,457
Apr 2015	339,700	99.98	339,782
Jul 2015	347,400	100.35	346,176
Oct 2015	366,700	100.35	365,436
Jan 2016	357,000	100.28	355,990
Apr 2016	357,900	101.09	354,056
Jul 2016	358,800	101.52	353,446
Oct 2016	364,900	102.16	357,188
Jan 2017	374,800	102.84	364,443
Apr 2017	376,900	103.03	365,833
Jul 2017	373,200	103.51	360,532
Oct 2017	399,700	104.32	383,147
Jan 2018	374,600	105.13	356,316
Apr 2018	378,400	105.79	357,693
Jul 2018	392,900	106.22	369,900
Oct 2018	384,000	106.63	360,114
Jan 2019	375,500	106.82	351,520
Apr 2019	376,700	107.75	349,623
Jul 2019	382,700	108.09	354,057
Oct 2019	384,600	108.79	353,513
Jan 2019	383,000	109.06	351,170
Apr 2020	374,500	108.21	346,092
Jul 2020	397,800	109.45	363,445
Oct 2020	403,900	110.11	366,810
Jan 2021	403,600	111.13	363,181

Source: Average house price is from the U.S. Census Bureau and the Consumer Price Index is compiled by the Bureau of Labor Statistics. Both data series are available at Federal Reserve Bank of St. Louis website: https://fred.stlouisfed.org/.

FIGURE 2.8 ■ Nominal and Real Average House Prices for the United States, 2015–2020



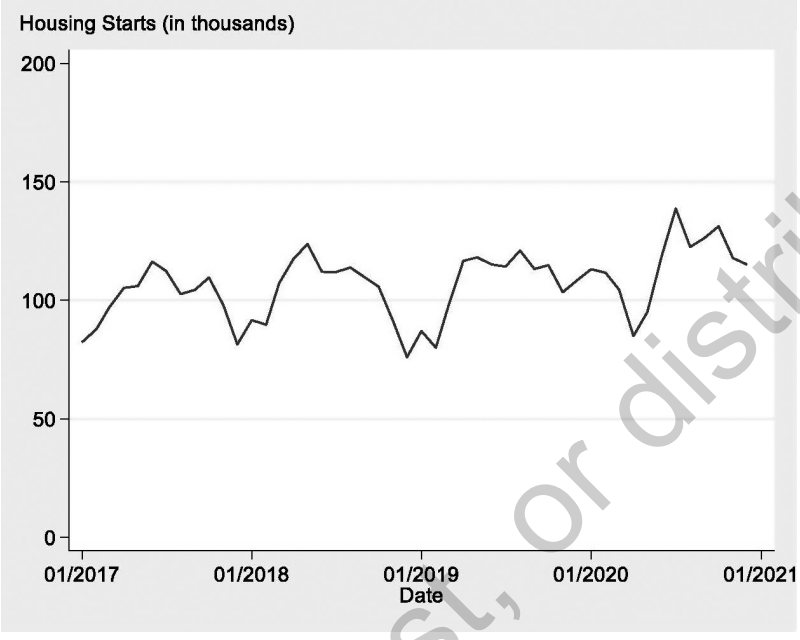
Sources: U.S. Census Bureau (average house price) and Bureau of Labor Statistics (Consumer Price Index). Both data series are available at Federal Reserve Bank of St. Louis website: https://fred.stlouisfed.org/

Other Types of Data Conversions: Adjusting for Seasonality

Many time-series variables exhibit seasonal variation or seasonality. Campaign spending, retail sales, electricity use, and accidental deaths are but a few examples. When time-series variables are seasonal, they exhibit regular and predictable patterns at time intervals of less than one year. This variation is often noticeable on a simple time-series line graph. The graph will show a cycle as low points and high points that appear at roughly the same time year after year. Consider Figure 2.9.

The figure reports housing starts for the United States compiled by the U.S. Department of Commerce (USDOC) from a survey of homebuilders. The USDOC counts a start when the construction begins on the footings or foundations of a residential structure. Analysts consider such starts as important indicators of future economic activity. However, we can see that the starts have a strong seasonal component. Housing starts are lower in colder months (December–February) than in warmer months for the entire data set. As such, the comparison we typically make using time-series data—comparing the prior month with the current month—is misleading. A drop in housing starts in December compared to November may not indicate economic weakness but rather the typical reduction in activity caused by cold weather.

FIGURE 2.9 ■ Housing Starts in the United States (in thousands), 2017–2020



Source: U.S. Census Bureau. The data series is available at Federal Reserve Bank of St. Louis website: https://fred.stlouisfed.org/

One simple method to correct for seasonality in data is to calculate annualized percentage changes from month to month. That is, rather than compare December to November we compare this December to December in the prior year. Thus, we complete the following calculation:

$$\text{Annualized Percentage Change} = \frac{Housing\ Starts_t - Housing\ Starts_{t-12}}{Housing\ Starts_{t-12}} \times 100 \quad (2.3)$$

where *t* indicates the current month and *t-12* indicates the same month one year prior. Table 2.9 shows the outcome of this calculation using monthly data for the 2017–2020 period. The table reports data for 2016 housing starts (Column 3) as these figures are necessary to calculate the annualized percentage changes from month to month. Column 4 shows the 12-month lag for housing starts, and consequently, the January 2016 figure from Column 3 for housing starts appears as the January 2017 observation in Column 4, and so on. Column 5 reports the outcome of the calculation shown in Equation 2.3 using the data from Columns 3 and 4.

Figure 2.10 shows the month-to-month annualized percentage changes in U.S. housing starts for the period 2017–2020. From the figure, we can see that there is a large surge in housing starts in late 2019 that extends into early 2020. Following this surge, housing starts plummet in March 2020, likely as a result COVID-related shutdowns. Housing starts begin to show recovery in June 2020. This suggests a surge in housing demand that pre-dates the COVID

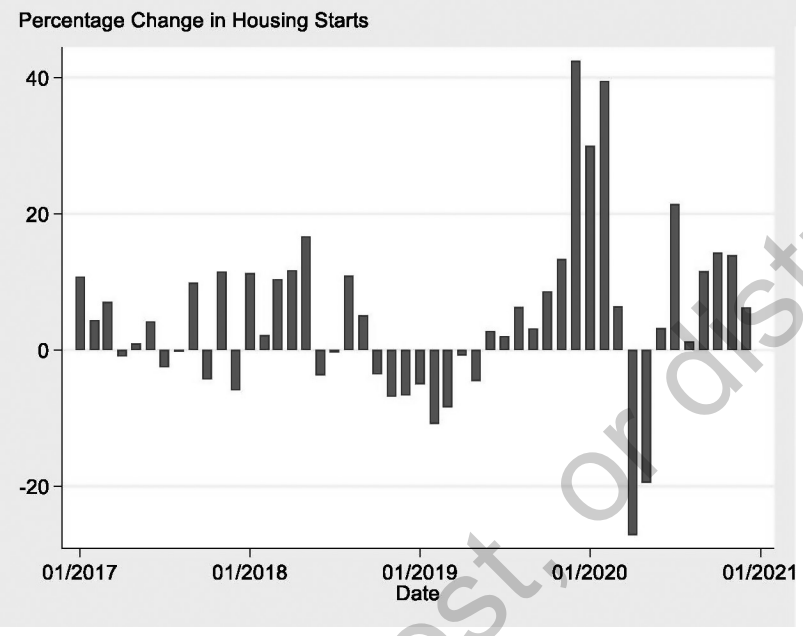
TABLE 2.9 ■ Housing Starts (in thousands) and Percentage Change in Housing Starts for the United States, 2016–2020

Date	Month Number	House Starts (in thousands)	House Starts Lag12	Percentage Change in House Starts
Jan-16	1	74.3		
Feb-16	2	84.1		
Mar-16	3	90.7		
Apr-16	4	106.2		
May-16	5	105		
Jun-16	6	111.6		
Jul-16	7	115.2		
Aug-16	8	102.8		
Sep-16	9	95		
Oct-16	10	114.5		
Nov-16	11	87.8		
Dec-16	12	86.5		
Jan-17	13	82.3	74.3	10.8
Feb-17	14	87.8	84.1	4.4
Mar-17	15	97.1	90.7	7.1
Apr-17	16	105.2	106.2	–0.9
May-17	17	106	105	1
Jun-17	18	116.3	111.6	4.2
Jul-17	19	112.3	115.2	–2.5
Aug-17	20	102.6	102.8	–0.2
Sep-17	21	104.4	95	9.9
Oct-17	22	109.6	114.5	–4.3
Nov-17	23	97.9	87.8	11.5
Dec-17	24	81.4	86.5	–5.9
Jan-18	25	91.6	82.3	11.3
Feb-18	26	89.7	87.8	2.2
Mar-18	27	107.2	97.1	10.4
Apr-18	28	117.5	105.2	11.7
May-18	29	123.7	106	16.7
Jun-18	30	112	116.3	–3.7

Date	Month Number	House Starts (in thousands)	House Starts Lag12	Percentage Change in House Starts
Jul-18	31	111.9	112.3	–0.4
Aug-18	32	113.8	102.6	10.9
Sep-18	33	109.7	104.4	5.1
Oct-18	34	105.7	109.6	–3.6
Nov-18	35	91.2	97.9	–6.8
Dec-18	36	76	81.4	–6.6
Jan-19	37	87	91.6	–5
Feb-19	38	80	89.7	–10.8
Mar-19	39	98.2	107.2	–8.4
Apr-19	40	116.6	117.5	–0.8
May-19	41	118.1	123.7	–4.5
Jun-19	42	115.1	112	2.8
Jul-19	43	114.2	111.9	2.1
Aug-19	44	121	113.8	6.3
Sep-19	45	113.2	109.7	3.2
Oct-19	46	114.8	105.7	8.6
Nov-19	47	103.4	91.2	13.4
Dec-19	48	108.3	76	42.5
Jan-20	49	113.1	87	30
Feb-20	50	111.6	80	39.5
Mar-20	51	104.5	98.2	6.4
Apr-20	52	84.9	116.6	–27.2
May-20	53	95.1	118.1	–19.5
Jun-20	54	118.8	115.1	3.2
Jul-20	55	138.7	114.2	21.5
Aug-20	56	122.5	121	1.2
Sep-20	57	126.3	113.2	11.6
Oct-20	58	131.2	114.8	14.3
Nov-20	59	117.8	103.4	13.9
Dec-20	60	115.1	108.3	6.3

Source: Housing starts is from the U.S. Census Bureau. The data series is available at Federal Reserve Bank of St. Louis website: https://fred.stlouisfed.org/.

FIGURE 2.10 ■ Annualized Percentage Change (month-to-month) in Housing Starts in the United States, 2017–2020



Source: U.S. Census Bureau. The data series is available at Federal Reserve Bank of St. Louis website: https://fred.stlouisfed.org/

shutdowns was reversed by the shutdowns. Importantly, for our purposes, it is difficult to spot these changes in a figure that simply shows housing starts over time (Figure 2.9).

Question for Review 2.6: Name three variables or measures that you expect would show seasonality. Explain why you expect them to show seasonality.

Other Types of Data Conversions: Adjusting for Noise

In the case of housing starts, we sought to remove the seasonal effects from the data to obtain a clearer picture of the more fundamental forces at work. This clearer picture improved our insight on the future performance of the housing market and the broader economy. Of course, not all factors that obscure fundamental forces are the result of seasonal factors. Some factors that make it difficult to extract meaningful information from data may be varied and poorly understood. Asset prices, or more specifically stock prices, are a case in point. Stock prices may change on a given day because of computer trading algorithms, dividend payments, or differences in beliefs. Each of these factors can obscure the more fundamental forces and

make it more difficult to observe trends in stock prices. We often refer to these factors that obscure fundamental forces as noise. We should remove the noise when the variable shows consistent increases followed by offsetting decreases that are a high percentage of the value of the variable.

One common method to remove noise from a time-series is a simple moving average. A simple moving average (SMA) adds the n most recent values of the variable and divides by n. The average "moves" because it deletes the oldest of the n observations in the subsequent period. Because the calculation reduces the effect of outliers, we refer to a SMA as "smoothed." Consider the data in Table 2.10. The table reports daily values of a stock index

TABLE 2.10 ■ The Dow Jones Home Construction Index April 2021 to May 2021

Date	Dow Construction Index	7-day Moving Average of Dow Construction Index
1-Apr-21	12392.5	
5-Apr-21	12554.8	
6-Apr-21	12629	
7-Apr-21	12460.3	
8-Apr-21	12406.4	
9-Apr-21	12731.9	
12-Apr-21	12777.5	12601.03
13-Apr-21	12682.8	12611.8
14-Apr-21	12661.6	12652.05
15-Apr-21	12714.1	12713.59
16-Apr-21	13041.5	12775.5
19-Apr-21	12933	12806.6
20-Apr-21	12574.3	12784.91
21-Apr-21	12708.6	12794.31
22-Apr-21	12683	12788.08
23-Apr-21	12889.8	12757.75
26-Apr-21	13030.6	12777.26
27-Apr-21	13108	12883.99
28-Apr-21	13080.6	12958.39
29-Apr-21	13304.2	13082.64

(*Continued*)

TABLE 2.10 ■ The Dow Jones Home Construction Index April 2021 to May 2021 *(Continued)*

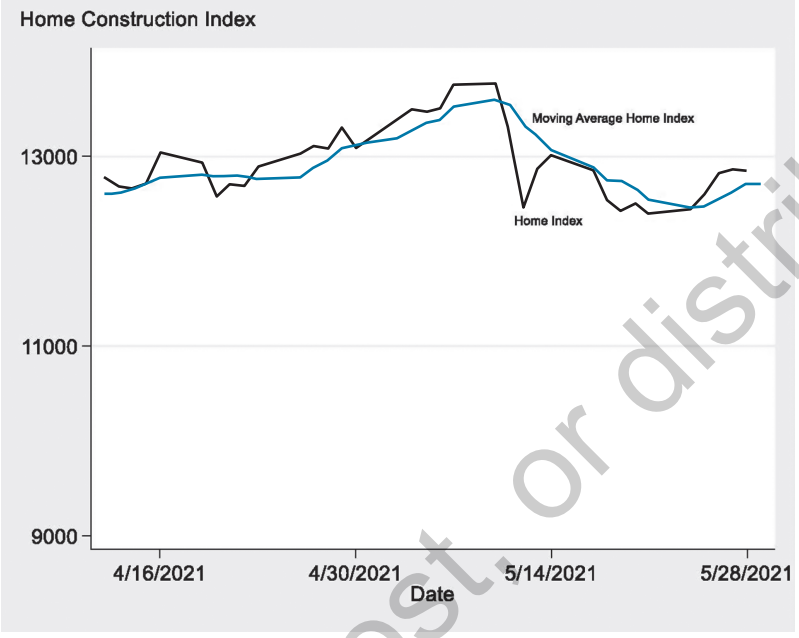
Date	Dow Construction Index	7-day Moving Average of Dow Construction Index
30-Apr-21	13086.1	13121.89
3-May-21	13386.4	13193.06
4-May-21	13494.7	13270.41
5-May-21	13467	13347.69
6-May-21	13499	13386.65
7-May-21	13755.5	13520.53
10-May-21	13771.1	13597.48
11-May-21	13231.9	13544.91
12-May-21	12448.2	13341.14
13-May-21	12868.3	13215.01
14-May-21	13012.1	13066.32
17-May-21	12859.2	12883.93
18-May-21	12536.8	12744.92
19-May-21	12424.1	12740.09
20-May-21	12500.4	12666.5
21-May-21	12391	12542.3
24-May-21	12441.8	12458.82
25-May-21	12606.6	12472.78
26-May-21	12817.9	12551.55
27-May-21	12863.2	12624.11
28-May-21	12844.6	12714.83

Note: Dow Construction Index is an index that tracks stock prices for firms in the residential construction sector. See https://www.barrons.com/market-data/indexes/djushb? countrycode=xx.

(i.e., Dow Jones U.S. Select Home Construction Index) that is intended to track stock prices for firms in the residential housing sector.[1] Suppose that we wish to calculate a 3-day SMA for April 6, 2021.

We first add the three most recent values 12629 (April 6) + 12554.8 (April 5) + 12392.5 (April 1) and then divide by 3. For April 7, we repeat the process but drop the April 1 value and include the April 7 value: [12460.3 (April 7) + 12629 (April 6) + 12554.8 (April 5)/ 3].

FIGURE 2.11 ■ Dow Home Construction Index Raw Data and Seven-Day Moving Average, April and May, 2021



Notes: Dow Construction Index is an index that tracks stock prices for firms in the residential construction sector. See https://www.barrons.com/market-data/indexes/djushb?countrycode=xx

Column 3 of Table 2.10 shows the seven-day moving average for the Dow Jones U.S. Select Home Construction Index for April 12 to May 28, 2021 (i.e., the Smoothed Home Construction Index). To show the net effect of calculating a moving average (or smoothing), we plot in Figure 2.11 the actual index against the smoothed index. The line graph clearly shows that it is easier to discern the trend in the smoothed data.

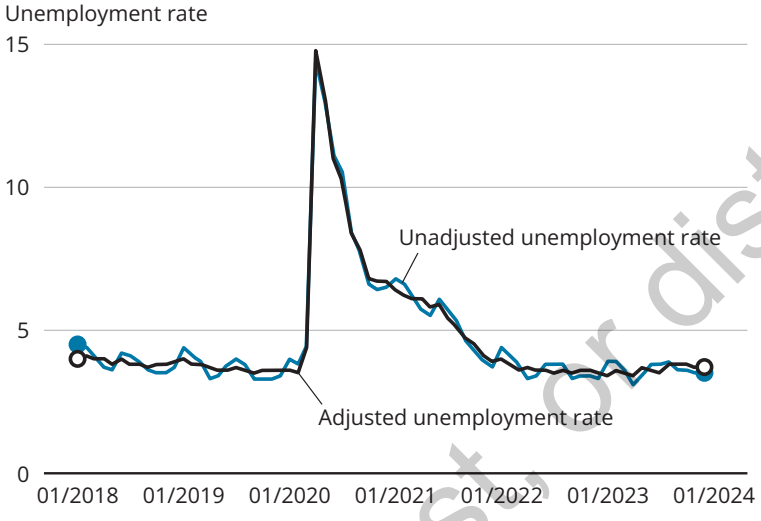
Questions for Review 2.7: During the Covid-19 pandemic, data on Covid-19 deaths and hospitalizations were often reported as a moving average (often, but not always, a seven-day moving average). What was the value of reporting the data in this way? What were the likely sources of noise in the Covid-19 data?

EXERCISES

1. This chapter offers a classification scheme for variable types: nominal, ordinal, interval, and ratio. What is the point of creating such a classification scheme? That is, how does it aid our ability to analyze data?

2. Consider the following figure that reports both adjusted and unadjusted unemployment rates for the United States (data from the U.S. Bureau of Labor Statistics: https://data.bls.gov/dataViewer).

FIGURE 2.Q2 ■ Adjusted and Unadjusted Employment



Source: Bureau of Labor Statistics

   (a) Is there a cyclical pattern in the unemployment rate? In which months is unemployment generally higher? Why? Why adjust these data?

   (b) How can we adjust the unemployment rate?

   (c) The chart shows a large spike in the unemployment rate that coincides with the onset of the Covid-19 pandemic in March 2020. What do you notice about the adjusted and the unadjusted unemployment rates in 2020? Does this suggest a flaw in the adjustment process? Explain.

3. Why must we order the *x*-variable data when we construct a line graph? Is this true for a scatter graph? Why?

4. Is it possible to use a scatter diagram to display panel data? Explain.

5. Can we use a scatter diagram to report a relationship between a nominal variable and ratio variable? Explain.

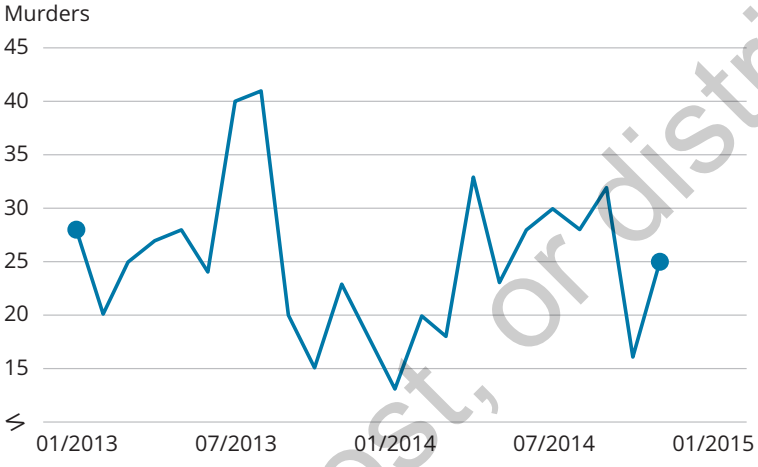
6. Consider the possible data displays that report relationships between a categorical (or nominal variable) and a ratio or a count variable. What type of figure normally produces a more informative display of such data? Why?

7. Identify the circumstances under which data should be corrected by (i) multiplying the variable by the ratio of price index values; (ii) converting the variable into an annualized

percentage change; and (iii) converting the variable into a n-observation moving average. Are there circumstances under which we should employ more than one of these corrections on the same variable?

8. **(a)** The Figure 2.Q8 displays murders in Detroit, MI over a two-year period. Explain why we might want to transform these data to increase their interpretability.

FIGURE 2.Q8 ■ Murders in Detroit 2013–2014



Source: Uniform Crime Reports

(b) Explain the changes we would make to increase their interpretability. Supply a mathematical operation that shows the change.