1

Types of Data

Steven Wright once joked that "42.7% of all statistics are made up on the spot."[1] One reason that his quip is effective is because there are good reasons to be suspicious of many of the statistics we encounter every day. Statistics are often reported as hard facts that cannot be argued with. This is not so. Statistics, and the data that the statistics are derived from, are generated by humans. Humans are not infallible and neither are the numbers reported from analyzing the data. As consumers of information, sometimes the statistics we encounter are just simply wrong or even nonsensical. There are examples of peer-reviewed publications reporting 200% reductions in some metric. Even reductions of 12,000% have been reported.[2] Without even glancing at the data analyzed in these studies, we know that such statistics are nonsense. You cannot decrease anything by more than 100%. Once you lose 100% of stuff, you are out of stuff. We tend to believe assertions when they are based on data. The problem is that we often do not look carefully at what type of data is being analyzed, how the data were gathered, and whether the results are valid. To be an active and informed citizen, you need to understand a bit about how statistics are generated and what they can tell us. It all starts with understanding the type of data being analyzed, which is the focus of this first chapter.

In the broadest terms, *statistics* is the science of collecting, analyzing, and interpreting data. One branch of statistics is concerned with how to describe and present data in useful ways (*descriptive statistics*) and the other branch is concerned with how to use samples of data to draw conclusions about unknown characteristics of a larger population (*inferential statistics*). In either case, the starting point is understanding a bit about data. Often, when students hear the term data or data analysis, they picture some geek crunching through endless columns of numbers in search for answers. The truth is that data are simply organized *information*. Data does not have to be numeric, and not all numeric

[1] He also has a line that "five out of four people have trouble with fractions."
[2] Pollack, L. and H. Weiss. (1984) "Communication satellites: Countdown for intelsat VI." Science 223(4636):553.

A Guide to Business Statistics, First Edition. David M. McEvoy.

data can be treated the same way. One great thing about the modern state of technology and connectivity is that we have access to incredible amounts of interesting, and often peculiar, datasets. For example, you can read the last words of every executed criminal in the state of Texas since 1982.[3] Or, if you think that is too morbid, you may be interested in the location, speed, age, and height of amusement park rollercoasters found all over the world.[4] Perhaps, you want to rank every character on the Simpsons by the number of words they spoke between season 1 and season 26.[5] The point is that there is so much data available to the public that the possibilities are endless. If you want to get weird, get weird.[6] You can let your imagination lead you to data, but let this book guide you on how to analyze it.

The important point is to recognize what type of data you are working with because that will dictate the way you analyze it. In this chapter, we consider the taxonomy of different data types. To begin, all data can be broadly classified as either *categorical* or *numerical*.

1.1 Categorical Data

Categorical data (also called *qualitative data*) have values described by words rather than numbers. Examples include gender, occupation, major, and location. Often, categorical data are represented with *codes* to make it easier to manage and manipulate. For example, a dataset that includes college majors may convert accounting = 1, economics = 2, and marketing = 3. The important distinction between these codes and numeric data is that the codes typically do not convey a ranking, they are just a way to organize categorical data. When data can be classified by two categories, we call that *binary* data. Examples include gender in which female = 1 and male = 0. Even when data have more than two categories, the qualitative data can often be represented in binary form. As an example, consider the three majors: accounting, economics, and marketing. If each observation in a dataset is a single student, then three binary variables (*accounting*, *economics*, and *marketing*) could be generated. When either of the three binary variables take a value of 1, it indicates that the student is majoring in the respective field. A 0, on the other hand, indicates that the student is not majoring in that field.

To illustrate the use of categorical data, consider the dataset in Table 1.1. The dataset includes the characteristics of students taking an undergraduate course in business statistics. The first two columns of data – *Student* and *Dorm* – are

[3] https://www.tdcj.state.tx.us/death_row/dr_executed_offenders.html

[4] https://www.statcrunch.com/app/index.php?dataid=1004405

[5] http://toddwschneider.com/posts/the-simpsons-by-the-data/

[6] An ambitious chap shared a dataset classifying every bowel movement he made over 2 years. There is even a histogram. http://imgur.com/a/n5Gm0

Table 1.1 Student characteristics from an undergraduate course in business statistics.

Student	Dorm	Floor	GPA	SAT rank
Barry	Hawthorne	5	3.98	1
Cindy	Whittier	3	2.87	10
Stan	Dickinson	1	1.98	9
Donna	Dickinson	−1	4.00	2
Drew	Whittier	−2	3.20	5
Wilbur	Fairchild	0	2.56	6
Frank	Hawthorne	4	2.98	8
Jose	Emerson	2	3.12	7
Paul	Hawthorne	1	3.45	4
Steve	Emerson	5	3.88	3

categorical. This includes the student's first name and the name of the dorm each student lives in on campus. While it may be possible to apply codes to these categorical variables (e.g., student ID's in place of names) those numbers would just be used as an alternative way to categorize data and would not reflect magnitudes or ranking.

The remaining three variables: *Floor*, *GPA*, and *SAT Rank* in Table 1.1 are numeric. The variable *Floor* denotes which floor they live on in their respective dorm. The numbers follow European conventions with 0 being the ground floor and negative numbers indicating floors below ground. The variable *GPA* is the student's grade point average capped at 4.0, and the variable *SAT Rank* ranks each student in terms of their SAT score with 1 being the student with the highest SAT score.

1.2 Numerical Data

Numerical, or quantitative, data result from some form of counting, measurement or computation. Numeric data are broken down into variables that are *discrete* or *continuous*. Discrete data are typically thought of as variables that are *countable*, in which fractions do not make sense. Often, these are integer values, and examples include the number of courses taken, number of credit hours earned, number of children, number of flights, and the number of absences. You may notice that the terminology "number of" often precedes the description of a discrete variable. In our dataset in Table 1.1, the variables *Floor* and *SAT Rank* are both discrete numeric variables. Clearly, the number of floors is countable and fractions of a floor do not make sense.[7] The variable *SAT Rank* is also discrete. The SAT rankings are integer values, can be counted, and are definitely not divisible.

In contrast, continuous variables can take on any value within an interval. Continuous data are not counted, and is usually measured. With continuous data "fractions make sense." Examples include weight, speed, height, distance, prices, and interest rates. Even if continuous data are rounded so that only integer values are reported, the data are still continuous. Age, for example, is typically reported in integer values. However, age can be measured very precisely by years, days, minutes, seconds, milliseconds, and so on. The same is usually standard with prices and other financial data. These are continuous measures that are rounded for convenience. They are not counted. The variable *GPA* in Table 1.1 is continuous.

In the later chapters, we sometimes blur the lines between discrete and continuous data. For example, the number of votes candidates receive in a presidential election is discrete. Why? Because votes are counted and fractions do not make sense. However, when the range of values is so large (e.g., millions of votes) that the difference between one unit (e.g., one vote) is so small, we sometimes treat discrete data to be continuous.

1.3 Level of Measurement

When data are categorical (or qualitative), the level of measurement is called *nominal*. Nominal data have no meaningful order and any numbers attributed to data values are simply for coding purposes. Denoting female observations with the number 1 and male observations with the number zero is an example. The numbers are not meaningful on their own and the numbers could be substituted with any other numbers without affecting the results. Dividing your classmates into geeks, dweebs, and nerds, for instance, would require nominal measurement. Simply coding students in one category, even if it is numeric, has no meaning in terms of relative rank. The level of measurement for the two categorical variables *Student* and *Dorm* in Table 1.1 is nominal.

Data that are *ordinal* in nature suggest that there is a meaningful ranking among the data, but there is no clear measurement regarding the distances between values. Placement in a race for instance could be denoted as first, second, third, and so on. Without additional clarifying data, the rankings are meaningful because we know that the second place runner finished before the third place runner, but we do not know how much faster the second place runner was relative to the third place runner. Another example is placement in an

[7] One exception is in the film "Being John Malkovich" in which many scenes took place on the 7.5 floor of the Mertin-Flemmer building.

Olympic event, where gold is better than silver that is better than bronze. However, those rankings do not convey how much better the gold medal winner was compared to the silver medal winner. Data on vehicle size could also be ordinal if it were classified as 3 = full size, 2 = compact, or 1 = subcompact. Clearly, $3 > 2 > 1$ in terms of size, but it is unclear how much bigger a full-size car is compared to a subcompact car. In Table 1.1, the variable *SAT Rank* is ordinal. The ranking indicates which student scored higher in the SAT exam (one indicating the highest grade), but it does not tell us how far the first highest score is from the second, and so on.

Interval data are numeric and have both a meaningful ranking and measurable distances between values. The defining feature of interval data is that there is no *true* zero. With interval data, a zero does not mean that the variable has no value. Temperature is the classic example. A temperature of zero degree Celsius does not mean there is an absence of temperature. Without a true zero, the numeric values cannot be divided or multiplied and still retain their meaning. A temperature of 20 degrees, for example, is not twice as warm as 10 degrees. The intervals between measures can be interpreted with precision (e.g., there is a 10-degree difference between 10 and 20 degrees), but we cannot say that 20 degrees is twice as warm. However, it is still possible to calculate an average with interval data (e.g., average temperature) and measures of variability. The variable *Floor* in Table 1.1 is interval data. A zero value does not mean the absence of a floor, it is simply a reference point. This reference point can change, for example in the United States, the ground floor of most buildings is typically a positive number. Interval data may be discrete or continuous.

The final category of measurement is *ratio*. Ratio data are like interval data except that there is a true zero. Examples include weight, height, speed, the number of children, number of classes, number of votes, calories, and grades. *GPA* is ratio data. Even though we do not observe a zero value for GPA, a value of zero is still meaningful. Ratio data may be discrete or continuous.

1.4 Cross-Sectional, Time-Series, and Panel Data

Another way to characterize data is by time period. When a dataset consists of observations from different individual units (e.g., people, businesses, and countries) in the same time period, we call that *cross-sectional data*. You can think of cross-sectional data as information taken from one single slice in time. US census data are cross-sectional since it consists of all individual households in a given year. The data in Table 1.1 are cross-sectional, because they consist of characteristics of 10 students in the same undergraduate business statistics course.

Time-series data, on the other hand, track observations over time. Often, time-series data follow one single individual unit (e.g., person, business, and country) over a time period. For example, tracking the daily Dow Jones industrial average over a period of 10 years would constitute a time-series dataset. Each observation is a different point in time (e.g., day, month, year, and decade). Another example is a dataset tracking temporal changes in a single company's stock price. Climate scientists rely on time-series data to understand trends in the average temperature of the earth and how those measurements interact with carbon emissions.

It is often useful to plot time-series data using a *line chart* to get a feel for specific trends, cycles, or seasons. To illustrate, consider the dataset in Table 1.2. The dataset includes voting results for every American presidential election after World War II. The data include the year, the candidate's name by party, total votes for both the democratic and republican candidates, and aggregate votes. The dataset in Table 1.2 can be considered to be time-series data. Each observation is from a different year, and the individual units are unique pairs of democratic and republican presidential candidates.

The data from Table 1.2 are plotted as a line chart in Figure 1.1. The Figure shows an increasing trend in the number of votes for candidates from both

Table 1.2 American presidential election voting results (in millions) post World War II.

Year	Democrat	Republican	Dem vote	Rep vote	Total vote
1948	Truman	Dewey	24.11	21.97	46.07
1952	Stevenson	Eisenhower	27.31	33.78	61.09
1956	Stevenson	Eisenhower	26.74	35.58	62.32
1960	Kennedy	Nixon	34.23	34.11	68.33
1964	Johnson	Goldwater	42.83	27.15	69.97
1968	Humphrey	Nixon	30.99	31.71	62.70
1972	McGovern	Nixon	28.90	46.74	75.64
1976	Carter	Ford	40.83	39.15	79.97
1980	Carter	Reagan	35.48	43.64	79.12
1984	Mondale	Reagan	37.45	54.17	91.62
1988	Dukakis	Bush Sr.	41.72	48.64	90.36
1992	Bill Clinton	Bush Sr.	44.86	38.80	83.66
1996	Bill Clinton	Dole	47.40	39.20	86.60
2000	Gore	Bush Jr.	51.00	50.47	101.46
2004	Kerry	Bush Jr.	58.89	61.87	120.77
2008	Obama	McCain	69.46	59.93	129.39
2012	Obama	Romney	65.92	60.93	126.85
2016	Hillary Clinton	Trump	65.85	62.99	128.84

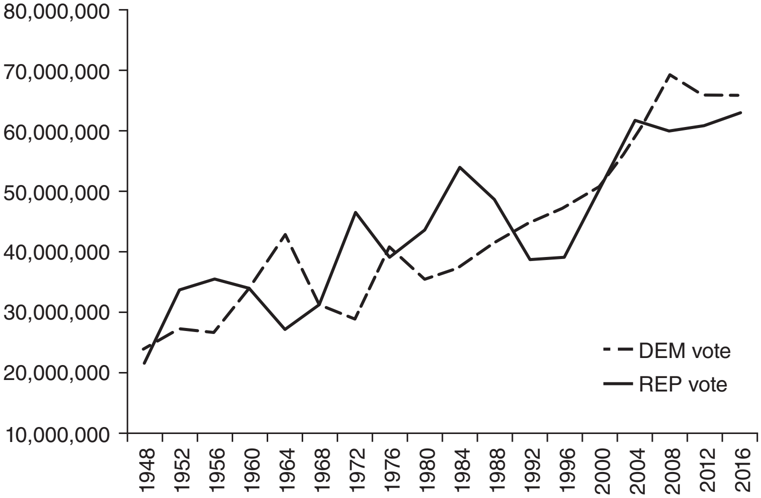

Figure 1.1 Number of votes for each party in U.S. presidential elections after World War II.

parties over time. Since the population is growing, it is unsurprising to see an increase in the total number of votes. What is more interesting is how the Figure shows repeated cycles in which one party votes more than the other.

When a dataset has multiple individual units and observations are taken at different points of time, we call that *panel data*. Tracking the stock price for multiple companies over a 5-year period would be panel data. Another example would be data on the number of regular season wins over a span of 15 years for all 30 teams in Major League Baseball.

1.5 Summary

The starting point with a course in statistics is understanding the differences in the types of data you may encounter. Data are categorical (qualitative) or numerical (quantitative). Categorical data are described by words rather than numbers. Measurement for these variables is classified as *nominal*, and they cannot be ordered in any meaningful way. Numeric data can be either discrete (countable – fractions do not make sense) or continuous (uncountable – fractions make sense). Measurement for numeric data can be *ordinal* – can be ordered, but there is no measurable distance between values, *interval* – can be ordered, distances between values can be measured, but there is no true zero, or *ratio* – like interval data, but there is a true zero. Finally, data taken from one point in time is cross-sectional, and data tracking values over a time period is time series. When a dataset includes both cross-sectional and time series, we call that a panel dataset.