

Chapter 1

Introduction to data analysis

This course is a data analysis course. The objectives of the course are for you to get an understanding of some concepts that you can use in analyzing the sounds of languages. These include concepts from phonetics, physics, psychology, and other fields that yield the data about speech sounds that you might want to analyze. And they include concepts from probability theory and statistics that you can use to analyze the data. The starting point then is the word *data*. What does this word mean?

1.1. Observation and inference

The *Concise Oxford English Dictionary* defines the word *data* as “facts of any kind.” The word *fact*, in turn, is defined as “a thing assumed as a basis for inference,” or as a “thing certainly known to have occurred or to be true, datum of experience.” The word *datum* in the second of these definitions of *fact* is the singular form of *data*, which is defined as a “thing known or granted ... from which inferences may be drawn.”

A key idea in this circle of definitions is the notion of *inference*. Data are facts that serve as the starting point for inferring things that you cannot observe directly. This starting point can be more or less solidly grounded in experience. You’ll feel most comfortable making inferences from facts that you are fairly certain to be true because they are grounded in careful observation of things that can be experienced directly. Such primary data can come from your own or others’ first-hand experiences, observed in contexts that let you trust the accuracy of the record. This is what it means to be a *datum of experience*. Science relies critically on this kind of trust in your own and others’ observations of experience.

How can you make your observations worthy of such trust? One way is to repeat an observation several times. For example, if you’re a skilled carpenter getting ready to cut a length of board, you might measure three times before making the cut. Another way is to make the observation from as many different angles as possible. For example, if you’re a reporter preparing a story about an accident, you might interview all of the people that you can find who were at the scene of the accident, asking each what he or she saw and heard, before writing the story.

Of course, getting trustworthy data is only the first step in making valid inferences. Consider the length of board that you measure three times. These three measurements probably will vary, and what you do next depends on how close they are to each other. If the variation is small relative to the width of the saw blade that you plan to use, you might decide to use the middle one of the three measurements as the cutting mark. Or if the variation is large and you notice that your tape measure is wobbling in the breeze, you might decide to make a new set of measurements using the shorter, but rigid yardstick in your toolbag.

Similarly for the accident, if the five different eyewitnesses that you found tell you quite different things, what you do next depends on how big the discrepancies are and whether you can make sense of the discrepancies in terms of other facts about the context, such as where each of the five was standing when the accident happened. You might decide that the discrepancies can be reconciled to support a news piece that simply reports “the facts” of the accident. Or you might decide instead to write a more complicated narrative that acknowledges and discusses the contradictions among the different eyewitness accounts, and propose several different versions of “the facts” of the accident, along with your estimation of the relative likelihoods of the different versions being true.

The same relationships between observed facts and inferred facts hold in science. Scientific data are trustworthy if your observations can be repeated successfully from different angles. Also, your summary statements or conclusions about the data should not be stronger than can be supported by a reliable chain of inferences. There are two modern subdisciplines of mathematics called *probability theory* and *statistics* which give you precise ways of quantifying both the trustworthiness of your recorded observations and the degree to which they support the inferences that you make from them.

1.2. Data types

It's in the nature of repeated observations that they vary from one record to the next. Because data are typically variable, the set of possible values for any type of data is often called a **variable**. There are two basic types of data, categorical variables and numerical variables.

A **categorical variable** is a set of recorded observations of group or category membership. For example, in describing a population of adult humans, you might categorize them as falling into one of the two mutually exclusive groups — “men” and “women” — which are the categories relevant for the variable “sex”. Similarly, in describing the consonant sounds that you pronounce for the letter ‘d’ in spellings of some particular list of English words (such as the *Merriam-Webster’s Online Dictionary* or all of the words that you learned to read by the time you were in sixth grade), you might categorize each example of ‘d’ as having a *pronunciation* that falls into one of the following three categories: (1) the sound **d** for the letter ‘d’ in various positions in spellings such as *deep, durable, davenport, audio, tedious, aboard, and loaded*, (2) the sound **dʒ** for the word-medial letter ‘d’ in spellings such as *schedule, graduate, and soldier*, and (3) the sound **t** for the word-final letter ‘d’ in spellings such as *forked, malnourished, and dachshund*. The distinct categories that you use in classifying categorical variables are called the **types** for the variable. To quantify your observations, then, you might count the number of observed instances for each of the types. These observed instances are called **tokens**. Thus, in talking about categorical variables, there are two distinct numbers involved: one is the number of category types and the other is the number of tokens for each type. So, to quantify your observations, you would count the number of tokens (individual people) who exemplify each of the two types “men” and “women” (for the variable “sex” in the first example) and you would count the number of tokens (individual instances of the letter ‘d’ in a list of words) that exemplify each of the three types of sound **d, dʒ, and t** (for the variable “pronunciations for the letter ‘d’” in the second example). Once you’ve quantified the variable in this way, you could rank the types by frequency, making such generalizations as “The sound **d** is by far the most common a pronunciation for the letter ‘d’ in English word spellings, as shown in Figure 1.1. When an English-speaking child first learns to read, he or she probably implicitly uses such quantitative generalizations to learn to read new words. However, the categorical variable is not inherently quantitative or ordered.

A **numerical variable**, by contrast, is inherently quantitative. It is a set of recorded observations that are inherently ordered on a scale of numbers. For example, in describing a population of adult humans, you might characterize them not just in terms of the two types for the variable sex, but also in terms of the much larger number of “types” that you can set up for the variable “height” obtained by measuring each person in the population using the same measuring device, such as a standard meter stick. In this variable, the individual measurements are the tokens, and the values that you get can be arranged on a continuum ranging from the measured height for the very shortest person to the measured height for the very tallest person.

Some numerical variables, such as height measurements, have **continuous** values. That is, while you could group the values into some set of named ranges, such as “short” (less 1.5 m) versus “average” (more than 1.5 m but less than 1.8 m) versus “tall” (more than 1.8 m), you don’t have to do so, and the primary data are not inherently grouped in this way even if you use countable units such as meters or feet. You can always subdivide meters or feet into smaller ones such as centimeters or inches, and if the context supports it, you can subdivide these smaller units even further, perhaps into millimeters or $\frac{1}{4}$ inches. So the number of distinct values in a population such as height measurements for a group of adult humans is in theory infinite (although finite in practice, given that any ruler that you might use to measure the heights will have marks on it that cannot be infinitely thin or close together).

Other numerical variables, by contrast, are inherently **discrete**. For example, in describing the population of English word spellings, you might characterize them in terms of the number of letters used to spell each word in a list. This is an **integer variable** that ranges from 1 letter for words such as *I* and *a* to 18 or more letters for words such as *antivivisectionist* and *counterintelligence*. See Table 1.1 for the range of lengths of words in the *Hoosier Mental Lexicon*, which is a list of English words with

pronunciations and familiarity ratings developed by David Pisoni and his colleagues so that psychologists could use it in studying what American English speakers know about the words of their native language.

Table 1.1. Range of lengths of spelled words (counted in number of letters) for the 19,321 words in the *Hoosier Mental Lexicon*, with number of words and examples for each length.

number of letters	number of words that long	example words
1	20	a, I
2	29	ad, is, me
3	474	ago, ink, moo, oak, she, the
4	1599	alms, idea, mild, oven, reek, toga
5	2354	album, inept, mirth, opera, ridge, torso
6	3052	aboard, indeed, myopia, oyster, ragged
7	3039	anemone, insular, soldier, tedious
8	2616	aardvark, diplomat, indoors, roughage
9	2316	adventure, davenport, knicknack
10	1682	alphabetic, derivative, roundhouse
11	1075	attenuation, dilapidated, forethought
12	567	appendicitis, recapitulate, malnourished
13	315	autobiography, kaleidoscopic, syllabication
14	106	aggrandizement, whippersnapper
15	50	Americanization, postconsonantal
16	16	arteriosclerosis, multimillionaire
17	8	indestructibility, ultraconservative
18	2	antivivisectionist
19	1	counterintelligence

1.3. Populations and samples

As our first example of a numerical variable, we characterized a population of adult men and women in terms of their measured heights. Because modern statistics developed from this kind of characterization of populations of people, animals, and plants, the word *population* is also used as a technical term to refer to the observable values of a variable that might characterize the population. For example, the set of measurable heights that we might use to characterize a population of adult humans is also a “population” in this technical sense, and so is the set of letter counts for that we might use to characterize a list of English word spellings.

We said earlier that one of the reasons that the term *variable* is used for populations such as measurable heights is because the recorded value of a repeated measure can vary from one observation to the next. A second, more profound reason why we use the term *variable* is that populations themselves vary. For example, in any population of adult humans that is larger than a pair of identical twins, there will be genetic differences that lead to differences in height, and even twins can differ greatly in height due to differences in nutritional history and so on. To characterize the height of the population, then, we might need to describe the *distribution* of the height measurements variable in some way that is analogous to the distribution of tokens for the different types for a categorical variable.

A further complication comes from variation in population membership over time. Any natural population, such as a group of adult humans, will necessarily change over time as older members die off or new members migrate into the group. With these changes, there will necessarily be changes to the “population” of measured heights for the group. This kind of variation over time can affect categorical variables just as naturally as they affect numerical variables. In a time of war, for example, there might be a large decrease in population that disproportionately affects the count of men. Since in *homo sapiens* (as in many other mammalian species), adult males are larger than adult females, this will change the

distribution of values for the numerical variable “height” at the same time that it changes the token counts for the categorical variable “sex” that is associated with the group. Moreover, even during periods when the members of the population of people is constant, there may be variation in the associated population of measured heights, as the oldest members shrink due to age-related conditions such as osteoporosis. Any characterization of a population, therefore, can only be a snapshot taken at a particular time. And if the population is large, the snapshot may take longer to develop than the time course of typical changes in group membership. If our population is “measured heights of U.S. citizens older than 18 years” what are the chances that the population of U.S. citizens older than 18 years will remain constant long enough for us to measure everyone’s height? If we want a reliable characterization of heights, our only recourse may be to select a much smaller *sample* of U.S. citizens, chosen in a principled way to lessen the chances that the sample is not representative of the population as a whole, and then measure the heights of only the sample. We can then characterize the distribution of values in the unchanging sample, to make inferences about the likelihood that the population as a whole will have certain characteristics.

Since spoken languages are similarly associated with natural groups of speakers, the categorical variables that we study for speech also vary over time. For example, when we counted the letter ‘d’ at the end of the word *dachshund* as a token of the category **t** above, we were observing the particular pronunciation assigned to that word by the particular speaker of American English who provided the phonemic transcriptions in the *Hoosier Mental Lexicon*, since that was the list we used. The *Merriam-Webster’s Online Dictionary* entry for this word that we consulted at <http://www.m-w.com/dictionary/dachshund> gives three different possible pronunciations, one of which has a final **d** rather than the very German-like final **t**. Similarly, when we counted the letter ‘d’ in *tedious* as a token of the pronunciation category **d** above, we were looking at the phonemic transcriptions in the *Hoosier Mental Lexicon*, but we also know some speakers who pronounce this word with a **dʒ**, as well as other speakers who vacillate between the pronunciation with **d** and the pronunciation with **dʒ** depending on the conversational partner and the formality of the context. This makes the population of pronunciations for the letter ‘d’ in English words inherently unstable, in the same way that the population of measured heights for U.S. citizens older than 18 years is inherently unstable. Again, our only recourse is to select a sample, trying to apply the same strategies for insuring that the sample is representative. In this sense, the *Hoosier Mental Lexicon* is a sample. It is one of the infinite number of possible lists of word pronunciations to examine for the distribution of pronunciations of ‘d’.

Moreover, we can think of each word in the list as another kind of sample. That is, since the pronunciation of the ‘d’ in *dachshund* or *tedious* varies from speaker to speaker and even from occasion to occasion within speakers, we can think of pronunciations for the sound of the final consonant in *dachshund* or the medial consonant in *tedious* as categorical variables that we can observe in one of several ways. For example, we could poll a large sample of speakers, asking each of them to pronounce the word, and then count the number of speakers (tokens) who produced each of the types **d** versus **dʒ**. This is similar to what John Kenyon and Thomas Knott did to get the information about variable pronunciations of words such as *dachshund* that *Merriam-Webster’s Online Dictionary* reports. Another thing we could do is to observe a smaller number of speakers for a longer period of observation (e.g., a one-hour interview) and then count the number of times each speaker produced each of the types **d** versus **dʒ** for this word and other words containing a medial ‘d’ in a similar context (e.g., *Indian, soldier, odious, idiot, stadium*). This is similar to what Edward C. Carterette and Margaret Hubbard Jones did to make generalizations about frequencies of different consonants and pronunciations of words in their study of *Informal speech* in the early 1970s.

1.4. Picturing your data

Whatever the data type you are using to represent your population, it’s always a good idea to try to find a way to draw a picture of the sample. We humans (like all other primate species) have a large portion of our brains dedicated to looking at things and to processing visual observations. It is far easier for our minds to process a picture than to process a bare list of numbers. So learning different ways to display

the distribution of values for a variable in a graph is a good shortcut to learning how to make informed decisions about what kinds of inferences about the population the data will support. Two of the oldest ways of picturing categorical data are the pie chart (as in Fig. 1.1) and the bar plot (as in Fig. 1.2).

First, there is no easy way to encode the size of the sample to warn the viewer not to infer too much if the sample is too small. The graph to the right in Fig. 1.1 would look the same if there were only 100 words with ‘d’ in the interviews with the first graders instead of more than a 1000 words. With a barplot, by contrast, you can use the y-axis dimension to represent the number of tokens instead of the proportions, as in Fig. 1.2. In this figure, you can see very quickly that the *Hoosier Mental Lexicon* is a much bigger sample of words than the Moe et al. list.

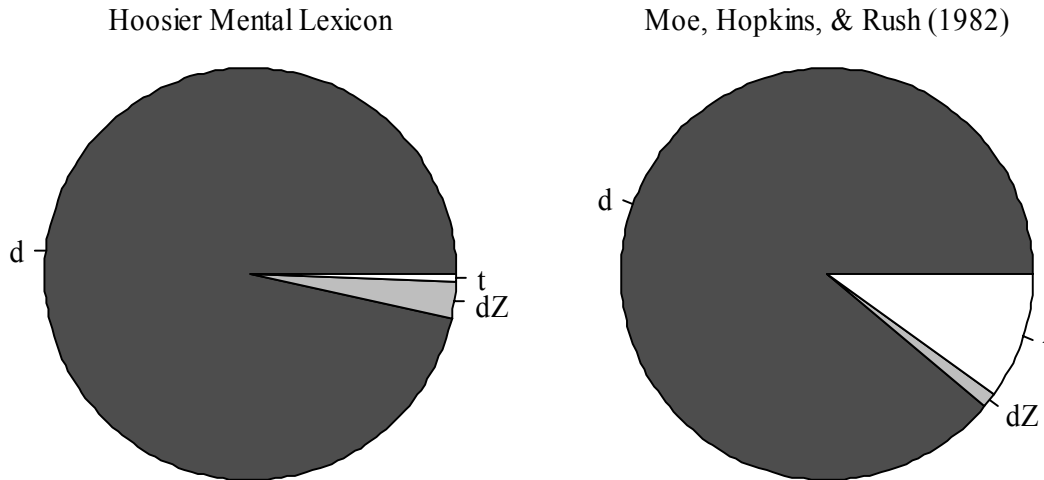


Figure 1.2. Pie charts showing the relative number of words containing the letter ‘d’ in which the ‘d’ is pronounced in each of the three ways described above. The chart on the left is for the words in the *Hoosier Mental Lexicon*. The chart on the right is for a list of 6366 words derived from an old study of children’s vocabularies that used recordings of conversations with first graders (Moe, Hopkins, & Rush, 1982).

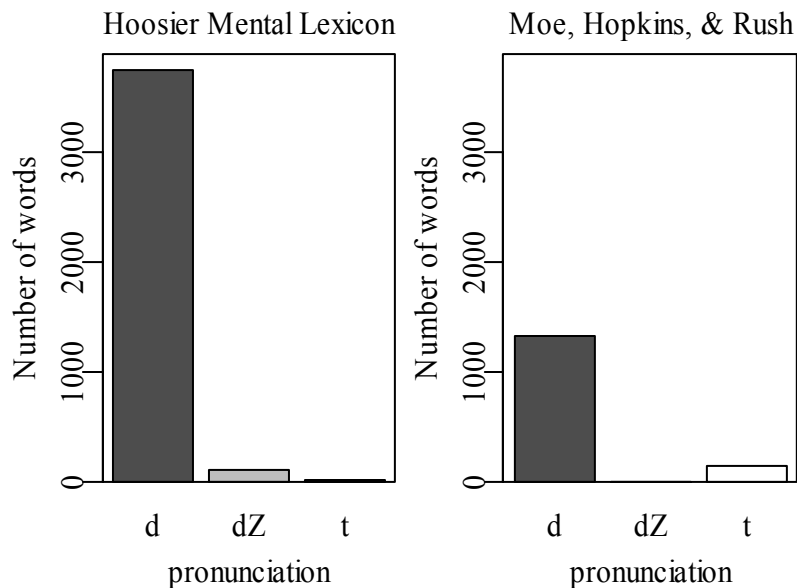


Figure 1.3. Bar charts showing the same information as in Figure 1.2.

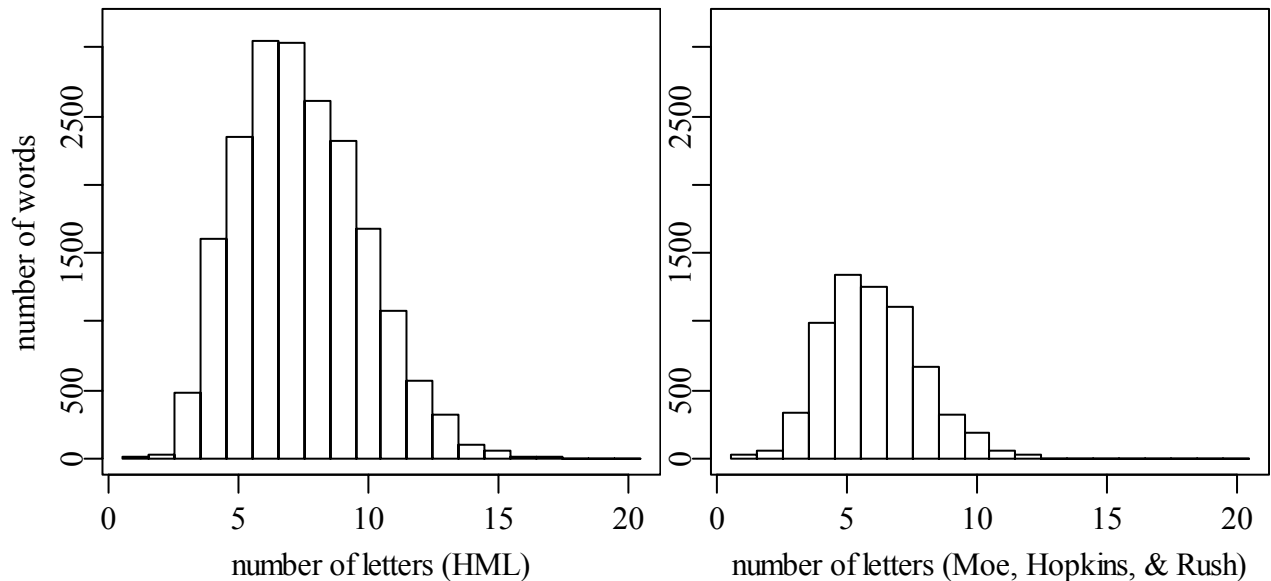


Figure 1.4. Histogram of the variable “orthographic word length” for the words in the *Hoosier Mental Lexicon* (left panel) and in the Moe, Hopkins, & Rush (1982) word list (right panel).

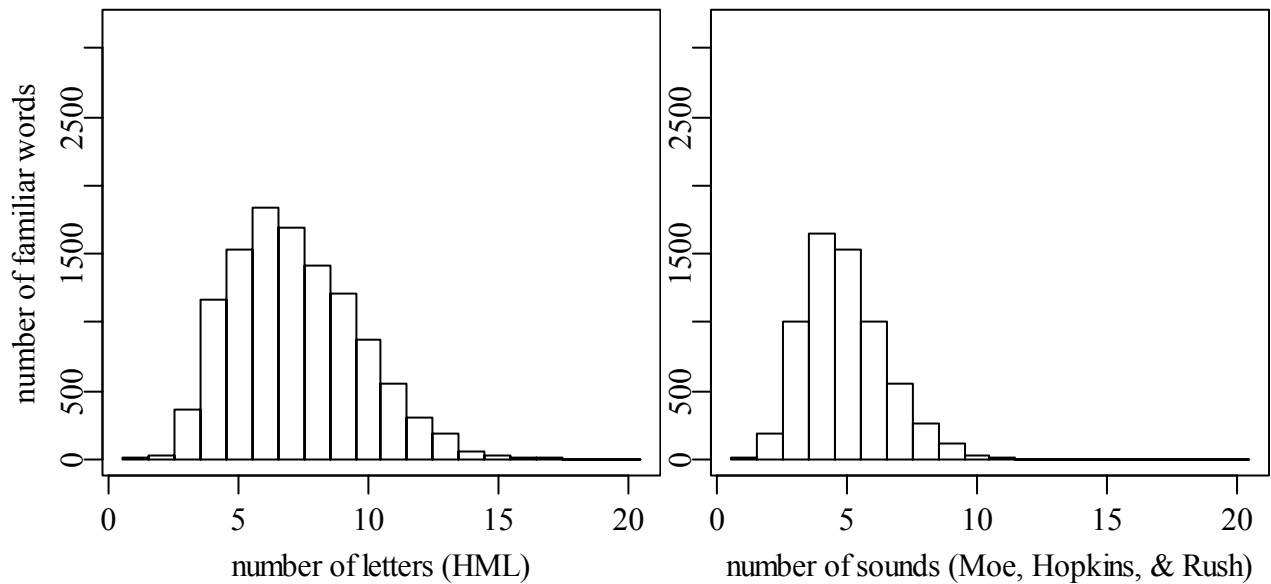


Figure 1.5. Histograms of “orthographic word length” for words in the *Hoosier Mental Lexicon* with familiarity rating of at least 6 (left) and for “phonemic word length” in the Moe, Hopkins, & Rush (1982) word list (right).

A second disadvantage of the pie chart is that it is one-dimensional. The only thing that varies is the arc lengths along the circumference of the circle. A barplot, by contrast, is two-dimensional, which means that you can use the x-axis in order show another aspect of the distribution. For example, in Fig. 1.2, the three categories are ordered by rank, from the most frequent (or most “typical”) pronunciation to the least frequent (least “typical”) pronunciation. This added dimension of the barplot means that you can also use a barplot to show the distribution of values for a numerical variable, as shown in the left-hand panel of Fig. 1.4 for the numbers in Table 1.1. This barplot shows the distribution of lengths of words in the *Hoosier Mental Lexicon*, and the different lengths are ordered on the x-axis from the shortest words, such as *a* and *I* (each of which contains only one letter) to the longest words, such as *antivivisectionist*.

That is, the “categories” plotted on the x-axis here are the inherently ordered possible values of word length in this list of pronunciations of English words. When a barplot is used in this way to show the distribution of values for a numerical variable, it is called a **histogram**.

The right-hand panel of Fig. 1.4 shows the distribution of the analogous variable for the Moe, Hopkins, and Rush (1982) wordlist. That is, this histogram plots the population of lengths of word spellings for the list of transcribed word types in these conversations with first graders. For comparison, Fig. 1.5 shows the distributions of two more variables related to word lengths in these word lists. The left-hand panel shows the number of letters in the words in the *Hoosier Mental Lexicon* that were rated as highly familiar in an experiment where many Indiana University undergraduate students were asked to rate each word in the list on a scale from 1 (for “I have never heard of this word”) to 7 (for “I know this word and use it fairly often”). The right-hand panel of the figure shows the word lengths for the Moe, Hopkins, and Rush (1982) word list, but this time counting length in terms of the number of vowel and consonant sounds in the dictionary pronunciation of the word instead of the number of letters in the word’s spelling.

Notice that each of these four histograms shows a single peak value. The peak is at 6 letters in the histograms in the two left-hand panels that plot the distributions for words in *Hoosier Mental Lexicon*. Although the peak value is the same, the shapes of the two histograms differ in a way that suggests that there are proportionally fewer long word spellings in the subset of words in the *Hoosier Mental Lexicon* that Hoosier undergraduates judged to be very familiar. By contrast, in the two right-hand panels for the Moe, Hopkins, and Rush (1982) wordlist, we see peaks at smaller values — at 5 letters in the histogram in Fig. 1.4 and at 4 sounds in the histogram in Fig. 1.5. We can infer very easily from looking at these two histograms a word’s pronunciation tends to have fewer vowel and consonant sounds than the number of letters that are used to spell the word. Why might this be? Looking at the example words in Table 1.1, we can see that some consonants, such as the consonants at the beginning of the words *she* and *the*, are typically spelled with more than one letter. Also, many words have vowels that are spelled with more than two letters. For example, the words *moo* and *oak* are both one-syllable words, just like *she* and *the*, and that means that the *oo* and *oa* are two-letter spellings of single vowels, just as *sh* and *th* are two-letter spellings for single consonants. In another chapter, we will talk about what we can say about a population from the location of peak values and from the shape of a distribution in general.

1.5. Summary

When we talk about data, we are talking about sets of reliable observations that we can use to reason numerically about the world. Some observations are inherently quantitative, derived by counting or measuring something. For example, we can measure word lengths by counting the number of letters or the number of sounds. Other observations are qualitative judgments about category membership. For example, we can group pronunciations of the letter ‘d’ in spelled words of English into at least three types. We can derive useful quantitative measures of such categorical variables by counting the number of tokens for each type. Plotting the numbers that we get in bar plots and histograms is a useful way to look for generalizations about the population.

References

- Edward C. Carterette & Margaret Hubbard Jones (1974). *Informal speech : Alphabetic and phonemic texts with statistical analyses and tables*. Berkeley, CA: University of California Press.
- John S. Kenyon & Thomas A. Knott (1944/1953). *A Pronouncing Dictionary of American English*. Springfield, MA: Merriam-Webster.
- Alden J. Moe, Carol J. Hopkins, & R. Timothy Rush (1982). *The vocabulary of first-grade children*. Springfield, IL: Thomas.
- David B. Pisoni, Howard C. Nusbaum, Paul A. Luce, and Louisa M. Slowiaczek (1985). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, 4, 75-95. (This paper is one of the earliest describing the *Hoosier Mental Lexicon*.)