Chapter 1

Data analysis and categorical variables

Contents:
1.1. How difficult is it to learn English writing?
1.2. Observation and inference
1.3. Categorical variables
1.4. Categorizing and quantifying
1.5. Counting types and tokens for categorical variables
1.6. Populations and samples
1.7. Making inferences from the type and token counts
1.8. Picturing your data
1.9. Answering our research questions
1.10. Summary
1.11. R code
1.12. Exercises
1.13. References and data sources
1.1. How difficult is it to learn the English writing system?

Perhaps you’ve heard of the word *ghoti* (pronounced *ffʃ*) as an example of how seemingly dysfunctional the English writing system is. If you hadn’t heard of it before, *ghoti* is an old orthographic puzzle that was popularized by the playwright and spelling reform advocate George Bernard Shaw. The earliest known reference to this puzzle is in a letter written in 1855 by the British publisher Charles Ollier, who said:

> My son William has come up with a new way of spelling Fish. As thus: – G.h.o.t.i. *Ghoti*, fish. Nonsense! say you. By no means, say I. It is perfectly vindicable orthography. You give up? Well then, here is the proof. *Gh* is *f*, as in *tough*, *rough*, *enough*; *o* is *i* as in *women*; *ti* is *sh*, as in *mention*, *attention*, etc.

The mid-19th century was a time of enormous popular interest in spelling reform and there are other contemporary spelling puzzles to illustrate how inconsistent English spelling is. For example, in a book published in 1845, Alexander Ellis decried the inconsistencies of English orthography, which made it so much more difficult for British children to learn to read and write than for German, Spanish, or Finnish children. To help make his point, he provided two tables.

Ellis’s first table listed all of the different ways he could find to write each English vowel or consonant sound. For example, the consonant sound *f* can be written with the single letter ‘f’ (as in the word *ffʃ* spelled *fish*). Or it can be written with the letter sequences ‘gh’ or ‘ph’ (as in the word *taʃ* spelled tough or the word *fon* spelled *phone*). We will call such a letter sequence that spells a single sound a *digraph*, to distinguish it from a letter sequence that writes a sequence of sounds, such as the ‘g’ plus ‘h’ that spells the sequence of sounds *g* followed by *h* in the word *doghouse*. We will use the cover term *grapheme* to refer to both single letters and digraphs, and we will call the different graphemes that can be used to write any particular sound, the *sound-to-letter correspondences* for that sound.

Recall that in this book, we are writing examples of ordinary *spelled forms* of words in the roman alphabet, setting them off from the surrounding text by using italics, as in *fish* versus *phish*, and we are writing individual letters using ordinary type but enclosing them in single quotes, as in the letters ‘f’ and ‘i’ and the letter sequence ‘sh’ that make the spelled form of the word *fish*. By contrast, we write examples of *spoken forms* of words in the International Phonetic Alphabet, setting them off by using boldface. And we write individual vowel and consonant sounds in the same way, as in the vowel ɪ and consonants *f* and *ʃ* that make up the spoken form *ffʃ* that is shared by the words *fish* and *phish*. You can consult the figures and tables in the preface chapter if you don’t know how to pronounce the vowel and consonant sounds that we write with the IPA and it isn’t obvious from the context.

Ellis’s second table listed all the ways that he could find to read each of the 26 letters and 176 digraphs that he identified for written words of (his 19th century British variety of) Eng-
lish. For example, the letter ‘o’ by itself can be read as any one of several vowel sounds. Ellis listed nine, including the vowel sound $\text{a}$ (in the word $\text{hop}$), the vowel sound $\text{o}$ (in the word $\text{go}$), the vowel sound $\text{u}$ (in the word $\text{do}$), and the vowel sound $\text{u}$ (in the word $\text{woman}$), as well as the vowel sound $\text{i}$ (in the word $\text{women}$) that is the intended reading in the contrived spelling $\text{ghoti}$ for the word $\text{fish}$. Ellis also listed 6 ways of reading the digraph ‘gh’, including a “silent” reading (in words such as $\text{though}$) as well as several readings as consonant sounds, including the consonant sound $\text{g}$ (in the word $\text{ghost}$) and the consonant sound $\text{p}$ (in the older spelling $\text{hiccough}$ for the word that we now spell $\text{hiccup}$), as well as the consonant sound $\text{f}$ (in the word $\text{laugh}$) that is the intended reading in the contrived spelling $\text{ghoti}$. We will call the ways of reading a grapheme its letter-to-sound correspondences.

Referring to his first table, then, Ellis calculated that by combining all the sound-to-letter correspondences for the sounds in the spoken word $\text{scissors}$, there were 1,745,226 distinct ways that the word could be written, including $\text{schiesourrhce}$. Applying the same reasoning to the letter-to-sound correspondences for the ‘gh’, ‘o’, and ‘ti’ graphemes that Ollier’s son used to create the new way of spelling the word $\text{fish}$, Ellis would have calculated 108 ways of reading $\text{ghoti}$ as a possible one-syllable word of English. That is, in his second table, Ellis lists 6 ways of reading the digraph ‘gh’ and 9 ways of reading the letter ‘o’ (as noted above) and also 2 ways of reading the digraph ‘ti’ (specifically, the consonant sound $\text{jf}$ in words such as $\text{question}$ and the consonant sound $\text{f}$ in words such as $\text{action}$). So by Ellis’s way of calculating the number of ways of reading a spelled form, there would be at least $6 \times 9 \times 2 = 108$ ways to read $\text{ghoti}$ as a possible word of English, including the pronunciations of real words $\text{tf}$ (itch) and $\text{puf}$ (push) as well as the intended joke pronunciation $\text{tf}$. And this does not begin to count the ways to try to pronounce $\text{ghoti}$ if ‘gh’ and ‘ti’ are read as letter sequences instead of as digraphs.

Notice that Ellis’s approach assumes that all of the different sound-to-letter correspondences for a given consonant or vowel sound are equally likely and that all of the different letter-to-sound correspondences for a given grapheme are possible ways of pronouncing a letter or letter sequence in any word position. But it seems very unlikely that children would become confused about how to pronounce the vowel sound $\text{o}$ in words such as $\text{profile}$ and $\text{totem}$, after learning how to spell the word $\text{women}$, which is the only word of English in which the grapheme ‘o’ is pronounced the same as the vowel sound $\text{i}$ in $\text{fish}$. And it seems quite unlikely that they would think of pronouncing the ‘ti’ sequence at the end of $\text{spaghetti}$ as the consonant sound $\text{s}$ after learning to read the digraph that way in words such as $\text{action}$, $\text{attention}$, and $\text{initials}$, where the digraph is followed by the letters ‘o’ or ‘a’ that spell the vowel sound $\text{a}$ in the suffixes $\text{tion}$ ($\text{fan}$) and $\text{tial}$ ($\text{fel}$). In other words, Ellis’s approach to showing how the English writing system is inconsistent seems to grossly exaggerate the difficulty that children face. We want a better way to estimate how inconsistent the writing system is by counting up not just how many different ways letters and sounds can be paired, but also...
calculating the likelihood of each pairing in the words that children are learning to read and write. And it seems that we need to estimate this likelihood separately for different positions in the word.

Where should we start? That is, should we start by estimating the relative likelihood that a sound will be written with a particular grapheme (e.g., the likelihood that the consonant sound \textit{s} will be spelled with the digraph ‘sc’ as it is at the beginning of the word \textit{scissors})? Or should we start by estimating the relative likelihood that a particular grapheme will be read in a particular way (e.g., the likelihood that the digraph ‘gh’ will be pronounced as the consonant sound \textit{f} as it is in the word \textit{tough})? In order to choose between the two starting points, we first need to decide what the more specific question is that we are trying to address when asking about the inconsistencies of English orthography.

In a chapter in the aptly named book, \textit{Learning to spell}, Bosman & VanOrden cite evidence that in languages with simple sound-to-letter correspondences, where most sounds are written in just one way, children learn to write sooner, and with fewer errors than in languages that have many ways of writing individual sounds. So, if we are concerned about difficulties for young children when they are first learning to write English, it makes sense to ask questions about what sound-to-letter correspondences they are most likely to need to know to write the spelled forms of spoken words that they already know.

Conversely, there are studies suggesting that English-speaking children are slower to learn to read than Finnish-, Italian-, or German-speaking children because the larger number of letter-to-sound correspondences in English compared to these more predictable writing systems. For example, Frith, Wimmer, and Landerl (1998) did a reading experiment comparing a group of German-speaking children and a group of English-speaking children, and found that the German-speaking children were consistently faster and more accurate at reading unfamiliar German words than their English-speaking peers were at reading unfamiliar English words. So, if we are concerned about the difficulties children will face in learning to read, we should ask questions about the letter-to-sound correspondences that they are likely to need to know in order to recognize the spelled forms of words they know and to pronounce the many new words that they will encounter for the first time in books, newspapers, web pages, and so on.

In this chapter, we will explore ways of answering the second set of questions – i.e., how difficult is it to learn to read and pronounce English words? We will do this by looking at a small subset of all of letter-to-sound correspondences for English, to address questions such as: How typical is it for the digraph ‘gh’ to be pronounced as the consonant sound \textit{f}? And does it make a difference where the ‘gh’ is in the spelled form of a word?

We will use these questions to illustrate important concepts about visualizing and describing categorical variables; distributions, populations, and samples; quantifying categorical
1.2. Observation and inference

The starting point for answering the questions in this chapter (and in this course as a whole), is the word data. What does this word mean? The 5th edition of 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. This singular form has its own entry in the Concise Oxford English Dictionary where it 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 facts that you cannot observe directly. This starting point can be more or less solidly grounded in experience. You make inferences from primary data when the facts are careful records of things that you or someone else experienced firsthand, in contexts that let you trust the accuracy of the observations. 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.

How can you make your observations worthy of such trust? One way is to repeat an observation several times. For example, if you’re an experienced 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 an experienced reporter preparing a story about a car 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 will probably vary, and what you do next depends on how close the numbers are to each other and what you are planning to do with the board after you have made the cut. If the variation is small, you might decide to use the middle one of the three measurements as the cutting mark. On the other hand, if the variation is large, but you’re using the board to replace a broken board in a boardwalk, you may infer that the measurement that gives you the longest piece to insert into the walk is the correct place to cut, because you can then plane the edge to be flush with the edges of the two other boards that are next to it.

In a similar way, if you’re reporting about an accident, and the five different eyewitnesses that you found tell you quite different things, what you do next depends on whether you can make sense of the discrepancies, for example, in terms of where each of the five was standing when the accident happened. You might decide that the discrepancies can be recon-
ciled to support a news piece that simply reports ‘the facts’ of the accident. Or you might decide instead to write a more complex article that discusses the contradictions among the different eyewitness accounts, along with your estimation of the likelihood that any ‘true’ account can be inferred.

The same relationships between observed facts and inferred facts hold in science. Scientific data are trustworthy if your observations can be repeated successfully (reliability) and if they can be reconciled with observations made from other angles (validity). Also, your conclusions about the data should not be stronger than can be supported by a reliable chain of inferences. There are two modern sub-disciplines 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. Later chapters will make extensive use of concepts from these two sub-disciplines of mathematics.

1.3. Categorical variables
Repeated observations usually vary from one record to the next. Because data typically vary, the set of possible values for any kind of data is often called a variable. A variable usually refers to one dimension along which the observed data vary. You could say that a variable is the characteristic that is noted or measured in each act of observing. There are two basic kinds of variables, categorical variables and numerical variables, depending on whether the types of possible values are categories or numbers. In this chapter, we will define and illustrate categorical variables, and in the next two chapters, we will define and illustrate different kinds of numerical variables.

A categorical variable is a set of recorded observations of group or category membership. Categorical variables are not inherently ordered on a number scale. For example, in describing a population of adult humans, you might categorize them as “male” and “female”, which are some of the relevant categories for the variable “sex”. While it is true that there may be more than one genetic combination of X and Y chromosomes underlying the legal interpretation of the two categories, the U. S. government only recognizes two types, for example, as they did in restricting voting to members of the “male” category before the passage of the 19th Amendment to the Constitution, and as they do for the purposes of TSA registration, and recording population characteristics in the census. So, the number of possible types depends on how you define category membership.

Similarly, suppose you are describing the set of words that you might see if you were to perform the experiment in (1). You might categorize each of the words that contain the letter ‘f’ as belonging in one of the two categories in (2).

(1) Experimental task: Get copies of all of the books that are used to teach reading in a primary school near you, and make a list of all the different words in the books that contain the letter ‘f’ in their spelled forms.
Say each word on the list to yourself and transcribe its spoken form in the International Phonetic Alphabet. Then decide which sound or sounds the letter ‘f’ is spelling.

(2) Possible categories for the letter-to-sound correspondences for the letter ‘f’.

a. The letter ‘f’ corresponds to the sound /f/ in words such as *fish*, *fold*, *office*, *stuff*, *defense*, *coffee*, and *roof*.

b. The letter ‘f’ corresponds to the sound /v/ in the word *of*.

We will call the categorical variable in (1) “letter-to-sound correspondences for the letter ‘f’”, and the categories for this variable are given in (2a,b).

1.4. Categorizing and quantifying

The categorical variable “letter-to-sound correspondences for the letter ‘f’” is one of many such letter-to-sound correspondence variables that we would want to use in analyzing data that can help us answer the question that we asked at the beginning of this chapter. Broadly, we asked how we might quantify the degree of inconsistency of English orthography. There are two general ways of describing data. A qualitative assessment is one that describes using language, abstract ideas, relationships, and so forth. A quantitative assessment is one that describes using numbers, measurements, counts, etc. So, to quantify something is to describe it in numerical terms, which are more precise and suited to data analysis than a qualitative assessment, which might describe something in more general terms.

In order to quantify the degree of inconsistency of English orthography, we have isolated one overarching question: “How difficult is it for young children to learn the patterns of English orthography in order to recognize the spelled forms of familiar words such as *ghost*, *laugh*, and *through*, and to learn new words such *ghoul*, *trough*, and *bough*?” And to begin to answer this “big-picture” question, we have identified more specific sub-questions about letter-to-sound correspondences such as: “How typical are the letter-to-sound correspondences in the “word” *ghoti*?” For example, how typical is it for the letter sequence ‘gh’ to be pronounced as the consonant sound /f/? And how typical is it for the letter ‘o’ to be pronounced as the vowel sound /ɪ/?” And finally, how typical is it for the letters ‘ti’ to spell the sound /ʃ/? It is this kind of more specific sub-question that we would address with the experiment in (1) that yields data for the categorical variable in (2). Additionally, we will examine the distribution of these letter-to-sound correspondences across different word positions. That is, how likely is it that ‘gh’ is pronounced as the sound /f/ at the beginning of a word versus the end of a word?

Of course, to estimate the degree of consistency across the whole system, we would want to look at letter-to-sound correspondences not just for these three graphemes, but also for all
the combinations of letters that represent all of the consonant and vowel sounds of the language. This could be especially challenging when you think about how, as Peter Ladefoged points out in Chapter 3 of his book *Vowels and Consonants*, the number of vowel sounds differs dramatically from one dialect to another. So the number of letter-to-sound correspondence variables will also differ. The many different pronunciations for several sounds across English dialects complicate the spelling reformers’ plea for reform. Whose pronunciation would we base the new spelling system on, and would this be disadvantageous for the majority who don’t speak that dialect? We won’t address these questions here, but they are worth thinking about.

1.5. **Counting types and tokens for categorical variables**

The distinct *categories* that you use in classifying something that is a categorical variable 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.

For example, to quantify observations of the categorical variable “sex” that we introduced at the beginning of section 1.3, you would count the number of tokens (individual humans) who are members of the category, or *type*, “male”, and of the *type* “female”. Recall that there might be more than two *types*, depending on how we define the categories, but we are just using the U. S. government’s definition for this example.

Similarly, for the categorical variable “letter-to-sound correspondences for the letter ‘f’” that we illustrated in (2), “the sound $f$” and “the sound $v$” are the two types. The individual words in (2a) are seven of the possible tokens of the type “‘f’ is pronounced as the sound $f$,” the word in (2b) is the one token we found of the type “‘f’ is pronounced as the sound $v$”

When we describe the individual words in (2a) as being seven of the many possible tokens of the type “‘f’ is pronounced as the sound $f$,” we are using a somewhat sloppy shorthand way of stating the following technically more correct description:

> Each of the seven individual instances of the sound $f$ in the words in (2a) is one of the many possible tokens of the “‘f’ is pronounced as the sound $f$” type.

In most cases, the two descriptions are equivalent. You can appreciate the difference if you think about how many tokens of this type there are in the words *alfalfa*, *fifteen*, and *fluffy*.

Why would we want to quantify a variable in this way? One reason would be to evaluate our intuitions about the *distribution* of types in a population. The distribution of a variable refers to how tokens are distributed among the possible values. That is, we might want to compare the relative number of tokens for different types of a variable. For example, we recently
looked at the faculty distribution of sex in two departments at the Ohio State University, and counted 50 men versus 45 women in Department A and 37 men versus 4 women in Department B. One of these departments is in the College of Arts and Humanities. The other is in the College of Engineering. Can you guess which department is which?

In a similar way, we may want to count the relative number of tokens to get a more sensitive measure of consistency of English spelling. After all, not all types are equally likely. While our results in (2) show two different types for the variable letter-to-sound correspondences for ‘f’, each of the types has a very different likelihood. For example, you can probably think of dozens of examples of tokens for the type “‘f’ is pronounced as the sound f” in (2a) but can you think of a single additional token for the type “‘f’ is pronounced as the sound v” in (2b)?

Let’s compare the types listed in (2) for the variable “letter-to-sound correspondences for the letter ‘f’”, and “letter-to-sound correspondences for the digraph ‘gh’” in (3) below.

(3) Possible categories for the letter-to-sound correspondence for the digraph ‘gh’ (which is another way of spelling the sound f, as in ghoti).

a. The digraph ‘gh’ corresponds to the sound g in words such as ghost, spaghetti, and Pittsburgh.

b. The digraph ‘gh’ corresponds to the sound f in words such as laugh and cough.

c. The digraph ‘gh’ corresponds to no sound (i.e., it is “silent”) in words such as neighbor, bough, through, high, and sight.

d. The digraph ‘gh’ corresponds to the sound ə in the word Edinburgh (pronounced edinbure).

When we count the number of types for these two variables, we can see that the letter-to-sound correspondences for the digraph ‘gh’ are less consistent than the letter-to-sound correspondences for the letter ‘f’. Whereas the ‘f’ letter spells two different sounds according to the type count in (2), the letter-to-sound correspondences for the digraph ‘gh’ in (3) shows four different ways that the digraph ‘gh’ can be pronounced. If we go on to also compare how the tokens are distributed across the types for each variable, we can see that the difference is actually much more dramatic than this difference between 2 types versus 4 types. As we noted, we could find only 1 token of the type in (2b), but the category in (3d) is also sparsely populated. If we use a larger sample, we can see just how big the difference in token counts for each type really is across the language.

Examples (2) and (3) could help explain why no one has any trouble pronouncing the name of Goff Street in Boston, but tourists and newcomers never know how to pronounce the homophonous name of Gough Street in San Francisco. The distribution of token counts for the different types in (3) also could help to explain why ghoti is such a lasting spelling puzzle. It is a funny puzzle, in part because our intuition is that ‘gh’ is far
from the most typical way to spell the sound \( f \), especially when it is at the beginning of the word. To find out if our intuition is correct, we can count the tokens for the different letter-to-sound correspondence types for the grapheme ‘gh’, as well as for ‘o’ and ‘ti’. But where can we find tokens that we can count? We need a sample.

1.6. Populations and samples

Before we can talk about the samples we will use for our questions, we need to discuss the term population. In the previous section, we characterized two groups of university faculty in terms of their very different token counts for two types of the categorical variable “sex”. As we also noted, this is a variable that the U.S. government also uses in the national census that happens every ten years, in which they try to gather information about the population of the United States. Modern statistics developed from this kind of characterization of populations of people, animals, and plants. Therefore, the word population is also used as a technical term to refer to all of the observable values of a variable that might characterize the population. For example, the total token counts for the two types, “male” and “female”, of the categorical variable “sex”, that the U.S. government publishes after each census is a “population” in this technical sense. In this way, so is the set of all the possible tokens of the different types for the letter-to-sound correspondence for the grapheme ‘gh’ that we might collect if we included every word in the English language that contained that grapheme.

We said earlier that a group of possible values for any set of data is called a variable because the values typically 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 most populations of humans that are greater than four or five people, there are likely to be tokens of both types of the categorical variable “sex”, and the relative number of tokens for the two types can vary from one population to another, as in the example of the faculty in the Departments A and B given earlier. To characterize the variable “sex” for any particular population, then, we might need to describe the distribution of token counts across the different possible types. As noted above, the distribution of a variable refers to how tokens are distributed among the possible values.

Populations vary in many other ways. For example, population membership can vary 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. This will also change the “population” of the variable sex for the group. In a time of war, for example, when population sizes for the warring groups tend to decrease overall, the number of men in a population may decrease much more than the number of women. This is what happened in the United Kingdom during World War I, to give a relatively recent example. Any characterization of a population, therefore, can only be a snapshot taken at a particular time.

In addition to the variation in populations, it is often not feasible to get measurements from the whole population. For exam-
ple, if we want a reliable characterization of the distribution of males and females in every country in the world, it would not be practical to go out and try to count all of the men and all of the women. Rather, we can select a much smaller group from each country that is representative of the population as a whole, and then count the members of each category in only the smaller group. This smaller group is called a sample. Since the sample represents the population, we need to make sure that our sample does not differ in obvious ways from the population. For example, given the different life expectancies of men and women, a sample representing the distribution of sexes in the U. S. should not come only from observations of people in nursing homes. Instead, we would want to include people of all ages.

Since spoken languages are associated with groups of speakers, the variables that we study about speech also vary. For example, pronunciations of words may vary depending on the dialect of the speaker and they may also change over time. For example, Ellis’s spelling of *schiesourrhce* for the word *scissors* is based on older pronunciations in a dialect in Southeast Britain. He justifies the spelling with the following letter-to-sound correspondences for the first and last sounds: s = ‘sch’ in *schism*, pronounced sizəm; z = ‘ce’ in (the verb) *sacrifice*, pronounced sækrɪfəz. This kind of variation makes the populations of letter-to-sound correspondences inherently unstable, in the same way that the distribution of types of sexes is inherently unstable. So what are we to do?

Again, our only recourse is to select a sample, trying to apply strategies for insuring that the sample is representative. In this sense, any list of words that you could get by doing the experiment described in (1) is a sample – that is, it is one of the infinite number of possible lists of word spellings that you can examine to yield a distribution of letter-to-sound correspondences for the letter ‘f’. But is this a good representative sample? This depends on what our research question is. Let us reframe the set of questions we asked earlier in the chapter about some of the difficulties that children face when learning to read English words:

(4) How often will a child learning to read encounter certain letter-to-sound correspondences? How typical are the letter-to-sound correspondences in the “word” ghoti? Specifically, how typical is it for the letter sequence ‘gh’ to be pronounced as the consonant sound f? And how typical is it for the letter ‘o’ to be pronounced as the vowel sound i?” And finally, how typical is it for the letters ‘ti’ to spell the sound ʃ?

(5) What is the distribution of letter-to-sound correspondences across different word-positions? For example, how likely is it that ‘gh’ is pronounced as the sound f at the beginning of a word versus the end of a word? And how likely is it that the letters ‘ti’ spell the sound ʃ at the end of a word as opposed to elsewhere in the word?
These are the questions that we had in mind when we wondered whether Ollier’s *ghoti* example (or Ellis’s *schiesourrhce* example) was really a valid case for spelling reform. If these are your questions, then the sample that you get by running the experiment in (1) might be a good start, but a larger sample can give us more confidence about the reliability of our estimates.

To answer these questions about primary school children, then, we will use two different, larger samples. The first is the Moe, Hopkins, and Rush (MHR) database. This is a list of 6366 words that Alden Moe and his colleagues developed in their research aimed at helping American children learn to read. They got this list by recording conversations with 329 children and then transcribing the recordings. There were 285,627 spoken words transcribed in these conversations yielding the 6366 different words in the MHR. We added IPA transcriptions of the spoken forms of these words, as they would be pronounced in the dialect of the children who were interviewed.

The second sample we found that is relevant to these questions is the *Children’s Printed Word Database* (CPWD). The CPWD is a list of all 12,452 different words that Jackie Masterson, Maureen Dixon, Morag Stuart, and Philip Quinlan compiled from books used in the first few grades of primary schools in Britain. For each different word, Masterson and colleagues provide the British English pronunciation written in the International Phonetic Alphabet.

Taken together, the two wordlists should give us reliable estimates of the type and token counts for the categorical variable that is relevant for questions (4) and (5). The estimates should be reliable because the datasets are both pretty large samples and because they were compiled in different countries. The first sample lets us address the questions from the angle of children who speak American English. The second lets us address the questions from the angle of children who speak British English. So if we observe the same distribution of letter-to-sound correspondences in the two samples, then we can be fairly confident of our quantitative assessment of this aspect of the English writing system as a whole.

### 1.7. Making inferences from the type and token counts

We used the two samples to get counts that are relevant to our questions. We first looked at letter-to-sound correspondences for the ‘gh’ digraph. We counted only the same three types in the MHR as we found in example (3). (In one of the exercises at the end of this chapter, we give you the full list of words in the MHR that include ‘gh’ in their spellings so that you can practice analyzing this kind of data. When you have done the exercise, see if you agree with us about the type count, as shown in Table 1.1 and Figure 1.1.) As you can see in Figure 1.1, ‘gh’ is most often “silent”, with only 10 words where it is pronounced as the sound *f*, and 5 words where it is pronounced as the sound *ɡ*. It is worth pointing out that the only two words beginning with ‘gh’ in this sample are *ghost* and
ghosts. Additionally, compared to the number of tokens of words containing the letter ‘f’ (518, listed in Table 1.3 in the exercise section), ‘gh’ is not a very frequent grapheme.

In the CPWD, on the other hand, there are more types of letter-to-sound correspondences for the ‘gh’ grapheme, because the British English pronunciation listed for the word ugh is ʊk, making it rhyme with book and took instead of with bug and tug, as it does in many dialects of American English. So there is an added type k. In Table 1.2, we show you the token counts for each of the four types for this categorical variable in the CPWD word list. Notice that we have divided the token counts into two groups, according to the position of the digraph in the spelled word form. In the shaded box below, we list some of the generalizations that you can make from these numbers.

### Table 1.2. Token counts for types of letter-to-sound correspondences for the ‘gh’ digraph in different word positions in the Children’s Printed Word Database (CPWD).

<table>
<thead>
<tr>
<th>IPA</th>
<th>WORD-INITIAL</th>
<th>MEDIAL, FINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>7 (100%)</td>
<td>2 (1%)</td>
</tr>
<tr>
<td>f</td>
<td>0 (0%)</td>
<td>16 (11%)</td>
</tr>
<tr>
<td>silent</td>
<td>0 (0%)</td>
<td>131 (87%)</td>
</tr>
<tr>
<td>k</td>
<td>0 (0%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>7</td>
<td>150</td>
</tr>
</tbody>
</table>

Table 1.1. Token counts for the types of letter-to-sound correspondences for the digraph ‘gh’ in the Moe, Hopkins, and Rush (MHR) wordlist, as shown in Figure 1.1. The ‘-’ represents the type “unpronounced” (or “silent”).

| IPA | MHR  |%
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>68</td>
</tr>
<tr>
<td>f</td>
<td>10</td>
</tr>
<tr>
<td>g</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>83</td>
</tr>
</tbody>
</table>

Figure 1.1. Letter-to-sound correspondences for the digraph ‘gh’ in the Moe, Hopkins, and Rush wordlist.
Generalizations about the numbers in Table 1.2 that are displayed in Figure 1.2 that show the distribution of letter-to-sound correspondences for the letters ‘gh’

1. The sound $g$ is the only pronunciation for the digraph ‘gh’ in word-initial position in the British English pronunciations of words containing this digraph in the CPWD.

2. The sound $f$ is 8 times more frequent than the sound $g$ as the value for the categorical variable letter-to-sound correspondences for the digraph ‘gh’ in word-medial or word-final position in these words.

3. However, the value “silent” (or “unpronounced”) is the most common value for this variable in word-medial or word-final position, being 8 times more frequent than the value $f$.

4. So all in all, ‘gh’ is highly predictable. In initial position, it is always pronounced as the sound $g$. Medially and finally, it is usually “silent”, with a handful of exceptions that children will have to learn.

We also asked about the letter-to-sound correspondences for the grapheme ‘ti’ in the Moe, Hopkins, and Rush database. This sequence of letters occurs 217 times in the spelled words in the database. However, the actual sample of the grapheme is much smaller, because in 183 of these instances, the letters are a sequence of two graphemes rather than the grapheme ‘ti’. That is, these are cases such as tickle, attic, Tina, and spaghetti where the grapheme ‘t’ spells the consonant sound $t$ and the following grapheme ‘i’ spells a vowel sound such as $ɪ$ (in tickle and attic) or $i$ (in Tina and spaghetti). The remaining 34 instances are examples of the grapheme ‘ti’ pronounced either as the consonant sound $tf$ (in the word question) or as the consonant sound $f$ (in words such as vacation, station, and initials). However, none of these instances of the ‘ti’ digraph is in word-final position, and in all but one word, initials, the grapheme ‘ti’ is part of the suffix –tion, as in question, vacation, and station. A child learning to read only has to learn that the suffix –tion is pronounced as $ʃən$, except in a handful of words, where the suffix follows an ‘s’, as in bastion, combustion and question (where it is pronounced $ʃən$), and then the child will know not only how to pronounce the digraph ‘ti’ but also how to pronounce the letter ‘o’ that writes the vowel $ə$ in this type of suffix.

Notice, too, that once the child has learned about the suffix –tion, the child will also be able to make the reliable generalization that the most likely pronunciation for the letter ‘t’ before the letter ‘i’ in words that don’t contain this suffix is the consonant sound $t$. Of course, it will be more difficult to make robust generalizations about how to pronounce the letter ‘i’ in these other cases, but that’s a far bigger problem. Learning to read and write the English vowel sounds is quite difficult, because there are far fewer vowel letters than there are vowel sounds. This is true of all dialects of English, even though the exact number of vowel sounds varies quite a bit. For example, most U. S. dialects have about 15 vowels, but most British dia-
lects have more than 20. Whatever dialect they speak, children need to learn many more letter-to-sound correspondences for vowel sounds than for consonant sounds because of historical changes that occurred in the English language after the orthography became standardized. We will look at some of these changes in following chapters. In the meantime, to complete our analysis for the current chapter, let’s look at the letter ‘o’ when it is a grapheme all by itself – i.e., when it is not part a digraph such as the ‘oa’ of boat or the ‘ou’ of house. Looking through the cases of the grapheme ‘o’ (i.e., the single letter acting alone to spell some vowel sound), we found that the letter ‘o’ typically represents one of four vowel sounds: o as in ocean (ɒʃən); ɑ as in operate (ɑpərət); ʌ as in other (ʌðər), or ə as in compare (kəmper). There are a few other very uncommon letter-to-sound correspondences, such as the u vowel in wolf or the wa sequence in once. Of these other letter-to-sound correspondences, the rarest is the pronunciation of ‘o’ as the vowel sound i in women, which occurs only in that one single word. In short, of the seven types that we found for the variable “letter-to-sound correspondences for the grapheme ‘o’” to gauge the likelihood of different readings of the ‘o’ in ghoti, pronouncing that ‘o’ as the i in women is the least likely.

1.8. Picturing your data

Whatever sample you are using to represent your variable, it’s always a good idea to try to find a way to draw a picture of the distribution of values in 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 counts of different 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.

One of the oldest ways of picturing categorical data is the bar plot, like the two graphs we showed in Figure 1.2. Because those graphs showed token counts for the four different types of a categorical variable, the order in which we arrange the bars is arbitrary. When we plotted the token counts for the four types of letter-to-sound correspondence for the digraph ‘gh’ in the two panels of Figure 1.2, we could have arranged the bars in any order. We chose to arrange the bars for the word-medial and word-final tokens in order from the most frequent sound to the least frequent sound, and we used that same order for the bars for the word-initial tokens so that you could see at a glance that the distributions are very different between the two positions. While f is a more frequent pronunciation than g in word-medial and word-final position, it is not as frequent as the type where the ‘gh’ is not pronounced at all. On the other hand, the most frequent (in fact the only) value for this categorical variable in word-initial position is g. This made it very easy to see the basis for the generalizations that we list under Table 1.2.
Notice, too, that we adjusted the y-axis range so that this tallest bar is at the same height in the two graphs. This brings out very clearly that, while the absolute number of tokens for each type is smaller in the MHR, the pattern of relative counts tokens across the three types is identical. When we talk about the pattern of relative counts in this way, we are talking about the distribution of the types in the sample.

Again, the generalizations that we want to make are about these distributional patterns. They are the ones that can support our inferences from the data. From these generalizations, we can guess that any other sample of pronunciations of ‘gh’ in spelled word forms of English will show that pronouncing ‘gh’ as f is not the most frequent type overall, and in fact we did not observe that type even once in word-initial position. Word-initial position is exactly where ‘gh’ takes the value f in the spelling puzzle ghoti. This helps explain why the spelling puzzle is a “puzzle” and helps us decide that this kind of spelling puzzle is not a good argument for spelling reform.

1.9. Answering our research questions

The questions that we asked at the beginning of this chapter are about the inconsistency of the English spelling system. We wondered how relevant spelling puzzles such as ghoti are to measuring the difficulty that inconsistencies in letter-to-sound correspondences cause when English-speaking children learn to read and write. In particular, we asked, how typical is it for the digraph ‘gh’ to correspond to the sound f, particularly in word-initial position? And how likely is it that the letters ‘ti’ are pronounced as f, especially in word-final position? And what is the likelihood that the letter ‘o’ is pronounced as i in any position?

To answer these questions, we counted the tokens for each of the different values of the categorical variable “letter-to-sound correspondences for the ‘gh’ digraph” in word-initial position and in word-medial or final position in the CPWD word list. We found that the value f is far from the most frequent pronunciation for this digraph in words that primary school children in Britain are likely to encounter in their textbooks. In word-initial position, in fact, it is not attested at all. That is, while children will encounter the digraph ‘gh’ inconsistently spelling the sound f in some words, the sound g in some words, and spelling no sound at all in most words, the only letter-to-sound correspondence that they will experience word-initially, where ‘gh’ is in ghoti, is ‘gh’ spelling the sound g. Similarly, we found that ‘ti’ most often represents a combination of ‘t’ and ‘i’. Word-finally, this sequence is the only possibility, so a letter-to-sound correspondence for ‘ti as the sound j never occurs at the end of a word.

Other researchers have confirmed that English orthography is less consistent than the German, Spanish, or Finnish writing systems. There are also studies showing that this has consequences for education. For example, Philip Seymour, Mikko Arno, and Jane Erkine reported in an article in 2003 that English-speaking children not only take longer to learn to read, they tend to make many more errors in reading new
words after they have learned to read. However, the inconsistencies are nowhere near as extreme as Ellis (1845) suggested. The *ghoti* argument for spelling reform is an exaggeration. The distribution of letter-to-sound correspondences is fairly predictable, especially when you take into account where in the word they occur. Vowels are more variable, but the suggestion that you can generalize the *ɪ* in *women* to any other word is highly implausible. We rate the pronunciation of *ghoti* as *fɪʃ* as extremely unlikely.

### 1.10. Summary

When we talk about data, we are talking about sets of reliably recorded observations that we can use to reason numerically about the world. Some records are qualitative judgments about category membership. For example, we can group pronunciations of the digraph ‘gh’ in spelled words of English into (at least) three types. Because the values that we observe vary, such groups of values are called variables. If we can find representative samples of tokens of a variable, 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 is a useful way to look for generalizations about the population from which a sample is drawn.

#### Summary of key terms

**Bar plot:** A picture of the distribution of a variable.

**Categorical variable:** A set of recorded observations of group or category membership. Categorical variables are not inherently ordered on a number scale. Example: “sex”.

**Data:** Facts that act as the starting point for inferring facts that are not directly observable; often a set of observations.

**Distribution:** A description of how data are distributed across the possible values of a variable. Example: The data may have a fairly equal number of tokens for each type, or one type may have a disproportionately large number of tokens.

**Grapheme:** The smallest unit of writing. This can be a letter or character. In case of a single sound represented by multiple letters or characters (like the single sound represented by ‘sh’), that sequence of letters or characters is also considered a single grapheme.

**Inference:** Indirect observations made from logical deductions using primary data.

**Population:** An entire set of possible entities or values.

**Primary data:** Facts or evidence from first-hand experience.
**Quantify:** Literally, to count something. More generally, it means to put something in numerical terms, by measuring or counting. If you can state your research question in a way that lets you quantify something, you can provide evidence for your argument by using generalizations about the numbers, such as the relative counts of the different types of a variable.

**Reliability:** Also known as replicability, that is, whether performing the same experiment multiple times, or performing the same analysis on multiple samples will yield the same results. If your analysis is reliable, others should be able to duplicate your results.

**Sample:** A portion of a population chosen to represent that population.

**Tokens:** The number of observed instances for each type.
   Example: The number of men or women for the categories/types “men” and “women.”

**Types:** Distinct categories that you use in classifying variables. Example: “men” vs. “women” for the variable sex.

**Variable:** The dimension along which your data vary and the set of possible values along that dimension. Basically, the thing that you are interested in measuring about your data.

### 1.11 R code
R code and exercises

1.11. R code

As you remember from the preface, each of the chapters of this textbook includes a section like this one, showing code similar to that which produced some of the tables and figures in the chapter. The textbook web page (http://hdl.handle.net/1811/77848) also has script files that give this code in a format that lets you open each file within R, to work through the code interactively, reading each comment (a sequence of lines that is marked off by “#” at the beginning) and then running the line(s) of code that the comment describes. Remember that you can run the code directly from the script editing window, by highlighting the line(s) you want to run, and then pressing (simultaneously) the \texttt{command}+\texttt{enter} keys (on a Mac) or the \texttt{ctrl}+\texttt{r} (on a PC running Windows or Linux). We encourage you to download and install \texttt{R} and use these \texttt{R code} sections of the textbook or the associated script files to teach yourself how to use \texttt{R} sooner rather than later. (See the R code section in the Preface chapter for more details.) In this R code section for Chapter 1, you will learn how to make graphs that are similar to the ones in Figure 1.2, for the CPWD dataset. A copy of this R note is saved inside a script that is called \texttt{Chapter01.R} and can be found on the course website or directly at http://kb.osu.edu/dspace/bitstream/handle/1811/77848/Chapter01.R.

First, use the \texttt{c()} function to create a vector (a one-dimensional matrix or a simple series of data, where all of the data are of the same type) to tell R how many tokens of the digraph ‘gh’ in the Children’s Printed Word Database in wordmedial or word-final position are silent, and how many are pronounced as f, g, or k. Call this vector \texttt{gh}. That is, use the assignment operator “=” to assign this vector to that name. (Remember that you could also use the “<-“ assignment operator.)

\texttt{gh = c(131,16,2,1)}

This command has two parts. The \texttt{c()} part on the right of the \texttt{=} creates the vector, and the name on the left in combination with the \texttt{=} stores this vector in a variable with that name. From now on, when you type the name of the variable all by itself, what you will see on the R console screen is the series of
four numbers that you assigned to it. Try doing that. Type `gh` into the console window. You should see the following on your R console:

```
[1] 131 16 2 1
```

(We can ignore the [1] here for now. Remember that formats output as a kind of table and puts row numbers to help navigate when there is a long stream of values.)

Next create a vector of character strings representing the types of sounds associated with the ‘gh’ digraph, and assign it to a variable (`sounds`), so we can refer back to this vector without having to type it in again. (Notice that these values are character strings and not numbers. That is why we needed to enclose each of them within the "...") quotation marks.)

```
sounds = c("_", "f", "g", "k")
```

Next, use the `names()` function along with the ":=" assignment operator to assign names to the numbers in the `gh` vector. Use the same order as you did in the command above.

```
names(gh) = sounds
```

Again, this command has two parts. The part on the right of the = represents the vector of names. The part on the left of the = uses the `names()` function to say that the series of character strings represented by the vector on the right will be assigned as names to the series of numbers represented by the variable "gh". Once you have done this, when you enter `gh` into the console, like this:

```
gh
```

you will see the names of the vector, as well as the numbers, like this on your R console screen:

```
_   f   g   k
131 16 2 1
```

Now, you’re ready to use the `barplot()` function to make a bar plot where the heights of the bars are the numbers in the `gh` vector. This function takes a named vector (the vector we named `gh`) as its argument (the argument is what is inside the ( . . . ) parentheses) and draws the bar plot on the plotting window.

```
barplot(gh)
```

We should add the variable name to the x-axis, and what the height of each bar represents along the y-axis. For this, we add the arguments `xlab="..."` and `ylab="..."` inside the parentheses. These stand for x-label and y-label, respectively. Make it a habit to always label your axes, or your audience will not know what your units of measurement are (y-axis), or what is being measured (x-axis).

```
barplot(gh, xlab="letter to sound correspondences for 'gh'", ylab="number of tokens of that sound")
```
You can save the graph using “Save as” in the drop-down menu. Make sure that you have first highlighted the plotting window, then go to File > Save as…

Here are the commands to make the analogous bar plot for the counts in word-initial position.

\[
\text{gh2} = c(0, 0, 7, 0) \\
\text{names(gh2)} = c("_", "f", "g", "k") \\
\text{barplot(gh2, xlab="letter to sound correspondences for 'gh'", ylab="number of tokens of that sound")}
\]

1.12. Exercises

1. **Data and inferences.** Read the following descriptions of datasets gathered to answer some research question. For each dataset, do the following: (a) Identify the big research question and two of the more specific subquestions that are being asked to address the big question. (b) State what the data are by describing what the observations are for one of the two subquestions. (c) Explain what the person who gathers the dataset would have to do with these observations to draw inferences that can relate the answer to the subquestion to the larger research question.

**Dataset 1.** Sue has made a list of 10 common words that she finds difficult to spell. The list includes words such as *immediately, separately, and definitely*. She would like to find out whether these words are difficult to spell for other people her age as well. To do this, she asks all of her classmates to write them down without consulting a dictionary.

**Dataset 2.** Geoffrey is helping his five-year-old twin nieces Joanie and Jill write a letter to their grandfather about a recent trip to the zoo. Joanie writes, “Uncle Jiff took me and Jill to the zoo.” Jill adds, “We saw a baby jiraff.” Geoffrey wonders how Joanie and Jill might spell other words containing the \(dʒ\) sound. (The sound \(dʒ\) is the consonant that is spelled with ‘G’ at the beginning of Geoffrey’s name, but with ‘J’ at the beginning of his nieces’ names.) He gives them each another sheet of paper and asks them to write down the words *juice, jungle gym, giant, fudge, and orange.*

**Dataset 3.** Kim asks teachers at her local primary school to send her the text of all of the books that are used to teach reading there, and assembles a list of all the different words in the books that contain the letter ‘f’ in their spelled forms. She says each word on the list to herself and transcribes its spoken form in the International Phonetic Alphabet. Then she decides which sound or sounds the letter ‘f’ is spelling for each word.

**Dataset 4.** Shangfei is taking Mandarin Chinese classes at his university in order to be able to read his grandfather’s diary and old letters from China that his family
has saved over the years. Shangfei’s mother, who learned English as an adult immigrant to the U.S., tells him that learning to read English was much harder for her, because unlike in written Chinese, homophone pairs in written English are not differentiated. For example, you can’t just look at the written form bank and know whether it is the word for the financial institution or the word for the land at the edge of a river. Shangfei thinks of the homophone pair one versus won and wonders which type is more common. He reports this conversation in class the next day and the teacher is intrigued. She interrupts the Chinese lesson for ten minutes so that he can ask his classmates to write down as many English homophone pairs as they can think of.

2. Types and tokens. Read the following descriptions of subsets of observations in the datasets in 1. For each subset, answer the following questions: (a) What subquestion does this subset of the data address? [If there is more than one subquestion, say so and answer questions (b) through (c) separately for each subquestion.] (b) What is the variable that makes these observations relevant to the subquestion? (c) What are the types of the variable? (d) What are the tokens?

Dataset 1. After collecting her data, Sue finds out the following about the word definitely. Nineteen of her classmates spelled it correctly as definitely, eight spelled it as defitely, five spelled it as definitly and one spelled it as defanitely. Since definitely is a word that Sue spells correctly about half the time (and always spells it definitely when she makes a mistake), she decides to look carefully at all 10 words and see how many classmates gave incorrect spellings for each one.

Dataset 2. After collecting his data, Geoffrey finds out that Joanie spells the words as jus, jongel jim, jiyint, foj, and arinj, whereas Jill spells them as joos, jungljim, jaint, fuge, and orenge.

Dataset 3. Kim counts 879 words containing at least one instance of the letter ‘f’ and notices that many of these words are forms such as fish, sofa, awful, leaf, and of, where the letter ‘f’ occurs just once. However, 98 of these are words such as giraffe, office, and sniff, where there is a sequence of two letters ‘f’ in a row. Another 18 are words such as fifteen and fluff, where the letter ‘f’ occurs more than once, but is separated by other letters. So she counts a total of 879+98+18=995 instances of the letter ‘f’. She looks at what she has transcribed for the pronunciation of the single letter ‘f’ in words such as fish, sofa, awful, leaf, and of, and realizes that in nearly every case, she has pronounced the grapheme ‘f’ as the consonant sound f. The single exception is the word of, where she has pronounced the ‘f’ as the consonant sound v. She also realizes that the pronunciation of the doubled ‘ff’ in words such as giraffe, office, and sniff is the same as the pronunciation of the single ‘f’ in words such as sofa, awful, and leaf, so she decides to treat the 98 instances of ‘ff’ as a sequence of the grapheme ‘f’ in
which the first ‘f’ is pronounced as the consonant sound \( f \) and the second ‘f’ is “silent” (just as the ‘e’ in giraffe is “silent”).

**Dataset 4.** Shangfei’s classmates think of the following homophone pairs:

- bank (edge of a river) : bank (a pile of snow)
- fish : phish
- fish (verb) : fish (noun)
- walk (as in ‘I walk to school.’) : walk (as in ‘I walk the dog.’)
- fit (as in ‘My clothes don’t fit anymore.’) : fit (as in ‘He had a fit when he saw the mess.’)
- read (past tense of ‘to read’) : red
- scent : cent
- sent : cent
- sense : cents
- you : ewe
- bee : be
- feet : feat
- nose : knows
- rest (verb meaning ‘take a breather’) : rest (noun meaning ‘remainder’)
- light (‘not heavy’) : light (as in ‘lamp’)
- meet : meat
- post (verb meaning ‘to mail’) : post (noun for a part of a fence)

**stick** (verb meaning ‘to adhere’) : **stick** (a piece of wood)

**shoo** : **shoe**

3. **Populations and samples.** List which population each dataset in exercises 1 and 2 is intended to represent, and say whether it is a good, representative sample for the research question. Explain why or why not.

4. Use the R code in Section 1.11 to reproduce the bar plots in Figure 1.2.

5. Adapt the R code in Section 1.11 to reproduce the bar plots in Figure 1.1. (Hint: use the numbers provided in Table 1.1.)

6. Adapt the R code in Section 1.11 to plot the distribution of letter-to-sound correspondences for the letter ‘f’ for the two different datasets shown in Table 1.3.

**Table 1.3.** Token counts for the types of letter-to-sound correspondences for the letter ‘f’ in the Moe, Hopkins, and Rush (MHR) and the Children’s Printed Word Database (CPWD).

<table>
<thead>
<tr>
<th>IPA</th>
<th>MHR</th>
<th>CPWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>517</td>
<td>(99.8%)</td>
</tr>
<tr>
<td>( v )</td>
<td>1</td>
<td>(0.2%)</td>
</tr>
<tr>
<td>total</td>
<td>518</td>
<td></td>
</tr>
</tbody>
</table>

7. **Letter to sound correspondences**

Table 1.4 is the list of the 85 words in the Moe, Hopkins, and Rush (1982) word list that contain the sequence of letters ‘gh’ in their spelling. Go through these words and
determine what type of sound or sound sequence the ‘gh’ represents in each word’s spelling. Then answer the following questions.

(a) How many different types are there?

(b) Are all of these types relevant for the variable “letter-to-sound correspondences for the ‘gh’ digraph”? If not, eliminate any that are not relevant.

(c) What is the token count for each type you identified (after eliminating the irrelevant ones)?

(d) Describe the types.

(e) Then divide each set of tokens into the ones that are for ‘gh’ in word-initial position and ones that are in word-medial or word-final position.

(f) Adapt the code in Section 1.11 to make a pair of bar plots showing their token counts, analogous to the plots in Figure 1.2.

---

8. Think of as many different words as you can write down in fifteen minutes that contain the digraph ‘sh’ in their spellings. Try to think of words that have this digraph in each word position: initial, medial, and final. Then answer the following:

a) Count the number of different pronunciations that are represented by that spelling in these words. Then tally up how many instances of the digraph ‘sh’ are tokens of each pronunciation type.

b) Make a bar plot similar to the plots in Figure 1.1 showing the token counts.

---

Table 1.4. Words in the Moe, Hopkins, & Rush (1982) list that contain the sequence ‘gh’

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>afghan</td>
<td>bought</td>
<td>bright</td>
<td>brought</td>
<td>caught</td>
<td>cough</td>
<td></td>
</tr>
<tr>
<td>Cunningh</td>
<td>daughter</td>
<td>daughters</td>
<td>Deffenbaugh</td>
<td>doghouse</td>
<td></td>
<td>dough</td>
</tr>
<tr>
<td>m</td>
<td>eghead</td>
<td>eight</td>
<td>eighteen</td>
<td>eighth</td>
<td>eights</td>
<td>enough</td>
</tr>
<tr>
<td>fight</td>
<td>fought</td>
<td>fighters</td>
<td>frightened</td>
<td>fighting</td>
<td>fights</td>
<td>flashlight</td>
</tr>
<tr>
<td>fought</td>
<td>freighters</td>
<td>heightened</td>
<td>ghost</td>
<td>ghosts</td>
<td>gunfights</td>
<td></td>
</tr>
<tr>
<td>headlight</td>
<td>heights</td>
<td>high</td>
<td>highchair</td>
<td>higher</td>
<td>highest</td>
<td></td>
</tr>
<tr>
<td>highland</td>
<td>highway</td>
<td>knights</td>
<td>laugh</td>
<td>laughed</td>
<td>laughing</td>
<td></td>
</tr>
<tr>
<td>laughs</td>
<td>light</td>
<td>lighted</td>
<td>lighter</td>
<td>lightning</td>
<td>lights</td>
<td></td>
</tr>
<tr>
<td>midnight</td>
<td>midnights</td>
<td>might</td>
<td>mighty</td>
<td>naughty</td>
<td>neighbor</td>
<td></td>
</tr>
<tr>
<td>neighborhood</td>
<td>neighbors</td>
<td>night</td>
<td>nightgown</td>
<td>nights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nights</td>
<td>ought</td>
<td>overnight</td>
<td>Pittsburgh</td>
<td>right</td>
<td>rough</td>
<td></td>
</tr>
<tr>
<td>roughs</td>
<td>sight</td>
<td>sleigh</td>
<td>spaghetti</td>
<td>straight</td>
<td>taught</td>
<td></td>
</tr>
<tr>
<td>though</td>
<td>thought</td>
<td>through</td>
<td>tight</td>
<td>tighter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tonight</td>
<td>tough</td>
<td>oughie</td>
<td>weigh</td>
<td>weight</td>
<td></td>
<td></td>
</tr>
<tr>
<td>weights</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
c) Describe the distribution of letter-to-sound correspondences for ‘sh’. How many different pronunciation types are there for this digraph? What does the distribution look like? For example, is there one value of the letter-to-sound correspondences that is very frequent?

9. **Sound-to-letter correspondences**

Think of as many different words as you can write down in fifteen minutes that contain the consonant sound \(\textsf{ʃ}\). (This is the sound that is spelled with the digraph ‘sh’ in the words *fish* and *phish*, but with the digraph ‘ch’ in the word *chef*.) Then do the analyses in (a) and (b) and write a description to answer the questions in (c):

a) Count the number of different ways that this sound is spelled in these words. Then count the number of instances of the sound \(\textsf{ʃ}\) that are tokens of each sound-to-letter correspondence type.

b) Make a bar plot like the plots in Figure 1.1 showing the token counts.

c) Describe the distribution of sound-to-letter correspondences for the sound \(\textsf{ʃ}\). How many different ways of spelling this sound are there? What does the distribution look like? Is there one very frequent spelling and several much less frequent ones, or are the frequencies more evenly distributed?

10. The spoken forms of the nicknames *Rob* and *Cindy* have fewer syllables than the spoken forms of the corresponding formal names *Robert* and *Cynthia*. The spoken forms of the nicknames *Johnny* and *Annie*, on the other hand, have more syllables than the corresponding formal names *John* and *Ann*. And the spoken forms of the nicknames *Jack* and *Suzie* have exactly the same number of syllables as the corresponding formal names *John* and *Susan*. Download the data file Ch00.Textfile1.txt from the web site for this course at http://hdl.handle.net/1811/77848, or directly at http://kb.osu.edu/dspace/bitstream/handle/1811/77848/Ch00.Textfile1.txt (or recreate the data file using the sample of names and associated nicknames at http://www.cc.kyoto-su.ac.jp/~trobb/nicklist.html). Use this sample to evaluate whether one of the three types of relationship (nickname shorter than formal name versus nickname longer than formal name versus nickname equal in length to the formal name) is a more common pattern. That is, do the following things:

a. Determine what the types are and count the tokens for each type.

b. Make a bar plot showing the count of tokens for each of the types.

c. Write a sentence or two describing the distribution of tokens across the types in the bar plot. Then make an
inference about which (if any) of the patterns is the most typical.

11. Do the same thing that you did in exercise 10, but evaluate the relationships in terms of number of letters in the spelled forms instead of number of syllables in the spoken forms.

12. Many English nicknames, such as Ron, Jack, and Sue, are monosyllabic, meaning that the spoken form contains exactly one syllable. However, many other nicknames, such as Ronnie, Johnny, and Suzie, are longer than one syllable. These three examples of longer nicknames all end with a diminutive suffix i (spelled –y, –i, or –ie) which also occurs in “baby words” such doggy and kitty. Follow the same three steps as in Exercise 10 to use the data file from the preface Ch00.Textfile1.txt to evaluate the hypothesis that nicknames that are longer than one syllable are longer because they end in this diminutive suffix.

1.13. References and data sources

Information about the history of the ghoti puzzle is summarized in the following article:


The counts of Ellis’s letter-to-sound and sound-to-letter correspondences and the calculation of the 1,745,226 different ways to spell scissors are from the following book.


Here are the references for Ladefoged’s book, and the studies we cited.


The numbers in Table 1.1 and Figure 1.1, and the left panel of Table 1.3, are from the Moe, Hopkins, and Rush (1982) wordlist. This wordlist is taken from a study of words that first graders might know. It is a list of 6366 distinct words that were transcribed from recordings of conversations with first graders in Nebraska. The following book described the study and gave the list of words.

Springfield, IL: Thomas.

The numbers in Table 1.2 and Figure 1.2 and the right panel in Table 1.3, are from Version 1.3 of the Children’s Printed Word Database that was developed by Jackie Masterson, Maureen Dixon, Morag Stuart, and Philip Quinlan. You can access the database and query it on the database web site at http://www.essex.ac.uk/psychology/cpwd. The following article describes the database and how it was developed.


The Ch00.Textfile1.txt data file used in Exercises 9-11 is documented in the “References and data sources” section of the preface chapter.