Chapter 4

Counting combinations and permutations

Contents:

4.1. Why are typical words in some languages longer than in others?
4.2. Sets and subsets
4.3. The operators NOT (!=), OR (|), and AND (&)
4.4. Computing the number of different selections for VC and CVC
4.5. Sampling with and without replacement
4.6. Does order matter?
4.7. Understanding independence
4.8. Answering our research question
4.9. Summary
4.10. R code
4.11. Exercises
4.12. References and data sources
4.1. Why are typical words in some languages longer than in others?

In the previous two chapters, we talked about the relationship between a word’s familiarity or predictability and its length. Familiar, everyday words tend to have fewer consonant and vowel sounds than unfamiliar, erudite words. Also, when we say longer words, such as government, personal, and chocolate, in contexts where they might be predictable, we tend to shorten them a bit by reducing some of the vowels and consonants. Or we shorten them more radically, as in the abbreviations cell phone, cell, or phone for cellular telephone and PC for personal computer. Other languages also reduce or abbreviate long words that are used frequently. For example, the French word for ‘cellular telephone’ is téléphone portable, but French speakers usually just say portable. Similarly, the Japanese word for ‘personal computer’ ‘is paasonaru kompyuuta, but most of the time Japanese speakers shorten it to pasokon.

Given this tendency to shorten words when they are used frequently in everyday settings, we might wonder why languages have long everyday words such as government and computer in the first place. Where do long words like this come from? One important source of long words is the process of combining two or more words into compound forms to name new inventions, as in the English compound words cellular telephone and personal computer. The French words for cell phone and laptop computer also are compound words made by combining the adjective portable (meaning ‘portable’) with the nouns téléphone (meaning ‘telephone’) and ordinateur (meaning ‘computer’).

Another important source of long words is the tendency to incorporate words from other languages to name things that are imported from the cultures associated with those languages. For example, the Japanese word paasonaru kompyuutaa is a recent borrowing into the language of the English compound word personal computer. The individual words personal and computer, in turn, are considerably older borrowings into English. These came into the language from French at about the same time that the word government did. Words in French tend to be longer than typical English words, and this was true 400 years ago, as well. The French word that is the source of government was gouvernement, with four syllables. So despite being fairly long, at three syllables, the dictionary form of the English word today is shorter than its original form.

The same is true of the English word chocolate, which is listed in the Merriam-Webster online dictionary as having either two or three syllables. That is, the dictionary form given is tʃɔlɪt, with an optional medial ə vowel.¹ This word was originally borrowed into the English language in the 16th century as a four-syllable word from the Spanish word chocolata. Other long everyday words that came into English from Spanish include alligator, canary, embargo, pimento, and mosquito. The Spanish word chocolata, in turn, was borrowed

into the Spanish language from the Nahuatl word *chocolatl* when the Spanish started importing cocoa beans from Mexico along with the practice of making a hot drink from ground cocoa. Other long everyday words that came into English from Nahuatl via Spanish include *avocado, coyote, ocelot, mesquite,* and *tomato.* Words in Spanish and Nahuatl tend to have more syllables than words of English, so words borrowed into English from these two languages, like words borrowed even earlier from French, usually started out as “big” words by comparison to the typical length of English words. Figure 4.1 illustrates this difference in typical length for Nahuatl versus English.

Saying that these everyday “big” words come from a different language that has longer “small” words, however, raises a more general question. Why are there such differences across languages in the typical length of everyday words? One possible reason that typical word lengths differ is that languages can have different rules about the ways in which consonant sounds and vowel sounds can be combined in sequence to make syllables. For example, the three-syllable Nahuatl word *chocolatl* ends in the laterally released dental stop consonant *tɬ,* but syllables of Spanish cannot end in this stop consonant. So when Spanish speakers borrowed this word into Spanish, they not only substituted their simple dental stop *t* for the difficult Nahuatl consonant *tɬ,* they also added a following a vowel, to make the four-syllable Spanish word *chocolata.* Similarly, the three-syllable English word *personal* ends in the lateral approximant *l,* and Japanese syllables cannot end in this consonant. So when Japanese speakers borrowed this word into Japanese as the first part of the word *paasonaru kompyuutaa,* they not only substituted their simple dental flap sound *ɾ* for the English lateral consonant, they also added a following *u,* to turn it into a four-syllable word.

**Figure 4.1.** Word length measured in number of syllables (left panels) and in number of vowel and consonant sounds (right panels) in a Nahuatl dictionary of 513 words (Tummy, 2004) and in the words in the *Hoosier Mental Lexicon* (Nusbaum, Pisoni, & Davis, 1984) that have the 513 highest frequencies, as estimated by number of occurrences in the Brown Corpus (Kučera & Francis, 1967). Only these highest-frequency English words are included in panels (c) and (d) in order to try to match the type of words that are likely to be included in a dictionary as small as the Nahuatl dictionary that provided the data for panels (a) and (b).
Nettle (1995) suggests another possible explanation for why typical word lengths can differ across languages. Different languages have different numbers of distinct vowel sounds and consonant sounds. For example, the set of Nahuatl sounds includes only 4 different vowel sounds and 15 different consonant sounds, while many dialects of American English have 16 different vowel sounds and 24 different consonant sounds. So, even if the two languages both allowed exactly the same kinds of syllable types, the set of distinct syllables that could be made by combining 4 vowels with 15 consonants is smaller than the set that can be made by combining 16 vowels with 24 consonants. So, words (which are formed from unique combinations of syllables) might need to have more syllables in languages such as Nahuatl, which have fewer different vowels and consonants. If the languages like this have more syllables per word, they will probably be longer than languages that have more sounds, and thus more unique syllables to choose from, and consequently fewer syllables per word.

In this chapter, we will evaluate Nettle’s explanation for differences across languages in typical word length. That is, we will compare Nahuatl and English sound inventories, and we will ask: How many unique syllables can be made in each language? And how many syllables does a word need in each language in order to get a basic set of words? We will do that by calculating the number of distinct syllables that can be made by combining the consonants and vowels of Nahuatl versus the consonants and vowels of English to make the syllable structures that are shared between the two languages. Then we will calculate the number of two- and three-syllable words that can be made by combining that many different syllables into longer sequences of sounds. We’ll start by introducing some technical terms and concepts having to do with set theory. Then, we’ll work through the logic of how to calculate how many distinct ways there are to select a certain number of sounds from the sets of vowel and consonant sounds that Nahuatl and English use to make words, introducing concepts from another branch of mathematics known as combinatorics.

4.2. Sets and subsets

When we talked about the set of Nahuatl sounds, we were using a technical term that implies certain properties. A set is an explicitly-defined group or collection of items. The items can be sounds or people or properties or numbers or things of any kind that can be counted as distinct objects and differentiated from things not in the set by the way the set is defined. The objects in a set are called the members or the elements of the set. A subset is a smaller set that a) includes only some of the members of some larger set and b) does not include any items that do not belong to the larger set.

One way to define a set is by specifying some property that all of the objects that are members of the set share. This excludes objects from membership that do not share the defining property, and includes all objects that share the property. Another way to define a set is by explicitly listing all of its members. When a set is defined by listing its members, the list is usually
enclosed inside of curly brackets, as in (1) for the set of 19 different sounds of Nahuatl.

(1) Nahuatl sounds: \{ i, e, a, o, p, t, k, k^w, ts, tɬ, tf, s, f, m, n, l, w, j, h \}

The set defined by (1) includes all the sounds that are used in the Nahuatl language, and no sounds that are not used in the Nahuatl language. You might notice that even though the set of Nahuatl sounds includes one sound that is not found in English (tɬ), it does not include many of the sounds that are found in English. That is because we have defined this unique set as just being all of the sounds in the Nahuatl language.

In Chapter 3 of *Vowels and Consonants*, Peter Ladefoged defines a vowel sound as “the core of nearly every spoken syllable” and as “any sound occurring in the middle of a syllable, provided that it is produced without any kind of obstruction of the outgoing breath.” Consonant sounds, by contrast, occur at the edges of syllables, and are produced by obstructing the outgoing breath in some way. Using these properties as definitions of vowel and consonant, we can assign each of the sounds in (1) to one or the other of the two subsets defined in (2) and (3).

(2) Nahuatl sounds that are vowels: \{ i, e, a, o \}

(3) Nahuatl sounds that are consonants: \{ p, t, k, k^w, ts, tɬ, tf, s, f, m, n, l, w, j, h \}

Note that while (2) and (3) are subsets of (1), they are also sets in their own right. Remember that a subset is also a set.

The definitions that we used to create the sets of vowels and consonants relies heavily on the idea of a syllable, which must be arranged around a required, or obligatory, vowel. Syllables can take many forms, including (but not limited to): a single vowel (V), as in the English words *I* and *owe*, pronounced *az* and *o*; a single consonant plus a vowel (CV), as in the words *high* and *do*, pronounced *har* and *du*; and even syllables with many consonant sounds, such as (CVCCC) in the words *text* and *fifths*, pronounced *tekst* and *tfθs*. A vowel (or vowel-like) sound is required to form a syllable, with different arrangements of consonants around the vowel. But not all languages have all syllable types. Table 4.1 below shows some of the Nahuatl words in the small Nahuatl-Spanish dictionary that we consulted to make Figure 4.1. If we look closely at these words, we can see that the sounds in (2) and (3) can be selected only in certain sequences to yield Nahuatl words. These sequences are the four different syllable types listed in set (4).

(4) Nahuatl syllable types: \{ V, CV, VC, CVC \}

In this set, the “V” symbol stands for any of the vowel sounds in (2) and the “C” symbol stands for any of the consonant sounds in (3). Thus, the first type of syllable in set (4) is one that selects just a single sound from the subset in (2), as in the *a* which is the first syllable of *ayik*, the middle syllable of *kiawi*, and the last syllable of *altia*. The second and third types in set (4) are selections of one vowel from subset (2) and one consonant from subset (3) in one or the other possible order, as in the CV syllable *fi* in the word *xitini* and the VC syllable...
ble if in the word ixtlawak. And the last type of syllable in set (1) is the three-sound sequence CVC that combines the obligatory vowel with both a preceding initial consonant and a following final consonant, as in both of the syllables nah and wat that combine to make the name of the language. This CVC sequence is the most common syllable type in one-syllable words, and it is the second most common type (after CV) in two-syllable words.

English also allows CVC syllables, and CVC is, in fact, the most common syllable type for one-syllable words of English as well. This is shown in (5), which is the set of syllable types that are allowed in one-syllable words of English, listed in order of frequency in the HML.

(5) Syllable types for 1-syllable words in the Hoosier Mental Lexicon, in order of frequency:

\{ CV, CVC, CCV, CVCV, CV, CCV, CCCV, V, VCC, CVCC, CCCV, CVCVCC, V, VCCC \}

As noted already, the items in a set are called its members or its elements. Different sets can have shared elements. For example, each of the Nahuatl syllable types listed in (4) is also a member of the set of English syllable types listed in (5). Similarly, each of the vowel sounds a and i and each of the consonant sounds j and h is a member of the set of Nahuatl sounds listed in (1) and also a member of the set of American English sounds listed in (6).

(6) American English sounds: \{ i, ɪ, e, ɛ, æ, a, ʌ, o, ʊ, u, œ, ɑ, ɒ, ɔ, ʌ, ʊ, ʌɪ, aɪ, ʊɪ, p, b, t, d, k, ɡ, ʃ, ʤ, f, v, θ, ð, s, z, ʃ, ʒ, m, n, η, l, r, w, j, h \}

When two sets share members in this way, we refer to the subset of shared elements as the intersection of the two sets. Of course, if there are no shared members, the intersection of two sets can be empty. This is the case for the intersection of sets (2) and (3), which are both subsets of set (1). That is, a sound in Nahuatl is either a member of set (2) – i.e., a vowel –
or it is a member of set (3) – i.e., a consonant, but it cannot belong to both sets; therefore, the set of intersecting elements is an empty set.

Subsets (2) and (3) are also an exhaustive list of the sounds in set (1). That is, if we combine sets (2) and (3), we end up with every member of set (1). So, every sound that is a member of (1) and is not a member of (2) is necessarily a member of (3). Conversely, every sound that is a member of (1) and is not a member of (3) is necessarily a member of (2). When two subsets of a larger set share no members (i.e., the intersection is empty) but together constitute the entire membership of the larger set, we call each subset the complement of the other subset.

We defined the vowels in (2) as the subset of Nahuatl sounds that are obligatory and central to the syllable. And we defined the complement of vowels – i.e., consonants, in (3) – as a type of sound that is optional in a syllable, which can occur either before or after the obligatory vowel. Not every consonant sound can begin a syllable, and not every consonant sound can end a syllable. These two positions can define two more subsets of Nahuatl sounds, listed in (7) and (8).

(7) Nahuatl consonants that can begin a syllable: \{p, t, k, kʷ, ts, tʃ, tʃj, s, f, m, n, l, w, j\}

(8) Nahuatl consonants that can end a syllable: \{p, t, k, ts, tʃ, tʃj, s, f, m, n, l, w, h\}

Note that there is considerable overlap between the elements in sets (7) and (8). They have a large intersection. However, there are sounds that are not shared. Specifically, all consonants other than \{h\} can occur at the beginning of a syllable and all consonants other than \{kʷ, j\} can occur at the end.

Set and subset membership can be visualized using a Venn diagram, such as the diagram in Figure 4.2 below. Looking at a Venn diagram can help us visualize the logical operators that define relationships among sets and subsets, such as how we sneaky used the logical operators AND and NOT to define complements and intersections above. These relationships and operators are useful for understanding how to compute the possible ways in which elements can be selected and combined, so let’s review those next.

4.3. The operators NOT (!=), OR (|), and AND (&)

Figure 4.2 shows a Venn diagram of the set of all the 19 Nahuatl sounds listed in (1) and of the different subsets of sounds that are listed in (2), (3), (7), and (8), as well as other subsets that are defined by operations on these sets. Specifically, set (1) – i.e. all Nahuatl sounds – contains all elements enclosed by the largest black outer circle. Within this larger set of sounds, set (2) – i.e., the 4 sounds that are vowels – contains the elements inside the small red circle at the bottom. The consonant sounds in (3) are then the subset of 15 sounds that are NOT in this red circle. If set (2) does NOT contain any ele-
ments that are in set (3), those two subsets are the comple-
ments of each other, as already mentioned. The logical opera-
tor NOT, in this way, helps to define a set that is the comple-
ment of another set. Notice that the logical operator is set in 
all caps, to set it apart from the ordinary conjunction, not, 
which has a different set of meanings than NOT. In R, we can 
select a subset() of items that do NOT have a certain prop-
erty by using the symbols "!=" to mean NOT. For example, in 
a data frame that contains a column of all the sounds in Nahu-
atl, and an additional column containing labels that identify 
each sound as a vowel or a consonant, we could select the sub-
set of sounds that are not vowels (that would leave only the 
subset of consonants) with the following notation:

```r
consonants <- subset(Nahuatl, label != "vowel")
```

The subset of 15 consonants sounds is further divided into two 
subsets. Set (7) – i.e., the 14 consonant sounds that can occur 
at the beginning of the syllable – is the subset of elements in-
side the yellow circle. Set (8) – i.e., the 13 consonant sounds 
that can occur at the end of a syllable – is the subset of ele-
ments inside the light blue circle. All the consonant sounds 
can then be defined as the union of these two subsets – i.e., 
the set of elements that either occur at the beginning of a syll-
ble before the vowel OR that occur at the end of a syllable af-
ter a vowel, or the whole space encompassed by both the yel-
low and light blue circles. In this way, the logical operator OR 
defines a potentially larger set that is the union of two (or 
more) sets. In R, the logical operator OR is written using the 
pipe symbol, "|". So, if we specified a subset of sounds that 
are vowels OR sounds that are consonants in R, the union of 
these two sets would give us a set of all of the sounds of Nahu-
atl, like this:

```r
allsounds <- subset(Nahuatl, label == "vowel" | label == "consonant")
```

The set of all Nahuatl sounds in (1) can also be defined as the 
union of all three subsets that are enclosed by colored circles 
in the Venn diagram in Figure 4.2.

Within the set of 15 consonants, there are 12 that are both 
members of the subset of sounds that can occur at the begin-
ning of a syllable AND members of the subset of sounds that 
can occur at the end of a syllable. As we already said, this is 
known as an intersection of sets, which is the set defined as
all the members of one set, which are also members in another set. That is, this logical operator **AND** defines a potentially smaller set that is the intersection of two (or more) sets. In R, AND is written using the ampersand character “&” if we want to select a subset() that contains only elements that are shared in common across two different sets. The 11 sounds in the section with no shading in the middle of the blue and yellow circles is this intersection of subset (7) and subset (8).

We need the logical operators NOT, OR, and AND in order to define other sets that will help us answer our research question about possible number of syllables and word length. That is, we want to know: What is the set of all possible V syllables that can be made by selecting some vowel from set (2)? And what is the set of all possible VC syllables that can be made by selecting some vowel from set (2) and some consonant from set (8)? And what is the set of all possible CVC syllables that can be made by selecting some consonant from set (7) and some vowel from set (2) and some consonant from set (8)? These three sets are defined and (fully or partially) listed in (9:12).

(9) Set of all possible V syllables: select one element of the set \{i, e, a, o\}

That is, select \{i\} OR select \{e\} OR select \{a\} OR select \{o\}

(10) Set of all possible VC syllables: select one element of the set \{i, e, a, o\} AND select one element of the set \{p, t, k, ts, tʃ, s, l, f, m, n, w, h\}

That is, select \{i\} AND \{p\} OR \{i\} AND \{t\} OR \{i\} AND \{k\} OR ...

[See Table 4.2 for the full listing.]

(11) Set of all possible CV syllables: select one element of the set \{p, t, k, kʷ, ts, tʃ, tf, s, f, m, n, l, w, j\} AND one element of the set \{i, e, a, o\}.

That is, select \{p\} AND \{i\} OR select \{p\} AND \{e\} OR select \{p\} AND \{a\} OR select \{p\} AND \{o\} OR select \{t\} AND \{i\} ... OR select \{j\} AND \{o\}

(12) Set of all possible CVC syllables: select one element of the set \{p, t, k, kʷ, ts, tʃ, s, l, f, m, n, w, j\} AND one element of the set \{p, t, e, a, o\} AND one element of the set \{p, t, k, ts, tʃ, tf, s, l, f, m, n, w, h\}.

That is, select \{p\} AND \{i\} AND \{p\} OR select \{p\} AND \{i\} AND \{t\} OR select \{p\} AND \{i\} AND \{k\} ... OR select \{j\} AND \{o\} AND \{h\}

### 4.4. Computing the number of different selections for VC and CVC

As (9) shows, when a syllable is made by selecting only one element from the set of vowels, the number of possible different sets of selections is exactly the same as the number of elements in the set from which the elements are being selected.
There are 4 different vowel sounds listed in (2) that could be selected to be the obligatory V element in a syllable, and so there are exactly 4 different V syllables listed in (9).

As soon as the selection involves more than one element, however, the number of possible different results rapidly increases. Table 4.2 gives the full listing of syllables for (10). There are 52 different VC syllables listed here.

**Table 4.2.** The 52 possible VC syllables that can be made by selecting one of the four vowels in (5) and one of the 13 consonants in (7).

<table>
<thead>
<tr>
<th>p</th>
<th>t</th>
<th>k</th>
<th>ts</th>
<th>tʃ</th>
<th>s</th>
<th>l</th>
<th>f</th>
<th>m</th>
<th>n</th>
<th>w</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>ip</td>
<td>it</td>
<td>ik</td>
<td>its</td>
<td>itʃ</td>
<td>is</td>
<td>il</td>
<td>if</td>
<td>im</td>
<td>in</td>
<td>iw</td>
</tr>
<tr>
<td>e</td>
<td>ep</td>
<td>et</td>
<td>ek</td>
<td>ets</td>
<td>etʃ</td>
<td>es</td>
<td>el</td>
<td>ef</td>
<td>em</td>
<td>en</td>
<td>ew</td>
</tr>
<tr>
<td>a</td>
<td>ap</td>
<td>at</td>
<td>ak</td>
<td>ats</td>
<td>atʃ</td>
<td>as</td>
<td>al</td>
<td>af</td>
<td>am</td>
<td>an</td>
<td>aw</td>
</tr>
<tr>
<td>o</td>
<td>op</td>
<td>ot</td>
<td>ok</td>
<td>ots</td>
<td>otʃ</td>
<td>os</td>
<td>ol</td>
<td>of</td>
<td>om</td>
<td>on</td>
<td>ow</td>
</tr>
</tbody>
</table>

In theory, we could continue to count the number of possible syllables that can be made by combining the vowels and consonants of Nahuatl simply by listing all of the different sets of selections that can be made for each of the four different syllable types. However, as soon as we go beyond the simplest syllable type, this listing quickly becomes onerous and error-prone. Instead of counting by listing each of the possibilities, we want to be able to compute the number directly from the definitions of the set and the counts of different vowel and consonant types in (2), (7), and (8).

Let’s generalize from the number of different VC syllables. Table 4.2 lists 52 possible Nahuatl syllables of this type. This number is obtained by multiplying the number of rows in the table (for the 4 different sounds that could be selected as the obligatory vowel at the center of the syllable) and the number of columns in the table (for the 13 different sounds that could be selected for the following optional consonant at the end of the syllable). The formula for computing this product is given in (13).

\[
(13) \text{Number of possible VC syllables as defined in (10)} = \text{the number of elements in the set } \{ i, e, a, o \} \\
* \text{the number of elements in the set } \{ p, t, k, ts, tʃ, s, l, f, m, n, w, h \} \\
= 4 \times 13 = 52
\]

Although we do not show the full listing for (11) and (12), we can generalize from the formula for counting the number of VC syllables in Table 4.2 to calculate that there can be 56 different CV syllables, and 728 different CVC syllables in (14) and (15). Again, the number is the product of the number of elements in the set of sounds that could be in each of the positions in the syllable.

\[
(14) \text{Number of possible CV syllables as defined in (11)} = \text{the number of elements in the set } \{ p, t, k, kʷ, ts, tʃ, s, f, m, n, l, w, j \} \\
* \text{the number of elements in the set } \{ i, e, a, o \} \\
= 14 \times 4 = 56
\]
(15) Number of possible CVC syllables as defined in (12) =
the number of elements in the set \{ p, t, k, kʷ, ts, tɬ, tf, s, f, m, n, l, w, j \}
* the number of elements in the set \{ i, e, a, o \}
* the number of elements in the set \{ p, t, k, ts, tɬ, tf, s, l, f, m, n, w, h \}
= 14 * 4 * 13 = 728

The calculation of the number of different possible VC, CV, and CVC syllables in (13), (14), and (15) is relatively straightforward, but before we can talk about combining these different syllable types to make longer words, we need to talk about methods of sampling from a set, the order of items, and the difference between combinations and permutations.

4.5. Sampling with and without replacement
There are two different methods of sampling members of a set: sampling with replacement and sampling without replacement. **Sampling without replacement** means that when selecting an element from a set, that element that you’ve selected is no longer available for sampling in a later selection. Dealing cards in a game of blackjack is a commonly-cited example of sampling without replacement. In this game, the first card to be selected is not replaced in the deck. So the same card value cannot be selected more than once, and the likelihood that the second card will be an ace is different from the likelihood that the first card will be an ace, because there will be one card fewer to select from in the remaining deck after you’ve removed one. **Sampling with replacement** means that when selecting an element from a set, each element can be selected repeatedly. A commonly-cited example of sampling with replacement is rolling a pair of dice in a game of craps. When the shooter in a dice game rolls two dice, the same value can surface on each one. Let’s assume that one of the dice will come to a rest first. The likelihood that the second die will be a • is exactly the same as the likelihood that the first die will be a •, because each die has exactly the same six sides to sample from.

We can also apply this distinction to the selection of the two consonants for CVC syllables or for the word-medial CC sequence that can occur when putting together two CVCs to make a two-syllable CVCCVC word. The selection of consonants for the two different positions in a CVC syllable typically is more like the sampling with replacement in rolling dice. In English, for example, p is selected twice in the word pop and k is selected twice in the word kick. Similarly, in Nahuatl, w is selected twice in the first syllable of the word wiwtla (meaning ‘the next day’) and n is selected twice in the first syllable of the word nantzin (meaning ‘Mother’). This is true also of most languages that have CVC syllables. The selection of the two consonants in a CVC syllable is mathematically more similar to rolling a die twice rather than to dealing the first two cards in a game of blackjack.
The selection of two consonants for a CC sequence, on the other hand, can be like rolling a die twice or it can be like dealing two cards, depending on the language. Some languages allow sequences of two instances of the same consonant in a row in the middle of a word and distinguish that selection from selecting just one consonant at the same position. In Japanese, for example, the word *kanna* (meaning ‘Canna’, the name of a flowering plant) has two instances of *n* in a row in the middle. This makes the word structure be a sequence of a CVC syllable followed by a CV syllable – i.e., with the same structure as the word *kanda* (meaning ‘I bit’) where there is a sequence of *n* followed by *d*. The word *kanna* sounds different from the word *kana* (meaning ‘a letter in the Japanese writing system’) where there is only one *n* in the middle. The single *n* makes the word contain two CV syllables instead of a sequence of CVC plus CV. Judging from words such as *chilli* versus *tzili*, it seems that CC sequences in Nahuatlan are similar to Japanese in this regard. So in describing the set of word medial CC sequences in Japanese or Nahuatlan, we have to allow the possibility of consonants such as *n* or *l* being selected twice for CC clusters, just as we did in describing the possibilities for selecting the two consonants in a CVC syllable. This makes CC clusters in these languages more like rolling a die twice. That is, to compute the number of possible word-medial CC clusters in Nahuatlan, we can simply multiply the number of syllable-final consonants that can be the first C in a CC cluster by the number of syllable-initial consonants that can be the second C in a CC cluster to get $13 \times 14 = 182$.

In English, on the other hand, the two sounds selected for a word medial CC cluster are always different consonants. Remember that we are talking about sounds and not spelling. There are many English words that are spelled with a sequence of the same two letters, but these are not both pronounced. For example, the English word *cannon* is spelled with a sequence of two ‘n’ letters. But the word *cannon* is pronounced just same as the word *canon*, which is written with one ‘n’ – both words are written in IPA as *kænən*. So, despite being spelled with a sequence of two ‘n’s, *cannon* does not have a CC sequence, but the syllable structure CV CVC. In English and languages like English, then, the selection of consonants for CC clusters is more like drawing two cards from a deck. So to compute the number of possible word-medial CC clusters in English, we can start by multiplying the number of different syllable-final consonants (which is 21) by the number of different syllable-initial consonants (which is 23), to get an initial inflated estimate of $21 \times 23 = 483$, but this overestimates the actual number of possible clusters, because it includes cases where the first C and the second C are the same. (See exercise 6 for two ways to compute a more conservative number that does not include these cases.)

### 4.6. Does order matter?

A second important distinction is whether or not order matters in selecting the elements to be combined. For example,
standard draw poker, a hand is a selection of five particular cards from the 52 cards in a deck, but the order in which the five cards were dealt doesn’t make a difference between one poker hand and another. A hand that contains the ace of clubs, the 2 of clubs, the 3 of clubs, the 4 of clubs, and the 5 of clubs is a straight flush whether the ace was dealt before the 2 or after the 2. Because the order in which the cards were dealt does not make a difference, the number of different possible poker hands is considerably smaller than the number of different sequences of cards in dealing. By contrast, the order of the two consonants in a CVC syllable does make a difference. For example, the English CVC words pot and top are not the same word, even though both involve the selection of a p and the selection of a t.

A selection of elements for which order does matter is sometimes called a permutation. A selection of elements for which order does not matter is then called a combination. In Chapter 6, we will talk about how to weed out “duplicates” (selections that are identical except for order) in order to compute the number of different combinations of elements as opposed to the number of different permutations. When sounds are combined to make syllables, and then syllables are combined to make longer words, order does matter, so our focus in this chapter has primarily been on how to compute the number of permutations, which is the formula that we used to calculate the number of possible Nahuatl syllables in (13:15).

Another, more general distinction that can be made is whether the likelihood of selecting one element changes, or stays the same, depending on such things as the identity of other elements that are selected. If the likelihood of selecting each element is the same no matter what other selections are involved, then we can say that the individual elements are independent of each other, or we can say that there is independence in the selection process as a whole. Since the calculation is simpler when the possibilities for different individual elements do not depend on each other, we use this term as the basic term and make a kind of double negative to describe the opposite case. That is, when the selection possibilities of different elements do depend on each other in some way, we say that they are not independent or lack independence instead of that they “are dependent”.

The Venn diagram in Figure 4.2 showed subsets that are based on a selection process that lacks independence. Nahuatl has 19 sounds, but the set of sounds that are available for selection at any position in a word is always a smaller subset of these sounds, because of two aspects of the selection process for syllables that show lack of independence among the elements selected.

First, in selecting from the set of Nahuatl sounds to make a syllable, the set of sounds that can be chosen depends on the number of sounds in the syllable. If it is just one sound, then the set of possible choices is just the four vowel sounds in the red circle at the bottom. If the syllable is more than one sound, then the set of possible choices for the vowel part of
the syllable is again just the four vowel sounds, and the set of possible choices for each of the other sounds does not include the set of four vowel sounds. The lack of independence here is what defines the distinction between vowel sounds in the red circle in the diagram and consonant sounds in the complement to the red circle.

Second, even for syllables that contain exactly two sounds, the number of choices for the consonant sound depends on the order. If the vowel is selected first (in a VC syllable), then neither k nor j can be selected as the consonant. If the vowel is selected second (in a CV syllable), then h cannot be selected as the consonant. The lack of independence here is what defines the distinction between the 13 consonants in the blue circle in the diagram and the 14 consonants in the yellow circle in Figure 4.2.

In every spoken language, the selection of sounds to make words lacks independence in ways that are similar to these aspects of Nahuatl syllable formation. The details can differ dramatically across languages, of course. For example, in English, the set of consonants that can be selected for a CV syllable includes h but the set of sounds that can be selected for a VC syllable does not include h. That is, English h shows exactly the opposite pattern to Nahuatl h. It is this kind of lack of independence that prompted us to first talk about syllables before talking about how sounds can combine to make words. Let’s talk about that next.

4.8. Answering our research question

We calculated all the permutations of the four different syllable types in Nahuatl in (9), (13), (14), and (15). We can now add the numbers for the four different syllable types together to get the total number of different possible syllables of Nahuatl, as shown in (16).

\[(16) \text{Number of different Nahuatl syllables} = \text{number of V} + \text{number of CV} + \text{number of VC} + \text{number of CVC} = 4 + 56 + 52 + 728 = 840\]

That’s how many different possible one-syllable word forms there are. For comparison, we can use the set of English sounds in (6) to compute the number of different possible one-syllable word forms there would be in English, if English did not allow other syllable types such as CCV and CVCC. This count is shown in (17).

\[(17) \text{Number of different English syllables if English did not have more syllable types than Nahuatl} = \text{number of V} + \text{number of CV} + \text{number of VC} + \text{number of CVC} = 16 + (23 \times 16) + (16 \times 21) + (23 \times 16 \times 21) = 8448\]

We encourage you to go back and look at the sounds in (6) to be able to explain the numbers that are multiplied to get the number of CV types, the number of VC types, and the number of CVC types for English. That is, determine which 16 sounds in (6) are vowels and which 24 sounds are consonants, and then determine which 23 of the consonants can begin a syllable, and which 21 of the consonants can end a syllable. (Hint: It might be easier to find the one consonant that cannot begin
a syllable and the two consonants other than h that cannot end a syllable.)

Once we have the number of different syllable types, we can also compute the number of different two- and three-syllable words that can be made by selecting any two or any three of the syllables in a particular order. We will assume for now that selecting syllables is sampling with replacement – i.e., that the same syllable can be selected twice, as in the English words mama and tutu. The computations for Nahuatl are shown in (18).

(18) Number of different possible 1-syllable words in Nahuatl: 840
   Number of different possible 2-syllable words in Nahuatl: $840 \times 840 = 705,600$

Adding these to the number of different three-syllable words, we get the total count in (19), which is how many words there could be if Nahuatl words were no longer than three syllables.

(19) Number of different possible 3-syllable words: $840 \times 840 \times 840 = 592,704,000$
   Number of 1-, 2-, and 3-syllable words: $706,400 + 592,704,000 = 593,410,400$

If we guess that there might need to be as many as a million words to name all of the events and objects that have been relevant in the history of Nahuatl-speaking cultures, the set of words that could be made of just one or two syllables is not quite large enough, but the set of words with one, two, or three syllables is far more than large enough.

Applying the same reasoning to English, we can compute the number of different one- and two-syllable words that would be possible even if English did not have many more syllable types than Nahuatl. And we can compute the number of one-, two-, and three-syllable words that would be possible. These calculations are given in (20) and (21).

(20) Number of possible 1- and 2-syllable words in English:
   $8448 + 8448 \times 8448 = 8448 + 84482 = 71,377,152$

(21) Number of 1-, 2- and 3-syllable words:
   $8448 + 8448 \times 8448 + 8448 \times 8448 \times 8448 = 8448 + 84482 + 84483 = 602,994,188,544$

The numbers in (20) suggest that an upper length of two syllables would be long enough for English words, given the number of different vowel and consonant sounds in the language, particularly if we can include the other possible syllable types in the computation. Of course, the number of possible one- and two-syllable words that we calculated in (20) is far larger than the number of words in the Hoosier Mental Lexicon. Similarly, the number of one-, two-, and three-syllable words that we calculated in (19) is far larger than the 531 Nahuatl words that we included in making the histograms in the top row of panels in Figure 4.1. Still the much smaller numbers in (18) relative to the numbers in (20) suggests that the difference in typical word lengths that we saw in the top ver-

Note that these calculations overestimate the number by not taking into account the fact that the two consonants in a word-medial CC cluster have to be different consonants. See exercise 6 for a way to compute a more conservative number.
sus bottom row of panels in Figure 4.1 may be due in part to the smaller number of vowels and consonants in Nahuatl relative to English, just as Nettle (1995) suggested.

### 4.9. Summary

In describing and counting different sound patterns in a language, we can use concepts that have been developed extensively in calculating outcomes in games of chance. We explored concepts from set theory to be able to group related items and explore the interactions of logical operators with different kinds of subsets. These sets and subsets allowed us to list sounds based on certain properties, such as the set of English consonants which is a subset of the set of English sounds. We also looked at sampling with and without replacement and saw that both kinds of sampling can occur within syllables, with the two consonants of CVC syllables representing sampling with replacement, but with medial CC sequences in English representing sampling without replacement (even though we ignored this constraint in our calculations). We introduced the concepts of combinations and permutations; although we focused on permutations with replacement in our calculations. One of the most important concepts that we introduced is independence. For example, in counting the number of different syllables that are possible in a language, given the number of sounds in the language, we have to take into account the ways in which the choice of sound is not independent of position. In counting the number of words that could be possible for Nahuatl and English, we operated under the assumption that syllables are independent, but this is not actually true in any language, and we explained just a few reasons why this is the case.

#### Summary of key terms:

**Combination:** a set or subset of items, in any order.

**Complement:** given a subset A of a larger set B, the set that includes all of the members of B that are not in A and no other members.

**Elements:** items belonging to a set.

**Independence:** when the number of choices in selecting some elements does not change depending on the selection of other elements.

**Intersection:** given two sets, the subset of members of the first set that are also members of the second set.

**Members:** items belonging to a set; same as elements.

**Permutation:** a set or subset of items in which the order of the elements matters.
**Sequence:** an ordered set of items.

**Set:** a group or collection of items or numbers.

**Subset:** a smaller set that includes only some of the members (elements) of the larger set, and no items that are not members (elements) of the larger set.

**Union:** the set defined by combining all of the elements of two sets.

**Without replacement:** each of the items can only be selected once.

**With replacement:** being able to select the same item multiple times.

### 4.10 R code
**4.10. R code**

A copy of this R note is saved inside a script that is called `Chapter04.R` and can be found in the list on the course website, at [http://hdl.handle.net/1811/77848](http://hdl.handle.net/1811/77848), or individually at [http://kb.osu.edu/dspace/bitstream/handle/1811/77848/Chapter04.R](http://kb.osu.edu/dspace/bitstream/handle/1811/77848/Chapter04.R).

**Part 1.**

One new thing that you absolutely need for the computations in this chapter is `^` which is the exponent operator in R. This operator returns the value that is the number to its left to the power on its right. That is, typing the following:

```R
8448^2
```

in the R console returns the same results as typing the following:

```R
8448*8848
```

**Part 2.**

We also introduced the operators NOT (`!`), OR (`|`), and AND (`&`). To see how these operators work with the `subset()` function that we introduced in chapter 2, we will first read in a data frame with some information about Nahuatl vowels. In chapter 2 we used the `read.delim()` function to read in a data frame that is saved somewhere on your computer as a tab-delimited text file. Here, we will create a small data frame using several vectors and the `data.frame()` function. The first three lines of the R code below define three vectors, one listing all Nahuatl vowels, one listing the vowel height for each of the Nahuatl vowels, and one listing the backness of each of the vowels. The fourth line of R code creates a data frame by listing all three vectors as arguments of the `data.frame()` function and assigning the name `vowel_dataframe` to this data frame.

```R
Nahuatl_vowels = c("i", "e", "a", "o")
height = c("high", "mid", "low", "mid")
backness = c("front", "front", "central", "back")
vowel_dataframe = data.frame(
    Nahuatl_vowels, height, backness)
```

If you now type `vowel_dataframe` into the R console, you can see what the data frame looks like.

```R
Nahuatl_vowels height backness
1 i high front
```
Now we can use the NOT (!=) operator to make a subset of all the vowels that are NOT front vowels and call this subset `not_front`. For the `subset()` command, we first specify the data frame from which you want to create a subset, in this case the `vowel_dataframe` data frame that we created above. Then you need to list criteria that R should use to create a subset. Here, we want to exclude all the vowels that are front vowels. The information about whether a vowel is a front vowel or not is in the `backness` vector that we created above. So, we want to exclude all the Nahuatl vowels that are associated with the type "front" in the `backness` vector. We do this by providing the name of the vector, then the NOT operator, and then the type that we want to exclude (`backness != "front"`). Altogether, the command looks like this:

```
not_front = subset(vowel_dataframe, backness != "front")
```

Typing `not_front` into the R console shows the subset of vowels that are not front vowels:

<table>
<thead>
<tr>
<th>Nahuatl_vowels</th>
<th>height</th>
<th>backness</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 a</td>
<td>low</td>
<td>central</td>
</tr>
<tr>
<td>4 o</td>
<td>mid</td>
<td>back</td>
</tr>
</tbody>
</table>

Similarly, we can use the OR (|) operator to make a subset of all the vowels that are either mid vowels OR front vowels, like this:

```
mid_or_front = subset(vowel_dataframe, height == "mid" | backness == "front")
```

Typing `mid_or_front` into the R console shows the subset of vowels that are mid vowels or front vowels:

```
Nahuatl_vowels  height backness
1 i             high   front
2 e             mid    front
4 o             mid    back
```

As you can see by the result, all the rows are selected that satisfy either of the conditions on each side of the '|' operator. The AND (&) operator works similarly to the OR (|) operator, except it only chooses rows that meet both criteria on each side of the '&' operator. So if we replace '|' with '&' in the `subset()` command above, we get a subset of all the vowels that are both mid vowels AND front vowels, like this:

```
mid_and_front = subset(vowel_dataframe, height == "mid" & backness == "front")
```

Typing `mid_and_front` into the R console shows the subset of vowels that are mid vowels and front vowels:

```
Nahuatl_vowels  height backness
2 e             mid    front
```

136
Part 3.

We also talked about permutations and combinations in this chapter. Let’s look at calculation these in R. First, create a vector of consonant sounds that can begin a syllable in Nahuatl:

```
initial_consonants=c("p","t","k","kw","ts","tl","tS","s","S","m","n","l","w","j")
```

Then create a vector of vowels:

```
vowels=c("i","e","a","o")
```

Using the `expand.grid()` function will allow us to select (with replacement) all the different permutations of one C and one V from each of the vectors:

```
CVs=expand.grid(initial_consonants,vowels)
```

Which we can paste together in a third column, using this:

```
CVs$CV<-paste(CVs$Var1,CVs$Var2,sep="")
```

Then, typing the variable name CVs into the console will yield the list of all possible CVs:

```
CVs
     Var1 Var2
 1     p    i
 2     t    i
 3     k    i
 4    kw    i
 5    ts    i
 6    tl    i
 7    tS    i
 8     s    i
 9     S    i
10    m    i
11    n    i
12    l    i
13    w    i
14    j    i
15    p    e
16    t    e
17    k    e
18   kw    e
19   ts    e
20   tl    e
21   tS    e
22    s    e
23    S    e
24    m    e
25    n    e
26    l    e
27    w    e
28    j    e
29    p    a
30    t    a
31    k    a
32   kw    a
33   ts    a
34   tl    a
35   tS    a
36    s    a
37    S    a
38    m    a
39    n    a
```
There are 56 of them, just as we calculated.

Now we can combine these CVs with all the possible final consonants, by creating a vector of final consonants:

    final_consonants=c("p","t","k","ts","tl", "tS","s","S","m","n","l","w","h")

And using the the `expand.grid()` function to calculate how many CVC permutations there are of the CV column with the vector of final consonants:

    CVCs=expand.grid(CVs$CV,final_consonants)

If we query the number of dimensions of this new CVCs data frame, we will see that there are 728 permutations:

    dim(CVCs)

[1] 728  2

**Part 4.**

To calculate more complex combinations and permutations, we need to install the `gtools` package in R. To install a package, on a PC, click on “Packages” in the drop-down menu above the R console and choose “Install package(s)”. A pop-up window that asks you to choose a CRAN mirror will open. Choose a mirror that is near you. The next pop-up window that opens lists all the packages. Select the `gtools` package and click “OK”. R will now download the package. On an Apple device, click on “Packages & Data” and select “Package Installer” from the drop-down menu. The pop-up window will have a search field, into which you should type `gtools`. Alternatively, on either type of device, you can type directly into the console:

    install.packages(gtools)

You will then be prompted to pick a CRAN mirror site close to you.

    --- Please select a CRAN mirror for use in this session ---
Select a site close to you. (It goes a little faster if you pick a closer site, but you could arbitrarily pick any location.) R will download the package into its library. Then, in order to use the package, you need to tell R to load it, using the `require()` function:

```r
require(gtools)
```

To calculate how many possible selections of some number of elements from some set there are, we use the `permutations()` function if order matters. For example, if we wanted to calculate how many possible permutations of the numbers 1 through 6 there are when rolling a die twice, we use the `permutations()` function, like this:

```r
permutations(6, 2, repeats.allowed=TRUE)
```

The first argument of the `permutations()` function gives the number of elements in the set that we are sampling from. In this case, the set has six elements: the numbers 1 through 6 that are on a standard die. The second argument provides the number of elements that should be selected (sampled) from the set. In this case, we are selecting two elements, one for each roll of the die. The `repeats.allowed=TRUE` part of the function states that we are sampling with replacement. Typing the formula above into the R console yields the following result:

```
[,1] [,2]
[1,] 1 1
[2,] 1 2
[3,] 1 3
[4,] 1 4
[5,] 1 5
[6,] 1 6
[7,] 2 1
[8,] 2 2
[9,] 2 3
[10,] 2 4
[11,] 2 5
[12,] 2 6
[13,] 3 1
[14,] 3 2
[15,] 3 3
[16,] 3 4
[17,] 3 5
[18,] 3 6
[19,] 4 1
[20,] 4 2
[21,] 4 3
[22,] 4 4
[23,] 4 5
[24,] 4 6
[25,] 5 1
[26,] 5 2
[27,] 5 3
[28,] 5 4
[29,] 5 5
[30,] 5 6
[31,] 6 1
[32,] 6 2
[33,] 6 3
[34,] 6 4
```

139
If we look at the last row of the R output, the number in square brackets shows us that there are 36 permutations for rolling a die twice. To calculate how many permutations there are for drawing two cards out of a set of five cards, we could type the following into the R console:

```r
permutations(5,2,repeats.allowed=FALSE)
```

Here, the number of elements in the set that we are sampling from is 5 since we are drawing from a set of five cards. We are selecting two elements from the set of five cards. The `repeats.allowed=FALSE` part of the function states that we are sampling without replacement. That is, if the first card has been drawn, it cannot be drawn again. Typing the formula above into the R console yields the following result:

```
[,1] [,2]
[1,] 1 2
[2,] 1 3
[3,] 1 4
[4,] 1 5
[5,] 2 1
[6,] 2 3
[7,] 2 4
[8,] 2 5
[9,] 3 1
[10,] 3 2
[11,] 3 4
```

Again, the number in square bracket in the last row shows us the number of permutations, which is 20 in this example.

To calculate how many possible selections of some number of elements from some set there are, if order does not matter, we use the `combinations()` function. For example, even though there are two ways of getting a ‘1’ and a ‘2’, that is, (1,2) or (2,1), that could count as one combination, i.e., getting a ‘1’ and ‘2’ in any order. This function works just like the `permutations()` function. For example, if we wanted to calculate how many possible combinations of the numbers 1 through 6 there are when rolling a die twice, we use the `combinations()` function, like this:

```r
combinations(6,2,repeats.allowed=TRUE)
```

The output from the R console tells us that there are 21 possible combinations of rolling a die twice:
To calculate how many combinations there are for drawing two cards out of a set of five cards, we could type the following into the R console:

```r
combinations(5,2)
```

Typing the formula above into the R console tells us that there are 10 possible combinations of drawing two cards from a set of five:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Now that you are comfortable using the combinations and permutations functions, we can look at an easier way to do the calculations for finding the number of syllables in Rotokas. Remember that picking a C and a V is a kind of permutation, with replacements allowed.

```
permutations(6,2)
```

4.11. Exercises

1. Geoffrey’s nieces Joanie and Jill have found a deck of playing cards and a large bag of dice in their grandmother’s cabinet of puzzles and toys, and they are making up rules for new games to play that involve drawing...
cards from some “small deck” that includes only a subset of the 52 cards. For their games, the twins have made a pair of game boards that look like this:

<table>
<thead>
<tr>
<th>Joanie’s board</th>
<th>Jill’s Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>a   b   c</td>
<td>A   B   C</td>
</tr>
</tbody>
</table>

For each of the following games, say whether the selection is sampling with replacement or sampling without replacement, and then compute the number of different possible orders of cards or dice on a board.

a. The “up or down” game, version 1. The twins make a “small deck” of 12 cards by taking just the ace, 2, 3, 4, 5, and 6 of hearts (which Joanie chooses as her suit) and the ace, 2, 3, 4, 5, and 6 of diamonds (which Jill chooses as her suit). To play the game, one twin shuffles the deck and then draws cards from the top of the deck. If a card is a heart, she places it on Joanie’s board, with the first heart that surfaces going in space a, the second in space b, and the third in space c. If the card is a diamond, she places it on Jill’s board, again putting the cards down in the slots from A to C in the order that they are drawn. She continues to draw until both boards are filled. The winner is the twin whose board was filled first, unless the other twin’s board shows a perfect run up or down – i.e., if her board shows some sequence of three consecutive numbers in either order, such as { ace, 2, 3} or { 5, 4, 3 }.

b. The “up or down” game, version 2. This version again uses the “small deck” of 12 cards, but it also uses 6 dice. The twins take turns shuffling the 12 cards and drawing just the top card. If the card is a heart, Joanie puts a die in the next open space on her board, with the die surface up that matches the number on the card. If the card is a diamond, Jill puts a die in the next open space on her board, with the die surface up that matches the card. Again, play continues until both boards are filled. The winner is the twin whose board was filled first, unless the other twin’s board shows a sequence that is strictly increasing (i.e., the number on the die in the A space is less than the number on the die in the B space, which in turn is less than the number on the die in the C space) or a sequence that is strictly decreasing (i.e., the number on the A die > the number on the B die > the number on the C die).

c. The “evensies” (iVENZIZ) game. Joanie takes all of the hearts to make a “hearts deck” and Jill takes all of the diamonds to make a “diamonds deck”. To play, each twin shuffles her deck of 13 cards and they then draw cards simultaneously from the tops of their decks. If the card that one of them draws is an even number
(counting the queen as 12), then she places the card in the next open space on her game board, starting from space $A$. They continue to draw until both boards are filled. Again, the winner is the twin whose board was filled first, unless the other twin’s board shows a sequence that is strictly increasing.

d. The “high or low” game. The twins combine Joanie’s “hearts deck” and Jill’s “diamonds deck” and then remove the two aces, to make a “red cards deck” of 24 cards. To play, one twin shuffles the deck and then places the deck between the two boards and begins to draw cards from the top of the deck. If the card is a 2, 3, 4, 5, 6, or 7, it goes on Joanie’s board. If it is an 8, 9, 10, jack, queen, or king, it goes on Jill’s board. Again, the winner is the twin whose board was filled first, unless the other twin’s board shows three cards from the same suit.

2. Try to explain the numbers in formula (20). For example, why is the number of possible CV syllables 23 * 16? Hint: Think about how the set of English sounds in (6) would be grouped into subsets, if you applied definitions that are analogous to the ones that we used to group the Nahuatl sounds in (1) to make the subsets shown in Figure 4.2. In particular, think about which of the sounds in (6) are vowels and which are consonants and then think about where the following four consonants can occur: $ŋ$ (as in ring), $w$ (as in were), $j$ (as in year), and $h$ (as in her).

3. Table 4.3 lists a selection of English 1-syllable words with the syllable types CCV, CCVC, and CCVCC in regular English spelling. Use the words in the table to make a list of the word-initial CC clusters that exist in English for 1-syllable words. Use the IPA to write down the word-initial CC clusters. Then use your list to answer the following questions.

a. Is the selection of the first two consonants for these syllable types independent? Why or why not?

b. Which set of consonants can be selected for the first C and for the second C of the word-initial CC cluster? Is the selection of the second C independent of what has been selected for the first C? Explain why or why not.

c. Based on the restrictions that you have discovered for word-initial CC clusters, calculate how many possible 1-syllable words with the syllable type CCV could exist in the English language.

d. The English language also has 1-syllable words with the syllable types CCCV, CCCVC, and CCCVCC, which differ from the syllable types of the words in Table 4.3 in that they begin with a CCC cluster rather than a CC cluster. English allows only five such word-initial CCC clusters in 1-syllable words. Which CCC clusters does English allow? Provide an example 1-syllable word for
each cluster. Are the sounds that can be selected for
the second and third consonant in CCC clusters a
proper subset of the sounds that can be selected for
CC clusters? Explain why or why not.

Table 4.3. Some English 1-syllable words with the syllable
types CCV, CCVC, and CCVCC in regular English spelling.

<table>
<thead>
<tr>
<th>truck</th>
<th>black</th>
<th>cliff</th>
<th>blotch</th>
<th>flee</th>
<th>skunk</th>
<th>shrug</th>
</tr>
</thead>
<tbody>
<tr>
<td>twelve</td>
<td>spine</td>
<td>step</td>
<td>steak</td>
<td>train</td>
<td>plump</td>
<td>press</td>
</tr>
<tr>
<td>drool</td>
<td>crutch</td>
<td>dress</td>
<td>crest</td>
<td>brag</td>
<td>square</td>
<td>stance</td>
</tr>
<tr>
<td>scarce</td>
<td>smell</td>
<td>grip</td>
<td>school</td>
<td>please</td>
<td>swag</td>
<td>swoon</td>
</tr>
<tr>
<td>brisk</td>
<td>clock</td>
<td>blanch</td>
<td>drum</td>
<td>clean</td>
<td>thrash</td>
<td>sweet</td>
</tr>
<tr>
<td>sled</td>
<td>space</td>
<td>shred</td>
<td>twirl</td>
<td>sketch</td>
<td>smack</td>
<td>spud</td>
</tr>
<tr>
<td>crook</td>
<td>clam</td>
<td>crown</td>
<td>clasp</td>
<td>dwell</td>
<td>plague</td>
<td>green</td>
</tr>
<tr>
<td>stiff</td>
<td>France</td>
<td>phrase</td>
<td>gloom</td>
<td>grasp</td>
<td>spell</td>
<td>snark</td>
</tr>
<tr>
<td>bleak</td>
<td>flood</td>
<td>crab</td>
<td>breeds</td>
<td>blind</td>
<td>thrill</td>
<td>suite</td>
</tr>
<tr>
<td>skate</td>
<td>slime</td>
<td>slug</td>
<td>shrimp</td>
<td>sponge</td>
<td>swirl</td>
<td>thrust</td>
</tr>
<tr>
<td>flute</td>
<td>brine</td>
<td>blast</td>
<td>draft</td>
<td>drink</td>
<td>slang</td>
<td>pledge</td>
</tr>
<tr>
<td>grade</td>
<td>quite</td>
<td>phlox</td>
<td>free</td>
<td>Gwen</td>
<td>fruit</td>
<td>quack</td>
</tr>
<tr>
<td>stop</td>
<td>skimp</td>
<td>scum</td>
<td>sniff</td>
<td>slob</td>
<td>snooze</td>
<td>scotch</td>
</tr>
<tr>
<td>branch</td>
<td>creek</td>
<td>flame</td>
<td>brace</td>
<td>cringe</td>
<td>suede</td>
<td>frilled</td>
</tr>
<tr>
<td>flinch</td>
<td>trip</td>
<td>Christ</td>
<td>clown</td>
<td>broom</td>
<td>treat</td>
<td>groove</td>
</tr>
<tr>
<td>frog</td>
<td>glass</td>
<td>trunk</td>
<td>print</td>
<td>queen</td>
<td>crane</td>
<td>pram</td>
</tr>
</tbody>
</table>

4. Figure 4.3 shows the distribution of word lengths in a
5501-word dictionary of Rotokas, a language of Papua
New Guinea that has only the 5 vowel sounds and 6 con-
sonant sounds in set (22) below, and only the two sylla-
bble types in set (23).

(22) Vowels and consonants of Rotokas: \{ i, e, a, o, u, p, t, k, v, r, g \}

(23) Syllable types allowed in Rotokas: \{ V, CV \}

For comparison, the bottom panels of the figure show
the distribution of word lengths of the 5501 most fre-
quent words in the Hoosier Mental Lexicon.

a. Calculate the number of one-, two-, three-, four-, and
five-syllable words that could be made by combing the
sounds in (22) to make the syllable types in (23).
Then calculate the number of one-, two-, and three-
syllable words that would be possible in English, if
English only had the same two syllable types as Roto-
kas.

b. Describe the distribution of word lengths in Figure
4.3, focusing on the differences that you see between
panels (a) and (c). Then write a sentence saying
whether the differences that you see are what you
would expect given your calculations in a.
Figure 4.3. Word length measured in number of syllables (left panels) and in number of vowel and consonant sounds (right panels) in a Rotokas dictionary of 5501 words (Firchow & Firchow, 2008) and in the words in the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis, 1984) that have the 5501 highest frequencies, as estimated by number of occurrences in the Brown Corpus (Kučera & Francis, 1967).

5. Figure 4.4 shows the distribution of word lengths in a database of 1972 Ju|’hoansi roots (Miller-Ockhuizen, 2003), a language spoken in northeastern Namibia and northwestern Botswana that has more than 100 vowel and consonant sounds. For comparison, the bottom panels of the figure show the distribution of word lengths of the 1972 most frequent words in the Hoosier Mental Lexicon.

In particular, Ju|’hoansi has approximately 40 consonants, 48 click consonants and over 30 vowels. The vowels in Ju|’hoansi are i, e, a, o, u and some of these vowels can occur in short and long, may be nasalized, murmured, glottalized, pharyngealized, or epiglottalized. Ju|’hoansi also has vowel combinations and diphthongs. Not all consonants in Ju|’hoansi can combine with all of the vowels. The situation is a little complicated, so we will simplify it (considerably) for our purposes. We will assume that there are exactly 40 non-click consonants, 48 click consonants and 36 vowels. We will also assume that 19 of the non-click consonants can combine with all of the vowels and 21 of them can combine with only 22 of the vowels. And we will assume that 12 of the 48 click consonants can combine with all of the vowels and 36 of them can combine with only 22 of the vowels. Finally, we will assume that the only syllable type in Ju|’hoansi is CV.

a. Calculate the number of one-, two-, and three-syllable words that could be made with our simplifying assumptions above. Then calculate the number of one-, two-, and three-syllable words that would be possible in English, if English only allowed CV syllables.

b. Describe the distribution of word lengths in Figure 4.4, focusing on the differences that you see between
panels (a) and (c). Then write a sentence saying whether the differences that you see are what you would expect given your calculations in (a).

Figure 4.4. Word length measured in number of syllables (left panels) and in number of vowel and consonant sounds (right panels) in a database of 1972 Ju|’hoansi roots (Miller-Ockhuizen, 2003) and in the words in the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis, 1984) that have the 1972 highest frequencies, as estimated by number of occurrences in the Brown Corpus (Kučera & Francis, 1967).

6. In section 4.5 we pointed out that in Japanese, the two consonants in a word-medial CC cluster can be the same consonant, so that the CVCV word *kanna* differs from CVCCV word *kanda*. In English, on the other hand, the two consonants in a word-medial CC cluster have to be different consonants, so that the CVCV word *canny* (pronounced *kaeni*) differs from the CVCCV word *candy* (pronounced *kaendi*), but *kaenni* is not a possible CVCCV word of English. In this exercise, we will explore the consequences of this fact for the number of different possible CVCCV words of English and for the number of all two-syllable words if V, VC, CV, and CVC were the only syllable types in English.

a. Here is the list of English vowel and consonant sounds, repeated from (6):

(6) American English sounds: \{ i, ɪ, e, ɛ, ə, a, ʌ, o, ʊ, u, ʊ, ə, -æ, aɪ, aʊ, ð, ş, z, ʒ, m, n, ŋ, l, r, w, j, h \}

Make a Venn diagram like the one in Figure 4.2, showing the following subsets of the set of English sounds:

A) the 16 vowel sounds (comparable to the red circle in Figure 4.2)

B) the 24 consonant sounds (comparable to the union of the yellow and blue circles)

and the following subsets of B:

C) the 23 consonant sounds that can begin a syllable (the yellow circle)
D) the 21 consonant sounds that can end a syllable (the blue circle)

E) the x consonant sounds that are in the intersection of C and D

F) the y consonant sounds that can only begin a syllable

G) the z consonant sounds that can only end a syllable

b. We suggested in section 4.5 that we could calculate an initial estimate for the number of possible word-medial CC clusters by multiplying these two numbers:

\[ 21 \text{ (number of sounds in D) } \times 23 \text{ (number of sounds in C) } = 483 \]

Write a formula that begins with that inflated estimate of the number and then subtracts the number of CC sequences where the two consonants are the same sound. (Hint: use the number of sounds in subset E.)

c. Write another formula that calculates the same number as in b. by combining the counts of the following two types of CC sequences:

- clusters where the 1st consonant is an element of subset G and the 2nd consonant is an element of subset C
- clusters where the 1st consonant is an element of subset E and the 2nd consonant is either a different element of E or an element of F

d. Adapting the strategy in part c, write a formula that calculates the number of two-syllable words that can be made by stringing together syllables of the four types \{ V, CV, VC, CVC \}. That is, the formula should add up the counts for all of the following types of two-syllable sequences:

- V or CV followed by any of the 4 syllable types
- VC or CVC where the final C is a member of subset G followed by any of the 4 types
- VC or CVC where the final C is a member of subset E followed by V or VC
- VC or CVC where the final C is a member of subset E followed by CV or CVC where the initial consonant is different from the final consonant in the first syllable

4.12. References and data sources


