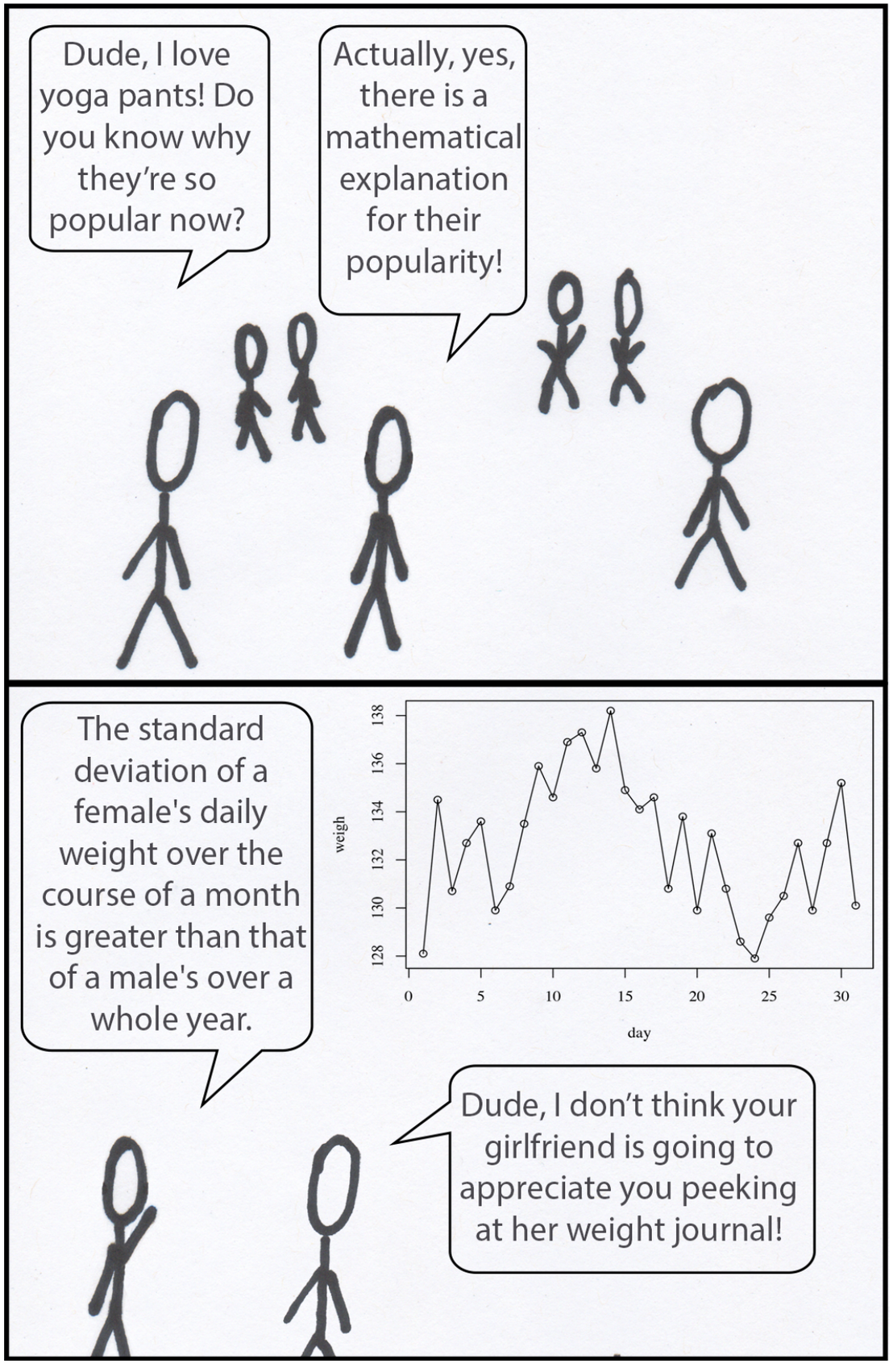
Chapter 9

Measures of dispersion

Contents:

9.1. Life's a *beach* where *sheet* happens
9.2. Relating variability to sources of deviation from the typical value
9.3. Measures of dispersion - Interquartile range
9.4. Using box plots to display quartiles
9.5. Measures of dispersion - Standard deviation
9.6. Robustness
9.7. Developing an *a priori* model of random variation
9.8. Using box plots to explore possibly non-random sources of spread
9.9. Answering the research question
9.10. Summary
9.11. R code
9.12. Exercises
9.13. References and data sources
Appendix 9A. Step-by-step through the IQR and standard deviation
Appendix 9B. Filled pause



9.1. Life's a *beach* where *sheet* happens. (Or, what counts as a "short" vowel?)

One of the oddities of English is the slightly reduced quality of short vowels such as **ʌ** and **ɪ**, even in careful speech. These **ʌ** and **ɪ** vowels are distinctive. They distinguish the words *cut* and *bit* from the words *caught*, *cot*, and *beat*, which contain long vowels **ɔː**, **ɑː**, and **iː**. Historically, the short vowels **ʌ** and **ɪ** were just shorter versions of the corresponding long vowels. That is, the difference was just a matter of duration. But today **ʌ** and **ɪ** also have a somewhat centralized quality when compared to their long counterparts. That is, when compared to the sounds that we transcribe with **ɔː** and **ɑ***ː* in *caught* and *cot*, the sound that we transcribe with **ʌ** in *cut* is a bit closer in quality to the sound of the reduced **ə** vowel in words such as *a* and *the*, and the same is true for the sound **ɪ** in *bit* when compared to the sound **iː** in *beat*. Look at the IPA chart of the vowels in Figure 0.2 in Chapter 0 to visualize what we mean by *centralized* when we compare these vowels.

Many languages use vowel duration distinctively. For example, in Japanese, the word **kita** with a short **i** sound in its first syllable means *north*, and contrasts with the word **kiːta** with a long **iː** in the same place, which means *heard*. There are fewer languages that use the vowel quality contrast between **ɪ** and **i** to differentiate words than languages that use the vowel duration contrast between **i** and **iː** to differentiate words. Researchers have shown that native speakers of English listen for this slightly centralized vowel quality as an extra cue that helps to distinguish the short vowels from the corresponding long vowels (e.g., Fox, Flege, and Munro, 1995). But what about native speakers of other languages? When people learn English as a second language, do they perhaps mishear the vowel quality difference as an optional casual-speech reduction effect? Could this lead them to pronounce the short vowel **ɪ** in careful speech in a way that makes it sound to native speakers as if they are saying the counterpart long vowel **iː** instead, as suggested in the title of this section? Second language learners might be especially susceptible to this kind of confusion if the English vowel duration differences are less consistent, which they might be because the distinction between short **ɪ** and long **iː** is carried by the vowel quality cue as well as by the duration cue. That is, since native speakers produce a slightly more centralized vowel quality to distinguish the word *bit* from the word *beat*, perhaps they don't need to be quite so careful to pronounce a distinctively shorter duration every time they say a word with the shorter vowel. If this is the case, then second language learners of English are faced with the task of learning to pay attention both to a subtle vowel quality cue and to a variable duration cue. This could make it even more difficult to learn to hear and then to pronounce the short vowel in words such *cut* and *bit*.

We would like to test the hypothesis that vowel durations are variable, and that they vary in ways that contribute to the difficulty that second language learners have in acquiring English short vowels. We saw some evidence in support of this hypothesis when we looked at the differences in duration between short and long vowels in Chapter 8 (the chapter on

measures of central tendency). There we saw that there is a lot of variation in the duration of vowels in speech, with some tokens of the short vowel **ʌ** being longer than some tokens of the long vowel **ɔ:** in the Hillenbrand et al. (1995) database. The question we want to ask, then, concerns the extent of this kind of variability. How much do the short vowels vary in duration? And how do the token-to-token differences within a given vowel type compare to the distinctive difference in duration between a short vowel and its long counterpart?

9.2. Relating variability to sources of deviation from the typical value

In order to answer this question, we first have to figure out a way of analyzing the spread, or dispersion, of values in a sample. In the chapter on measures of central tendency, we looked at the distributions of measured durations for several vowel samples. One was a sample of 488 "weak" **ə** vowels observed in tokens of the words *a* and *the* in an interview from the Buckeye Speech Corpus. Another was a sample of 1668 "strong" vowels observed in tokens of **h__d** words from the Hillenbrand et al. (1995) study, a sample that included the "short" strong vowels **ɪ** and **ʌ** as well as vowels such as **i:**, **ɔ:**, and **ɑ:** that are transcribed with the IPA length mark in Peter Ladefoged's table of American English vowel qualities. We noticed that these two distributions differed in interesting ways. There were clear peaks in the histograms that we made for both sets of vowels, but the peak in the histogram of the strong vowel tokens was at a much longer duration, and it was more symmetrical. We redraw the histograms in Figure 9.1 here, using the same set of bins[1] so that the peak locations and shapes can be compared directly.

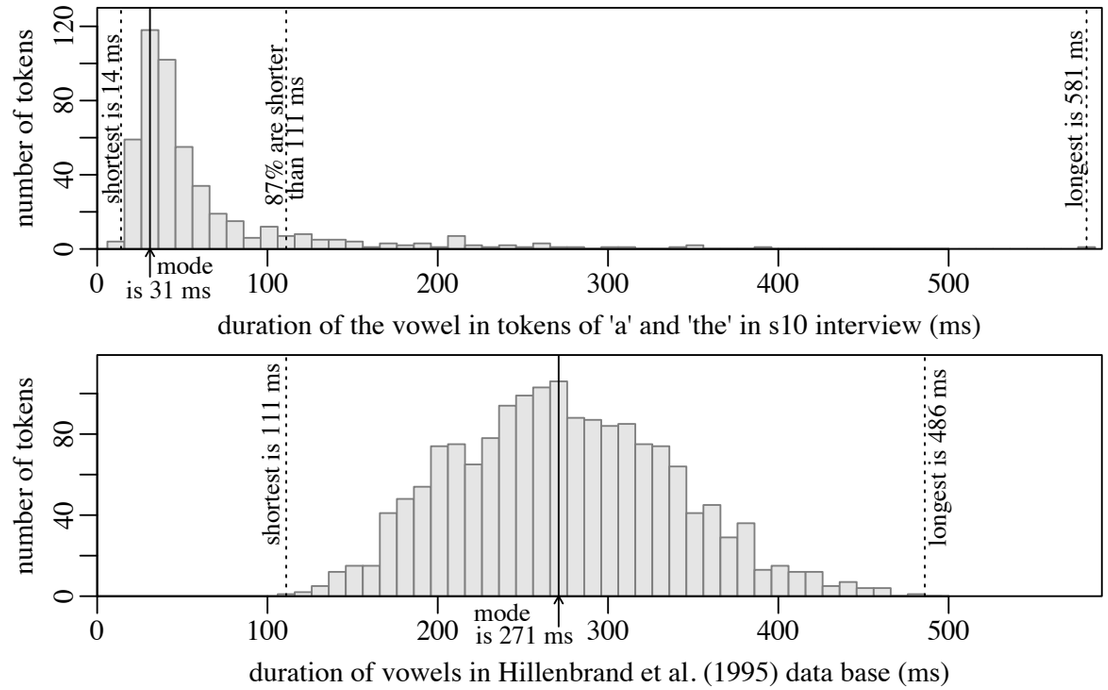


Figure 9.1. Histograms of measured durations of the reduced short vowel **ə** in all 488 tokens of *the* and *a* in speaker s10's interview in the Buckeye Speech Corpus (top panel) and of the full strong vowel in each of 1668 tokens of **h__d** words pronounced by the 139 speakers in the Hillenbrand et al. (1995) study (bottom panel). The bins are 10 ms wide and start at 6 ms.

[1] The bins we chose are 10 ms wide, as in Figure 8.7, but the lower edge of the first bin is at 6 ms, as in Figure 8.9, instead of at 0 ms. This admittedly arbitrary starting point positions the mode for the distribution in the bottom panel at the same value as the median. If we had positioned the bins to start at 0 ms, the modes would be 35 ms instead of 31 ms for the top panel and 265 ms rather than 271 ms for the bottom panel.

Another difference that comes out more clearly here, when we use the same bin widths and put the two histograms close together like this, is that the distribution of durations of strong vowel tokens in the carefully controlled sample of "lab speech" utterances in the Hillenbrand et al. (1995) study covers a smaller range than the distribution of the **ə** vowels from the connected speech corpus. "Lab speech" here means words spoken in isolation in a laboratory setting. Lab speech is more generally consistent as to duration, and has generally longer vowels than in the same words spoken in conversational speech. The minimum duration in the Hillenbrand et al. (1995) database is 111 ms and the maximum is 486 ms. By comparison, the minimum and maximum values in the sample of **ə** durations are 14 ms and 581 ms. So some of the tokens of the weak vowel type in the interview are actually longer than the longest tokens of the long vowels **iː**, **ɑː**, and **ɔː** in the lab speech sample.

Still, looking at the histograms, it's obvious that the typical weak vowel is much shorter than the typical strong vowel. That is, the values in the upper panel of the figure cluster very closely around the mode of 31 ms, which is 240 ms shorter than the mode of 271 ms that we got for the values in the lower panel.

In fact, 427 out of the 488 weak vowel tokens in the top panel of the figure — i.e., very nearly 90% — are shorter than the very shortest of the strong vowel tokens in the bottom panel. So the tokens of **ə** that are longer than the shortest strong vowel token are not "normal" for this vowel type. They are "outliers" in the distribution (like the extreme example of a skewed distribution that we gave in the last chapter where the outliers were the households with an annual income of $200,000 in a village where 90% of households have an annual income of $1000). And when we went back and listened to the sentences where these longer tokens occurred, we found that many of them were places where the speaker dragged out the vowel — i.e., where he was using the word *the* or *a* as another way of filling up a pause while he hesitated to find the next word, like the tokens of the **ʌ**-like vowel all by itself that the transcribers of the Buckeye Speech Corpus spelled as the "hesitation word" *uh*. (You can look at Figure 9.10 in Appendix 9B to see the distribution of the vowel durations for *uh* and its variant *um*.) In other words, when we look at the distribution of the **ə** vowel durations in the top panel of Figure 9.1, it's obvious what range of values to normally expect, and we can make a hypothesis about why these outlier tokens deviate in such an extreme way from the expected value for **ə** by considering one of the obvious non-random sources of ***sampling error*** (in this case hesitation) that we have to contend with in a sample like this.

So what about the deviation from the center of the distribution in the more controlled sample of strong vowel tokens? The total range of the distribution plotted in the bottom panel is much smaller than the range of the distribution plotted in the top panel of the figure (375 ms versus 567 ms) and there aren't obvious outliers. But the tails on either side of the highest bar fall off much more gradually, so the peak is not nearly

as peaky as in the distribution plotted in the top panel. Moreover, this appearance of a more gradual fall off is not just because the peak is not as tall in the bottom panel. This is clear from Figure 9.2, which is the same pair of histograms as in Figure 9.1, but plotted using a narrower bin width for the weak vowel and a wider bin width for the strong vowels, so that the tallest bars in the plots include an equivalent proportion of the tokens in the distributions. (The figure also zooms in on a narrower x-axis range, so that the longest outlier is no longer shown in the upper panel and the minimum and maximum values are very close to the edges of the view in the lower panel.)

Why is the peak in the histogram for the strong vowels so much less peaky? One reason is because strong vowels are longer than typical durations of the weak vowel **ə**. This means that they're further away from the boundary condition at 0 ms. So there is room for "random" error in both directions. We can see support for this interpretation in the fact that the distribution for the strong vowels is far more symmetrical than the distribution for the weak vowel. But we shouldn't assume that this is the only reason for the more spread-out shape. We can't automatically rule out other, non-random sources of the greater deviation from the center.

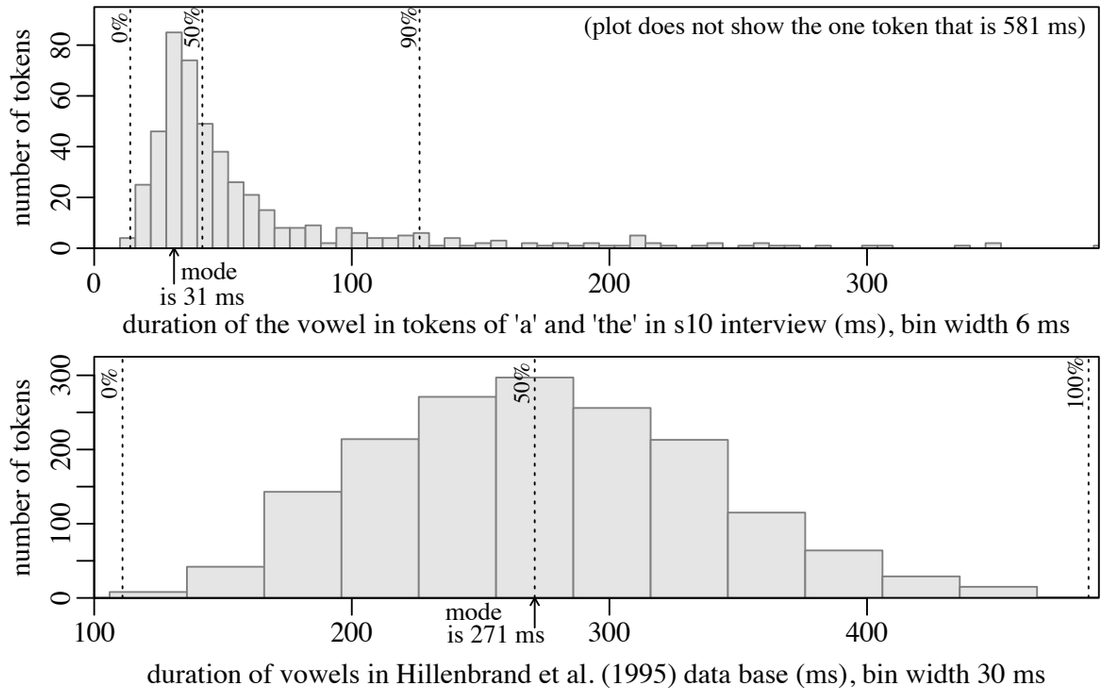


Figure 9.2. Histograms of the distributions in Figure 9.1, using different bin widths so that about 18% of the tokens are included in the tallest bar that is centered at the mode.

One possible non-random source of variation is the contrast between the distinctive vowel types that are included in the sample for the bottom figure. That it, whereas the histogram in the top panel of Figure 9.1 is the distribution for only one vowel type (the reduced vowel **ə** in *a* and *the*), the histogram in the bottom panel is for 12 different vowel types, including both the distinctively short strong vowels and the distinctively long strong vowels. We saw in the chapter on measures of central tendency that a histogram for a smaller subset of only two vowels in the Hillenbrand et al. study showed a bimodal distribution, with the taller peak (the mode proper) at the center of the values for the short vowel **ʌ**, and the shorter peak nearer to the center of values for the long **ɔː** vowel. Figure 9.3 shows

similar graphs of the vowels ɪ in *hid* and iː in *heed*, with the tokens of the two vowel types plotted in separate histograms in the two panels of the figure. The mode for the short vowel ɪ is the bar centered at 211 ms in the top panel. That's still 180 ms longer than the mode for ə in Figures 9.1 and 9.2, but it's also 60 ms shorter than the mode for the long vowel iː in the bottom panel of this figure.

On the other hand, even when we plot just one vowel type per histogram, neither of the histograms is as peaky as the histogram in the top panels of Figures 9.1 and 9.2. For example, there are a fair number of tokens of ɪ that are as long as the most typical tokens of the long iː vowel. Indeed, far fewer than 25% of the tokens of the short vowel are shorter than the shortest token of the long vowel, and far more than 75% of the tokens of the long vowel are shorter than the longest short vowel. That is, the two distributions have a lot of overlap, so that even the two tails that are the lower quarter and higher quarter of the values overlap to some degree. You should be able to see this easily because of the way that we've used dotted lines to mark the values at 25% and 75% of the observations, as well as at the minimum (0%), median (50%), and maximum (100%) in each panel of Figure 9.3.

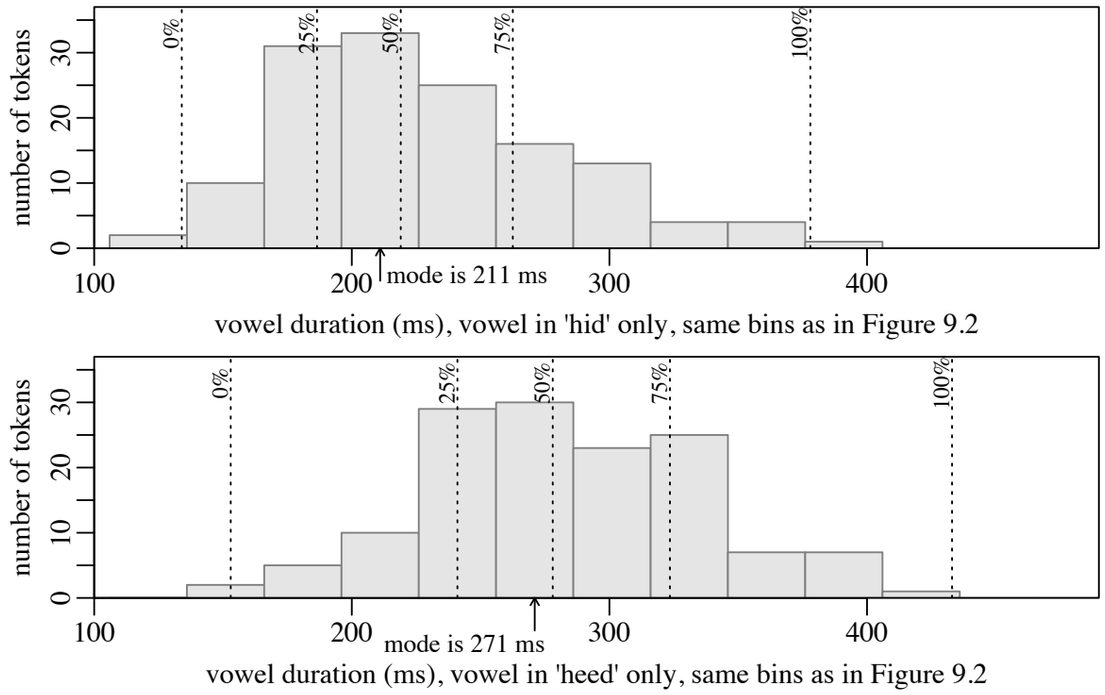


Figure 9.3. Histograms of measured durations of the vowel in the 139 tokens of *hid* (top) and the 139 tokens of *heed* (bottom) in the Hillenbrand et al. (1995) study.

So, how can we evaluate the difference between the expected value for the short vowel ɪ and the expected value for the long vowel iː, given the large average deviation from the center in these two samples? How can we gauge the reliability of the mean value relative to the pattern of deviations from this central value?

In this chapter, we will introduce some of the numerical concepts that we need in order to begin to answer these questions. One important concept is the idea that we can quantify the spread of the data around the center of the distribution, choosing one of two measures that have been developed in conjunction with the measures of central tendency that we in-

troduced in Chapter 8. The other important concept is the idea that, for many natural numerical variables, there is a normal pattern of random deviation from the expected value at the center that accounts for the symmetrical peak shape that we see in the lower panels of Figures 9.1 and 9.2. So we can evaluate differences across subparts of a sample using our expectations about this normal pattern of variation. We'll talk briefly about this idea toward the end of the chapter, and in more depth in several of the later chapters. But first, we'll introduce methods for measuring the spread of data.

9.3. Measures of dispersion - Interquartile range

The traditional way to evaluate the spread of the data is to calculate how tightly the observations cluster around the number that measures the central tendency, using some number that is small when the distribution is tightly clustered around the middle and large when the distribution is very spread out. This kind of number is called a ***measure of dispersion***. Just as there are different ways of measuring the center of a distribution, there are different ways of measuring dispersion.

The measure of dispersion that is most often used with the median value is the ***interquartile range*** (IQR). This is the spread of values that covers the middle half of the observations, excluding the lowest 25% and the highest 25%. The name "interquartile" refers to the fact that these two values and the median divide the data into quarters — i.e., into four equal groups. We've already illustrated the concept in the two panels of Figure 9.3. The IQR is the range of values between the dashed vertical lines at 25% and 75% on each histogram. That is, the 25% mark is the ***first quartile*** (Q1) and the 75% mark is the ***third quartile*** (Q3).

R Note 9.1 – interquartile range

Equation (A1) in Appendix 9A describes how to compute the IQR "by hand" for a dataset that is small enough to identify Q1 and Q3 just by looking at the sorted numbers. For larger sets, you can use the `quantile()` function to find these numbers. While ***quantile*** is a generic term that refers to any series of points that cuts the data into same-sized chunks (e.g., quartile (4 chunks), quintile (5 chunks), percentile (100 chunks), etc.), R uses this term to stand in for the points of a quartile. The `quantile()` function takes a vector of numbers as an obligatory argument, and returns the following 5 values: (1) minimum, (2) Q1, (3) median, which is the same as Q2, (4) Q3, and (5) maximum, which is the same as Q4. For example, if you read in the Hillenbrand et al. dataset:

```
hill = read.delim(Ch09.Textfile1.txt)
```

You could calculate those 5 values for the `duration` values in `hill`, by specifying the duration column of the hill table as the argument to the `quantile()` function, like this:

```
quantile(hill$duration)
```

The elements in the vector that is returned are named by the percentage of the data that is covered up to that "quantile"

value, so the command returns the following:

```
  0%  25%  50%  75% 100%
 111  227  271  319  486
```

You can refer to the elements in this returned vector by name using the `[…]` indexing brackets, which gives you the following convenient way to calculate the IQR:

```
quantile(hill$duration)["75%"] -
  quantile(hill$duration)["25%"]
```

Note that the value returned by this equation is:

```
75%
 92
```

The `75%` name is because the result of subtraction inherits the name of the first term in the equation. We can suppress this (confusing) name by embedding the equation in the `as.numeric()` function, like this:

```
as.numeric(quantile(hill$duration)["75%"]-
  quantile(hill$duration)["25%"])
```

to get the following unlabeled value:

```
[1] 92
```

You can compare the IQR values for the 2 subsets of the data in Figure 9.3 to the IQR value for the dataset as a whole by looking at Figure 9.4. This is a repeat of the lower panel of Figure 9.2, but this time with dotted lines marking off the IQR. That is, the dotted lines mark the first and third quartile values (**Q1** and **Q3**), which define the bottom and top of the interquartile range. As noted earlier, the median value for this sample of vowel durations is 271 ms. The Q1 value is 227 ms (44 ms less than the median) and the Q3 value is 319 ms (48 ms above the median). This is an IQR of 319 – 271 = 92 milliseconds. In this case, the distribution is symmetrical, so the median is about midway between Q1 and Q3.

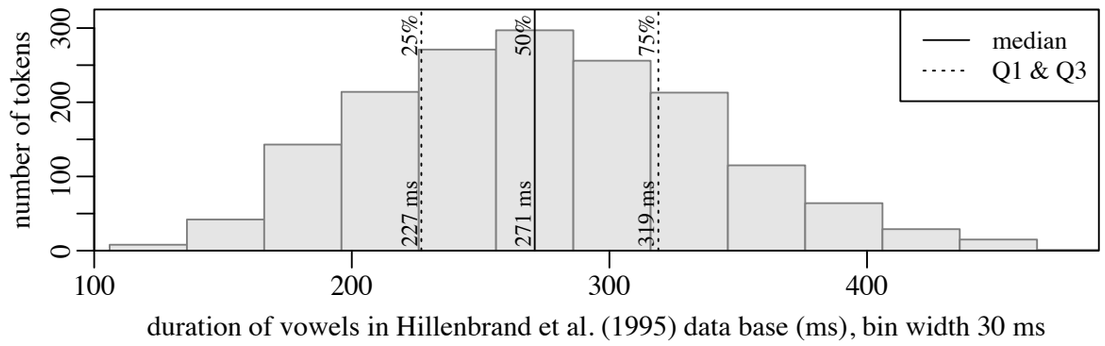


Figure 9.4. Same histogram as in the lower panel of Figure 9.2, with the median marked by a solid line and the 1st and 3rd quartiles marked by the dotted lines, as indicated in the legend.

We can see a somewhat different pattern in the plot of the short **ɪ** vowel tokens in the top panel of Figure 9.3. Here the IQR is smaller than the IQR for the strong vowels as a whole (76 ms as compared to 92 ms), but much of that reduction is in the interval from Q1 (at 186.5 ms) to the median (at 219 ms), a distance that is only 32.5 ms as compared to the 43.5 ms distance between the median and Q3 (at 262.5 ms). That is, the distribution shows some asymmetry. The shrinking of

the distance between the first quartile and the median reflects the positive skew.

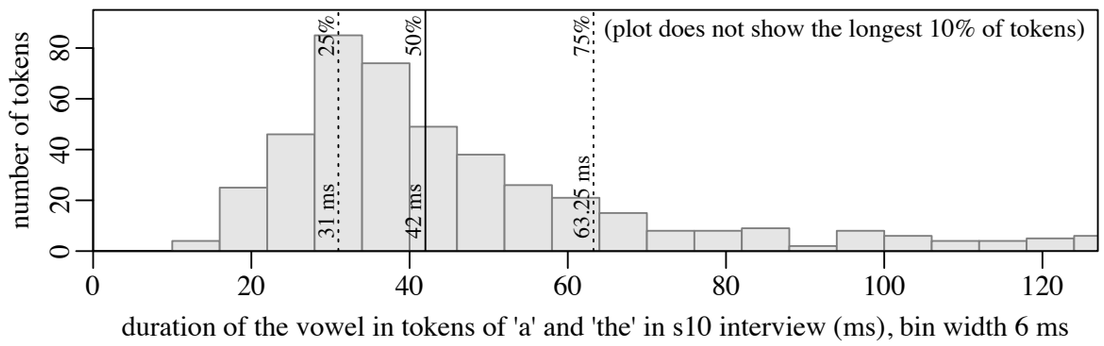


Figure 9.5. Same histogram as in the top panel of Figure 9.2, but with the x-axis zoomed in even further to not show the longest 10% of the data. The median is marked by a solid vertical lines, and Q1 and Q3 are marked by and dashed vertical lines.

Figure 9.5 shows an even more dramatic asymmetry. This figure repeats the histogram in the top panel of Figure 9.2, but with the x-axis zoomed in even further, on the range from 0 ms to 127 ms. That is, this plot shows the durations of **ə** in all of the words *a* and *the* in the interview with speaker s10 except for the longest 10% of tokens. The median and IQR that are shown are calculated over all 488 tokens, though, so that you can see where the central value and the main spread of the data lie. Even in this close-up view, the central value (i.e., the median) for all 488 tokens is still not at the center of the range of values covered in the figure. That is, the median duration of this weak vowel is 42 ms, well to the left of the middle value plotted.

Another thing to notice is the way that the IQR lines highlight the asymmetry in the distribution. First, note how the IQR is even smaller than the IQR for the short strong vowel in Figure 9.3 (32.5 ms as compared 56.5 ms). Then note how the "shrinking" is concentrated in the region to the left of the median, because of the positive skew in the distribution. The first quartile at 31 ms is only 11 ms below the median, whereas the third quartile at 63.25 ms is more than 21 ms above the median. Thus, the IQR is even less centered around the median than in the top panel of Figure 9.3 because the peak in this histogram is even closer to the boundary condition that a measured vowel duration cannot be less than 0 ms. In such skewed distributions, one or the other edge of the IQR will be further away from the median, in the direction of the skew. In this case, the right edge (Q3) is further away from the median because the distribution is skewed right.

9.4. Using box plots to display quartiles

These differences in degree of symmetry are especially obvious in the box plots in the right panel of Figure 9.6. A box plot is a special type of graph that uses a "box" (a divided rectangle) and "whiskers" (lines stretching out from the edges of the rectangle) to represent the quartile edges for a sample of values. A ***box plot*** is also sometimes called a "box-and-whisker" diagram. The left panel of Figure 9.6 illustrates how a box plot is laid out. This figure contains a horizontal box plot with the left and right sides of the box representing the first and third quartiles of the vowel duration values in the Hillenbrand et al. (1995) dataset. We've put a histogram of the same data, with vertical cursors drawn at the quartiles, as in Figure 9.4, below

the boxplot, so you can see how areas of increased density of the histogram contribute to the body of the boxplot. You can see how the cursors at the Q1 and Q3 run through the edges of the box, and also how the cursor at the median runs through the thick line in the center of the box. In this figure, the ends of the whiskers that extend leftward from the Q1 and rightward from the Q3 are the minimum and maximum. (The minimum, Q1, median, Q3, and maximum are also often called the "five number summary" of a sample.)

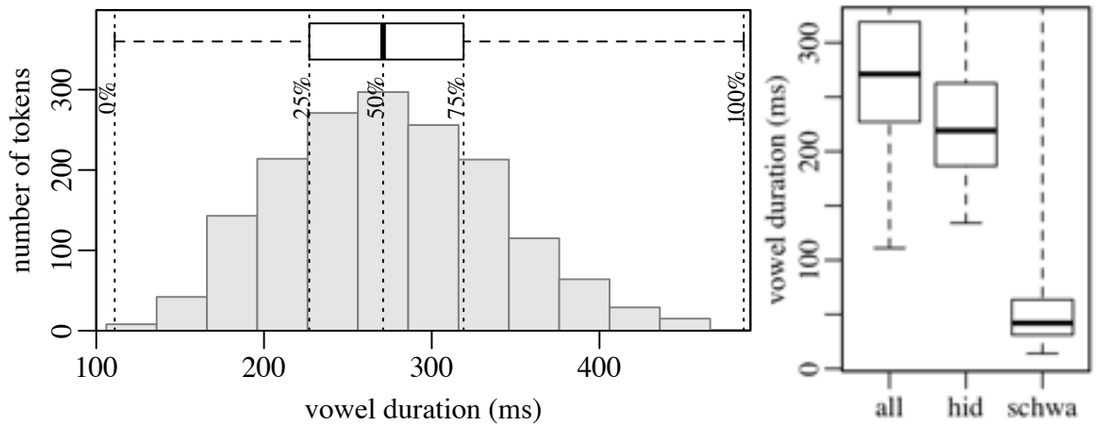


Figure 9.6. The left panel shows the same histogram as in Figure 9.4, with a horizontal box plot overlaid at the top. The right panel shows the same box plot for all the vowels, and also box plots for the durations of the vowel in *hid* (the data from the top panel of Figure 9.3) and for the durations of the reduced vowel in all of the tokens of *a* and *the* from talker s10 in the BSC (as in Figure 9.5).

While we have shown this box plot of all vowel durations horizontally so that you can see the relationship to the histogram, it is more usual to display box plots vertically, as shown in the right panel of Figure 9.6. The first box plot in this panel on the right is the same one as in the left panel, but rotated so that the left and right sides of the box are now the bottom and top of the box and the median divides the box into a lower part (for the range of values from Q1 to the median) and an upper part (for the range of values from the median to Q3). The other two box plots show the analogous summary values for the durations of the ɪ vowel in *hid* (the same data as in the top panel of Figure 9.3) and the durations of the ə vowel in *a* and *the* from talker s10 in the BSC (the same data as in Figure 9.5). By comparing the part of the box below the median line to the part of the box above the median, you can quickly see the degree to which the distribution is skewed. The lower and upper parts of the box are not as similar in height for the second box plot when compared to the first box plot, and the two parts are even less similar in height for the third box plot.

9.5. Measures of dispersion - Standard deviation

The measure of dispersion that is traditionally used in association with the mean value is the sample ***standard deviation***. The standard deviation evaluates the spread of the observations by calculating the average distance between the observation and the mean value, using formula (1):

$$(1)\quad \text{standard deviation} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

In this formula, n is the number of observations in the sample, x_i is the series of observed values from the 1st to the nth, the subscript i meaning each one in turn, and $\bar{x}$ is the mean

value for the n values in the sample. The formula says, essentially, to subtract the mean value from each observed value and square that difference (so that deviation from the mean is expressed as a positive number whether an observed value is higher than the mean or lower than the mean). The Σ then says, add up all those squared differences, and the fraction in front of it says to then divide by one less than the number of observations to get the average squared difference. This average squared difference, before we take the square root, is called the ***variance*** of the sample and it is an important number in its own right. Notice that unlike in the formula for the mean in Chapter 8, where we divided by the number of observations (n), here we divide by one less than the number of observations (n-1). This adjusts for the fact that in small samples, the sample variance tends to underestimate the true variance of the variable because a small sample is much less likely to include any extremely small or extremely large values. This formula is therefore sometimes called the "bias-corrected standard variance." The standard deviation of the sample is the square root of this variance — i.e., we "undo" the squaring in the formula for the variance in order to convert back to the original units of measurement.

Equation (A2) in Appendix 9A illustrates these steps for a very small sample where we can show all of the values in x, and there is also some R code there for calculating each step starting with $(x_i - \bar{x})$. In R Note 9.2, we give you the R function that does it all in one step.

Like the first and third quartile, the standard deviation can be used to mark off an estimate of the middle part of the range of the data. The dashed lines in Figures 9.7 and 9.8 illustrate this. In each figure, the central solid line marks the mean value for the distribution plotted, and the dashed lines mark the values that are one standard deviation away from the mean on either side. As you can see by comparing Figure 9.6 to Figure 9.3, and Figure 9.7 to Figure 9.4, the range of data points that is covered by the mean ± one standard deviation is larger than the IQR. It covers a bit more than the middle 68% of the data, whereas the IQR covers the middle 50% of the data. However, the range using standard deviation is, by definition, centered at the mean. Thus, as Figure 9.6 also shows, when the distribution is very skewed, one or the other edge of this range can be outside of the observed range of the data. Despite these differences between the two measures in terms of how well they capture the shape of the distribution, both are fairly robust measures of dispersion.

R Note 9.2 – standard deviation

The sequence of steps in (A2) in Appendix 9A illustrates how to calculate the standard deviation step by step, as in the description of equation (1) here. You can use R to do each of these steps, of course, and R Note 9.4 in Appendix 9A shows you how. The easiest way to get the standard deviation, though, is to use the built-in `sd()` function. This function takes a vector of values as its argument, like this for the named vector of durations that you read in for R Note 9.1:

```
sd(hill$duration)
```

This is the standard deviation that the function returns:

```
[1] 64.50101
```

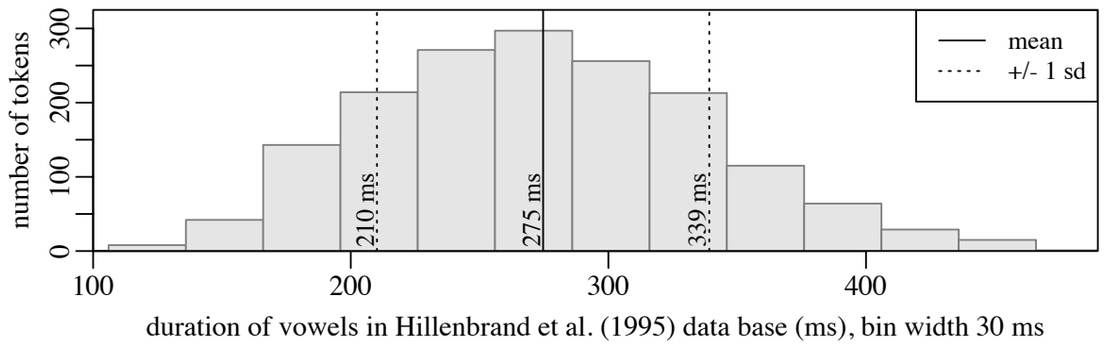


Figure 9.7. Same histogram as in Figure 9.4, but with the mean±1 standard deviation marked by the different vertical lines, as indicated in the legend.

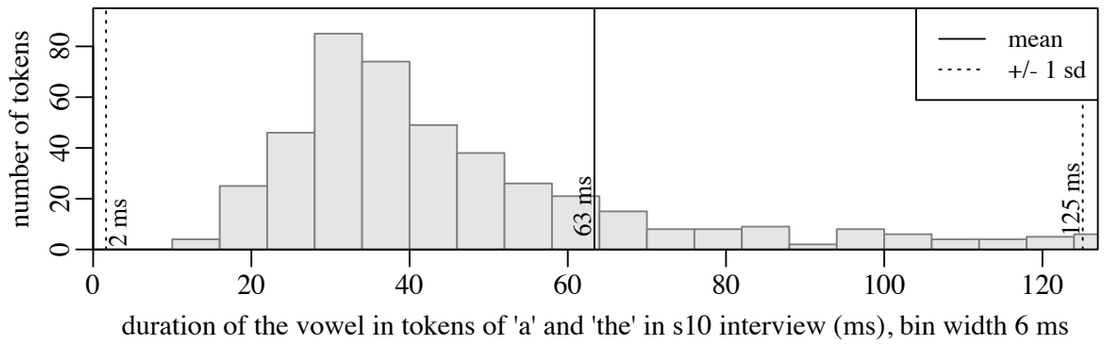


Figure 9.8. Same histogram as in Figure 9.5, but with the mean±1 standard deviation marked by the different vertical lines, as indicated in the legend.

9.6. Robustness

What do we mean by "robust"? We mentioned in the chapter on measures of central tendency that the values at the mode, at the median, and at the mean of a sample are very similar to each other whenever there is a single, clearly defined peak near the middle of the distribution. For example, in the sample distribution plotted in the lower panels of Figures 9.1 and 9.2 and again in Figures 9.4 and 9.7, these three measures of central tendency are all within 10 ms of each other. This is a convenient fact, because the median and the mean are more ***robust*** measures of central tendency than the mode. So what do we mean by "robust"?

A first criterion for "robustness" is whether a value is ***well-defined***. By this criterion, the mean and median are robust measures in that they can be calculated for any distribution, no matter what the shape. By contrast, the mode cannot even be defined for a distribution that has no clear peaks. An extreme example is any numerical variable that has a ***uniform distribution*** — a distribution where each possible value occurs equally often — such as the distribution of values from 1 to 6 for the number of dots that surfaces on a roll of a fair die in panel (f) of Figure 5.4 in Chapter 5. Since there is no peak, the mode cannot be defined. Figure 9.11 in Appendix 9B gives a more relevant example of a distribution for which the mode cannot be meaningfully defined.

A second criterion for deciding whether a statistic is "robust" is its ***stability*** or constancy. We showed in the chapter on central tendency that, for any given sample of observations, the mode can shift depending on how the observed values are binned in order to plot a histogram. By contrast, the mean and median for a sample stay the same no matter how we plot the observed values. The mean and median and the corresponding measures of dispersion (i.e., the standard deviation

and IQR, respectively) are also stable in that they are fairly consistent across samples drawn randomly from a population, given a large enough sample size.

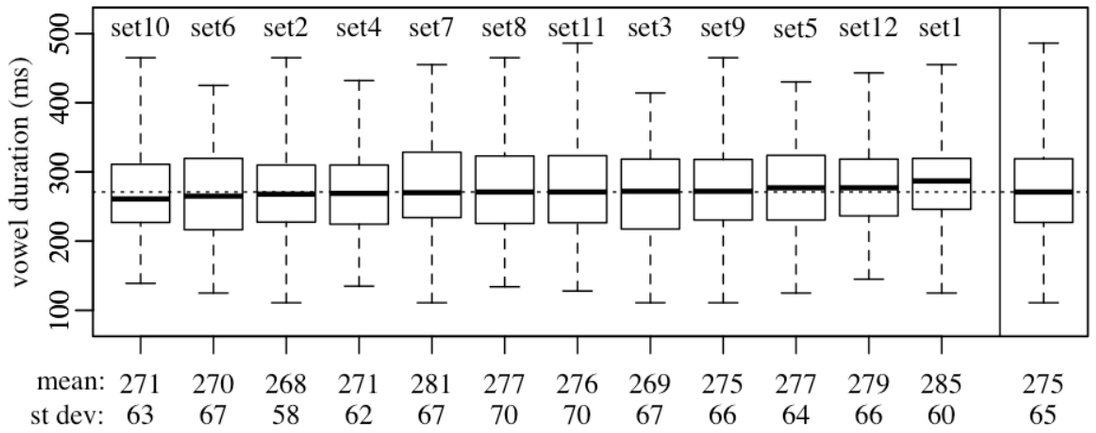


Figure 9.9. Box plots for the vowel duration values for 12 subsets of 139 values each, drawn randomly from the Hillenbrand et al. (1995) data (first 12 boxes), with horizontal dashed line at the median for the overall dataset (box at far right). The boxes for the random subsets are ordered by their median values. The numbers below the x-axis show the means and the standard deviations for the 12 subsets of values and for the sample as a whole.

This stability is illustrated by Figure 9.9, which shows summary measures for 12 random subsets of the vowel duration values in the Hillenbrand et al. (1995) database. Each of these random subsets contains 139 tokens, so it's the same sample size as if we had asked each of the 139 talkers in the database to produce exactly one vowel. The median and the IQR values for these subsets are indicated by the boxes in the plot. The number above each box indicates the order in which we drew that random sample from the dataset. We could have used this order to arrange the boxes, but we've chosen to arrange them in order of the size of the median. The two numbers below each box are the mean and standard deviation for the random sample of 139 duration values. The box at the far right is for the dataset as a whole, and the dashed horizontal line marks the median for the whole 1668-token sample. (The box plot for the whole dataset also reproduces the first of the three box plots in the right panel of Figure 9.8.)

The arrangement of this figure lets us compare the center (median line), spread (IQR box height), and overall range (whisker edges) of the distribution for each group (any one of the 12 boxes on the left) to the center, spread, and overall range of the sample as a whole (in the summary box on the far right). Also, by comparing among the 12 boxes themselves, we can compare medians, IQR, and full ranges across the different smaller samples that were drawn at random from the larger dataset. And by examining the numbers below the boxes, we can compare the means and standard deviations, too.

By making these comparisons, you can see how each of the summary measures for these random subsets remains relatively stable, clustering around the value for that summary measure for the dataset as a whole. For example, the median value is right on top of the dashed line for seven of the subsets. Moreover, the median for even the subset with the most extreme deviation — the 287 ms median for set 1 — is only 16 ms above the median for the dataset as a whole. The corresponding measure of dispersion shows a similar degree of stability. The IQR for the dataset as a whole is 92 ms. The IQR values for the random subset range from 74 ms for set 1 (i.e.,

18 ms smaller than the IQR for the whole dataset) to 101 ms for set 3 (11 ms bigger). And the other pair of summary measures is even more stable. The mean values for the random subsets are all within 10 ms of the 275 ms mean for the sample as a whole, and the standard deviations for the random subsets are all within 7 ms of the 65 ms standard deviation for the whole sample.

9.7. Developing an *a priori* model of random variation

One of the most important achievements in the history of the modern science of statistics was to prove that the mean and the standard deviation must be stable across large random samples, because of the logic of how probabilities for simple outcomes are combined in determining the probability of a more complex compound outcome. Remember the equations that we worked through for using the probability of the letter 'e' all by itself to calculate the probability that the letter 'e' occurs anywhere on a *Wheel of Fortune* puzzle board with a certain number of letter positions? The logic of those equations gives us a basis for calculating the degree to which the mean and the standard deviation should be stable for samples of different sizes drawn randomly from a population of values like the vowel durations in all of the figures in this chapter.

This achievement is called the *Central Limit Theorem* and we'll talk about it more in the coming chapters in this textbook. For now, we'll pre-figure that discussion by pointing out that the stability of the mean, the median, and the two corresponding measures of dispersion gives us an *a priori* model of the variation in a population of values like the duration values in Figures 9.3 and 9.6. When random samples are drawn from such a population, the summary values for the center and the spread should be similar to the summary values for the population as a whole, and also, therefore, similar to each other, as in Figure 9.8. If there is a systematic pattern of deviation from this *a priori* model, then we have a basis for suspecting that the samples are not random.

9.8. Using box plots to explore possibly non-random sources of spread

Figure 9.10 illustrates what we mean by a "systematic" pattern of deviation. Like Figure 9.9, this figure shows a set of box plots for the vowel duration values in the Hillenbrand et al. (1995) dataset. As in Figure 9.9, the box on the far right shows the quartile values for the dataset as a whole and the other 12 boxes are for 12 subsets of 139 tokens sampled from dataset. Instead of taking random samples, though, here the values are grouped into 12 subsets by the vowel type. Since the grouping variable that defines these subsets is a categorical variable, there is no inherent order that we need to use to arrange the boxes from left to right on the graph. So we can choose to order the boxes from the box for the subset with the shortest median to the box for the subset with the longest median. This makes it easier to evaluate the basis for the idea that some

strong vowels are inherently "short" vowels and others are "long" vowels.

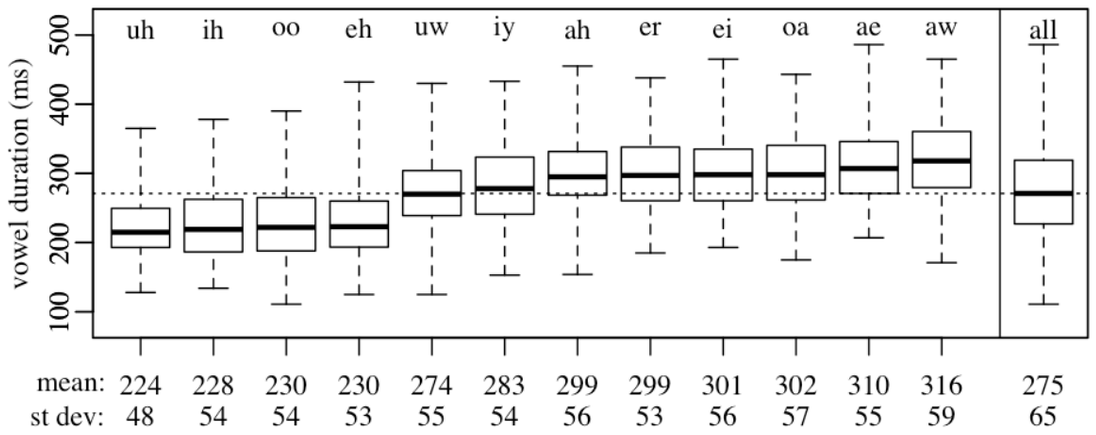


Figure 9.10. Box plots for the vowel duration values in the Hillenbrand et al. (1995) data, for subsets that are defined by "vowel type" (first 12 boxes) and for all the vowels in the dataset (box on the far right). The horizontal dashed line marks the median for the overall dataset. The rows of numbers below the x-axis are the means and standard deviations for the 12 subsets of values and for the sample as a whole.

The arrangement of this figure lets us compare the center, spread, and overall range of the distribution for each group (any one of the 12 boxes on the left) to the center, spread, and overall range of the population from which the samples were drawn (the summary box on the far right). We can also compare medians, IQR, and full ranges across the different smaller samples defined by the grouping variable. This makes it easy to see that the medians for only two of the vowels (the long high vowels **u:** and **i:**) are very close to the median for the dataset as a whole. All of the other vowel types are either substantially shorter on average or substantially longer on average than the median for all of the tokens. For example, the upper edge of the box (which represents the upper limit of the IQR) for each of the four short vowels **ʌ**, **ɪ**, **ʊ**, and **ɛ** lies below the median for the dataset as a whole, meaning that 75% of the tokens of each of these vowels are less than the expected value for this dataset.

Another thing to note is that all twelve of these boxes are flatter than the box for the dataset as a whole, meaning that the IQR for each of the subsets defined by the vowel type is smaller than the IQR for the dataset as a whole. As noted earlier, the IQR for the dataset as a whole is 92 ms. The IQR values for the 12 subsets in Figure 9.10 range from 57 ms for **ʌ** to 83 ms for **i:**. This pattern of deviation from the IQR for the sample as a whole suggests that at least part of the spread for the dataset (i.e., the relatively flat peak in Figure 9.6) is due to a systematic (that is, a non-random) pattern of variation in the expected durations for the different vowel types. In other words, the dataset as a whole may have a larger IQR simply because the typical durations for the four inherently short vowels are less than the typical durations for all of the other vowel types, which include the diphthongs **eɪ** and **oʊ** as well as four vowels that Ladefoged transcribes with the length marker (**u:**, **i:**, **ɑ:**, and **ɔ:**).

R Note 9.3 – making a box plot

If you're making just one box, as in the far right of Figure 9.9, the R function for making the box plot takes a single vector of values as its only required argument, like this:

```
boxplot(hill$duration)
```

If you're making a set of boxes, using an associated categorical variable that you could use to make subsets, such as the different types listed under `vowel`, as the grouping variable associated with each duration, you can use the "~" ("tilde" symbol) formula operator, like this:

```
boxplot(hill$duration ~ hill$vowel)
```

This command says, essentially, make a series of box plots for subsets of the `duration` values in the `hill` dataset, where the subsets are chosen using the category types in the vowel column of the `hill` dataset. Since there are 12 category types, this produces 12 box plots in a row.

We can contrast the pattern here with the pattern for the 12 random samples shown in Figure 9.9. In Figure 9.9, even the most extreme deviations from the overall median or mean are very small by comparison to the deviation downward for the medians of the four short vowels and the deviations upward of the longest of the long vowels. And even though we've arranged the box plots in Figure 9.9 by the medians, the tops and the bottoms of the boxes that mark off the interquartile ranges don't show any pattern of increase across the 12 random samples, because the IQR is similar to the IQR for the data set as a whole, and the direction of the difference (i.e., whether the spread is larger or smaller than the spread of the dataset) is independent of the direction of the difference between the median value of the smaller sample and the median value of the sample as a whole.

In contrast, Figure 9.10 shows a clear pattern for the placement of the boxes relative to the line that is drawn across the figure at the median value for the dataset as a whole. The size of the spread for each of the subsets of the dataset is consistently smaller than the IQR for the sample as a whole, so that the top of the box at Q3 for each of the four short vowels is below the median for the dataset as a whole, while the bottom of the box at Q1 for each of the long vowels lies above or on the median line for the dataset as a whole.

We can also compare the placement of pairs of boxes. By looking at where the top of the box for some short vowel lies relative to the bottom of the box for the corresponding long vowel, we can see, for example, that 75% of the tokens of the short vowel **ʌ** have a measured duration that is less than the measured duration of 75% of the tokens of the corresponding long vowel **ɔː**. Similarly, 75% of the tokens of the short vowel **ɛ** have a measured duration that is less than the measured duration of 75% of the tokens of the corresponding diphthong **eɪ**. That is, for these pairs of vowels (**ʌ** and **ɔː** as well as **ɛ** and **eɪ**), only about 25% of the measured duration of the tokens for one vowel in the pair overlap with the measured duration of the tokens for the other vowel in the pair. The durations measured for tokens of the short vowel **ɪ** do show more overlap with the durations measured for its long counterpart **iː**, but even for this pair, there is a substantial difference in the median values in the direction that we would expect if the duration of the vowel is a cue to the distinction between the more centralized short vowel and the less centralized long vowel.

9.9. Answering the research question

At the beginning of this chapter, we posed the following question about the observed differences between the typical durations of short vowels, such as **ɪ**, and the typical durations of long vowels, such as **iː** – Do long and short vowel durations vary in ways that might contribute to the difficulty that second language learners have in acquiring English short vowels, such as **ɪ** and **ɛ**? Then we looked at the distributions of long and short vowels in Figure 9.10, and saw that short vowels were indeed shorter than the long vowels; although, there was a good deal of variability in each of the vowel categories. So, how can we evaluate whether those observed differences are reliable? The difference between the typical duration of the weak vowel **ə** and the typical duration of the strong vowels as a group is glaringly obvious, given the clear separation between the modes of the histograms for the two samples that we've been comparing. By contrast, neither of the histograms for the durations of the **ɪ** and **iː** tokens in the carefully controlled lab speech sample that we looked at shows a mode that is anywhere near as peaky as the mode for the sample of **ə** vowels taken from the less controlled interview, and there was far more overlap between the longest tokens of the "short" **ɪ** vowel and the shortest tokens of the "long" **iː** vowel. How can we gauge whether the difference between the short **ɪ** vowel and the long **iː** vowel isn't just random deviation from a single expected value for strong vowels as a whole? We can pose this question as the pair of competing explanations in (2).

(2) Null Hypothesis (H_0, the default explanation): The deviations from the overall average across the 12 full vowel types are random.

Alternative Hypothesis (H_1, the alternative explanation): The deviations from the overall average are systematically related to the differentiation of the 12 vowel types.

As a first step toward developing a test of the alternative explanation of the variation within the sample of strong vowels, we extracted 12 sets of duration values by systematically taking all 139 values each of the vowel types. And we also extracted 12 sets of 139 values by drawing randomly from the same larger dataset. We then calculated measures of central tendency and of dispersion for all 24 groups of values and compared the patterns of deviation from the measures of central tendency and dispersion for the dataset as a whole. Figures 9.9 and 9.10 let us compare the two types of grouping.

In Figure 9.9, for the random subsets, we see that the median values cluster closely around the median for the dataset as a whole. This pattern contrasts with the pattern in Figure 9.10, where the median for the short vowel **ɪ** is 47 ms below the dashed line and the median value for the long vowel **iː** is 8 ms above the dashed line.

The mean values for the random subsets in Figure 9.9 also are quite stable, deviating from the mean for the dataset as a whole by no more than 10 ms in either direction. Again, this contrasts with the distribution for the groups in Figure 9.10,

where the means range from 224 ms for **ʌ** (51 ms shorter than the mean for the whole sample) to 316 for **ɔ:** (41 ms longer than the mean for the whole sample).

The IQR values for the random subsets in Figure 9.9 are also very similar to the IQR of 92 ms for the dataset as a whole. By contrast, the largest IQR in Figure 9.10 is the 82.5 ms IQR for **i:** and most of the vowels had even less spread. The 57 ms IQR for the short **ʌ** vowel was the smallest. While this is still considerably more than the 32 ms IQR for the sample of **ə** vowels, it is much less than the IQR for any of the random subsets.

The standard deviations of the random subsets in Figure 9.9 are especially stable, with none being more than 6 ms away from the standard deviation for the dataset as a whole. This helps us appreciate how the sample standard deviation is a "bias-corrected" measure of the variation in the "population" from which these random samples are drawn. Even though the box plots in Figure 9.10 show overlap between the longer values measured for the short vowel types and the shorter values measured for the long vowel types, the different standard deviations reflect the non-random nature of grouping the vowel durations by vowel type. That is, the standard deviation for this "population" (i.e., the larger dataset) is 65 ms, but most of the vowel types showed much smaller standard deviations, with the short **ʌ** vowel again showing the smallest, with a standard deviation of only 48 ms. Only the standard deviation of 59 ms for the **ɔ:** vowel is within 7 ms of the standard deviation for the larger dataset. In contrast, the standard deviations for the random subsets hover around the standard deviation for the larger dataset, deviating at most by 7 ms (for dataset 2).

These differences between the random subsets and the subsets defined by the vowel type strongly suggest that the short strong vowels are typically shorter than the long strong vowels even though they are all longer than the typical **ə**. And while the measures of dispersion show a larger spread of values than for the very short reduced vowel **ə**, there is a systematic difference between the range of durations measured for each short vowel and the range of durations measured for the corresponding long vowels. Even the vowel pair short **ɪ** versus long **i:** shows a substantial separation, so that only the longest 25% of the short vowel's values (the values above the top of the interquartile box in Figure 9.10) are longer than the median value for the long vowel. So, even if the overlap between long and short vowels makes the difference more difficult to learn for adult learners of English as a second language, native speakers do appear to be making a systematic difference. (In Chapter 11, we will show you how to use a statistical test that can tell if the mean and dispersion of two distributions, such as the durations of **i:** and **ɪ**, are significantly different from each other.)

9.10. Summary

Because many of the numerical variables that we are interested in reflect the complicated accumulation of many different random events, their values often tend to gravitate toward

the middle of the range of possible values. When a large sample can be drawn randomly from the "population" of values of a variable, the sample median and the sample mean are robust measures of the center of the distribution for the population as a whole. The interquartile range and the "bias-corrected" standard deviation of the sample, similarly, are robust measures of the typical variation around the median and mean. While a histogram is good for looking at the distribution of values in a single sample, it is easier to compare the distributions of several samples by picturing the median, IQR, and full range of values using box plots. By generating box plots for different subsets of values in a larger dataset, we can explore potentially meaningful differences in median values across groups of tokens that are differentiated by their values for another variable that characterizes the dataset. For example, a dataset of vowel duration values can be divided into subsets by the categorical variable "vowel type" in order to explore potential differences in typical duration values for short strong vowels versus long strong vowels.

Summary of key terms:

Box plot: A special type of graph that describes the distribution of a sample using a divided "box" to indicate the middle three quartile values (i.e., the median and the limits of the IQR) and "whiskers" to indicate the edges of some larger range, such as minimum to maximum.

Dispersion: The idea that we can describe the spread of the values in a sample, by measuring some proportional range of values that deviate from the expected value at the "center" that has been estimated by calculating one of the two robust measures of central tendency (i.e. the mean or the median).

Interquartile Range (IQR): The middle 50% of the data, a measure of dispersion that is used together with the median. For example, the interquartile range for the durations of the vowels in the Hillenbrand et al. (1995) study is 92 ms, ranging from 227 ms (the first quartile) to 319 ms (the third quartile).

Quantile: A generic term that refers to any series of points that cuts the data into same-sized chunks (e.g., quartile (4 chunks), quintile (5 chunks), percentile (100 chunks), etc.

Robustness: The reliability of a summary measure or statistical test, evaluated by the criteria that it is always (1) well-defined, and (2) stable (see definitions just below). For example, the sample variance is a robust measure of the population variance.

Well-defined: A statistical test or summary measure is well-defined if it can be calculated for a wide variety of distributions, and holds up in case of small violations of assumed parameters.

Stable: A statistical test or summary measure is stable if it does not vary in unpredictable ways from one sample to another.

Sample variance: The average *squared* distance between observations and the mean value.

Sampling Error: The idea that, in taking a sample of a population, not all members of the population are included, so not all possible values are included. High and low outliers are less likely to be included in a sample, but if they are, they may comprise a larger than normal proportion of the data. Either way, the inaccuracy of sampling must be accounted for. Larger samples generally produce less sampling error, but it cannot be completely avoided.

Standard deviation: The square root of the variance — i.e., the average distance between the observations and their mean value, calculated using the formula:

$$\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

where $\bar{x}$ is the mean value for the sample x, and n is the sample size. The standard deviation is a measure of dispersion that is used together with the mean. For example, the standard deviation for the durations of the vowels in the Hillenbrand et al. (1995) study is about 64.5. The standard deviation covers approximately the middle 68% of the data.

9.11 R code

R code and exercises

9.11. R code

Part 1 of this section gives R code that you can use to compute the standard deviation and the IQR, and to make a box plot for the duration values in the Hillenbrand et al. (1995) dataset as a whole, as in the far right panel of Figures 9.9 and 9.10. Part 2 shows how to use the vowel type of each token to group the duration values so as to add the other 12 boxes in Figure 9.10. Part 3 shows how you can use vowel type also to make a vector of medians, or of any other summary measure.

A copy of this R note is saved inside a script that is called `Chapter09.R` and can be found in the list on the course website, at http://hdl.handle.net/1811/77848, or directly at http://kb.osu.edu/dspace/bitstream/handle/1811/77848/Chapter09.R.

Part 1. IQR, standard deviation, and box plot for the whole sample

The data file `Ch09.Textfile1.txt` is a subset of the data in the `vowdata.dat` file that can be downloaded from Jim Hillenbrand's web page. We created this new data file so that we could add the grouping variable for you, and it can be downloaded at http://kb.osu.edu/dspace/bitstream/handle/1811/77848/Ch09.Textfile1.txt. Once you have this file, open R, set the working directory to where you put the `Ch09.Textfile1.txt` file, and read it into a data frame using the `read.table()` function. Use the = operator to assign this data frame to the variable `hill`.

```
hill = read.table("Ch09.Textfile1.txt",
  header=TRUE)
```

Notice that we have specified `TRUE` for the value of the `header` argument, since the other thing we did when we added the grouping variable for you was to give names to that column and to the columns of data from the original data file. As always, you can see what these names are with the `names()` function, like this:

```
names(hill)
```

This should return the following vector of character strings for the column names:

```
[1] "filename" "duration" "F1" "vowel"
```

(We'll explain the value in the "F1" column in Exercise 4.) Since the columns have names, you can refer to the vector of values in any column by name either by putting the column name after the comma in the indexing brackets or by using the $ operator. That is, the following two commands are synonyms:

```
hill[,"duration"]

hill$duration
```

They both refer to the column of vowel durations. That means that you can refer to that vector by name as an argument for the sd() function, like this:

```
sd(hill$duration)
```

This returns the following value, which is the standard deviation for the whole dataset:

```
[1] 64.50101
```

As with any number, you can round it, like this:

```
round(sd(hill$duration))
```

(That's how we got the number "65" for the "st dev" row in Figures 9.9 and 9.10.) You can also use the named vector of durations as an argument for the quantile() function, like this:

```
quantile(hill$duration)
```

This returns the following five-element vector for (1) the minimum duration, (2) the Q1 value, (3) the median, (4) the Q3, and (5) the maximum:

```
  0%  25%  50%  75% 100%

 111  227  271  319  486
```

As noted in R Note 9.2, because the elements of this vector are named, the value that is returned by an equation using the vector can inherit the names of the elements. For example, the value returned by the following equation for the IQR is named:

```
quantile(hill$duration)["75%"]-quantile(
  hill$duration)["25%"]
```

Try typing it to see what you get. The value returned should be:

```
75%

 92
```

This name is misleading, of course, because this value of "92" ms is not the Q3 (the 75% mark), but the IQR. You can suppress that misleading name if you want, using the as.numeric() function, like this:

```
as.numeric(quantile(hill$duration)["75%"]-
  quantile(hill$duration)["25%"])
```

You can also use the named column of duration values as the argument for the `boxplot()` function, in order to get the box plot that we put at the far right in Figure 9.8, like this:

```
boxplot(hill$duration, range=0, at=13,
  xlim=c(0,14))
```

Note that we have specified "`0`" as the value for the `range` argument. This tells R that we want the whisker going down from Q1 to extend to the minimum value in the dataset and the whisker going up from Q3 to extend to the maximum value, so as to indicate the full range of the data. While all box plots use the conventions that the box itself marks of the middle two quartiles, there is not a single fixed convention for what the whiskers mark off. To understand the range of different meanings for what the ends of the whiskers mark, you can research the topic, "box and whisker charts" or just "box plots." For example, you could read the Wikipedia article on box plots at http://en.wikipedia.org/wiki/Box_plot and then type the following R command to see what the different values that you can specify for the `range` argument do:

```
help(boxplot)
```

Note also that we specified "`13`" as the value of the `at` argument. This specification, along with the specification of the vector `c(0,14)` for the `xlim` argument, is how we placed the box at the far right — specifically, at position 13 along an x-axis that has room to show 13 boxes. We did this in order to leave room for the twelve other boxes that we want to add to the picture.

Part 2. Using a grouping variable to make a set of box plots

If you type the `names(hill)` command again, you can see that we've added the grouping variable for `vowel`. Remember that R uses the term `factor` to refer to categorical variables and the term `level` to refer to the different types. You can use the `is.factor()` function to confirm that this added column is a factor, like this:

```
is.factor(hill$vowel)
```

And you can use the `levels()` function to see a list of the types, like this:

```
levels(hill$vowel)
```

The first of these commands returns the following vector of names for the vowel types:

```
 [1] "ae" "ah" "aw" "eh" "ei" "er" "ih" "iy" "oa" "oo" "uh" "uw"
```

These are the ASCII encodings that Hillenbrand and his colleagues used for the 12 vowel types, which we've used as one set of labels for the first 12 boxes in Figure 9.10. You can look at the IPA symbols below the boxes to see how the ASCII encoding corresponds to the IPA symbols that Peter Ladefoged used. (See also Table 8.1 in the chapter on measures of central tendency.) You can use the categorical variable `vowel` as the grouping factor in the box plot, to add the 12 other boxes to your box plot figure, like this.

```
boxplot(hill$duration ~ hill$vowel,
  range=0, at=1:12, add=TRUE)
```

Note how we've specified "TRUE" as the value for the add argument, so that this command doesn't erase the earlier plot. You could also make the 12 boxes first and then add the box for the whole data set afterwards by reversing the order in which you type the two commands and specifying add = TRUE when you make the box for the whole dataset instead, like this:

```
boxplot(hill$duration ~ hill$vowel,
  range=0, at=1:12, xlim=c(0,14))

boxplot(hill$duration, range=0, at=13,
  add=TRUE)
```

It doesn't matter which order you use, as long as you specify xlim=c(0,14) in the first command so as to make the x-axis range wide enough for all of the boxes you want to show, and specify add=TRUE in the second command so as not to erase the part you drew first. Once you have the graph, you can add the horizontal line at the median for the whole dataset, using the abline() function that you used to draw a line at the median value on a histogram, but now specifying the value for the h (i.e., horizontal line) argument rather than the v (vertical line) argument:

```
abline(h=median(hill$duration), lty=2)
```

Part 3. Using the grouping variable to make tables of summary measures

The levels of a factor can also be used to specify groups for vectors of summary values, using the tapply() function. tapply() is kind of a shortened form of the words table and apply, and it basically means to take a function or operation and apply it to each of the levels of a factor, as in a table. This function has three obligatory arguments: (1) a numerical variable, such as the measured duration of each vowel token in the Hillenbrand et al. dataset, (2) an associated categorical variable, such as the vowel type for each vowel token, which can be used as the grouping variable, and (3) an R function that should be applied to each group of values to return a number, such as the mean() function that returns the mean value in a sample or the sd() function that returns the standard deviation of the sample. We used this function to get the group means and standard deviations that we listed below the boxes in Figure 9.10, like this:

```
tapply(hill$duration, hill$vowel, mean)

tapply(hill$duration, hill$vowel, sd)
```

Well, actually we used the tapply() function inside of the round() function, like this:

```
round(tapply(hill$duration, hill$vowel,
  mean))

round(tapply(hill$duration, hill$vowel,
  sd))
```

The first of these two commands returned the following (named) vector of mean values for the 12 samples:

```
ae  ah  aw  eh  ei  er  ih  iy  oa  oo  uh  uw
310 299 316 230 301 299 228 283 302 230 224 274
```

And the second of these two commands returned the following (named) vector of standard deviation values for the 12 samples:

```
ae ah aw eh ei er ih iy oa oo uh uw
55 56 59 53 56 53 54 54 57 54 48 55
```

Some esthetics. You probably noticed that the box plot you created in Parts 1 and 2 of the R code section does not quite look like Figure 9.10. In particular, the 12 box plots that you added to the graph in Part 2 are in a different order than the first 12 box plots in Figure 9.10. To reorder these box plots, we took advantage of the fact that the names of the elements of the returned vector are the names of the levels (the types) in the grouping variable. This allowed us to easily get a list of the vowel types sorted by their median values, like this:

```
vowelsbydur = names(sort(tapply(
  hill$duration, hill$vowel, median)))
```

That is we defined the vector vowels to be the names of the sorted median values that the `tapply()` returned. In other words, `vowelsbydur` takes the median values that `tapply` returns, sorts these median values, and gives us the names associated with the sorted median values. Here's what `vowelsbydur` contains:

```
[1] "uh" "ih" "oo" "eh" "uw" "iy" "ah" "er" "ei" "oa" "ae" "aw"
```

We then redefined the grouping variable so that the types would be ordered by their median values instead of alphabetically by their names, like this:

```
hill$vowel = factor(hill$vowel, levels=
  vowelsbydur)
```

Once we have reordered the list of names of the types like that, we can retype the R code for the box plot in Parts 1 and 2, and the boxes for the 12 groups that were defined by the vowel type in Figure 9.10 will automatically be in the order that we want for the figure.

9.12. Exercises

1. Calculate the IQR and standard deviation for each of the 3 sets of data in a, b, and c. Then answer the questions in d and e.

   a. The following vector lists the durations (in ms) of all of the tokens of the vowel sound **æ** that occurred in the first syllables of the 3 tokens of *absentee* in Exercise 1 in chapter 8, and in 24 other word tokens that were produced in the first five minutes of the conversation with talker s02 in the Buckeye Speech Corpus.

{ 80, 116, 118, 80, 197, 60, 54, 84, 115, 163, 41, 168, 194, 103, 121, 123, 41, 40, 149, 82, 159, 36, 151, 111, 100, 203, 92 }

Note that tokens 1, 2, 10, 11, 13, 14, 15, 17, 18, 20, and 22 are in the function word *that*.

b. The following vector lists the durations (in ms) of all of the tokens of the vowel sound **ə** that occurred in the second syllables of the 3 tokens of *absentee* and in 39 other word tokens in the first five minutes of the conversation with talker s02.

{ 35, 74, 43, 42.826, 68, 46, 33, 35, 32, 57, 36, 40, 80, 30, 29, 23, 19, 32, 57, 26, 83, 92, 30, 39, 50, 57, 30, 26, 55, 43, 34, 53, 40, 81, 107, 47, 32, 30, 39, 23, 30, 19, 22 }

Note that there are 43 values here because the vowel occurs twice in the word *Columbus*.

c. The following vector lists the durations (in ms) of all of the tokens of the vowel sound **i:** that occurred in the third syllables of the 3 tokens of *absentee* and in 37 other word tokens in the first five minutes of the conversation with talker s02.

{ 79, 105, 31, 210, 430, 58, 80, 40, 49, 50, 171, 135, 80, 31, 46, 102, 50, 40, 47, 48, 67, 60, 30, 90, 48, 109, 99, 71, 90, 73, 89, 74, 90, 129, 37, 50, 56, 46, 96, 30, 114 }

Note that there are 41 values here because the vowel occurs twice in the word *repealing*. Also, the fifth token of this vowel is in a token of the word *the* that is prolonged as the talker hesitates and searches mentally for the next word. So it is pronounced as **ði:** (i.e., with the same vowel as in *see*) instead of as the more usual reduced form **ðə**.

d. Which of the three vowel types in a, b, and c shows the smallest IQR? Which shows the smallest standard deviation? Is it the same vector? If it is, use what you've learned in this chapter to explain why this vector shows the smallest dispersion. If it is not the same vector, use what you've learned in this chapter to explain why there is this difference in patterning for the IQR and standard deviation.

e. Which of the three vowel types in a, b, and c shows the biggest IQR? Which shows the biggest standard deviation? Is it the same vector? If it is, use what you've learned in this chapter to explain why this vector shows the biggest dispersion. If it is not the same vector, use what you've learned in this chapter to explain why there is this difference in patterning for the IQR and standard deviation.

2. Download the data file `Ch09.Textfile1.txt` from the website for this textbook. Use the code in Part 1 of Section 9.11 to read the data file into R and assign it to the variable `hill`. Refresh your understanding of how

to use the `hist()` function by remaking the histogram of the vowel durations shown in Figure 9.3. Use the following command to add vertical lines at the minimum, the Q1 value, the median, the Q3 value, and the maximum.

```
abline(v=quantile(hill$duration),lty=3)
```

Open a new plotting window and use the `hist()` function again to make a second copy of the histogram. Then use the `mean()` function, the `sd()` function, and the `abline()` function to add a vertical line at the mean, at the mean minus one standard deviation, and at the mean plus one standard deviation.

3. As in exercise 2, download the data file `Ch09.Textfile1.txt` from the website for this textbook and use the code in Part 1 of Section 9.11 to read the data file into R and assign it to the variable `hill`. Refresh your understanding of how to use the `hist()` function and then do the following exercises.

   a. Use the `subset()` command to make paired histograms of the durations of the short vowel **ɪ** in *hid* versus the counterpart long vowel **i:** in *heed*. Adapt the `abline()` code given in exercise 2 to add the same set of five cursors to each of the histograms, to reproduce the top and bottom panels in Figure 9.2. (The ASCII code conventions that Hillenbrand et al. used for the different vowels are given in Table 8.1 in Chapter 8. The relevant strings are "`ih`" for the vowel **ɪ** in *hid* and "`iy`" for the vowel **i:** in *heed*.)

   b. Repeat the above for the short vowel **ʌ** in *cut* versus the counterpart long vowel **ɔ:** in *caught*. (These are "`uh`" and "`aw`" in the Hillenbrand et al. ASCII conventions.)

   c. Repeat the above for the short vowel **ɛ** in *bet* versus the counterpart long vowel **eɪ** in *bait*. (These are "`eh`" and "`ei`" in the Hillenbrand et al. ASCII conventions.)

   d. Repeat the above for the short vowel **ʊ** in *book* versus the counter long vowel **u:** in *boot*. (These are "`oo`" and "`uw`" in the Hillenbrand et al. ASCII conventions.)

   e. Compare the four pairs of histograms and decide which of the short versus long pairs have the least overlap in the tails of their distributions and the most separation between their medians. Look at the box plots in Figure 9.8 and explain whether this was easier to decide from the box plots or from the histograms and why.

4. The physical property that we listen to when we hear the distinct quality of a vowel is a resonance, like the deeper resonance that distinguishes the sound of a cello from the sound of a violin even when they are playing the

same note on the musical scale. Resonances are caused by sound waves bouncing off of surfaces that increase their amplitude at a certain pitch, or frequency. In a vowel, those resonances are caused by sound waves bouncing off of the inside of the vocal tract, and are called formants. Formants are numbered such that the lowest formant is called the first formant (or F1, as in the third column of the Hillenbrand text file we are looking at), the next highest is called the second formant (F2), and so on. F1 usually decreases for vowels where the tongue is in a higher position in the oral cavity, F2 usually indicates vowel frontness, and so on. The main distinction in vowel quality between the vowel **iː** in *heed* and the vowel **æ** in *had* is that **iː** has extremely low first formant values (because the tongue is in a very high position) and **æ** has extremely high first formant values (because the tongue is in a very low position). And the first formant values for the "indistinct" reduced vowel **ə** are "neutral" – i.e., about halfway in between the values for the two "extreme" vowels.

In this exercise we will explore the following statement that Peter Ladefoged makes on p. 37 of his book *Vowels and Consonants*: "If a language uses a still larger number of vowels [larger than 7], then they may be distinguished not only by differences in their first formant values but also by other differences such as length. English has several vowels that are kept distinct in this way. For example, the vowel in *heed* is different from that in *hid* not only by having a lower first formant, but also by being longer."

a. Use the code in Section 9.11 to reproduce the duration box plots for the 12 vowel types in Figure 9.9.

b. Then adapt the code to make box plots for the first formant values in this dataset. That is, use vowel as the grouping variable not to explore the variation in duration across the twelve vowel types, but to explore the variation in first formant (F1), like this:

```
vowelsbyF1 = names(sort(tapply(hill$F1,
    hill$vowel, median)))

hill$vowel = factor(hill$vowel, levels=
    vowelsbyF1)

boxplot(hill$F1 ~ hill$vowel, range=0,
    at=1:12, xlim=c(0,14))

boxplot(hill$F1, range=0, at=13, add=TRUE)
```

c. Compare the two sets of box plots. Look particularly at the box plots for the vowels **ɪ** and **iː** in order to explain whether the distributions of their duration values and of their first formant values support Ladefoged's statement.

5. In this exercise, you will compare the distributions of spelled word lengths in three subsets of the *Hoosier Mental Lexicon*.

a. Read in the data for the *Hoosier Mental Lexicon*, and calculate a vector of word lengths, just as you did for Exercise 5 in Chapter 8 on measures of central tendency (`Ch08.Textfile1.txt`).

```
hml$size = "medium"
```

The first line of code creates a new variable called `size` to the dataset and assigns the value `medium` to every element in the variable. This default allows us to replace it for the `small` and `big` words, and leave it for all other words.

Use the == operator (for identifying subsets of data) and the logical operator "&" to add a new level of familiarity:

```
hml$size[hml$familiarity==7 &
    hml$frequency > 50] = "small"
```

This line of code changes the value from `medium` to `small` for all elements in the `size` column that have a `familiarity` of 7 (the highest familiarity rating) and a `frequency` greater than 50 (to exclude well-known but little-used words).

```
hml$size[hml$familiarity < 3] = "big"
```

This changes the value from `medium` to `big` for all elements in the variable that have a `familiarity` smaller than 3.

```
hml$size = factor(hml$size, levels=
    c("small","medium","big")
```

This reorders the levels of the `size` variable from the default alphabetical order to one that makes more sense: `small`, `medium`, and `big`.

b. Make a box plot of word lengths (use the `nchar()` function on the `HMLbet` column) that has boxes for each of the three subsets of word size types (i.e. the `size` variable). Also add a box plot for the distribution of word lengths in the *HML* as a whole.

c. Use `tapply()` to calculate the mean and standard deviation for each of the three subsets and for the *HML* as a whole.

d. Describe the patterns that you see. Are there differences among the mean lengths for the three subsets? Is the median for any of the word types substantially shorter than or substantially longer than the median length for the *HML* as a whole? Are the standard deviations for the subsets similar to the standard deviation for the *HML* as a whole or are they different? Explain how what you observe makes sense linguistically.

9.13. References and data sources

The numbers in Figures 9.3, 9.4, 9.6, 9.9, 9.10, and in the bottom panel of Figure 9.1, as well as in exercises 2 through 4 are from the Northern Cities vowel study described in:

James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97 (5), 3099-3111.

The quotations from Ladefoged are from his book:

Peter Ladefoged (2005). *Vowels and Consonants: An Introduction to the Sounds of Languages*. 2nd Ed. Blackwell.

The numbers in the top panel of Figure 9.1 and in Figures 9.5 and 9.7 are from the Buckeye Speech Corpus, and specifically from the data in the Tyson (2005) study, both of which were described in Chapter 3.

Exercise 5 uses data from the *Hoosier Mental Lexicon*, which is described in Chapter 2.

Some of the many experiments showing that native speakers of English listen to the vowel quality and not just the duration in differentiating the short vowel **ɪ** in *hid* from the long vowel **iː** in *heed* are reported in the following paper.

Robert Allen Fox, James Emil Flege, and Murray J. Munro (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *Journal of the Acoustical Society of America*, 97(4): 2540-2551

Appendix 9A. Step-by-step through the IQR and standard deviation

In this appendix, we'll work through the logic and the arithmetic for calculating the two measures of dispersion that we introduced in this chapter using a very small sample so that you're not overwhelmed by a long stream of numbers. The sample we'll use is the vector of measured durations of the vowel and consonant sounds in the first tokens of the word government in the interview with speaker s10 in the Buckeye Speech Corpus. Figure 9.11 repeats the first panel of Figure 8.1 from the chapter on measures of central tendency, to remind you what the pattern of variation is in this sample.

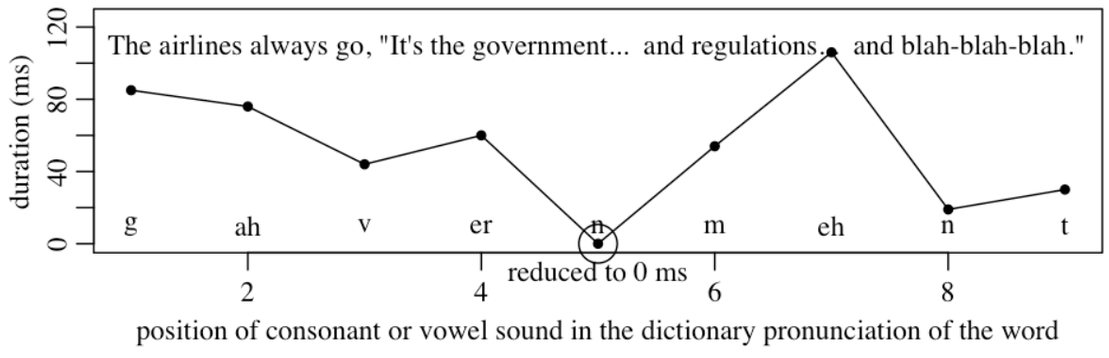


Figure 9.11. (= the top panel of Figure 8.1 in Chapter 8) Time plot of the measured durations of the vowel and consonant sounds in the first tokens of the word *government* in speaker s10's interview in the Buckeye Speech Corpus.

Part 1. Interquartile range

In order to calculate the interquartile range, you first need to find the first and third quartiles. To see where the first and third quartile are, start by sorting the vector, just as you did to see where the median is. The displays in (A1) illustrate this, using the sample of 9 values that includes the 0 ms for the **n** that the speaker did not pronounce, and then using the sample of 8 values that excludes that deleted consonant's duration. That is, as in identifying the median value, the exact method differs depending on whether the number of elements (the ***n*** of the sample) is an odd number, as in (A1a), or an even number, as in (A1b).

(A1) a. Start by sorting the vector of durations of sounds in the first token of government in the interview with speaker s10 in the Buckeye Speech Corpus.

{ 0 , 19 , 30 , 44 , 54 , 60 , 76 , 85 , 106 }
| **30** | **76** |

The minimum is the 1st value, median value is the 5th value, and the maximum is the 9th value. The first quartile (Q1) is in the middle between the minimum and median values, that is, in the middle between the 1st and 5th values. This middle value, that is, the first quartile, is the 3rd value. The third quartile (Q3) is in the middle between the median and maximum values, that is, in the middle between the 5th and 9th values. This middle value, that is, the third quartile, is the 7th value. So, the first and third quartile values are the 3rd and 7th values out of 9 values. The interquartile range is 76-30=46 ms.

b. The same vector as in (a), but excluding the 0 ms duration of the deleted **n**.

{ 19 , 30 , | 44 , 54 , 60 , 76 **|** , 85 , 106 }
| **40.5** 57 **78.25** |

The minimum value is the 1st value, median value is ½ of the way between the 4th and 5th values (at 57), and the maximum value is the 8th value. The first quartile (Q1) is in the middle between the minimum and median values, that is, in the middle between the 1st value and ½ of the way between the 4th and 5th values. This middle value, that is, Q1 is ¾ of the way between the 2nd and 3rd values. The third quartile (Q3) is in the middle between the median and maximum values, that is, in the middle between ½ of the way between the 4th and 5th values and the 8th value. This middle value, that is, Q3 is ¼ of the way between the 6th and 7th values. The interquartile range is 78.25-40.5=37.75 ms.

Part 2. Standard deviation

In (A2), we work through the various parts of the equation for calculating the standard deviation, working outward from the $(x - \bar{x})$ inside the summation, just as we did when we de-

scribed the standard deviation measures in Section 9.3, in the paragraph just under equation (1).

(A2) We start with the vector of numbers, which is:
$\boldsymbol{x}$ = { 85, 76, 44, 60, 0, 54, 106, 19, 30 }

1. The first thing we need to do is to get the number of elements $\boldsymbol{n}$ and the mean $\bar{x}$.

   n = 9, $\bar{x}$ = (85+76+44+60+0+54+106+19+30)/ 9 = 474/9 ≈ 52.66667 ≈ 53

2. Then we can plug the appropriate numbers into formula (1) which we repeat here:

standard deviation:

$$= \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{(9-1)} * \left((85-53)^2 + (76-53)^2 + (44-53)^2 + (60-53)^2 + (0-53)^2 + (54-53)^2 + (106-53)^2 + (19-53)^2 + (30-53)^2\right)}$$

3. Then we simplify:

$$= \sqrt{\frac{1}{(9-1)} * \left((32)^2 + (23)^2 + (-9)^2 + (7)^2 + (-53)^2 + (1)^2 + (53)^2 + (-34)^2 + (-23)^2\right)}$$

4. Simplify again:

$$= \sqrt{\frac{(1024 + 529 + 81 + 49 + 2809 + 1 + 2809 + 1156 + 529)}{(9-1)}}$$

5. Simplify one last time:

$$= \sqrt{\frac{8987}{8}}$$

$$= 33.51679$$

6. So our standard deviation is about 33.

R Note 9.4 – standard deviation, step by step

Here is how you can use the arithmetic operators of R to do the above step by step calculations. First assign the vector to a variable and its mean to a variable, so that you can refer to them by name:

```
dur = c(85,76,44,60,0,54,106,19,30)

dur.m = mean(dur)
```

Here is the inner part of the summation, i.e. $(x_i - \bar{x})^2$:

```
(dur - dur.m)^2
```

Plug that into the `sum()` function to get the $\sum_{i=1}^{n}(x_i - \bar{x})^2$ part:

```
sum(dur - dur.m)^2)
```

Use the `length()` function to count the number of elements in the vector so that you can divide the sum calculated in the previous by (n – 1):

```
sum(dur - dur.m)^2)/(length(dur)-1)
```

Embed that as an argument of the `sqrt()` function to take the square root:

```
sqrt(sum(dur - dur.m)^2)/(length(dur)-1))
```

Of course, you can also just use plug the vector of durations directly into the `sd()` function, which we said in R Note 8.2 takes a vector of values as its argument, like this:

```
sd(c(85,76,44,60,0,54,106,19,30))
```

This is the value that the function returns:

```
[1] 33.51492
```

This is not quite the same as the value we calculated by hand, because we rounded the mean in order to make it easier to work with.

Appendix 9B. Filled pause

In making Figure 9.5, we chose an x-axis range that excluded the bars of the histogram covering the durations for the longest 10% of the tokens in the sample of the **ə** vowel from occurrences of the words *the* and *a* in the interview with speaker s10. We suggested that these tokens were outliers because they were places where the speaker was using the word as another way of filling up a pause while he hesitated to find the next word.

In order to evaluate this explanation, we decided to compare the distribution of durations measured for these excluded tokens to the distribution of durations measured for "words" that were transcribed as "true" cases of the filled pause. The top panel of Figure 9.12 zooms in on the outliers. The bottom panel is a histogram of the durations of all tokens of the **ʌ**-like vowel in the interview with speaker s10 that were identified as a filled pause — i.e., everything that the transcribers of the Buckeye Speech Corpus spelled out as one of the "true" hesitation words uh and um. In both cases, the values are quite spread out, showing an extremely flat distribution with no central peak at all.

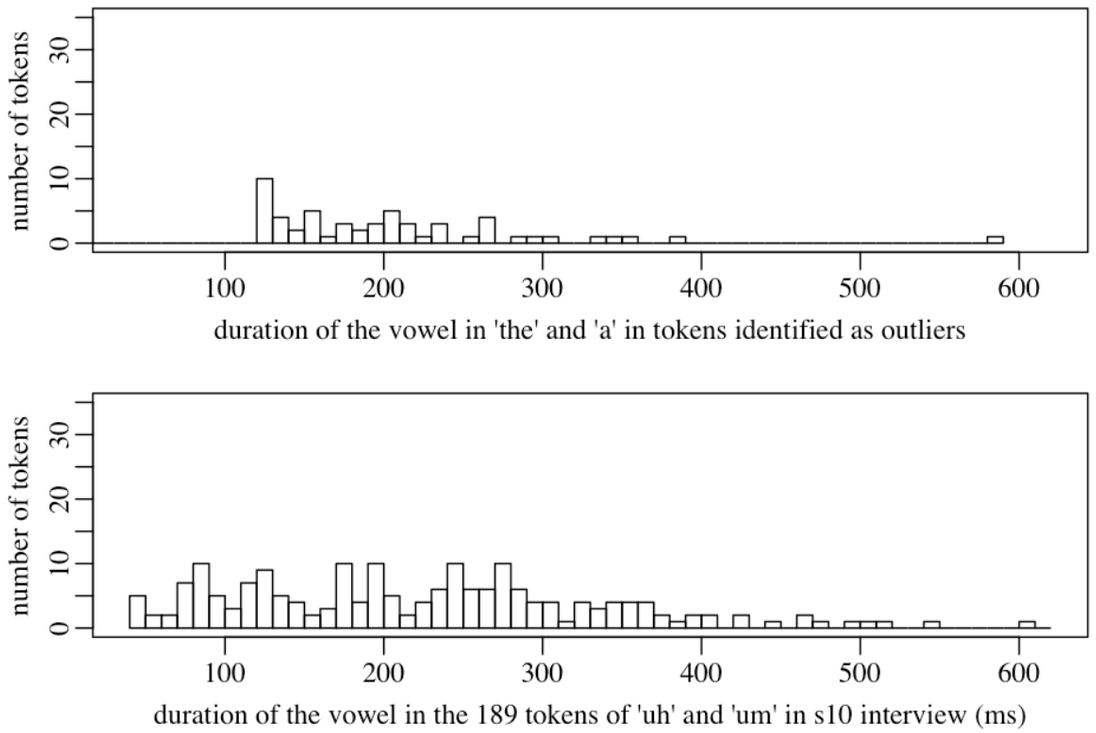


Figure 9.12. Durations of the vowel tokens excluded from Figure 9.4 (top) compared to the durations of the vowels that were transcribed as filled pauses *uh* or *um* in the interview with speaker s10 in the Buckeye Speech Corpus (bottom).

This distribution makes sense if we consider the function (the "meaning") of hesitation words such as *uh* and *um*. The word indicates that the speaker is pausing the sentence momentarily as s/he figures out what to say next. The vowel in the word doesn't contrast with any other word that could have occurred in that place in the sentence, in the way that the vowel ʌ in *cut* contrasts this word to the word *caught*, or the way that the vowel ɪ in *hid* contrasts this word to the word *heed*. The duration of such a filled pause is constrained only by how long the speaker needs to pause until he or she figures out what to say next. We can compare this function (i.e. figuring out what to say next) to the function of the duration of the vowels in words such as *cut* versus *caught* or *hid* versus *heed* (which distinguishes similar-sounding words). The Fox, Flege, and Munro (1995) study suggests that the vowels in these words are kept distinct partly by differences in their durations, in addition to the vowel quality cue. See exercise 4.