SLHS 1302

Chapter 7
Measures of Dispersion

Background knowledge

• Without knowing something about how data is dispersed, measures of central tendency may be misleading.

For example, a residential street with 20 homes on it having a mean value of $200,000 with little variation from the mean would be very different from a street with the same mean home value but with 3 homes having a value of $1 million and the other 17 clustered around $60,000.

• Measures of dispersion provide a more complete picture. **Dispersion measures include the range, average deviation, variance, and standard deviation**.

Measures of Dispersion Defined

- **Range**

  The simplest measure of dispersion is the **range**.

  The range is calculated by simply taking the difference between the maximum and minimum values in the data set.

  However, the range only provides information about the maximum and minimum values and does not say anything about the values in between.

Measures of Dispersion Defined

- **Average Deviation**

  Another method is to calculate the average difference between each data point and the mean value, and divide by the number of points to calculate the **average deviation** (mean deviation).

  However, performing this calculation will result in an average deviation of zero since the values above the mean will cancel the values below the mean. If this method is used, the absolute value of the difference is taken so that only positive values are obtained, and the result sometimes is called the *mean absolute deviation*.

  The average deviation is not very difficult to calculate, and it is intuitively appealing. However, the mathematics are very complex when using it in subsequent statistical analysis. Because of this complexity, the average deviation is not a very commonly used measure of dispersion.

Measures of Dispersion Defined

- **Variance and Standard Deviation**

  A better way to measure dispersion is to square the differences before averaging them. This measure of dispersion is known as the variance, and the square root of the variance is known as the standard deviation.

  The standard deviation and variance are widely used measures of dispersion.

Standard deviation and variance

- A commonly used measure of dispersion is the **standard deviation**, which is simply the square root of the **variance**. $\sqrt{\frac{1}{(N-1)}\sum_i (x_i - \bar{x})^2}$
- The variance of a data set is calculated by taking the arithmetic mean of the squared differences between each value and the mean value. Squaring the difference has at least three advantages:
  - Squaring makes each term positive so that values above the mean do not cancel values below the mean.
  - Squaring adds more weighting to the larger differences, and in many cases this extra weighting is appropriate since points further from the mean may be more significant.
  - The mathematics are relatively manageable when using this measure in subsequent statistical calculations.

Exercise

- Calculate variance of the following sample
  - 68, 55, 98, 87, 89, 92, 56
  - x = c(68, 55, 98, 87, 89, 92, 56)
  - v = var(x)
- Calculate standard deviation for this data set
  - s = sd(x)

Research Question

- How long can a short vowel be before it should be considered a long vowel?
  - **Life's a *beach where sheet happens*.**
  - In Japanese, vowel duration is also very important
    - Obasan: aunt
    - Obaasan: grandma

Data on vowel durations in English

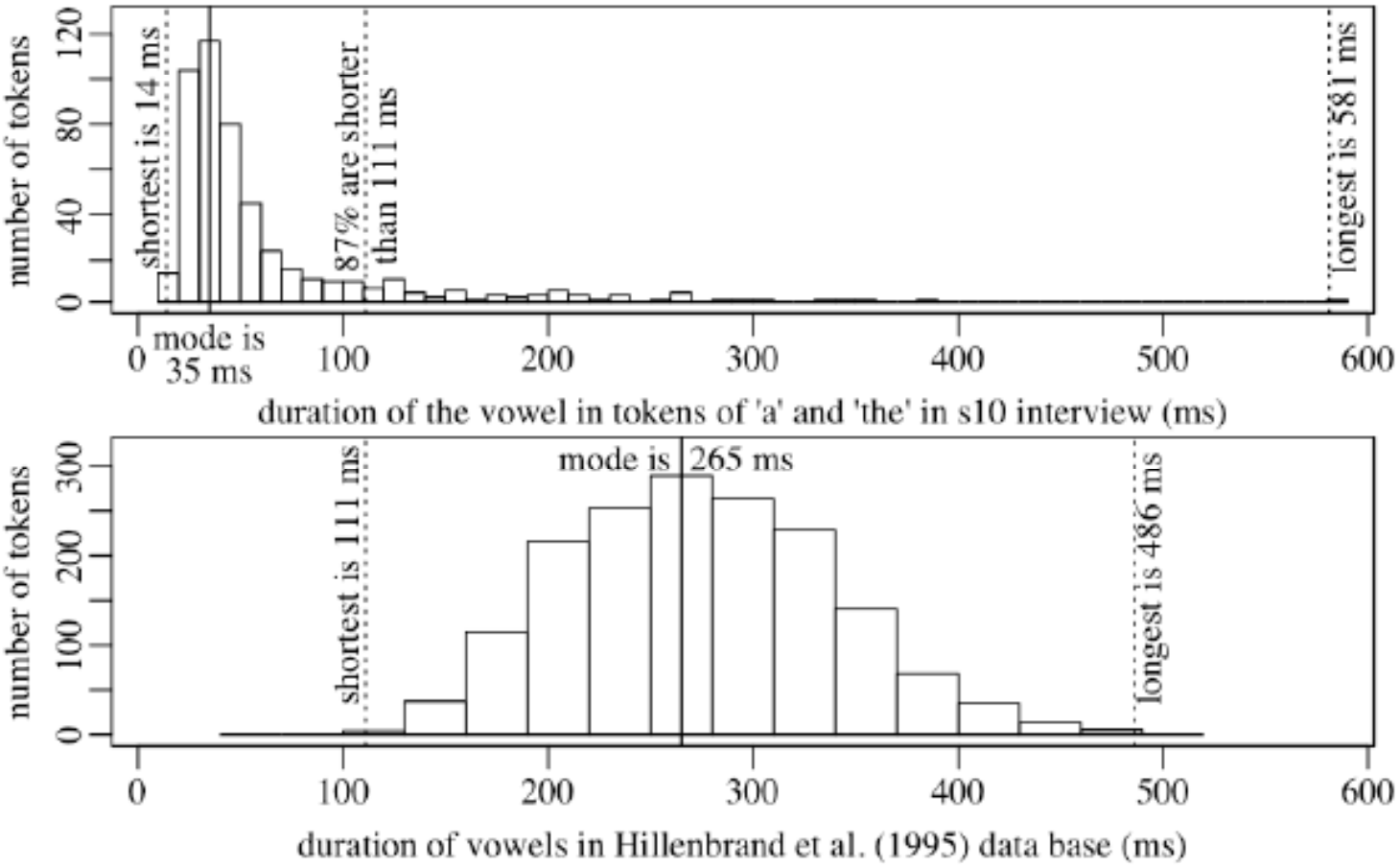

Figure 7.1. Histograms of measured durations of the vowel in all 488 tokens of *the* and *a* in speaker s10's interview in the Buckeye Speech Corpus (top panel) and in 1668 tokens of **h_d** words pronounced by the 139 speakers in the Hillenbrand et al. (1995) study (bottom).

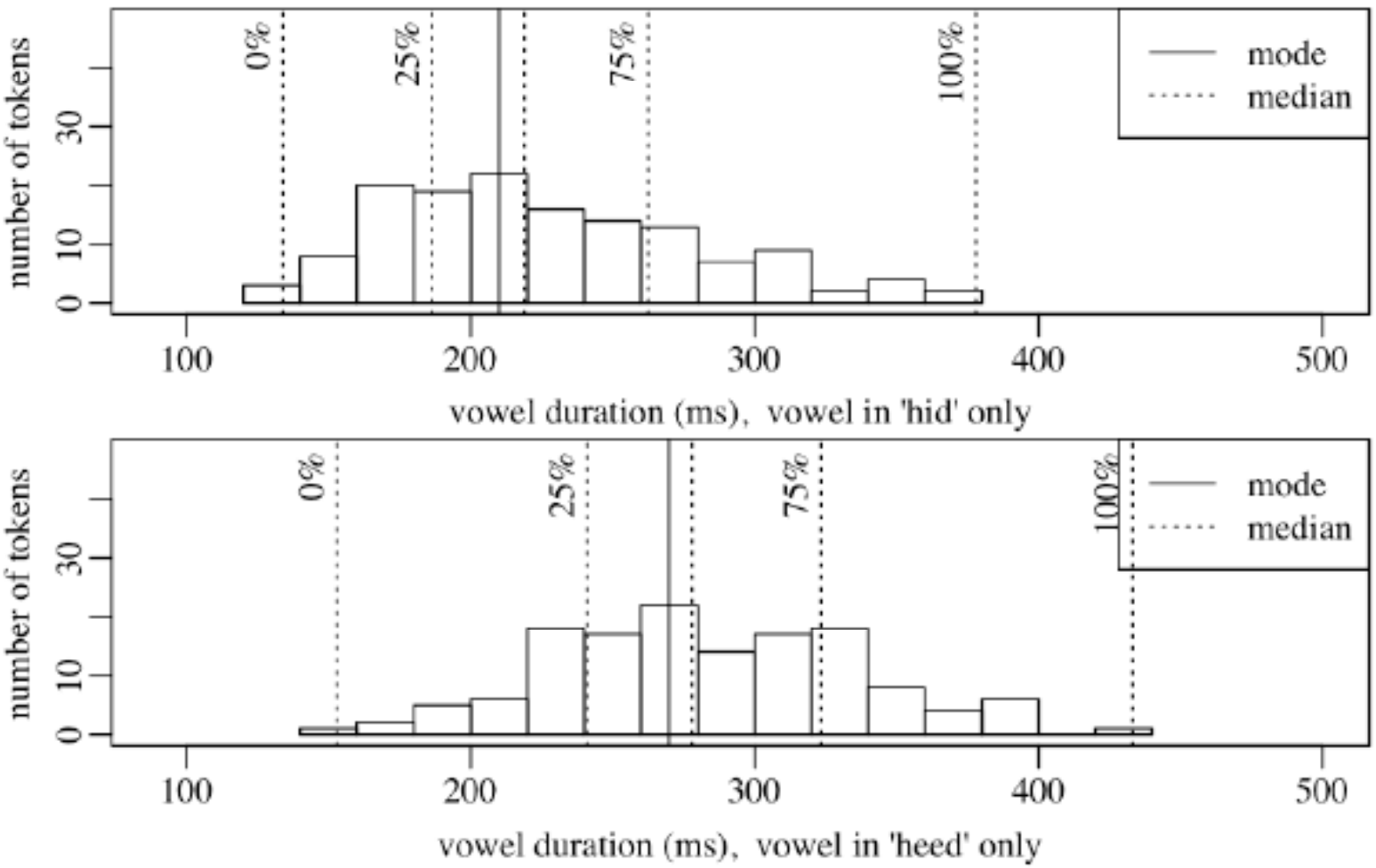

Figure 7.2. Histograms of measured durations of the vowel in the 139 tokens of *hid* (top) and the 139 tokens of *heed* (bottom) in the Hillenbrand et al. (1995) study.

Range and Interquartile Range

- ***Interquartile range (IQR)*** is the spread of values that covers the middle half of the observations, excluding the lowest 25% and the highest 25%.

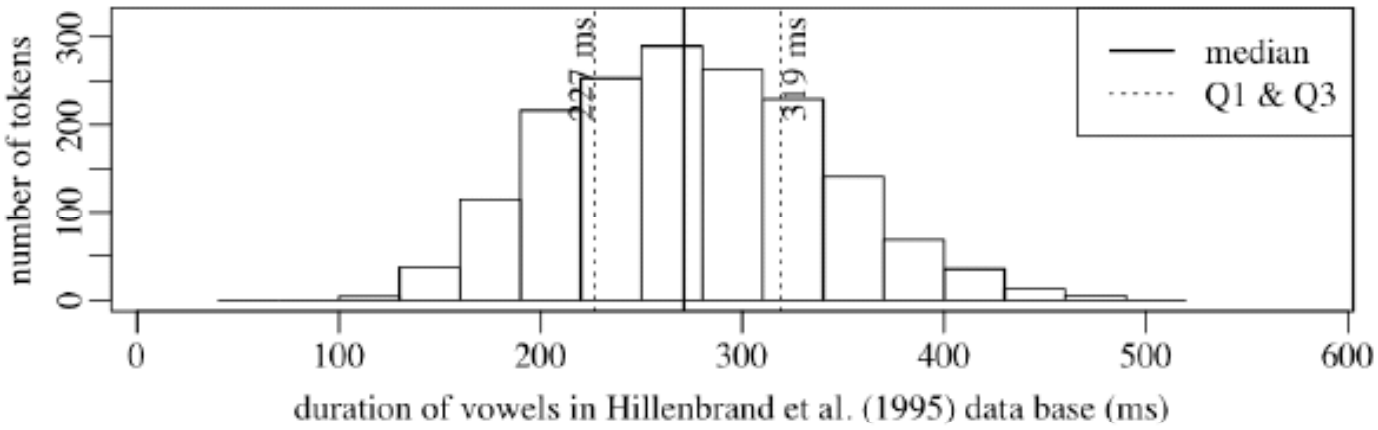

Figure 7.3. Same histogram as in the lower panel of Figure 7.1, with the median and 1st and 3rd quartiles marked by the different vertical lines, as indicated in the legend.

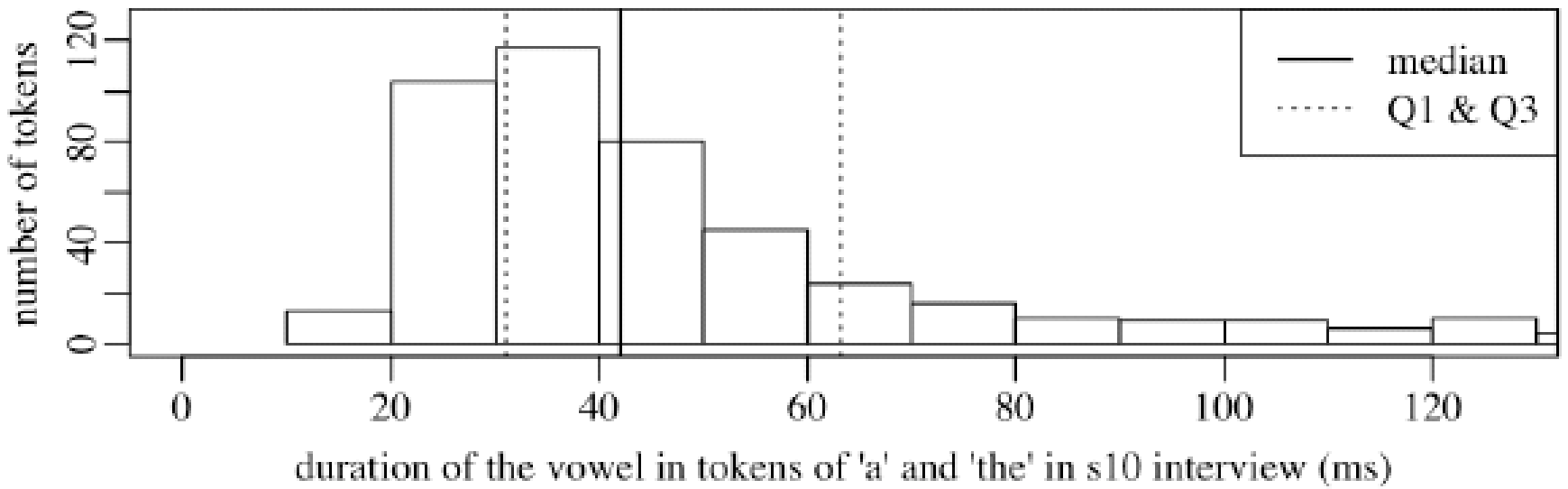

Figure 7.4. Same histogram as in the top panel of Figure 7.1, but with the x-axis zoomed in to exclude the longest 10% of the data, and the median and 1st and 3rd quartiles marked by the solid and dashed vertical lines, as indicated in the legend.

Standard deviation

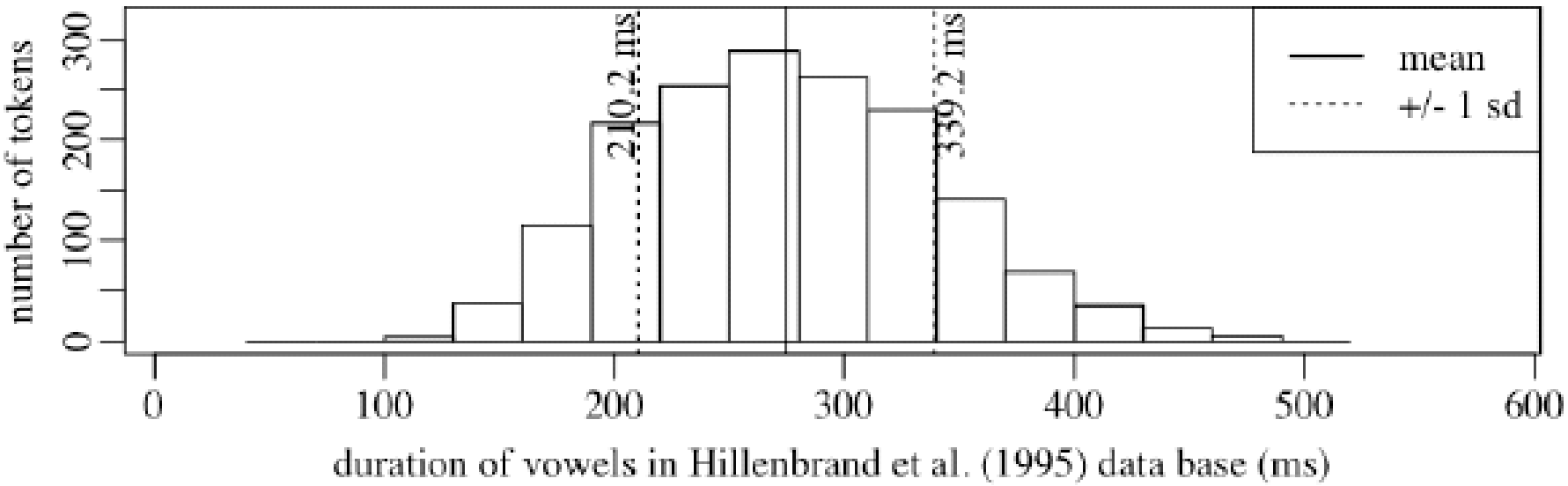

Figure 7.5. Same histogram as in Figure 7.3, but with the mean±1 standard deviation marked by the different vertical lines, as indicated in the legend.

Boxplot

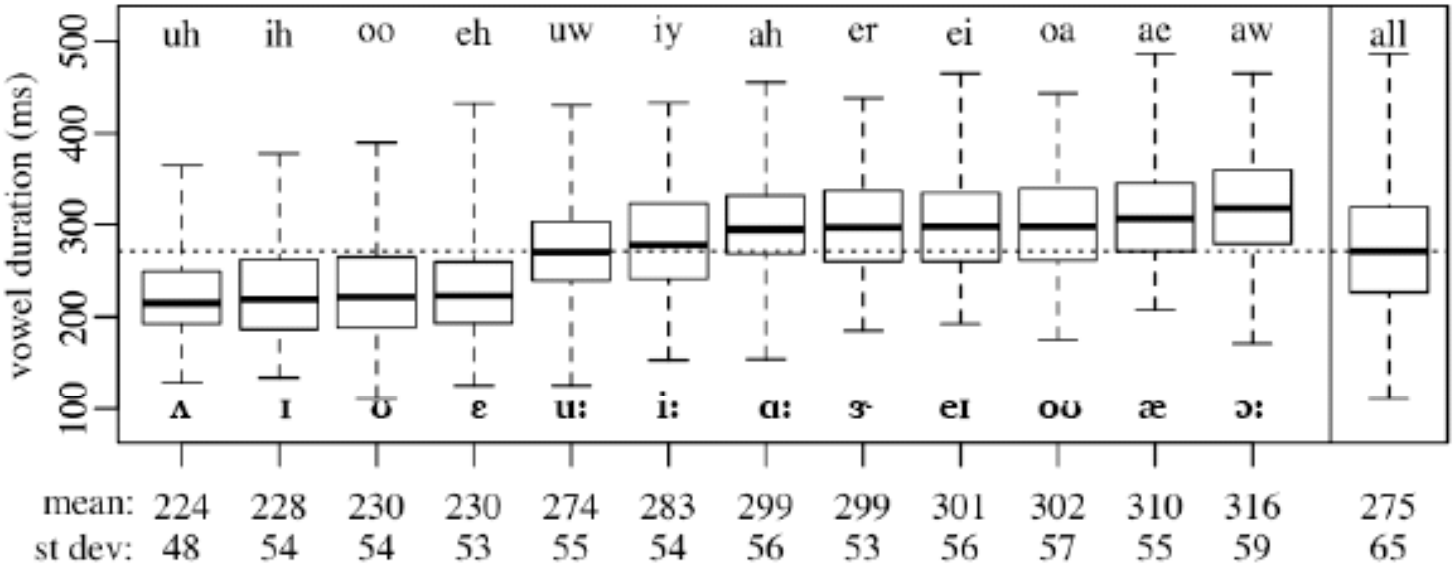

Figure 7.7. Box plots for the vowel duration values in the Hillenbrand et al. (1995) data, for subsets that are defined by "vowel type" (first 12 boxes) and for all of the dataset (box on the far right). The horizontal dashed line marks the median for the overall dataset. The rows of numbers below the x-axis are the means and standard deviations for the 12 subsets of values and for the sample as a whole.

Summary

- To describe the distribution of values in a dataset, measures of central tendencies and measures of dispersion are equally important.
- To test whether the data are significantly different from expectation, we can conduct hypothesis testing using different statistical tools, including binomial test, t-test, ANOVA, etc.

R code

- quartile (x) gives the values for 0%, 25%, 50%, 75%, 100% of the x data set.
  - Interquartile range would be

    lower bound: quartile (x) [“25%”]

    upper bound: quartile (x) [“75%”]
- Standard deviation
  - sd (x)
- Boxplot
  - boxplot (x)