SLHS 1302

Chapter 2
Counting Numerical Data

Key terms

- *Quantitative: A way of describing data using numbers (quantities).*
- *Qualitative: A way of describing data using descriptions and concepts (qualities), rather than* numbers.
- *Numerical variable: A set of recorded observations that are inherently ordered on a scale of* numbers. Example: height (for people) or number of letters (for spelled forms of words).
- *Continuous values: Values that you can always subdivide into smaller units. The number of* distinct values for a continuous variable is in theory infinite. Example: height. If your range is 70-170 cm, you may have an infinite number of possible fractions or decimals, such as the possible values {72.5, 72.49, 72.486458, 72 ….}

◘ *Discrete values: Values that are always integers and can't be subdivided into fractions. For a* definite range of values, you will have a finite number of possible values. Example: number of letters. If your range is 1-15 letters, you can only have 15 possible values {1,2,3,4...15}.

◘ *Histogram: A special type of bar plot used to represent the frequencies for different values of* a discrete numerical variable. Specifically, the bars of a histogram must be arranged in strict numerical order and there must be a space left for an "invisible" bar for any value where there are no tokens within the range of values for the sample.

Introduction

- Research question
  - How do the sizes of 'big words' and 'small words' compare to the typical sizes of English words overall?"

- Big words, small words
  - What determines the perceived "size"?
    - Aver, onus, rive, wily
    - Caterpillar, refrigerator, strawberry, kindergarten

Concepts

◻ Categorical (nominal) vs. Ordinal

- ◻ Categorical variable: a set of recorded observations of group or category membership
  - ◻ Example: Digraph gh has for possible pronunciations: /f/, /g/, /p/, and ∅
- ◻ Ordinal variable: ranked categorical variables
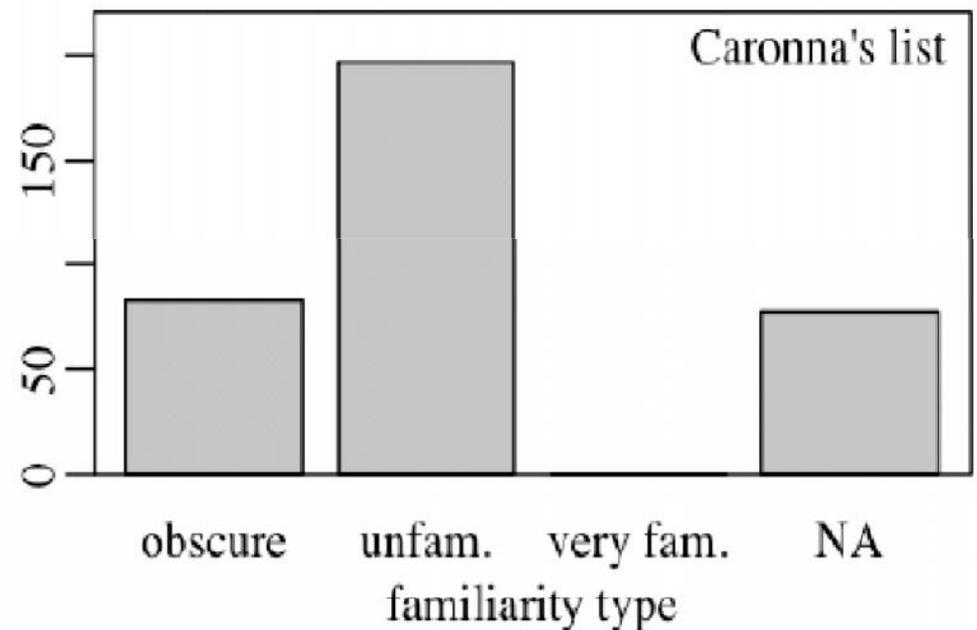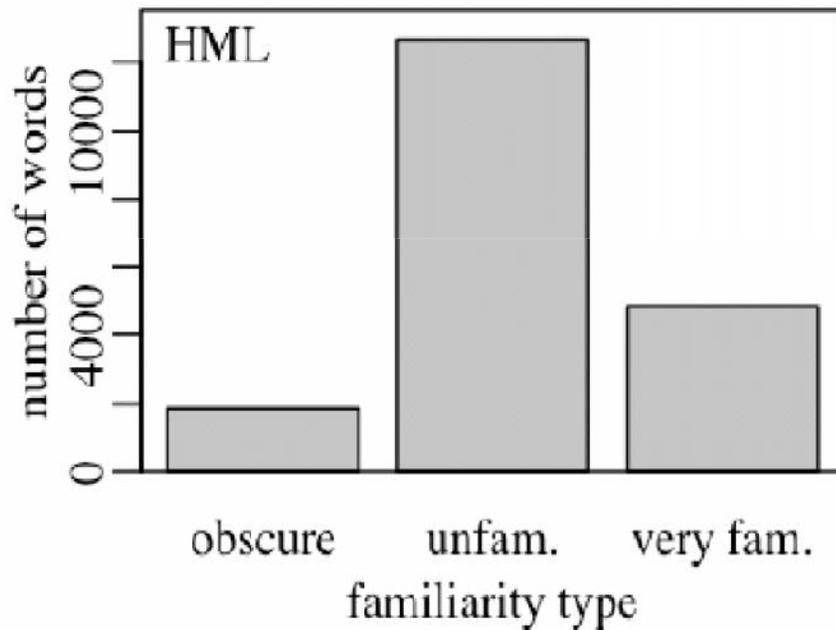 - ◻ Example: familiar, unfamiliar, unknown
  - ◻ Ranking can be expressed in numbers.
    - ◻ Example: familiarity rating for words (1-7)
    - ◻ Example: speech accent rating (1-5)

big words against ordinary words in terms of familiarity rating

Hoosier Mental Lexicon
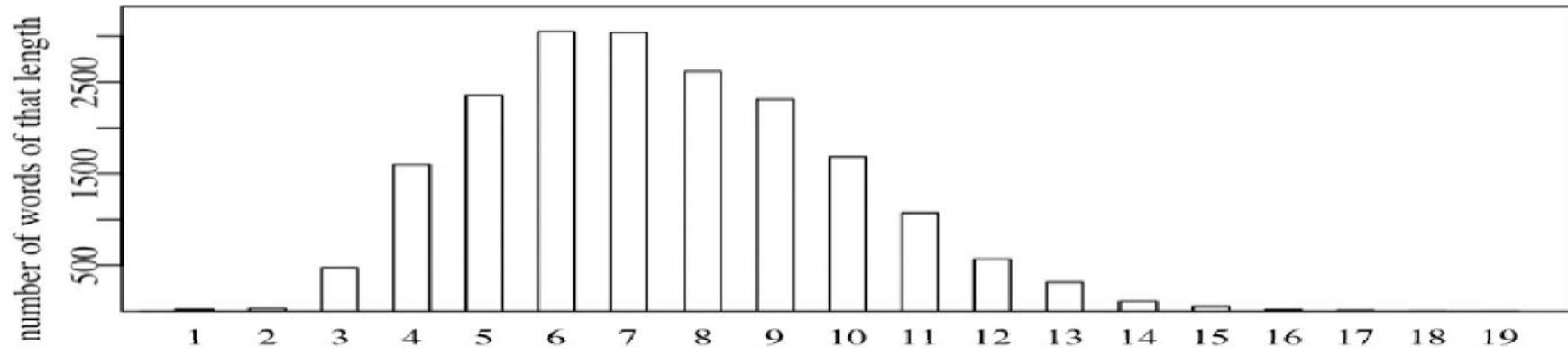
big words against ordinary words in terms of word length

Figure 2.2. Histogram of word lengths (in number of letters) in the *Hoosier Mental Lexicon*.
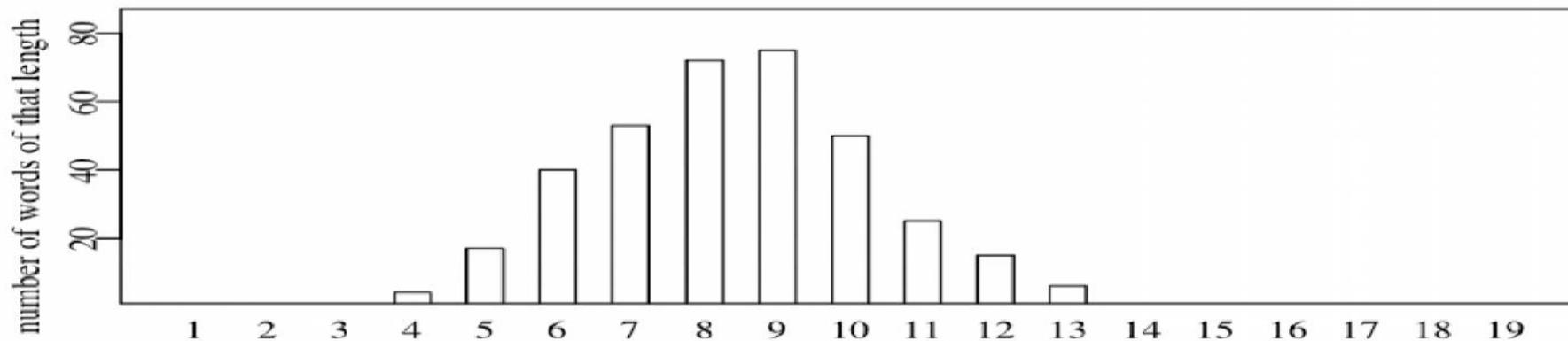

Figure 2.3. Histogram of word lengths in Carol Caronna's collection of "big words."
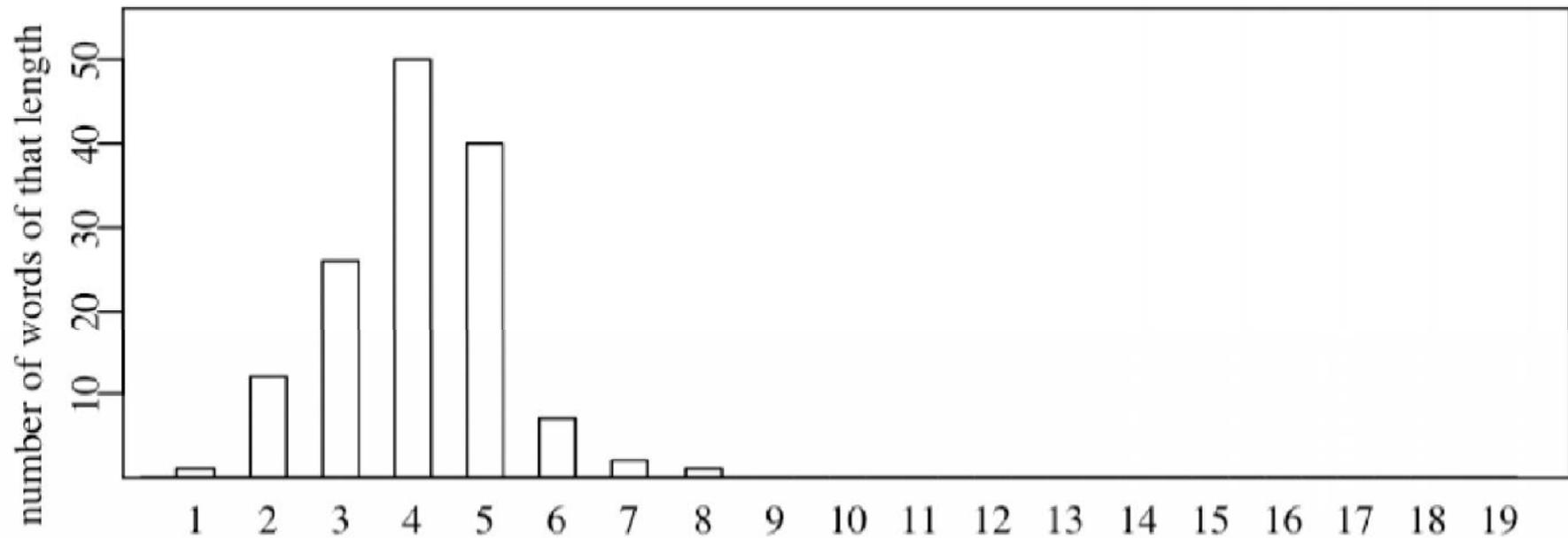
Word length on small words

Figure 2.4. Histogram of word lengths in the title and first four paragraphs of Richard Lederer's essay on "The Case for Short Words" with each different word counted only once.

Answering our research question

◻ Nine letters is the most common size of our "bigwords" sample, whereas four letters is the most common size for our "small words" sample.

◻ The largest bar in the histogram showing word lengths in the *Hoosier Mental Lexicon is 6 letters, closely followed* by the bar for 7 letters.

Summary

◻ When we talk about data, we are talking about sets of reliably recorded observations that we can use to reason numerically about the world. Some records are inherently quantitative, derived by counting or measuring something.

◻ We can derive useful quantitative measures of such categorical variables by counting the number of tokens for each type.

◻ Plotting the numbers that we get in bar plots and histograms is a useful way to look for generalizations about the population.