Ling 105
Sounds of Language

Thursday, October 31, 2024

Kevin Ryan

Reading for this week

- A short write-up of Wilcoxon tests (last time) and correlation tests (this time) can be found in Douglas et al.'s (2024) *An introduction to R* §6.1–2, online at **intro2r.com**
- If you want a reference for regular expressions (today), you can see *R for data science (2e)* chapter 15, but you only need to know what is in today's slides

Heads-up: scientific notation

- Say $p = 4.52\text{e-}6$
- p cannot be greater than one!
- $= 0.00000452$
- Any value printed with "e-" (i.e. $\times 10^{-x}$) will be close to zero and thus highly significant

select()

- Keep only the specified columns

  select(Word, Duration)

- Or drop any column(s) indicated as “negative”

  select(−Duration)

mutate()

- Change the data frame
- Add a column by specifying it

  mutate(log_dur = log(Duration))
- If the specified column already exists, it will be overwritten

  mutate(Duration = log(Duration)

if_else()

- Dichotomize or recode a variable
- if_else(CONDITION, OUTCOME_IF_TRUE, OUTCOME_IF_FALSE)
- word_size = if_else(Duration > mean(Duration), “long”, “short”)

str_detect()

- Useful not just for filtering, but for defining groups
- Define a new column indicating whether a word is vowel-initial (in pronunciation)
- Vowels are encoded as *i*, *u*, *I*, *U*, etc.
- Condition: str_detect(Surface, "^[iuIUQ{2645E13VPHF]")
- The search string "^[iuIUQ{2645E13V]" is a **regex** (regular expression)
- ^ is the left anchor ("words beginning with")
- [...] is a disjunction (*i* or *u* or *I*...)
- If any of the elements in the disjunction is multiple characters, use (...) with bars, e.g.
  1. (ch|th|wh|ph|sh) is the same as
  2. [ctwps]h

Regexes continued

- In a regex, . matches any character

❶ Get every orthographic word of three letters that begins with *b* and ends with *d*

- * means any number (including zero) of the preceding element

❷ Get every orthographic word that begins with *b* and ends with *d*

❸ Get every word containing at least three *b*s

- For negation (the complement set), use [^...]

❹ Get every word pronounced without any vowel

Regex substitutions

- We can also alter strings using regexes
- str_replace(SOURCE, TARGET, REPLACEMENT)
- (Or str_replace_all() to replace all matches)
- TARGET and REPLACEMENT are both regexes

1. What are all attested word-initial onsets?
2. Which are three consonants? What phonological generalization can we draw?

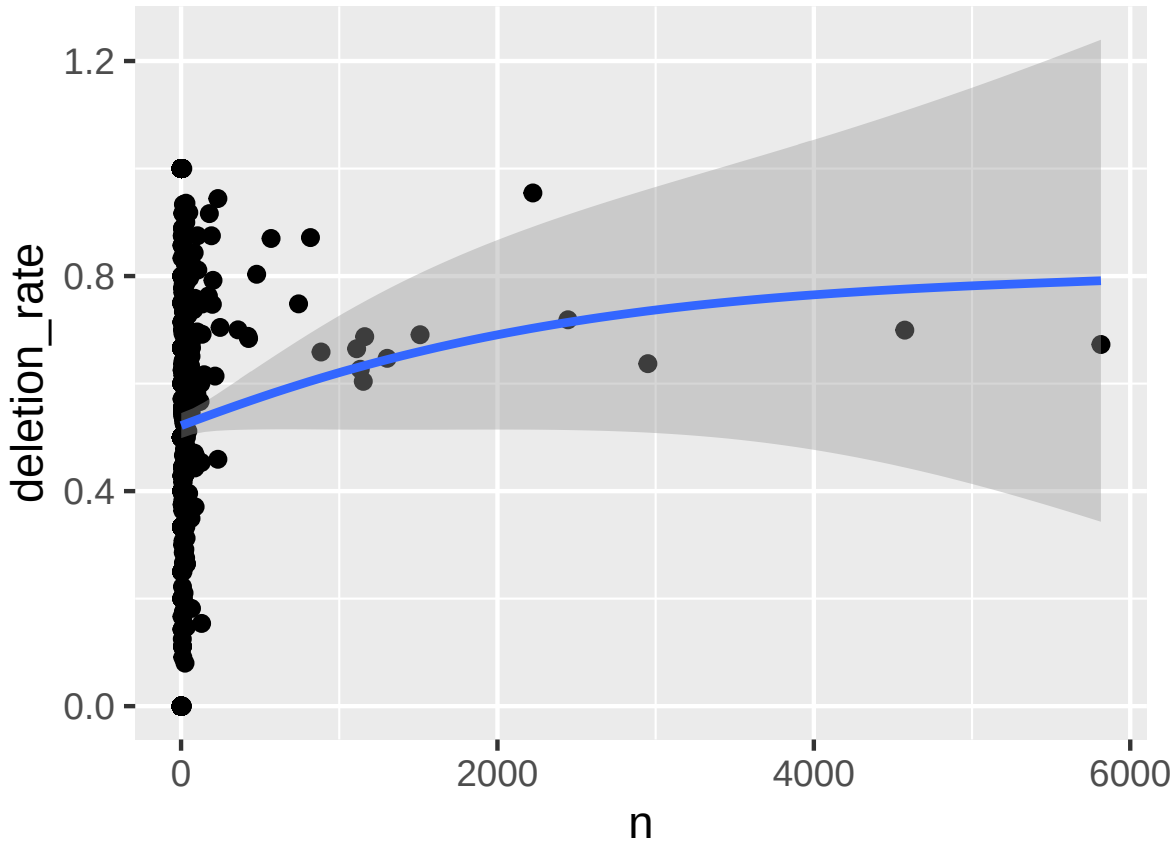
Phonological processes

- If the citation form (**Ideal**) differs from the pronounced form (**Surface**), we can say some process (e.g. deletion) has applied

1. Get all tokens in which final *t* deletes (or otherwise changes)
2. What percentage of the time does *t* delete?
   - First, filter down to *t*-final citation forms
   - Optionally, make a new column indicating whether *t* deletes (1 if yes, 0 if no)
   - Get the mean of the new column

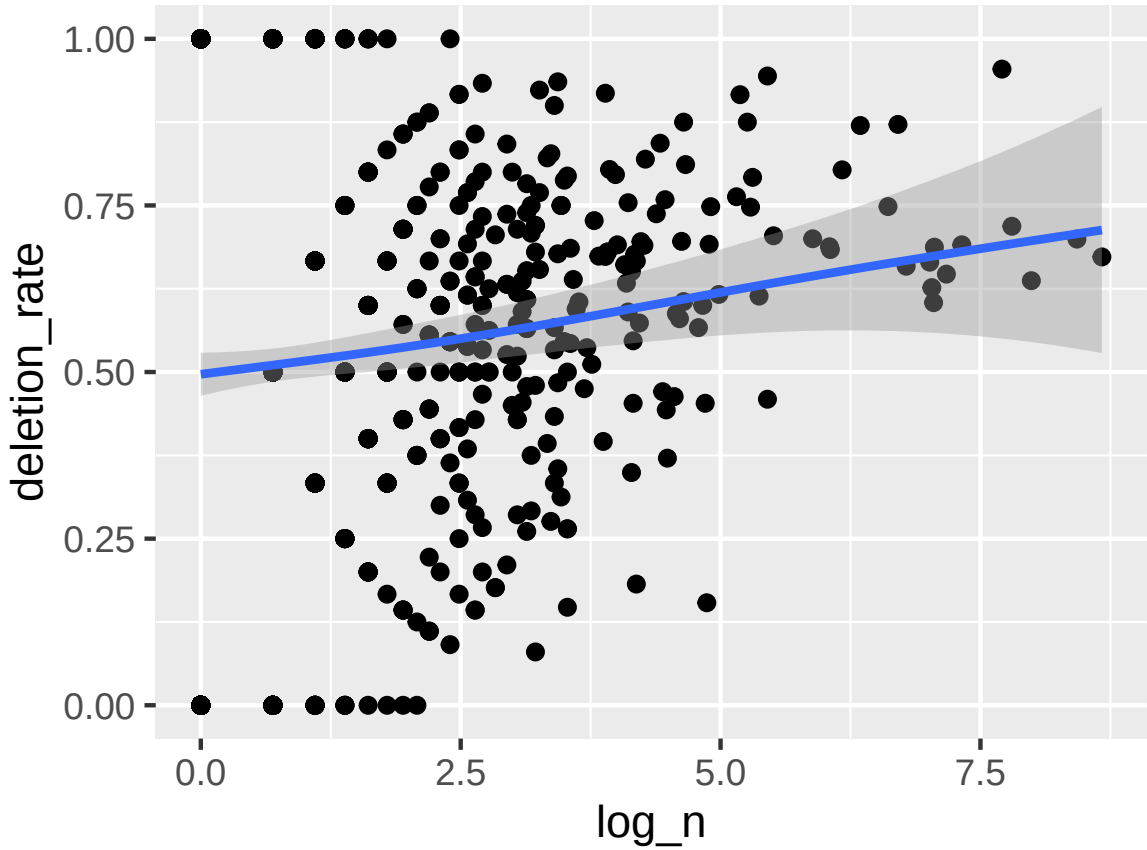
Final *t*-deletion by word

- Now get each word's deletion rate
- Plot against word frequency (via count in corpus)



Final *t*-deletion by word

- Frequency is usually "logged" (i.e. take the logarithm) to compress high values: log(n)



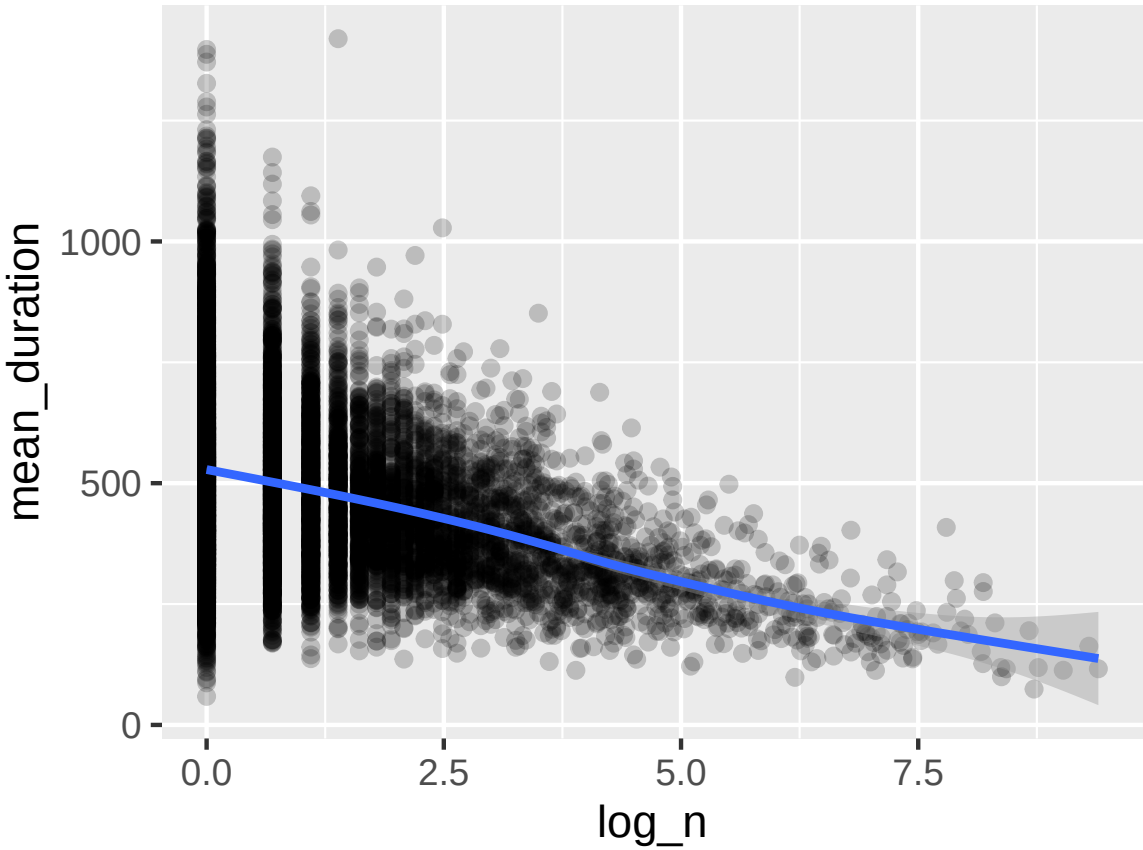
Correlation test

- Usually accompanies a scatterplot
- Characterizes whether there is a significant linear relationship between the two variables
- cor.test(x, y)
- Specify x and y using the notation *data-frame-name$column*, e.g.

  cor.test(deletion_rates$deletion_rate, deletion_rates$log_n)
- Report the p-value (as always, if ≤ 0.05, it's significant) and the test statistic, in this case, **Pearson's correlation coefficient** r (which R labels "cor")
- A (significant) correlation can be positive or negative
  1. If r is positive, y tends to increase as x increases
  2. If r is negative, y tends to decrease as x increases

Word frequency vs. duration

- Log frequency on x
- Mean word duration on y

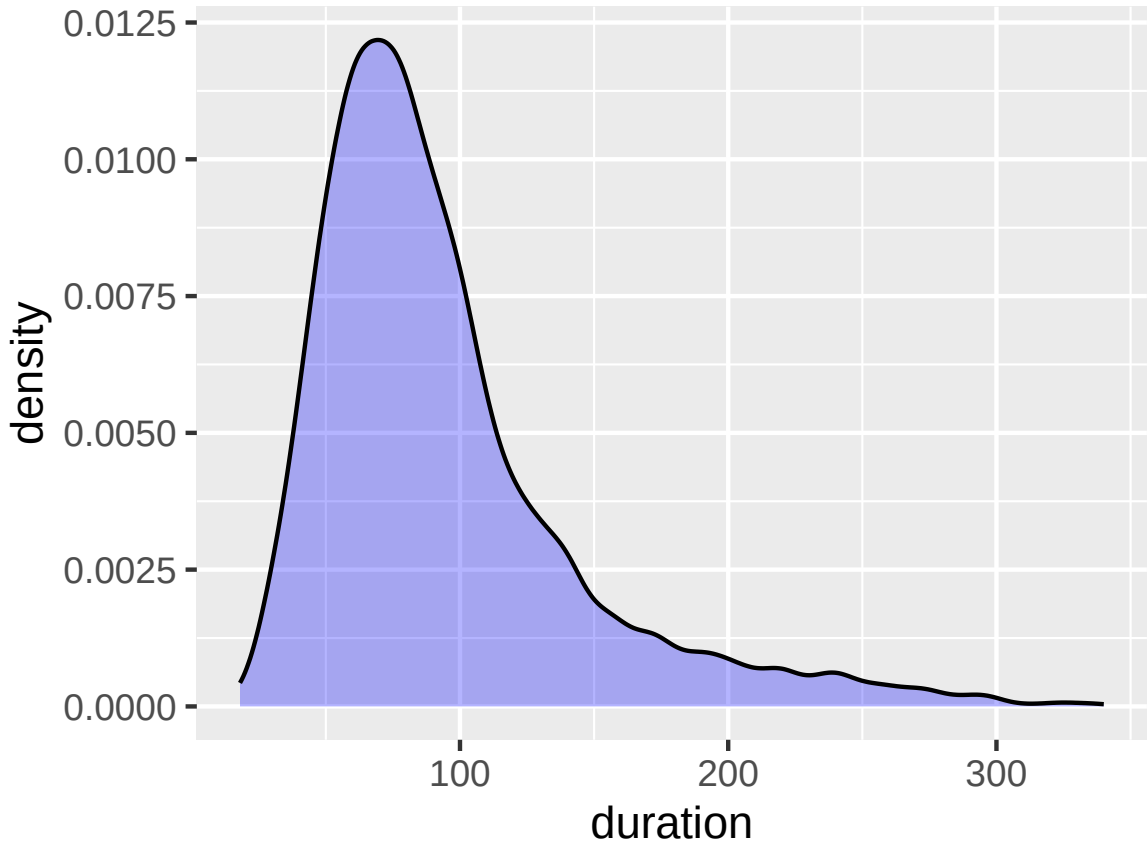


Word frequency vs. duration

- Significant *negative* correlation
- $r = -0.37$, $p < 0.0001$
- (If the p-value is extremely small, it's typical just to say it's less than a round value such as 0.0001)

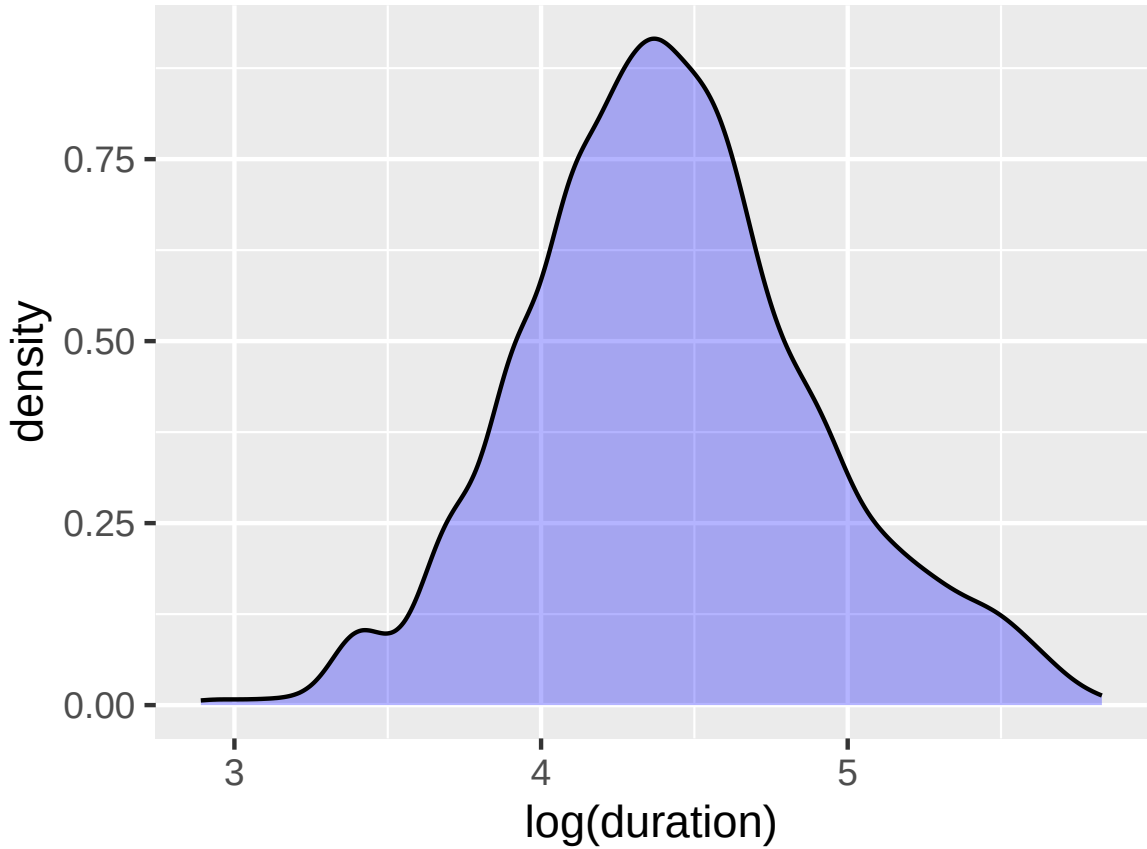
Vowel duration

- Duration (here, of [i]) is right-skewed



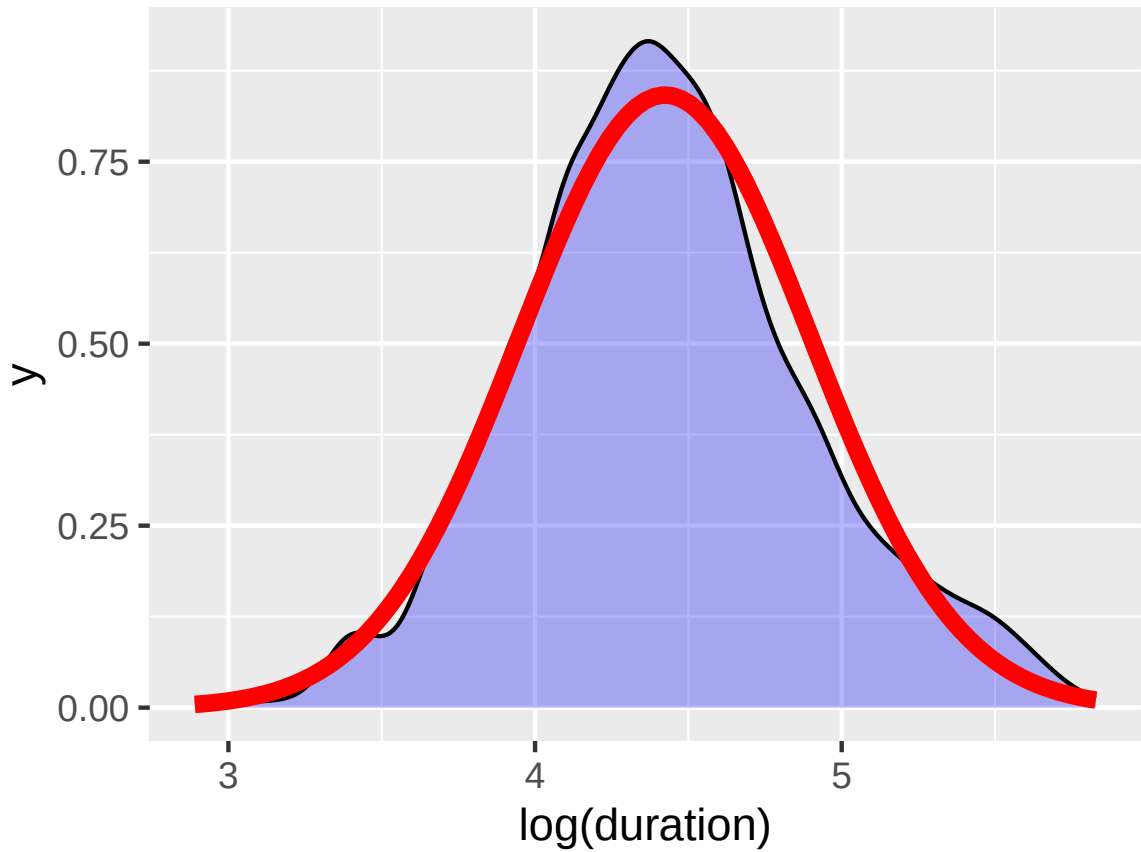
Vowel duration

- Logging unskews it
- Thus, one often sees **log duration**
- Why is an unskewed distribution useful?



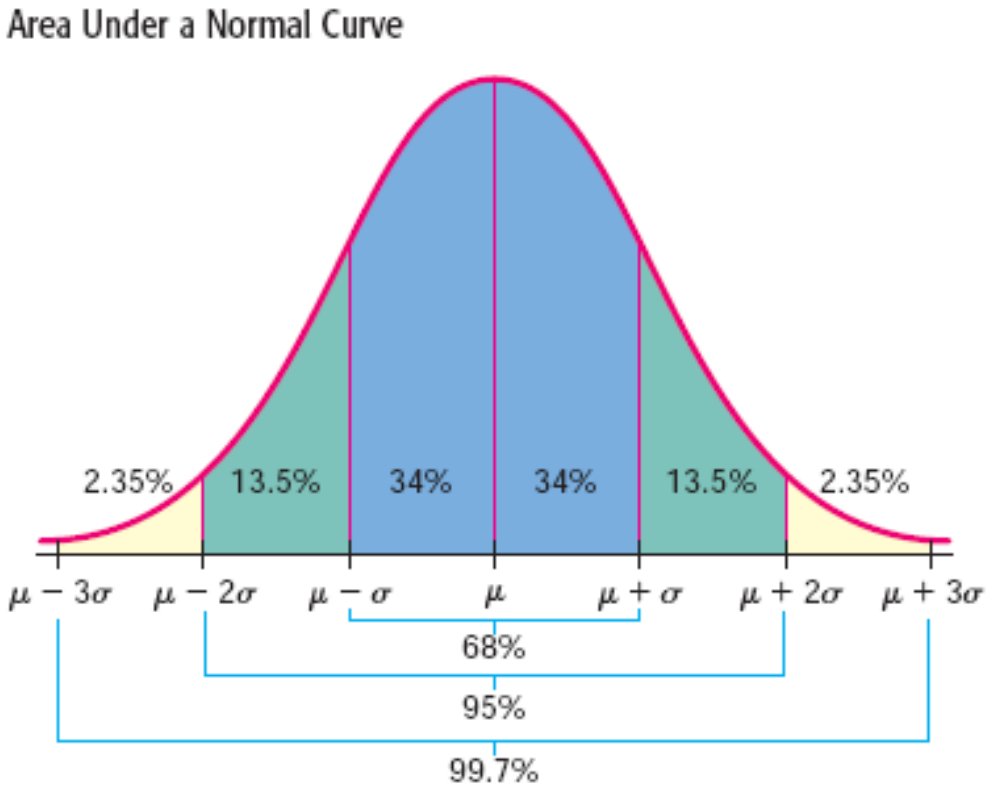
Vowel duration

- Log duration maps better onto the **normal distribution** (a symmetric bell curve)
- Here, superimposing the best-fitting normal



Normal distribution

- We can now meaningfully refer to standard deviations (SDs) from the mean
- Above, the mean is 4.4 and SD is 0.5
- Thus, 95% of measurements are contained by [3.4, 5.4]



Summary

- Wilcoxon test vs. correlation test
  1. Wilcoxon to test whether the medians significantly differ between two groups
  2. Correlation to test whether two variables relate to each other
- In either case, $p \leq 0.05$ suggests a reliable conclusion
- The logarithm is often applied to measures like frequency and duration because
  1. It compresses high values (including outliers)
  2. It unskews the distribution (allowing for more meaningful statistics, such as interpretable standard deviations)