CHAPTER 11

QUANTITATIVE ANALYSIS

KYLE GORMAN AND
DANIEL EZRA JOHNSON

A sociolinguist who has gathered so much data that it has become difficult to make sense of the raw observations may turn to graphical presentation, and to *descriptive statistics*, techniques for distilling a collection of data into a few key numerical values, allowing the researcher to focus on specific, meaningful properties of the data set (see Johnson in press).

However, a sociolinguist is rarely satisfied with a mere snapshot of linguistic behavior, and desires not just to describe, but also to evaluate hypotheses about the connections between linguistic behavior, speakers, and society. The researcher begins this process by gathering data with the potential to falsify the hypotheses under consideration (e.g., Lucas, Bayley, & Valli 2001: 43). A sociolinguist who suspects that women and men in a certain speech community differ in the rate at which they realize the final consonant of a word ending in <ing> with coronal [n] rather than velar [ŋ] would collect tokens of these words in the speech of women and men, recording which variant was used. While this data, in the form of a descriptive statistic or an appropriate graph, could suggest that women differ from men in the rate at which they use these competing variants, these techniques cannot exclude the possibility that this difference is due to random fluctuations. *Inferential statistics* allow the researcher to compute the probability that a hypothesized property of the data is due to chance, and to estimate the magnitude of the hypothesized effect.

Statistical inferences may not be valid, however, if the assumptions such techniques make are inappropriate for the data. This chapter compares inferential methods for sociolinguistic data in terms of these assumptions.

The Elements of Quantitative Analysis

The sample. The data under investigation is necessarily finite. If it comes from the spontaneous speech of a speech community, a single interview makes up only a tiny fraction of any speaker's lifetime of language, and there are usually many more speakers who could have been interviewed but were not. The same concerns apply to experimental data gathering, where there are always more possible subjects to run or stimuli to present. Inferential statistics uses the finite *sample* gathered by the researcher to generate a model of the *population* of all relevant linguistic behavior in a speech community.

Hypothesis testing. Because of the variable nature of linguistic phenomena, it is always possible that the sample differs quantitatively from the population, even under careful random sampling. The sociolinguist seeks to infer whether the patterns observed in the sample are likely to generalize to the population, but the women in a sample, for instance, may not be representative of the women in the population. The possibility that a pattern, usually an observed difference, in the sample does extend to the population is called the *alternative hypothesis*, whereas the opposing view that there is no real difference to be discovered in the population is the *null hypothesis*. For example, if a sociolinguist is interested in the association between gender and speech rate, then the null hypothesis is that speech rate is constant across genders, and the alternative hypothesis is that the speech rate differs between women and men. Inferential methods provide a way to summarize the sample data as a *test statistic* (e.g., a Z-score, t-statistic, F-statistic, or chi-square statistic), then compute the probability, henceforth the p-value, that a test statistic as large or larger would have occurred under the null hypothesis (i.e., no difference in the population).

Although this threshold is arbitrary, a result where $p < 0.05$ is generally labeled *statistically significant* in the social sciences, meaning that the null hypothesis is rejected. When comparing two sample means, $p < 0.05$ indicates that a difference of such size and consistency would be observed in no more than 5 percent of samples if it were actually spurious with respect to the population.[1] In the foregoing example, the alternative hypothesis only requires that there be *some* difference between groups, but in practice it is common to use the difference estimated from the sample as a measure of the population-level difference.

This notion of statistical significance, since it is sensitive to the amount of data as much as to the magnitude of the effect, does not always mean the result should be of interest, as the label "significant" might suggest. Researchers who discover a large effect that falls short of the significance threshold may modify the alternative hypothesis for later statistical testing, or they may choose to forgo further investigation of an effect that is statistically significant but which has a vanishingly small effect on the outcomes.

Some Frequently Violated Assumptions

An inferential statistical model relies on a set of assumptions that allow the researcher, generally with help from a computer, to calculate a test statistic and p-value from a set of data; the responsibility of making assumptions that are appropriate for the data falls to the researcher.

The random sample. In sociolinguistic studies, the contents of the sample are shaped by convenience factors, such as speakers' willingness to be interviewed or participate in an experiment. When the presence or absence of a particular type of speaker or subject is correlated with some other factor of interest—for instance, a researcher interested in stigmatized speech may unfortunately discover that low-prestige speakers are the least likely to agree to an interview with a stranger—then the sample will not provide a good estimate of the rate at which the stigmatized variant is used in the speech community. If such information is desired, the researcher may deploy *proportional stratified sampling* (e.g., Cedergren 1973); if the population consists of middle-class speakers, who account for 25 percent of the population, and working class speakers, accounting for the remaining 75 percent, the researcher ensures that this 1:3 ratio of middle- to working-class speakers (and tokens) is also found in the sample.

The omitted variable problem. No one predictor is ever sufficient to fully determine all the variation observed in a language sample (Bayley 2002: 118). While it is in some sense impossible to include every predictor that might be relevant to the outcomes of interest, a statistical model is of little use for inferring a causal connection between predictors and outcomes if one or more important predictors have been omitted. For instance, consider a study that attempts to assess the relative influences of grammatical category and phonological context on a variable process of consonant deletion. If the researcher tests the grammatical category and phonological context separately, and finds that both are significant, it does not entail that both these predictors are independently affecting the rate of deletion.

Regression models, discussed below, are perpetually popular tools in sociolinguistics because they provide an easy way to control for this effect by specifying multiple predictors for a model. It is common to find that two predictors are both significant predictors of the outcome by themselves, but when they are combined in the same regression model, only one of the two (e.g., phonological context) is significant the other predictor (e.g., grammatical category) is said to have been suppressed (e.g., Tagliamonte & Temple 2005). Such a situation could arise if the two predictors are correlated, for example, if certain grammatical categories tend to co-occur with certain phonological contexts (e.g., Bybee 2002: 275f.), but if grammatical category itself has no additional effect on the rate of deletion.

Multicollinearity of predictors. It may however be the case that multiple predictors stand in a causal relationship with the outcome (e.g., both phonological context and grammatical category increase rate of deletion), and this must be distinguished from the above scenario. Unfortunately, carelessly including every available predictor is not a helpful for drawing this distinction. Multivariate statistical methods assume that the predictors are "orthogonal," that is, fully independent of each other. The parameters of a model that includes *multicollinear* (i.e., strongly nonorthogonal) predictors are highly unstable and greatly influenced by small fluctuations in the data. Gorman (2010) gives an example of a spurious sociolinguistic finding due to multicollinearity between measures of socioeconomic status, and demonstrates the method of residualization, one way to eliminate multicollinearity among predictors.

Independence of outcomes. Ordinary regression models make a strong assumption that once the predictors are taken into account, the outcomes themselves are mutually independent. Since it is standard, both in the field and the laboratory, to gather many data points from each speaker or subject, this assumption is frequently violated in practice. The question of whether an effect of gender in the sample is generalizable to the population is potentially of great sociolinguistic interest. To determine this fact, it is necessary to distinguish between a gender effect in the population and the presence in the sample of a few speakers who just happen to be male and furthermore are "outliers" from the rest of the sample; erroneously rejecting the null hypothesis in this latter case is known as Type I error. These two possibilities cannot easily be teased apart unless the effects of gender and speaker can be modeled simultaneously.

Insofar as speakers belonging to the same speech community may differ in the rates at which they use different variants, even after gender, age, and social status are taken into account (Guy 1980, 1991: 5), speaker identity is a strong predictor of linguistic behavior, one that is desirable to model. Yet, all tokens of a single speaker collected at a single time are also tokens of the same gender, etc.: every token from "Celeste S." also has the same value for the gender predictor ("female"), age (45), etc., and thus speaker identity fully determines these other predictors. Random effects, described later, provide a principled solution to the problems created by this nesting, without giving rise to multicollinearity.

Dichotomization and categorization. It is all too frequent that a researcher gathers observations—whether predictors or outcomes—on a continuous or integer scale, but converts these values to a few-valued (often binary) coding before performing statistical analysis. While there is occasionally a good reason to treat data that are naturally many-valued as a few-valued scale,[2] it usually increases the chance of Type II error, the error of failing to reject the null hypothesis in the case when this null hypothesis is in fact false (Cohen 1983). If a researcher posits a sound change in progress in a speech community, then a 78-year-old speaker should be less advanced with respect to this change than a 60-year-old speaker, but if these two speakers are placed together into the "60 years of age and older" bin, this trend is treated as noise rather than being credited to the alternative hypothesis of an age effect.

This example highlights another point: binning usually requires the researcher to arbitrarily choose the number and location of the cutpoint(s) between bins, and these decisions have unpredictable effects on the results that obtain. One reason this binning is so commonly seen in sociolinguistics is the "founder effect" of VARBRUL and its descendants, which require both outcomes and predictors to be categorical. However, it is incorrect to assume that VARBRUL's feature set delimits the set of possible sociolinguistic analyses, and the use of continuous predictors and/or outcomes in sociolinguistics dates back at least as far as Lennig's (1978) study of variation in the Parisian vowel system.

Another reason that some researchers are willing to bin continuous data is that the most basic use of a continuous predictor in regression assumes that the predictor and the outcome stand in a relationship that is monotonic, and more specifically, linear. A clear example of a relationship that violates this assumption is the one that holds between the use of stable sociolinguistic variables and social class, which a number of studies have found to be curvilinear, with interior social classes using the highest rates of a nonstandard variant of a stable linguistic variable (Labov 2001: 31f.). In such cases, the appropriate response to this problem, though, is not ad hoc dichotomization, but rather for the researcher to explore the relationships observed in the data (e.g., by plotting the predictor and outcome), and choosing appropriate "transformations" of the data so that the linearity assumption is satisfied.

In many cases, the hypothesis under consideration will determine an appropriate transformation. For example, the exemplar theory of lenition (e.g., Bybee 2002) predicts a relationship between the logarithm of word frequency and the rate of lenition, and thus a researcher who wishes to evaluate this hypothesis must convert word frequency to a log scale before modeling. Harrell (2001: 16–26) provides a useful discussion of transformations for regression modeling.

Summary

Inferential analysis allows for hypothesis testing, but there are many common pitfalls. The rest of the chapter outlines what we consider the best practices for analyzing the most common types of sociolinguistic data. The next two sections describe the analysis of binary and *multinomial* outcomes (categorical variables with more than two values). The following section considers methods for *continuous* outcomes, with a focus on acoustic measurements of vowels. The concluding section discusses some recent trends in the field of statistics of relevance to sociolinguists.

Methods for Binary Variables

Interpreting Cross-Tabulations

Many quantitative sociolinguistic studies compare two distinct, discrete, semantically equivalent variants in complementary distribution.

The chi-square test. In November 1962, William Labov elicited tokens of the phrase "fourth floor" from employees in three Manhattan department stores for the purpose of studying the social stratification of post-vocalic *r* realization in New York City. While this original study (Labov 2006: chapter 4), first published in 1966, does not include any inferential statistics, the cross-tabulation of the data (e.g., *r*-full vs. *r*-less tokens by store) lends itself to a simple statistical test. Consider the null hypothesis that there are no differences between the employees of the three department stores, chosen to represent the class spectrum in New York. The employees at middle-class Macy's pronounce post-vocalic *r* in 125 tokens, and do not in 211 tokens; *r* is present 37 percent of the time (= 125/336). At working-class department store S. Klein's, employees only have 21 tokens of post-vocalic *r* and 195 tokens where it is not realized, for a 10 percent rate of *r* presence. Saks, the department store representing the upper class, has a 48 percent rate of *r* presence. To compute the probability this effect is due to chance, these counts are used to compute a test statistic called Pearson's chi-square: the value obtained is 73.365. We then compute the probability of a test statistic of this size or larger being obtained for a sample of this size simply by chance using the two-tailed *chi-square distribution*. The *p*-value representing this possibility is p = 1.1e-16, indicating that there is good reason to reject the null hypothesis that there are no differences in the *r*-realization among the different department stores, and the average rates of *r* presence just calculated indicate that post-vocalic *r* is realized more often by speakers from higher social classes.

Fisher's exact test. The chi-square test is not very appropriate for small amounts of data, since it is based on an approximation that is true under the obviously false assumption of an infinite sample; the accuracy of this test is worse as the sample grows smaller. For this reason, we favor a related technique known as Fisher's exact test, which computes the "exact" (i.e., correct) *p*-value even for small data sets. As is sometimes the case, the Fisher *p*-value is somewhat smaller than the Pearson chi-square p-value (p = 1.4e-18), but it is always more precise. The Fisher p-value is often difficult to compute by hand, but since it can be computed for huge data sets by a modern computer in the blink of an eye, it should always be used rather than the chi-square test. Table 11.1 shows the results of applying the chi-square and exact test to two other contrasts in Labov's data. First, Labov feigned misunderstanding after the first "fourth floor," usually causing the speaker to repeat him- or herself, to obtain more data in a more careful speech style. Secondly, Labov recorded whether each token comes from "fourth" or "floor." These results are summarized in table 11.1; word and department store are significant predictors, but the repetition contrast is not.

Simple Logistic Regression

Because of the potential for omitted variable bias discussed above, it is preferable whenever possible to consider the relative contributions of multiple predictors in a single model. While the department-store data is relatively balanced, the p-values obtained from using a *univariate* method like the Fisher exact test may be inaccurate when this is not the case. *Logistic regression*, which predicts binary outcome using one or more independent predictor(s), and which will be familiar to many readers as the model underlying VARBRUL, is the appropriate model in this case.

What to include. In the logistic regression model, the outcome is either r or zero; the predictors, all categorical, are word ("fourth" vs. "floor"), repetition (first vs. second), and store (Saks vs. Macy's vs. S. Klein's). Modern regression software also allows the user to include what are generally called *interaction* effects, predictors that are derived from the combinations of other predictors.

Table 11.1. New York City department store (r) cross-tabulation, chi-square, and Fisher exact test

	# r	# zero	% r	p-value (chi-square)	p-value (Fisher exact)
S. Klein's	21	195	9.7	1.2e-16	1.4e-18
Macy's	125	211	37.2		
Saks	85	93	47.8		
"fourth"	87	295	22.8	1.4e-07	1.2e-07
"floor"	143	204	41.2		
first repetition	136	322	29.7	0.187	0.162
second repetition	94	177	34.7		

In this case, an interaction between word and department store allows the researcher to probe whether, in addition to any differences between "fourth" and "floor" and the different department stores, there is any difference in the difference between "fourth" and "floor" across the different department stores. For example, is "fourth" versus "floor" at Saks different from "fourth" versus "floor" at S. Klein's? There is no obvious reason to hypothesize such an interaction in this case, but it is included for the purpose of demonstration. The results from fitting this model, which reports numbers in a form that will be familiar both to users of VARBRUL and other software packages (who may know log-odds as *betas*, *coefficients*, or *estimates*) are given in table 11.2.

In this model, absence of r is treated as rule application, so an increase in the log-odds or the weights indicates fewer r's. Just as was the case for the univariate tests mentioned earlier, there is strong support for differences between stores and the two words. The effect of repetition is approaching significance, with the second repetition being more likely to contain an overt r than the first, but is just short of the standard threshold of 0.05. Among the interaction terms, which taken together are nonsignificant, there is one suggestive trend: "fourth" has more r than "floor" at S. Klein's, but the pattern is reversed at the other two department stores.

This raises an important question: how does one decide which predictors to include and which to omit? A useful procedure, adapted from Gelman and Hill (2007: 69), is as follows. The initial model should include any predictors

Table 11.2. New York City department store (r) fixed-effects logistic regression

	log-odds	weight	p-value
(intercept)	0.910	0.713	8.3e-19
S. Klein's	1.304	0.787	1.2e-19
Saks	−0.875	0.294	
Macy's	−0.428	0.395	
"floor"	−0.444	0.391	
"fourth"	0.444	0.609	8.2e-09
first repetition	0.166	0.541	0.065
second repetition	−0.166	0.459	
S. Klein's and "fourth"	−0.239	0.441	0.341
S. Klein's and "floor"	0.239	0.559	
Macy's and "fourth"	0.061	0.515	
Macy's and "floor"	−0.061	0.485	
Saks and "fourth"	0.177	0.544	
Saks and "floor"	−0.177	0.441	

the experimenter has recorded and thinks might influence the outcomes. After the model is fit, the predictors are assessed in the following manner:

1. If a predictor is not statistically significant, but the estimate (or factor weight) goes in the expected direction, leave it in the model.
2. If a predictor is not statistically significant, and the estimate goes in an unexpected direction, consider removing it from the model.
3. If a predictor is statistically significant, but the estimate goes in an unexpected direction, reconsider the hypothesis and consider more data and input variables.
4. If a predictor is statistically significant, and the estimate goes in the expected direction, leave it in the model.

The resulting regression model supports the Fisher exact test observations in the sense that the same predictors are significant, but we see that the department-store p-value is now even smaller. Indeed, in the absence of multicollinearity, the p-value of a given variable usually becomes smaller when other relevant predictors are taken into account.

On stepwise techniques. This technique of allowing prior assumptions to guide variable selection, and potentially reporting non-significant effects, contrasts with the use of automated *stepwise* model selection techniques, such as is found in VARBRUL, which may be familiar to many sociolinguists but which are the target of derision by many statisticians (e.g., Harrell 2001: 56, 79f.). Step-up procedures are subject to the problem of omitted variable bias previously discussed. Step-down procedures do not suffer from this problem, as they begin with a full model (containing all the predictors), but there is no compelling reason the researcher shouldn't stop there. If a predictor actually has a small effect, it is beneficial, and if it does not, it does no harm. In contrast, the coefficients of any marginally significant predictor that are retained by stepwise methods are biased upwards in comparison.

Nesting and regression. The previous model measured a sociolinguistic variable's distribution according to department store, the grammar-internal effects of different phonological context ("fourth" vs. "floor"), and contrasts with respect to style (repetition). Since there are no more than four tokens per speaker, and 264 speakers in the sample, there is no reason to believe that some speaker outlier is driving the trend; even if some speakers in this sample do differ drastically from the rest of the population in their usage of post-vocalic *r*, one can no more detect these outliers in this data than one could reasonably assess whether a coin is or is not fair after flipping it only four times, since even a fair coin will come out all heads or all tails 12.5 percent of the time.

As mentioned above, it is generally understood that speakers may differ from each other in their overall rates of usage of different variants. What has not been as widely acknowledged is that this means when there are many tokens per speaker in the sample, that the differences between

speakers must be modeled in order to satisfy the assumption of independent outcomes (see above). As already mentioned, the above fixed-effects logistic regression models do not provide any appropriate solution to the nesting between speaker and other demographic factors. One method to deal with this problem is to compute separate models by speaker, and then perform inference over the coefficients of the individual models (e.g., Gelman & Hill 2007: chapter 12; Rousseau & Sankoff 1978; Guy 1980), but this does not allow us to constrain speakers from the same speech community to behave the same with respect to grammatical constraints on variation, despite our strong bias that speakers from the same community share these constraints (Guy 1980).

Mixed-effects Regression

Mixed-effects models (Pinheiro & Bates 2000) are a recent innovation in regression which allow for, in addition to the familiar stratum of fixed-effects predictors, a set of predictors called *random effects* providing a natural solution to the nesting problem. An advantage of the mixed-effects model is that in most cases it returns more accurate p-values compared to a fixed-effects model that ignores nesting.

Random intercepts. The simplest type of mixed-effects model augments a standard regression with a *random intercept*, which is a predictor consisting of many levels (such as unique identifiers for the different speakers in the sample). During model fitting, the variance attributable to different levels of the random intercept is estimated, and each level of the random-effects predictor is mapped onto this normal distribution in a way that preserves the essential insight that speakers are otherwise the same. This is particularly useful for measuring the differences between speakers when a researcher is interested in social factors like gender or ethnicity in a nesting relationship with speaker identity.

Another application of random intercepts is to model word-level effects. One may have a null hypothesis that once phonological context, grammatical effects, and so on, are controlled for, there is no effect of word identity on sociophonetic variables, but there are many reports of purely lexical effects in variation (e.g., Neu 1980: 50). However, words and grammatical category may be in a nesting relationship, making word identity a good candidate for a random intercept.

An advantage of the mixed-effects model is that in many cases, it returns reduced, and more accurate, significance levels (i.e., smaller p-values) compared to a fixed-effects model that ignores by-subject and by-word grouping. This can be illustrated using data on the English of adolescent Polish immigrants in the United Kingdom collected by Schleef, Clark, and Meyerhoff (2011); here, the focus is a subset of their sample gathered in London. The data consists of 925 tokens of the variable (ing) from 21 speakers and representing 123 word types. This variable concerns the realization of the final consonant in word-final <ing> sequences. In addition to the velar nasal and coronal nasal articulations included here, the data also contains a third category, where the variable is realized with an oral velar stop (e.g., [iŋk]). Henceforth, this final variant is ignored, leaving 718 tokens from 21 speakers.

Despite the modest size of the data set, a fixed-effects regression identifies three significant between-word predictors (preceding phonological segment, grammatical category, and lexical frequency) and three significant between-speaker predictors (gender, English proficiency level, and friendship network), summarized in table 11.3. The fixed-effects model finds all six of these predictors highly significant. One surprising effect is that, while a higher degree of English proficiency results in a higher rate of the coronal variant, speakers with a mostly Polish friendship network are also more likely to use the coronal variant than those with a mixed or mainly English network. One might have expected that both these predictors were imperfect measures of the speaker's contact with first-language English, and thus would pattern together.

The effect of the remaining between-speaker predictor, gender, is also surprising. Whereas men generally use more of a stigmatized variant than women (Labov 2001: 264), Polish women (22 percent coronal tokens from 12 females) favor the stigmatized coronal variant more than men (9 percent coronal tokens from nine men). While this different in rate is somewhat small in absolute terms, the fixed-effects model treats gender as significant (p = 0.015).

The addition of random intercepts for speaker and lexical item results in somewhat different patterns of significance. All three of the between-word predictors are still found to be significant (the reported significance levels are now roughly p = 0.001 instead of several orders of magnitude smaller), indicating that the effects are unlikely to be due only to properties of individual words in the sample. However, as shown in table 11.3, none of the three between-speaker predictors reaches significance, and one cannot reject the null hypothesis that they have no effect on (ing).

Random slopes. Whereas the random intercepts used above adjust the model's predictions for any speaker or word, mixed-effects models can also include random slopes. These can be used, for example, to allow speakers not only to differ in the rate at which they use a variant, but also to differ in the size of the effect of between-word constraints, such as phonological context. While there may be a null hypothesis that such differences are not present in the speech community once per-speaker intercepts are properly accounted for, mixed-effects models are capable of testing this null hypothesis without elevating it to the level of a potentially dangerous assumption. Random intercepts and slopes are of particular use for modeling the results of laboratory experiments where

Table 11.3. Polish English (ing) in London fixed-effects and mixed-effects logistic regression

	Fixed-effects model		Mixed-effects model	
	log-odds	*p*-value	log-odds	*p*-value
(intercept)	−2.828	4.7e-14	−3.106	8.1e-08
lexical frequency (log)	0.978	2.0e-08	1.215	3.29e-05
noun	−0.532	6.3e-05	−0.716	0.004
verb	0.857		1.214	
gerund	0.001		0.173	
adjective	0.924		−1.204	
preposition	0.198		0.158	
discourse marker	−1.446		−0.622	
preceding apical consonant	−0.530	1.7e-05	−0.592	7.7e-04
preceding dorsal consonant	1.002		1.215	
other preceding consonant	−0.472		−0.622	
male	−0.547	1.9e-04	−0.381	0.185
female	0.547		0.381	
little English proficiency	−0.187	1.5e-06	−0.132	0.260
good English proficiency	−0.678		−0.584	
very good English proficiency	0.865		0.717	
mostly Polish friendship network	0.648	0.029	0.683	0.445
mixed friendship network	0.266		0.286	
mostly English friendship network	−0.914		−0.969	

subjects, stimuli, and conditions may all interact (e.g., Baayen, Davidson, & Bates 2008; Gorman 2009).

Summary

This section has described the application of univariate and multivariate techniques to modeling the predictors of the classic variety of sociolinguistic variable, binary outcomes in complementary distribution.

Methods for Multinomial Variables

For binary outcomes, logistic regression is the tool of choice. However, a sociolinguistic variable may be categorical, but have more than two variants in competition, as is the case with many consonantal variables. In some cases, a prior theory of the variable may make it reasonable to model these alternatives with separate binary logistic regressions. However, if the hierarchical structure of the variable is not absolutely clear, then the appropriate tool is *multinomial logistic regression.* In its most common implementation, this method does nothing more than fit multiple logistic models to the data simultaneously. However, if there is a natural ordering to the variants, and additional assumptions are reasonable, it is possible to fit a more constrained (and thus more powerful) model, *ordinal logistic regression.*

To illustrate these assumptions, we consider 8071 tokens of post-vocalic *r* gathered in Gretna, Scotland, one of the four communities investigated by the Accent and Identity on the Scottish-English Border (AISEB) project (Llamas 2010). The quality of the *r* sound was given a narrow transcription, but here we collapse the observations into three categories: taps/trills, approximants, and zero.[3] Since post-vocalic *r* is disappearing in apparent time in Gretna—moving away from a Scottish standard and toward an English one—the change can be thought of as a lenition process with a natural ordering: tap/trills > approximants > zero, and thus is a candidate for ordinal logistic regression.

To check the assumptions of a proportional odds ordinal logistic regression model, an unordered multinomial logistic regression is applied to the data. This model includes three binary external predictors: age group (older, 57–82, vs. younger, 15–27), gender (female vs. male), and social class (middle class vs. working class). The 40 Gretna speakers are a balanced sample of these external predictors, with five speakers belonging to each of the eight combinations of these three external predictors. Internal predictors relating to syllable stress, speech style, and the identity of the preceding and following segments are also included. These are *nuisance variables*, meaning that we wish to control for their effects to prevent omitted variable bias, but they are not the focus of this investigation and their effects will not be discussed. For the three external predictors, the multinomial regression produces two intercepts and six coefficients. Since each speaker in the sample produces many tokens, a mixed-effects model with a per-subject intercept would be ideal, but at the time of writing we are unaware of any software that fully supports mixed-effects multinomial models. For this reason, table 11.4 reports the log-odds, but not the potentially misleading *p*-values.

The three-valued outcome has one baseline category, here the most conservative variant: taps/trills. Each predictor is associated with two coefficients, one for approximants, and one for zeros. The first coefficient, for approximants,

Table 11.4. Gretna (r) unordered multinomial logistic regression (external effects only)

	Log-odds (approximant vs. tap/trill)	Log-odds (zero vs. tap/trill)
(intercept)	2.601	3.209
younger	0.605	1.377
older	−0.605	−1.377
male	−0.457	−0.502
female	0.457	0.502
working class	−0.024	−0.052
middle class	0.024	0.052

represents the estimated adjustment, in that environment, to the log-odds of an approximant occurring instead of a tap or trill. The coefficient for zero represents the adjustment to the log-odds of a zero occurring instead of a tap or trill.

The model output contains two intercept terms, 2.601 for approximants and 3.209 for zeros. The numbers are related to the raw proportions of the response categories—6.3 percent taps/trills, 33.3 percent approximants, and 60.3 percent zero—but adjusted to represent the mean over all the possible cells formed by the predictor variables.

The two coefficients for female gender, 0.457 for approximants and 0.502 for zeros, indicate that these two variants are approximately equally favored by females as opposed to taps/trills. A cross-tabulation shows the same fact: overall, females produced only 4.3 percent taps/trills while males produced 8.4 percent. (Note that gender has little effect on the contrast between approximants and zeros, a fact which will be important later.)

The coefficients for the younger age group, 0.605 for approximants, and 1.377 for zeros, indicate that younger speakers favor approximants over taps/trills even more than older speakers do, and that younger speakers favor zeros over taps/trills much more than older speakers do. The coefficients for social class show only a small effect in the expected direction: the middle class favors more advanced, lenited forms, while the working class preserves more traditional variants.

Some of the coefficients of this model suggest the data does not satisfy the *proportional odds* assumption of the ordinal logistic regression model. The ordinal regression divides the three-outcome variation into two cut-points: taps/trills versus {approximants, zeros} and {taps/trills, approximants} versus zeros. An ordinal model with proportional odds assumes that the predictors affect both of these cut-points identically, so there will be only one coefficient for each binary predictor, rather than *k–1* for *k* response categories, as in the unordered multinomial model.

Under proportional odds, the difference between male and female speakers must have the same effect at the two cut-points, but this is not the case here: speakers' gender has quite an effect on "first step" of lenition, from taps/trills to one of the other categories, but has little effect at the "second step," from one of the first two categories to zeros. At the first cut-point, we see 174 taps/trills versus 3894 approximants/zeros for women, and 338 taps/trills versus 3665 approximants/zeros for men. Women favor the more lenited variants by 95.7 percent to 91.6 percent, a difference of 0.725 log-odds. At the second cut-point, there are 1569 taps/trills/approximants versus 2499 zeroes for women, and 1634 taps/trills/approximants versus 2369 zeroes for men. Here, women favor the more lenited variant by 61.4 percent to 59.2 percent, a difference of only 0.094 log-odds. The proportional odds assumption does not hold with respect to gender.

For this reason, it would be inappropriate to force the Gretna post-vocalic *r* data into a proportional odds ordinal regression, although this does not indicate that the three variants are truly unordered, or that the two different stages of lenition are independent phenomena.

Ordinal logistic regression is better suited to model data on /ay/-diphthongization in Waldorf, Maryland, reported by Bowie (2001). This data set, consisting of 4038 tokens collected from 25 speakers, was originally coded with a three-way response variable: "monophthong" (9.8 percent) versus "weak glide" (12.9 percent) versus "full glide" (77.3 percent). However, the distinction was collapsed into a binary one—monophthong versus diphthong (i.e., weak and full-glide tokens)—before multivariate analysis, because the distinction between weak and full diphthongs was "not found to produce meaningful results" (342).

Bowie (2001) provides a full discussion of all the predictors analyzed, including stress, style, following phonological environment, and syntactic environment, but here, the focus is on two external predictors, age and gender. There is a clear effect of age, with younger speakers more likely to use the diphthong. Similarly, females lead in the use of the standard (diphthongal) variant. Under the hypothesis that this is an ordered process—monophthong > weak glide > full glide—then the two-way choice analyzed in Bowie (2001) is the first cut-point of an ordinal regression. The second cut-point separates monophthongs and weak diphthongs, on the one hand, from full diphthongs, on the other. We first validate the proportional odds assumption with an unordered multinomial regression model, and find that age and gender affect both steps of the diphthongization process to a similar degree, in contrast to what was observed in the Gretna sample. There are also formal tests for validating a proportional odds assumption, but Harrell (2001: 335) reports that these tests too-frequently reject the null hypothesis of proportional odds, and thus an informal approach is sufficient. We then fit an ordered multinomial model to the data; the between-speaker results are summarized in table 11.5. The results suggests that the weak/full glide distinction is indeed meaningful in this data. This is expected if monophthongization proceeds gradually, and what were recorded as weak and full glides are not natural categories but rather a useful categorization assigned to a continuous variable, such as glide length.

Table 11.5. Waldorf /ay/-monophthongization ordered multinomial logistic regression (external effects only)

	log-odds
Strong vs. weak glide/monophthong	1.323
Strong/weak glide vs. monophthong	2.579
male	0.336
female	−0.336
born before 1920	1.858
born 1920–1939	0.371
born 1940–1949	0.366
born 1950–1959	−0.404
born 1960–1969	−0.412
born 1970–1979	−0.954
born after 1980	−0.825

If one is dealing with an ordered response and the data conforms to the proportional odds assumption, there are two main advantages to using an ordinal method. First, the coefficient estimates should be more accurate in the sense that the model will better describe the underlying population and be more useful for the prediction of future data. The second advantage to an ordinal method is that is it lowers the likelihood of Type II error (failing to reject a null hypothesis when the alternative hypothesis is true), while avoiding the problems inherent in making multiple comparisons over the results of separate binary regressions. If we have reason to believe that a multiple-variant outcome reflects an underlying ordering, then some form of ordinal modeling is desirable.

Methods for Continuous Variables

Often the variables of interest can be measured on a continuous scale, such as acoustic measures extracted from a recorded speech signal. This section compares several modern methods used to study continuous outcomes. The methods described are used here to study vowel formants in the F1 x F2 space, but such techniques apply naturally to continuous outcomes of other types.

Bigham, White-Sustaíta, and Hinrichs (2009) administered word lists to 52 Anglo-American, Mexican American, and African American speakers in Austin, Texas; here we look at paired tokens of *bot* and *bought*, and *hod* and *hawed*. *Bot* and *hod* represent a vowel (written LOT, following Wells 1982) that is etymologically distinct from the vowel of bought and *hawed* (written THOUGHT), but these vowel classes have merged or are in the process of merging for many North American English speakers.

Simple Two-Sample Tests

These two-sample tests are univariate methods that can be used to test the null hypothesis that the two etymological classes are acoustically identical at the population level.

The t-test. The *t*-test is a class of methods for testing the null hypothesis that two subsamples have identical means. These samples can either be *paired*, in which case each observation from one subsample stands in a one-to-one relationship with an observation from the other subsample, or *unpaired*, when this does not hold. The Bigham et al. data consists strictly of minimal pairs, so it is natural to pair tokens of *hod* and *hawed*, and *bot* and *bought*, respectively. Even when the pairing requires the researcher to exclude words that are not one part of a minimal pair, Herold (1990: 73) and Johnson (2010: 108) argue that unpaired *t*-tests are a poor tool for quantifying merger, since the tests frequently result in assigning a significant effect for vowel class even to speakers who are judged by the researcher to be merged in production. This is likely caused by the omission of phonological context, which happens to be associated with vowel class membership.

Variance (equal to the standard deviation squared) is a standard measure of how far away individual values in a sample or population are from the mean. When two subsamples have the same variance, they are said to be *homoscedastic*, and heteroscedastic otherwise. For this data, the assumption of homoscedasticity is not strictly true: THOUGHT has lower variance for both F1 and F2. Heteroscedasticity between two vowel classes undergoing merger has been observed in other studies (e.g., Johnson 2010: 113, 128); this may indicate that the speaker is style-shifting towards, or away from, merger. The *unequal-variance* varieties of the *t*-test, which do not assume homoscedasticity, are the default choice for most tasks and it is this type that is used here. The results for the two formants find a difference in F1 ($p = 0.0024$) and in F2 ($p = 1.9\text{e-}05$), and inspection of the means and medians shows that LOT is lower and more front than THOUGHT.

The Wilcoxon test. The *t*-test used above does not assume the classes share the same degree of variance, but it does assume that the two classes are normally distributed. Since this assumption is often violated in practice, it is often preferable to use the family of Wilcoxon tests, especially when communicating with other fields (such as other social sciences) where the *t*-test has been replaced by this family of tests, which are free of assumptions of homoscedasticity or normally distributed data. Whereas the *t*-tests compare means, and therefore can be greatly influenced by outlying data points, the Wilcoxon tests focus on medians, for which the influence of outliers is minimal. The test used here, the *Wilcoxon signed rank* test, evaluates the null hypothesis that the paired sets of vowels have the same median formant values; the resulting *p* values for F1 ($p = 0.002$), and and F2 ($p = 8.7\text{e-}05$) are now somewhat smaller.

Tests with Multiple Predictors

Both the *t*-test and the Wilcoxon test found a significant difference between LOT and THOUGHT for F1 and F2. However, these univariate tests are of less use for looking at the demographic predictors of merger, since they only allow the data to be partitioned into two subsamples. In many cases, the data is unbalanced according to the various other predictors (because, for instance, it was collected from spontaneous speech), which means that a failure to control for demographic factors or grammatical factors can undermine the attempt to determine whether the two vowels are underlyingly different.

Mixed-effect regression. *Linear regression* is the classic technique for one or more predictors of continuous outcomes; the most basic case is not illustrated here. Linear regression also permits random effects to be included as predictors. While linear regression by default assumes homoscedasticity between binary predictors, it is also possible to allow for heteroscedasticity between, for instance, the two vowel classes.

To compute such a model over the whole sample, F1 and/or F2 values are the outcomes, and vowel-class identity and the following consonant (/t/ or /d/) are the fixed effects. To these models it is possible to add in per-speaker predictors that address the role of ethnicity, gender, and age on participation in the merger; these three are treated as fixed-effects interactions with vowel class. These interaction terms are simply the "predictor" vectors derived from the combination of vowel class and ethnicity, so that the model estimates the effect of vowel class for the whole population, but also the effect of vowel class for each ethnic group. The final components to this model are per-speaker intercepts and a random-effect interaction (or random slope) between speaker and vowel identity. The former controls for physiological differences between speakers which influence formant measures (i.e., it is a form of normalization), and the latter allows speakers to differ on their participation in the merger. Table 11.6 reports the subset of the F2 model that pertains to age and ethnicity. The column marked "estimate" reports the predicted change in F2 in Hz.

The F2 model finds a small but significant difference between the F2 of LOT and THOUGHT; for Anglo speakers, the predicted size of the contrast is approximately 70 Hz. However, there is a strong interaction between vowel class and the other two ethnicities: African American speakers have twice as large a contrast, whereas the contrast is almost completely neutralized for Mexican American speakers.

Table 11.6. Austin LOT/THOUGHT F2 heteroscedastic mixed-effects regression

	estimate	*p*-value
(intercept)	1216.47	2.2e-16
LOT	35.57	0.001
THOUGHT	−35.57	
male	−74.56	3.6e-05
female	74.56	
Anglo-American	−0.69	0.773
African American	11.21	
Mexican American	−10.52	
Anglo-American and LOT	−0.71	0.011
Anglo-American and THOUGHT	0.71	
African American and LOT	35.40	
African American and THOUGHT	−35.40	
Mexican American and LOT	−34.68	

Tests for Multivariate Outcomes

So far, F1 and F2 have been treated separately, focusing on F2's more robust separation of the vowel classes. Modeling the two formants separately makes the results difficult to interpret, since there may be some correlation between the two, especially near the bottom of the vowel space, (and of course, F2 is defined in such a way that it is always greater than F1). Just as multiple predictors are needed to deal with complex causal structures, tests for multivariate outcomes are a necessity when the outcome itself exists on more than one dimension. The designs considered in this section all convert the data into a per-speaker measure of the separation between the two vowel classes.

Euclidean distance. One way to compute a distance between two vowel classes is to compute the *Euclidean* (or Cartesian, or Pythagorean) *distance* (e.g., Gordon et al. 2004:145). This is simply the length that would be obtained by measuring the distance between two points in F1 x F2 plane with a ruler. The Euclidean distance between the points is given by the Pythagorean theorem: it is the square root of the sum of two quantities, the squared difference in mean F1 and the squared difference in mean F2. While this measure is intuitive, there are potential problems with it. First, the relative contribution of F1 and F2 to the ultimate distance measure is fixed to be equal, which may be undesirable when one of the acoustic measures has a larger range or different variance. Secondly, we have not addressed the possibility of correlations between F1 and F2; any correlative structure will be artificially inflated when they are combined in this fashion.

Multivariate analysis of variance. In a study of vowel merger, Hay, Warren, and Drager (2006) fit a type of multivariate outcome model called MANOVA to each speaker, using vowel class as the main predictor and F1 and F2 as outcomes. From these per-speaker models, Hay et al. compute a quantity called the *Pillai score* (or *trace*), which is simply the proportion of multivariate variance accounted for by the vowel class predictor. The Pillai score is near zero when no variance is accounted for by vowel class, and if all variance is due to vowel class, the Pillai score is one. This method also controls for any correlation between F1 and F2. Figure 11.1 plots the vowel-class medians of the two speakers in the Austin data with the highest, and lowest, vowel-class Pillai scores.

As can be seen, the tokens of the low Pillai score speakers are not well separated by vowel class, consistent with merger and their low score, and the speakers with the highest scores are well separated by vowel class, though these two speakers have very different acoustic targets.

Both Pillai score and Euclidean distance for the LOT-THOUGHT contrast as produced by the speakers in the *Atlas of North American English* (Labov, Ash, & Boberg 2006) is plotted in figure 11.2. The speakers are plotted by

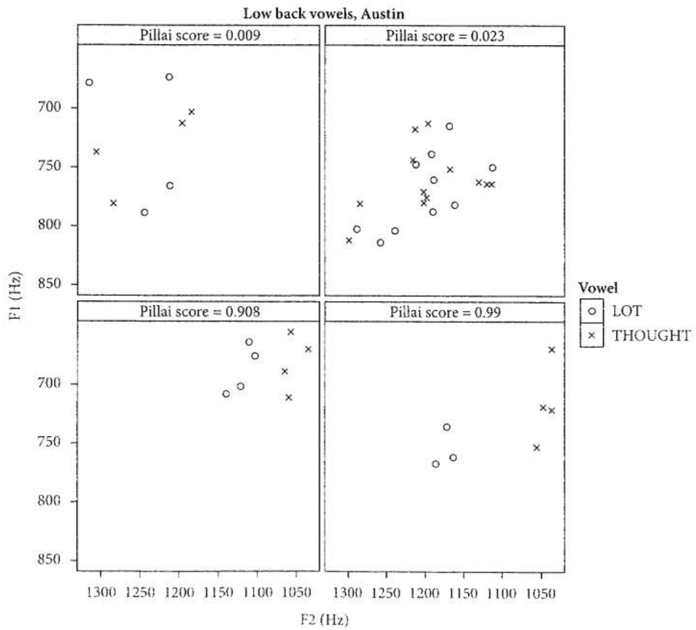

Figure 11.1. The two Austin speakers with the highest and the lowest Pillai scores for vowel class, respectively, and their vowel tokens in the F1 x F2 space

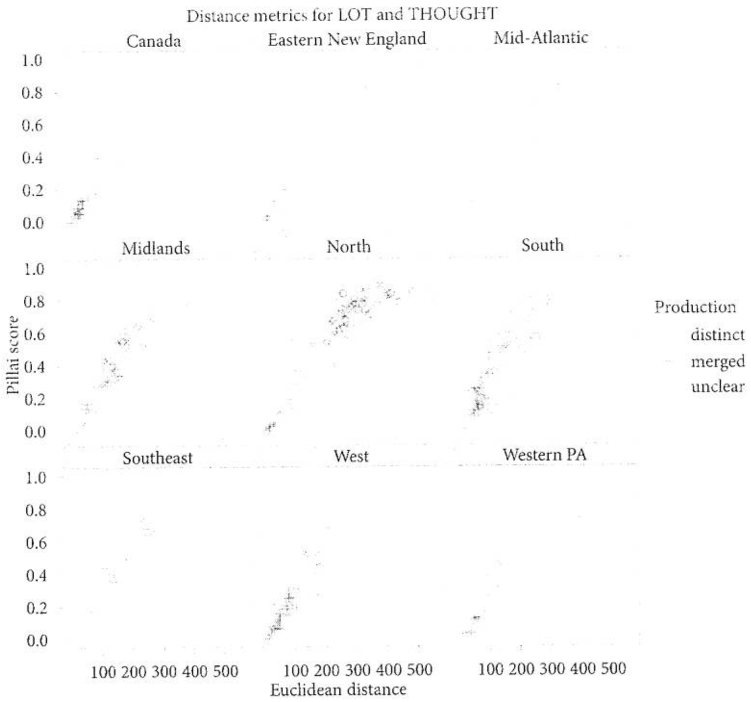

Figure 11.2. North American English LOT/THOUGHT distance by region and speaker

region (with the exception of the five speakers from New York City), and the shapes indicate the interviewers' impressionistic coding of the degree to which the speaker was perceived to be merged. As can be seen, the two measures are highly correlated in all nine regions. At least in this case, the Pillai score and Euclidean distance metrics produce results so similar that they can be used interchangeably. The one caveat is that Pillai score may not be appropriate when the number of observations per speaker is very small.

Summary

Sociolinguists can deploy a rich variety of methods in the study of continuous outcomes, including paired univariate methods with and without the assumptions of heteroscedasticity and normality (t-tests and Wilcoxon tests), mixed-effect linear regression, and models for correlated multivariate outcomes (MANOVA). As always, it is crucial to attend to the assumptions inherent in statistical techniques.

Conclusion

Having shown the effects that assumptions about the data make on the results of inferential analysis, the one assumption that remains to be considered is the *frequentist* paradigm itself.

A Bayes new world? Frequentism is the name given to the traditional approach to statistics that coalesced in the early twentieth century around the statisticians Egon Pearson, Ronald Fisher, and their collaborators, and implicitly assumed in this chapter; it stands in contrast with a second paradigm known as *Bayesianism*, after eighteenth-century minister Thomas Bayes, which has developed only in the past few decades. Whereas frequentist analysis is concerned with the probability of rejecting a null hypothesis, the Bayesian approach focuses on the change in probability of a null hypothesis before and after performing data collection and statistical testing. The following example illustrates this contrast.

Frequentism and Friday effects. A sociolinguist's data collection is very much influenced by prior knowledge about the speech community, universals of language variation, and so on. The null hypotheses that are ultimately subject to testing are generally quite likely to be false; in Bayesian terms, the prior probability of the null hypothesis is quite low, and consequently, its rejection is not a particular surprise. This logic has more interesting consequences when one considers a null hypothesis that is quite likely to be true, such as the hypothesis that New Yorkers produce the same rates of the variants of post-vocalic *r* on Fridays as during the rest of the week. If however a statistical test reports a significant Friday effect (e.g., $p = 0.03$), frequentist principles require the researcher to take this result seriously, even in the absence of a mechanistic explanation for any component of this correlation between days of the week and phonetic variation.

However, the Bayesian theory has a different take on this kind of unlikely, but significant, result. It is only rational that to reject such a strongly believed null hypothesis demands extraordinary evidence to cause us to shift our beliefs. Jeffreys (1939) describes a Bayesian method to integrate our prior beliefs about the non-existence of Friday effects with the result of the experiment. Before the experiment, the researcher must specify a prior probability of the null hypothesis. While Labov's study of New York City post-vocalic *r* discussed earlier has several replications over the last half century, none of the studies mention day-of-week effects, nor are they discussed in any sociolinguistic work of which we are aware. Given this, one might somewhat arbitrarily say that the prior probability of the null hypothesis is 0.99; that is, it is unlikely that the null hypothesis is false and there really is a Friday effect. The posterior probability (i.e., the probability that the null hypothesis is true after the statistical test) is given by dividing the p-value of the statistical test by the sum of the following two terms: the product of the p-value and the prior probability, and the product

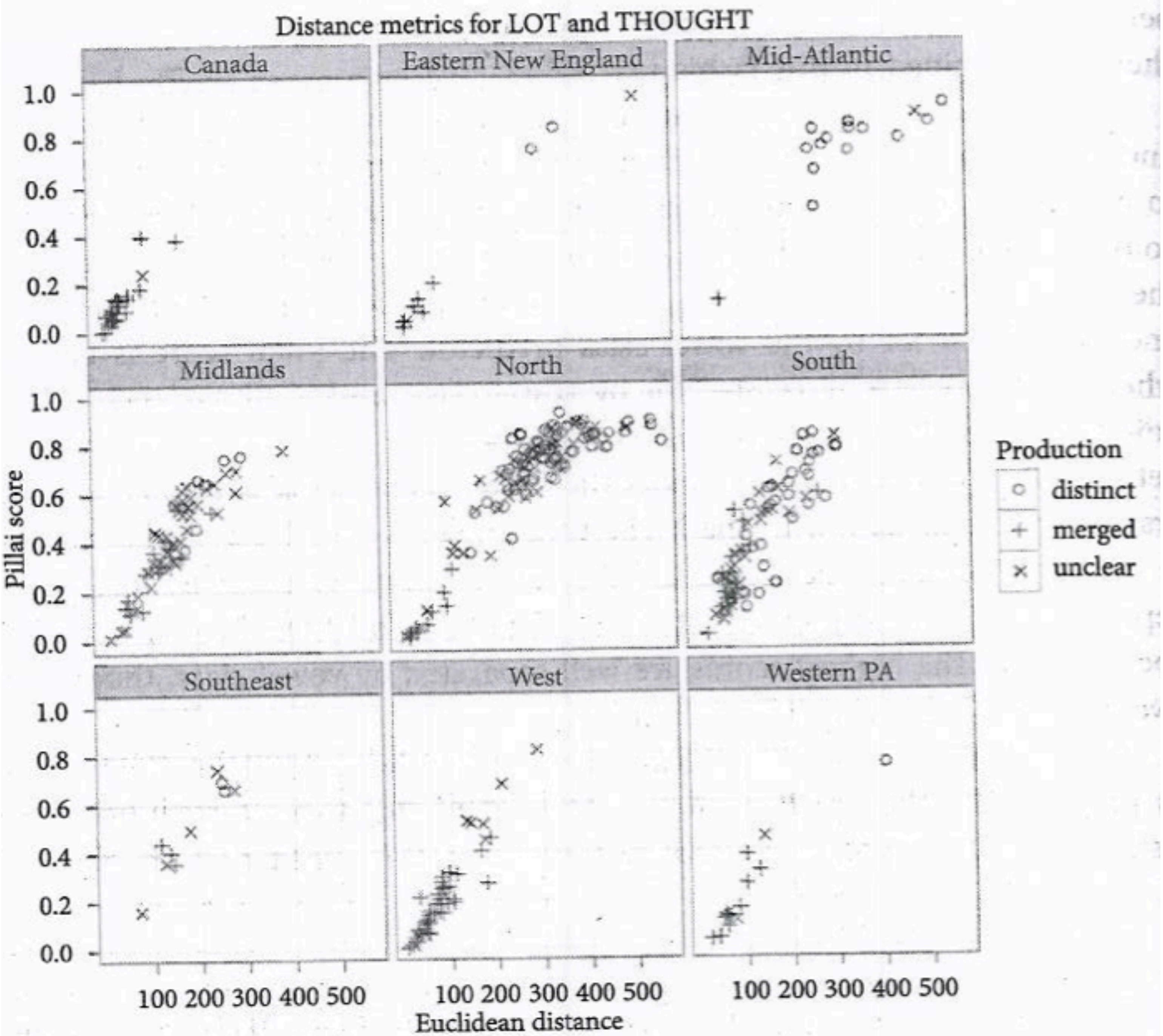

Figure 11.2. North American English LOT/THOUGHT distance by region and speaker

of the one minus the p-value and one minus the prior. For this example, the denominator is (0.03 x 0.99) + (0.97 x 0.01) = 0.039, and thus the posterior probability is 0.03/0.039 = 0.761. The change from the prior probability of 0.99 to the posterior of 0.7614 indicates that this single test has not deeply shaken the faith in the null hypothesis of no Friday effect. While sociolinguistics has not generally used such explicit computation of posterior probabilities, it seems unlikely a single Fridays effect study would take the field by storm, simply because of the a priori unlikeliness of such an effect. After all, events of probability 0.03 do occur by chance—3 percent of the time, to be precise—so a p-value of 0.03 should not lead to abject certainty in the veracity of the alternative hypothesis.

Proving the null. Conversely, when we apply statistical analysis under the frequentist approach and the data fails to provide strong evidence to reject a null hypothesis, it does not necessarily mean that the null hypothesis is true: the failure to reject the null hypothesis may be due to too little data or failing to control for nuisance variables, and it does not rule out the existence of some other alternative hypothesis that would result in rejection of the null. Bayesian statistical analysis tools allow the researcher to estimate the probability of the null hypothesis (e.g., Gallistel 2009). The finding of a non-effect (like Fridays) or non-interaction (for instance, between style and internal constraints on variants, as proposed by Sankoff & Labov 1979) may itself be of considerable sociolinguistic import, and only Bayesian methods are capable of identifying them.

Against a Statistical Monoculture

The collaboration between statisticians and sociolinguists in the 1970s was a fruitful one, but advances in statistics since then have been slow to diffuse into sociolinguistic practice. Mixed-effects models provide sociolinguists with an important new tool to excise the assumption that speakers or words, for instance, do not behave differently once appropriate demographic or grammatical constraints have been taken into account. While mixed-effects models provide a reasonable way to test this intuitively reasonable null hypothesis and identify when it is false, using fixed-effects models that fail to address these concerns may produce spurious inferences (Gorman 2010).

Cross-fertilization. Sociolinguists have recently availed themselves of sophisticated psycholinguistic paradigms (Loudermilk, chapter 7 this volume), and the potential for collaboration is clear. In psycholinguistics, subject and word effects have been addressed for decades (e.g., Clark 1973), and the leading *Journal of Memory and Language* recently dedicated an issue (volume 59, no. 4) to best practices in statistical analysis, which recommends mixed-effects models for these purposes. A shared statistical vocabulary will only strengthen the alliance between sociolinguists and psycholinguists (among other research communities).

Some new tools. Sociolinguists have long benefited from free software packages like VARBRUL (and descendants like GoldVarb). In the past, those who wished to use other methods (such as more-general regression models) had no choice but to pay large sums for proprietary statistical software. Sociolinguists can now avail themselves of a huge library of statistical methods using the free, cross-platform environment known as R, which has become the lingua franca for quantitative analysis of every stripe. R, of course, is capable of emulating the features of VARBRUL.[4]

There is a learning curve associated with a new statistical interface, and R is no exception. The second author is the creator of Rbrul (Johnson 2009), which provides a guided interface for regression modeling with random effects, continuous predictors and/or responses, and interactions, but this is only a stopgap measure, and researchers wishing to avail themselves of the full power of modern statistics will need the full power of modern statistical software. While this may seem daunting, it is important for sociolinguists to rise to the statistical challenges posed by their complex and meaningful data in furtherance of science.

Acknowledgment

Thanks to all those who provided the data analyzed in this chapter, especially Miriam Meyerhoff, Lynn Clark, Dominic Watt, David Bowie, Douglas Bigham, and Lars Hinrichs. We would also like to thank Douglas Bates and other members of the R-sig-ME mailing list for their wisdom.

Notes

1. In this chapter, we use p-values to report the "exact level of significance" (Gigerenzer, Krauss, & Vitouch 2004), and thus do not limit ourselves to statements like "$p < 0.05$."
2. One such case was brought to our attention by Robert Bayley. Lucas, Bayley, and Valli (2001), in a large-scale survey of variation in American Sign Language (ASL), translate the age of a signer into a three-level predictor, under the reasonable hypothesis that the true relevance of age to variation in their population is that informants of different ages encountered different administrative policies toward ASL in the classroom. Recent work by these authors and collaborators (McCaskill, Lucas, Bayley, & Hill 2011), which focuses on ASL in the African American community, groups signers into those who attended school before and after integration of the schools for Deaf children in the southern United States.

3. The models presented in this section depend on having data on the level of the observation, making it possible to control for any grammatical predictors on opposing variants. If all that is available are the percentages of variants used by each speaker, an appropriate method is *compositional data analysis* (Aitchison 2003).
4. All the analyses and graphics in this chapter were created in R. Some of the code used in this chapter is available at http://ling.upenn.edu/~kgorman/papers/handbook/.

REFERENCES

Aitchison, J. 2003. *The statistical analysis of compositional data*, 2nd ed. Caldwell, NJ: Blackburn Press.

Baayen, R. H., Davidson, D., & Bates, D. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412.

Bayley, R. 2002. The quantitative paradigm. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *Handbook of language variation and change*, 117–41. Oxford: Blackwell.

Bigham, D., White-Sustaíta, J., & Hinrichs, L. 2009. Apparent-time low vowels among Mexican-Americans and Anglos in Austin, Texas. Paper presented at NWAV 39, University of Ottawa.

Borowsky, T. 1986. Topics in the lexical phonology of English. Ph.D. dissertation, University of Massachusetts, Amherst. Published by Garland, New York, 1991.

Bowie, D. 2001. The diphthongization of /ay/: Abandoning a southern norm in southern Maryland. *Journal of English Linguistics* 29: 329–45.

Bybee, J. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14: 261–90.

Cedergren, H. 1973. The social dialects of Panama. Ph.D. dissertation, Cornell University.

Clark, H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12: 335–59.

Cohen, J. 1983. The cost of dichotomization. *Applied Statistical Measurement* 7(3): 249–53.

Gallistel, R. 2009. The importance of proving the null. *Psychological Review* 116: 439–453.

Gelman, A., & Hill, J. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press..

Gigerenzer, G., Krauss, S., & Vitouch, O. 2004. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (ed.), *Sage handbook of quantitative methodology for the social sciences*, 391–408. Thousand Oaks, CA: Sage.

Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A., & Trudgill, P. 2004. *New Zealand English: Its origins and evolution*. Cambridge: Cambridge University Press.

Gorman, K. 2009. Hierarchical regression modeling for language research. Technical Report 09–02, Institute for Research in Cognitive Science, University of Pennsylvania. http://repository.upenn.edu/ircs_reports/202/.

Gorman, K. 2010. The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. *University of Pennsylvania Working Papers in Linguistics* 16(2): 66–75.

Guy, G. R. 1980. Variation in the group and the individual: The case of final stop deletion. In W. Labov (ed.), *Locating language in space and time*, 1–35. New York: Academic Press.

Guy, G. R. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3: 1–22.

Harrell, F. 2001. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hay, J., Warren, P., & Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34: 458–84.

Herold, R. 1990. Mechanisms of merger: The implementation and distribution of the low back merger in Eastern Pennsylvania. Ph.D. dissertation, University of Pennsylvania.

Jeffreys, H. 1939. *Theory of probability*. Oxford: Oxford University Press.

Johnson, D. E. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–83.

Johnson, D. E. 2010. *Stability and change along a dialect boundary: The low vowels southeastern New England*. Publications of the American Dialect Society 95. Durham, NC: Duke University Press.

Johnson, D. E. In press. Descriptive statistics. In R. Podesva & D. Sharma (eds.), *Research methods in linguistics*. Cambridge: Cambridge University Press.

Labov, W. 2001. *Principles of linguistic change: Social factors*. Malden, MA: Wiley-Blackwell.

Labov, W. 2006. *The social stratification of English in New York City*, 2nd ed. Cambridge: Cambridge University Press.

Labov, W., Ash, S., & Boberg, C. 2006. *The atlas of North American English: Phonetics, phonology, and sound change*. Berlin: Mouton de Gruyter.

Lennig, M. 1978. Acoustic measurements of linguistic change: The modern Paris vowel system. Ph.D. dissertation, University of Pennsylvania.

Llamas, C. 2010. Convergence and divergence across a national border. In C. Llamas & D. Watt (eds.), *Language and identities*, 227–36. Edinburgh: Edinburgh University Press.

Lucas, C., Bayley, R., Valli, C. 2001. *Sociolinguistic variation in American Sign Language*. Washington, DC: Gallaudet University Press.

McCaskill, C., Lucas, C., Bayley, R., & Hill, J. 2011. *The hidden treasure of Black ASL: Its history and structure*. Washington, DC: Gallaudet University Press.

Neu, H. 1980. Ranking of constraints on /t, d/ deletion in American English: A statistical analysis. In W. Labov (ed.), *Locating language in space and time*, 37–54. New York: Academic Press.

Pinheiro, J., & Bates, D. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.

Rousseau, P., & Sankoff, D. 1978. Advances in variable rule methodology. In D. Sankoff (ed.), *Linguistic variation: Models and methods*, 57–69. New York: Academic Press.

Sankoff, D., & Labov, W. 19679. On the uses of variable rules. *Language and Society* 8: 189–222.

Schleef, E., Clark, L., & Meyerhoff, M. 2011. Teenagers' acquisition of variation: A comparison of locally-born and migrant teens' realisation of English (ing) in Edinburgh and London. English World-Wide 32: 206–36.

Tagliamonte, S. & Temple, R. 2005. New perspectives on an ol' variable: (t, d) in British English. Language Variation and Change 17: 281–302.

Wells, J. C. 1982. *Accents of English*. 3 volumes. New York: Cambridge University Press.

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
The Oxford handbook of sociolinguistics / edited by Robert Bayley, Richard Cameron, and Ceil Lucas.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-19-974408-4 (alk. paper)
1. Sociolinguistics—Handbooks, manuals, etc. I. Bayley, Robert, 1943–
II. Cameron, Richard, 1953– III. Lucas, Ceil.
IV. Title: Handbook of sociolinguistics.
P40.O94 2013
306.44—dc23
2012020046

ISBN 978-0-19-974408-4

1 3 5 7 9 8 6 4 2
Printed in the United States of America
on acid-free paper