Words and Numbers:
Statistical analysis in sociolinguistics

Gregory R. Guy
New York University

1. Introduction.

The origins of linguistics lie mainly in the humanities, but humanistic disciplines have traditionally been little concerned with quantitative analysis. Consequently, naïve observers, and even many linguists, may wonder about the place of numbers in a discipline focused on words. The importance of statistics for linguistic research emerges most clearly when one investigates questions that lack categorical answers.

Much of linguistic theory, at least since the Neogrammarians and their exceptionless sound laws, has focused on the categorical: the generalization that admits no exception. When properties of language are categorical, no quantitative analysis is necessary. Thus English speakers do not require statistical studies of English NPs to be convinced that articles never follow nouns. A phrase like *the dog* is grammatical, while the alternative **dog the* is not, and never occurs, and this generalization is so comprehensive that English speakers would likely find it silly to count up instances to confirm it.

However, there are many interesting facts about language that involve relative, not absolute, properties – that involve relations of *more* and *less* rather than relations of *either/or*. Such properties are necessarily described in quantitative terms. In some approaches, quantification is only implicit: thus optimality theory sees some constraints as more determinative than others, but the only quantification that OT admits is an ordinal scale of constraint ranking. Other approaches adopt explicit quantification; one of these is variation theory in sociolinguistics, which was developed to describe and model quantitative patterns in everyday language use.

In their everyday lives, speakers systematically adapt their pronunciation, grammar, lexicon, and discourse strategies to address different hearers or to serve different communicative ends. Also, individual speakers and groups defined by social characteristics and practices, differ systematically in usage. To adequately characterize this rich patterning of speakers' sociolinguistic knowledge requires statements of more and less: older speakers use an innovative form less than younger speakers, upper class speakers use more of a prestige variant than working class speakers; everyone in a community typically uses more prestige variants in their more careful speech styles.

Consequently, if linguistics is going to faithfully and adequately describe and model the social and psychological processes that give rise to any of these systematic, but non-categorical alternations in language, it must adopt a quantitative apparatus. Since we are not the first discipline to do this, it behooves us to study the prior discoveries in this area of other disciplines, principally mathematics and statistics. This chapter will seek to provide the basic elements of such a study, an introductory lesson that we might call 'numbers for wordsmiths'.

2. Quantitative approaches to generalization

Perhaps the most essential strategy of linguistic analysis is generalizing: attempting to move beyond individual cases to formulate a general rule, principle, or pattern. This is, of course, a strategy that is basic to human intelligence. There is even an English aphorism describing failures of generalization: 'he can't see the forest for the trees'. In quantitative analysis, many of the basic methods could be characterized as quantitative generalizations, techniques for perceiving the shape of the forest through all the trees. Let us briefly consider some of these: the calculation of percentages and ratios, the definition of variables, and measurements of central tendencies.

2.1 Ratios. Consider the following data from a study of relative clauses in Australian English, comparing two speakers from different social class backgrounds. There is of course a three-way variation in relativizing strategies in English: relative clauses can be introduced with a *wh-* word, *that*, or zero: *Here's the book {which, that, ø} I mentioned*). In this study, the working class speaker was found to use 2 tokens of a *wh-* relative (i.e. *who*) in subject position with a [+human] antecedent, while the middle class speaker used 49 tokens in the same context. If we assumed for the moment that these speakers are representative of their social classes, would it be reasonable to hypothesize that middle class speakers use *wh-* forms more often than working class speakers?

The difference between the two figures – 2 vs. 49 – is substantial, so the unreflecting impulse might be to answer yes, but based only on this information, the thoughtful answer should be a thundering NO! What we have so far is only a fragment of the information necessary for inferring a pattern. To conclude that the MC speaker uses more *wh-* forms based on this information would be like going out to look at trees, reporting that you saw five oaks, and then concluding that you had found an oak forest. Such inferences are only valid, or even possible, if we have more information about the technique that produced these numbers: in the forest case, how many trees did you look at and what other kinds did you find? Did you find nothing but oaks, or did you walk past hundreds of maples in order to find your quintet of oaks? To make a reasonable generalization requires knowing all of the relevant observations, not just a selectively reported subset. In the relative clause study, some crucial additional information would be, for example, how many opportunities did each person have to produce *wh-* forms, and what other types of relativizers were used?

A crude measure of opportunities would be observation time: for how many minutes or hours of speech was each subject observed or recorded. If you recorded the working class speaker for 10 minutes, and while the other speaker was recorded for 20 hours over a period of several days, then these data might lead us to the opposite conclusion. What we need, in short, to formulate hypotheses or to see the makeup of the forest, is information on **data quantity**, and the **ratio** of the various kinds of observations. A hypothesis will be more credible when it is based on more data, and it will go in the direction suggested by the ratio, rather than the raw counts of samples of different sizes.

In the study at hand, it happens that the MC speaker was actually recorded for 50% more time than the WC speaker: 3h45m as opposed to 2h30m. But this is a crude measure of opportunities to produce a given relativizer; if one speaker was talkative and the other taciturn, the number of relative clauses each produced might be expected to differ dramatically even if they were recorded for the same amount of time. Hence the most straightforward approach to formulating an estimate of the speakers' relative usage of *wh-* forms is to count up all their relative clauses and look at the ratio of *wh-* to *that* forms, or the percentage of *wh-* as a fraction of all relative clauses, or some similar measure.

For the relative pronoun study, we can provide the missing information as follows: in addition to the *wh-* forms, the middle class speaker produced 33 cases of the relativizer *that*, while the working class speaker produced 9 such cases. (Note that we are still restricting our attention to human antecedents and subject-position relatives; hence there were no cases of Ø relativizers.) The data can now be summarized as follows:

	MC speaker	WC speaker
who	49	2
that	33	9
% wh-	60%	22%

Given this information, we now know the total number of relative clauses observed (in the specified context), and we can now compare the ratios of usage of *wh-* as opposed to other alternatives. The crucial statistic for present purposes is the percentage of use of *wh-* forms, and we can now make a meaningful statement that indeed these two speakers do differ in *wh-* usage, at least in subject position with human antecedents. The MC speaker uses *wh-* forms more often in this context, but we say this not because he used a total of 49 of them, but rather because 60% of his relative clauses were introduced by *wh-* forms, as opposed to the 22% usage of the WC speaker. It is the fractions we are comparing, not the numerators of these fractions. Had we found 450 cases of *that* for the MC speaker, he would have been using *wh-* forms only 10% of the time, and we would come to the opposite conclusion.

Now, what about generalizing further, beyond the context in which these data were observed. Based on these data, limited to human antecedents and subject relatives, can we justifiably infer that the MC speaker generally uses more *wh-* forms? Whether or not such reasoning is valid depends on the context of observation and how representative it is, or is thought to be, of the total range of possible contexts. In this case, given what is known about English relatives, the selected context is arguably quite unrepresentative, so that a broader generalization based just on this data is unjustified. In the first place, subject relatives often disallow the zero relativizer, so the cited context eliminates one of the three alternatives of this variable.[1] In addition, there is a high level of linguistic insecurity among many English speakers regarding the case-marking of the relative pronoun *who/whom*. This case marking is prescriptively favored in standard English, but it is absent from the productive grammar of most modern English speakers, and is of course absent from all other possible relativizers in English (there is no case marking of *that, which, when, where, why,* or *Ø*). Consequently, some speakers appear to adopt an avoidance strategy to escape the necessity of figuring out whether *who* or *whom* is the prescriptively appropriate form: instead, they use *that* or *Ø* forms with human antecedents.

Therefore, before generalizing about these speakers based on the data I have presented so far, it would be wiser to examine their usage in other contexts that allow a full range of variants and are not muddied by linguistic insecurity. Consider for example the following data on their usage in object relatives with non-human antecedents:

	MC speaker	WC speaker
which	9	5
that	20	6
Ø	27	12
% wh-	16%	22%

In this context, the WC speaker actually uses a higher percentage of *wh-* forms than the MC speaker, reversing what was found in the previous context. Across all contexts other than the subject relatives with human antecedents, the total figures were 36% *wh-* forms for the MC speaker, and 33% for the WC speaker – nearly identical. Hence overall there is no evidence for a systematic difference in rate of *wh-* use between these two speakers, which shows the importance of evaluating the contexts from which data are drawn before making generalizations.

2.2. The definition of variables. The above example also serves to illustrate some additional points about working with variables: the kind of relationship that is assumed to obtain between the variables under consideration, and how they are defined on a quantitative-qualitative axis. The first of these is often theoretically illuminating. In the relative pronoun example we were investigating an association between two different dimensions: a linguistic dimension (the use of relativizers) and a social dimension (speaker's social class). Each of these may be treated as a variable, in the mathematical sense. This in itself is a preliminary generalization. The various relativizers are treated as constituting alternative ways of fulfilling the same syntactic function, and the speakers in the study are treated as instantiating alternative values of an attribute (social class), that all speakers in this society are presumed to possess. Hence MC and WC are possible values of the variable 'class' and *which, that,* and *Ø* are possible values of the variable 'relativizer'.

In this example, we also have a clear sense that these two variables have a different status in our understanding of how the world works. Most commonly, sociolinguists think of the rate of use of the linguistic feature as somehow caused, influenced, or determined by the social feature. Applying the terminology used in statistics for such a distinction, we would call the choice of relativizer a "dependent variable", while the speaker's social class would be considered an "independent variable". Such terminology implies a dependency or even causal relationship between the two dimensions. In the example at hand, this would be justified on the basis of an expectation that a speaker's momentary decision to use or not use a *wh-* relativizer could not influence their social standing on a class scale which is ordinarily seen as a function of more durable social traits like educational and occupational level. A Ph.D. who utters "ain't" on some occasion does not thereby lose their post-graduate educational history. The same would be true of other social dimensions: sex and race are normally permanent characteristics of a person's identity, and should therefore be independent of their momentary linguistic choices. Rather, it should be the linguistic choices that are determined (or perhaps restricted) by the speaker's social identity and prior linguistic experience.

However, the terminology of dependent and independent variables should not make us lose sight of the fact that this distinction is highly conceptual, deriving, in effect, from some 'theory' of the world. Adopting a different point of view may shift the relationship between a given pair of variables. Thus if we construe social identity as a construct based on an individual's performance of certain practices, including linguistic practices, then we might talk about the use of prestige linguistic variants as an independent variable that speakers use when they are seeking to construct an identity as an educated, middle class, person. Gender identities – masculinity and femininity – as well as ethnic identities, and social relationships like boss, teacher, friend, etc. are all at least partially constructed through linguistic practice, such that making certain linguistic choices contributes to the establishment and maintenance of the social identity. From this perspective, one might reasonably construe the linguistic variables as independent, and the social identities as dependent.

The same trade-off between dependent and independent can be encountered within the domain of the linguistic variables, insofar as they are interconnected. In vernacular Brazilian Portuguese, for example, verbs rarely agree with post-posed subjects; what is the direction of dependency? Research on this topic has most commonly assumed that the word order is prior, and the agreement is dependent, but it could be conceptualized as the other way around: perhaps agreement blocks postposition of the subject. Or perhaps both agreement and word order are triggered by something else: raising of a subject across the verb or AGR position. The general point here is that what is dependent and what is independent is not given by statistical methods, but by one's prior assumptions or theories. It is sometimes illuminating to test those assumptions by exploring alternative possible dependency relations between the variables under investigation.

The second point to note about variables is their place in a typology of quantitative vs. qualitative, continuous vs. discrete. This characterization affects the kinds of statistics that can be utilized in one's analysis. In the case of English relativizers, the linguistic variable has three possible realizations, but each is discretely different from the others: they do not form a continuum, and there are no possible intermediate values of this variable – e.g. nothing that is, say, one-quarter of the way between *which* and *that*. This is what is termed a **nominal** variable. Such variables label particular categories that are treated as qualitatively distinct from the other categories in the analysis. Other linguistic variables which are usually treated as nominal are word order (a clitic might precede or follow a verb, but can't be any place else), deletion or non-deletion of a segment (there's no 'partial deletion'), grammatical categories such as person, number, gender, case (e.g. English nouns are either singular or plural; there is no number scale with intermediate values of 'partially plural'), etc. Social variables may also be nominal: speaker's sex, nationality, L1 vs. L2 status, etc. are normally construed as nominal variables in which each possible value of the variable is a discrete, qualitatively different category, without intermediate values.

However, it is clear that other variables that linguists work with do not have this nominal, discontinuous nature. In the social domain for example, a speaker's age is an intrinsically continuous scale with an infinite variety of intermediate values. The same is true of income, years of schooling, length of residence in a particular country or dialect region, etc. In the relativizer example cited above, the social variable of social class was instantiated by only two individuals, who could be thought of as constituting a nominal variable, but many sociolinguistic studies treat class as a continuum, with fuzzy, non-discrete boundaries between the points on the class scale. Often in such work social class or ('socioeconomic' class) is operationally defined in terms of some kind of quantified scale, such as the nine-point scale used by Labov in his path-breaking New York City study (1966).

Among linguistic variables, the same is true. Vowel articulations, for example, are notoriously continuous. Although we phonemicize them, labeling particular articulations as tokens of the category /i/ and others as tokens of the category /I/, the articulatory and acoustic regions over which these vowel sounds are defined have no hard boundaries, and in the course of linguistic change, they are continuously deformed into each other; intermediate articulations are not only possible, but occur frequently. Other phonetic/ phonological properties with this continuous character include pitch, stress, voice-onset time, etc. Acquisition (both L1 and L2) is ordinarily thought of as a continuous variable, along which speakers could be infinitely differentiated.

These variables which are not nominal are typically treated as quantifiable in some dimension: there is some trait which a given realization may have more or less of in a quantifiable way. For age, a speaker has a given quantity of years and days of life; for vowels, a given articulation can be defined in terms of the formant frequencies that are produced, which vary continuously within a range of approximately 200-5000 Hz. But it is worth noting additional distinctions among the non-nominal, quantitative variables. One important type with particular properties are the **ordinal** variables. These are values that form a scale with a rank order – a weak quantification in which there is a particular directionality, so that transitivity obtains (if A>B and B>C, then A>C), but in which there is no defined sense of the distance between any points on the scale. Such a variable can be contrasted with **interval** or continuous variables in which the separation between given points has a measurable definition. These two types can be contrasted by considering the results of an Olympic marathon. The order in which the contestants finish is an ordinal scale: the first place finisher gets the gold, second place silver, and so on. But this doesn't tell us how much time or distance separated each finisher; the gold-medal winner might have been just a half step in front of the silver medallist, who in turn could have been 500 meters in front of the third-place finisher. Providing this additional information requires an interval scale, for example a list of the times that each contestant took to run the race. In linguistics, a currently well-known example of an ordinal scale is the constraint hierarchy of Optimality Theory. In its orthodox version, this is purely ordinal: we can state that constraints X, Y, and Z are ranked in that order, but the theory makes no provision for concepts like X being just a little bit higher than Y, but Z falling way behind.

It should be noted that in analytical practice, it is often possible to choose whether to treat a given variable as discrete or continuous. Thus speaker's age could be treated as a continuous variable, or the age continuum can be segmented into groups, such as adolescent, younger adult, older adult, which effectively makes it a nominal variable. The same is true of linguistic variables. For example, studies of lax vowel lowering in Canadian English such as De Decker & Mackenzie 2000 and Hoffman 1998 classified tokens of /I/ and /E/ auditorily as either lowered or not, a binary, nominal variable. But vowel height is, of course, a continuous function of the vowel's first formant. Accordingly, De Decker 2002 treats the same phenomenon using normalized acoustic measurements of F1 and F2.

2.3. Central tendencies. What are the statistical consequences of these distinctions? Particular statistical methods are associated with particular types of variables. For example, consider the common statistical techniques for identifying a 'central tendency' – a value which attempts to characterize some kind of 'center' in a collection of data. The best known such measure is the **mean** or arithmetic average. The mean is a statistic that is only relevant for quantitative scales (normally interval or continuous scales, although it is occasionally used with ordinal variables). Thus if we had a group of speakers, we could calculate their average age or average income, and if we had a set of productions of the vowel /i/, we could calculate average F1 and F2 values. The procedure is to sum all the values and divide by the number of cases. In the relativizer example above, we can calculate the mean percentage of *wh-* forms for the two speakers (41% giving each speaker equal weight, or 55% if we weight them by data quantity). But what we cannot do in that example is calculate the average relative pronoun used by each speaker. There is no 'average' of *which, that,* and Ø forms. Neither could we calculate the average sex or nationality of a sample of speakers, nor the average grammatical case of a collection of pronouns. These are nominal variables, and the mean is undefined for such cases (although one occasionally finds it used metaphorically, such as in statements like "the average nursing home resident in the United States is a white female").

A mean measures the 'center' of a set of scalar values in the sense that it constitutes the number which differs from all the values in the set by the same total amount in both positive and negative directions. In other words, the difference between the mean and all the values higher than it totals up to the same amount as the difference between the mean and all the values smaller than it. This is a very common measure of central tendency which everyone is familiar with in everyday life. But it has a well-recognized defect for certain kinds of distributions: it is powerfully affected by extreme values. Thus if ten speakers in a sample used some variable at a rate of 1%, while one other speaker used it a a rate of 100%, the average percentage of usage would be 110/11 = 10% (assuming all speakers were observed the same number of times.) What does this number mean? It is likely that the one speaker who uses this form all the time is anomalous, an 'outlier' in the graphical sense, so if we ignored him or her as atypical, the rest of the group would have an obvious central tendency of 1% usage. The mean figure of 10% doesn't characterize anyone in the sample; in fact, it doesn't even come close to any observed value. Therefore, we might wish for an alternative measurement of central tendency that avoids such a distorted outcome. One measure that achieves this is the **median**. It can be used for any quantitative variable, and consists of simply the middle value in a list of values ordered by size. When there is an odd number of cases, it is the middle case (e.g. the third value in a list of 5, or the 6th value in a list of 11 items), while with an even number of cases the median is the average of the two middle-most values (e.g. the third and fourth in a list of 6). In the example just given, it would be 1%, because counting from the lowest to the highest rates of use of the variable, the 6th speaker would use 1%, (as would the first through tenth speakers).

Such measurements can describe the central tendency of a set of quantitative variables, whether interval, ordinal, or continuous. Is there any way to describe the central tendency of a set of observations of a nominal variable? One useful device in such situations is the **mode**. This is quite simple: it is just the value that occurs most frequently. If we were looking at pronoun cases in a text, and observed 85 pronouns in the nominative case (*he, I, we*, etc.), 47 in the accusative (*him, me, us*), and 16 genitive (*his, mine, ours*), then the modal case would be the nominative, as there are more items marked with this case than any other. This is what is really meant when people make statements about the "average" value of a nominal variable: "the average nursing home resident is a white female" really means "the modal nursing home resident is a white female", i.e. there are more white females than white males or non-white males or females living in nursing homes.

Note that, although the mode is the only central tendency measure that can be meaningfully used with nominal scales, it is also useful in some circumstances for analyzing interval or ordinal scales. A set of measurements of the F1 of the nucleus of the vowel /æ/ might find, in an /æ/-raising dialect like Chicago, that there were more tokens at 450 Hz than any other frequency, and this would be a valid measure of one kind of central tendency. The same study could easily find a median value of, say, 475 Hz, and a mean value of 500 Hz, and all three of these measures would be valid descriptions of different kinds of central tendency. Furthermore, it is possible to find more than one mode, if there are several clusters of data points at different values. This is in itself an important finding about a data set which is obscured by the median or mean, which always yield just a single value, even if the data are distributed like the weights on a barbell.

3. Statistical inference: the significance of significance

The discussion so far has dealt with what are known as descriptive statistics, which are ways of describing the forest as a whole, based on information about the trees within it. But there are other things that the analyst would like to be able to do with quantitative analysis, like making informed decisions based on a knowledge of the odds, the way a poker player knows not to draw to an inside straight. When we have only partial information rather than complete, omniscient knowledge of a situation (which is normally the case), we would like to be able to extrapolate from the part to the whole in a reasonable way. Statistical methods for doing these sorts of things are called 'inferential statistics'. A central concept in this field is that of statistical significance, a much misunderstood term.

Statistical significance is essentially a way of estimating how likely it is to get a given distribution of data given certain assumptions about the nature of the source from which the data are drawn. In scientific studies, the available data are almost always a subset of the total possible data set, in other words, a sample drawn from a 'universe'. Thus a study of relative pronoun choice in English cannot possibly hope to investigate all the relative pronouns uttered by English speakers, nor can a study of the variable realizations of coda /-s/ in Spanish study all the times that Spanish speakers utter this sound in this syllabic position. Therefore, we investigate a sample, and seek to draw inferences about the statistical patterning of the universe from this sample. A sample, of course, can easily deviate from the universe in various ways. For example, we know from studying coins that they should have a 50% chance of coming up heads when flipped, and in a very large sample of coin flips, the observed percentage of heads should converge on 50%. But if we flip a coin twice, do we necessarily expect one head and one tail? Clearly not. Indeed, we might not even be surprised to get 4 or 5 heads in a row, because in a universe where heads were randomly but evenly balanced with tails, 5 consecutive heads should occur once every thirty-two (2^5) trials. But if we flipped a coin 200 times, and got heads every time, we would begin to wonder whether the universe from which those flips were drawn really did have an equal likelihood for heads and tails, because the chance of 200 consecutive heads would be equal to 1 in 2^{200}, a number so small as to render our starting assumption highly unlikely. Consequently, we might begin to entertain alternative assumptions, such as that the coin in question has heads on both sides!

Statistical tests of significance work in this way; they provide standard reference values which can be tested against known distributions to evaluate the likelihood that the observed data come from such a distribution. They are most commonly phrased with reference to the "null hypothesis", which always states that nothing is going on, the source distribution is normal, the independent variables do not influence the dependent variables, etc. In the case of coin flips, it would state that neither heads nor tails is more likely to occur. The ultimate significance statistic is usually stated in terms of the probability that the null hypothesis is true; this value is conventionally represented as *p*. If this number is small, meaning that the null hypothesis is very unlikely, then the results are said to be "statistically significant", meaning that it is reasonable to entertain some other hypothesis about the nature of the universe. "Small", in this context is generally taken to mean a value less than .05 or .01; in other words, if there is less than a 5% or 1% chance that the data are drawn from a universe in which the null hypothesis is true, this means that there is a 95% or 99% chance that the source universe really has a different distribution of the data, such as a real and significant effect of some independent variable on your dependent variable.

Consider a sociolinguistic example. Cedergren & Sankoff (1974) report that in Montreal French, the complementizer *que* is variably deleted: sometimes it is present and sometimes absent. Furthermore, in a set of sociolinguistic interviews with 16 speakers, the rate of absence appeared to be correlated to the social status of the speaker. Some figures are given in Table 1.

Table 1. Complementizer *que* in Montreal French[2] (from Cedergren & Sankoff 1974)

	Que absent	*Que* present	% absent
Working Class	28	90	23.7%
Professional Class	3	130	2.3%

The overall percentage of que-absence is appreciably higher for the working class speakers. But can we infer that this is true of the universe? If it were possible to study all utterances containing complementizer *que* by all speakers of Montreal French, then we could answer this question definitively; but since that is not possible, we can only draw inferences from the data that we do have. What inference is reasonable?

The null hypothesis, in this case, would state that, no, class DOESN'T have anything to do with use of *que* by Quebecois(e). Opposed to this would be the "experimental hypothesis", stating that yes, it does. Given the data in Table 1, either hypothesis is possible. The higher rate of *que* presence among professional class speakers might be due to simple sampling error: we merely happened by chance to encounter more utterances with retained *que* among the professional class speakers who were interviewed for the study, and if we had recorded them longer, or added other speakers, the apparent class difference would disappear. So we wish to move beyond statements of what is possible to statements about what is likely, and this is what statistical significance permits. We can compare the distribution in the sample with known facts about the distribution of samples drawn from populations in which the null hypothesis is true, and make a statement about whether the null hypothesis is likely to be true about the universe (i.e. Montreal French) from which these tokens of complementizer *que* were drawn.

3.1. Chi-square. One useful procedure for doing this is called the chi-square test. We begin this test by arranging the data in a contingency table, as in Figure 2, where each possible combination of 'contingencies' is given a separate cell. In our case, that means one square for each of the combinations of: utterances with *que* present produced by working class speakers, those produced by professional class speakers, and then utterances with *que* absent produced by working class speakers, and *que*-absent tokens produced by middle class speakers. In this table there are two variables, *que* realization and social class of speaker. Each has two possible values, giving a two-by-two contingency table, which is the four-cell core of Table 2. Note that we might assume that class is the independent variable, but this has no bearing on the procedure for calculating the chi-square statistic.

Table 2. Contingency table for Montreal French example

	Que absent	*Que* present	Totals
Working Class	28	90	118
Professional Class	3	130	133
Totals	31	220	251

Also given in Table 2 are the 'marginal totals' for the contingency table – the totals in each row and in each column, and the grand total. This is a preliminary requirement for calculating chi-square. The marginal totals represent the total number of items found for each value of each variable: i.e. all tokens collected from professional class speakers (in this example, there were 133), all tokens collected from working class speakers (118), all cases of absent *que*, regardless of who said them (31 tokens) and all cases where *que* was present (220). The grand total is the total N for the entire corpus, in this case, 251.

The marginal totals are required for the chi-square test because the logic of the test involves considering other ways that the same data might have been distributed across the cells, while preserving the same marginal totals. One way to see how this works is to consider a hypothetical case of some sociolinguistic variable with two possible realizations, A or B, which is examined in speakers belonging to two age groups, older and younger. Suppose that we collected one hundred tokens from each of the two age groups, and found that they were evenly divided between variants A and B. We would set up a contingency table as in 3.1, with marginal totals of 100 in each row and column, and a grand total of 200. Now, what distributions of items in the cells would preserve the row and column totals, and what conclusions would they suggest about any possible relationship between speakers age and the use of this variable? If it were the case that each cell had 50 tokens, as in Table 3.2 we would probably conclude that there was no association between age and this variable. This, in fact, is the distribution that the null hypothesis predicts. But in Table 3.3 we see another extreme: there are 100 tokens in the two cells on one diagonal, and zero in the other two. This preserves the same marginal totals, but shows a categorical association: older speakers use only variant A, while younger speakers use only variant B. Encountering such a distribution, most linguists would conclude that there is a rapid change in this community, with B supplanting A in apparent time. This would involve the rejection of the null hypothesis. Note that in both of these tables, the marginal totals are the same, even though they lead us to opposite conclusions. Consequently, in constructing a statistical test, we take those values as given; in this case, we could say they are determined by the size of the sample (exactly 100 tokens were collected from each age group) and by the overall rate of use of the variants in the community (which is 50% usage of each variant).

Table 3. Examples of contingency tables showing different degrees of association between speaker's age and a linguistic variable

3.1 Marginal totals

	A	B	Total
Younger			100
Older			100
Total	100	100	200

3.2 No association

	A	B	Total
Younger	50	50	100
Older	50	50	100
Total	100	100	200

X^2=0, p=1

3.3 Categorical association

	A	B	Total
Younger	0	100	100
Older	100	0	100
Total	100	100	200

X^2=800, p=0

3.4 Slight association

	A	B	Total
Younger	45	55	100
Older	55	45	100
Total	100	100	200

X^2=2, p>.20

3.5 Strong association

	A	B	Total
Younger	30	70	100
Older	70	30	100
Total	100	100	200

X^2=32, p<.001

Now in real data, one rarely runs into such extreme cases. More commonly we will encounter intermediate cases, like 3.4 and 3.5. Table 3.4 is only slightly different from the null hypothesis case, while 3.5 goes robustly in the direction of a strong but non-categorical association between age and usage. The task we wish our statistical test to address is, in a sense, to quantify where a given observed distribution falls on the continuum of possible distributions between the two extremes illustrated by 3.2 and 3.3. The question is phrased in terms of how likely it is to get whatever distribution we observe from a universe in which the data are distributed in a way analogous to 3.2. If that is highly unlikely, we will tend to conclude that the universe is not so constructed; in this example, we would conclude that there is a significant association between age and usage, perhaps because of ongoing linguistic change in the community.

The chi-square figures given for tables 3.2-3.5 illustrate how this statistic fulfills this task. For the data quantity and marginal totals given in the example, chi-square values range from a low of zero for the case showing no association between the variables (3.2), to a high of 800 for the case showing categorical association (3.3). At these extremes, the p value – the probability that the null hypothesis is true – approaches 1 when chi-square is 0, and approaches 0 when chi-square is very large (in this case, 800). But the statistic really does its job in the middle ranges. Table 3.4 gives a chi-square of 2, corresponding to a p somewhat higher than .2. This means that such a distribution could be drawn more than 20 times in a 100 trials from a universe in which the data were actually distributed evenly across these two variables, like in Table 3.2. Getting such a result, a researcher would ordinarily be unjustified in considering it as valid evidence for the existence of an age difference in this hypothetical community. But the pattern in Table 3.5 is much less likely to be drawn randomly from a null-hypothesis universe. The chi-square of 32 corresponds to a p value of less than .001, meaning there is less than one chance in a thousand that the null hypothesis is true in this case. (Actually, the chance is much less than this, more like one in 100,000.) Given such results, we would be nearly certain that there was in fact a real association between age and the use of this variable. Thus the chi-square statistic gives us, in effect, a quantification of distributions along the continuum from the balanced, null hypothesis distribution to the categorically imbalanced distribution; using this quantification, we can draw informed inferences about the universe based on the sample.

What, exactly, is the procedure for calculating chi-square? Returning to the Montreal French example, we now need to calculate what the distribution would be if there were no association between class and *que* presence or absence, but the marginal totals remained the same. This is equivalent to calculating the distribution where both social classes use the same percentage of *que*-deletion. This is easily done. The overall percentage of *que*-deletion is equal to 31 (total cases of deletion – the column total for the *que*-absent column) divided by the grand total of 251, which gives 12.35%. We multiply this figure by the N for each social class (i.e. each row), and get the figures in Table 4. These are called the EXPECTED values for the table, in contrast with the OBSERVED values that appeared in Table 2. This means 'expected if the null hypothesis were true', and each class had exactly 12.35% deletion.

Table 4. Expected values for Montreal French example, under null hypothesis

	Que absent	*Que* present	Totals
Working Class	14.57	103.43	118
Professional Class	16.43	116.57	133
Totals	31	220	251

The chi-square statistic is now going to be a function of the similarity or difference between the figures in Tables 2 and 4, between the observed and expected values. We start by subtracting the one from the other: for each cell we compute (observed-expected). This will necessarily yield some positive and some negative numbers, and to obtain all positive numbers we square them (because the square of any real number is a positive value). (This value, called the squared-difference is a common intermediate step in many statistical calculations. Note that in a two-by-two table like this one, it is the same value for all four cells, which is a reflection of the fact that this table has only one degree of freedom, as we will discuss below. In larger tables, this will not be the case.) Next, we need to make these values proportionate to a common reference point, so that cells with large N's don't contribute excessively to the statistic merely because they have more data. For this purpose we use the expected value in each cell, dividing the squared difference by the expected value.

Table 5. Chi-square calculations

a. differences between observed and expected values (observed-expected)

	Que absent	*Que* present
Working Class	13.43	- 13.43
Professional Class	- 13.43	13.43

b. squared difference = $(\pm 13.43)^2 = 180.36$

c. chi-square computation for each cell: $(\text{obs-exp})^2/\text{exp}$

	Que absent	*Que* present
Working Class	180.36/14.57=12.38	180.36/103.43=1.74
Professional Class	180.36/16.43=10.98	180.36/116.57=1.55

Total chi-square across all cells = 12.38+10.98+1.74+1.55=26.65

The result of these calculations is the chi-square statistic for each cell. We sum all the values for all the cells, and obtain a total chi-square for the entire table. In other words, the figure is computed as $\Sigma\ (\text{observed-expected})^2/\text{expected}$. This number will increase to the extent that any observed distribution diverges from a null hypothesis distribution. In the Montreal example, the figure is 26.65.

Now, it is possible to calculate for any given degree of divergence from a null distribution the probability of getting such a distribution by randomly sampling from a universe in which the null hypothesis is in fact true. Statisticians have done these calculations, and compiled tables summarizing the distributions of the relevant statistics in such circumstances. For present purposes, we are interested in the distribution of the chi-square statistic. A fragment of such a table is reproduced below in Appendix 1. Note several things about that table. In the body of the table are values of the chi-square statistic, and across the top are various specific percentage points, or decimal fractions. These are equivalent to the percentage of times in a large number of trials that one would expect to obtain a distribution which gave a certain chi-square value even though the universe from which those trials were obtained was governed by the null hypothesis. Note that as the chi-square value gets bigger, going from left to right in the table, the chance of finding such a distribution in a null-hypothesis universe gets smaller and smaller. The specific figure for these 'chances' are what we call the p-values, the significance statistic.

In our example, we are effectively asking what is the chance of getting the figures in Table 2 by sampling from a universe in which class and *que* deletion had no systematic relationship. This will be the significance, or p-value, of Table 2. We computed the chi-square figure of 26.65 for the table. What is the corresponding value of *p*?

To answer this, we must consider which row of the table to search in. These rows correspond to varying 'degrees of freedom' in the data. What are these? There are several ways of thinking about this concept. Basically it is a measure of the complexity of the data set or the analysis. In this case we had a simple two-by-two contingency table: there were only two possible realizations of *que*, and only two social classes. But what would we do if there were four social classes investigated, and we were studying a variable that had three realizations, like the English relativizers with their alternation among *wh*-words, *that*, and zero? In such a case the data would be more complex, and the table would have more cells, making for a bigger chi-square. The degrees of freedom are a measure of this. For contingency tables, it is computed by the formula (number of rows – 1) X (number of columns – 1). In a two-by-two table, this means 1X1=1 degree of freedom, but for four classes and three relativizers, it would be (4-1)X(3-1)=3X2=6.

Accordingly, for the Montreal *que* data, we look in the first row of the table. The chi-square calculated is greater than the rightmost value given in the first row; this means that the corresponding value of *p* is less than .001. In other words, if we drew data randomly from a universe in which the null hypothesis is true, such a chi-square value would be obtained less than one time in a thousand. Hence, we would report this as a significant result. Since there is an extremely small chance that the null hypothesis is true of class and *que*-deletion in Montreal, we conclude in favor of the alternative, namely that class IS in fact associated with the rate of use of this linguistic variable.

Two points should be noted about this procedure. First, if the chi-square value obtained falls within the range found in the table, we search for the value in the table that is closest to but still smaller than the one we obtained. Thus if we had calculated a figure of 5.5 for some table with one degree of freedom, this value would fall between the fourth and fifth columns of row 1: it is greater than 3.84, but less than 6.64. Therefore the closest smaller value is 3.84, and hence we would report the significance – the p value – as p<.05 (recall that as we proceed from left-to-right across the table, the chi-square values increase but the p values decrease). Second, it should be emphasized that the chi-square test does not tell us anything about the DIRECTION of a significant association. In this case, the direction is that the professional class speakers retain *que* more often and the working class speakers delete it. But if the direction of association had been the opposite, and the figures in Table 2 were exactly the reverse (so that the professional class speakers had used 30 absent and 90 present tokens, and the working class had the 3 absent and 130 present), the chi-square statistic would be unchanged. Hence, with chi-square, the statistic tells us only if the association between the variables is significant, while the nature of the association must be determined by inspection of the original values.

Now, given that the p-value falls on a continuum, at what point do we conclude that some finding is significant? The normal practice in statistical work is to set some 'criterial value' for significance, such as .05 or .01, which means rejecting the null hypothesis when it has less than one chance in twenty (p<.05) or one chance in a hundred (p<.01) of being true. In social science research, p<.05 is the most commonly accepted criterial value (and this is the default criterion for significance in Varbrul programs like Goldvarb and MacVarb). Why choose .05 as our criterial value? A five percent chance of the null hypothesis being true is a pretty small chance. Why not say a significant result was found when p<.10 or .25? Why not conclude in favor of the experimental hypothesis any time p<.50 – which gives a less than 50-50 chance of the validity of the null hypothesis? The answer is that .05 is merely a convention – a fairly widespread convention in social sciences and other fields, but a convention nonetheless. The choice of a criterion really depends on what one wishes to do with the information. Anything better than a 50-50 chance might be enough to win money gambling, in the long run, but the only cost of being wrong is some money. But what if the stakes were higher? Imagine a case where clinical trials of a new drug reveal that the death rate in the population that took the drug, although higher than in the control population, is associated with a p=.60. In other words, the difference between the two populations has a 60% chance of being random, leaving 'only' a 40% chance that taking it will raise your risk of dying. Would you take such a medication?

The consequences of drawing erroneous conclusions in linguistic research are unlikely to be either fatal or unprofitable, but the conclusions of a study are likely to enter the body of knowledge about a subject and inform future hypotheses, theories, and conclusions. Consequently, we would like to be fairly confident of their accuracy, and conservative about the conditions under which we reject the null hypothesis. The .05 figure is therefore a reasonable value to adopt as our cutoff for significance. But it is not a magic number. When doing a study of something that for other reasons we strongly believe to be significant, if we get a result that is near but not beyond .05, like, say, .08 or .10, we might not abandon our interest in the phenomenon. Instead, we might be better advised to investigate it further or collect more data. On the other hand, if something appears in one study to be significant at the .04 level, say, and in several other studies to be highly insignificant, giving p values like .40 and .60, it is good to remember that the .04 result will be expected once in 25 trials, even when the null hypothesis is true! This is particularly relevant in a study that does multiple significance tests. If in the course of writing your dissertation, you have performed 30 or 40 chi-square tests using the .05 criterion, it is likely that you have one or two 'false positives' among the results.

3.2. T-test. The chi-square test is defined for contingency tables: in other words, co-occurrence relations among nominal variables. For interval or continuous variables, other types of tests of significance must be used. A full treatment of these may be found in statistical textbooks; for our purposes here I will illustrate the issue with one common and useful approach, called the t-test. This test is an inferential statistic that is used to compare the means of two sets of quantitative variables (normally interval or continuous variables; for ordinals other tests are appropriate). Thus if we wanted to compare two contexts for their effect on the formant values of some vowel – say, the effect of following nasal vs. non-nasal environments on the raising of English /æ/ – then we might measure a number of tokens in the two contexts and compare their mean F1 values using this test. The test returns a p-value, which in this case will mean the probability of the two data sets being drawn by random sampling from the same distribution, or the probability that the independent variable that defines the two different sets has no effect on the measured value.

What measures would influence our estimates of this probability? Before looking at the mathematics, let's consider the problem logically. Clearly, if the two means we are comparing are close together, it is more likely that there is no significant difference between them, while if they are far apart, this should increase the significance of the results. Accordingly, the t-test begins with a calculation of the difference between the means of the two populations. But in addition, whether we consider the means to be 'close together' or 'far apart' will depend in part on how clustered or dispersed the values in each set are. If all the values in one set clustered closely around a mean value of, say, 700 Hz, and all the values in the other set were grouped tightly around a mean of 600 Hz, we would be more inclined to think that the difference between them was significant than if the populations had wide dispersions which overlapped extensively but happened to have different means.

A good measure of how clustered or dispersed the values of a quantitative variable are is a number called the **variance**, which is roughly the average of the squared differences between the mean and the individual values. Specifically, it is calculated thusly: for all tokens of x_i, x_j, etc. in a set, total up the values of $(x_i- x_{mean})^2$, and divide this total by (n-1), where n is the number of tokens in the sample. Thus the variance is large when the numbers in the sample are widely spread out, and small when they are close together, independent of the mean. Another commonly used measure of clustering in a distribution is called the **standard deviation**; this is simply the square root of the variance. Mathematically, we can represent these values by the following expressions. The sum of the squared differences is[3]:

$$\sum (x_i - \overline{x})^2$$

and the variance is

$$\frac{\sum (x_i - \overline{x})^2}{n-1}$$

and the standard deviation is:

$$\sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

To illustrate this, consider two hypothetical sets of measurements of the F1 of some vowel: in one context, the measurements are 400, 500, 600 Hz, and in the other they are 490, 500, 510. Both sets have a mean of 500 Hz, but the first is loosely distributed around this mean while the second is tightly clustered. The variances illustrate this: for the first set it is $(100^2 + 0^2 + 100^{2)}/3\text{-}1 = 10{,}000$, while for the second it is $(10^2 + 0^2 + 10^{2)}/3\text{-}1 = 100$. The associated standard deviations are 100 and 10 respectively; (in this simple case, these are the amounts by which the two extreme values diverge from the middle, mean value).

So, in the variance and/or the standard deviation we have measures that we can use to adjust the difference between the means of the sample populations in order to provide our scale of clustering or dispersion. We do this by dividing the difference between the means by a figure which approximates an average standard deviation adjusted for sample size. The precise formula for this denominator is the square root of the sum of the variance divided by n_1 and the variance divided by n_2; in other words (where s is the standard deviation), the denominator is:

$$\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$$

and the full formula for the test statistic t is:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

One complication that arises is that the two populations ordinarily have two different variances and two different standard deviations, while the test assumes that they have the same standard deviation. Therefore, the values of variance and standard deviation that are observed for the two samples are treated as separate estimates of the assumed common values, and are combined to produce a sort of weighted average variance and standard deviation, as if the two populations were one. The formula for this is the sum of the squared differences for population one plus the sum of the squared differences for population two divided by n_1 plus n_2 minus 2, i.e. (where s.s.d. represents ‘sum of squared deviations’):

$$s = \sqrt{\frac{s.s.d._1 + s.s d._2}{n_1 + n_2 - 2}}$$

This yields the value of s which is to be plugged into the t-test formula above. (Note that if you have at hand the variance or the standard deviation for the populations, you can use these to work backwards to the sum of squared deviations. The s.s.d. for a population with standard deviation s is equal to $(n\text{-}1)s^2$.)

Now let us apply these formulae to an example. Returning to the /æ/ raising example, suppose we had the following sets of measurements of F1 in pre-nasal and pre-oral contexts.

following [-nasal] consonant	following [+nasal] consonant
636, 683, 691, 700, 705, 707, 710, 722, 755	597, 610, 612, 627, 644, 650, 666, 669, 672, 678, 703
sum of squared differences = 8059	sum of squared differences = 10488
n = 9, mean = 701, variance = 1007.4, standard deviation = 31.7	n = 11, mean = 648, variance = 1048.8 standard deviation = 32.4

combined variance and standard deviation for both populations

combined variance: 8059 + 10488 / 9 + 11 – 2 = 1030.4
combined standard deviation: square root of 1030.4 = 32.1

difference in means (numerator of t-test) = 701-648 = 53

denominator of t-test: square root of (1030.4/9 + 1030.4/11)
= square root of (114.5 + 93.7)
= 14.43

t = 53/14.43 = 3.67

Again, we must compare this result with a table of the known distribution of values of the t statistic for random samples drawn from a population in which the null hypothesis is true. A fragment of such a table is given in Appendix 2. Here, the degrees of freedom are calculated as n_1+n_2 – 2, which in our example amounts to 9+11-2=18. Since 18 falls between the rows for 15 and 20 d.f., we look across these two rows, and find that our result is closest to the values in the column for p = .002. It is greater than the t-value of 3.55 corresponding to p=.002 for 20 degrees of freedom, but less than the 3.73 at 15 d.f. The value for 18 d.f. is somewhere in between these two, and therefore will be fairly close to our figure of 3.67. Hence the p-value for our result is close to .002, and certainly less than .01 or even .005. This means that there are approximately two chances in one thousand that such results could come from a universe in which following nasal and non-nasal consonants had identical effects on the raising of /æ/, and that our two sets of measurements are really equivalent. If we were using a .05 or .01 criterion, we would report the difference in means as statistically significant, and conclude that the F1 of /æ/ is indeed influenced by whether the following context is nasal or not.

Note that in this case, the difference in standard deviations (just 0.7) between the two populations was small compared to the difference in means (53), so the step that we took to obtain an averaged standard deviation and variance had little effect on the outcome. Had we merely used one or the other of the unadjusted variances, the t statistic would have been computed to be either 3.64 or 3.72, which would have yielded essentially the same estimate of p. But where the differences in standard deviations and variances between the samples is large, this will not be the case, and the validity of the test might be called into question.

4. Correlations

Of the tests that we have considered so far, the chi-square test looked at relations between two nominal variables, while the t-test is used in the analysis of a quantitative dependent variable and a nominal independent variable. What do we do if we have two quantitative variables? This issue arises frequently in sociolinguistic research. For example, studies of change in progress often wish to investigate the relationship between age and frequency of use of a possibly innovative form, and studies of social stratification may wish to analyze a quantitative measure of some linguistic variable in terms of a quantified, scalar model of social class. The appropriate statistical approach in such cases is to look at a measure of **correlation**. Correlation is a measure of the extent to which the value of one variable covaries with or predicts the value of the other. For example, among children, age and various measures of language acquisition (such as vocabulary size, and mean length of utterance) will tend to increase together: older children know more words, and produce longer sentences. Hence there is a positive correlation, and we might be able to derive a mathematical relationship, so that knowing a child's age would allow us to predict with some degree of accuracy some quantitative measures of acquisition.

Other types of correlation are also possible. One is a negative correlation, where one number goes down as the other goes up. This would be the case, for example, in a linguistic change, if we looked at measures of age and use of some innovation: as age of speaker increases, the use of the innovation decreases. And finally, we may encounter cases where there is no correlation at all: one number does not predict or imply anything about the value of the other. This absence of a correlation would obtain between measures like the F2 frequency of a vowel and the lexical frequency of the word containing it: all vowel phonemes, both front and back (with both high and low F2s) occur in both high and low frequency lexical items.

How can we express such relationships statistically? If the two quantitative measures exhibit some degree of positive or negative correlation, then it is in principle possible to derive some mathematical function that relates them: one is a multiple or a fraction or a power of the other, or differs from the other by some constant amount, or is a negative multiple of the other, etc. Many such functions are possible, and looking at a given data set we may not have any *a priori* expectation about what exactly that function is, so the best strategy is usually to start with the simplest relationship, and look for a linear function. If we label our two variables x and y, the general abstract form of a linear relationship is:

1. $y=ax + c$

Now, from our experience of the world, it is clear that variables can show a general tendency to correlate without being very precise or exact. There are usually other variables that affect the precise values of the measures we look at. During childhood and adolescence, for example, age and height are correlated: everybody gets taller as they get older, which is why we talk about "growing **up**". But it is hardly a perfect correlation: neither of these numbers is a perfect predictor of the other. One's height at age 3 should be greater than age 2, but it also depends on genetics, nutrition, health or disease, etc. So if our data set consisted only of speakers' age and height, an equation like "height equals some coefficient times age plus some constant" will not be exact, although it will be roughly correct. Hence what we need is some measure of how 'rough' or 'exact' the hypothesized linear function is for a given data set. The statistic that does this is known as the coefficient of correlation (technically "Pearson's product-moment coefficient of linear correlation"), and is represented as r.

The mathematics for calculating r will not be addressed here; interested readers are referred to statistical texts such as Woods, Fletcher & Hughes 1986. In any case, it is a tedious computation to do by hand; as a practical matter, most users will prefer to have this statistic computed for them by a statistical program package like SPSS, or by the statistical functions in a spreadsheet like Excel. Rather, I will focus here on its properties and interpretation. Basically, r measures how well a linear function captures the relationship between the two variables. It ranges in value between –1 and +1. An r of +1 describes a perfect linear correlation. Given an equation like (1) above, the value of *y* would for all data points be precisely equal to ax + c, and the coefficient *a* would be a positive number, so that both x and y increase together. An r of –1 means the same thing, except that the relationship is inverse, and *a* is a negative number, so that *y* goes down when *x* goes up. An r-value of 0 means there is no discernible linear correlation between the two measures. Both occur over a range of values, but knowing the value of one does not help to predict the value of the other (at least, as a linear function; an r of 0 does not rule out the possibility of certain non-linear mathematical relationships between the variables).

Most of the time, of course, we will be dealing with r-values that are neither –1, 0, nor +1. How do we interpret these? Essentially they are measures of how precise or how weak the correlation is. A number close to ±1 means a strong correlation; r values of .5, or .6, or -.5 or -.6 are usually considered reasonably good correlations, and r values in the neighborhood of ±.7 or ±.8 are very strong correlations. But values close to 0 indicate the absence of a relationship between the variables: .1 or .2 is nothing to get excited about. The mathematical definition of r depends on the variance in the sample as a function of the linear relationship. If the data points were plotted on a graph, and the linear equation defines a straight line drawn through that plot, how close to the line do the data points fall? If they were all on the line, r would be equal to 1 or -1, while if they were all near the line but not necessarily on it, r would have an absolute value close to 1. Overall, r tends towards 0 as more and more of the points are farther and farther from the line. In fact, there is a specific mathematical definition of this relationship: the square of r expresses the percentage of variance in the results of one value that are predicted by the other value. Hence an r of .8 means that 64% ($.8^2 = .64$) of the variance in *y* is predicted by the value of *x*.

As usual, the interpretation of the r-value depends on the amount of data available. The minimum sample size for this statistic is 3, because it effectively involves seeing how close the points are to a straight line, and a perfectly straight line can be drawn to connect any two points, rendering the result vacuous. Even with 3 or 4 or 5 data points, this statistic continues to be fairly unilluminating. But with any sample size, the statistical significance of any r-value can be found by consulting statistical tables of the r distribution (not reproduced here, but available in any standard statistical reference text). As with the other tests discussed above, the result is to provide a value of *p*, which in this case will be an estimate of the likelihood that the sample is drawn from a universe in which there is no correlation between the two numerical variables under consideration. By way of illustration, an r of ±.6 is significant at the .05 level for a sample size of 11 or greater, an r of ±.5 is significant at the .05 level for a sample of 16 or more, and an r of ±.4 is significant at this level when there are more than about 25 data points in the sample.

To take a real example of this statistic, consider Figure 1, from Guy & Boyd 1989. This study examined rates of deletion of –t,d from final consonant clusters in English (where speakers will often say a phrase like *wes' side* with the /t/ deleted from *west*). In one morphological category, namely irregular past tense forms like *kept, left, told*, the speakers in this study showed an inverse correlation between age and rate of deletion. The figure is a scattergram which plots all the speakers' factor weights for deletion in irregular past verbs against their respective ages; the points cluster roughly along a line declining from left to right (i.e. deletion rates decline from younger to older speakers). The line drawn in the diagram is a linear regression line that most closely captures this overall relationship. The coefficient of linear correlation for this data set was reported in that study as r=-.72. Given that there were 34 data points in the sample, this is significant well beyond the .001 level.

<insert Figure 1 about here>

5. Conclusions.

This chapter has provided a necessarily brief introduction to some of the basic points about quantitative analysis. The selection of topics is guided by the author's experience in what from this domain proves most useful, necessary, and illuminating for linguistic research. Clearly this cannot pretend to be a complete treatment of a subject which is an entire discipline in its own right; for this the reader must consult a real statistician, take courses in statistics, or read a statistical textbook. What I hope to have provided here is some useful background material for linguists who wish to conduct quantitative research, as well as a point of departure for readers of this book who will confront more advanced subjects in later chapters. In short, the chapter may be seen as an attempt to write some words about the numbers which are written about words, in the pursuit of a deeper understanding of both of these symbolic systems.

Notes

[1] In fact, zero subject relatives do occur, especially in presentative or existential constructions like *There's a guy (Ø) lives down the street from me*. However, these are much rarer than zero relatives in nonsubject position.

[2] These are tokens occurring in postvocalic position. Cedergren and Sankoff report a higher rate of deletion in post-consonantal position; those tokens are omitted here to simplify the example and eliminate another intervening variable.

[3] The mean value of a population of quantitative values, represented in the text above as x_{mean}, is indicated in these formulae by the conventional mathematical representation: $\overline{x}$

References

De Decker, P. & S. Mackenzie, 2000. "'Slept through the ice': a further look at lax vowel lowering in Canadian English." In G. Easson, ed., *Toronto Working Papers in Linguistics, vol. 18. Special Issue in Social Dialectology*. Toronto: University of Toronto, Dept. of Linguistics.

Guy, G.R. & S. Boyd, 1989. "The development of a morphological class." *Language Variation and Change* 2.1-18.

Hoffman, M. 1998.

Labov, W. 1966. *Social stratification of English in New York City*.

Cedergren, H. & D. Sankoff, 1974. Variable rules: performance as a statistical reflection of competence. *Language* 50.335-355.

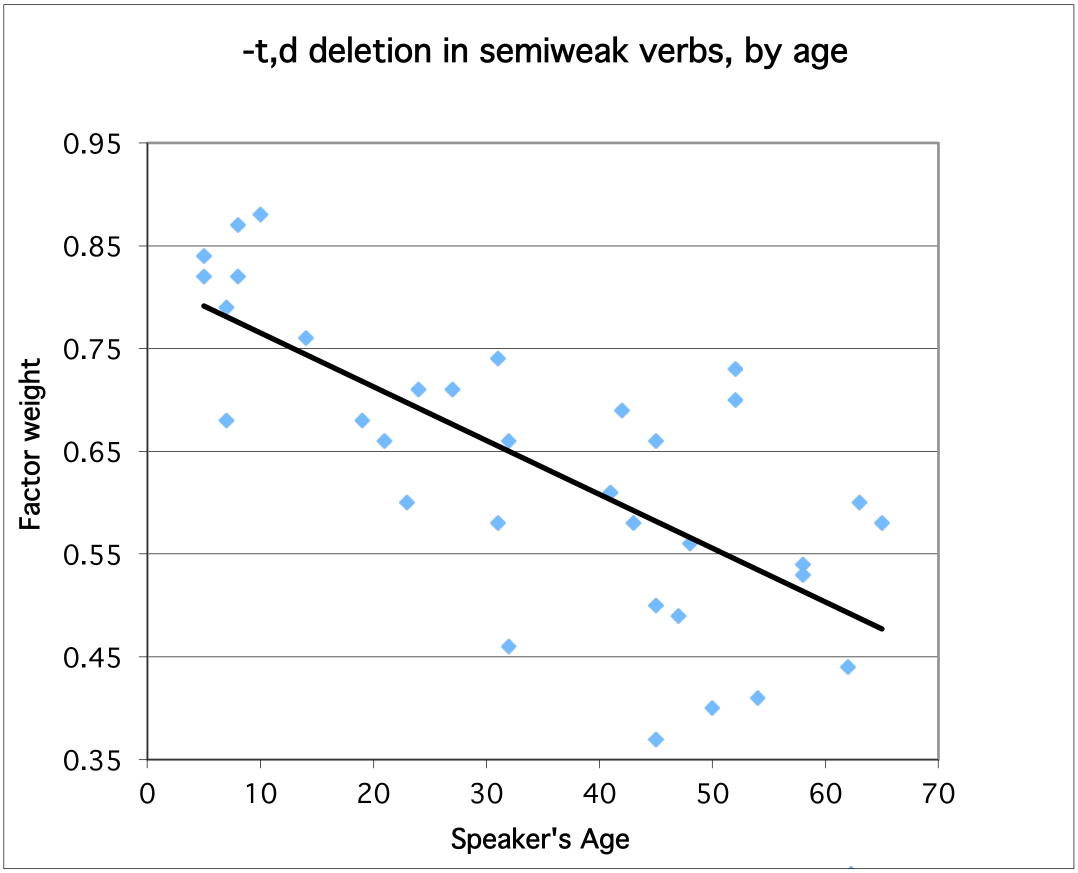
Woods, A., P. Fletcher & A. Hughes, 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.

Appendix 1. Partial chi-square distribution

	p=					
Degrees of freedom	.95	.50	.10	.05	.01	.001
1	.0039	.45	2.71	3.84	6.64	10.8
2	.103	1.39	4.61	5.99	9.21	13.8
5	1.15	4.35	9.24	11.1	15.1	20.5
10	3.94	9.34	16.0	18.3	23.2	29.6

Appendix 2. Partial t-distribution

	p=					
Degrees of freedom	.50	.10	.05	.01	.002	.001
1	1.00	6.31	12.7	63.7	318	637
2	.82	2.92	4.30	9.92	22.3	31.6
5	.73	2.02	2.57	4.03	5.89	6.87
10	.70	1.81	2.23	3.17	4.14	4.59
15	.69	1.75	2.13	2.95	3.73	4.07
20	.69	1.72	2.09	2.85	3.55	3.85


-t,d deletion in semiweak verbs, by age
Factor weight
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0
10
20
30
40
50
60
70
Speaker's Age