14 Descriptive statistics

Daniel Ezra Johnson

1 Introduction

When we have a small amount of data, we can avoid statistics completely. In such cases, we can inspect and discuss each and every *observation* or data point. For example, if we measured the fundamental frequencies (F0) of three siblings' speech, we might observe that Betty's voice was 25 Hz lower than Sue's, but 100 Hz higher than Frank's. It would probably be uninteresting to report a statistic like the average pitch of the family. With a larger dataset, like F0 measurements taken from 1,000 men and 1,000 women, the situation is reversed. It is no longer possible to discuss each data point individually, and while it can still be useful to make graphs that display every observation, we will usually be less interested in individual points and more interested in the patterns or trends formed by groups of points.

This is where *descriptive statistics* come in. Descriptive statistics generally constitute the second step in a quantitative analysis. The first step is to display the data in a tabular or graphical format, using a histogram, bar chart, scatterplot, cross-tabulation, or other method. This will reveal any peculiarities of the data that will shape further analysis. For example, a severely skewed dataset may motivate a transformation, or the use of non-parametric statistics. The second step is the descriptive statistics themselves, which distill the complexities of the data down to a small, manageable set of numbers, abstracting away from details (and noise) in order to describe the basic overall properties of the data. This process can suggest the answers to existing questions or inspire new hypotheses to be tested.

So if we take a single variable like voice pitch, we can talk about its distribution (are all pitches equally common or are there one or more "peaks" at certain frequencies?), its central tendency (what is the most typical pitch for a woman's voice?), its dispersion (how much do men's voices vary in pitch?), as well as higher-order properties like skewness and kurtosis. If we take two variables at once, we can report on their association or correlation (e.g., what is the relationship between voice pitch and the age of the speaker?).

Descriptive statistics describe samples of data, but they do not attempt to answer questions (make inferences) about the larger populations from which the samples are drawn. So if we measured the pitch of twenty English speakers and twenty German speakers, descriptive statistics might tell us that the English sample had an average F0 that was 10 Hz higher than the German sample. If we wanted to know what to make of this result – in particular, whether the difference could be due to mere chance (sampling error) – we could perform a statistical test called a *t-test*. But in doing so, we would be leaving the domain of descriptive statistics and entering the realm of inferential statistics (Chapter 15).

Different types of variables often call for distinct statistical methods; these are discussed in Section 2. Data distributions are covered in Section 3, and the following three sections discuss how to describe distributions: beginning with measures of central tendency or "averages" in Section 4, continuing with measures of dispersion or "spread" in Section 5, and concluding in Section 6 with higher-order descriptive statistics. In Section 7, we discuss how to quantify the extent to which variables relate to one another: association and correlation. Since the chapter will have been concerned primarily with continuous, numeric variables up to this point, Section 8 turns its attention to descriptive statistics for categorical variables. The chapter concludes with Section 9.

2 Types of variables

The most basic descriptive statistic of all refers to the type of variable under consideration. Until we identify the type of variable, we do not know which other statistics are appropriate to apply. Linguistic variables, collected through acoustic analysis, impressionistic judgment, experimental measurement, questionnaire categories, counting within corpora, and more, run the gamut of variable types.

The most fundamental division here is between continuous and categorical variables. *Continuous variables* are numeric measurements that can theoretically take on any value, or at least any value within a certain range. F0 is an example of a continuous variable; in principle it can take on any positive value, even though in practice no one has a mean F0 of 5 Hz or 500 Hz. Formant measurements, reaction times, and lexical frequencies are other examples of continuous variables. For truly continuous variables, no two observations are ever identical. However, we can sometimes treat more granular numeric variables, like frequency counts, ratings on a scale, or values that have been rounded, as if they were continuous. Continuous variables are the input to linear regression (see Chapter 16).

It is sometimes important to distinguish between *interval-scale* and *ratio-scale* continuous variables. Interval-scale variables do not have a natural zero point, so it is meaningless to perform multiplication, division, and certain other mathematical and statistical operations. For example, on the Fahrenheit scale, it is not meaningful to take a ratio of temperatures, and say that 80 degrees is twice as hot as 40 degrees. However, we can compare intervals, and say that an increase of

20 degrees is twice as large as an increase of 10 degrees. On the Kelvin scale, though, where absolute zero is defined meaningfully, not only can we compare intervals, but we can also take ratios. For example, we can indeed say that 400 K is twice as hot as 200 K. Here and throughout this chapter, we will sometimes employ non-linguistic examples in order to make concepts or arguments clearer. Here, we have shown how interval-scale and ratio-scale variables can measure temperature, with the difference lying in the choice of a relatively non-meaningful (Fahrenheit) vs meaningful (Kelvin) zero point. A related issue arises when we use a subject's date of birth as an independent variable. We could use "1900," "1925," "1950," "1975," or "0," "25," "50," "75" for the same four speakers, and while the means will be interconvertible and the standard deviations will not change, the second approach gives more useful coefficients in regression, since we will not be making any predictions about 0 A.D.

Unlike continuous variables, *categorical variables* have values that fall into two or more distinct categories, rather than having a range of intermediate possibilities. If there are more than two categories, we can make a distinction between ordinal and nominal variables. For ordinal variables, the categories have a natural order; the categories of nominal variables have no natural order. Classic examples of ordinal sociolinguistic variables are the contraction and deletion of the African-American English copula (*he is tall*, *he's tall*, *he tall*) and the lenition of coda /s/ in Spanish, first to [h] and then to zero (*los libros*, *loh libroh*, *lo libro*). Examples of nominal variables are the alternation among *that*, *which*, and zero in introducing a relative clause (*the cake that I prefer*, *the cake which I prefer*, *the cake I prefer*), or whether a quotation is introduced with *say*, *go*, *be like*, or some other variant. In these cases, there is no obvious ordering of the possibilities.

If there are only two categories, then we are dealing with a *binary* (or dichotomous) variable. This type of variable is very common in linguistics, in both phonology and syntax. Binary variables can involve the presence vs absence of some element (e.g., the word-final coronal stop in *last chance* or the negative *ne* in French). More generally, binary variables can capture any alternation between two possibilities, as in the (ing) variable (*gone fishing* vs *gone fishin'*), the dative alternation (*he gave John the book* vs *he gave the book to John*), or the particle alternation (*she took out the trash* vs *she took the trash out*). Binary variables are the usual input to logistic regression (Chapter 16).

In this chapter, we will mainly discuss descriptive statistics as applied to continuous variables. We will cover descriptive statistics for categorical variables, including binary variables, in Section 8.

3 Distributions

When we have a variable, especially a continuous one, one of the first things we should do is examine its *distribution*. The temptation is to skip ahead to summary statistics like the mean and standard deviation. These do describe the distribution in an overall way, but as always, a picture is worth a handful of numbers. A distribution refers to the frequency of the values of a variable. It asks how often the variable took on particular values as opposed to others.

```
2 | 2599
3 | 000022234444444444555555555566666666777788888888889999999
4 | 00000000000011111122222222222222222333333333444444444444555555555555556666667777778888999
5 | 00000111222222333444455555556667788888888999999
6 | 000001112222222333333334444555555555556666666677778999
7 | 00000111111222222222222223333333333333344444444444445555555555555566666666666677777777888888888 9999999
8 | 000000011111222222223333334444445677
```

Figure 14.1. *Stem-and-leaf plot of daily temperatures for Albuquerque in 2010*

This question applies to linguistic variables of whatever sort. Sometimes, the distribution is expected (or hoped) to fit a particular shape called *normal* (see below), enabling the use of more powerful parametric statistics instead of having to rely on less powerful but equally useful non-parametric statistics.

Suppose our variable is the average daily temperature in Albuquerque in 2010 (ADTA 2011). Naturally, the data consist of 365 measurements. We can display it in raw form as follows: 30, 35, 36, 33, 34, . . ., 39, 40, 35, 37, 22 (this only shows the first five and the last five days of the year). This format is not very useful. If we were interested in 2010 for its own sake, we might want to make a plot of temperature against time, showing how the temperature changed over the course of the year (very roughly speaking, it went up and then down!). This would be one version of a *bivariate* (two-variable) distribution. But if we are more interested in how 2010 measures up against other years, then we want to describe the *univariate* distribution of the 2010 data. For example, we might want to know how many days were below 30 degrees. (Four.) And how many days were above 90 degrees. (None.)

The *stem-and-leaf plot*, popularized by Tukey (1977), is one way of showing a univariate distribution. For the 2010 Albuquerque temperatures, if we divide the data into 10-degree ranges, we obtain the stem-and-leaf plot in Figure 14.1.

Each temperature is split up into a "stem" and a "leaf" – for example, 29 is split into 2 (shown on the left) and 9 (shown on the right). The plot shows that there were 4 days in the 20s (22, 25, 29, 29), and that there were more days in the 40s and 70s than in the 50s and 60s, and so on. Once you know how to read it, a stem-and-leaf plot is more immediately revealing than a conventional table of frequencies, such as Table 14.1. The table shows absolute frequency (number of days in each temperature range) and relative frequency, the latter expressed as a percentage (number of days in each range divided by the total number of days, 365, multiplied by 100). Annual temperature data have a fixed denominator of 365 (or 366), but if we were going to compare distributions with different N (the total number in a distribution is usually called N) then the relative frequency is much more useful.

Table 14.1 *Frequency table of daily temperatures for Albuquerque in 2010*

Temperature range	Absolute freq. (days)	Relative freq. (100 * days/365)
20–29	4	1%
30–39	54	15%
40–49	82	22%
50–59	45	12%
60–69	50	14%
70–79	96	26%
80–89	34	9%

The most common way to display a univariate distribution of a continuous variable is neither a stem-and-leaf plot nor a frequency table. It is the type of graph that Pearson (1895) called a *histogram.* A histogram is a kind of bar chart (sometimes called a column chart, since the bars are vertical), with the value of the variable shown on the x-axis and its frequency shown on the y-axis. We must break the continuous x-axis into categories called bins, as we have already been doing. The bins can be of any width, although the histogram gives less useful information if they are too wide or too narrow. Figure 14.2 is a histogram of the Albuquerque temperature data. Note that the histogram is essentially an upright stem-and-leaf plot, minus the detailed information about the exact temperatures. The height of each bar is equal to the number of days where the average temperature fell into that bin. We see that the distribution has peaks in the 40s and 70s, as noted earlier. Distributions with two peaks are called *bimodal* (a frequency peak is called a *mode*; see Section 3). We also see that there are no *outliers*, that is to say, no days where the temperature was noticeably higher or lower than any other day. This distribution is not noticeably *skewed* to the left or to the right. If there had been a few days with temperatures in the 10s, a few in the 0s, and 1 or 2 days below zero, that would be a left-skewed distribution: a distribution with a long left tail. Similarly, if there were a long right tail, that would be called a right-skewed distribution (see Section 6).

In reporting linguistic research, distributional plots should be used more often than they are. They can be used in two main ways: at the outset of analysis, to reveal the shape of the data (and at the same time, revealing what simplifications or distortions are involved with taking means, standard deviations, etc.); or applied at the end, to the *residuals* (or *error terms*) of a linguistic model, to verify that the variation not accounted for by the model is not strongly correlated with any of the variables in the model, which would indicate a lack of fit of the model.

In Section 1, we discussed vocal pitch of men and women in a hypothetical way. A real dataset with F0 information is the classic Peterson and Barney (1952) study of American English vowels. Peterson and Barney recorded thirty-three men, twenty-eight women, and fifteen children reading a set of ten words, twice each. The words contained a range of vowels, all in the same consonantal environment: *heed*, *hid*, *head*, and so on. Taking the mean F0 for each adult speaker leads to the histogram of Figure 14.3, which shows the men in white and the women in grey. As we might expect, the distribution of F0 in Figure 14.3 is strongly bimodal, with a peak around 125 Hz representing the most typical men and one around 205 Hz representing the most typical women. It also looks like both men and women, especially women, have right-skewed distributions (with longer right tails).

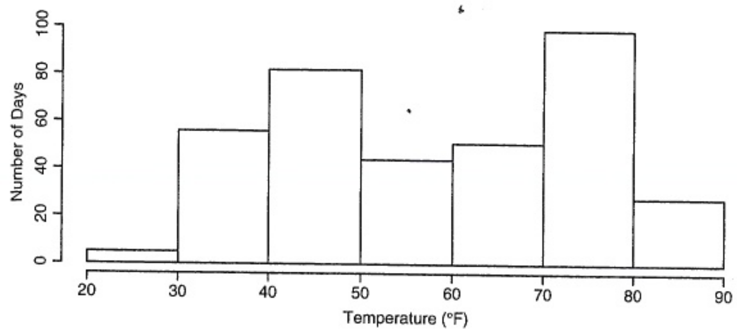

Figure 14.2. *Histogram of 2010 Albuquerque temperatures*

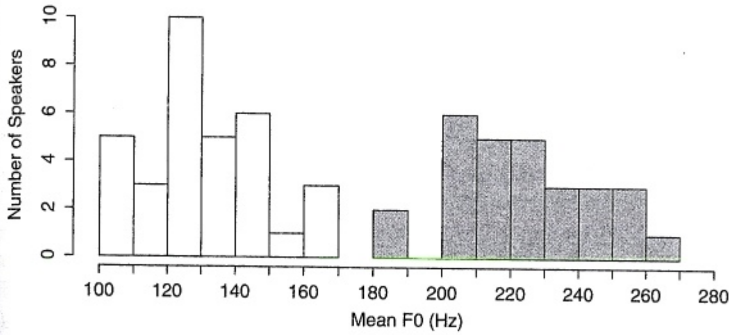

Figure 14.3. *Histogram of men's and women's mean F0. Johnson, based on Peterson and Barney 1952*

We can reduce the skew of this data by performing a logarithmic *transformation* (usually using the natural logarithm, but it does not matter). This makes a great deal of sense for F0 data, because pitch is perceived logarithmically: doubling the frequency makes the pitch go up one octave; quadrupling it makes it go up two octaves. It is therefore natural to log-transform F0 (and arguably higher formant frequencies as well). We see the result of this transformation in Figure 14.4, where the male and female distributions are still somewhat right-skewed, but less so.

Besides its natural applicability to pitch data, the log transformation is often employed to change the distribution of other skewed datasets so that they are closer to a *normal distribution*. Normal (or Gaussian) distributions are a particular family of bell-shaped curves, as illustrated in Figure 14.5. They are defined by two

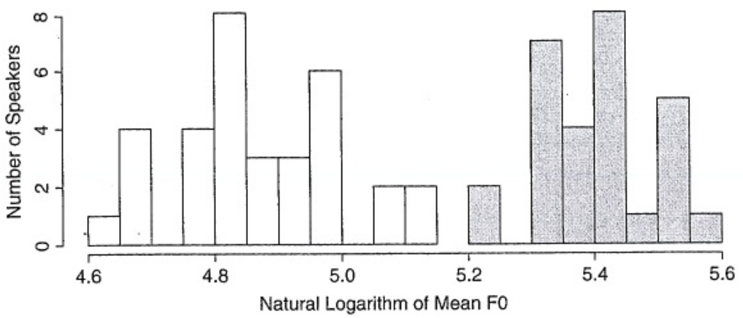

Figure 14.4. *Histogram of men's and women's natural-log-transformed F0. Johnson, based on Peterson and Barney 1952*

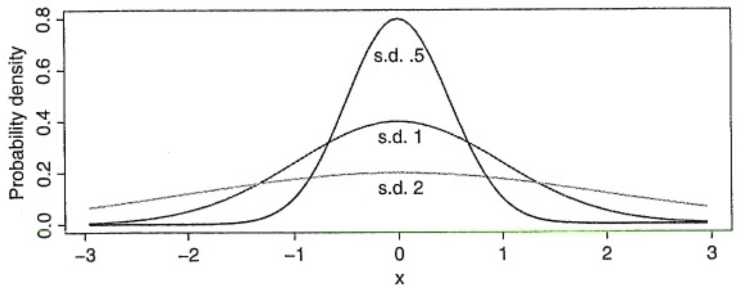

Figure 14.5. *Three normal distributions: mean = 0, standard deviations = {0.5, 1, 2}*

parameters, the mean and the standard deviation (see Sections 3 and 4 below). Figure 14.5 shows the standard normal distribution (standard deviation 1) as well as a narrower normal distribution (standard deviation 0.5) and a wider one (standard deviation 2). The y-axis is probability density; the total area under each curve is 1. A property of normal distributions is that 95 percent of the values fall between −1.96 and +1.96 standard deviations from the mean, regardless of what the standard deviation is. Continuous variables often follow normal distributions quite naturally, because a large number of factors cause them to vary, and the sum of a large number of random variables always follows a normal distribution (this is called the Central Limit Theorem). Other continuous variables – reaction time measurements being one example – are usually log-transformed to make them more normal.

Real data will never be precisely normal, and besides inspecting the data with a histogram, there are several other ways to estimate how close to normal a dataset is, including other graphical methods like the quantile-quantile (or Q-Q) plot, and formal tests like the Shapiro-Wilk test, as discussed in Chapter 15. *Parametric* statistics, which require that data be distributed normally or according to some other probability density function (see Chapter 15), make assumptions about the distribution of the data. However, data do not have to be precisely normal in order to perform most statistical analyses. Methods are called *robust* to the extent they can tolerate deviations from assumptions like normality. *Non-parametric* methods are a class of robust statistics that make no assumptions about data distribution, so they can be used with highly skewed data. Non-parametric methods are also often the most appropriate choice for analyzing ordinal and nominal data.

4 Central tendency

If we needed to describe a variable and could only use a single number, we would surely report a measure of *central tendency*. The central tendency is a "best estimate" of the value of the variable; different definitions of "best" result in different measures, such as the mean, median, and mode. It is almost always essential to calculate central tendency, as it is the principal number that gets reported for a distribution, or compared between groups.

By far the most commonly used measure of central tendency with continuous variables is the *arithmetic mean*, or simply the *mean*. The arithmetic mean is the sum of all the values of the variable, divided by N, the number of observations. The mean is informally called the average, but this term can be ambiguous and should be avoided. When it is appropriate, the calculation of a mean (and the comparison of means) is the powerhouse of descriptive and parametric statistics. There is also a *geometric mean* – the Nth root of the product of all the values – best used when (a) the quantities being compared are on different scales, or (b) when a logarithmic/exponential relationship exists. For example, the geometric mean of 1, 10, and 100 is 10, which depending on the details of the situation may be a more sensible mid-point than the arithmetic mean of 37. A third type of mean, the *harmonic mean* – the reciprocal of the arithmetic mean of the reciprocals of the values – is often used when the quantities are ratios or rates. So if one travels from point A to point B at 50 miles per hour, and returns at 100 miles per hour, the average speed (total distance / total time) is the harmonic mean of 50 and 100, or 66.6 miles per hour (not the arithmetic mean, 75, or the geometric mean, 70.7). While there are few clear applications of the harmonic mean in linguistic research, note that in the field of pattern recognition, the F1 score is defined as the harmonic mean of precision and recall.

The *median* is defined quite differently. If the values of the variable are placed in order from smallest to largest, the median is the value in the middle. (If N is even, we take the mean of the two middle values.) Outliers – unusually small or large values – will affect the mean, but will have little or no effect on the median, so the median is preferred when large numbers of (valid) outliers exist. Also, if the distribution is very skewed (see Section 6), the mean can be misleading. In the million-word Brown Corpus of English, there are 45,215 word types, which occur between 1 and 69,836 times each. The *mean* word frequency is 22, which would point to words like *refund*, *sphere*, and *Florida* as typical in frequency. But in reality only 10 percent of word types are this frequent or more so. On the other hand, the *median* word frequency is 2, exemplified by rarer words like *kelp*, *starchy*, and *Tchaikovsky*. Some 58 percent of word types are this frequent or more so, showing that the median, not the mean, successfully represents something like the mid-point of word frequency. In the case of an ordinal variable, such as the five-point survey's popular "strongly agree, agree, neither agree nor disagree, disagree, strongly disagree," there is no possibility of calculating a mean response, because we only have information on ordering, not distance, between the categories. Ordinal variables therefore call for medians and median-based statistics, including non-parametric methods.

The third measure of central tendency is the *mode*, the most common value in a distribution. In the Brown Corpus example, the modal frequency for word type would be 1, since more word types have a frequency of 1 (19,130) than any other value. A variable always has a single mean and a single median, but it can have more than one mode, if more than one value is equally frequent. A variable with two modes is bimodal, but as we saw above, the term bimodal can be applied more broadly whenever the frequency distribution has two peaks, even if they are not equally frequent. For a nominal variable, with unordered categories (e.g., noun, verb, adjective, preposition), we cannot establish a mean or a median; the mode is the only central tendency that is defined.

Household income is more tangible than most linguistic variables, and is a classic way to explore the differences between the mean, median, and mode. We will look at household incomes under $200,000 in the United States in 2009 (US Census Bureau 2011a). The histogram in Figure 14.6 reveals a right-skewed distribution of income (with a longer right tail), and the mean, median, and mode are labeled. The mean, $57,990, is equal to the total income of all the households, divided by the number of households. This answers the question, "If all the income were redistributed equally among the households, how much would each household make?" This is an interesting question, but we are usually more interested in reporting the actual income of a typical household. We can do this with the median or the mode. The median, $47,500, would be the midpoint of all the households, if they were sorted by income. In other words, half the households made less than $47,500 and half made more than $47,500 (besides those that made $47,500). This answers the question, "What is the income of the middle household?" The mean is the most commonly used measure of central tendency, but it is often the median that tells us what we are more interested in knowing. The relative position of the mean and median is related to skewness (see Section 5). In a right-skewed distribution, like this one, the mean is usually greater than the median. In a left-skewed distribution the mean is usually less than the median. The mode is the income bin with the most households in it; this is $22,500. The mode answers the question, "If we choose a household at random, what is its income most likely to be?" More households made $22,500 than any other amount. Despite the appeal of the mode, it is rarely reported as a measure of central tendency (and the mode is not necessarily a central value, just the most common value). For household income, it is most common to report the median.

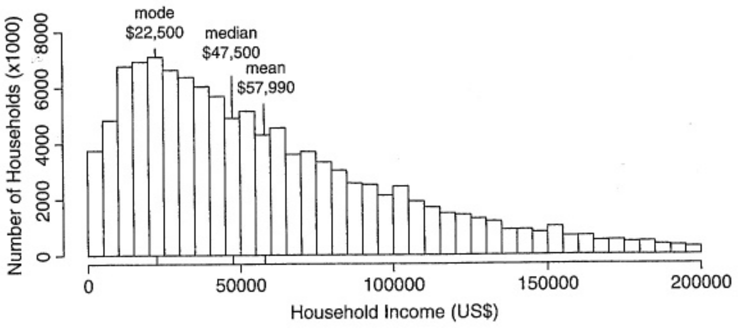

Figure 14.6. *Histogram of 2009 household income, with central tendencies labeled*

In linguistics, a common right-tailed distribution is the *Zipf's Law* relationship, where, in a corpus for example, token frequency is inversely proportional to type rank: the most common word occurs twice as often as the second-most-common word, and so on. These distributions follow a *power law* function of the general form y = 1/x, where the mean, median, and mode are far apart, a distribution much more skewed than any set of acoustic or articulatory measurements are likely to be. As a general rule, we expect repeated measurements to approximate an unskewed *normal distribution*, where the mean, median, and mode are quite close together.

Returning to the Albuquerque temperature data, the mean temperature is 58.2 degrees (we can imagine dividing all the degrees equally among all the days). The median temperature is 58.9 degrees (182 days were colder, 182 were warmer). And there are two modes: 5 days were 44.8 degrees and 5 days were 74.6 degrees.

For the Peterson and Barney pitch data, the mean of the speaker F0 values (each of which is itself a mean of 20 individual observations) is 173 Hz overall, 131 Hz for men and 223 Hz for women. The median values are 163 Hz overall, 126 Hz for men and 223 Hz for women. The generally higher values for the means reflect the right-skewed distribution of the untransformed F0 data. The male data had two modes, as three men had F0s of 122 Hz and three more were at 126 Hz. The female data had four modes, with two women each at 201, 207, 231, and 252 Hz. Recall that for continuous variables, no two values are underlyingly identical, so the result for the mode will always depend somewhat on how the values are binned (the F0 measurements were rounded to the nearest Hz, the temperatures to the closest 1/10 of a degree; the household incomes were placed in $5,000-wide bins).

The median (like the mode) is relatively immune to the presence of outliers and other extreme values, while the mean is more affected by them. A few unusually high values will pull the mean up noticeably, and a few extremely low values will pull it down. Since such outliers may represent measurement errors or other "bad data," we may prefer to use the median, or a more robust version of the mean such as the truncated or Winsorized mean (see Erceg-Hurn and Mirosevich 2008).

Above, we have graphically displayed the distribution of variables by using histograms. When comparing two or more distributions, the box plot (or box-and-whiskers plot; Tukey 1977) is especially useful. See Chapter 15 for more details.

5 Dispersion

A measure of central tendency describes the average, middle, or most typical value of a variable. A measure of *dispersion* tells us how much the values vary on either side of the central tendency. For example, a variable where all the values are clustered near the mean would exhibit low dispersion, while a widely ranging variable would show high dispersion. Dispersion is an essential part of the description of any variable's distribution. Furthermore, a given difference in central tendency means more in the context of low dispersion than high dispersion. For example, words that are twice as long as the mean might be fairly common in English and even more so in German – but people with twice the average number of toes are an extreme rarity.

A common application of dispersion in sociophonetics is to help determine if two vowel clouds represent merged or distinct categories. One can carry out separate *t*-tests for each formant, or calculate the position of each data point along a single (diagonal) axis and perform one *t*-test, or use more complex methods (e.g., *Hotelling's T-squared*, *Pillai's trace*). In all cases, the greater the dispersion, the greater a difference in mean position is required to support the hypothesis of distinct categories. Another use of dispersion is in normalizing vowel formants across speakers (e.g., the Lobanov method). Speakers differ in their mean formant frequencies, but also in their dispersion, so both must be equalized.

For a continuous variable, the easiest dispersion statistic to calculate is the *range*, which is simply the maximum value minus the minimum value. This measurement is obviously very sensitive to outlying values. When there are no real outliers, it can be useful. Our daily temperatures in Albuquerque stretched from 22 to 87, so the range is 65 degrees. We usually report the range alongside the median, which is 59 degrees (rounded to the nearest degree).

The concept of *quantiles* helps us to define a more robust and more frequently used measure of dispersion called the *interquartile range*. Quantiles are the dividing points obtained when you divide the data values into equally sized subsets or bins. Here, the number of observations is equal across bins, not the width of the bins. For example, *percentiles* result from dividing the data into 100 equal bins. The 50th percentile is the same as the median. The 25th, 50th, and 75th percentiles are otherwise known as the 1st, 2nd, and 3rd *quartiles* (the break points from dividing the data into four equal bins). The difference between the 1st and 3rd quartiles is the interquartile range (IQR), a good measure of dispersion. The values within the IQR comprise the middle half of the data. The IQR also forms the "box" part of a box-and-whiskers plot (see Chapter 15). The "whiskers" of a standard box plot stretch at most +/– 1.5 IQR out from the ends of the box; any data point further away is considered to be an outlier. For the Albuquerque temperatures (median 59 degrees), the IQR is 31 degrees. For Peterson and Barney's male speakers' F0 (median 126 Hz), the IQR is 22 Hz. For the female speakers' F0 (median 223 Hz), the IQR is 25 Hz.

By far the most commonly used measure of dispersion is the *standard deviation*, a quantity derived from the *variance*. The variance is the sum of the squared distances between each data point and the mean, divided by the number of observations, N. So for the dataset (1, 3, 4, 5, 6, 7, 9), the mean is 5, the distances from the mean are (–4, –2, –1, 0, 1, 2, 4), and the squared distances are (16, 4, 1, 0, 1, 4, 16). N is 7, making the variance $(16 + 4 + 1 + 0 + 1 + 4 + 16) / 7 = 42 / 7 = 6$. (By showing formulas and calculations, this chapter sometimes goes over math that in practice is done by a computer running a statistics package. However, it is useful to understand what is going on inside statistical operations and tests, which can otherwise become "black boxes.") The standard deviation is the square root of the variance, or in this case, $\sqrt{6} = 2.45$. Taking the square root ensures that the units of the standard deviation are the same as the units of the original data. This makes the standard deviation easier to interpret than the variance, which will often be expressed in unnatural units such as square degrees, square dollars, or square Hz.

When the data are a sample drawn from a larger population – like the Peterson and Barney F0 data, but not the Albuquerque temperature data or the US household income data – we must replace N with N – 1 in the variance and standard deviation formulas. The sample variance above would be 42 / 6 = 7, and the sample standard deviation would be $\sqrt{7} = 2.65$. (The reason we use N – 1 instead of N in the divisor, called *Bessel's correction*, is because we would otherwise be underestimating the variance and standard deviation by using the distances of each point from the sample mean instead of the population mean.)

Two distributions can have similar means but very different standard deviations (and vice versa). We recall that the mean of the 2010 Albuquerque temperature distribution was 58.2 degrees. The standard deviation of these 365 temperatures is 16.5 degrees. In San Francisco during the same year, the mean daily temperature was 57.5, almost the same as in Albuquerque. But in San Francisco, the standard deviation was only 6.0 degrees, reflecting the much smaller seasonal temperature variation in that city.

The standard deviation for the F0 of the Peterson and Barney male speakers is 17.0 Hz, and for the female speakers it is 20.5 Hz. We can see that for these data, whether we use IQR (22 vs 25) or standard deviation (17 vs 20.5) as a measure of dispersion, we find the value for the women is slightly higher than for the men. Figure 14.7 illustrates the dispersion of the Peterson and Barney F0 measurements, separated between men and women. For each group, the figure shows a box plot, which identifies the median and the IQR, and a histogram labeled with the mean and +/– 1 and +/– 2 standard deviations from the mean.

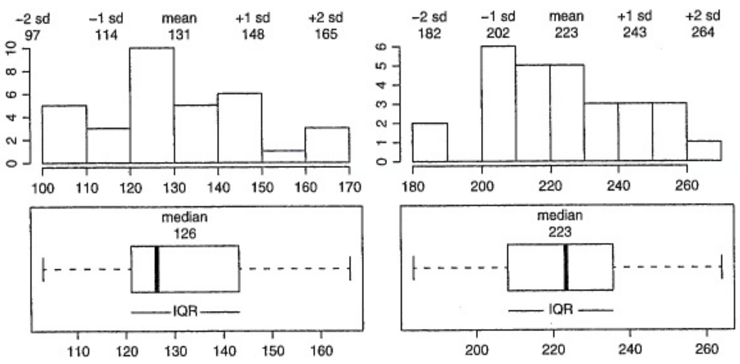

Figure 14.7. *Dispersion of Peterson and Barney F0 for men and women*

In analyzing a continuous variable, we usually choose between reporting the mean and standard deviation, on the one hand, or the median and interquartile range, on the other. If there are significant outliers, or if the data are quite skewed, the median is preferred. Median-based statistics are also preferred if the variable is ordinal. If the variable is nominal, only the mode is well defined.

Although this chapter does not cover tests for statistical significance (see Chapters 15 and 16), such tests make use of the kinds of descriptive statistics discussed thus far. Non-parametric tests, for example, refer to medians (e.g., *Mood's median test*) or ranks (e.g., the *Mann-Whitney test*), while parametric statistical tests (e.g., the *t*-test) employ means and standard deviations. In inferential statistics, much use is made of the fact that 95 percent of the values of any normally distributed dataset will fall between −1.96 and +1.96 standard deviations from the mean.

The measures of dispersion discussed above are all expressed in the same units as the variable itself. There are also *dimensionless* measures of dispersion, which are useful for comparing dissimilar datasets. A parametric example is the *coefficient of variation*, the absolute value of the standard deviation divided by the mean. A non-parametric example is the *quartile coefficient of dispersion*, the IQR (difference between first and third quartiles) divided by the sum of the first and third quartiles. Using these measures, we could demonstrate that the US household incomes are more dispersed than the Albuquerque temperatures.

6 Higher-order descriptive statistics

In this section, we will discuss *skewness* and *kurtosis*. These properties of a distribution are not as basic as central tendency and dispersion, but they are important nonetheless. Two distributions could match exactly in central tendency and dispersion, but be quite different according to these higher-order measures.

We have already referred to the skewness of a distribution in informal terms. Left-skewed distributions have a longer left tail, while right-skewed distributions have a longer right tail. Calculating skewness is a formal way of describing where a distribution lies along this dimension. Recall that the variance is the average squared difference from the mean of a variable's values. To calculate skewness, we take the average *cubed* difference from the mean, and divide this by the cube of the standard deviation. If the distribution has many values well above the mean, when these are cubed it will create large positive terms in the skewness formula. If the distribution has many values well below the mean, there will be large negative terms in the skewness formula. All in all, positive skewness means a distribution is right-skewed, and negative skewness means it is left-skewed.

Unlike the mean and standard deviation, skewness is a dimensionless quantity, without units. Any symmetric distribution has a skewness of zero, because the left and right tails are mirror images of one another. Symmetric distributions include – though of course are not limited to – normal distributions. For this reason, skewness is one measure of non-normality, while the absence of skewness is no guarantee of normality.

Above, we observed that the distribution of American incomes is noticeably skewed to the right, with a long tail of higher values. The calculated skewness for 2009 United States household incomes is 0.99. We can make an interesting contrast between the United States and Canada in this respect, if we compare 2009 personal incomes between $5,000 and $100,000 (Statistics Canada 2011; US Census Bureau 2011b). The means (US: $33,008; Can.: $35,045) and standard deviations (US: $22,027; Can.: $22,519) are quite similar between the two countries. However, the skewness figures are more noticeably different (US: 0.92; Can.: 0.84). This reflects a greater inequality of wealth in the United States, a difference which would show up even more strongly if we included higher incomes.

In Section 2, we observed informally that the Peterson and Barney pitch distributions were skewed to the right for both men and women. We can now quantify this skew: the men's data have skewness of 0.46, the women's have skewness of 0.16. As noted, one way to reduce this skewness is the log transformation, which reduces it to 0.22 for men and –0.02 for women. (The base of the logarithm used does not affect the change in skewness.)

Any distribution following Zipf's Law is inherently skewed to the right. Zipf's Law says that the frequency of a word is inversely proportional to the frequency rank of the word. So, for example, the second most common word should be half as frequent as the most common word, and the third most common word should be one-third as frequent as the most common word. We can see a pattern like this in the Brown Corpus of American English (Francis and Kucera 1964), where the 10th most common word occurs 9,801 times, the 100th most common word occurs 904 times, and the 1,000th most common word occurs 104 times. For the distribution as a whole, the skewness is a whopping 95.6. Log-transforming the frequencies reduces the skewness to 1.45, although the transformed distribution is still one big right tail, certainly far from normal.

Kurtosis measures the extent to which a distribution has a pointy peak (leptokurtic) or a rounded peak (platykurtic). We can graphically assess kurtosis by comparing our variable to a normal distribution with the same standard deviation. (In fact, all normal distributions have the same kurtosis.) The formula for kurtosis is the same as for skewness, except we substitute the fourth power for the cube in both the numerator and denominator. To calculate the *excess kurtosis* (usually just called kurtosis), we subtract 3 to account for the kurtosis of a normal distribution. After this correction, leptokurtic distributions have positive kurtosis, platykurtic ones have negative kurtosis.

Normally distributed data has zero (excess) kurtosis, although the converse is not true: zero kurtosis does not guarantee normality. Like skewness, kurtosis is a dimensionless quantity, making it easy to compare across different variables.

Our temperature distribution is platykurtic, with a rounded "peak" (actually two peaks). Its (excess) kurtosis is −1.37. Our pitch data, with men and women combined, has a similarly wide double peak; its kurtosis is −1.25. On the other hand, our household income distribution has a pointier peak; it is slightly leptokurtic, with a kurtosis of 0.41. Our word frequencies are extremely leptokurtic, having a very sharp peak at 69,836 (representing the word "the"), while most of the values are less than 10. The kurtosis for this dataset is 11,877!

Skewness and kurtosis are underused in the linguistics literature, but it is better to calculate and report them than to compare the shapes of distributions informally. The analysis of vowel formant clouds usually relies on means, with standard deviations employed for difference-of-means testing and normalization, but the acoustic analysis of some consonantal features like fricatives has found spectral skewness and kurtosis to correlate with key perceptual distinctions.

7 Association

The previous sections mostly dealt with one variable at a time. They described various properties of distributions, like central tendency and dispersion. They also compared variables taken from different datasets (e.g., showing that a particular income distribution is more skewed than a particular F0 distribution). This section will compare variables taken from the same dataset. So if we were talking about the physical traits of a certain set of people, we might discuss the relationships among their heights, weights, and eye colors.

In linguistics, a great deal of research involves identifying the associations between variables. For example, in sociolinguistics we might want to know which of a set of social and linguistic variables might affect the phonetics of a sound, the rate of occurrence of a phonological rule, or a choice between morphological or syntactic structures. In experimental research the purpose is very often similar: to establish the existence and strength of the relationship between an independent and a dependent variable. For example, in a lexical decision experiment, we might measure the effect on reaction time between various types of potential primes. The accurate assessment of an association can be complex, especially when there are many other variables to be controlled for, and/or repeated measurements from subjects and from items. One flexible approach is *mixed-effects regression* (see Chapter 16). This section will cover only much simpler statistics.

If knowing the value of variable X does not help you predict the value of variable Y, then the two variables are *independent*. If the values are related in any way, then the variables are *dependent* or *associated*. Associations can take many forms, but to the extent that an association is linear – "if X goes up by a certain amount, then Y goes up or down by a certain amount" – we can measure it with a statistic called the Pearson *correlation*.

If we compared the heights and weights of a large group of people, we would find a strong positive correlation. Knowing someone's height helps you to predict their weight (not precisely, of course, but to a large extent). Taller height goes along with heavier weight, which makes the correlation positive. On the other hand, eye color is independent of both height and weight. Knowing someone's eye color does not help you predict their height or weight.

The relationship of association or independence between two variables is always a two-way street. If height can help us predict weight (association), then weight can help us predict height. And if eye color does not predict height (independence), then height does not predict eye color. Famously, correlation (two-way) does not imply *causation* (usually one-way, if it exists at all).

Figure 14.8 is a plot showing a non-linear association, between the 2010 Albuquerque temperature data (on the y-axis), and the day of the year (on the x-axis). Of course, we know that temperature is highly dependent on the time of year, but we have an up-and-down trend, not a straight-line trend. If we plot the data over 5 years, as in Figure 14.9, we see a cyclical trend. We might try to model this relationship with a sine wave or similar function, but certainly not with a straight line.

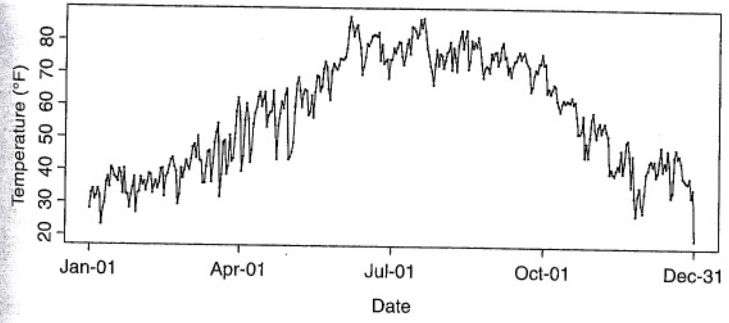

Figure 14.8. *Plot of 2010 Albuquerque temperatures by date*

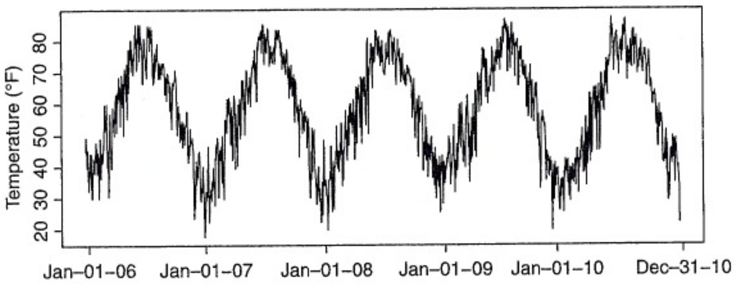

Figure 14.9. *Plot of 2006–2010 Albuquerque temperatures by date*

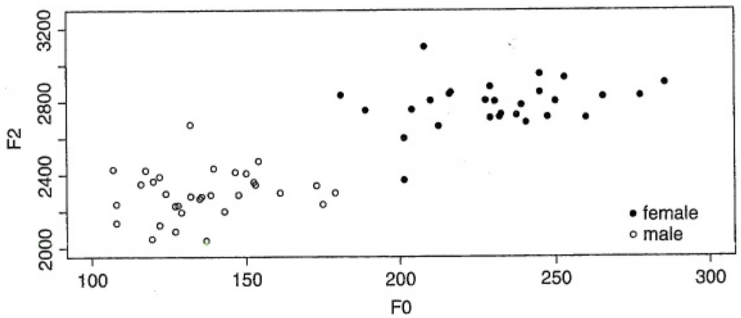

Figure 14.10. *F2 vs. F0 for* heed *in Peterson and Barney data*

As with any statistic, it is important to graph data before attempting to calculate a Pearson correlation. If the relationship between two variables is not basically linear, then the Pearson correlation coefficient can be very misleading. For example, the correlation between date and temperature for the 2006–10 Albuquerque data is only 0.04, even though we know – and can see – that temperature is highly dependent on time. The association is simply not linear.

To illustrate a more appropriate use of the Pearson correlation, suppose we want to know if there is a relationship between the fundamental frequency of a speaker's voice (otherwise known as F0 or pitch) and the higher formant frequencies observed in vowel production. As a quick test using the Peterson and Barney data, we can plot F2 against F0 for the word *heed*, averaging the two observations of the word made for each speaker, and shading the points according to sex, as in Figure 14.10. The figure shows almost no overlap between the men's and women's points. Women clearly have higher F0 and higher F2 than men; therefore the variables are associated. The points lie roughly on a line, so we can go ahead and calculate a Pearson correlation. (The upward-sloping relationship seems to be less strong if we look at the men's or women's data separately; see below.)

Correlations always range between –1 and +1. For the Pearson correlation, –1 means that all the points fall exactly on a downward-sloping line, and +1 means that they all fall exactly on an upward-sloping line. We expect to see a positive correlation between F0 and F2 here, since the points fall close to – but not right on – an upward-sloping line.

The Pearson correlation is defined as the *covariance* of the two variables divided by the product of their standard deviations. To understand covariance, let us consider the first five male speakers. Their F0 values are (173, 148, 108, 153, 134), with a mean of 143. Their F2 values are (2340, 2290, 2240, 2345, 2280), with a mean of 2299. For each speaker, we take the difference between their F0 and the F0 mean and multiply it by the difference between their F2 and the F2 mean. This gives us (30 * 41, 5 * –9, –35 * –59, 10 * 46, –9 * –19) * = (1230, –45, 2065, 460, 171). The covariance is the mean of these products. For these five speakers it is 776, but for the whole dataset it is 12,287 (the unit is squared Hz). The standard deviation for F0 is 52.5 Hz, and for F2 it is 277.4 Hz, making the Pearson correlation coefficient (12,287 / (52.5 * 277.4)) = 0.844. The symbol for the Pearson correlation, a dimensionless quantity with no units, is *r*.

If we square *r* = 0.844, we get *r-squared* = 0.712. The value of r-squared, which always falls between 0 and 1, has a very useful interpretation. It is the proportion of the variance in F2 that is accounted for by F0. That is, knowing F0 decreases our error in predicting F2 by 71 percent. R-squared is most often used this way, to summarize the fit of a *model*: how much of the variance in the dependent variable is accounted for by the independent variable(s). On the other hand, *r* is more often used to measure the correlation between two variables when we are not thinking of one as the predictor and the other as the predicted variable.

A related number is the slope of the *regression line*, the best-fitting straight line drawn through the points (see Chapter 16). The *regression slope* is the correlation multiplied by the standard deviation of the y-axis variable, and divided by the standard deviation of the x-axis variable. Here we have 0.844 * (277.4 / 52.5) = 4.46. This means that F2 increases 4.46 Hz for each 1-Hz increase in F0. Looked at the other way round, the Pearson correlation *r* is a standardized version of the regression slope. It says that F2 increases by 0.844 standard deviations for every 1-standard-deviation increase in F0.

Although there is a high correlation (0.844) between F0 and F2 for the men and women combined, the correlations for men alone (0.160) and for women alone (0.245) are much lower. Although it is a general principle that correlations are smaller when variables are observed over a restricted range, the decrease here is extreme. We conclude that F2 is associated with sex more than it is with F0. This is why we see greater F2 variability between the sex groups and fairly little within them. (A regression analysis, of the sort covered in Chapter 16, would tell us that F0 is no longer a significant predictor of F2 once sex is included in the model.)

As a parametric statistic, the Pearson correlation works best when both variables are roughly normally distributed. The Pearson method is also sensitive to outliers. If our data deviate greatly from normality, and especially if there is a nonlinear relationship between the variables, it is better to use a non-parametric measure of correlation such as *Spearman's rho* or *Kendall's tau.*

Spearman's rho is calculated using the same method as *r* (covariance divided by product of standard deviations), but the data are transformed into ranks first. Ranks just look at the ordering of the numbers, not their values, so (10, 3, 6, 1, 100) and (8, 0, 7, –100, 1,000) would both become (4, 2, 3, 1, 5). Non-parametric methods often involve using ranks, which convert continuous data to an ordinal scale. This makes the methods less powerful – more data are often required to observe an effect – but more robust against outliers and skewed or multimodal distributions. While Pearson's *r* quantifies the linearity of a relationship, Spearman's rho assesses its monotonicity. In a perfectly *monotonic* relationship, as one variable increases, the other consistently increases or decreases (but not both). If both variables consistently move in the same direction, we have rho = 1, and if they consistently move in opposite directions, rho = –1.

For example, suppose that x = (1, 2, 3, 4, 5). If y = x, Pearson's *r* is 1, because the points fall exactly on a straight line. If $y = x^2$ = (1, 4, 9, 16, 25), Pearson's *r* is 0.98; the points are close to a straight line, but not quite. If $y = x^3$ = (1, 8, 27, 64, 125), r = 0.94. If $y = 10^x$ = (1, 10, 100, 1,000, 10,000), r = 0.76. Whenever the data follow a curve rather than a straight line, the Pearson correlation will depart from 1. However, in all four cases, Spearman's rho is still 1, because the relationship is perfectly monotonic; in each case, as x goes up, y always goes up. This relationship is demonstrated in Figure 14.11.

Kendall's tau is another non-parametric correlation coefficient, which has a fairly simple geometric interpretation (Noether 1980). If we make a scatterplot of our variables, pick any two points at random, and join them with a line, then Kendall's tau is the probability that this line will have a positive (upward) slope, minus the probability that it will have a negative (downward) slope. We can see that this quantity will fall in the familiar range between –1 and +1, and that a perfect monotonic relationship between x and y will again result in a coefficient of +/– 1. Kendall's tau tends to be smaller than Spearman's rho, but the two are similar.

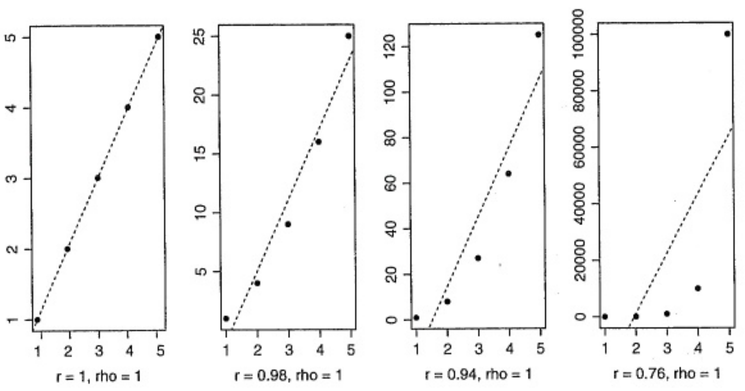

Figure 14.11. *The relationship between Pearson's* r *and Spearman's rho*

Since both Spearman's rho and Kendall's tau disregard the numerical distance between the values in determining a correlation coefficient, both methods are also appropriate for use with ordinal data, where the concept of distance between values does not exist. We will discuss descriptive statistics for ordinal and nominal data in the next section.

8 Descriptive statistics for categorical data

In the sections above, we have discussed descriptive statistics for continuous variables, defined broadly as numeric measurements made to some reasonable level of precision. We may have rounded our temperatures to the nearest degree and recorded our pitch measurements as the closest Hz, but it did not stop us from treating them as continuous variables.

This section will discuss descriptive statistics appropriate for the three main types of categorical variable: ordinal, nominal, and binary. The categories of ordinal variables have a natural order (e.g., a *Likert scale*: strongly disagree, disagree, neither agree nor disagree, agree, strongly agree). The categories of nominal variables have no natural order (e.g., type of tree: elm, ginkgo, maple, oak, pine). Binary variables, with only two categories, can behave in some ways like continuous variables. For example, we can take the mean of a binary variable, but this is not possible for ordinal or nominal variables.

Linguistic investigations often employ categories as independent variables, while the dependent variables are continuous; our analysis of voice pitch by sex was an example of this. It is also common for dependent variables to be categorical. Responses to experimental scales, such as acceptability judgments, identity reports, and ratings of speech samples (guises) along personality dimensions are ordinal variables, though they can sometimes be treated as continuous. Articulatory judgments can be ordinal – front, central, back; raised, canonical, lowered – or nominal, as in rating /r/ as a trill, tap, approximant, or uvular sound. Binary linguistic dependent variables include many morphological and syntactic alternations, and some phonological ones. The VARBRUL/GoldVarb method (a type of logistic regression) was developed for binary alternations (see Chapter 20). The methods given here are simpler ways of describing and quantifying distribution and association.

To assess the distribution of an ordinal or nominal variable, we typically use a *bar chart* (the term "histogram" should be reserved for continuous variables). Figure 14.12 is a bar chart showing the distribution of quotative variants taken from a corpus collected in York (UK) in 2006 (Durham et al. 2012). We see that over 60 percent of the tokens are *be like*, with *say* coming in a distant second place, under 20 percent, and *go* and zero each comprising about 10 percent of the data.

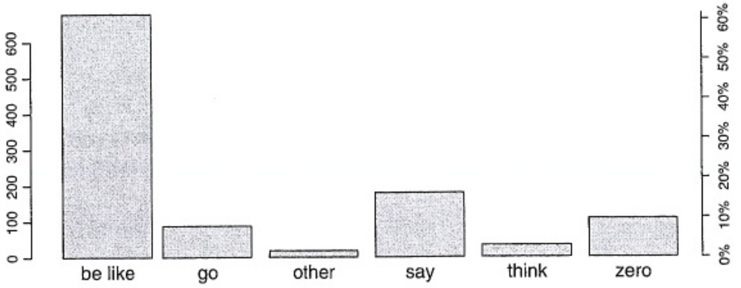

Figure 14.12. *Counts and proportions of quotative variants in 2006 York corpus*

For nominal data like these, the order of the bars is arbitrary (in Figure 14.12, it is alphabetical) and the concept of the distance between bars is undefined. This means that we cannot calculate a mean or a median for a nominal variable with three or more categories. The mode or most frequent value, however, is well-defined: here it is *be like*. The standard deviation is also meaningless in this context. To report the dispersion of a nominal variable, we can use the *index of dispersion*, which is close to 0 if most of the data fall in a single category, and is equal to 1 if the data are equally distributed among all the categories. If N is the total count, k is the number of categories, and f is a vector of the counts for each category, then the index of dispersion = $(k * (N^2 - \text{sum}(f^2))) / (N^2 * (k\text{-}1))$. For the quotative data overall, the index of dispersion is 0.69. For the female speakers, the index is 0.65, while for the males it is 0.76. This is a concise way of saying that the males used a more diverse array of quotative forms, although *be like* is in the majority for both groups (females 65 percent, males 56 percent).

With ordinal variables, a greater range of descriptive statistics can be used. The values of an ordinal variable have a meaningful order, so concepts like "more," "less," "highest," and "lowest" are well defined. This allows us to use the median and some of the measures related to it, like the interquartile range. However, unless a variable has a large number of categories, this is not always very useful. Variables measured on a discrete scale are often best treated as ordinal, although treating them as continuous is a common practice. With some types of scales, an ordinal analysis is necessary because the spacing may not be even (slightly agree, agree, strongly agree). We will now examine data from an experiment where subjects rated sentences on an eleven-point scale.

The first experimental item is the syntactically questionable sentence, "Mary has had more drinks than she should have done so." This was rated by 335 subjects. The bar chart in Figure 14.13 displays the range of rating categories on the x-axis, from 0 to 10. The number of responses in each category is measured on the y-axis. A chart like this is a good way to visualize and begin to interpret the results of acceptability judgment tasks (Chapter 3), as well as responses from questionnaires (Chapter 6) or experiments (Chapter 7). We can see from Figure 14.13 that the distribution skews toward the right, and the mode is the lowest possible rating (0 = completely impossible). A few responses are up toward the high end of the scale (10 = perfectly natural). The range of the variable is 10 – 0 = 10. The median rating – the 50th percentile, or middle value – is 2. As far as dispersion, the 25th percentile is 1, and the 75th percentile is 4, making the interquartile range 4 –1 = 3. The index of dispersion is 0.94.

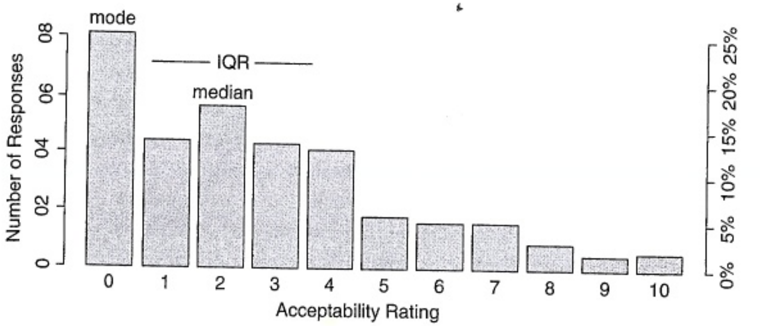

Figure 14.13. *Distribution of 335 ratings for* "Mary has had more drinks than she should have done so" *(0 = completely impossible, 10 = perfectly natural)*

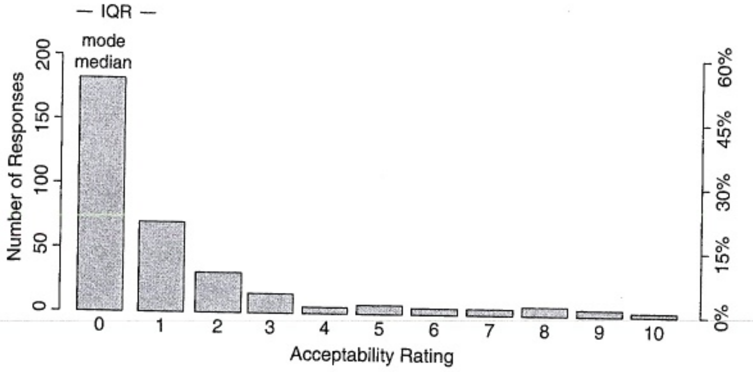

Figure 14.14. *Distribution of 335 ratings for* "Who did John see George and?" *(0 = completely impossible, 10 = perfectly natural)*

Figure 14.14 shows the distribution of a more clearly unacceptable sentence, "Who did John see George and?" This distribution is much more skewed. The total range is still 10, and the mode is still 0, but now the median value is 0 as well. The interquartile range is 1 – 0 = 1, reflecting a less widely dispersed set of scores. Accordingly, the index of dispersion is considerably less: 0.71.

When a variable is binary (also called *dichotomous*), we can report a kind of mean. For example, if the variable is "yes" or "no" votes, we would count each "yes" as 1, each "no" as 0, and calculate an ordinary mean using these numbers. So 30 "yes" votes and 20 "no" votes would be reported as 30 / 50 = 0.60 = 60 percent yes. This measure of central tendency is called the *mean of a proportion*, or p. (We would be unlikely to talk about the median or mode of a proportion.) To measure dispersion for a binary variable, we can take these 1s and 0s and calculate a standard deviation, but the result is not independent from the mean. If the mean of a proportion is p, the standard deviation is $\sqrt{(p * (1 - p))}$. For this reason, the standard deviation of a proportion is not very useful as a statistic.

We now turn to measures of association for categorical data. In discussing association above, we introduced several correlation coefficients for continuous variables. Of these, the Spearman and Kendall coefficients are most appropriate for ordinal variables (or if we have one ordinal and one continuous variable).

Suppose we want to check a possible correlation in the sentence rating task. We want to know if the same subjects who gave high ratings to the Mary sentence were also more lenient in judging the John/George sentence. We find a Kendall tau of 0.27, indicating that there is indeed a small degree of correlation. This value of tau means that if we pick two of the 335 subjects at random, the probability of the pair being *concordant*, minus the probability of the pair being *discordant*, is 0.27. A concordant pair of subjects agreed in their ranking of the two sentences. A discordant pair of subjects disagreed in their ranking.

When there are a lot of ties in the data (i.e., a given pair of subjects gave one or both sentences the same rating), as there are here, it is preferable to use a variant called *Goodman-Kruskal gamma*. The numerator for gamma is the same as for tau: the number of concordant pairs minus the number of discordant pairs. The denominator is smaller: the total number of pairs, not counting ties. So gamma will always be at least as large as tau; here it is 0.35.

If one variable is binary and the other is continuous, we describe association with the *point-biserial correlation coefficient*, r_{pb}, which can be calculated like an ordinary Pearson coefficient. So if we were wondering if there was an association between a subject's sex and their rating of the sentence about Mary drinking too much, there is probably none ($r_{pb} = -0.05$). Note that this example treats the rating as a continuous variable.

If both variables are binary, we report their association with the *phi coefficient*. Again, this can be calculated like a Pearson coefficient – covariance divided by the product of the standard deviations – though the coefficient will fall within a restricted range, not the full –1 to +1 range available for continuous variables.

We can illustrate the phi coefficient with data on the 2,201 people aboard the *Titanic*. Of 1,731 men, only 367 survived (21 percent). Of 470 women, 344 survived (73 percent). There was clearly a very different survival rate for men and women; the question is how to quantify it. Here the phi coefficient comes out as 0.46. The corresponding phi for survival vs age is only 0.10, indicating the lesser importance of age for survival. But these are only bivariate correlations; phi for survival vs sex does not take age (or class) into account. In order to cover all these bases at once, we would use multiple logistic regression (Chapter 16). This method gives a corrected number for the odds of survival for women vs men.

Table 14.2 *Cross-tabulations for survival vs sex and survival vs age on the* Titanic

	SEX (phi = 0.46)		
SURVIVED	female	male	total
yes	344	367	711
no	126	1364	1490
total	470	1731	2201
	AGE (phi = 0.10)		
SURVIVED	adult	child	total
yes	654	57	711
no	1438	52	1490
total	2092	109	2201

Table 14.3 *Cross-tabulation of York quotative variants by grammatical person, observed*

	VARIANT					
PERSON	*be like*	*go*	*say*	*think*	other	total
first	376	25	68	30	9	508
third	302	62	111	3	9	487
total	678	87	179	33	18	995

When one or both of our variables is nominal, we begin to assess their association using a *contingency table*, otherwise known as a *cross-tabulation* or *cross-tab*. Just as we make scatterplots to explore continuous data, a good first step with categorical data is to make a cross-tab. A cross-tab is simply a matrix using the categories of one variable for the columns and the categories of the other variable for the rows. Each cell is filled with the number of observations or cases for that combination of categories. So if one variable had three categories (red, blue, green) and the other had four (triangle, square, circle, star), we would have a 3×4 table, and each of the twelve cells would contain a number representing the quantity of that particular colored shape. Table 14.2 shows cross-tabs for the *Titanic* data discussed above.

We usually want to know if two variables are actually associated, and if they are, the strength of the association. The first question is answered using a significance test; indeed, all of the correlations discussed above have their corresponding significance tests (see Chapter 15.)

The second question can be answered with *Cramer's V*, which ranges from 0 (no association) to 1 (perfect association). Cramer's V is a useful metric that can be applied to nominal data regardless of the shape of the table. If the table is 2×2, Cramer's V equals the absolute value of phi; otherwise we derive it from chi-squared. To understand chi-squared, we return to the York quotative data.

Table 14.4 *Cross-tabulation of York quotative variants by grammatical person, expected (if no association)*

	VARIANT					
PERSON	*be like*	*go*	*say*	*think*	other	total
first	346.2	44.4	91.4	16.8	9.2	508
third	331.8	42.6	87.6	16.2	8.8	487
total	678	87	179	33	18	995

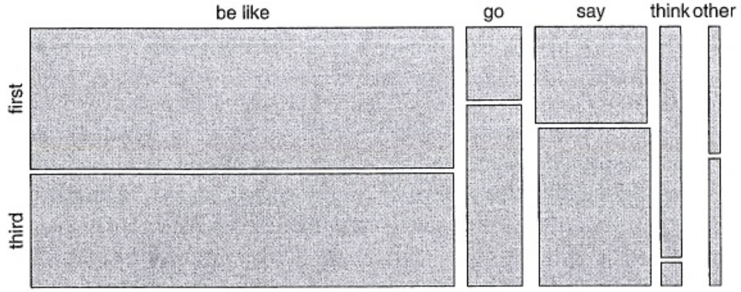

Figure 14.15. *Mosaic plot of York quotative variants by grammatical person*

We can use Table 14.3 – or the corresponding *mosaic plot* in Figure 14.15 – and see that the first-person context has slightly more *be like* and much more *think*, while the third-person context has more *go* and *say*. We suspect an association: knowing the grammatical person of a quotative sentence helps predict the quotative variant.

If there were no association between person and verb, but still the same overall proportions within the person and verb categories (these are called the *marginal frequencies*), the contingency table would look like Table 14.4.

To obtain chi-squared, we subtract each "expected" frequency E (in Table 14.4) from the corresponding "observed" frequency O (in Table 3), square the result, divide by the expected value E, and then take the overall sum, by adding each cell. This formula for chi-squared, which represents how dependent the two variables are, appears in (1). In this case, chi-squared is 55.8.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

While we would certainly be interested in whether quotative use differs significantly between first- and third-person subjects, we leave the details of significance tests like these for Chapter 15.

To obtain Cramer's V – which measures the *strength* of an association (here, between quotative choice and grammatical person) – we divide chi-squared by N (the total count) times k-1 (the number of categories of the variable with fewer categories minus one), and then take the square root, as summarized in (2). In the case of grammatical person and quotative choice, Cramer's $V = \sqrt{(55.8/(995 * (2-1)))} = 0.23$. This is not a very strong association, but it is larger than that between quotative choice and gender, where Cramer's V = 0.12. Gender is less associated with quotative choice than grammatical person is.

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}} \qquad (2)$$

Another approach to organizing categorical data in linguistics is *implicational scaling*; elsewhere, the same concept is often called *Guttman scaling*. It is a procedure employed with a number of binary variables or questions, if they can be placed in a consistent order, where the answer to one implies the answer to others. The original use of implicational scaling in linguistics was by De Camp (1971) in a study of Jamaican Creole. De Camp showed that if a speaker used, for example, the form *nyam* for "eat," then they definitely also used *nanny* for "granny" – but not necessarily the other way around. Similarly, using *nanny* implied using *pickni* for "child," but not vice versa.

If a set of linguistic variables is found to form an implicational scale, this means there is a strong type of association between them. The values of the variables do not co-occur freely, which would lead to 2^n combinations for n binary variables; instead, they are constrained by the scaling, allowing as few as n + 1 combinations. Implicational scaling typically scales linguistic variables relative to each other (in the horizontal dimension) as well as speakers relative to each other (in the vertical dimension). With implicational scales, varieties can be compared not only in terms of the ordering of linguistic features and speakers, but also in terms of the overall *scalability*, or goodness of fit, of the scaling model. Implicational scaling has been found to be particularly useful in relation to questions concerning individual variation, as opposed to statistical approaches that aggregate data for individuals into groups. For this reason, implicational scales continue to be used, especially in studies of creoles, bilingualism, and second language acquisition (see Rickford 2002).

9 Conclusion

If a dataset is "exploratory" – gathered based on an idea, but not a specific hypothesis – then descriptive statistics can suggest hypotheses to test. With "confirmatory" data, we will already have one or more hypotheses in mind. In testing them, we want to know how our sample relates to a larger population: inferential statistics (significance tests in Chapter 15; regression in Chapter 16).

When two subgroups of our data (males and females, first person and third person, treatment and control, etc.) differ on some descriptive statistic, we often want to know the probability that the two samples could actually derive from the same underlying population, despite the surface difference. In other words, we see what looks like an effect: a difference between groups. We want to estimate the size of the effect (descriptive statistics), but also decide whether it is a real, replicable, significant effect, or potentially a mere fluke (inferential statistics).

There are two situations when descriptive statistics can be enough, and inferential statistics are unnecessary or even inappropriate. The first, as mentioned above, is when the purpose of a piece of research is purely exploratory, designed to raise questions rather than answering them. The other situation is when the data are not a sample from a larger population. When a candidate wins an election, we do not ask about the statistical significance of the victory margin. Assuming there were no voting improprieties, the candidate with more votes – even one more vote – is the winner. And if we were studying the speech of a small village, with no plan to compare it to any other place – and if we interviewed every person in the village – we would not have to worry about any observed age or gender differences generalizing to a larger group of people. (However, we would still have to worry about analyzing a large enough sample of speech from each person to generalize about that individual's habits.)

Descriptive statistics are especially valuable when datasets are large, when it would be overwhelming to try to visualize or describe the patterns in the raw data. Descriptive statistics are a valuable set of simplifications that allow us to capture the essence of a dataset – and compare it to other datasets – using a few numbers, most of which have a simple derivation and interpretation.

References

ADTA (Average Daily Temperature Archive). 2011. Source data from National Climatic Data Center. Available at: http://academic.udayton.edu/kissock/http/Weather (accessed June 27, 2013).

DeCamp, D. 1971. Implicational scales and sociolinguistic linearity. *Linguistics* 9.73: 30–43.

Durham, M., B. Haddican, E. Zweig, D. E. Johnson, Z. Baker, D. Cockeram, E. Danks, and L. Tyler. 2012. Constant linguistic effects in the diffusion of *be like*. *Journal of English Linguistics* 40.4: 316–37.

Erceg-Hurn, D. M. and V. M. Mirosevich. 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist* 63.7: 591–601.

Francis, W. N. and H. Kucera. 1964. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, RI: Brown University. Available at: www.archive.org/details/BrownCorpus (accessed June 27, 2013).

Noether, G. E. 1980. Why Kendall tau? *Teaching Statistics* 3.2: 41–3. Available at: www.rsscse-edu.org.uk/tsj/bts/noether/text.html (accessed June 27, 2013).

Pearson, K. 1895. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A* 186: 343–414. Available at: http://visualiseur.bnf.fr/CadresFenetre?O=NUMM-55991&I=427&M=tdm (accessed June 27, 2013).

Peterson, G. E. and H. L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175–84.

Rickford, J. R. 2002. Implicational scales. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds. *The Handbook of Language Variation and Change*. Oxford: Blackwell, 142–67.

Statistics Canada. 2011. CANSIM. Table 202–0402. Distribution of total income of individuals, 2011 constant dollars, annual. www5.statcan.gc.ca/cansim/pick-choisir?lang=eng&id=2020402&pattern=2020402&searchTypeByValue=1 (accessed June 27, 2013).

Tukey, J. C. 1977. *Exploratory Data Analysis*. New York: Addison-Wesley.

US Census Bureau. 2011a. Current Population Survey. Annual Social and Economic Supplement. Table HINC-06. Income distribution to $250,000 or more for households: 2009. www.census.gov/hhes/www/cpstables/032010/hhinc/new06_000.htm (accessed June 27, 2013).

US Census Bureau. 2011b. Current Population Survey. Annual Social and Economic Supplement. Table PINC-11. Income distribution to $250,000 or more for males and females: 2009. www.census.gov/hhes/www/cpstables/032010/perinc/new11_000.htm (accessed June 27, 2013).

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Published in the United States of America by Cambridge University Press, New York

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107696358

First published 2013

Printing in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-01433-6 Hardback
ISBN 978-1-107-69635-8 Paperback

Research Methods in Linguistics

EDITED BY

ROBERT J. PODESVA
Stanford University

AND

DEVYANI SHARMA
Queen Mary University of London

CAMBRIDGE
UNIVERSITY PRESS