CHAPTER 9

Statistical Data Analysis

Learning Objectives:

- How do the graphical and mathematical techniques of data analysis help to achieve the four scientific goals of description, prediction, explanation, and control?
- What are descriptive statistics, and what are the main properties of data sets we might want to describe?
- What is a statistical relationship, and what are some of the different types of relationships?
- What are inferential statistics, and what is the basic logic of their two main approaches, estimation and hypothesis testing?
- What are some of the properties of geospatial data that make their analysis particularly rewarding and particularly difficult?

Data analysis is the set of display and mathematical techniques, and attendant logical and conceptual considerations, that allow us to extract meaning from systematically collected measurements of our phenomena of interest and communicate it to others. Data analysis thus helps us achieve the four scientific goals described in Chapter 1: description, prediction, explanation, and control. Data analysis helps to efficiently identify and describe patterns in large amounts of data, patterns that improve our predictions about the unknown, explain the causes of relationships, and allow us to exert influence over phenomena we wish to control.

Although we treat them in separate chapters, in practice the display and mathematical techniques of data analysis are not entirely separate but make up a single set of procedures for extracting and communicating meaning. In this chapter we discuss the techniques of statistical data analysis; we cover the display techniques, such as mapping and graphing, in Chapter 10. We must emphasize at the outset, however, that the present chapter is *not* intended to be a detailed tutorial on how to statistically analyze data. Any number of textbooks provide detailed treatments of

data analysis in geography and other disciplines (we list a few in the Bibliography), and in any case, all geography students need to take courses specifically focused on data analysis, both the statistics-based and display-based versions. Our purpose in this chapter is to provide an introduction and overview that describe the conceptual forest of data analysis rather than its technical trees.

Mathematical data analysis in geography is usually treated as being statistical (probabilistic or stochastic) in nature. As we pointed out in Chapter 2, geographers in most subfields assume their data sets do not reflect constructs and their interrelationships in a simple deterministic manner but in a complex and imperfect way that is influenced by factors difficult to control completely, in all the senses of control discussed in Chapter 7. There are three primary reasons most geographers treat data in a statistical rather than deterministic fashion. First is that data are usually considered to be an incomplete sample of a larger population of data of interest; different samples from the same population thus vary from each other (the sampling error we mentioned in Chapter 8). Second is that measurement in all scientific fields is imperfect, necessarily involving error; thus, the values we have in our data set vary at least a little from the true values they would have if they were perfect reflections of our constructs of interest. The third reason geographers treat analysis statistically is that their phenomena of interest are typically expressions of complex sets of many interacting variables, all of which are not likely to be identified or measured in any single piece of research. A statistical approach allows geographers to interpret the meaning of one or a few variables in this background of complex multivariate reality; it provides an approach to "statistical" control, as we discussed in Chapter 7. In fact, 20th-century developments in our understanding of "chaotic systems" suggest that some unpredictability is inevitable in complex systems and will never be avoided, even in theory.

Statistical Description

As we discussed in Chapter 8, if we were feasibly able to measure the entire target population of cases in which we were interested, we would do so. We would then interpret the resulting data set, looking for patterns of high and low values in the variables we measured, typical and atypical values, values measured on the same cases that seem to vary together in some systematic way, and so on. That is, we would want to describe patterns in our data. How would we do that? We could simply print out an unorganized list of raw data measurements or view them on a computer screen. Surely that would be a difficult way to see patterns, however. We could organize the raw data into a matrix with columns that represent variables and rows that represent cases. But it would still be difficult to see patterns in this matrix, especially if the data set was larger than a few rows and columns.

Instead, we would go further by displaying the data in graphs, tables, maps, and other displays. And we would try to summarize various potentially relevant properties of our data by calculating summary indices designed to reflect these properties in an efficient manner. This is **descriptive statistics** (statistical description). We would describe patterns in our set of population measurements by displaying the data in various ways and by calculating summary indices of the population data called

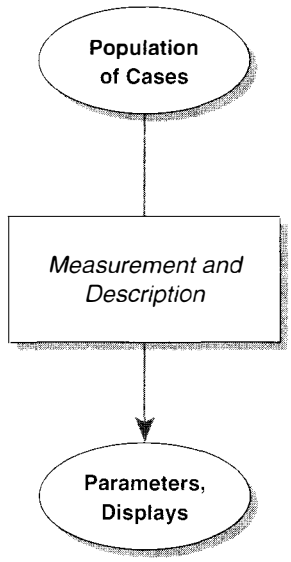

Figure 9.1 Data collection and statistical analysis with populations of cases. If possible, researchers would always take this direct approach to learn about parameters.

parameters (see Figure 9.1). We would interpret the patterns in these displays and calculated parameters by combining what we see in our data with what we know about our cases, our constructs, our measures, the places and times we are studying, and so on. Then we would tell other people what we found and what we interpreted it to mean. Then we would be finished.

What are the properties we might find relevant or interesting to summarize in our population data? The first one that comes to most people's mind is a measure of the average or most representative value in the data, a descriptive property known as the **central tendency**. There are many different ways to identify or calculate a central tendency. They reflect the "central value" in somewhat different ways conceptually, and they are more or less appropriate depending on the measurement level that characterizes your data (see Chapter 2). The three most common are the mode, median, and mean. The **mode** is the most frequently occurring value in your data set. It is a good measure of central tendency for nominal variables, such as soil type or religious affiliation. The **median** is the middle value in your data set; half the scores in the set are higher than the median and half are lower. It is a good measure of central tendency for ordinal variables, such as rivers ranked by their discharge volume or cities ranked by their position in an urban hierarchy. The **mean** is the "average" value in your data set. The standard type is the **arithmetic mean** (there are several others useful in geography), calculated by adding up all the scores and dividing the sum by the number of scores. It is a good measure of central tendency for interval and ratio (metric) variables, such as counts of tree rings or the lengths of straight segments in street networks in different regions. However, metric variables are sometimes strongly skewed—unevenly distributed towards high or low values—because of extreme values in one direction, high or low (we discuss skew, a property

of the form of a distribution, below). When this is so, the median is a better descriptive measure of central tendency.[1] Monetary variables, such as real estate prices in a region, provide a good example. The mean of four houses valued at $82,000, $56,000, $133,000, and $77,000 is $87,000. If a single million-dollar property were added to this set, the five houses would have a mean of over $269,000, which is not very representative of the set of houses. The median of $82,000 is better.

After the central tendency, we would probably want to know how scores in our data differ from the central tendency, a property known as **variability** or dispersion. After all, none of the measures of central tendency tell us anything about how close the individual scores are to the central value or to each other. Variability is interesting not only in its own right, but because the variability of a variable helps in interpreting its central tendency; for example, a measure of central tendency is more representative when the variability is low. Three common measures of variability are the range, variance, and standard deviation. The **range** is a simple index reflecting the distance between the highest and lowest values in the data set. It is a good measure of variability for ordinal or metric variables, but it is limited insofar as it is based on only two scores in the distribution. The **variance** is based on an average of **deviations from the mean**, which are calculated for each individual score by subtracting the arithmetic mean from it. Thus, high scores will have large positive deviations, low scores will have large negative deviations, and scores near the mean will have small deviations. To calculate the variance, the deviations are squared, summed, and divided by the number of scores; that is, the variance is the "mean squared deviation from the mean." Because the deviations are squared (if they were not, the deviations from the mean would always sum to zero), the variance is "blown up" relative to the original distribution of scores, so its square root is often used instead for descriptive purposes.[2] That is the **standard deviation**.

A third property we might want to describe about our data is their overall **form**, essentially the shape of the distribution. Form is easier to understand if you think of data in terms of a graph rather than a list or table. There are a variety of form properties we might want to describe about our data. **Modality** refers not to the value of the most commonly occurring score, which is central tendency, but to the number of "local" modes a distribution has; a local mode is not necessarily the most common in the entire distribution, but it is more common than values just below and above it. As we mentioned above, **skewness** is a property of form that describes "unevenly" distributed scores. A distribution with "positive" skew has mostly low and medium scores with a few extremely high scores that are not balanced by an equal number of extremely low scores; "negative" skew is the opposite pattern

[1] We emphasize "descriptive" here. Especially when calculating various inferential indices, the mean is often favored over the median, even for skewed variables, because of various statistical reasons beyond the scope of this book. Again, check out an advanced statistics course.

[2] Again, we emphasize descriptive. As in footnote 1, the variance is typically preferred over the standard deviation for inferential indices, for statistical reasons beyond the scope of this book.

(see Figure 9.2). Skewness contrasts with **symmetry**, a property of nonskewed distributions whose two sides around the central tendency are mirror-image reflections of each other. One common symmetric distribution form is the unimodal **bell-shaped distribution**. A particular bell-shaped distribution that has a specific proportion of scores within any given range from the central tendency is the important **normal distribution**.[3]

Central tendency, variability, and form are properties of entire data sets. Another approach to describing data is the calculation of **derived scores** that describe properties of individual scores by expressing their value relative to the rest of the data set. A simple one is to express scores in terms of their rank within the data set; for example, the highest score is "1," the next highest is "2," and so on. A more sophisticated version expresses scores in terms of their **percentile** rank. This is the percentage of the data set that is less than the score in question. The highest score is at the 99th percentile (a score cannot be greater than 100% of the scores). The median is thus at the 50th percentile. Sometimes it makes sense to convey raw scores in terms of ratios or proportions, which can be expressed in terms of a decimal fraction from 0 to 1 or a percentage from 0% to 100%. Finally, derived scores are sometimes calculated from metric-level data so that the values of individual scores are expressed by dividing their deviation from the mean by the standard deviation of their data set. This ***z* score** expresses the score in terms of standard deviation units above or below the mean of the data set.

A final property of data we often want to describe concerns pairs of variables or larger sets, rather than single variables. It is the property of **relationship**, which is when cases show systematic (consistent) patterns of high or low values across pairs of variables, each variable measured on a common set of cases. The simplest form of relationship is a **linear relationship**, which comes in one of two types (see Figure 9.3). A **positive** (or direct) **relationship** between two variables means that cases with high values on one variable—high relative to that variable's central tendency, that is—tend to have high values on the other; cases with low values on one variable tend to have low values on the other. A **negative** (or indirect) **relationship** between two variables means that cases with high values on one variable tend to have low values on the other, and vice versa. **Relationship strength** is the strength of these patterns, whichever direction they take. Weak relationships tend only weakly to show the systematic pattern, whereas strong relationships tend strongly to show it. *No relationship* describes when there is no systematic tendency for high or low values on one variable to go with high or low values on the other variable: a case with a high value on one variable is just as likely to have a high value on the other variable as it is to have a medium or low value.

---

[3]The normal distribution is so important because many variables naturally approximate such a distribution; the heights of adult female German Americans and of Jack Pine trees are two examples. But even more important, it has been proven that both random measurement errors and the sampling error for statistics as estimates of parameters (see the section on Statistical Inference in this chapter) are normally distributed. In many situations, however, other theoretical distributions besides the normal are useful to geographers.

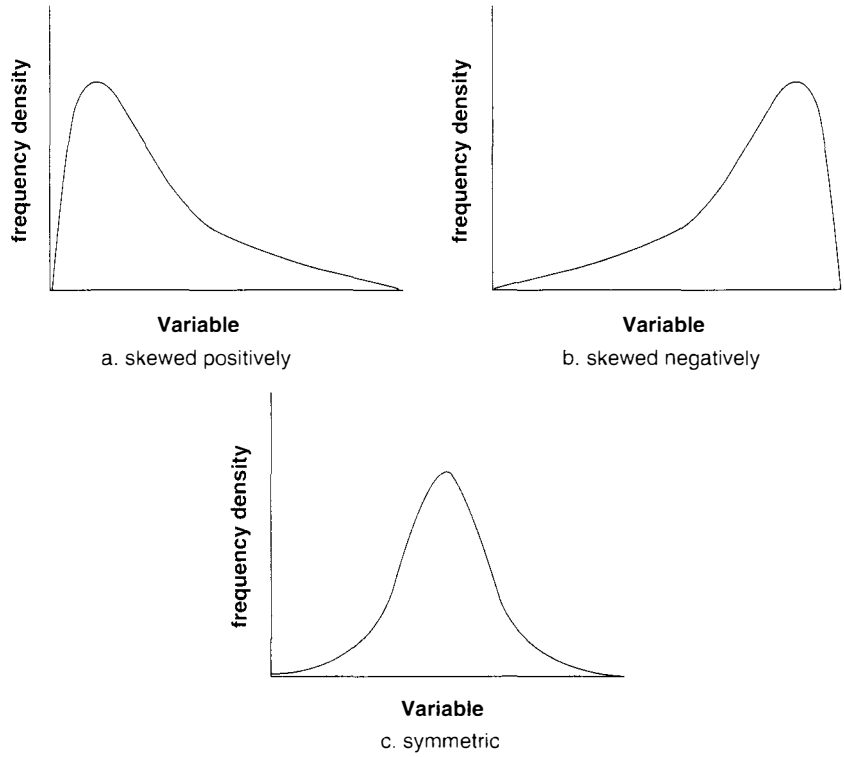

Figure 9.2 Distributions that are (a) skewed positively, (b) skewed negatively, and (c) symmetric. This symmetric distribution is a normal distribution, which is bell-shaped.

Relationship is usually quantified by some type of **correlation coefficient** (strictly speaking, the square of the correlation is usually the index of relationship strength). When dealing with linear relationships, the correlation coefficient is calculated by a formula designed to produce a value of 1.0 when there is a perfect positive relation, –1.0 when there is a perfect negative relation, 0.0 when there is no relationship at all, and some intermediate value when there is a relationship of intermediate strength, which is pretty much always true with actual data. But there are some other statistical indices of relationship. In areas of geographic research where true experiments are conducted (see Chapter 7), relationships are often statistically revealed by comparing the central tendency (usually the mean) of one variable across discrete levels or experimental conditions of another variable. For example, there is a relationship between "Cloud Seeding" and "Rainfall" if the mean rainfall produced by a cloud seeding procedure is higher or lower than that produced when no seeding is carried out (in fact, to the best of our knowledge, the evidence for cloud seeding is rather equivocal).

When researchers are especially interested in the *form* of the relationship in addition to its strength, which is usually the situation, they often apply a statistical technique called **regression analysis** (we discuss the reason it is called "regression"

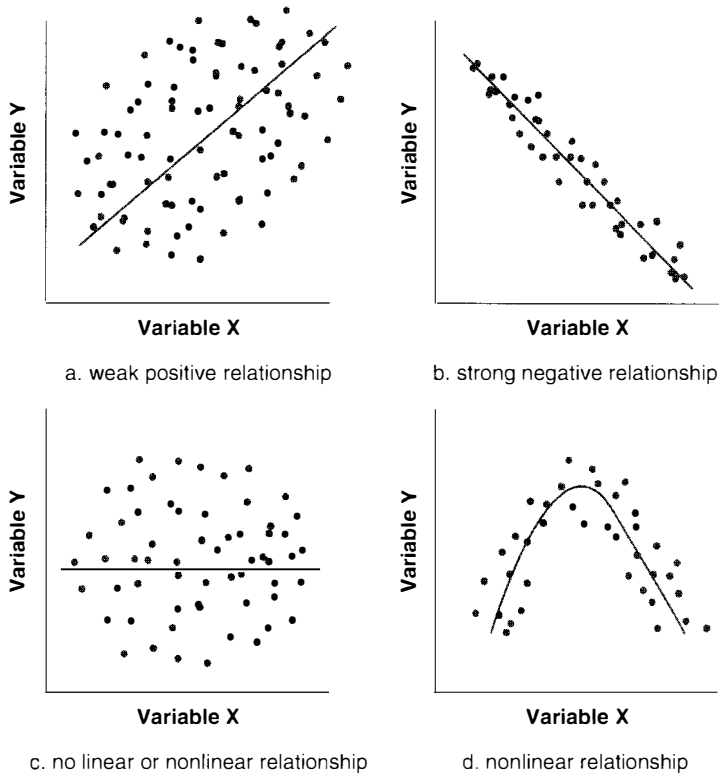

Figure 9.3 Scatterplots and regression lines depicting strong and weak relationships, either positive or negative, between two variables. The absence of a linear relationship may or may not indicate there is no relationship of any form.

in Chapter 11). In regression analysis, a statistical model of relationship form is investigated. This model is a simplified representation of the form of the relationship between two variables, or among larger sets of variables.[4] It expresses relationship as an equation that predicts the values of one or more variables (the **criterion variables**, usually labeled Y) as a function of the values of one or more other variables (the **predictor variables**, usually labeled X). When possible values for the predictor variable are "plugged in to" the regression equation, the resulting predicted values for the criterion variable can be graphed, resulting in a picture of the modeled relationship form. The simplest case is the linear relationship we spoke of above, such as graphs *a* and *b* in Figure 9.3. These relationships are expressed by the equation for a straight line:

$$\hat{Y}_i = a + bX_i$$

[4]Relationships between *pairs* of variables and the statistical techniques for their analysis are called **univariate.** Relationships and techniques for larger sets of variables are called **multivariate.** There are many multivariate techniques used in geographic research, including principal components analysis, multidimensional scaling, and cluster analysis.

This is a simple linear equation resulting from ordinary least squares (OLS) regression. Values of the Y variable are predicted to fall along a straight line created when values of the X variable are multiplied by a weighting constant b and added to another constant a. The i subscripts indicate that the individual scores for each variable are to be entered and output one at a time. As is familiar from high-school math, the weighting constant is the slope of the line, and the additive constant is the place where the line intercepts the Y axis. The "hat" symbol "^" is placed over Y to indicate that the outputs of the equation are predicted values for the criterion, not its actual values determined empirically. In Figure 9.3, the predicted Y values fall along the straight or curved lines, whereas the actual Y values are the data points that fall above or below the lines. The vertical distance of the points from the predicted line is the **error of prediction** of the model, which is assumed to be random from case to case. If the errors of prediction are not random, the model has been inappropriately formulated.

In some areas of research, geographers are not particularly interested in the specific values for the constants of the linear equation, perhaps because the quality or quantity of sampling or measurements is insufficient to justify faith in the specific values. Instead, they may just be interested in whether the relationship does or does not tend to go in only one direction, either up or down, even if it is not exactly a straight line. Such a relationship is called **monotonic**. In contrast, in many other situations, geographers are interested not just in linear relationships but in **nonlinear relationships**. Regression analysis and other statistical techniques for studying relationships can investigate a potentially unlimited variety of such nonlinear relationships that do not follow straight lines (a "quadratic relationship" is shown by graph *d* in Figure 9.3). Alternatively, nonlinear relationships can sometimes be "linearized" by subjecting the variables to a **transformation** that applies some mathematical operation to each of the raw scores; logarithmic, trigonometric, and square-root operations are examples. But these important topics go beyond our scope here; get to those statistics courses!

Statistical Inference

Now that we have calculated parameters to describe our population of data, displayed it with the techniques of Chapter 10, looked at it over and over in different ways, and finally communicated it to others using the techniques of Chapter 13, we are finished—except for one problem. We don't *have* a population of data. We're scientists who want more general truths than our specific data set by itself reveals directly, so we have only a sample. This leads us to the thorny (you may apply your favorite invective here) topic of **inferential statistics** (statistical inference): drawing informed guesses about likely patterns of data in a population on the basis of evidence from samples drawn from that population. It's thorny because it's conceptually quite difficult, as compared to descriptive statistics; we know this not only intuitively but from years of teaching statistics. Even worse, it's thorny because it makes

our interpretation of data and the decisions we make about our research ideas far more uncertain than they would otherwise be. Ultimately, in fact, this need to sample makes interpretation and decision-making in science fundamentally and irrevocably uncertain. Like we said—a problem.

So we are not finished just yet. Given the data set we obtained by sampling from our population, we will again want to describe patterns, by calculating summary indices of the sample data called **statistics**[5] (see Figure 9.4). These are conceptually identical to the population parameters we covered above, but in many cases, the statistics are calculated with slightly different formulas than their corresponding parameters[6] (the sample indices of central tendency are important exceptions; they are calculated in exactly the same way as the parameters, although different symbols are used for them). The typical difference in the formulas for statistics arises from the fact that whereas parameters are considered to reflect properties of populations that are fixed, statistics reflect properties of samples that fluctuate from sample to sample. We learned in Chapter 8 that this fluctuation is called **sampling error**. Perhaps "sampling variability" would be a better term insofar as sampling error is not a *mistake* that one can even potentially avoid if one is sampling. The necessity of sampling error and the resulting fact that we can never know for sure how representative our sample is of our population is the root of the problem we lamented above.

Given our sample statistics, we next use them to infer our parameters. Here we should probably balance our negativity about inferential statistics with a little appreciation for what we can actually do with them. This is where the real beauty and power of the statistical theory developed over the last couple centuries really shines. You see, we don't simply make guesses about parameters from sample statistics. We use statistical theory *to assign probabilities to our guesses*, probabilities that we are right or wrong, or at least probabilities that we are close to being right. That is the heart of inferential statistics, and it is a very big deal. It allows to us to optimize our reasoning under necessarily uncertain circumstances. We appreciate that some of you may not readily grasp the incredible implications of this. Think of

---

[5]This is a narrow technical meaning of the term "statistics." The term is frequently used broadly as shorthand for the entire set of logical and mathematical techniques of statistical data analysis that is the topic of this chapter. Some people (not us) use "statistics" colloquially to refer to the raw data set, as in "the weather forecaster failed to predict the weather accurately even though she had all the temperature and pressure statistics."

[6]The way to calculate the sample standard deviation, for example, is based on the population formula, but with one change. Instead of dividing the sum of squared deviations from the mean by the sample size, as is done when calculating the population standard deviation, you divide by the sample size *minus 1*. This is called the **degrees of freedom**. If it were not done this way, the sample standard deviation would be biased low as an estimate of the population standard deviation; on average, its expected value would be a little less than the actual population standard deviation.

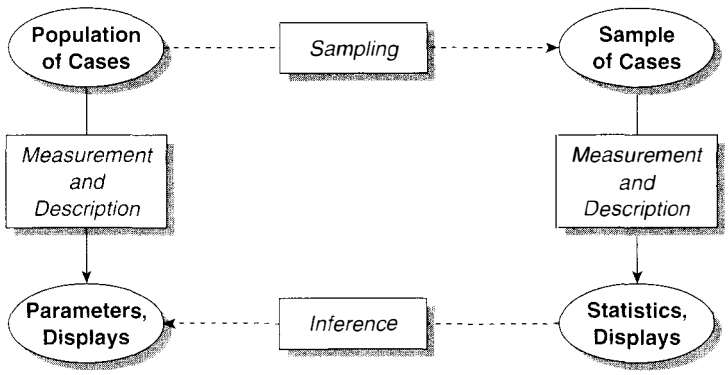

Figure 9.4 Data collection and statistical analysis with samples of cases. Almost always, researchers, especially those doing basic science, must take the indirect approach of sampling and performing inferential analyses to learn about parameters.

it like a gamble. Suppose your friend Martha tells you about a company that is publicly announcing a big business decision in two days. She strongly urges you to invest $10,000 in the company right away, because their stock is "very likely to go way up in value soon." Would you find it useful to know that her "very likely" actually means a 42% chance, and that in addition, there is a 27% chance that the stock will *drop* in value? We thought you would.

So both the power and the difficulty of inferential statistics comes from deriving probabilities about how likely it is that sample patterns, which vary from sample to sample, reflect population patterns, which are constant. The probabilities come from a distribution of the different sample patterns that describe all the possible samples of a given size that could be sampled from a population (strictly speaking, from a sampling frame—see Chapter 8). This is called the **sampling distribution** for a given statistic. For example, if I took a sample of 20 cases and calculated its mean, I would have one guess about the mean of the entire population of cases. If I took another sample of 20 that had at least one different member, I would have another sample mean that would be at least a little different than the first. I could keep doing this, recording each new sample mean as I calculated it, until I had managed to look at every different sample of size 20 that could be drawn from the population. I would then have a sampling distribution of sample means. This distribution would show the entire range of values that the sample mean could take, given the population distribution and the sample size in question. It would also show that certain values of the sample mean would occur more often than other values; for example, a sample with a mixture of high, medium, and low scores is more likely than a sample with all high scores or all low scores. Thus, the sampling distribution would show the probability that a single sample of size 20 would have a mean within some given range of values. We use these probabilities to determine the probabilities for the inferences we make from samples to populations.

If we actually constructed our sampling distributions "manually," as in our example, we would know for sure exactly what the probabilities would be for certain values of our sample statistics. But of course, that's not only impossible to do in most cases but stupid.[7] Impossible because there are far too many different samples that would have to be taken in any population of a decent size. Given a population of only 1●● cases, for example, one could draw over 5●● *quintillion* (5 followed by 2● zeroes) different samples of size 2●. And this is just a tiny population of a hundred. That leads us to the reason we say that manually deriving the sampling distribution would be stupid. In order to generate and measure all possible samples, you would need to be able to access and measure all members of the population. If you can reasonably do that, however, by all means do it and skip this inference stuff. With a population of a hundred, you should certainly do that. But as we discussed in Chapter 8, one almost never has such small populations, and often the populations in basic science are extremely large and indefinite in size.

Luckily, we don't have to use a manual approach to generating sampling distributions. Instead, we use statistical theory. For example, the "central limit theorem" says the sampling distribution of the sample mean will be a normal distribution with a mean (**expected value**) equal to the actual population mean and a standard deviation (**standard error**) equal to the population standard deviation divided by the square root of the sample size. There are a variety of other theoretical ideas about sampling distributions for other statistics.[8] However, the theories that allow generation of sampling distributions involve assumptions for their valid application. Such assumptions include:

1. ***Distributional assumptions.*** These are assumptions about properties of the population from which the sampling distribution is generated. They include normality and "homogeneity of variance"—the assumption that the variances of the populations from which separate experimental samples are taken are equal. Actually, these assumptions apply only to so-called **parametric statistical tests**. There are tests called **nonparametric** that do not require these assumptions because they assume data are only nominal or ordinal, not metric. We return to this distinction in Chapter 11.

2. *Independence of scores.* This is the assumption that individual data values from separate cases or measurement events are independent of each other. Independence

[7]Some of our colleagues actually do generate sampling distributions in this manual fashion (with a computer anyway), and they are not stupid. They do this because they are statistical researchers who specialize in developing and testing ideas about statistical tests.

[8]In practice, most inferential statistics are not done directly with the raw sample statistics but with derived indices called **test statistics**. These are typically indices that combine the statistics in question with estimates of their variability in the sampling distribution, in order to increase the interpretability of sample patterns as reflections of population patterns. For example, inferences about means are often evaluated with "*t* scores," calculated by dividing the mean by its standard error.

is the property that separate scores in a data set cannot be predicted from each other—they are uncorrelated. As we discuss below, data in geography often show nonindependence as a function of spatial relations on the earth's surface (as a function of temporal relations too).

3. ***Correct specification of models.*** These are assumptions about the appropriateness of the statistical models fitted to data. For example, do the data being fitted with a linear model really follow a linear pattern? Are all the relevant predictor constructs included in the model?

We return to the statistical assumptions in Chapter 11, where we will see that some of the assumptions are more important than others for the valid use and interpretation of inferential statistics.

Estimation and Hypothesis Testing

There are two major approaches to inferring parameters from statistics. The first is **estimation**. This is the inferential approach of choice when you do not have a particular value that you want to evaluate for the parameter but only want to make your best guess of the parameter value. Estimation has two parts, the **point estimate** and the **confidence interval** around the point estimate. The point estimate is the guess about the specific parameter value. For example, our point estimate of a population mean is usually the sample mean. The confidence interval is a range of values that is distributed, usually symmetrically, around the point estimate; it is a guess, with a specified probability of confidence, that the true value of the population parameter falls somewhere within the range of values. This confidence probability is most often .95 (95%). When pollsters tell you that candidate *X* is favored by 44% of likely voters, "plus or minus 3%," the 44% is the point estimate, and the "plus or minus 3%" is a confidence interval from 41% to 47% that you should be 95% confident contains the actual percentage of likely voters in the population who favor candidate *X*. Of course, given a particular confidence probability, we prefer narrow intervals to wide intervals; that is, we prefer greater **precision of estimation**. We do that by collecting more data or by reducing random noise in the data.

The second major approach to inference is **hypothesis testing**. This is the inferential approach of choice when you *do* have a particular value you want to evaluate for the parameter. That value is expressed in the **null hypothesis**, symbolized H_0. H_0 is a hypothesis about the exact (point) value of a parameter, or set of parameters. In statistical hypothesis testing, we use our sample statistics to make an inference about the probable truth of our null hypothesis. If we decide that our sample data indicate that the probability of the null being true is too low, we accept that the **alternative hypothesis** must instead be true. The alternative, symbolized H_A, is the hypothesis that the parameter in question does *not* equal the exact value hypothesized

by the null; the alternative thus always hypothesizes a range rather than an exact value.[9]

So hypothesis testing is used to evaluate the probability that the null hypothesis is true. The alternative hypothesis is not directly tested, even though it's typically what we believe is actually true in the population. That may sound a little confusing, but there is a good logical reason we directly test the hypothesis we do not believe is true. It's because hypothesis testing is a statistical version of a classic form of logical reasoning that allows us to disprove a tested hypothesis but not prove it. This form of logical reasoning is called **modus tollens**. In classic form, modus tollens is the conditional logic ("syllogism") of consequences. In the abstract:

–If A is true, then B is true

–B is not true

–Therefore, A is not true

A is the "antecedent" proposition, and B is the "consequent" proposition. For example, A might be the statement that "it is raining today" and B could be "I take my umbrella to work today." So this logical argument states the two premises that "If it is raining today, then I take my umbrella to work today" and "It is not true that I have taken my umbrella to work today." The valid conclusion is that "Therefore, it is not raining today." Now consider what happens if the second premise is true rather than false:

–If A is true, then B is true

–B is true

–Therefore, ??

[9]Typically, the alternative specifies the entire range of values that does not include the null value. Thus, evidence against the null is provided whenever the sample statistic is far lower *or* higher than the null value. This typical nondirectional test is called "two-tailed." If you have a prior reason to be certain the true parameter value differs from the null in one direction but not the other, some texts will recommend a directional "one-tailed" test that proposes an alternative covering the range either lower or higher than the null value, but not both. This has some effect, usually very little, on the probability with which you can reject the null hypothesis. It can work to your benefit as long as you predict the proper direction for the difference from the null, but it also prevents finding differences in the other direction if you predict incorrectly. Some researchers consider one-tailed tests to be little more than a way to help salvage weak results.

What do you conclude, now that you know that the consequent has been found to be true? If you conclude, "Therefore, A *is* true," you are making a very common and widespread logical error, the **fallacy of affirming the consequent**. In fact, the valid conclusion is:

–Therefore, no inference can be drawn

To see this, consider that although I do take my umbrella to work every day that it is raining, I also take my umbrella to work in anticipation of rain that never comes. Or I might take my umbrella to make a demonstration in class about the logic of hypothesis testing. None of this violates the truth of the two premises. In other words, this common form of conditional reasoning is not "bidirectional" with respect to the truth of the consequent. However, it is often used incorrectly in everyday reasoning as if it were bidirectional in this way.

Hypothesis testing employs modus tollens conditional logic, with the statistical twist that H_0 plays the part of A, and R_0, the likely range of the sampling distribution if the null is actually true, plays the part of B. Also, because we now must deal with statistical uncertainty, as we discussed above, we must draw conclusions that are at best "probably" rather than "definitely" true.

–If H_0 is true, then R_0 is likely	–If H_0 is true, then R_0 is likely
–R_0 is not true	–R_0 is true
–Therefore, H_0 is probably not true	–Therefore, no inference can be drawn

Thus, because of its reliance on modus tollens logic, hypothesis testing is useful for disconfirming ("disproving" is too strong a word for probabilistic reasoning) null hypotheses but not for confirming them. This fact is usually not a severe problem, however. That's because the null hypothesis is very often a guess that there is no relationship (the relationship is "null") in the population. For example, one might hypothesize that a correlation or the difference between two means in a population is zero. These hypotheses are equivalent to saying there is no relationship in the population. The alternative hypothesizes that there is a relationship.

In Table 9.1, we overview the steps of statistical hypothesis testing. In all such tests, an empirically obtained test statistic, calculated from sample data, is compared to a range of values we consider to be likely if the null were really true. That range of values is the portion of the null sampling distribution around the most likely value—the null hypothesized value. This portion can vary from test to test, but by convention it is most often the portion that ranges over 95% of the distribution around the null value; it's computed just like the 95% confidence interval in estimation, but with the null value in the center rather than the sample estimate. Every hypothesis test ends with one of two decisions. If the empirical test statistic is quite far from the null value, outside the null range, we consider the null hypothesis to be unlikely. We reject the null hypothesis and accept the alternative. This is

Table 9.1 Steps of Statistical Hypothesis Testing

1. State your null (H_0) and alternative hypotheses (H_A).
   a. null is equality of a parameter or set of parameters to a point value; for example, H_0: *Population mean = 120*.
   b. H_0 is never about the value of a statistic—you can just look at your data to find that out; H_0 is also never about "getting significance"—you will know that for sure at the end of your test.
   c. alternative is all other values; for example, H_A: *Population mean = 120*.
2. Determine your appropriate test statistic and its sampling distribution if H_0 is true (the "null sampling distribution").
   a. appropriate test statistic depends on which parameters are in your hypotheses; for example, "*t* scores" are often used to test hypotheses about one or a pair of population means.
   b. the value hypothesized by H_0 becomes the expected value of the null sampling distribution.
   c. a middle portion of the null sampling distribution, often the middle 95% like the confidence probability in estimation, is considered the "likely" range (R_0)
3. Calculate the test statistic from your sample data.
4. Compare the empirically obtained test statistic to the null sampling distribution.
   a. if the test statistic is so different from the null expected value that it falls outside the likely range, then conclude that the null is probably false ("reject the null") and the alternative is probably true ("accept the alternative")—statistical significance at a given rejection *p* level.
   b. if the test statistic is close enough to the null expected value that it falls within the likely range, then conclude that either the null or alternative might be true or false ("fail to reject the null" or "retain both hypotheses")—statistical nonsignificance at the given *p* level.

called statistical **significance** at the rejection ***p* level** of "1 minus the likely probability"; given the common likely probability of 95%, the *p* level is usually 5%. On the other hand, if the empirical test statistic is not so far from the null value and is within the null range, we conclude that we cannot reject either the null or the alternative hypotheses. We "fail to reject the null" or "retain both hypotheses." *We do not accept the null*—that would be the fallacy of affirming the consequent. This is called statistical **nonsignificance** at the applied *p* level.

In Figure 9.5, we show that when you perform a hypothesis test, you always end up making either a correct inference or a mistake. Of course, since that depends on the true value of the parameter in the population, you can never know with certainty whether you have made a mistake. But you can put a number on the chances that you have made a correct decision or a mistake. As Figure 9.5 shows, when the null hypothesis is actually true, the probability of mistakenly rejecting it, called a **Type I error**, is just the rejection *p* level, also symbolized by the Greek letter **alpha** (α). The probability of correctly retaining the null (and the alternative) in this situation is thus $1 - \alpha$. When the null hypothesis is actually false, the probability of

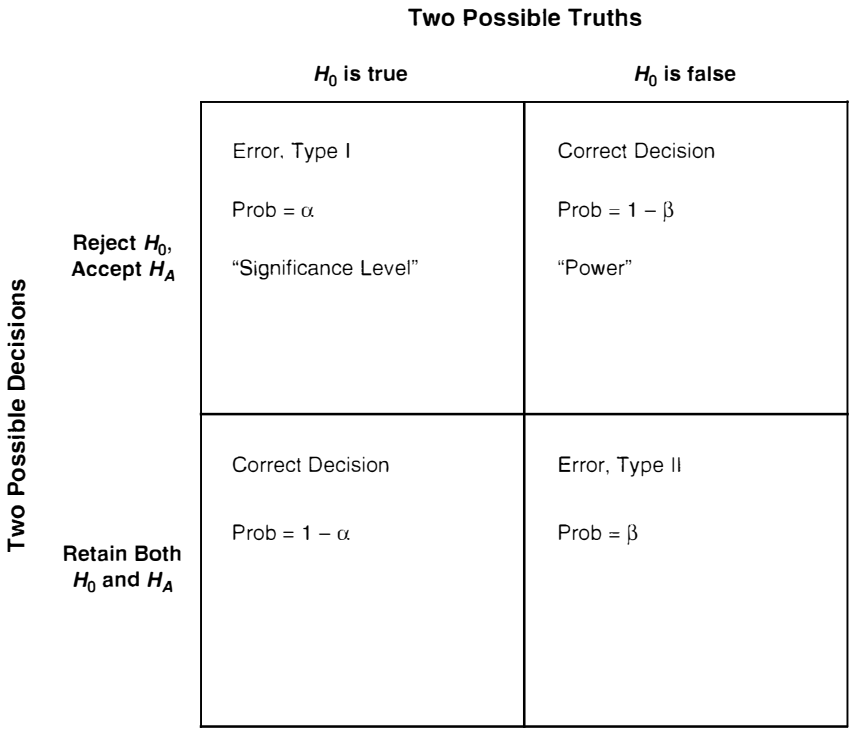

Figure 9.5 When conducting hypothesis tests, two decisions are possible and two actual truths are possible. Thus, four outcomes are possible in hypothesis testing, two of which are correct decisions and two of which are errors.

mistakenly retaining it, called a **Type II error**, is symbolized by the Greek letter **beta** (β). The probability of correctly rejecting the null (accepting the alternative) in this situation is thus 1 – β; this probability is charmingly referred to as **power**. Unlike α, which is set by convention, β and power are determined by the size of α, by the size of one's sample, by the amount of random noise in the data, and by the actual difference of the parameter in question from the value hypothesized by the null. In other words, it is mostly *not* just set by convention or choice, and is rather difficult to estimate accurately. The techniques of "power analysis" mentioned in Chapter 8 can be used to estimate power.

Perhaps a concrete example will help clarify all of this. Imagine that you want to know whether a particular coin is "fair" for flipping. Your null hypothesis would be that the probability of heads coming up equals the probability of tails: .5 or 50%. Suppose you sample the coin's fairness by flipping it six times, and it lands "heads" every time. The probability of that happening if the coin really is fair is roughly .015 or 1.5%. From the perspective of hypothesis testing, that is rather unusual if the null hypothesis is actually true. At an α-level of .05, you would reject the null and conclude the coin was unfair. According to Figure 9.5, if the coin were actually fair, you would be making a Type I error by rejecting it; this is sort of like unfairly

convicting an innocent defendant in a court trial. If the coin were actually unfair (one side was weighted), you would be making a correct decision on this test, like appropriately convicting a guilty defendant. Suppose instead that you flip the coin six times and it lands "heads" four times and "tails" two times. The probability of that happening if the coin really is fair is roughly .234 or 23.4%. That is not so unusual if the null hypothesis is actually true. At an α-level of .05, you would retain both hypotheses and conclude you did not show the coin was unfair. If in fact the coin were actually fair, you would be making a correct decision by retaining the null, like appropriately acquitting an innocent defendant. If the coin were actually unfair, you would be making a Type II error on this test, akin to finding a guilty person to be not guilty.

A final observation about hypothesis testing is in order. Although the null hypothesis most often states that there is no relationship in the population, such a hypothesis is rarely if ever going to be true for any variables in any population in any domain of reality. After all, a population correlation of .01 means the null hypothesis that the correlation is .00 is wrong; remember that positive correlations range from .01 to 1.00. In practice, however, researchers always have limited power in their data, so that a tiny population relationship is usually found to be non-significant. In most areas of geographic research, therefore, sample relationships are not considered real in the population unless they are suitably large. For example, with a sample size of 30 independent data points (considered small in some research areas, adequate in others), you would need a sample correlation of between .30 and .40 to conclude that there really was a relationship in the population. In essence, null hypothesis testing usually does not just identify false null hypotheses—it identifies null hypotheses that are *quite* false. This is a good thing.

Data in Space and Place: Introduction to Geospatial Analysis[10]

Data in geography often have a property that notably differentiates them from many data in other sciences—they are spatially distributed, and that spatiality is theoretically relevant. Geographic features have location, extent or size, shape, pattern, connectivity, and more. As we discuss further in Chapter 12, geographic data represent natural and human earth-surface features and processes, their properties, and their

[10]The large and diverse literature on geospatial data analysis tends to use the term "spatial statistics" when the features being studied are conceived of as discrete entities—points, lines, or polygons. Thus, it is the common term for most applications in human geography and other social sciences. In contrast, "geostatistics" is used when the phenomena being studied are conceived of as fields—continuous two-dimensional surfaces. It is the commonly used term in physical geography and other earth sciences. As we discussed in Chapter 8, and return to in 12, however, the discrete-continuous distinction certainly does not map perfectly on to the distinction between human and physical geosciences, and analytic techniques from each tradition have application in the other.

spatial distributions. Even when other sciences have spatial data, which they do more often than they sometimes acknowledge, that spatiality is typically not a focus of interest like it often is in geography. For example, "central place theory" explains where cities of different sizes are located, or should be located, as a function of the influence of distance on the interactions of economic agents (such as shoppers or retailers) with the economic institutions of the cities. To analyze the spatiality in data, a variety of descriptive spatial indices can be calculated that are analogous to the nonspatial descriptive statistics we overviewed above. Spatial central tendency, variability, form, and so on can be examined. For example, spatial means or medians can be calculated, as can indices of feature clustering that are somewhat analogous to variability measures. Other spatial properties can be analyzed that do not have obvious nonspatial analogues. For example, geographers sometimes want to analyze whether spatial patterns take particular shapes, such as hexagons or circles.

Spatiality is often important to geographic data analysis even when it is not the focus of interest, because its existence in data strongly influences the accuracy of inferential statistical analyses of nonspatial variables. In particular, geographic data usually exhibit a fairly robust pattern of dependence as a function of their spatial arrangement on the earth's surface (often their arrangement in time too). That is, features or processes at one place are more (or less) similar to those at other places than would be expected by chance. Such patterns of spatial dependence are known as **spatial autocorrelation**. Most often, closer places are more similar, an expression of **distance decay** or the so-called First Law of Geography: *Everything is related to everything else, but near things are more related than distant things.* This common pattern of spatial dependence is known as "positive" spatial autocorrelation. The less common reverse pattern, where closer things are less similar, is "negative" autocorrelation. And any number of hybrid or more complex patterns of autocorrelation are at least conceptually possible, if rarely investigated. Whatever its specific pattern, such spatial dependence constitutes a violation of the important statistical assumption of independence we introduced above. Independence of scores is an important requirement for the accuracy of statistical significance testing, as it is normally done. Spatiality usually violates independence in a big way that radically distorts statistical inference. It must be accounted for when spatially interpolating fields from discretely sampled data, discussed in Chapter 8. Much of the effort of geospatial analysis goes toward identifying patterns of spatial autocorrelation and accounting for them in statistical tests.

Consider, for example, a network of rain gauges set up to measure rainfall. Rainfall measurements are strongly and positively spatially autocorrelated. If it is raining in my back yard, I can say with a high degree of confidence that it is raining in my neighbor's back yard, but my level of confidence that it is raining across town is lower, and my level of confidence that it is raining 300 miles away is even lower still. This distance decay of statistical relatedness is often characterized by a **variogram** (or semi-variogram). One characteristic of spatial autocorrelation identified by a variogram is the "range" or distance over which relatedness operates. If this range for rainfall were 10 miles, it would mean that knowing it is raining at a particular place would provide me with some information about the likelihood of

rain within a radius of 10 miles from that place. Beyond 10 miles, my estimate of whether or not it was raining would return to near chance—random guessing.

The various ways that spatiality plays a role in geographic data analysis often inspires the cliché geographic witticism that "spatial data are special." And special they are—sometimes a special difficulty. Identifying and dealing with the specific pattern of autocorrelation in data can be quite challenging. As another example, distance frequently plays a leading conceptual role in geospatial analysis, but there are many ways to conceptualize and quantify distance,[11] even if it is restricted just to physical separation rather than temporal, monetary, or other forms of separation, a restriction that is often unacceptable. Other difficulties include the fact that several descriptive spatial indices are insensitive to the overall pattern of the data, such as the **Gini coefficient**, which is used to calculate economic or demographic segregation. Some indices depend in theoretically uninteresting ways on the size of the surrounding area used to delimit, typically somewhat arbitrarily, the scope of a problem. Some delimitation is required; after all, we can't use the entire earth surface or the universe as the spatial backdrop for every test we carry out.

One of the special difficulties of spatial analysis arises from questions about which areal units should be used to analyze geographic data. It is common in geography to have data organized into areal units (zones or regions) that are at least partially if not totally arbitrary with respect to the researcher's phenomenon of interest. As we saw in Chapter 6, this almost always characterizes census data. Researchers usually don't believe that the causal factors underlying their phenomenon of interest operate at the census tract level, but that's one of the common ways the data are packaged. Most other secondary sources of data are like this too, and we have already noted that geographers use a lot of secondary data. For most questions of interest in basic research, units like nature preserves or U.S. states are pretty arbitrary, but that is what's available. And in a disturbingly large number of situations, researchers will take data off of a continuous representation like a map, but break the map into areal pieces so that they can apply discrete spatial analysis techniques (this is analogous to the discretization discussed in Chapter 8 that geographers carry out in order to sample from fields). This practice is even more dubious given that researchers are typically ignorant about how the data used to make the map were obtained and treated in the first place. How's an honest researcher to choose an appropriate way to break up the space when there are an infinite number of possibilities?

The rub with all this is that changing the number, size, shape, and/or location of areal units can change the results of analyses, often dramatically. The phenomenon known as **gerrymandering** provides a great example of this. Gerrymandering refers

[11]There are numerous (potentially infinite) "metric geometries" that calculate distance differently. Some of these have been applied in geography, including the city-bock and spherical metrics (after all, the earth is very nearly a sphere). If that weren't enough, consider that physical separation is often best considered in terms of such things as counts of the number of stops a subway train makes, rather than metric distances.

to the design of electoral-district boundaries to concentrate certain types of voters into certain districts so as to give particular candidates an advantage in the election.[12] Both racial and political-party gerrymandering are alive and well in the United States. For example, the state of Pennsylvania in the year 2000 had a majority of voters registered as Democrats; however, resulting in part from gerrymandering of the congressional districts, the majority of congressmen elected from the state of Pennsylvania were Republican (it has certainly worked out for the Democrats in other situations). Figure 9.6 demonstrates the profound effect that redesigning district boundaries can have on election outcomes. The effect that theoretically arbitrary areal geometries can have on the results of geographic analyses is known as the **Modifiable Areal Unit Problem**, or MAUP for short.

We mentioned size as an aspect of areal units that contributes to the MAUP. The question of the proper size for the areal units to be used in geographic analyses really goes beyond just the MAUP. It is an aspect of the fundamental issue of scale in geography. In Chapter 2, we pointed out that scale concerns time and theme as well as space. We also distinguished between phenomenon and analysis scale (and also discussed cartographic scale), noting that geographic phenomena are very often scale-dependent. That is, theories or models often apply at one scale, or a range of scales, and not at others. In order to observe and study a phenomenon accurately, researchers must match their scale of analysis to the actual scale of the phenomenon. That is, researchers must identify the scale of a phenomenon so they can collect and organize their data in units of that size.[13]

Given spatial units of a particular size, one can readily aggregate or combine them into larger units; it is not possible without additional information or theory to disaggregate them into smaller units. A great deal of geographic data is aggregated from data gathered at a finer spatial resolution. U.S. census data again provide a good example. As we saw in Chapter 6, census data are summarized at several levels, from the whole country to blocks, yet these levels of analysis are derived by aggregation from individual responses to the census form. The level at which the data are aggregated can seriously influence statistical patterns identified in the data and the ultimate conclusions drawn about their meaning. As a general rule, the correlation between two geographically distributed variables increases with their level

---

[12]Gerrymandering was named by the artist Gilbert Stuart (his portrait of George Washington is on the U.S. dollar bill) in an 1812 political cartoon depicting the complexly shaped Massachusetts voting district designed by Governor Elbridge Gerry to concentrate his Federalist opponents in one district. Stuart depicted the sinuous arching district as a salamander—or "gerry-mander." It looks more like a winged dragon, if you ask us.

[13]Using data at one scale to make inferences about phenomena at other scales is known as the **cross-level fallacy**. A specific instance of this is making inferences in aggregated form from data that were measured on individual people; this was identified as the "ecological fallacy" in a classic paper: Robinson, W. S. (1950). Ecological correlation and the behavior of individuals. *American Sociological Review, 15,* 351–357. The reverse error might be called the "atomistic fallacy."

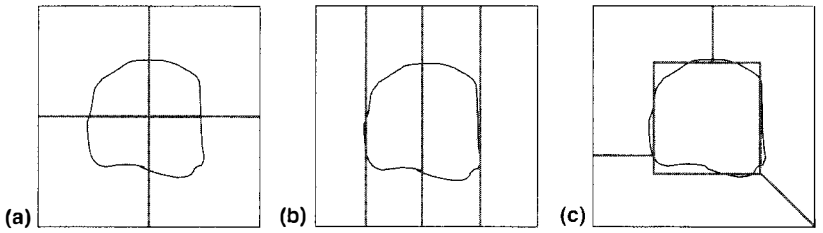

Figure 9.6 Gerrymandering as an example of the MAUP. The gray region represents an area in which the residents are predominantly of the same group, such as members of a racial or ethnic group, or a particular political party (the surrounding white region would be made up of residents that are not members of the group). The straight lines represent alternative voting district boundaries that could be designed. In (a), the group is broken up so that it is a distinct minority in each of the four districts. In (b), the group is broken up so that it is completely missing from two districts but constitutes nearly half of each of the other two districts. In (c), the group is not broken up at all, so it is completely missing from three districts but constitutes a strong majority of the fourth district. Of course, none of this matters for an election unless members of the group are in fact more likely to vote for the same candidates than would be expected by chance (and the candidates must be different than those the residents in the white region would vote for)

of aggregation. For example, assume you run a simple regression between median household income and education level for several different levels of spatial aggregation: for example, states, counties, and census tracts. The highest correlation will likely occur at the state level and the lowest at the tract level. But the potential severity of this aggregation effect can be even more disturbing than that, especially when other aspects of the MAUP apply in such situations as well. Given the right arrangement and aggregation level of data units, the correlation between two variables distributed across the earth's surface can be calculated to take on almost any value from +1.0 to −1.0!

Problems of the scale and arrangement of data units are deep and fundamental in many domains of geographic research, and it appears that many of them will not be unambiguously solved any time soon. Ideally, one would like a theory that specifies the scale and arrangement of spatial (as well as temporal and thematic) units at which structures and processes actually exist. Unfortunately, such theories are few and far between. Lacking this, as geographers typically do, it is often recommended that empirical "trial-and-error" approaches be used to try to identify the appropriate units at which a phenomenon should be analyzed. Computer tools exist that allow analysts to try many alternative regionalizations of the data space and hope for similar results across the regionalizations, or at least identify systematic changes across regionalizations that can be characterized.

However, the best way to deal with these problems remains contested in the field today.[14]

Review Questions

- How does data analysis contribute to the four scientific goals of description, prediction, explanation, and control?
- Why is data analysis in geography usually conceptualized in statistical (probabilistic) terms?

Statistical Description

- What are the following properties that describe distributions of data, and what are some specific indices for expressing each one: central tendency, variability, form, derived scores, relationship?
- What do we mean by the strength and form of statistical relationships?
- What are monotonic, linear, and nonlinear relationship forms?

Statistical Inference

- What is the purpose of statistical inference, and why are statistical inferences necessarily and ultimately uncertain?
- Why do we say that sampling error is not an *error* in the sense of a mistake?
- What are sampling distributions, and how do they relate to population and sample distributions? In scientific research, sampling distributions are generated by statistical theory rather than actual repeated sampling; why is this so?

Estimation and Hypothesis Testing

- What are the similarities and differences between the two inferential procedures of estimation and hypothesis testing?
- What are point estimates and confidence intervals in estimation?
- Why is hypothesis testing ultimately useless for confirming null hypotheses?

[14]Tobler offered the view that MAUP effects result from inappropriate methods of analyses, and that spatial analysts should find "frame-independent" analytic tools that produce the same results regardless of the partitioning of the data space. See Tobler, W. R. (1990). Frame independent spatial analysis. In M. Goodchild & S. Gopal (Eds.), *Accuracy of spatial databases* (pp. 115–122). London: Taylor & Francis. This solution is not very satisfying, at least with respect to scale aspects of the MAUP, because many of us consider the frequent scale-dependent nature of geographic phenomena to be a substantive reflection of their true nature rather than just an artifact of human analysis.

- What are the steps of hypothesis testing, including the two possible decisions that it can lead to?
- What are the two types of correct decisions and two types of errors possible when hypothesis testing?

Introduction to Geospatial Analysis

- What are some spatial properties of phenomena that might be of interest when analyzing data?
- What is spatial autocorrelation, what forms can it take, and why is it so important to geographic data analysis?
- What is the MAUP, and what is a specific example of MAUP? How is gerrymandering an example of MAUP?

Key Terms

alpha (α): the probability of mistakenly rejecting the null hypothesis in hypothesis testing when it is actually true; the same as the significance "*p* level"

alternative hypothesis: the hypothesis that expresses all possible values for the parameter not specified by the null hypothesis; symbolized H_A, it is always a range guess about the value of a parameter or set of parameters

arithmetic mean: the most common type of mean, calculated as the sum of the values of all scores divided by the number of scores

bell-shaped distribution: distribution of data that has the form of being symmetric and unimodal

beta (β): the probability of mistakenly retaining the null hypothesis in hypothesis testing when it is actually false

central tendency: the descriptive property of the average or most representative value in a data set

confidence interval: one's guess of the range of values the parameter probably has in estimation, with some level of confidence probability

correlation coefficient: index of relationship strength; when dealing with linear relationships, it is calculated to equal 1.0 when there is a perfect positive relation, –1.0 when there is a perfect negative relation, and 0.0 when there is no linear relationship at all

criterion variables: the variables chosen to be predicted by the values of the predictor variables in a regression model, usually designated *Y*

cross-level fallacy: drawing inferences about phenomena at one scale from measurements made at smaller (the ecological fallacy) or larger (the atomistic fallacy) scales

data analysis: the set of display and mathematical techniques, and attendant logical and conceptual considerations, used to extract meaning from data and communicate it to others

degrees of freedom: in data analysis, the number of independent pieces of information represented by a given sample data set; usually just a little fewer than the sample size

derived score: way to describe properties of individual scores in a data set by expressing their value relative to the rest of the data set

descriptive statistics: the branch of statistical data analysis that uses mathematical and display techniques to describe patterns in data sets

deviations from the mean: the basis for the variance and standard deviation, calculated for each score by subtracting the arithmetic mean from the score

distance decay: positive spatial autocorrelation, wherein similarity between phenomena decreases as distance between them increases; famously referred to as the First Law of Geography

distributional assumptions: assumptions about properties of the population from which the sampling distribution is generated, including normality and homogeneity of variance; theoretically must be valid for the valid conduct of parametric tests

error of prediction: the vertical distance between data points and the predicted line in a regression model of relationship; assumed to be random from case to case

estimation: one of two major approaches to inferential statistics (hypothesis testing being the other), appropriate when you do not have a particular value that you want to evaluate for the parameter but want only to make your best guess of its value

expected value: the arithmetic mean of the sampling distribution of a given statistic

fallacy of affirming the consequent: common logical mistake in modus tollens reasoning in which a true consequent, the second premise of the argument, is taken as evidence for a true conclusion

form: the descriptive property of the shape of the distribution of a data set, easier to grasp when the data are graphed

geospatial analysis: data analysis that explicitly takes account of the spatiality in geographic data; variously called "geostatistics" or "spatial statistics"

gerrymandering: the design of electoral-district boundaries to concentrate certain types of voters into certain districts so as to give particular candidates an advantage in an election; expression of the MAUP in an electoral context

Gini coefficient: statistical index of the concentration of some property within areal units (regions), as compared to other units; typically used to quantify residential segregation based on wealth or ethnicity

H_0: common symbol for the null hypothesis in hypothesis testing

H_A: common symbol for the alternative hypothesis in hypothesis testing

hypothesis testing: one of two major approaches to inferential statistics (estimation being the other), appropriate when you have a particular value that you want to evaluate for the parameter

independence of scores: the property that individual scores in a data set cannot be predicted from each other—they are uncorrelated; it is an important assumption in inferential statistics that is nonetheless frequently violated in geography

inferential statistics: the branch of statistical data analysis that attempts to infer patterns in populations of data from the evidence of samples of data; in addition to generating accurate guesses, inferential procedures generate probabilities that the guesses are correct or close

linear relationship: the simplest form of relationship, in which the values of two variables tend to follow a straight line when graphed against each other

mean: descriptive index of central tendency calculated as the average of all values in a data set

median: descriptive index of central tendency calculated as the middle value in a data set

modality: descriptive index of form calculated as the number of "local" modes in a data set, which are values that occur more commonly than values just below or above them

mode: descriptive index of central tendency calculated as the most frequently occurring value in a data set

Modifiable Areal Unit Problem: the effect that theoretically arbitrary areal geometries can have on the results of geographic analyses; "MAUP" is its common acronym

modus tollens: classic form of the conditional logic of consequences that provides the basis for statistical hypothesis testing

monotonic: approximate relationship form that goes in only one direction, either up or down, but does not necessarily follow an exact straight line

multivariate: relationships and the techniques for analyzing them that involve more than a pair of variables

negative relationship: linear relationship in which high values on one variable tend to go with low values on the other variable, and low values tend to go with high values; also called "indirect" relationship

nonlinear relationship: any relationship form that follows a pattern other than a straight line

nonparametric statistical tests: class of inferential statistical tests appropriate for nonmetric data and for metric data that violate the distributional assumptions; they are less powerful and less flexible than parametric tests

nonsignificance: the outcome of hypothesis testing that does not allow you to reject the null hypothesis but forces you to retain both hypotheses at a given rejection probability p

normal distribution: particular and important bell-shaped distribution of data that has a specific proportion of its scores within any given range from the central tendency

null hypothesis: the hypothesis that is tested when doing hypothesis testing; symbolized H_0, it is always a point guess about the value of a parameter or set of parameters

parameters: summary statistical indices calculated to describe properties of population data

parametric statistical tests: class of inferential statistical tests for metric data that satisfy the distributional assumptions

percentile: derived score calculated as the percentage of the data set that is less than the score in question; for example, the median is at the 50th percentile

***p* level:** the rejection probability in a hypothesis test, which equals "1 minus the null likely probability (confidence probability)"; set by convention, it is most often 5%

point estimate: in estimation, one's guess of the precise value of the parameter

positive relationship: linear relationship in which high values on one variable tend to go with high values on the other variable, and low values tend to go with low values; also called "direct" relationship

power: the probability of correctly rejecting the null hypothesis in hypothesis testing when it is actually false; equal to $1 - \beta$

precision of estimation: the width of the confidence interval in estimation, given a particular confidence probability

predictor variables: the variables chosen to predict the values of the criterion variables in a regression model, usually designated X

range: descriptive index of variability calculated as the difference between the highest and lowest values in a data set

regression analysis (model): statistical model of the form of the relationship between two or more variables

relationship: descriptive property of how values of pairs (or larger sets) of variables vary across cases in systematic ways

relationship strength: the degree to which systematic relationship patterns hold across all cases

sampling distribution: distribution of a sample statistic based on all possible samples of a given size taken from a given population; unless one is teaching or doing research specifically *on* statistical analysis, the sampling distribution is created from theory rather than actual repeated sampling

sampling error: the way different samples from the same population vary from each other

significance: the outcome of hypothesis testing that allows you to reject the null hypothesis and accept the alternative hypothesis at a given rejection probability p

skewness: descriptive index of form that is the deviation from symmetry caused by an uneven distribution of extreme score values to the high or low side; "positive" skew has extreme scores to the high side, "negative" skew has extreme scores to the low side

spatial autocorrelation: nonindependence among measurements of phenomena as a function of their location relative to other phenomena; commonly observed in geographic data, it may be positive, negative, or some combination of the two

standard deviation: descriptive index of variability calculated as the square root of the variance

standard error: the standard deviation of the sampling distribution of a given statistic

statistical data analysis: approach to mathematical data analysis that treats data as reflecting phenomena of interest in a probabilistic way, rather than a deterministic way

statistics: summary statistical indices calculated to describe properties of sample data and infer properties of population data

symmetry: descriptive index of form for distributions of data sets with two sides that are mirror images around their central tendency; a symmetric distribution has no skew

test statistic: derived indices used to conduct inferential statistics, calculated by combining a particular statistic with estimates of the variability in the data

transformation: mathematical operation applied to each raw score in a data set to produce a new data set that has particular desired properties; for example, a nonlinear relationship of a certain form can be linearized by taking the logarithm of each score

Type I error: mistakenly rejecting the null hypothesis in a hypothesis test when it is actually true

Type II error: mistakenly retaining the null hypothesis in a hypothesis test when it is actually false

univariate: relationships and the techniques for analyzing them that involve a single pair of variables

variability: descriptive property of how values in a data set differ from the central tendency or each other; also called "dispersion"

variance: descriptive index of variability calculated as the sum of squared deviations from the mean, divided by the number of deviation scores; the square of the standard deviation

variogram: graphical display used to identify patterns of spatial autocorrelation; it shows average dissimilarities between measurements as a function of distance between them

***z* score:** derived score for metric-level data, calculated by dividing a score's deviation from the mean by the standard deviation of its data set

Bibliography

Clark, W. A. V., & Hosking, P. L. (1986). *Statistical methods for geographers.* New York: Wiley.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography.* Thousand Oaks, CA: Sage.

Games, P. A., & Klare, G. R. (1967). *Elementary statistics: Data analysis for the behavioral sciences.* New York: McGraw-Hill.

Haining, R. (1990). *Spatial data analysis in the social and environmental sciences.* Cambridge, U.K.: Cambridge University Press.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics.* New York: Oxford University Press.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem.* Norfolk, U.K.: Geo Books.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.