Peter A. Rogerson

STATISTICAL METHODS *for* GEOGRAPHY

A STUDENT'S GUIDE

1

INTRODUCTION TO STATISTICAL METHODS FOR GEOGRAPHY

1.1 INTRODUCTION

The study of geographic phenomena often requires the application of statistical methods to produce new insight. The following questions serve to illustrate the broad variety of areas in which statistical analysis has been applied to geographic problems:

1. How do blood lead levels in children vary over space? Are the levels randomly scattered throughout the city, or are there discernible geographic patterns? How are any patterns related to the characteristics of both housing and occupants? (Griffith et al. 1998).

2. Can the geographic diffusion of democracy that has occurred during the post–World War II era be described as a steady process over time, or has it occurred in waves, or have there been 'bursts' of diffusion that have taken place during short time periods? (O'Loughlin et al. 1998).

3. What are the effects of global warming on the geographic distribution of species? For example, how will the type and spatial distribution of tree species change in particular areas? (MacDonald et al. 1998).

4. What are the effects of different marketing strategies on product performance? For example, are mass–marketing strategies effective, despite the more distant location of their markets? (Cornish 1997).

These studies all make use of statistical analysis to arrive at their conclusions. Methods of statistical analysis play a central role in the study of geographic problems – in a survey of articles that had a geographic focus, Slocum (1990) found that 53% made use of at least one mainstream quantitative method. The role of statistical analysis in geography may be placed within a broader context through its connection to the 'scientific method', which provides a more general framework for the study of geographic problems.

1.2 THE SCIENTIFIC METHOD

Social scientists as well as physical scientists often make use of the *scientific method* in their attempts to learn about the world. Figure 1.1 illustrates this method, from the initial attempts to organize ideas about a subject, to the building of a theory.
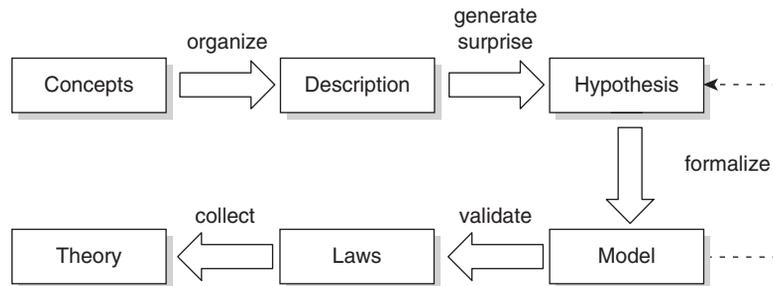

Figure 1.1   The scientific method

Suppose that we are interested in describing and explaining the spatial pattern of cancer cases in a metropolitan area. We might begin by plotting recent incidences on a map. Such descriptive exercises often lead to an unexpected result – in Figure 1.2, we perceive two fairly distinct clusters of cases. The surprising results generated through the process of description naturally lead us to the next step on the route to explanation by forcing us to generate hypotheses about the underlying process. A 'rigorous' definition of the term *hypothesis* is a proposition whose truth or falsity is capable of being tested. We can also think
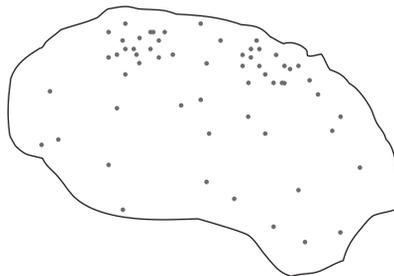

Figure 1.2   Distribution of cancer cases

of hypotheses as potential answers to our initial surprise. For example, one hypothesis in the present example might be that the pattern of cancer cases is related to the distance from local power plants.

To test the hypothesis, we need a *model*, which is a device for simplifying reality so that the relationship between variables may be more clearly studied. Whereas a hypothesis might suggest a relationship between two variables, a model is more detailed, in the sense that it suggests the nature of the relationship between the variables. In our example, we might speculate that the likelihood of cancer declines as the distance from a particular power plant increases. To test this model, we could plot cancer rates for a subarea versus the distance from the subarea centroid to the power plant. If we observe a downward slop-ing curve, we have gathered some support for our hypothesis (see Figure 1.3).

Figure 1.3   Cancer rates versus distance from power plant

Models are validated by comparing observed data with what is expected. If the model is a good representation of reality, there will be a close match between the two. If observa-tions and expectations are far apart, we need to 'go back to the drawing board', and come up with a new hypothesis. It might be the case, for example, that the pattern in Figure 1.2 is due simply to the fact that the population itself is clustered. If this is true, the spatial pat-tern of cancer then becomes understandable; a similar rate throughout the population generates apparent cancer clusters because of the spatial distribution of the population.

Though models are often used to learn about particular situations, more often one also wishes to learn about the underlying process that led to it. We would like to be able to *generalize* from one study to statements about other situations. One reason for studying the spatial pattern of cancer cases is to determine whether there is a relationship between cancer rates and the distance to *specific* power plants; a more general objective is to learn about the relationship between cancer rates and the distance to *any* power plant. One way of making such generalizations is to accumulate a lot of evidence. If we were to repeat our analysis in many locations throughout a country, and if our findings were similar in

all cases, we would have uncovered an empirical generalization. In a strict sense, *laws* are sometimes defined as universal statements of unrestricted range – that is, the statement may be applied to *any* place, during *any* time period. In our example, we might not expect our generalization to have an unrestricted range, and we might want, for example, to confine our generalization or empirical law to power plants and cancer cases in the country of interest.

Einstein called theories 'free creations of the human mind'. In the context of our diagram, we may think of theories as collections of generalizations or laws. The collection is greater than the sum of its parts in the sense that it gives greater insight than that produced by the individual generalizations or laws alone. If, for example, we generate other empirical laws that relate cancer rates to other factors, such as diet, we begin to build a theory of the spatial variation in cancer rates.

Statistical methods occupy a central role in the scientific method, as portrayed in Figure 1.1, because they allow us to suggest and test hypotheses using models. In the following section, we will review some of the important types of statistical approaches in geography.

1.3 EXPLORATORY AND CONFIRMATORY APPROACHES IN GEOGRAPHY

The scientific method provides us with a structured approach to answering questions of interest. At the core of the method is the desire to form and test *hypotheses*. As we have seen, hypotheses may be thought of loosely as potential answers to questions. For instance, a map of snowfall may suggest the hypothesis that the distance away from a nearby lake may play an important role in the distribution of snowfall amounts.

Geographers use spatial analysis within the context of the scientific method in at least two distinct ways. *Exploratory* methods of analysis are used to *suggest* hypotheses; *confirmatory* methods are, as the name suggests, used to help confirm hypotheses. A method of visualization or description that led to the discovery of clusters in Figure 1.2 would be an exploratory method, while a statistical method that confirmed that such an arrangement of points would have been unlikely to occur by chance would be a confirmatory method. In this book, we will focus primarily upon confirmatory methods.

We should note two important points here. First, confirmatory methods do not always confirm or refute hypotheses – the world is too complicated a place, and the methods often have important limitations that prevent such absolute confirmation and refutation. Nevertheless, they are important in structuring our thinking and in taking a rigorous and scientific approach to answering questions. Second, the use of exploratory methods over the past few years has been increasing rapidly. This has come about as a result of a combination of the availability of large databases and sophisticated software (including GIS), and a recognition that confirmatory statistical methods are appropriate in some situations and

not others. Throughout the book, we will keep the reader aware of these points by point-ing out some of the limitations of confirmatory analysis.

1.4 PROBABILITY AND STATISTICS

1.4.1 Probability

Probability may be thought of as a measure of uncertainty, with the measure taking on a value ranging from zero to one. Experiments and processes often have many possible out-comes, and the specific outcome is uncertain until it is observed. If we happen to know that a particular outcome will definitely not occur, that outcome has a probability of zero. At the other extreme, if we know that an outcome *will* occur, it is said to have a probabil-ity of one. At these two extremes, there is no uncertainty regarding whether the outcome will occur. When we are maximally uncertain about the outcome of an event, its probabil-ity is equal to 0.5 (i.e., 1/2); there is a '50–50' chance that it will occur. A major focus of the study of probability is the study of the likelihood of various outcomes. How likely or probable is it that a town will be struck by two hurricanes in one season? What is the probability that a resident in a community who is 4 km from a new grocery store will become a new customer?

Probabilities may be derived in a variety of ways, ranging from subjective beliefs, to the use of relative frequencies of past events. When guessing whether a coin will come up heads when tossed, you may choose to believe that the probability is 0.5, or you could actually toss the coin many times to determine the proportion of times that the result is heads. If you tossed it 1000 times, and it came up heads 623 times, an estimate of the prob-ability of heads that relied on relative frequency would be $623/1000 = 0.623$.

The study of probability has its origins, at least to some degree, in questions of gambling that arose in the 17th century. In particular, correspondence between Pascal and Fermat in 1654 concerned how to properly resolve a game of chance proposed by deMéré that had to be terminated before its conclusion. Suppose that the first player with three wins is declared the overall winner, and can lay claim to the prize of 64 pistoles (Pistole is the French word given at the time to a Spanish gold coin). How should the prize money be divided, given that the game had to be terminated, and given that one player had two wins, and the other had one win? DeMéré argued that the first player should receive 2/3 of the prize, because that person had two-thirds of the wins. Since 2/3 of 64 is 42.67, the first player would receive the remaining 21.33 pistoles.

Fermat wrote to Pascal with an alternative solution. Pascal agreed, and in a letter to Fermat suggested a more compact way of stating the solution. In particular, he argued that the two players should consider what could happen if they continued. With probability equal to ½, the player with only one win would win the next round, and they could then split the pot of money (each receiving 32 pistoles), since they would then have an equal chance of winning

the contest. With probability also equal to ½, the player with two wins would win the next round, and consequently the entire prize of 64 pistoles. Pascal argued that this player's fair share was an average of these two, equally likely outcomes; his rightful share was therefore (32+64)/2 = 48 (and not 42.67, as DeMéré had suggested). Pascal's reasoning, which is based upon probabilities and possible outcomes, forms the basis of modern probability.

What is the difference between *probability* and *statistics*? The field of probability provides the mathematical foundation for statistical applications. Year-long courses in probability and statistics are often subdivided into a first-semester course on probability, and then a second-semester course in statistics. Probability is discussed in more detail in Chapters 3 and 4; in the next section, we describe in more detail the field of statistics.

1.4.2 Statistics

Historically, a 'statist' was a word for a politician, and statistics was 'the branch of political science concerned with the collection, classification, and discussion of facts bearing on the condition of a state or community' (Oxford English Dictionary, online at www.oed.com/view/Entry/189322?)). A good example of this usage that survives to the present is the term 'vital statistics' – used to describe the collection and tabulation of information on a region's rates and numbers of births and deaths.
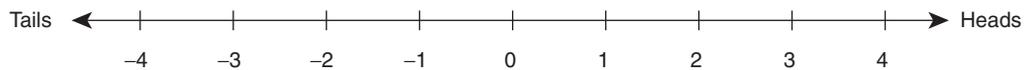

Figure 1.4   One-dimensional space for random walk

McGrew and Monroe (2000) define statistics as 'the collection, classification, presentation, and analysis of numerical data'. Note that this definition contains both the historical function of collection, classification, and presentation, but also the *analysis* of data. Modern definitions have in common the objective of inferring from a sample of data the nature of a larger population from which the sample was drawn. Statistics is often subdivided into two general areas – *descriptive statistics* are used to summarize and present information and this is in keeping with the more historical definition of the field; *inferential statistics*, as the name implies, allow inference about a larger population from a sample.

1.4.3 Probability Paradoxes

The following paradoxes are described for both mild amusement and to show that, although the use of probability to answer questions can often lead to intuitive outcomes, careful consideration is sometimes required to think through seemingly counterintuitive results.

1.4.3.1 A spatial paradox: random movement in several dimensions

This paradox is taken from Karlin and Taylor (1975). Consider a number line as in Figure 1.4, and suppose that our initial position is at the origin. We flip a single coin to govern our movement; if it is heads, we move one step to the right, and if it is tails, we move one step to the left. If we flip the coin many times, it is certain that we will, at some point, return to the origin (implying that, at that point in time, the total number of heads is equal to the total number of tails). This should not be surprising – it accords with our intuition that the number of heads and tails exhibited by a fair coin should be roughly equal.

Now consider generalizing the experiment to two dimensions (Figure 1.5), where the outcome of two coin flips governs the movement on the two-dimensional grid. One coin governs movement in the vertical direction and the other governs movement in the hori-zontal direction (for example, go one step up and one step to the right if both coins are heads, and one step down and one step to the left if both are tails). Again, it is possible to show that, although the path will certainly wander in the two-dimensional space, it is cer-tain that there will be a return to the origin.

Finally, extend the procedure to three dimensions; each of three coins governs movement in one of the three dimensions. Movement starts at the origin and proceeds to lattice points within a cube. It now turns out that a return to the origin is no longer guaranteed! That is, there is a probability greater than zero that the random path will wander away from the origin and never return! This conclusion is also true for random walks in all dimensions greater than three. This is an example where the process of induction fails – what is true here in one and two dimensions cannot be generalized to higher dimensions. As the number of dimensions increases, the volume of space is increasing faster than the linear increase in the number of coins we have to govern our movement in the space. The example highlights the
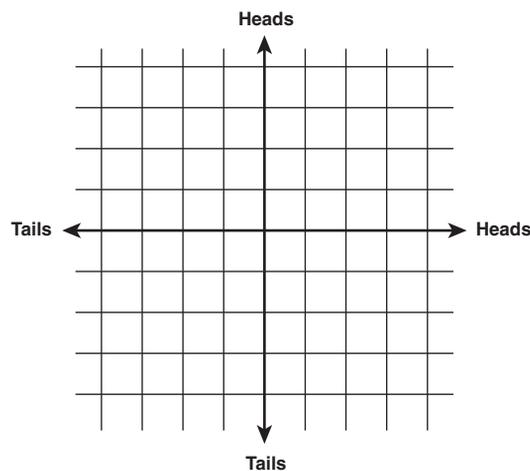

Figure 1.5   Two-dimensional space for random walk

fact that, while our intuition is often good, it is not perfect. We need to rely not only on our intuition about probability, but on a firmer foundation of the theory of probability.

1.4.3.2 A non-spatial paradox: quality pie

This probability paradox is taken from Martin Gardner's 'Mathematical Games' section of *Scientific American* (March 1976).

Consider an individual who goes into a diner each day to have a piece of pie. The diner always has apple and cherry pie, and it sometimes has blueberry pie. The quality of the pies is rated on a scale of one (lousy) to six (excellent), and the daily variability in the quality of each is summarized in Figure 1.6. For example, the cherry pie is either very good (it has a rating of five 49% of the time) or barely palatable (it has a rating of one 51% of the time). The person wishes to make a choice so as to maximize the proportion of time he or she ends up with the best pie. (With the exception of the blueberry pie, the person does not know the quality of the pie until after they have ordered!)
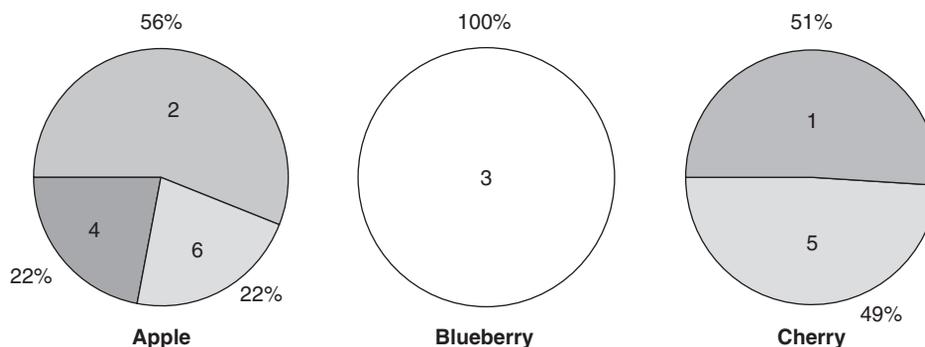

Figure 1.6   The relative frequency of pie quality

Consider first the decision that will be made on days where there is no blueberry pie. The possibilities are given in Table 1.1 (the best choice for the day is in bold).

The probabilities represent the proportion of times that particular combination of pie qualities will occur. If the person chooses apple pie, they will be getting the best pie the diner has to offer about 62% of the time (0.1078 + 0.1122 + 0.1122 + 0.2856 = 0.6178). If they choose cherry, they will get the best pie only 38% of the time (0.1078 + 0.2744 = 0.3822). The choice is clear – go with the apple pie.

Now let's examine what happens when the diner also happens to have blueberry pie. The possibilities are now given in Table 1.2. Here, apple is best about 33% of the time (0.1078 + 0.1122 + 0.1122 = 0.3322), cherry is best about 38% of the time (0.1078 + 0.2744 = 0.3822),

and blueberry is best almost 29% of the time (it is only best on days when apple has a rating of two and cherry a rating of one − which occurs 28.56% of the time).

Table 1.1   Pie qualities and probabilities: apple and cherry

Apple	Cherry	Probability
6	5	.22 × .49 = .1078
6	1	.22 × .51 = .1122
4	5	.22 × .49 = .1078
4	1	.22 × .51 = .1122
2	5	.56 × .49 = .2744
2	1	.56 × .51 = .2856

Table 1.2   Pie qualities and probabilities: apple, blueberry, and cherry

Apple	Blueberry	Cherry	Probability
6	3	5	.22 × .49 = .1078
6	3	1	.22 × .51 = .1122
4	3	5	.22 × .49 = .1078
4	3	1	.22 × .51 = .1122
2	3	5	.56 × .49 = .2744
2	3	1	.56 × .51 = .2856

The best choice is now to go with the cherry pie. Thus, we have a rather bizarre sce-nario. The optimal strategy should be for the individual to ask the waiter or waitress if they happen to have blueberry pie; if they don't, the person should choose apple, and if they do, the person should choose cherry!

Recall that the objective here was to maximize the proportion of times that one would choose the best pie. A more common objective, employed in economic theory, is to maximize expected utility, which in this case would mean making a choice to maximize average pie quality. Apple has an average quality of $(6 \times 0.22) + (4 \times 0.22) + (2 \times 0.56) = 3.32$. Cherry has an average quality of $(5 \times 0.49) + (1 \times 0.51) = 2.96$, and blueberry has an average quality of 3. Using this objective, one should choose apple if they don't have blueberry (as before); if they have blueberry, one should still choose apple, since it has the best average quality. The economist's objective of maximizing expected utility leads to consistent results; other objectives can possibly lead to counterintuitive results.

As is pointed out in the original article, the example with pie is entertaining, but the example takes on more significance if one now imagines the information in Figure 1.6 representing the effectiveness of three alternative drugs in treating an illness. Imagine that

there are two drugs, A and C, and their effectiveness is summarized by the 'pie' charts on the left and right sides in Figure 1.6 (where for example the '6' implies that drug A is found to be highly effective for 22% of all patients it is administered to). To obtain the best out-come the most frequently, one should choose drug A. Now suppose that drug B enters the market (the middle pie chart in Figure 1.6); the result is that in a three-way comparison of the available drugs, most frequently it is drug C that is best.

1.4.4 Geographical Applications of Probability and Statistics

This section provides examples of geographical applications of probability and statistics. The first two are what may be described as traditional, common applications, of the type we will address later in the book. The second two are illustrative of the unique and novel ways in which probability and statistics can be used to address geographical questions.

1.4.4.1 Are housing prices lower near airports?

An important objective in urban geography is to understand the spatial variation in hous-ing prices. Housing characteristics such as lot size, the number of bedrooms, and the age of the house have a clear influence on selling prices. Characteristics of the neighborhood can also influence prices; whether a house is situated next to an industrial park or a rec-reational park is likely to have a clear effect on the price!

A nearby airport could potentially have a positive impact on prices, since accessibility is generally desirable. However, owning a home in the flight path of an airport may not nec-essarily be positive when the noise level is taken into consideration. We could take a sample of homes near the airport in question; we could also find a sample of homes that are not near the airport with similar characteristics (e.g., similar number of bedrooms, floor space, lot size, etc.). Suppose we find that the homes near the airport had a mean selling price that was lower than the homes not located near the airport. We need to decide whether (a) the sample reflects a 'true' difference in housing prices, based upon location with respect to the airport, or (b) the difference between the two locations is not significant, and the observed sample difference in prices is the result of sampling fluctuations (keep in mind that our samples represent a small fraction of the homes that could potentially be sold; if we went out and collected more data, the mean difference in selling price would likely be different). This is a problem in inferential statistics, based upon a desire to make an inference from a sample. We will return to this problem later in the book, and we will discuss how a critical difference threshold may be set: if the observed difference is below this threshold, we settle on conclusion (b); if the difference is above the threshold, we decide on option (a) above.

1.4.4.2 Do two places differ in terms of air quality?

Suppose we are interested in comparing the particulate matter in two cities. We collect daily data on PM10 (particles of 10 micrometers or less in size). Further suppose that we

collect five daily samples in city A and five in city B, and these are designed to estimate the 'true' mean in each city. Table 1.3 presents the results.

The sample mean in city B is clearly higher than the mean in city A. But keep in mind that we have only taken a sample; there is certainly fluctuation from day to day, and so the 'true' means could possibly be the same (that is, if we took a very large sample over a very large number of days, the means could be equal).

We should not necessarily immediately conclude that city B has a higher 'true' mean particulate count; our results could be due to sampling fluctuations. Instead, we need to weigh the observed difference in the sample means against the difference in sample means that we might expect from sampling variation alone (when the true means are equal). If the observed difference in sample means is small relative to the difference that might be expected even when the true means are equal, we will accept the possibility that the true means are equal. On the other hand, if the observed difference in sample means is larger than the differences we would expect from such sampling fluctuations, we will conclude that the two cities have different levels of particulate matter. Details of problems like this (including the difference thresholds that must be set to distinguish between accepting and rejecting the idea that the means are equal) are covered in Chapter 5, which deals with questions of statistical inference.

Table 1.3   Hypothetical PM10 readings (units are micrograms per cubic meter)

City A	City B
40	45
38	41
52	59
35	34
26	25
Sample Mean:	**Sample Mean:**
38.2	40.8

1.4.4.3 Buffon's needle and migration distances

In the United States very little data are collected on the distances people move when they change their residential address. Yet this is a very basic measurement pertaining to a very important geographic process. Information is collected on the proportion of people changing their county of residence, and this may be used, together with concepts of prob-ability, to estimate migration distances.

We begin with the work of Buffon, a 17th-century naturalist. Buffon was interested in many topics, ranging from subjects in botany to the strength of ships at sea. He was also

interested in probability, and embedded in a supplement to the 4th volume of his 24–volume treatise on natural history is the following question.

Suppose we have a set of many parallel lines, separated by a constant distance, s. Now toss a needle of length L onto the set of parallel lines (see Figure 1.7). What is the probability that the needle will cross a line? Clearly, this probability will be higher as the length of the needle increases, and as the distance between the parallel lines decreases. Buffon found that the prob–ability (p) of a randomly tossed needle crossing the lines was $p = 2L/(\pi s)$. Buffon's needle was actually used at the time to estimate π; if a needle of known length is tossed many times onto a set of parallel lines separated by a known distance, one can calculate p as the ratio of the number of crossings to the number of tosses. The only remaining unknown in the equation is π. Beckmann (1971), for example, refers to a Captain Fox, who passed at least some of his time this way while recovering from wounds received in the US Civil War. Unfortunately,
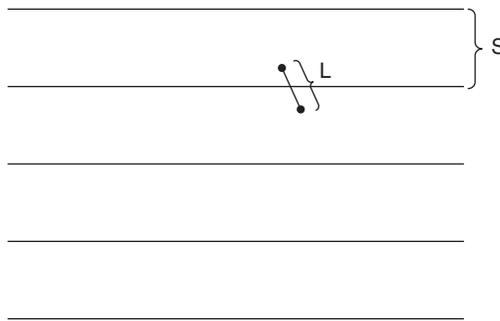

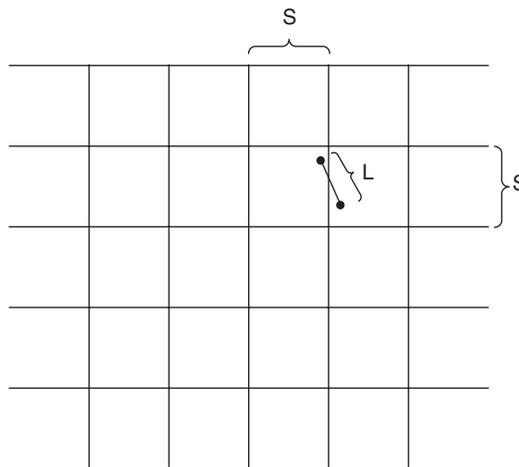**Figure 1.7**   Buffon's needle on a set of parallel lines

Figure 1.8   Buffon's needle on a square grid

the needle must be tossed a very large number of times to estimate π with any reasonable level of accuracy, and so such needle-tossing never developed into a popular pastime. Laplace generalized this to the case of a square grid (see Figure 1.8). When the side of a square is equal to s (and $L<s$), the probability of crossing a line is now equal to $p = (4Ls - L^2/(\pi s^2))$.

Let us now turn to the connection to migration distance estimation. Define the ends of the needle as the origin and destination of a migrant; we wish to estimate this unknown needle length (L), which corresponds to migration distance. We will make the assumption that counties are approximately square, and that they are all the same size (i.e., we assume that a county map will look roughly like a square grid). We can estimate the side of a square, s, as the square root of the average area of a county. We may also estimate p if we have data on the proportion of all migrants who change their county of residence when they relocate. Finally, we of course now know that the value of π is 3.14……. We can solve Laplace's equation for the unknown migration distance:

$$L = 2s - s\sqrt{4 - p\pi}$$

Using data from the United States, $p = 0.35$ and $s = 33$ miles, and therefore we estimate the average migration distance as approximately $L = 10$ miles. Despite the perception that long-distance moving is perhaps the norm, the majority of individuals move a short distance when they relocate.

Although the assumption of square counties of equal size is of course unreasonable, a primary objective of a model is to simplify reality. We do not expect or contend that counties are equal-sized squares. We could be more exact by performing an experiment where we toss needles of a given size down onto a map of US counties; by trying different needle sizes, we will eventually find one that gives us county-crossing probabilities equal to about $p = 0.35$. We do not evaluate this assumption in greater detail here, but it turns out that the assumption of square counties of equal size is relatively *robust* – the conclusion doesn't change much when the assumption does not quite hold. Instead, such an assumption allows us to get a reasonable estimate of average migration distances.

1.4.4.4 Why is the traffic moving faster in the other lane?

Almost all would agree that traffic seems to always move faster in the other lane. There have been several statistical explanations for this. These explanations include:

(a) Redelmeier and Tibshirani (2000) created a simulation where two lanes had identical characteristics, in terms of the number of vehicles and their average speed. The only difference in the two lanes was the initial spacing between vehicles. In the simulation, the hypothetical vehicles would accelerate when traveling slowly, and would decelerate when approaching too closely to the vehicle in front of them. Not surprisingly, while moving quickly, vehicles were relatively far from one another; while

moving slowly, they were closer together. Since the average speed in each lane was the same, and the number of cars in each lane was identical, each vehicle was passed by the same number of vehicles it had passed. However, the number of one-second time intervals during which a vehicle was passed was greater than the number of one-second intervals during which the vehicle was passing another vehicle. Thus, more time is spent being passed by other vehicles than is spent in passing vehicles (fast cars are spread out, and they are the ones overtaking you … you are passing the slow cars, which are bunched up, so that it doesn't take long to pass them).

(b) Bostrom (2001) has, on the surface, a simpler answer to the question – cars in the other lane *are* moving faster! If cars in the fast lane are more spread out, the density of cars will be greatest in the slow lane. Now if you randomly choose a car at any time, there is a relatively high probability it will be from the slow lane, since that is where the density of cars is highest. So, at any given time, most drivers are in fact in the slow lane, and cars in the other lane *are* in fact moving faster.

(c) Dawson and Riggs (2004) note that if you are traveling at just under or just over the speed limit, and if you accurately observe the speeds of the vehicles passing you as well as the speed of the vehicles you are passing, there will be a misperception of the true average speed. In particular, drivers traveling at just under the average speed will perceive traffic to be going faster than it really is, while drivers traveling at just over the average speed will perceive traffic to be slower than it really is. The reason has to do with the selection of vehicles whose speeds are being observed – this sample will be biased in the sense that it will include many of the very fast and very slow vehicles, but not many of the vehicles going at your own speed. Although Dawson and Riggs do not mention this, if the distribution of speeds is skewed such that more than half of the vehicles are going slower than the mean speed (a likely assumption), then more than half of the vehicles will perceive traffic to be faster than it really is.

1.5 DESCRIPTIVE AND INFERENTIAL METHODS

A key characteristic of geographic data that brings about the need for statistical analysis is that they may often be regarded as a sample from a larger population. *Descriptive* statistical analysis refers to the use of particular methods that are used to describe and summarize the characteristics of the sample, while *inferential* statistical analysis refers to the methods that are used to infer something about the population from the sample. Descriptive methods fall within the class of exploratory techniques, while inferential statistics lie within the class of confirmatory methods. Descriptive summaries of data may be either visual (e.g., in the form of graphs and maps), or numerical; the mean and median are examples of the latter.

To begin to better understand the nature of inferential statistics, suppose you are handed a coin, and you are asked to determine whether it is a 'fair' one (that is, the likelihood of a 'head' is the same as the likelihood of a 'tail'). One natural way to gather some information would be to flip the coin a number of times. Suppose you flip the coin ten times, and you observe heads eight times. An example of a descriptive statistic is the observed proportion of heads – in this case $8/10 = 0.8$. We enter the realm of inferential statistics when we attempt to pass judgment on whether the coin is 'fair'. We do this by *inferring* whether the coin is fair, on the basis of our sample results. Eight heads is more than the four, five, or six that might have made us more comfortable in a declaration that the coin is fair, but is eight heads really enough to say that the coin is *not* a fair one?

Table 1.4 Hypothetical outcome of 100 experiments of ten coin tosses each

No. of heads	Frequency of occurrence
0	0
1	1
2	4
3	8
4	15
5	22
6	30
7	8
8	8
9	3
10	1

There are at least two ways to go about answering the question of whether the coin is a fair one. One is to ask what *would* happen if the coin *was* fair, and to simulate a series of experiments identical to the one just carried out. That is, if we could repeatedly flip a known fair coin ten times, each time recording the number of heads, we would learn just how unusual a total of eight heads actually was. If eight heads comes up quite frequently with the fair coin, we will judge our original coin to be fair. On the other hand, if eight heads is an extremely rare event for a fair coin, we will conclude that our original coin is not fair.

To pursue this idea, suppose you arrange to carry out such an experiment 100 times. For example, one might have 100 students in a large class each flip coins that are known to be fair ten times. Upon pooling together the results, suppose you find the results shown in Table 1.4. We see that eight heads occurred 8% of the time.

We still need a guideline to tell us whether our observed outcome of eight heads should lead us to the conclusion that the coin is (or is not) fair. The usual guideline is to ask how likely a result equal to or more extreme than the observed one is, *if* our initial, baseline hypothesis that we possess a fair coin (called the *null* hypothesis) is true. A common practice is to not rule out the null hypothesis if the likelihood of a result at least as extreme as the one we observed is more than 5%. Here we would not rule out the null hypothesis of a fair coin if our experiment showed that eight or more heads occurred more than 5% of the time.

Alternatively, we wish to reject the null hypothesis that our original coin is a fair one if the results of our experiment indicate that eight or more heads out of ten is an uncommon event for fair coins. If fair coins give rise to eight or more heads less than 5% of the time, we decide to reject the null hypothesis and conclude that our coin is not fair.

In the example above, eight or more heads occurred 12 times out of 100, when a fair coin was flipped ten times. The fact that events as extreme, or more extreme than the one we observed, will happen 12% of the time with a *fair* coin leads us to accept the inference that our original coin is a fair one. Had we observed nine heads with our original coin, we would have judged it to be unfair, since events as rare or more rare than this (namely where the number of heads is equal to 9 or 10) occurred only four times in the 100 trials of a fair coin. Note, too, that our observed result does not prove that the coin *is* unbiased. It still *could* be unfair; there is, however, insufficient evidence to support the allegation.

The approach just described is an example of the *Monte Carlo method*, and several examples of its use are given in Chapter 10. A second way to answer the inferential problem is to make use of the fact that this is a *binomial* experiment; in Chapter 3, we will learn how to use this approach.

1.6 THE NATURE OF STATISTICAL THINKING

The American Statistical Association (1993, cited in Mallows 1998) notes that statistical thinking is:

   (a) the appreciation of uncertainty and data variability, and their impact on decision making, and

   (b) the use of the scientific method in approaching issues and problems.

Mallows (1998), in his Presidential Address to the American Statistical Association, argues that statistical thinking is not simply common sense, nor is it simply the scientific method. Rather, he suggests that statisticians give more attention to questions that arise in the beginning of the study of a problem or issue. In particular, Mallows argues that statisticians should: (a) consider what data are relevant to the problem; (b) consider how relevant data can be obtained; (c) explain the basis for all assumptions; (d) lay out the arguments on all

sides of the issue; and only then (e) formulate questions that can be addressed by statistical methods. He feels that too often statisticians rely too heavily on (e), as well as the actual use of the methods that follow. His ideas serve to remind us that statistical analysis is a comprehensive exercise – it does not consist of simply 'plugging numbers into a formula' and reporting a result. Instead, it requires a comprehensive assessment of questions, alternative perspectives, data, assumptions, analysis, and interpretation.

Mallows defines statistical thinking as that which 'concerns the relation of quantitative data to a real-world problem, often in the presence of uncertainty and variability. It attempts to make precise and explicit what the data has to say about the problem of interest.' Throughout the remainder of this book, we will learn how various methods are used and implemented, but we will also learn how to interpret the results and understand their limitations. Too often, students working on geographic problems have only a sense that they 'need statistics', and their response is to seek out an expert on statistics for advice on how to get started. The statistician's first reply should be in the form of questions: (1) What is the problem? (2) What data do you have, and what are their limitations? (3) Is statistical analysis relevant, or is some other method of analysis more appropriate? It is important for the student to think first about these questions. Perhaps a simple description will suffice to achieve the objective. Perhaps some sophisticated inferential analysis will be necessary. But the subsequent course of events should be driven by the substantive problems and questions of interest, as constrained by data availability and quality. It should not be driven by a feeling that one needs to use statistical analysis simply for the sake of doing so.

1.7 SPECIAL CONSIDERATIONS FOR SPATIAL DATA

Fotheringham and Rogerson (1993) categorize and discuss a number of general issues and characteristics associated with problems in spatial analysis. It is essential that those working with spatial data have an awareness of these issues. Although all of their categories are relevant to spatial *statistical* analysis, among those that are most pertinent are:

(a) the modifiable areal unit problem;

(b) boundary problems;

(c) spatial sampling procedures;

(d) spatial autocorrelation or spatial dependence.

1.7.1 The Modifiable Areal Unit Problem

The modifiable areal unit problem refers to the fact that results of statistical analyses are sensitive to the zoning system used to report aggregated data. Many spatial datasets are aggregated

into zones, and the nature of the zonal configuration can influence interpretation quite strongly. Panel (a) of Figure 1.9 shows one zoning system and panel (b) another, where the solid lines represent regional boundaries. The arrows represent the migration flows of individuals, and they are identical in each panel. The tail of an arrow represents a migrant's origin and the head represents a migrant's destination. In panel (a), no interzonal migration is reported, while a regional interpretation of panel (b) would lead to the conclusion that there was a strong southward movement, since five migrations from one zone to another would be reported. More generally, many of the statistical tools described in the following chapters would produce different results had different zoning systems been in effect.
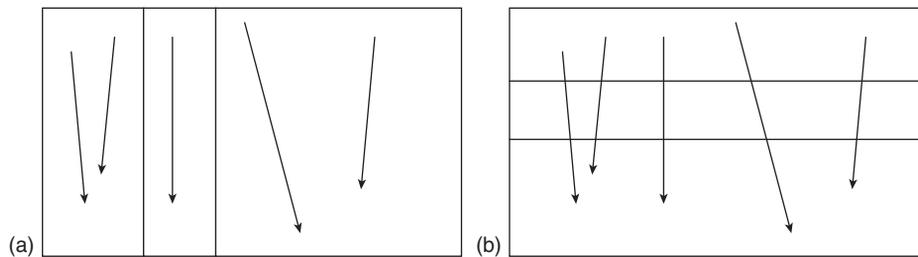
Figure 1.9   Two alternative zoning systems for migration data (note: arrows show origins and destinations of migrants)

The modifiable areal unit problem has two different aspects that should be appreciated. The first is related to the placement of zonal boundaries, for zones or subregions of a given size. If we were measuring mobility rates, we could overlay a grid of square cells on the study area. There are many different ways that the grid could be placed, rotated, and oriented on the study area. The second aspect has to do with geographic scale. If we replace the grid with another grid of larger square cells, the results of the analysis would be different. Migrants, for example, are less likely to cross cell boundaries in the larger grid than they are in the smaller grid.

As Fotheringham and Rogerson (1993) note, GIS technology now facilitates the analysis of data using alternative zoning systems, and it should become more routine to examine the sensitivity of results to modifiable areal units.

1.7.2 Boundary Problems

Study areas are bounded, and it is important to recognize that events just outside the study area can affect those inside of it. If we are investigating the market areas of shopping malls in a county, it would be a mistake to neglect the influence of a large mall located just outside the county boundary. One solution is to create a buffer zone around the area of study to include features that affect analysis within the primary area of interest. An example of the use of buffer zones in point pattern analysis is given in Chapter 10.

Both the size and shape of areas can affect measurement and interpretation. There are a lot of migrants leaving Rhode Island each year, but this is partially due to the state's small size – almost any move will be a move out of the state! Similarly, Tennessee experiences more out-migration than other states with the same land area in part because of its narrow rectangular shape. This is because individuals in Tennessee live, on average, closer to the border than do individuals in other states with the same area. A move of given length in some random direction is therefore more likely to take the Tennessean outside of the state.

1.7.3 Spatial Sampling Procedures

Statistical analysis is based upon sample data. Usually, one assumes that sample observations are taken randomly from some larger population of interest. If we are interested in sampling point locations to collect data on vegetation or soil, for example, there are many ways to do this. One could choose *x*- and *y*-coordinates randomly; this is known as a *simple random sample*. Another alternative would be to choose a *stratified* spatial sample, making sure that we chose a predetermined number of observations from each of several subregions, with simple random sampling within subregions. Alternative methods of sampling are discussed in more detail in Section 5.7.

1.7.4 Spatial Autocorrelation

Spatial autocorrelation refers to the fact that the value of a variable at one point in space is related to the value of that same variable in a nearby location. The travel behavior of residents in a household is likely to be related to the travel behavior of residents in nearby households, because both households have similar accessibility to other locations. Hence, observations of the two households are not likely to be independent, despite the requirement of statistical independence for standard statistical analysis. Spatial autocorrelation (or spatial dependence) can therefore have serious effects on statistical analyses, and hence can lead to misinterpretation. This is treated in more detail elsewhere in this book, including discussions in Chapters 5 and 10.

1.8 THE STRUCTURE OF THE BOOK

Chapter 2 covers methods of descriptive statistics – both visual and numerical approaches to describing data are covered. Chapters 3 and 4 provide the useful background on probability that facilitates understanding of inferential statistics. Inference about a population from a sample is carried out by first using the sample to make estimates of population characteristics. For example, a sample of individuals may be asked questions about their income; the sample mean provides an estimate of the unknown mean income of the entire population under study. Chapter 5 provides details on how these sample estimates can be used – to both construct confidence intervals that contain the true population value with

a desired probability, and to formally test hypotheses about the population values. The chapter also contains details on the nature of sampling and the choice of an appropriate sample size.

Chapter 5 also contains descriptions of hypothesis tests designed to determine whether two populations could conceivably have the same population characteristic. For example, the two-sample difference of means test focuses upon the possibility that two samples could come from populations that have identical means (this objective was illustrated in the examples in sections 1.4.4.1 and 1.4.4.2). Chapter 6 covers the method of analysis of variance, which extends these two-sample tests to the case of more than two samples. For example, data on travel behavior (e.g., distance traveled to a public facility, such as parks or libraries) may be available for five different geographic regions, and it may be of interest to test the hypothesis that the true mean distance traveled was the same in all regions. In Chapter 7, we begin our exploration of methods that focus upon the relationship between two or more variables. Chapter 7 introduces the methods of correlation, and Chapter 8 extends this introduction to the topic of simple linear regression, where one variable is hypothesized to depend linearly on another. Regression is the most widely used method of inferential statistics, and it is given additional coverage in Chapter 9, where the linear dependence of one variable on more than one other variable (i.e., multiple linear regression) is treated.

One of the basic questions geographers face is whether geographic data exhibit spatial patterns. This is important both in its own right (where, e.g., we may wonder whether crime locations are more geographically clustered than they were in the past), and in addressing the fundamental problem of spatial dependence in geographic data when carrying out statistical tests. With respect to the latter, inferential statistical tests almost always assume that data observations are independent; this, however, is often not the case when data are collected at geographic locations. Instead, data are often spatially dependent – the value of a variable at one location is likely to be similar to the value of the variable at a nearby location. Chapter 10 is devoted to methods and statistical tests designed to determine whether data exhibit spatial patterns. Chapter 11 returns to the topic of regression, focusing upon how to carry out analyses of the dependence of one variable on others, when such spatial dependence in the data is present.

Finally, it is often desirable to summarize large datasets containing many observations and large numbers of variables. For example, it is often difficult to know where to begin when using census data for many different subregions (e.g., census tracts) to summarize the nature of a geographic region, in part because there are so many variables and many different subregions. Chapter 12 introduces factor analysis and cluster analysis as two approaches to summarizing large datasets. Factor analysis reduces the original number of variables to a smaller number of underlying dimensions or factors, and cluster analysis places the observations (i.e., the data for particular geographic subregions) into categories or clusters. The Epilogue contains some closing thoughts on new directions and applications.

1.9 DATASETS

1.9.1 Introduction

Some of the exercises at the end of each chapter make reference to various datasets. These datasets are available at the website associated with this book. Clearly the exercises constitute only a very small number of the many questions that could be asked, and the reader is encouraged to explore these datasets in more detail. The subsequent subsections describe these datasets.

1.9.2 Home Sales in Milwaukee, Wisconsin, USA in 2012

The City of Milwaukee (no date) maintains a database of home sales that is available to the public. It is accessible at http://assessments.milwaukee.gov/mainsales.html .

The dataset used here is an *Excel* file consisting of the 1,449 home sales reported in calendar year 2012. Each record, or row of the dataset, consists of the following column variables:

1. number of stories;
2. size of house, as measured by the number of finished square feet;
3. number of bedrooms;
4. number of bathrooms;
5. size of lot (in square feet);
6. date of sale;
7. sales price, in dollars;
8. political district (based on alderman district and represented by a number from 1 to 15;
9. age of house;
10. whether there is a full basement – 1 = yes; 0 = no;
11. whether there is an attic – 1 = yes; 0 = no;
12. whether there is a fireplace – 1 = yes; 0 = no;
13. whether the house is air-conditioned – 1 = yes; 0 = no;
14. whether the house has a garage – 1 = yes; 0 = no;
15. x-coordinate of house location;
16. y-coordinate of house location.

The complete online dataset covers the period from 2002 to present (2013), and includes information on a small number of other variables as well (taxkey number, address, style, type of exterior, number of half and full baths, and neighborhood number).

1.9.3 Singapore Census Data for 2010

This dataset consists of various census variables for the 36 planning areas of Singapore. The variables, which constitute the columns of the dataset, are as follows:

Name of planning area, total population, male population, female population, population over 65; Chinese population, Malay population, Indian population; population 5 and over, percent 5 and over who speak English, students 5 and over, students 5 and over with a long commute; population 15 and over, population over 15 who are in the labor force, unemployment rate, illiteracy rate among those 15 and over, population in various religious categories: not religious, Buddhist, Tao, Islam, Hindu, Sikh, Catholic, Other Christian, percent of those over 15 with no schooling, percent with university degree; population of working persons with monthly income under 1,000 Singapore dollars, population of working persons with monthly income over 8,000 Singapore dollars, fraction of those in the labor force with a long commute; total households, and percentage of householders who rent.

Many more variables are accessible at http://www.singstat.gov.sg/statistics/browse_by_theme/geospatial.html .

1.9.4 Hypothetical UK Housing Prices

This file is an *SPSS* formatted file consisting of 500 cases (rows) and 16 variables (columns). Each row represents a hypothetical home that has been sold, and variables consist of a mixture of regional location, housing attributes, and census attributes for the subregion that the house is located in.

1.9.4.1 Definition of variables

1. **region**   a number between one and six, representing the region in which the house is located.

2. **price**   price the house sold for in £.

3. **garage**

a dummy variable which takes the value:

1 if a garage is present

0 if a garage is not present.

4. **bedrooms**

number of bedrooms.

5. **bathrooms**

number of bathrooms.

6. **datebuilt**

year in which house was built.

7. **floor area**

floor area of the house in square meters.

8. **detached**

a dummy variable which takes the value:

1 if the house is detached

0 othewise.

9. **fireplace**

a dummy variable which takes the value:

1 if a fireplace is present

0 if a fireplace is not present.

10. **age<15**   percentage of subregional population age less than 15.

11. **age65+**   percentage of subregional population age 65 and over.

12. **nonwhite**   percentage of subregional population non-white.

13. **unemploy**   percentage of subregional population unemployed.

14. **ownocc**   percentage of subregional housing that is owner-occupied.

15. **carsperhh**   average number of cars per household in subregion.

16. **manuf**   percentage of subregional population employed in manufacturing.

1.9.5 1990 Census Data for Erie County, New York

A 235 × 5 data table was constructed by collecting (from the 1990 US Census) and deriving the following information for the 235 census tracts in Erie County, New York (variable labels are in parentheses):

(a) Median household income (medhsinc).

(b) Percentage of households headed by females (femaleh).

(c) Percentage of high-school graduates who have a professional degree (educ).

(d) Percentage of housing occupied by owner (tenure).

(e) Percentage of residents who moved into their present dwelling before 1959 (lres).