GEOGRAPHIC INFORMATION ANALYSIS

David O'Sullivan and
David Unwin

JOHN WILEY & SONS, INC.

Appendix A

The Elements of Statistics

A.1. INTRODUCTION

This appendix is intended to remind you of some of the basic ideas, concepts, and results of classical statistics, which you may have forgotten. If you have never encountered any of these ideas before, this is *not* the place to start—you really need to read one of the hundreds of introductory statistics texts or take a class in introductory statistics. If you have taken any introductory statistics class, what follows should be reasonably familiar. Most of the information in this appendix is useful background for the main text, although you can probably survive without a detailed, in-depth knowledge of it all. A good geographical introduction to many of these ideas, which also introduces some of the more spatial issues that we focus on in this book, is Peter Rogerson's *Statistical Methods for Geography* (2001).

We may introduce different terminology and symbols from those you have encountered elsewhere, so you should get used to those used here, as they appear in the main text. Indeed, the presentation of this book may be more mathematical in places than you are used to, so we start with some notes on mathematical notation. This is not intended to put you off, and it really shouldn't. Many of the concepts of spatial analysis are difficult to express concisely without mathematical notation. Therefore, you will get much further if you put a little effort into coming to grips with the notation. The effort will also make it easier to understand the spatial analysis literature, since journal articles and most textbooks simply assume that you know these things. They also tend to use slightly different symbols each time, so it's better if you have an idea of the principles behind the notation.

Preliminary Notes on Notation

A single instance of some variable or quantity is usually denoted by a *lowercase italicized* letter symbol. Sometimes the symbol might be the initial letter of the quantity we're talking about, say h for height or d for distance. More often, in introducing a statistical measure, we don't really care what the numbers represent, because they could be *anything*, so we just use one of the commonly used mathematical symbols, say x or y. Commonly used letters are x, y, z, n, m, and k. In the main text, these occur frequently and generally have the meanings described in Table A.1. In addition to these six, you will note that d, w, and $\mathbf{s}$ also occur frequently in spatial analysis. The reason for the use of boldface type for $\mathbf{s}$ is made clear in Appendix B, where vectors and matrices are discussed.

A familiar aspect of mathematical notation is that letters from the Greek alphabet are used alongside the Roman alphabet letters that you are used to. You may already be familiar with mu (μ) for a population mean, sigma (σ) for population standard deviation, chi (χ) for a particular statistical distribution, and pi (π) for ... well, just for pi. In general, we try to avoid using any more Greek symbols than these. The reason for introducing symbols is so that we can use mathematical notation to talk about related values or to indicate mathematical operations that we want to perform on sets of values. So if h (or z) represents our height values, h^2 (or z^2) indicates height value squared. The symbols are a very concise way of saying the same thing, and that's very important when we describe more complex operations on data sets.

Two symbols that you will see a lot are i and j. However, i and j normally appear in a particular way. To describe complex operations on sets of

Table A.1 Commonly Used Symbols and Their Meaning in This Book

Symbol	*Meaning*
x	The "easting" geographical coordinate or a general data value
y	The "northing" geographical coordinate or a general data value
z, a, b	The numerical value of some measurement recorded at the geographical coordinates (x, y)
n, m	The number of observations in a data set
k	Either an arbitrary constant or, sometimes, the number of entities in a spatial neighborhood
d	Distance
w	The strength or weight of interaction between locations
$\mathbf{s}$	An arbitrary (x, y) location

values, we need another notational device: *subscripts*. Subscripts are small italic letters or numbers below and to the right of normal mathematical symbols: The i in z_i is a subscript. A subscript is used to signify that there may be more than one item of the type denoted by the symbol, so z_i stands in for a series or set of z values: z_1, z_2, z_3, and so on. This has various uses:

- A set of values is written between braces, so that $\{z_1, z_2, \ldots, z_{n-1}, z_n\}$ tells us that there are n elements in this set of z values. If required, the set as a whole may be denoted by a capital letter: Z. A typical value from the set Z is denoted z_i and we can abbreviate the previous partial listing of z's to simply $Z = \{z_i\}$, where it is understood that the set has n elements.
- In spatial analysis, it is common for the subscripts to refer to locations at which observations have been made and for the same subscripts to be used across a number of different data sets. Thus, h_7 and t_7 refer to the values of two different observations—say, height and temperature—at the same location (i.e., location number 7).
- Subscripts may also be used to distinguish different calculations of (say) the same statistic on different populations or samples. Thus, μ_A and μ_B would denote the means of two different data sets, A and B.

The symbols i and j usually appear as subscripts in one or other of these ways. A particularly common usage is to denote summation operations, which are indicated by use of the $\sum$ symbol (another Greek letter, this time capital sigma). This is where subscripts come into their own, because we can specify a range of values that are summed to produce a result. Thus, the sum

$$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 \tag{A.1}$$

is denoted

$$\sum_{i=1}^{i=6} a_i \tag{A.2}$$

indicating that summation of a set of a values should be carried out on all the elements from a_1 to a_6. For a set of n "a" values this becomes

$$\sum_{i=1}^{i=n} a_i \tag{A.3}$$

which is usually abbreviated to either

$$\sum_{i=1}^{n} a_i \tag{A.4}$$

or to

$$\sum_{i} a_i \tag{A.5}$$

where the number of values in the set of 'a's is understood to be n. If instead of the simple sum we wanted the sum of the squares of the a values we have

$$\sum_{i=1}^{n} a_i^2 \tag{A.6}$$

instead. Or perhaps we have two data sets, A and B, and we want the sum of the products of the a and b values at each location. This would be denoted

$$\sum_{i=1}^{n} a_i b_i \tag{A.7}$$

In spatial analysis, more complex operations might be carried out *between* two sets of values, and we may then need two summation operators. For example,

$$c = k \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i - z_j)^2 \tag{A.8}$$

indicates that c is to be calculated in two stages. First, we take each z value in turn (the outer i subscript) and sum the square of its value minus every z value in turn (the j subscript). You can figure this out by imagining first setting i to 1 and calculating the inner sum, which would be $\sum_j (z_1 - z_j)^2$. We then set i to 2 and do the summation $\sum_j (z_2 - z_j)^2$, and so on, all the way to $\sum_j (z_n - z_j)^2$. The final double summation is the sum of all of these individual sums, and c is equal to this sum multiplied by k. This will seem complex at first, but you will get used to it.

In the next section you will see immediately how these notational tools make it easy to write down operations like finding the mean value of a data set. Other elements of notation will be introduced as they are required and explained in the appendices and main text.

A.2. DESCRIBING DATA

The most fundamental operation in statistics is describing data. The measures described below are commonly used to describe the overall characteristics of a set of data.

Population Parameters

These are presented without comment. A *population mean* μ is given by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{A.9}$$

population variance σ^2 is given by

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (a_i - \mu)^2 \tag{A.10}$$

and *population standard deviation* σ is given by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (a_i - \mu)^2} \tag{A.11}$$

These statistics are referred to as *population parameters.*

Sample Statistics

The statistics above are based on the entire population of interest, which may not be known. Most descriptive statistics are calculated for a *sample* of the entire population. They therefore have two purposes: First, they are *summary descriptions* of the sample data, and second, they serve as *estimates* of the corresponding population parameters. The *sample mean* is

$$\bar{a} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{A.12}$$

the *sample variance* s^2 is

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (a_i - \bar{a})^2 \tag{A.13}$$

and the *sample standard deviation* s is given by

$$s = \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2} \tag{A.14}$$

In all these expressions the ^ ("hat") symbol indicates that the expression is an *estimate* of the corresponding population parameter. Note the different symbols for these sample statistics relative to the corresponding population statistics.

Sample statistics may be used as *unbiased estimators* of the corresponding population parameters, and a major part of inferential statistics is concerned with determining how good an estimate a sample statistic is of the corresponding population parameter. Note particularly the denominator $(n-1)$ in the sample variance and standard deviation statistics. This reflects the loss of one *degree of freedom* (df) in the sample statistic case because we know that $\sum(a_i - \bar{a}) = 0$, so that given the values of $\bar{a}$ and $a_1 \cdots a_{n-1}$, the value of a_n is known. The expressions shown, using the $(n-1)$ denominator, are known to produce better estimates of the corresponding population parameters than those obtained using n.

Together, the mean and variance or standard deviation provide a convenient summary description of a data set. The mean is a *measure of central tendency* and is a typical value somewhere in the middle of all the values in the data set. The variance and standard deviation are *measures of spread*, indicative of how dispersed are the values in the data set.

Z-Scores

The z-*score* of a value a_i relative to its population is given by

$$z_i = \frac{a_i - \mu_A}{\sigma} \tag{A.15}$$

and relative to a sample by

$$z_i = \frac{a_i - \bar{a}}{s} \tag{A.16}$$

The z-score indicates the place of a particular value in a data set relative to the mean, standardized with respect to the standard deviation. $z = 0$ is equivalent to the sample mean, $z > 0$ is a value greater than the mean and $z < 0$ is less than the mean. The z-score is used extensively in determining *confidence intervals* and in assessing *statistical significance*.

Median, Percentiles, Quartiles, and Box Plots

Other descriptive statistics are based on sorting the values in a data set into numerical order and describing them according to their position in the ordered list. The first *percentile* in a data set is the value below which 1%, and above which 99% of the data values are found. Other percentiles are defined similarly, and certain percentiles are frequently used as summary statistics. The 50th percentile is the *median*, sometimes denoted M. Half the values in a data set are below the median and half are above. Like the mean, the median is a measure of central tendency. Comparison of the mean and median may indicate whether or not a data set is *skewed*. If $\bar{a} > M_A$, this indicates that high values in the data set are pulling the mean above the median; such data are *right skewed*. Conversely, if $\bar{a} < M_A$, a few low values may be 'pulling' the mean below the median and the data are *left skewed*.

Skewed data sets are common in human geography. A good example is often provided by ethnicity data in administrative districts. For example, Figure A.1 is a *histogram* for the African-American percentage of population in the 67 Florida counties as estimated for 1999. The strong right skew in these data is illustrated by the histogram, with almost half of all the counties having African-American populations of 10% or less. The right skew is confirmed by the mean and median of these data. The median percent African-American is 11.65%, whereas the mean value is higher at 14.17% and the small numbers of counties with higher percentages of African-Americans pull the mean value up relative to the median. The

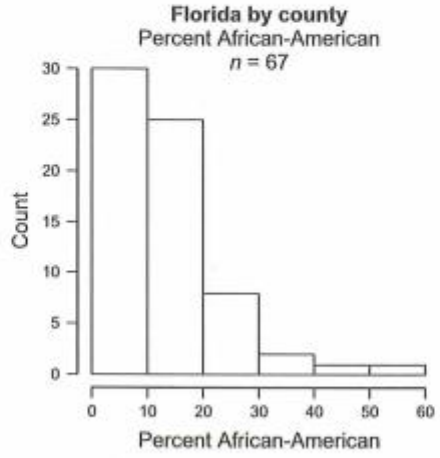

Figure A.1 Example of right skewed data in human geography.

median often gives a better indication of what constitutes a typical value in a data set.

Two other percentiles are frequently reported. These are the *lower* and *upper quartiles* of a data set, which are the 25th and 75th percentiles, respectively. If we denote these values by Q_{25} and Q_{75} respectively, the interquartile range (IQR) of a data set is given by

$$\text{IQR} = Q_{75} - Q_{25} \tag{A.17}$$

The interquartile range contains half the data values in a data set and is indicative of the range of values. The interquartile range, as a measure of data spread, is less affected by extreme values than are simpler measures such as the range (the maximum value minus the minimum value) or even the variance and standard deviation.

A useful graphic that gives a good summary picture of a data set is a *box plot*. A number of variations on the theme exist (so that it is important to state the way in which any plot that you present is defined), but the diagram on the left-hand side of Figure A.2 is typical. This plot summarizes the same Florida percent African-American data as in Figure A.1. The *box* itself is drawn to extend from the lower to the upper quartile value. The horizontal line near the center of the box indicates the median value. The *whiskers* on the plot extend to the lowest and highest data values within one-and-a-half IQRs below Q_{25} and above Q_{75}. Any values beyond these limits, that is, less than $Q_{25} - 1.5(\text{IQR})$ or greater than $Q_{75} + 1.5(\text{IQR})$, are regarded as *outliers* and marked individually with point symbols. If either the minimum or maximum data value lies inside the 1.5 IQR limits, the *fences* at the ends of the whiskers are drawn at the minimum or maximum value as appropriate. This presentation gives a good general picture

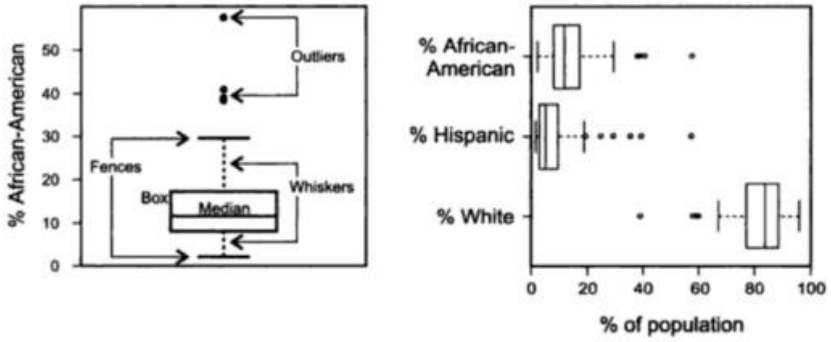

Figure A.2 Box plots of Florida ethnicity data.

of the data distribution. In the example illustrated on the left, the minimum value is greater than $Q_{25} - 1.5(\text{IQR})$, at around 2%, and is marked by the lower fence. There are four outlier values above $Q_{75} + 1.5(\text{IQR})$, three at around 40%, and one at around 55%.

Several data sets may be compared using parallel box plots. This has been done on the right-hand side of Figure A.2, where Florida county data for percent Hispanic and percent white have been added to the plot. Note that box plots may also be drawn horizontally, as here. This example shows that the African-American and Hispanic population distributions are both right skewed, although the typical variability among African-American populations is greater. The white population distribution is left skewed, on the other hand, and has higher typical values.

A.3. PROBABILITY THEORY

A great deal of statistics depends on the ideas of probability theory. Probability theory is a mathematical way of dealing with unpredictable events. It enables us to assign probabilities to events on a scale from 0 (will never happen) to 1 (will definitely happen). The most powerful aspect of probability theory is that it provides standard ways of calculating the probability of complex composite events—for example, A and B happening when C does not happen—given estimates for the probability of each of the individual events A, B, and C happening on its own.

In probability theory, an *event* is a defined as a collection of observations in which we are interested. To calculate the probability of an event, we first *enumerate* all the possible observations and count them. Then we determine how many of the possible observations satisfy the conditions for the event we are interested in to have occurred. The probability of the event is the number of outcomes that satisfy the event definition, divided by the total number of possible outcomes. For example, the probability of you winning the big prize in a lottery is given by

$$P(\text{lottery win}) = \frac{\text{number of ways you win}}{\text{number of combinations that could come up}} \qquad \text{(A.18)}$$

Note that the notation $P(A)$ is read as "the probability of event A occurring." Since the number of possible ways that you can win (with one ticket) is 1, and the number of possible combinations of numbers that could be drawn is usually very large (in the UK national lottery, it is 13,983,816), the probability of winning the lottery is usually very small. In the UK national lottery it is

$$P(\text{lottery win}) = \frac{1}{13{,}983{,}816} = 0.000000071511 \tag{A.19}$$

which is practically 0, meaning that it probably won't be you.

Here are some basic probability results with which you should be familiar. For an event A and its complement NOT A, the total probability is always 1:

$$P(A) + P(\text{NOT } A) = 1 \tag{A.20}$$

You can think of this as the *something must happen rule*, because it follows from the fact that a well-defined event A will either happen or not happen, since it can't "sort of" happen. This rule is probably obvious, but it is useful to remember, because it is often easier to enumerate observations that *do not* constitute the event of interest occurring than those which do, that is, to calculate $P(\text{NOT } A)$, from which it is easy to determine $P(A)$. An example of this is: What is the probability of any two students in a class of 25 sharing a birthday? This is a hard question until you realize that it is easier to calculate the opposite probability—that no two students in the class share the same birthday. Each student after the first can have a birthday on one of only 365, then 364, then 363, and so on, of the remaining "unused" days in the year if they are not to share a birthday with a student considered previously. This gives the probability that *no* two students share a birthday as $(365/366) \times (364/366) \times \cdots \times (342/366) \approx 0.432$, so that the probability that any two students *will* share a birthday is $1 - 0.432 = 0.568$.

For two events A and B, the probability of *either* event occurring, denoted $P(A \cup B)$ is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{A.21}$$

where $P(A \cap B)$ denotes the probability of *both A and B* occurring. In the special case where two events are *mutually exclusive* and cannot occur together, $P(A \cap B) = 0$, so that

$$P(A \cup B) = P(A) + P(B) \tag{A.22}$$

For example, if A is the event "drawing a face card from a deck of cards", and B is the event "drawing an ace from a deck of cards," the events are mutually exclusive, since a card cannot be an ace and a face card simultaneously. We have $P(A) = \frac{3}{13}$, $P(B) = \frac{1}{13}$, so that $P(A \cup B) = \frac{4}{13} = 30.8\%$. On the other hand, if A is the event "drawing a red card from a deck of cards" and B is as before, A and B are no longer mutually exclusive, since there are

two red aces in the pack. The various probabilities are now $P(A) = \frac{26}{52} = \frac{1}{2}$, $P(B) = \frac{1}{13}$, and $P(A \cap B) = \frac{2}{52} = \frac{1}{26}$, so that $P(A \cup B)$, the probability of drawing a card that is either red or an ace is $\frac{1}{2} + \frac{1}{13} - \frac{1}{26} = \frac{7}{13} = 53.8\%$.

Conditional probability refers to the probability of events given certain preconditions. The probability of A given B, written $P(A : B)$ is given by

$$P(A : B) = \frac{P(A \cap B)}{P(B)} \tag{A.23}$$

This is obvious if you think about it. If B must happen, $P(B)$ is proportional to the number of all possible outcomes. Similarly, $P(A \cap B)$ is proportional to the number of events that count, that is, those where A has occurred given that B has also occurred. Equation (A.23) then follows as a direct consequence of our definition of probability.

A particularly important concept is *event independence*. Two events are independent if the occurrence of one has no effect at all on the likelihood of the other. In this case,

$$\begin{aligned} P(A : B) &= P(A) \\ P(B : A) &= P(B) \end{aligned} \tag{A.24}$$

Inserting the first of these into equation (A.23), gives us

$$P(A) = \frac{P(A \cap B)}{P(B)} \tag{A.25}$$

which we can rearrange to get the important result for independent events A and B that

$$P(A \cap B) = P(A)P(B) \tag{A.26}$$

This result is the basis for the analytic calculation of many results for complex probabilities and one reason why the *assumption of event independence* is often important in statistics. As an example of independent events, think of two dice being rolled simultaneously. If event A is "die 1 comes up a six" and event B is "die 2 comes up a six," the events are independent, since the outcome on one die can have no possible effect on the outcome of the other. Thus the probability that both dice come up six is $P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = 2.78\%$.

Calculation of Permutations and Combinations

A very common requirement in probability calculations is to determine the number of possible *permutations* or *combinations* of n elements in various situations. Permutations are sets of elements where the *order* in which they are arranged is regarded as significant, so that *ABC* is regarded as different from *CBA*. When we are counting combinations, *ABC* and *CBA* are equivalent outcomes. There are actually six permutations of these three elements: *ABC*, *ACB*, *BAC*, *BCA*, *CAB*, and *CBA*, but they all count as only one combination. The number of permutations of k elements taken from a set of n elements, without replacement, is given by

$$P_k^n = \frac{n!}{(n-k)!} \tag{A.27}$$

where $x!$ denotes the *factorial* of x and is given by $x \times (x-1) \times (x-2) \times \cdots \times 3 \times 2 \times 1$, and $0!$ is defined equal to 1. The equivalent expression for the number of combinations of k elements, which may be chosen from a set of n elements is

$$C_k^n = \binom{n}{k} = \frac{n!}{k!\,(n-k)!} \tag{A.28}$$

These two expressions turn out to be important in many situations, and we will use the combinations expression to derive the expected frequency distribution associated with complete spatial randomness (see Chapter 3).

Word of Warning about Probability Theory

The power of probability theory comes at a price: We have to learn to think of events in a very particular way, a way that may not always be applicable. The problem lies in the fact that probability theory works best in a world of *repeatable* observations, the classic examples being the rolling of dice and the flipping of coins. In this world we assign definite probabilities to outcomes based on simple calculations (the probability of rolling a six is $\frac{1}{6}$, the probability of heads is $\frac{1}{2}$) and over repeated trials (many rolls of the die, or many flips of a coin) we expect outcomes to match these calculations. If they don't match, we suspect a loaded die or unfair coin. In fact, this is a very particular concept of probability. There are at least three distinct uses of the term:

1. *A priori* or *theoretical*, where we can precisely calculate probabilities ahead, based on the "physics." This is the probability associated with dice, coins, and cards.
2. *A posteriori* probability is often used in geography. The assumption is that historical data may be projected forward in time in a predictive way. When we go on a trip and consult charts of average July temperatures in California, we are using this type of probability in an informal way.
3. *Subjective probability* is more about hunches and guesswork. "The Braves have a 10% chance of winning the World Series this year," "Middlesbrough has a 1% chance of winning the FA Cup this season," or whatever.

There are, however, no hard-and-fast rules for distinguishing these different "flavors" of probability.

In the real world, especially in social science, data are once-off and observational, with no opportunity to conduct repeated trials. In treating sample observations as typical of an entire population, we make some important assumptions about the nature of the world and of our observations, in particular that the world is stable between observations and that our observations are a representative (random) sample. There are many cases where this cannot be true. The assumptions are especially dubious where data are collected for a localized area, because then the sample is only representative locally, and we must be careful about claims we make based on statistical analysis.

A.4. PROCESSES AND RANDOM VARIABLES

Probability theory forms a basis for calculation of the likely outcomes of processes. A process may be summarized by a random variable. Note that this does not imply that a process is random, just that its outcomes may be modeled as if it were. A random variable is defined by a set of possible outcomes $\{a_1, \ldots, a_i, \ldots, a_n\}$ and an associated set of probabilities $\{P(a_1), \ldots, P(a_i), \ldots, P(a_n)\}$. The random variable is usually denoted by a capital letter, say A, and particular outcomes by lowercase letters a_i. We then write $P(A = a_i) = 0.25$ to denote the probability that the outcome of A is a_i.

When A can assume one of a countable number of outcomes, the random variable is *discrete*. An example is the number of times we throw a six in 10 rolls of a die: the only possibilities are none, one time, two times, three times, and so on, up to 10 times. This is 11 possible outcomes in total. Where A can assume *any* value over some range, the random variable is continuous. A set of measurements of the height of students in a class can be regarded as a continuous random variable, since potentially any specific height in a range from, say, 1.2 to 2.4 m might be recorded. Thus, one student might be 1.723 m tall, while another could be 1.7231 m tall, and there are an infinite number of exact measurements that could be made. Many observational data sets are approximated well by a small number of mathematically defined random variables, or *probability distributions*, which are frequently used as a result. Some of these are discussed in the sections that follow.

The Binomial Distribution

The binomial distribution is a discrete random variable that applies when a series of trials are conducted where the probability of some event occurring in each individual trial is known, and the overall probability of some number of occurrences of the event is of interest. A typical example is the probability distribution associated with throwing x "sixes" when throwing a die n times. Here the set of possible outcomes is

$$A = \{0 \text{ sixes}, 1 \text{ six}, 2 \text{ sixes}, \ldots, n \text{ sixes}\} \tag{A.29}$$

and the binomial probability distribution will tell us what the probability is of getting a specified number of sixes.

The binomial probability distribution is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{A.30}$$

where p is the probability of the outcome of interest in each trial, there are n trials, and the outcome of interest occurs x times. x may take any value between 0 and n. For example, for the probability of getting two sixes in five rolls of a die, we have $n = 5$, $x = 2$, and $p = \frac{1}{6}$, so that

$$
\begin{aligned}
P(\text{two sixes}) &= \binom{5}{2}\left(\frac{1}{6}\right)^2\left(1-\frac{1}{6}\right)^{5-2} \\
&= \left(\frac{5!}{2!3!}\right) \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^3 \\
&= \left(\frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)}\right) \times \left(\frac{1}{6} \times \frac{1}{6}\right) \times \left(\frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}\right) \\
&= \left(\frac{120}{12}\right) \times \left(\frac{1}{36}\right) \times \left(\frac{125}{216}\right) \\
&= \frac{15{,}000}{93{,}312} \\
&= 0.16075
\end{aligned}
\tag{A.31}
$$

Figure A.3 shows the probabilities of the different numbers of sixes for five rolls of a die. You can see that it is more likely that we will roll no sixes or only one six than that we will roll two.

For statistical purposes it is useful to know the mean, variance, and standard deviation of a random variable. For a binomial random variable these are given by

$$
\begin{aligned}
\mu &= np \\
\sigma^2 &= np(1-p) \\
\sigma &= \sqrt{np(1-p)}
\end{aligned}
\tag{A.32}
$$

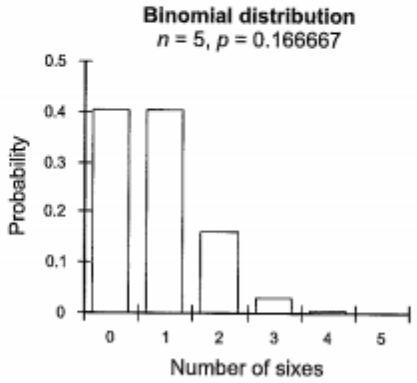

Figure A.3 Histogram showing the probabilities of rolling different numbers of sixes for five rolls of a die.

Applying these results we find that the mean or expected value when throwing a die 5 times and counting 'sixes' is

$$\mu = np = 5 \times \frac{1}{6} = 0.8333 \tag{A.33}$$

with a standard deviation of

$$\sigma = \sqrt{np(1-p)} = \sqrt{5 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{\frac{25}{36}} = \frac{5}{6} = 0.8333 \tag{A.34}$$

Note that we would never actually observe 0.833 six. Rather, this is the long-run average that we would expect if we conducted the experiment many times.

The Poisson Distribution

The Poisson distribution is useful when we observe the number of occurrences of an event in some fixed unit of area, length, or volume or over a fixed time period. The Poisson distribution has only one parameter λ, which is the average intensity of events (i.e., the mean number of events expected in each unit). This is usually estimated from the sample. The probability of observing x events in one unit is given by

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{A.35}$$

which has parameters

$$\begin{aligned} \mu &= \lambda \\ \sigma^2 &= \lambda \\ \sigma &= \sqrt{\lambda} \end{aligned} \tag{A.36}$$

Figure A.4 shows the probabilities for a Poisson distribution with $\lambda = 2$. This distribution is important in the analysis of point patterns (see Chapters 3 and 4).

Continuous Random Variables

Both the binomial and Poisson distributions are discrete variables, in which the meaning of the probability assigned to any particular outcome is

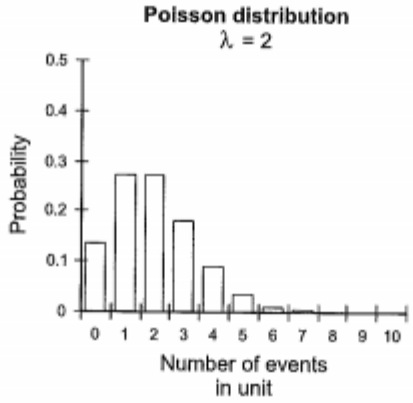

Figure A.4 Histogram of the Poisson distribution for $\lambda = 2$.

obvious. In the continuous case it is less so. For example, the chance that any student in a class will have a height of *precisely* 175.2 cm is very small, almost zero, in fact. We can only speak of a probability that a measurement will lie in some range of values. Continuous random variables are therefore defined in terms of a *probability density function*, which enables the calculation of the probability that a value between given limits will be observed.

The Uniform Distribution

In the uniform distribution, every outcome is equally likely over the range of possible outcomes. If we knew that the shortest student in a class was 160 cm tall, and the tallest 200 cm, and we thought that heights were uniformly distributed, we would have the continuous uniform distribution shown in Figure A.5. As shown in the diagram, the probability that a student's height is between any particular pair of values a and b is given by the area under the line between these values. Mathematically, this is expressed as

$$P(a \leq x \leq b) = \int_{x=a}^{x=b} f(x)\,\mathrm{d}x \tag{A.37}$$

where $f(x)$ is the probability density function. The units for the probability density therefore depend on the measurement units for the variable, and the area under the line must always total to 1 since something must occur with certainty. This is the only time you will even see an integration ($\int$) symbol in this book. The calculus to determine the area under standard continuous probability functions has already been done by others, and is

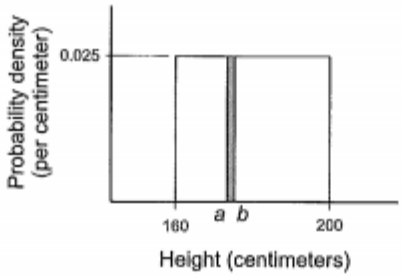

Figure A.5 Uniform distribution. The probability of a measurement between a and b is given by the area of the shaded rectangle.

recorded in statistical tables. In the next sections, two of the most frequently encountered and therefore most completely defined continuous random variables are described.

The Normal Distribution

It is unlikely that student heights are distributed uniformly. They are much more likely to approximate to a *normal distribution*. This is illustrated in Figure A.6. A particular normal distribution is defined by two parameters: its mean μ and its standard deviation σ. The probability density function is given by

$$\mathrm{N}(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{A.38}$$

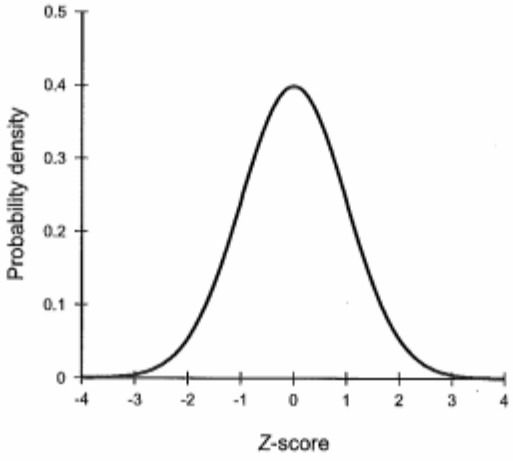

Figure A.6 Normal distribution.

where x is a particular value that the variable might take. The standardized form of this equation for a normal distribution with mean of 0 and standard deviation of 1 is denoted N(0, 1) and is given by

$$\mathrm{N}(z, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{A.39}$$

which shows how the probability of a normally distributed variable falling in any particular range can be determined from its z-score alone. Tables of the normal distribution are widely available that make this calculation simple. This, together with the central limit theorem (see Section A.5), is the reason for the importance of this distribution in statistical analysis. It is useful to know that 68.3% of the area under the normal curve lies within one standard deviation, 95.5% within two standard deviations, and 99.7% within three standard deviations of the distribution mean.

The Exponential Distribution

Many natural phenomena follow an approximately exponential distribution. A good example is the lengths of time between catastrophic events (earthquakes, floods of given severity). The formulas for the exponential distribution are

$$\begin{aligned} f(x) &= \frac{e^{-x/\theta}}{\theta} \\ \mu &= \theta \\ \sigma &= \theta \end{aligned} \tag{A.40}$$

where θ is a constant parameter that defines the distribution. The probability that a value higher than any particular value will be observed is conveniently calculated for the exponential distribution, according to

$$P(x \geq a) = e^{-a/\theta} \tag{A.41}$$

A.5. SAMPLING DISTRIBUTIONS AND HYPOTHESIS TESTING

We now come to one of the key ideas in statistics. A set of observations is often a sample of the population from which it is drawn. A voter survey is a good example. Often, a sample is the only feasible way to gather data. It

would not be practical to ask all the voters in the United States which way they intended to vote, on a daily basis, in the runup to a presidential election. Instead, polling organizations ask a sample of the population which way they intend to vote. They then determine statistics (mean, variance, etc.) from their sample in order to estimate the values of these parameters for the entire population.

If we imagine taking many different samples from a population and calculating (say) $\bar{a}$ in order to estimate the population mean μ, we get a different estimate of the population parameter from each different sample. If we record the parameter estimate $\hat{\mu} = \bar{a}$ from many samples, we get numerous estimates of the population mean. These estimates constitute a *sampling distribution* of the parameter in question, in this case the *sampling distribution of the mean*.

The Central Limit Theorem

The *central limit theorem* is a key result in statistics and allows us to say how good an estimate for a population parameter we can make given a sample of a particular size. According to the central limit theorem, given a random sample of n observations from a population with mean μ and standard deviation σ, then for sufficiently large n, the sampling distribution of $\bar{a}$ is normal with

$$\mu_{\bar{a}} = \mu \tag{A.42}$$

and

$$\sigma_{\bar{a}} = \frac{\sigma}{\sqrt{n}} \tag{A.43}$$

A number of points are significant here:

- The distribution of the population *does not matter* for large enough n. For almost any population distribution, $n \geq 30$ is sufficient to ensure that the sampling distribution is close to normal with the mean and standard deviation above.
- The sample *must be random*, so that every sample of size n has an equal chance of being selected.
- Since $\sigma_{\bar{a}} = \sigma/\sqrt{n}$, our estimate of the population mean is likely to be closer to the actual population parameter with a larger sample size. However, because the relationship goes with $\sqrt{n}$, it may be necessary to increase our sample size considerably to get much improvement in

the parameter estimate, since we must take a sample of $4n$ observations to halve the sampling distribution standard deviation.

- The central limit theorem applies to a good approximation to other measures of central tendency (such as the median). It does not necessarily apply to other statistics, so you may have to check statistics texts for other cases.

For the data below (a sample of $n = 30$ incomes), we get an estimate of the population mean income of $\hat{\mu} = \bar{a} = 25{,}057$.

16511	14750	21703	16496	32311	25186
32379	17822	17992	22862	39907	39043
15324	19889	32632	24706	38480	25227
34878	17898	16867	18644	20630	16132
36463	28714	18346	28398	25613	35908

We can estimate the population income standard deviation using the standard deviation of the sample (with $n - 1$ as denominator, remember) as $\hat{\sigma} = s_{\bar{a}} = 8147.981$. From the central limit theorem, we then know that the sampling distribution of the mean has a standard deviation given by

$$\sigma_{\bar{a}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{8147.981}{\sqrt{30}} = 1487.611 \tag{A.44}$$

Using this information, we know that if we were to take repeated samples of income data for this population, about 95% of our estimates of the mean income would fall within about two standard deviations of the estimate. In fact, we can say that the mean income of the population is $25{,}057 \pm 1.96 \times 1487.6 = 25{,}057 \pm 2916$ with 95% confidence. In other words, we are 95% confident that the mean income of the population we have taken this sample from is between 22,141 and 27,973.

Hypothesis Testing

The result above is the basis of a good deal of statistics. Given some question concerning a population, we formulate a *null hypothesis* that we wish to test. We then collect a sample from the population and estimate from the sample the population parameters required. The central limit theorem then allows us to place *confidence intervals* on our estimates of the population parameters. If our parameter estimates are not in agreement with the null hypothesis, we state that the evidence does not support the hypothesis at

the *level of significance* we have chosen. If the parameter estimates do agree with the null hypothesis, we say that the evidence is insufficient to reject the null hypothesis.

For example, in the income example above, we might have hypothesized that the mean income of the population is over 30,000. The evidence from our sample does not support this hypothesis, since we are 95% confident that the population mean lies between 22,141 and 27,973. In fact, given the sampling distribution of the mean, we can say how likely it is that the population mean is over 30,000; 30,000 has a z score of $(30{,}000 - 25{,}057)/1487.6 = 3.323$ relative to the sample mean. The probability of a z score of 3.323 or greater may be determined from tables of the normal distribution and is extremely low, at just 0.045%, or roughly 1 chance in 2000. We can say that the null hypothesis is rejected at the $p = 0.00045$ level. It is important to note that what we are really saying is that if we repeatedly take samples of $n = 30$ incomes from this population, only one in 2000 of the samples would give us an estimate of mean income greater than 30000. From this we deduce that it is extremely unlikely that the population mean really is 30000 or greater.

Note what would happen if our original sample were larger, with (say) 120 observations. If the sample mean and standard deviation were the same, a *fourfold* increase in sample size *halves* the standard deviation of the sampling distribution of the population mean, allowing us to halve the width of our 95% confidence interval on the population mean. This makes our estimate of the population mean more precise.

The procedure outlined here is the basis of most statistics. We have a question that we want to answer using whatever appropriate data we can obtain. Generally, it is impractical to collect complete data on the entire population, so we set up a hypothesis and gather a sample data set. The key statistical step is to determine the probabilities associated with the observed descriptive statistics derived from the sample. This is where the various probability distributions we have discussed come in, because many sampling distributions conform to one of the standard distributions discussed in Section A.4. In other cases it may be necessary to perform computer simulations to produce an empirical estimate of the sampling distribution. This is common in spatial analysis, where the mathematical analysis of sampling distributions is often very difficult, if not impossible. Having determined a sampling distribution for the descriptive statistics we are interested in, we can determine how likely the actual observed sample statistics are, given our hypothesis about the population. If the observed statistic is unlikely (usually, meaning less than 5% probability), we reject the hypothesis. Otherwise, we conclude that we cannot reject the hypothesis.

A.6. EXAMPLE

The ideas of data summary, postulating a null hypothesis, computation of a test statistic, location of this on an assumed sampling distribution assuming the null hypothesis to be true, and then making a decision as to how unusual the observed data are relative to this null hypothesis are best appreciated using a simple worked example. The chi-square (χ^2) test for association between two categorical variables illustrates the concepts well and can be developed from basic axioms of probability. It puts the null hypothesis up front and also leads easily to the distribution of the test statistic. This test is used in some approaches to point pattern analysis and hot-spot detection (see Chapters 4 and 5). Almost all other standard tests follow a similar approach, and details can be found in any statistical text. Chi-square just happens to be easy to follow and useful in many practical situations.

Step 1: Organize the Sample Data and Visualization

Suppose that as a result of a sample survey, you have a list of 110 cases, each having two *nominal attributes* describing it. A simple example might be the results of a point sampling spatial survey where at each point we recorded the geology and the maximum valley side angle of slope. The attribute "geology" is recorded simply as "soft rock" and "hard rock," and the slope angle, although originally measured in the field using a surveying level to the nearest whole degree, is coded as gentle (angles from flat to 5°), moderate (5° to 10°), and steep (>10°). Hence, a typical observation consists of codes for each of these attributes; for example, observation 1 had attributes H and M, indicating that it was hard rock with a moderate slope. The complete survey records the combined geology and slope for 110 sample locations, each determined by the spatial equivalent of simple random sampling. Our interest is to determine if slope varies with geology. The obvious research idea we might have is that harder rock is associated with the steeper slope angles, and vice versa. All 110 observations can be organized and visualized by forming a *contingency table,* in which rows represent one of the attributes (slope) and columns the other (geology). Each cell entry in the table is the count of cases in the sample that have that particular combination of unique conditions:

Observed Frequencies

	Geology		
Slope	*Soft rock*	*Hard rock*	*Totals*
Gentle	21	9	30
Moderate	18	11	29
Steep	18	33	51
Totals	57	53	110

Note that these entries are whole-number counts, not percentages, although most statistical analysis packages allow these to be calculated. There are, for example, just 21 cases in the data where the attributes are "gentle slope" and "soft rock," 9 cases of "gentle slope" and "hard rock," and so on. Of particular importance are the row and column totals, called the *marginal totals*. There were, for example, 30 cases of gentle slopes in the sample irrespective of geology and 57 sites on soft rock. By inspecting this table, we can see that there is a tendency for steeper slopes to be on hard rock, and vice versa. It is this idea that we test using the chi-square test of association for two qualitative variables.

Step 2: Devise a Test Statistic or Model

The key to the procedure adopted lies in the idea of testing not this rather vague hypothesis, which after all, does not say *how* the two variables are associated, but a more specific null hypothesis which says that the two variables are *not* associated. The idea is that we propose a *null hypothesis* and hope that we will disprove it. If it is disproved, the alternative research hypothesis (that the categories are associated) can be regarded as proven. It is conventional to use H_0 ("H nought") to refer to the null hypothesis and H_1 ("H one") for the alternative.

The chi-square statistic computes a measure of the difference between the cell frequencies in the observed contingency table and those that would be expected as long-run averages if the null hypothesis H_0 were true. It is here that basic probability theory comes into play. Consider the first cell of our table showing the number of cases where there was a gentle slope on soft rock. If our null hypothesis were true, what would we expect this number to be? We know from the laws of probability (see Section A.3) that if the two categories are independent, the probabilities involved should be multiplied together. In this case we need to know the probability of a case being on soft

rock and the probability of it being a gentle slope. In fact, we know neither, but we can *estimate* the probabilities using the marginal totals. First, we estimate the probability of a sample being on soft rock as the total observed (irrespective of slope) on soft rock, divided by the grand total as

$$P(\text{soft rock}) = 57/110 = 0.518 \tag{A.45}$$

Similar logic leads to the probability of a sample having a gentle slope:

$$P(\text{gentle slope}) = 30/110 = 0.273 \tag{A.46}$$

So if the two are really independent, as H_0 suggests, their *joint* probability is

$$\begin{aligned} P(\text{soft rock} \cap \text{gentle slope}) &= P(\text{soft rock}) \times P(\text{ gentle slope}) \\ &= 0.518 \times 0.272 \\ &= 0.1414 \end{aligned} \tag{A.47}$$

We can repeat this operation for all the remaining cells in the table. Given the total number of observed cases, this gives us a table of *expected* counts in each table cell, as follows:

Expected Frequencies

	Geology		
Slope	*Soft rock*	*Hard rock*	*Totals*
Gentle	15.55	14.45	30
Moderate	15.03	13.97	29
Steep	26.43	24.57	51
Totals	57	53	110

Because the expected values are long-run expected averages, not actual frequencies, they do not have to be whole numbers that could actually be observed in practice. Notice that calculation of the expected frequencies is easy to remember from the rule that the entry in a particular position is the product of the corresponding row and column totals divided by the total number of cases.

Now we have two tables, one the observed frequencies in our sample, the other the frequencies we would expect if the null hypothesis were true. The first we call the table of *observed frequencies*, the second the table of *expected frequencies*. Intuitively, if these tables are not much different,

we will be unable to reject the null hypothesis, whereas large discrepancies will lead us to reject the null hypothesis in favor of H_1, concluding that the variables are associated. What the formal statistical test does is to quantify these intuitions. An obvious thing to do if we are interested in the difference between two sets of numbers is to take their differences. This will give us some negative and some positive values, so we actually take the *squared* differences between the observed and expected frequencies $(O_{ij} - E_{ij})^2$, to get:

Squared Differences

	Geology	
Slope	*Soft rock*	*Hard rock*
Gentle	29.7205	29.7205
Moderate	8.8209	8.8209
Steep	71.0649	71.0649

These numbers sum to 219.18, but a moment's thought reveals that the this total depends as much on the number of cases in the sample as it does on the cell differences, so that the larger n is, the greater will be this measure of the difference between the observed and expected outcomes. The final step in constructing the chi-square test statistic is to divide each squared difference by its expected frequency, giving $(O_{ij} - E_{ij})^2/E_{ij}$, to standardize the calculation for any n.

$(O_{ij} - E_{ij})^2/E_{ij}$

	Geology	
Slope	*Soft rock*	*Hard rock*
Gentle	1.910	2.056
Moderate	0.587	0.675
Steep	2.689	2.892

The sum of these cell values, 10.809, is our final chi-square statistic. Given a set of individual cell observed frequencies O_{ij} and expected frequencies E_{ij}, the statistic is defined formally by the equation

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{A.48}$$

You should be able to see that this is an obvious and logical way of deriving a single number (a statistic) to measure the differences between two sets of frequencies. There is nothing mysterious about it.

Step 3: Study the Sampling Distribution of Our Test Statistic

Intuition tells us that big values of chi-square will indicate large discrepancies between observed and expected, and small numbers the reverse, but in terms of whether or not we reject the null hypothesis, how big is "big"? Now we come to the sampling distribution of chi-square. Note first that this cannot be a normal distribution because it must have a lower limit of zero, since no negative numbers are possible, and if the two sets of frequencies are the same, we sum a series of zeros. If we *assume* that the distribution of *differences* between a typical observed frequency and the expected value for the same cell is normal and then square these values, provided that we standardize for the numbers involved by dividing by the expected frequency, we can develop a standard chi-square distribution for a single cell. In fact, this is exactly the basis for tabulated values of the chi-squared statistic. Since the chi-square statistic is calculated by summing over a set of cells, it is parameterized by a related whole-number (integer) value called its *degrees of freedom*. For a contingency table, this is calculated from

$$\text{Degrees of freedom} = \text{df} = (\text{no. rows} - 1) \times (\text{no. columns} - 1) \tag{A.49}$$

As the table below shows, there is a different distribution of chi-square for each different value of this parameter. In this table, each value is the chi-square that must be exceeded at the given probability in order for any null hypothesis to be rejected. The listed value of χ^2 would occur with the listed probability, if H_0 were true.

	Probability		
df	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.34	16.27
4	9.49	13.28	18.46
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
10	18.31	23.21	29.59
20	31.41	37.57	45.32
30	43.77	50.89	59.70

Step 4: Locate the Observed Value of the Test Statistic in the Assumed Sampling Distribution and Draw Conclusions

So what can we conclude about the association between the hardness of the rock and the slope angles? Recall that our data formed a 3×2 contingency table and generated a chi-square test statistic of 10.809. The table has $(3 - 1) \times (2 - 1) = 2$ degrees of freedom. We now relate these numbers to the standard chi-square distribution. With $df = 2$, the critical value at $\alpha = 0.05$ is 5.99. This means that if there were no association between the attributes (i.e., if the null hypothesis were true) this value would be exceeded only five times in every hundred. Our value is higher than this. In fact, it is also higher than the $\alpha = 0.01$ critical value of 9.21. This tells us that if the null hypothesis were true, the chance of getting a chi-square value this high is low, at less than 1 in 100. Our choice is either to say, despite this, that we still think that H_0 is true, or more sensibly, to say that we must reject the null hypothesis in favor of the alternative (H_1). We have pretty good evidence that rock hardness and slope angle are associated in some way.

This may seem a rather formal and tedious way of proving something that is obvious right at the outset. Indeed, it is. However, the hypothesis test acts as a check on our speculations, providing a factual bedrock on which to build further, perhaps more scientifically interesting ideas (such as "why do we get steeper slopes on harder rocks?").

REFERENCE

Rogerson, P. A. (2001). *Statistical Methods for Geography*. London: Sage.