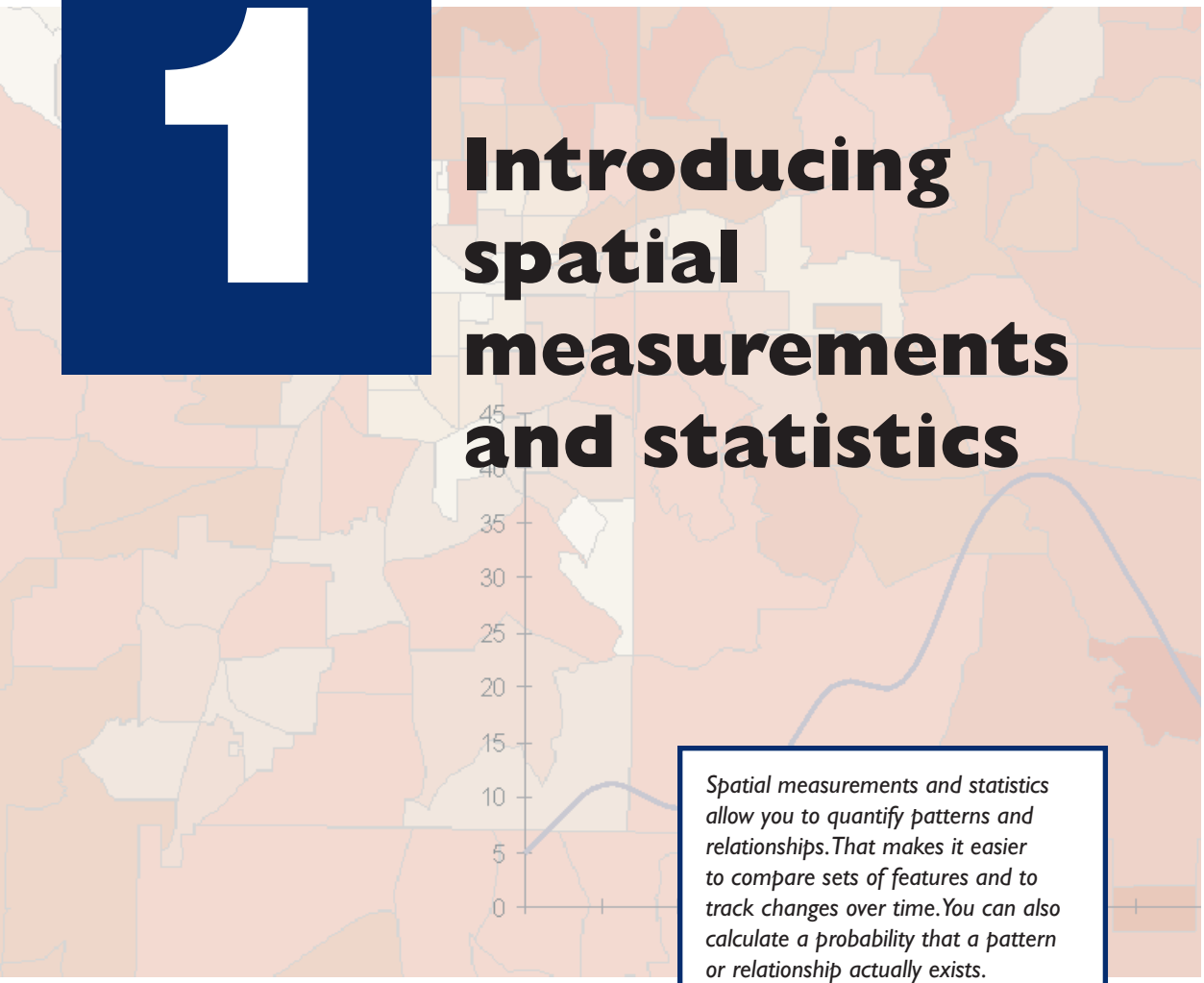
1 Introducing spatial measurements and statistics



Spatial measurements and statistics allow you to quantify patterns and relationships. That makes it easier to compare sets of features and to track changes over time. You can also calculate a probability that a pattern or relationship actually exists.

In this chapter:

- *What are spatial measurements and statistics?*
- *Geographic analysis with statistics*

GIS analysis is about getting answers to questions so you can make intelligent decisions. The previous book in this series showed you how to do GIS analysis with maps. In some cases, the map was the analysis. In other cases, you used GIS tools and methods to create new data that was then displayed on a map so you could analyze it and draw conclusions.

Sometimes, making a map may be enough to get the answers you need. But trying to draw conclusions from a map isn't always easy. How you classify and symbolize features and values on a map can obscure the information, and humans see patterns and relationships everywhere—even sometimes when they don't really exist.

Over the past 50 years or so, geographers, regional scientists, ecologists, economists, and others have developed tools to describe the distribution of a set of features, to discern patterns, and to measure relationships between features.

These tools rely on statistics to cut through the map display and get right at the patterns and relationships in the data. Space is a fundamental component of these statistics. That's what sets them apart from traditional statistics used to analyze aspatial data (tables of data values). The locations of the features and in many cases the spatial relationship between them (distance, for example) are considered, along with the attribute values associated with the features. If you just tried to analyze the attribute values by themselves, using traditional statistics, you'd get a false picture of what's occurring.

What if you could find the center of a group of influenza cases without guessing? Or clearly see the overall direction of a set of storm tracks? What if the GIS could identify clusters of burglaries for you?

Spatial statistics tools can help you perform these tasks, and others—tasks you may already be doing with maps. But spatial statistics open up a new set of questions you could be asking, to get even better information and be even more confident in your decisions: How sure am I that the pattern I'm seeing isn't simply due to a random occurrence? To what extent does the value of a feature depend on the values of surrounding features? How well does the value of one attribute predict the value of another? What are the trends in the data?

Statistics describe or summarize large amounts of data, useful in geographic analysis where you're often dealing with large datasets. Having a summary statistic—such as the center of features or the directional trend—makes it easy to compare sets of features or track changes over time without having to guess.

Statistics also let you derive information from a sample of features, and apply your conclusions to the whole set of features in your study area. If a sample of a plant species creates a clustered pattern, you can conclude that the species generally appears in clusters.

Statistics help you predict unknown values from known sample values. If you establish a relationship between feature values, you can predict where certain other values will occur. Knowing that landslides have occurred on slopes of a certain angle, soil moisture, and vegetation cover lets you find other slopes with these values and zone them as susceptible to landslides. The query capability of GIS—finding areas that match a set of criteria—lets you put the predictions to work.

Maybe most importantly, statistics let you verify your conclusions. You can assign a probability that your conclusions are true, and thus know how confident you can be in the decisions you make.

So why haven't people been using statistics for geographic analysis all along? One reason is that statistics are, after all, statistics—they're perceived as hard to understand and to use. Another is that statistical tools haven't been available in commercial GIS software. Many of the statistical tools have been used primarily in academic research, or limited to use in a specific discipline. People had to write their own software routines to perform their analyses. Recently, though, spatial statistics packages—such as CrimeStat® and SpaceStat™—have begun to appear. Spatial statistics tools are also appearing in comprehensive statistics software such as SAS® and S+®, and in commercial GIS, including ArcGIS® 9.

Because of the heretofore limited availability of these tools, most GIS users have not been aware of them and how they can be applied. That's where this book comes in. We want to introduce you to the most commonly used spatial statistical tools—and those most helpful to GIS users—and show how they can be applied in a range of disciplines, from crime analysis to habitat conservation. The ultimate goal of this book is to help you extract information that is already in your GIS database (in which you've undoubtedly already invested substantial amounts of time and money), but that might not be obvious simply by creating a map.

In this book we've identified some common questions that spatial statistics can answer.

How are the features distributed?

Statistics can describe the characteristics of a set of features, including the center of the features, the extent to which features are clustered or dispersed around the center, and any directional trend. Analyzing the distribution of features is useful for studying change over time—for example, to see where the center of cases of a particular disease moves over the course of several months—or for comparing two or more sets of features.

What is the pattern created by the features?

You can use statistics to measure whether—and to what extent—the distribution of features creates a pattern. If you find that cases of a disease form a clustered pattern, there are likely local sources (perhaps ponds harboring infected mosquitoes).

You can also identify patterns in the distribution of attribute values associated with the features. For example, you might calculate the degree to which student test scores in a city are clustered. If attendance areas with similarly high or low scores occur together, it may mean money and other resources are not being distributed equally.

Where are the clusters?

Finding individual clusters is useful when you need to take immediate action or when you want to find the cause of the cluster, so you know what action to take. A public health department would take immediate action to notify people living where a flu cluster has been identified to watch for symptoms. They could then try to identify the source of the outbreak—if it's a school, they would know to begin inoculating the children.

You can also use statistics to identify clusters of features with similar values. A tax assessor could create neighborhoods by identifying clusters of block groups with similar median house values.

What are the relationships between sets of features or values?

While the first three questions focus on the distribution of features in a single layer, this question deals with the relationships between two or more layers. You can determine if the features—or values associated with the features—occur together, and measure the strength of the relationship. A public health analyst could determine the extent to which economic or demographic factors and the quality of infant health are related in neighborhoods across a county. Once you've identified a relationship, you can predict where features or particular attribute values will occur.

The book assumes you have little or no knowledge of statistics, but some familiarity with GIS. Four sections that deal with general statistical concepts applicable throughout the book appear between chapters: "Understanding data distributions," "Testing statistical significance," "Defining spatial neighborhoods and weights," and "Using statistics with geographic data."

The emphasis in this book is on applying the statistical tools to get meaningful results, rather than on the mathematical theory behind the statistics. However, enough background and context is presented to understand the concepts behind the tools.

There are many spatial statistical tools and methods available, more than are discussed in this book. The ones included are those that are widely used and applicable to GIS analysis across a range of disciplines. Researchers continue to improve existing tools, and develop new ones, to better capture how geographic phenomena behave. The tools presented here represent current published versions. The references at the end of each chapter contain additional information on these tools, as well as others you might find useful.

A couple of related fields beyond the scope of this book are also worth exploring. One involves predicting values in spatially continuous data from a set of sample points (a field known as geostatistics). Geostatistics has primarily been used to study air pollution and soil contamination, and to explore for oil and natural gas, but has many other applications. Another related field involves measuring the shape and form of individual features—for example, by comparing areal extent to length of boundary for an area feature, such as a patch of forest. Measures of shape and form have often been used in landscape ecology and biogeography to study potential wildlife habitat areas and corridors.

Another related area is spatial modeling, encompassing everything from suitability models you can build in a GIS, to mathematical models that predict the behavior of fires and floods, to research models that predict the behavior of people or animals. The methods discussed in this book are often incorporated into spatial models.

Geographic analysis with statistics uses mathematical equations to draw conclusions about the characteristics, patterns, and relationships of geographic data. The process is similar to the statistical analysis you'd do with aspatial data, although using statistics with geographic data entails additional considerations.

Frame the question

You start an analysis by figuring out what information you're trying to get. In descriptive statistics this usually takes the form of a question: Where is the center of crimes? What is the overall direction of the storm tracks? In inferential statistics, the analysis is stated as a hypothesis: Burglaries are more clustered than auto thefts. Landslides in this area tend to occur more frequently on slopes over 30%. To ensure impartiality, statisticians structure the analysis assuming that the inverse of the hypothesis is true—burglaries are not more clustered than auto thefts; landslides are equally likely to occur on any type of slope. They then set out to decide whether to reject this null hypothesis, or not. (See "Testing statistical significance.")

Understand your data

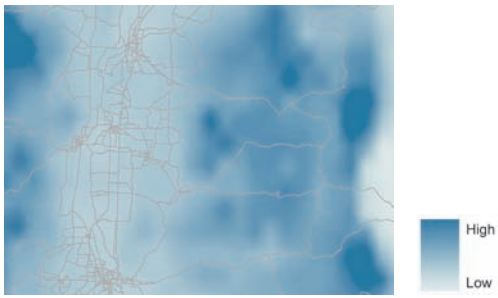
In general, you can analyze features using location alone, or using location influenced by an attribute value. The geographic representation of the data and the type of attribute values will influence the statistical methods you use.

Geographic features are either discrete or spatially continuous.

Discrete features can be points, lines, or areas. Points are used to represent either stationary features (stores and pollution monitoring stations) or events that occur at a specific place and time (burglaries and earthquake epicenters). Lines can be disjunct (elk migration routes) or connected in a network. Discrete areas are usually distinct and separate, but may share a border or even overlap as fire boundaries often do.

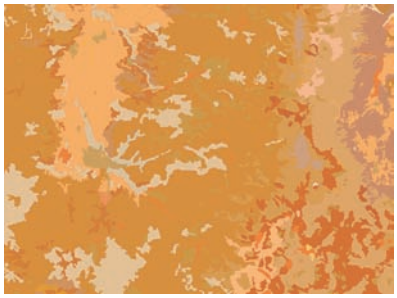
Burglaries (points), bighorn sheep migration routes (lines), and bobcat habitat (areas) are examples of discrete features.

Spatially continuous features—temperature and precipitation are oft-cited examples—are found and can be measured anywhere and everywhere. This type of data is also referred to as a continuous field, and is usually represented as a surface.



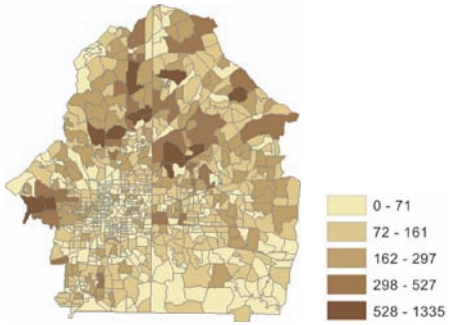
Average annual rainfall (shown with roads) is an example of spatially continuous data represented as a surface.

Spatially continuous categorical data, such as land-cover types, is represented as contiguous areas defined by boundaries.



Land-cover categories are represented using contiguous areas.

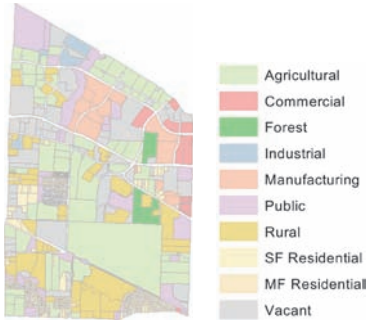
Summarized data is also represented by contiguous areas. Rather than representing what's occurring at any given location inside the area, though, the attributes associated with the areas are summaries of what's inside them—the number of senior citizens in each census tract, or the percentage of harvestable timber in each watershed. The value applies to the entire area, not any specific location within it.



Census block groups color coded by the number of senior citizens in each is an example of summarized data.

Attribute values include nominal, ordinal, interval, and ratio data.

Nominal (categorical) data describes features of a similar type. For example, you can categorize parcels by their land use, or crimes by whether they're burglaries, assaults, thefts, and so on. You might look for the center of all crimes (an entire layer) or the center of burglaries (a subset).



PARCEL ID	ACRES	LANDUSE
1N2230001200	47.90	PUB
1N223DD0520	10.06	PUB
1N223DD0510	14.29	VAC
1N225BB00100	16.61	IND
1N225BB00300	11.20	IND
1N225BB00200	4.08	VAC
1N225BB00400	7.45	PUB

Parcels color coded by land-use categories

Ordinal (ranked) data describes data that is ordered from high to low. You only know where a feature falls in the order—you don't know how much higher or lower a value is than another value. For example, you know that a city with a livability rank of 3 is lower than one ranked 2 and higher than a 4, but you don't know how much lower or higher.

SOIL CODE	RANK
11C	3
35E	6
28E	4
43	3
28E	4
33E	6
56E	7

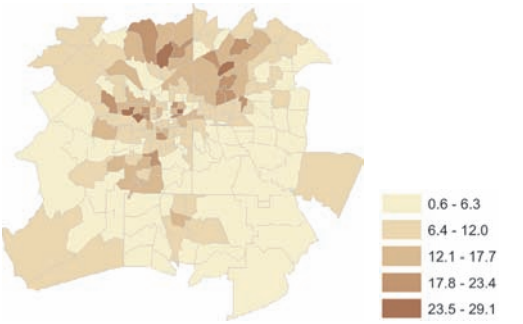
Soil types ranked by suitability for agriculture

Interval data (quantities), on the other hand, does tell you relative magnitude—you know that a house with a value of $400,000 is worth twice as much as one with a value of $200,000. Interval data can be the number of something—the number of employees at each business, the number of twelfth graders in each attendance area—or a value representing a magnitude—soil pH, store revenue.

FIPS	MEDIAN VALUE
41051004300	66600
41051004101	37800
41051007000	143500
41051004200	36400
41051004001	38600

Census tracts color coded by median house value

Ratios show you the relationship between two quantities, and are created by dividing one quantity by another for each feature. For example, dividing the number of people in each census tract by the number of households gives you the average number of people per household in each tract. Proportions (what part of a total each value is—often represented as a percentage) and densities (the quantity per unit area) are types of ratios.



TRACT	POPULATION	%65+
0211	7834	9.7
021702	10696	8.7
0100	8190	17.7
021403	6439	10.3
031103	10766	8.7
021809	7583	11.1
0098	7393	18.7
0099	5012	12.4

Census tracts color coded by percentage of the population age 65 and over

Interval and ratio data are continuous values—each feature potentially has a unique value anywhere in the range between the highest and lowest values. Knowing some basic characteristics about your data, such as whether there are any extreme values (outliers) and how bunched or spread out the values are, will help you draw correct conclusions from your analysis results. (See "Understanding data distributions.")

Choose a method

There may be more than one method you can use to answer your original question or analyze your hypothesis. Your choice depends on the type of data you have, your analysis, and—in some cases—the differences between methods.

Certain types of data are appropriate for certain types of analysis. For discrete features, you can analyze the distribution of the features themselves (whether grocery stores are clustered), or the distribution of an attribute associated with the features (whether stores with high revenue are clustered).

For spatially continuous phenomena, the distribution of values associated with the phenomenon is what you're interested in—analyzing the distribution of cells in a surface or a set of contiguous areas tells you nothing. The same is true of summarized data—you're interested in the distribution of values associated with the areas; since the areas themselves cover the entire region, no information is gained by analyzing their distribution.

If you're analyzing the distribution of feature values, you'll be dealing for the most part with continuous values. When identifying patterns and clusters, you'll want to use ratios—especially when analyzing data summarized by contiguous areas—because ratios even out the differences between large and small areas. That's important if you're interested in the concentration of features or values—a large census tract may have a large number of seniors, but they may be spread out over a wide area.

Calculate the statistic

While the emphasis in this book is on the application of the statistics, we do present the equations for each statistic along with an explanation. Since the GIS software performs the calculations, you could use a statistic without knowing the math behind it; by understanding how the statistic is derived you'll be better able to decide which statistic is best for your analysis as well as avoid drawing incorrect conclusions from your results.

Some of the statistical tools require you to provide parameters. For example, you may need to specify the nature of the influence of features on each other, such as the distance within which the prices of surrounding homes influence the assessed value of a house. (See "Defining spatial neighborhoods and weights.")

Interpret the statistic

Descriptive statistics calculate a value that can be displayed on a map—the center is an x,y coordinate location; a directional trend can be displayed using an ellipse.

Other statistics calculate a number that tells you whether there is a pattern or relationship. Often the number is within a range—the position in the range indicates the nature of the pattern (whether clustered or dispersed, for example) or the relationship, and its strength. But, to really know whether the result is meaningful, you need to test its statistical significance.

Test the significance of the statistic

The null hypothesis essentially states that there is no pattern or relationship. Significance tests help you decide whether you should or should not reject the null hypothesis.

You first decide the risk you are willing to accept for being wrong. This degree of risk, often referred to as the confidence level, is expressed as a probability ranging from 0 to 1.

Statisticians accept the null hypothesis unless there is a very small chance that they would be wrong to reject it. If you're deploying police officers, knowing that you can be 80% sure (0.20 confidence level) that burglaries are clustered in a particular area may be enough to decide to send them there. If, however, you're trying to pinpoint the cause of an outbreak of disease, you'd probably want to be at least 95% sure (0.05 confidence level) that the clusters you have identified did not occur simply by chance.

Usually, the software you're using runs the appropriate test when it calculates the initial statistic. The test calculates a statistic that you compare to a critical value—based on the confidence level—to determine whether the results are significant at that confidence level. If the test statistic exceeds the critical value, you'd be right to reject the null hypothesis. The results are said to be statistically significant at that level. (See "Testing statistical significance.")

Question the results

Finally, even if the results of your analysis are statistically significant, you'll want to question them. The geographic scale you're working at, where the study area boundaries fall, the type of data you're using, the quality of the data, and how you define proximity between features all influence the results. For example, your results may be very different if you specify straight-line distance as opposed to travel time when defining how close features are to each other.

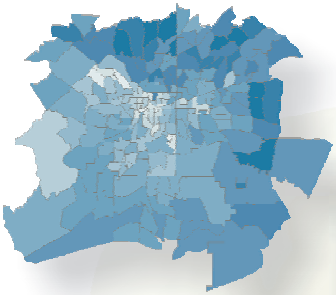
In many cases, you'll want to compare the results for the features you're analyzing to a control group. Crimes may form clusters just because those are the locations where people live; however, a cluster of crimes occurring where the population density is low may indicate a true hot spot. There are often local and regional trends in geographic data, so the outcome of your analysis may be predetermined, to some extent. (See "Using statistics with geographic data.")

The conclusions you draw from your analysis should be used in conjunction with other information, including your knowledge of the features, when making a decision. You may want to use alternate methods to confirm the results of your analysis. Statistical analysis is only one of several inputs—along with economic and political factors—to the decision-making process.

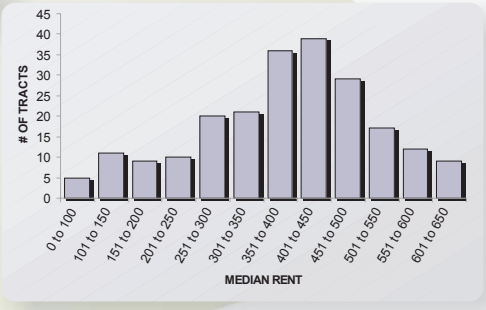
Understanding data distributions

To identify trends, patterns, and relationships, spatial measurements and statistics analyze distributions of features. Understanding the characteristics of data distributions will help you reach the correct conclusions from your analysis. Also, knowing how your data is distributed is useful before even starting the analysis—for example to spot extremely high or low values that might throw off the results of your analysis.

A type of bar chart known as a histogram shows the number of features having a particular value—the frequency distribution. For continuous values (interval and ratio data), each feature can potentially have a unique value, so ranges of values are used.

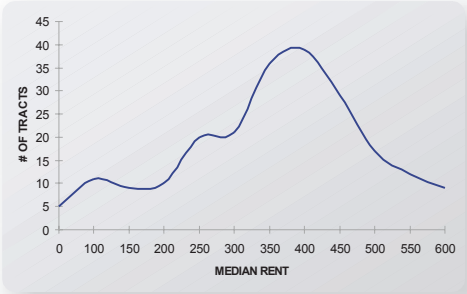


TRACT	MEDIAN RENT
0097	750
021810	614
021806	483
021402	491
021601	459
0096	462
031202	521
021703	675
021902	504
009401	813
0095	597
021401	480
⋮	⋮



Histogram for median rent by census tract, showing the number of tracts (y-axis) in each value range

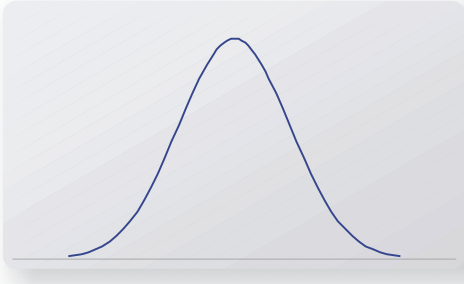
From the histogram you can create a frequency curve, which eliminates the ranges and presents the values as continuous along the x-axis. You can use a spreadsheet to create a histogram or a curve. Many GIS software packages also let you create these charts.



Frequency curve for median rent by census tract

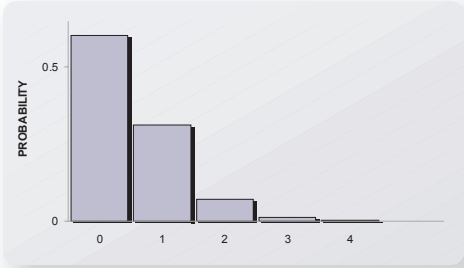
Certain frequency distributions occur routinely, with a variety of data. Mathematicians have identified and described their ideal form. Two in particular are used often in spatial statistics.

The normal distribution often occurs for phenomena where the values are similar, but some are higher and some lower, such as annual rainfall over a number of years. Most years will have close to the average amount of rainfall, but there will be a few very wet years and a few very dry ones. The frequency curve for a normal distribution is the classic symmetrical bell curve in which most values cluster in the center of the curve, and there are as many values on the left side of the curve as on the right. Given enough readings over a period of time, most values will be close to the mean.



Frequency curve for a normal distribution

The Poisson distribution, named after the French mathematician Siméon Poisson who described the distribution in the late 1830s, occurs when there are extreme events in space and time, like large-magnitude earthquakes. If the events occur randomly, there will be many time periods in which few or no events occur and very few time periods in which many events occur. In these cases, the mean will often be less than one, and the probability of no events occurring during the time period



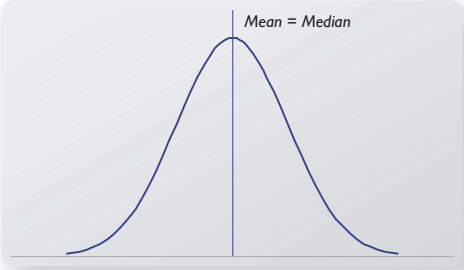
An example of a Poisson distribution, with a mean of 0.5. The y-axis shows the probability of a given number of events occurring.

will be high. In cases where the mean is greater than one (that is, on average more than one event occurs during the time period), the distribution will look different—the probability of one or more events occurring will be greater than the probability of none occurring.

Describing frequency distributions

Frequency distributions are described by measuring their characteristics.

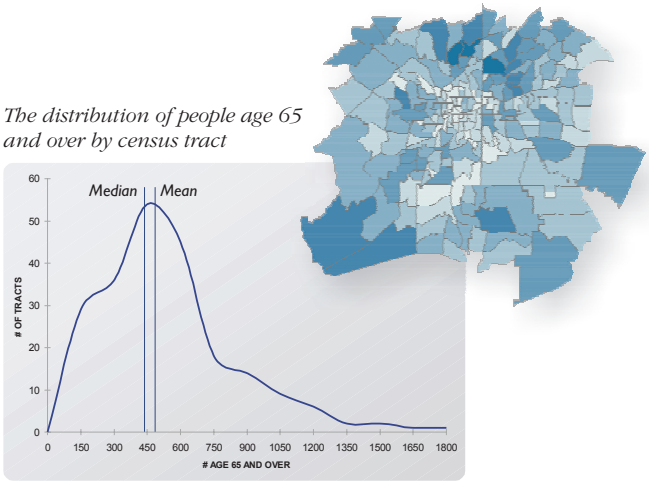
The mean and median are both measures of the central value. The mean is the average value—the attribute values are summed and divided by the number of values. The median is the middle value—half the values are higher, and half lower. In a normal distribution, the mean and median are equal.



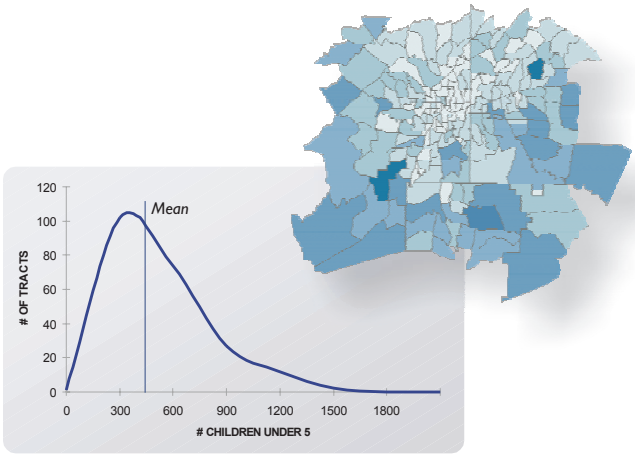
In a normal distribution, the mean and median are equal.

Most geographic distributions do not fit the normal curve. In this example, the curve is skewed toward the high values, creating a tail. The high values cause the mean to be higher than the median.

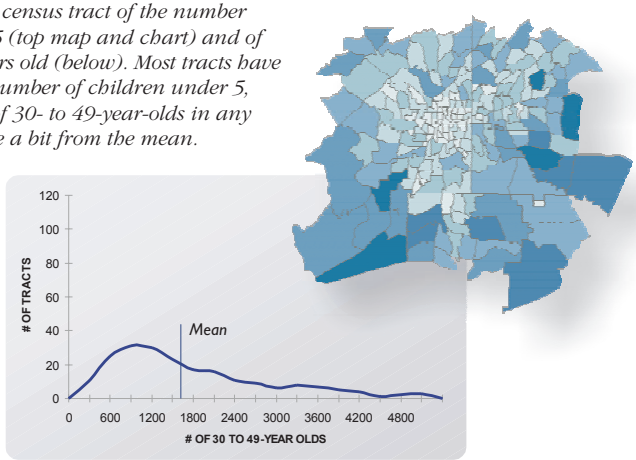
The distribution of people age 65 and over by census tract



Another characteristic of a distribution is the extent to which values vary from the mean, known as the variance. The larger the variance, the more dispersed the values are around the mean.



The distribution by census tract of the number of children under 5 (top map and chart) and of adults 30 to 49 years old (below). Most tracts have close to the mean number of children under 5, while the number of 30- to 49-year-olds in any tract can vary quite a bit from the mean.



The variance is calculated by subtracting the mean from each value, summing these differences, and dividing by the number of values. The difference for values less than the mean will be negative, so all the differences are squared to make sure they're positive before they're summed.

$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$

The variance is calculated by subtracting the mean from each value, squaring the difference, summing all the resulting squares, and dividing by the total number of values in the set.

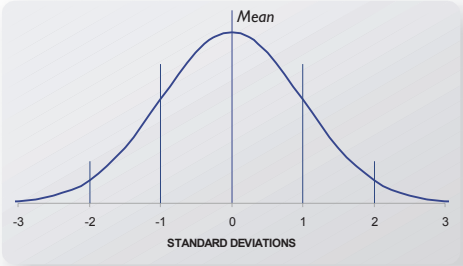
Since the difference from the mean is squared to calculate the variance, the values are in squared units rather than the original units. By taking the square root of the variance, the values are calculated back into the original units (feet, meters, inches, dollars, or whatever) required by some statistics. This measure is known as the standard deviation.

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

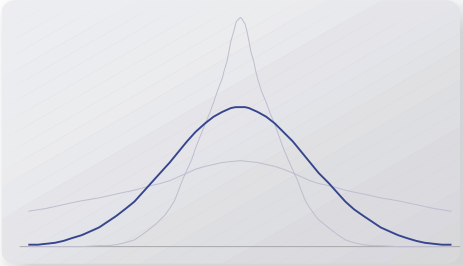
The standard deviation is the square root of the variance.

In a normal distribution, a certain proportion of the values will be within a certain number of standard deviations (plus or minus) of the mean.

- Plus or minus one standard deviation from the mean will contain about 68% of values.
- Plus or minus two standard deviations will contain about 95% of values.
- Plus or minus three standard deviations will contain over 99% of values.

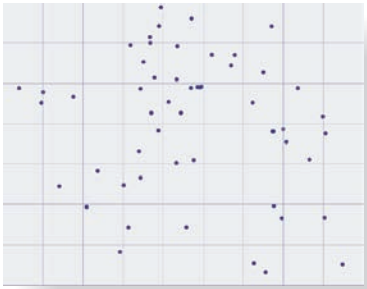


Distributions in which the mean and median are equal may still not be normal if they don't match the expected dispersion around the mean. These distributions will appear flatter or more peaked than a normal distribution.



Spatial distributions

Traditional (nonspatial) statistics deal with distributions of values that describe something—the test scores of students or the revenues generated by a chain of stores. Spatial measurements and statistics deal with distributions of these descriptive values (stored as attribute values for features in a GIS database) but also with distributions of values derived from the spatial arrangement of features. If you measured the distance between each burglary and the point at the center of the burglaries, you could analyze the distribution of distance values to get a measure of how concentrated the burglaries are.

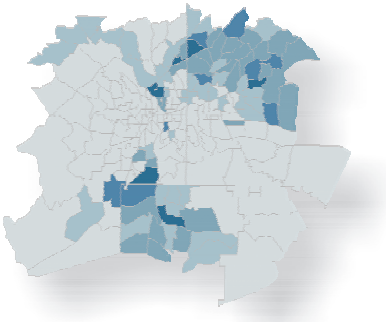


# points	# cells
0	34
1	11
2	7
3	5
4	4
5	1
6	1

Suppose you hypothesize that assaults in a city are spread evenly and ubiquitously throughout. If you impose a grid structure on top of the city and count the actual number of assaults in each cell, you're on your way to testing this hypothesis. The histogram for this data would have the lowest count to the highest count along the x-axis, while the y-axis would show the number of cells with a given count. This distribution would be compared to the distribution had the assaults been spread evenly across the city.

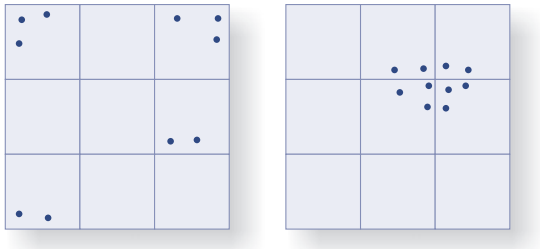
Similarly, you could measure the distance between each assault and the one closest to it, and create a frequency distribution of the distances. The mean distance between nearest neighbors can be calculated and compared to what the mean distance between neighbors would be if the assaults were randomly distributed across the city.

To analyze the spatial distribution of attribute values, the frequency distribution represents the frequency of the values influenced by the distance between features. Suppose you wanted to know whether census tracts with a high percentage of an ethnic group were clustered or not. You'd weight the percentage by the distance between tracts, create the frequency distribution of these weighted values, and compare that to the distribution of the same set of values assigned randomly to the tracts.



Spatial clusters are identified as nearby features that have similar values.

In the examples above, superimposing a grid and counting the number of assaults within each cell to measure the distribution starts to address the spatial nature of the data. However, the frequency distribution for two sets of features could be identical, even though their patterns are very different.



Though these distributions both have the same number of cells containing zero, two, and three points, their patterns are quite different.

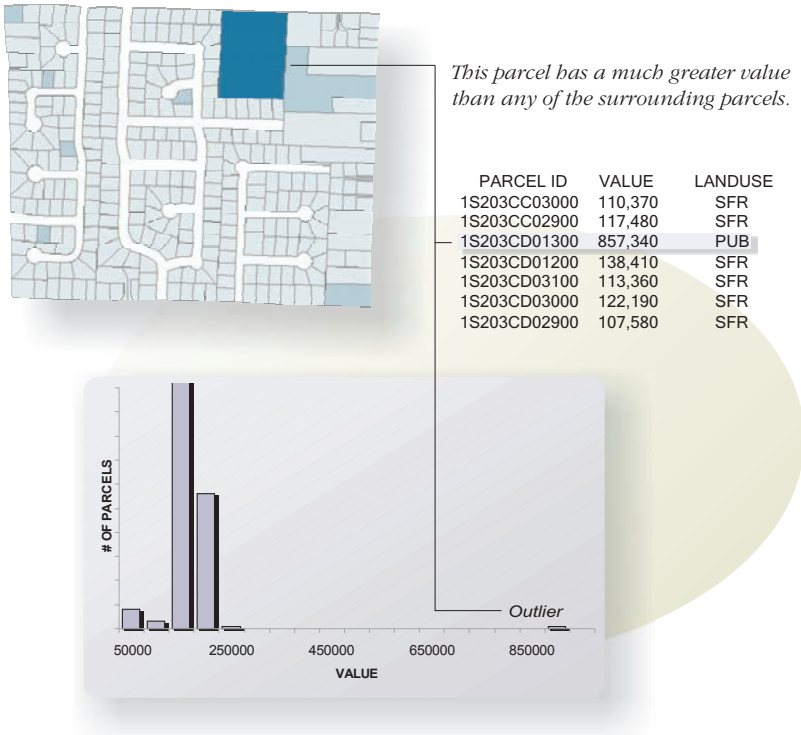
Spatial statistics let you compare the spatial distribution of a set of features to a hypothetical random spatial distribution to perform true spatial pattern analysis. In most cases, the dispersion of values around the mean (variance or standard deviation) is used as the basis of comparison. To the extent that your distribution differs from a random distribution, there is a trend or a pattern in the data. See "Testing statistical significance" and "Using statistics with geographic data" for more on comparing spatial distributions.

Outliers

Outliers are exceptionally high or low values, beyond what you'd expect even with a skewed distribution. Since outliers can throw off the results of your analysis, you need to know if there are any in your data. The frequency curve might give you a hint there is an outlier, but to be certain you need to sort the values in a table, or plot the individual data values on a graph.

Outliers often represent data errors. Values can be entered incorrectly in the database or associated with the wrong feature. Once you've identified any outliers, you need to check your original data sources to make sure the values for those features are correct. Missing values often show up as outliers. If you can't obtain a valid value for the feature, you may need to remove it from the dataset before performing your analysis.

Outliers are not always data errors. They may, in fact, represent valid but unexpected values, like a mansion in an otherwise modest neighborhood, which will have a much higher value than its neighbors. Or, an outlier might reflect a previously unknown condition, which alters assumptions of your analysis or changes your approach. (A block group having an unexpectedly high number of crimes, for example, might point to a previously unknown drug-trafficking hot spot.)



When outliers represent valid data values, you'll want to measure the influence these outliers have on your analytical results. One way to do this is to run the analysis with and without the outliers. If the results are very close to each other, the outliers aren't having a strong impact on your results. If the results deviate strongly, you may need to find analysis methods that will not be as sensitive to the presence of any outliers.

Your data may also contain spatial outliers—features that are far from other features. As with value outliers, a spatial outlier could be a valid feature. Or it could be a feature that doesn't really exist and needs to be removed from the database. Even more likely, the outlier could represent a feature that is in the wrong place—perhaps an address that was geocoded to the wrong location.

References

Burt, James E., and Gerald M. Barber. *Elementary Statistics for Geographers.* Guilford, 1996.

Earickson, Robert J., and John M. Harlin. *Geographic Measurement and Quantitative Analysis.* Macmillan, 1994.

Ebdon, David. *Statistics in Geography.* Blackwell, 1985.