2 Exploratory Spatial Data Analysis Tools and Statistics

THEORY

Learning Objectives

This chapter deals with

- The notion of exploratory spatial data analysis
- The presentation of descriptive statistics
- Spatial statistics and their importance in analyzing spatial data
- Analyzing univariate data
- Simple exploratory spatial data analysis tools such as histograms, boxplots and other visual methods utilized for deeper insight of spatial datasets
- Bivariate analysis
- Correlation and pairwise correlation
- Normalization, rescaling and adjustments
- Introducing basic notions of statistical significant tests
- The importance of hypothesis setting in a spatial context
- The importance of normal distribution in classic statistics and how it is integrated into spatial analysis

After a thorough study of the theory and lab sections, you will be able to

- Have a solid knowledge of descriptive statistics
- Use descriptive statistics for univariate analysis
- Understand and use exploratory spatial data analysis techniques to map and analyze variables attached to spatial objects
- Create plots, link them to maps and identify interesting data patterns
- Conduct bivariate analysis and identify whether two variables are linearly related; use plots to further examine their relation
- Rescale data to make comparisons between variables easier and also allow for better data handling
- Apply ESDA tools through ArcGIS and GeoDa

2.1 Introduction in Exploratory Spatial Data Analysis, Descriptive Statistics, Inferential Statistics and Spatial Statistics

Definitions

Exploratory Spatial Data Analysis (ESDA) is a collection of visual and numerical methods used to analyze spatial data by

(a) Applying classical nonspatial descriptive statistics that are dynamically linked to GIS maps and spatial objects
(b) Identifying spatial interactions, relationships and patterns, through the use of a spatial weights matrix (defined by the appropriate conceptualization method), hypothesis testing and various metrics

ESDA methods and tools are used to

- Describe and summarize spatial data distributions
- Visualize spatial distributions
- Examine spatial autocorrelation (i.e., trace spatial relationships and associations)
- Detect spatial outliers
- Locate clusters
- Identify hot or cold spots

Descriptive statistics is a set of statistical procedures that summarize the essential characteristics of a distribution through calculating/plotting:

- Frequency distribution
- Center, spread and shape (mean, median and standard deviation)
- Standard error
- Percentiles and quartiles
- Outliers
- Boxplot graph
- Normal QQ plot

Inferential statistics is the branch of statistics that analyzes samples to draw conclusions for an entire population.

Spatial statistics employ statistical methods to analyze spatial data, quantify a spatial process, discover hidden patterns or unexpected trends and model these data in a geographic context. Spatial statistics are largely based on inferential statistics and hypothesis testing to analyze map patterns so that spatially varying phenomena can be better modeled (Fischer & Getis 2010 p. 4). Unlike nonspatial methods, spatial statistics use spatial properties such as location, distance, area, length and proximity directly in their mathematical formulas (Scott & Janikas 2010 p. 27). Spatial statistics quantify and further map what the human eye and mind intuitively see and do when reading a map that depicts spatial arrangements, distributions, processes or trends (Scott & Janikas 2010 p. 27).

Why Use Descriptive Statistics and ESDA

Describing a dataset is usually the first task in any analysis. This quickly provides an understanding of data variability and allows for the identification of possible errors (e.g., a value that is not acceptable, omissions [blank cells] or outliers [scores that differ excessively from the majority]). To describe a dataset, we use **descriptive statistics** (also called "summary statistics"). Typical questions that **descriptive** statistics may address in a geographic context include the following: What is the average income in a neighborhood? What is the percentage of people having graduated from the university in a postcode? How many customers of a specific coffee shop live within a distance of less than 10 minutes walking time? What is their purchasing power, and what is the standard deviation of their income?

Descriptive statistics are useful for calculating specific characteristics (e.g., average or standard deviation), thus providing insights for data distributions. However, they do not provide linkages among the results and the spatial objects arranged in a map. The main characteristic of ESDA tools is that they are dynamically linked to maps in a GIS environment. For example, when a point in a scatter plot is brushed (selected), a spatial object is also highlighted on the corresponding map. Likewise, brushing spatial objects on the map, the relevant points/areas/bars are highlighted in the graphs. The basis of exploratory spatial data analysis is the notion of spatial autocorrelation (see Chapter 4), whereby spatial objects that are closer tend to have similar values (in one or more attributes). As such, ESDA offers a more sophisticated analysis, as it discovers patterns in data through mapping and statistical hypothesis testing (see Chapter 3).

ESDA's strength rests on two major features (Dall'erba 2009; Haining et al. 2010 p. 209):

- ESDA extracts knowledge based on its data-mining capacity, as the information that the attribute values carry is relevant to the location of data. This is extremely useful when no prior theoretical framework exists – for example, in many interdisciplinary social science fields.
- ESDA utilizes a wide range of graphical methods combined with mapping, making the analysis more accessible to people who are not accustomed to model building.

Descriptive statistics are used in conjunction with ESDA tools. Sometimes, the boundaries between them are unclear, at least for simple tools. For this reason, many books include histograms, scatter plots or boxplots in descriptive statistics and others in ESDA. The distinction is not of major importance as long as one understands how each tool works. In essence, the only difference with simple ESDA tools (e.g., histograms, scatter plots, boxplots) is that they offer the ability to link graphs to spatial objects, which enhances their power when used in research analysis (Fischer & Getis 2010 p. 3). In this book, simple tools such as histograms, scatter plots and boxplots are presented from the spatial

analysis perspective and are linked to GIS maps. More advanced ESDA topics that focus on both spatial and attribute association (e.g., point patterns analysis, spatial autocorrelation) are presented in Chapters 3 and 4. Broadly, simple ESDA tools can be used prior to the modeling phase, and advanced ESDA tools can act as model builders to identify spatial relationships and hidden patterns in spatial data (Fischer & Getis 2010 p. 3).

Why Use Spatial Statistics

Spatial statistics can be considered part of various spatial analysis methods such as ESDA, spatial point pattern analysis, spatial clustering and spatial econometrics. Spatial statistics are mainly used to

- Analyze geographic distributions through centrographic measures (see Chapter 3). In a way similar to descriptive statistics, geographic distributions can be measured to analyze their mean center and standard distance. Spatial statistics are calculated based on the location of each feature; this is a major difference from their homologous descriptive statistics, which refer solely to the nonspatial attributes of the spatial features. Although spatial statistics related to measuring geographic distributions can be weighted using an attribute value, the results refer to a spatial dimension. The spatial features used are typically points and polygons (centroids).
- Analyze spatial patterns. Spatial statistics can be used to analyze the pattern of a spatial arrangement. When this arrangement refers to point features, then the analysis is called point pattern analysis. Through such analysis, we determine whether a point pattern is random, clustered or dispersed (Chapter 3). The analysis of the spatial pattern that the attribute values (of spatial features) form in space is part of the spatial autocorrelation analysis examined in Chapter 4.
- Identify spatial autocorrelation, hot spots and outliers (see Chapter 4).
- Perform spatial clustering (see Chapter 5).
- Model spatial relationships. Spatial statistics can also be used to identify the associations and relationships between attributes and space; examples include spatial regression methods and spatial econometric models (analyzed in Chapters 6 and 7).
- Analyze spatially continuous variables such as temperature, pollution, soils, etc. In general, the type of spatial statistical analysis dealing with continuous field variables is named "geostatistics" (O'Sullivan & Unwin 2010 p. 115). Geostatistics focus on the description of the spatial variation in a set of observed values and on their prediction at unsampled locations (Sankey et al. 2008 p. 1135).

Spatial statistics are built upon statistical concepts, but they incorporate location in terms of geographic coordinates, distance and area. They extend classic statistical measures and procedures and offer advanced insights in analyses of

data. In geographical analysis, spatial statistics are not used separately from statistics but are complementary. However, there is a fundamental difference between classical and spatial statistics. In classical statistics, we make a basic assumption regarding the sample: it is a collection of independent observations that follow a specific, usually normal, distribution. Contrariwise, in spatial statistics, because of the inherent spatial dependence and the fact that spatial autocorrelation exists (usually), the focus is on adopting techniques for detecting and describing these correlations. In other words, in classical statistics, observation independence should exist while, in spatial statistics, spatial dependence usually exists. Classical statistics should be modified accordingly to adapt to this condition.

2.2 Simple ESDA Tools and Descriptive Statistics for Visualizing Spatial Data (Univariate Data)

This section presents the most common ESDA techniques and descriptive statistics for analyzing univariate data (only one variable of the dataset is analyzed each time; bivariate data analysis is examined in the next section). These include

- Choropleth maps
- Frequency distributions and histograms
- Measures of the center, spread and shape of a distribution
- Percentiles and quartiles
- Outlier detection
- Boxplots
- Normal QQ plot

2.2.1 Choropleth Maps

Definition

Choropleth maps are thematic maps in which areas are rendered according to the values of the variable displayed (Longley et al. 2011 p. 110).

Why Use

Cloropleth maps are used to obtain a graphical perspective of the spatial distribution of the values of a specific variable across the study area.

Interpretation

The first task when spatial data are joined to nonspatial data (i.e., attributes from a census) is to map them, creating "choropleth maps." For example, population, population density and income per capita can be rendered in a choropleth map. There are two main categories of variables displayed in choropleth maps: (a) spatially extensive variables and (b) spatially intensive

variables. In spatially extensive variables, each polygon is rendered based on a measured value that holds for the entire polygon – for example, total population, total households or total number of children. In the spatially intensive category, the values of the variable are adjusted for the area or some other variable. For example, population density, income per capita and rate of unemployment are spatially intensive variables because they take the form of a density, ratio or proportion. Some argue that the first category is not always appropriate and that variables should be mapped using the second category (Longley et al. 2015 p. 111) because, as each polygon has a different size, mapping real values directly might be misleading. For example, a very small and very large polygon with identical populations will be rendered in the same color if we map population in absolute values. However, adjusting population to polygon area, thus depicting population density, would lead to rendering these two polygons in different colors, as their density values are very different. Thus, the type of variable used to create a choropleth map clearly depends on the problem and on the message one wants to communicate through the mapping.

Through choropleth maps, we visually locate where values cluster or whether they exhibit similar spatial patterns. We may describe such formations using expressions such as "In the western part of the study area, variable *X* has low scores, while, in the northern part, scores are higher," or "High scores of variable *X* are clustered in the city center." This is a descriptive way of reading a map and the related symbology. There are no statistics yet, but it communicates a great deal. It may even be better than many statistical analyses, since maps often speak for themselves, provided that the maps and symbols are accurate. Nevertheless, scientific analysis must always be accompanied by statistical analysis in order to prove the findings in a statistically sound way. The next step in mapping variables through a choropleth map is to apply descriptive statistics to summarize the data and use exploratory spatial data analysis methods to visualize the values associated with locations.

Discussion and Practical Guidelines

One might present findings through maps and graphs in an inappropriate way and give the wrong impression or even a misleading message (Tufte 2001). Sometimes this happens through ignorance, and sometimes it is done deliberately to mislead. For example, the choice of colors, scale, map projection or even map title might be misleading (see Box 2.1). As professionals, we have to create accurate maps and graphs and always use solid statistics to back up our findings.

Box 2.1 The Mercator projection was invented by Mercator in 1569 to help explorers navigate to the sea. Since the earth's shape is approximated

Box 2.1 (*cont.*)

by an ellipsoid, every two-dimensional map induces distortions in either area, length or angle direction. Mercator created a projection that keeps the angle right whether in two-dimensional or real-earth terms. By drawing a line on this map between two points and calculating the angle from north, ships could go directly to the destination with no divergence. However, this projection does not preserve area. In this projection, Brazil seems to have the same area as Alaska. In fact, Brazil is five times the size of Alaska. This does not mean that the map is wrong. It is just used wrongly, as its purpose is to map angles correctly. To compare areas, we have to use other projections. To avoid misleading interpretations, a map should be used for the purposes it was created for.

2.2.2 Frequency Distribution and Histograms

Definitions

Frequency distribution table is a table that stores the categories (also called "bins"), the frequency, the relative frequency and the cumulative relative frequency of a single continuous interval variable (de Vaus 2002 p. 207; see Table 2.1).

The **frequency** for a particular category or value (also called "observation") of a variable is the number of times the category or the value appears in the dataset.

Relative frequency is the proportion (%) of the observations that belong to a category. It is used to understand how a sample or population is distributed across bins (calculated as *relative frequency* = *frequency*/*n*)

Table 2.1 Frequency distribution table. Example for $n = 15$ postcodes and their population. Five (frequency) postcodes have population between 800 and 899 (bin) people, which is 33.33% (relative frequency = 5/15) of the total postcodes. Overall, 66.67% (cumulative relative frequency = 13.33% + 20.00% + 33.33%) of the postcodes have a population of at least 899 people.

Population range/bins	Frequency	Relative frequency %	Cumulative relative frequency %
600–699	2	13.33	13.33
700–799	3	20.00	33.33
800–899	5	33.33	66.67
900–999	3	20.00	86.67
1000–1199	2	13.33	100.00
n =	15	100.00	

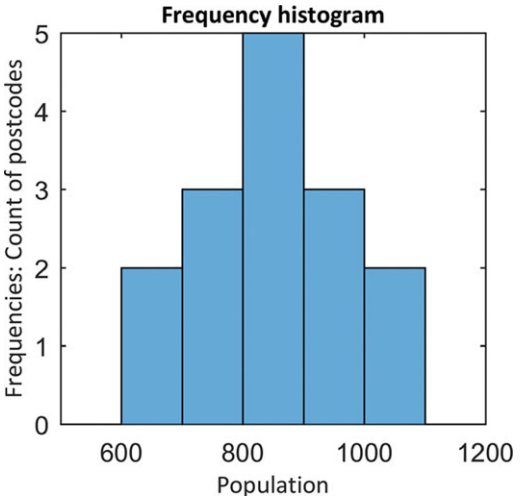

Figure 2.1 Frequency distribution histogram for the population variable in Table 2.1. Each bar depicts the number of postcodes (frequencies in the *y*-axis) for each population bin (in the *x*-axis).

The **cumulative relative frequency** of each row is the addition of the relative frequency of this row and above. It tells us what percent of a population (observations) ranges up to this bin. The final row should be 100%.

A frequency distribution histogram is a histogram that presents in the *x*-axis the bins and in the *y*-axis the frequencies (or the relative frequencies) of a single continuous interval variable (de Vaus 2002 p. 207; see Figure 2.1).

A **probability density histogram** is defined so that

(a) The area of each box equals the relative frequency (probability) of the corresponding bin
(b) The total area of the histogram equals 1

Why Use

Frequency distribution tables and histograms are used to analyze how the values of the studied variable are distributed across the various categories. The histogram can also be used to determine if the distribution is normal or not. Additionally, it can be used to display the shape of a distribution and examine the distribution's statistical properties (e.g., mean value, skewness, kurtosis). Interesting questions can then be answered that may assist spatial analysis or spatial planning (e.g., about how many postcodes have a population of less than a specific value; see Table 2.1). Histograms should not be confused with bar charts, which are used mainly to display nominal or ordinal data (de Vaus 2002 p. 205).

Discussion and Practical Guidelines

To calculate frequencies, we first set the number of bins in which the values will be grouped by dividing the entire range of values into sequential intervals (bins). The choice of the appropriate number of bins as well as their range depends on the project at hand and the scope of the analysis; it should be meaningful. A trial-and-test method is an appropriate approach for choosing how many bins to use. Using many bins is suitable if there is high variance and a large population, but using too many bins makes interpretation difficult. On the other hand, a relatively small number of bins might conceal data variability (for a small number of bins, we can use other graphs, such as pie graphs). As a rule of thumb, a value that falls on the boundary of two bins should be placed on the upper bin. For example, if the intervals for the variable "age" are set for every five years (e.g., 0–5, 5–10 and so on), then a five-year-old child should be grouped with the 5–10 bin. More mathematically, bins should be set as 0 to $<$ 5, 5 to $<$ 10 and so on.

After defining the bins and ranges, we count how many values (observations) lie within each bin. This count is the frequency. If we add all frequencies, we should obtain a total that equals the sample (n). A frequency distribution table also includes the relative frequency (percentage) and the cumulative relative frequency (cumulative percentage; see Table 2.1). All relative frequencies have to add up to 100%. For large frequencies, the plots are not well presented; some bins might be much larger than others. By using relative frequency, we change our scale on the y-axis to 0–1 (0%–100%), but all essential characteristics of the frequency distribution – such as location, spread and shape – are unchanged (see next section; Peck et al. 2012 p. 28). Likewise, the final cumulative relative frequency should be 100%.

Based on the frequency distribution table, a frequency distribution histogram can also be plotted (see Figure 2.1). In this histogram, each frequency is centered over the corresponding bin and is represented by a rectangle. For same-width bins, the area of the rectangle is proportional to the corresponding frequency. Histograms can also be created for the relative frequency. Each bar is centered on the same values in the x-axis as in the frequency distribution histogram, and the height equals the relative frequency, with the sum of the heights of all bins equaling 1. The relative frequency can be considered as the probability of a value occurrence.

Another option is to normalize (divide) the relative frequency by the width of each bin. In this case, we have on the y-axis the relative frequency per unit of the variable on the x-axis. This type of histogram is called probability density histogram. In this case, the bins are not of the same width. In fact, the area of each bin equals the probability of occurrence of a specific value or range of values and the area of all bins should equal 1. As the sample size increases, creating more bins for the same range of values, the density histogram can be fitted by a continuous function called a "probability density function" (PDF) (see Figure 2.2). The PDF is used to find the probability that a

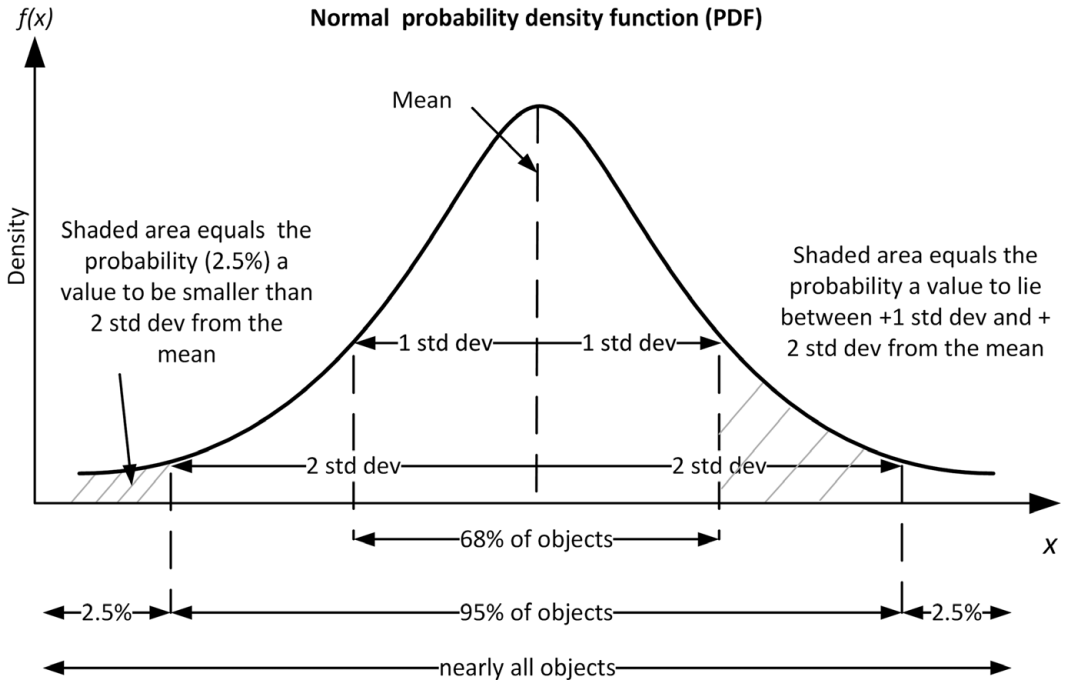

Figure 2.2 Standard deviation of a normal distribution; 68% of objects lie within one standard deviation from the mean (34% in each direction).

value *X* falls within some interval. A PDF is widely used in statistics and spatial statistics. Most of the times the units on the *x*-axis is the standard deviation of the variable.

For a normal distribution, the normal PDF is (Illian et al. 2008 p. 53, see Figure 2.2)

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.1}$$

where

μ is the mean of the population
σ is the population standard deviation
σ^2 is the population variance
x is the value of the variable

The probability distribution function of normal distribution takes two parameters: the population mean and the population standard deviation. In such a distribution, we expect that 68% of the values lie one standard deviation from the mean and that 95% of the values lie inside two standard deviations from the mean (in both directions; see Figure 2.2). The area between the intervals and the curve equals the probability by which we anticipate a value

to appear in our dataset. For example, the probability that a value will range between +1 standard deviation and +2 standard deviations in a normal distribution is 13.5% (the shaded area on the right). Values larger than two standard deviations from the mean are expected at less than 2.5% (right tail). In other words, the probability of obtaining a value larger than two standard deviations from the mean is less than 2.5%. Likewise, values smaller than two standard deviations from the mean, are expected at less than 2.5%. Some of the most commonly used significance tests in spatial statistics (as explained later) make use of the normal PDF to test if a hypothesis is true by calculating the probability under a specific interval (significance level).

Normal distribution (also called Gaussian distribution) is the most commonly used distribution in statistics, as many physical phenomena are normally distributed (e.g., human weight and height). In a normal distribution, the values of a variable are more likely to be closer to the mean, while larger or smaller scores have low probabilities of occurring. Normal distribution is used in many statistical tests to draw conclusions regarding the distribution studied. It has a zero mean, one unit standard deviation, symmetrical histogram and a bell-shaped shape (see Figure 2.3A). Not all bell-shaped histograms reveal a normal distribution, as a normal distribution decreases from the top to the tails in a certain way (see Figure 2.3G; for more on this, see Section 2.6).

In general, any distribution can be described by three important essential features: center, spread and shape. Analyzing these features provides information for (a) center, (b) extent, (c) general shape, (d) location and number of peaks and (e) the presence of gaps and outliers, as discussed next.

2.2.3 Measures of Center

Definitions

Measures of central tendency provide information about where the center of a distribution is located. The most commonly used measures of center for numerical data are the mean and the median (mode is another measure of center and is the value that occurs most often in a sample).

The **mean** is the simple arithmetic average: the sum of the values of a variable divided by the number of observations (calculated for interval data), as in (2.2):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.2}$$

where

n is the total number of observations
x_i is the score of the i*th* observation
Σ is the symbol of summation (pronounced sigma)
$\bar{x}$ is the sample mean value

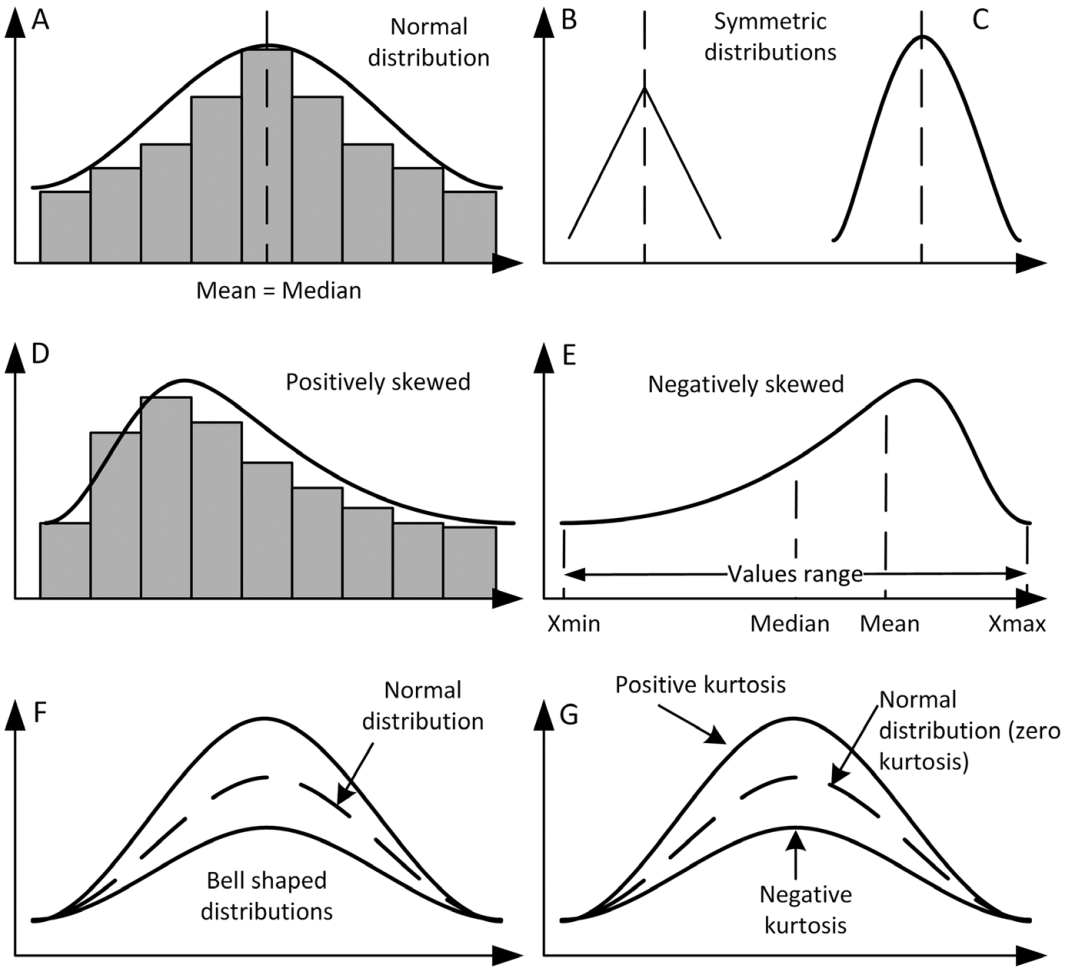

Figure 2.3 (A),(B),(C) Symmetric histograms. (D) Positively skewed. (E) Negatively skewed. (F) Different types of bell shaped distributions. (G) Curves with various kurtosis values.

The **median** is the value that divides the sorted scores form smaller to larger in half. It is a measure of center.

Why Use

The mean is used to describe the center of a distribution, while the median is used to split the frequency distribution histogram into two equal area parts.

Interpretation

If we list scores from smallest to largest, the middle score is the median. It cuts scores in two equal parts. Fifty percent of the objects have values larger than the median, and 50% of the objects have values less than the median. In

addition, the median splits a frequency distribution histogram in two equal area parts, while the mean is the balance point of the distribution histogram (the sum of the values at the left of the mean equals the sum of the values at the right).

Discussion and Practical Guidelines

We should be cautious when we interpret the mean because (a) the same mean might be the result of completely different distributions and (b) extreme values or outliers change the mean value and inflate the skewness of a distribution significantly. When we exclude outliers or extreme values, the mean is likely to change significantly. In a normal distribution, the mean is zero (calculated for standard deviations) and is located in the center of the distribution (see Figure 2.3A). The median overcomes the outlier problem, as it is based on ranked positions and not on real values. When *n* is odd, there is a single median. When *n* is even, there are two "middle values," and we take their average to obtain the median. The median is located in the center of a normal distribution and coincides with the mean (see Figure 2.3A). In other types of distributions, the median tends to deviate from the mean. Typically, but not always (depending on the values), the median lies at the left of the mean in a negatively skewed distribution (see Figure 2.3E) and at its right in a positively skewed one. The median can be calculated for both ordinal and interval data.

2.2.4 Measures of Shape

Definitions

Measures of shape describe how values (e.g., frequencies) are distributed across the intervals (bins) and are measured by skewness and kurtosis.

Shape of the distribution is the curved line (sometimes a straight line) that approximates the middle top of each bin in a continuous way. The *x*-axis closes the shape, and the area can be calculated. If a shape is symmetrical around a vertical line, the part of the histogram to the right is the mirror of the left part (see Figure 2.3A–C).

Skewness is the measure of the asymmetry of a distribution around the mean.

Kurtosis, from the graphical inspection perspective, is the degree of the peakedness or flatness of a distribution.

Why Use

Skewness is used to identify how values are distributed around the mean, while kurtosis reveals how stretched a distribution is on the *y*-axis compared to the normal distribution (see Figure 2.3G). Peakedness and flatness are actually based on the size of the tails of a distribution. For this reason,

kurtosis is prone to outliers, as outliers tend to stretch the distribution tails and significantly change the mean. Thus, the upper hill of a curve may move upward or downward according to the strength and location of an outlier.

Interpretation

If a histogram is not symmetric, it is skewed. If the right tail (part) of the histogram tends to stretch considerably more than the left tail, this histogram is named "positively skewed" or "right skewed" (see Figure 2.3D). If the left tail is stretched, we call the histogram "negatively skewed" or "left skewed" (see Figure 2.3E). Skewness of greater than 1 or less than -1 typically indicates a nonsymmetrical distribution (de Vaus 2002 p. 225). Values higher than 1.5 or less than -1.5 indicate large skewness, meaning that the data tend to stretch away from the mean in some direction. For example, suppose that Figure 2.3D depicts the frequency distribution (y-axis) of annual per capita income (x-axis). In this case, the distribution of income is positively skewed. Only a few people have very high income (lie in the right tail further from the mean) while the majority has income less than the mean (which is left of the median). Income is unequally distributed: More have less and less have more.

A zero kurtosis indicates a near-normal distribution peakedness. A negative kurtosis indicates a more flat distribution (lower than normal), while a positive kurtosis reveals a distribution with a higher peak than the normal distribution. Strictly speaking, the kurtosis for a normal distribution is 3 (de Vaus 2002 p. 227). Most statistical software subtract 3 from the final figure to adjust to the zero definition. This provides a quicker understanding, as the positive or negative values are directly interpreted as distributions over or under a normal distribution. Some popular software, such as Matlab and ArcGIS, do not follow the zero definition and regard a kurtosis of 3 as the normal distribution

2.2.5 Measures of Spread/Variability – Variation

Definitions

Measures of spread (also called measures of variability, variation, diversity or dispersion) of a dataset provide information of how much the values of a variable differ among themselves and in relation to the mean. The most common measures are as follows (de Smith 2018 p. 150):

- Range (Peck et al. 2012 p. 185)
- Deviation from the mean
- Variance
- Standard deviation
- Standard distance (see Section 3.1.4)
- Percentiles and quartiles (see Section 2.2.6)

A **range** is the difference between the largest and smallest values of the variable studied, as in (2.3):

$$Range = x_{max} - x_{min} \tag{2.3}$$

where x_{max} is the maximum value of a variable, and x_{min} is the minimum value of the same variable (see Figure 2.3E).

Deviation from the mean is the subtraction of the mean from each score, as in (2.4):

$$Deviation = (x_i - \bar{x}) \tag{2.4}$$

where

x_i is the score of the ith object
$\bar{x}$ is the sample mean value

The sum of all deviations is zero (sometimes, due to rounding up, the sum is very close to zero), as in (2.5):

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{2.5}$$

Sample Variance is the sum of the squared deviations from the mean divided by $n - 1$ (sample variance) as in (2.6) (see Table 2.2):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \text{ (sample variance)} \tag{2.6}$$

Squared values are used to turn negative deviations to positive. To calculate the variance for the entire population, denoted by σ, we simply divide by n, as in 2.7:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \text{ (population variance)} \tag{2.7}$$

Standard deviation is the square root of variance (2.8, 2.9) (see Table 2.2).

$$s = \sqrt{s^2} \text{ (sample standard deviation)} \tag{2.8}$$

$$\sigma = \sqrt{\sigma^2} \text{ (population standard deviation)} \tag{2.9}$$

Table 2.2 Sample and population statistical symbols. Sample statistics are denoted by Latin letters and population parameters by Greek letters.

Measure	Sample statistic symbol	Population parameter symbol
Mean	$\bar{x}$ pronounced: ex bar	μ: pronounced: mu (miu)
Variance	s^2 pronounced: es squared	σ^2 pronounced: sigma squared
Standard deviation	s_x pronounced: es of ex	σ pronounced: sigma

Why Use

Range is used to assess the variation of values in a variable, while deviation from the mean is used to calculate how far away a score lies from the mean. Variance is used to measure the spread of values in a variable. Standard deviation indicates the size of a typical deviation from the mean. In essence, variance and standard deviation reflect the average distance of the observations from the mean (de Vaus 2002 p. 224). Standard deviation is easier to interpret than variance, as it is measured at the same unit of the variable studied.

Interpretation

The greater the range, the more variation in the variable's values, which might also reveal potential outliers. Large values of s^2 (variance) reveal a great variation in the data, indicating that many observations have scores further away from the mean. If the variation is large, we may cut off the top and bottom 5% or 10% of the dataset to produce a more compact distribution. This typically happens in satellite image analysis for color enhancement.

A positive standard deviation value indicates the number of standard deviations above the mean, and a negative value indicates the number of standard deviations below the mean. Standard deviation is used to estimate how many objects in the sample lie further away from the mean in reference to the z-score (e.g., 1 or 2) (see Section 2.5.5). In any normal distribution and for a specific variable:

- Approximately 68% of all values fall within one standard deviation of the mean (z-score = 1).
  [($\bar{x} - 1^*$standard deviation) up to ($\bar{x} + 1^*$standard deviation)]
- Approximately 95% of all values fall within two standard deviations of the mean (z-score = 2).
  [($\bar{x} - 2^*$standard deviation) up to ($\bar{x} + 2^*$standard deviation)]
- Nearly all values fall within three standard deviations of the mean (z-score = 3).
  [($\bar{x} - 3^*$standard deviation) up to ($\bar{x} + 3^*$standard deviation)]

Discussion and Practical Guidelines

It is important to note that variation is not the same as variance (a synonym for variation is variability). Variation and variability are not some specific quantities. They are typically used as general terms expressing fluctuations in values. These fluctuations are calculated through the measures of spread.

Sample variance is the total amount of the squared deviation from the mean divided by the sample (n), but not exactly; it is divided by the sample (n) minus 1. Why minus 1? If it was just n, it would be the average of the total amount of the squared deviation, which makes more sense. In fact, minus 1 is necessary. It has been observed that variation tends to be underestimated when we use samples. Overestimating variance is better than underestimating it (Linneman 2011 p. 90). In more advanced statistics, the term $n - 1$ in this formula reveals

the degrees of freedom (*df*). Degrees of freedom generally equal the sample (*n*) minus the number of parameters estimated. It is actually the number of objects (of the sample) that are free to vary when estimating statistical parameters. "Free to vary" means that these objects have the freedom to take any value (inside the set in which the function is defined), while others are constrained by restrictions. If we are interested in the standard deviation for the entire population, denoted by σ, we simply divide by n (2.7).

Selecting between the sample statistic and the population parameter formula (see Table 2.2) depends on the nature of our analysis and the available data. Suppose we want to calculate the standard deviation of income for a specific city. If we have data for the entire population of this city (through census), we should apply the population standard deviation formula. The results are not an estimate but a real population calculation. If we want to estimate the standard deviation of income for the entire country, based only on this city sample (we infer from the city to the country), then we should apply the sample standard deviation formula (see more on inferential statistics in Section 2.5).

Finally, by combining the standard deviation and z-scores (see Section 2.5.5), we can describe how objects (and their values) lie within a distribution. For example, suppose that the mean value of incomes in 30 postcodes of a city is 15,000 US dollars and the standard deviation is 2,000 US dollars. The standard deviation of 2,000 US dollars means that, on average, incomes vary away (in both directions) from the mean by 2,000 US dollars. If the distribution of income follows a normal distribution (in practice, it does not), approximately 68% of the postcodes (nearly 20) would have incomes in the range of [*15,000 – 1*standard deviation up to 15,000 + 1*standard deviation*], or between 13,000 and 17,000 US dollars.

Additional questions using the standard deviation and z-score can be asked. For example, how many postcodes are likely to have incomes higher than 19,000 US dollars? This is an important type of questions, especially when we focus on certain subpopulations. We first calculate the z-score: *(value-mean)/(standard deviation) = (19,000−15,000)/2,000 = 2* standard deviations away (see Eq. 2.21). This value means that a postcode with an income of 19,000 US dollars lies two standard deviations above the mean. As mentioned above, only 5% of objects lie more than two standard deviations from the mean in case of normally distributed variable. This is 2.5% in each direction. In the preceding example and to answer the original question, 2.5% of the postcodes (that is one postcode) have income larger than 19,000 US dollars.

2.2.6 Percentiles, Quartiles and Quantiles

Definition

A **percentile** is a value in a ranked data distribution below which a given percentage of observations falls. Every distribution has 100 percentiles.

The **quartiles** are the 25th, 50th and 75th percentiles, called "lower quartile" (Q1), "median" and "upper quartile" (Q3) respectively.

The **interquartile range (IQR)** is obtained by subtracting the lower quartile from the upper quartile as in 2.10:

$$IQR = Upper\ quartile - Lower\ quartile = \mathrm{Q3} - \mathrm{Q1} \tag{2.10}$$

Quantiles are equal-sized, adjacent subgroups that divide a distribution.

Why Use

Percentiles are used to compare a value in relation to how many values, as a percentage of the total, have a smaller or larger value. The lower quartile (Q1), the upper quartile (Q3) and the median are commonly used to show how scores are distributed for every 25 percentiles. The interquartile range provides a measure of the variability of the 50% of the objects around the median. Quantiles are often used to divide probability distributions into areas of equal probabilities. In fact, percentiles are quantiles that divide a distribution to 100 subgroups.

Interpretation

If the 20th percentile of a distribution is 999, then 20% of the observations have values less than 999. If a student's grade lies in the 80th percentile, the student achieved better grades than did 80% of his/her classmates. The 50th percentile is the median. As mentioned, the median is the score that splits ranked scores in two; thus, 50% of the objects have higher scores, and 50% have lower ones.

Discussion and Practical Guidelines

Percentiles and quartiles are not prone to outliers, as they are based on ranks of objects. For example, the maximum score would lie in the last percentiles whether it is an outlier or not. Quartiles provide an effective way to categorize a large amount of data into a mere four categories. Finally, GIS software uses quantiles to color and to symbolize spatial entities when there are many different values.

2.2.7 Outliers

Definition

Outliers are the most extreme scores of a variable.

Why Use

They should be traced for three main reasons:

- Outliers might be wrong measurements
- Outliers tend to distort many statistical results
- Outliers might hide significant information worth being discovered and further analyzed

Interpretation (How to Trace Outliers)

For a univariate distribution, outliers can distort the mean and the standard deviation. In bivariate and multivariate analyses, many statistics – such as

correlation coefficient, trend lines and regression analysis – will provide false results if outliers exist. The most common way of tracing outliers is by graphical representation through a histogram or a boxplot. In case of histograms, if there is an isolated bar in the far left or far right part of the histogram, it is a serious indication of having outliers in the data. If skewness is very large (positive or negative), it is also an indication of outliers' presence. Another approach is to regard outliers as those scores lying more than 2.5 standard deviations from the mean. In fact, it is not easy to set a specific number of standard deviations in order to identify an outlier. When we calculate how many standard deviations from the mean a potential outlier is, we have to consider that the outlier itself raises the standard deviation and also affects the value of the mean. A scatter plot (see Section 2.3.1) is another effective way to locate an outlier in bivariate analysis.

There is also a set of methods of tracing outliers by analyzing the residuals in regression analysis (e.g., standardized residuals), but we will not refer further to these methods in this book. We should be cautious about labeling an object as an outlier because removing it from the dataset leads to new values for the mean, standard deviation, and other statistics. Outliers should be eliminated only if we comprehend why they exist and whether it is likely that similar values reappear. Outliers often reveal valuable information. For example, an outlier value for a room's temperature indicates the potential for a fire, allowing preventive action. An outlier value for credit card use may reveal a different location than those commonly used (e.g., in a different country) and thus potential fraud (Grekousis & Fotis 2012). Thus, defining a value as an outlier depends on the broader context of the study, and the analyst should decide to eliminate or include it in the dataset carefully.

Discussion and Practical Guidelines

We can handle traced outliers based on the following guidelines:

- Scrutinize the original data (if available) to check whether the outliers' scores are due to human error (e.g., data entry). If scores are correct, attempt to explain such high or low value, as it is unlikely to be just a random phenomenon.
- Transform the variable. Still, data transformation does not guarantee outliers' elimination. In addition, it may not be desirable to transform the entire dataset for only a small number of outliers.
- Delete outlier from the dataset or change its score to be equal to the value of three standard deviations (de Vaus 2002 p. 94). The choice depends on the effect it will have on the results, but deletion is preferred. In either case, the deleted or changed score should be reported.
- Temporarily remove the outlier from the dataset and calculate the statistics. Then include the outliers again in the dataset for further analysis. For example, suppose we study the socioeconomic profiling of postcodes. Some postcodes might have extremely high incomes per capita relative to others, but they also carry additional socioeconomic information that might not include outlier values. If we completely remove the postcodes

with outlying incomes, we will lose valuable information (regarding the other variables). For this reason, it is wiser to temporarily remove the income outliers only for those statistics that have a distorting effect and include them again for further analysis later.

2.2.8 Boxplot

Definition

A **boxplot** is a graphical representation of the key descriptive statistics of a distribution.

Why Use

To depict the median, spread (regarding percentiles) and presence of outliers.

Interpretation

The characteristics of a boxplot are as follows (see Figure 2.4):

- The box is defined by using the lower quartile Q1 (25%; left vertical edge of the box) and the upper quartile Q3 (75%; right vertical edge of the box). The length of the box equals the interquartile range $IQR = Q3 - Q1$.
- The median is depicted by using a line inside the box. If the median is not centered, then skewness exists.
- To trace and depict outliers, we have to calculate the whiskers, which are the lines starting from the edges of the box and extending to the last object not considered an outlier.
- Objects lying further away than $1.5 \times IQR$ are considered outliers.
- Objects lying more than $3.0 \times IQR$ are considered extreme outliers, and those between ($1.5 \times IQR$ and $3.0 \times IQR$) are considered mild outliers. One may change the 1.5 or 3.0 coefficient to another value according to the study's needs, but most statistical programs use these values by default.
- Whiskers do not necessarily stretch up to $1.5 \times IQR$ but to the last object lying before this distance from the upper or lower quartiles.

If a distribution is positively skewed, the median tends to lie toward the lower quartile inside the box of a boxplot (see Figure 2.5A). A boxplot with the median line near to the center of the box and with symmetric whiskers slightly longer than the box length tends to represent a normal distribution (see Figure 2.5B). Negatively skewed distributions tend to look like graph C (see Figure 2.5C). Outliers might lie in any direction. They can also be traced as isolated bins (see Figure 2.5A and C).

Discussion and Practical Guidelines

Apart from describing a single distribution, boxplots can be used to compare distributions of the same variable but for different groups. To compare such distributions, we use parallel boxplots (see Figure 2.6). In this case, boxplots are

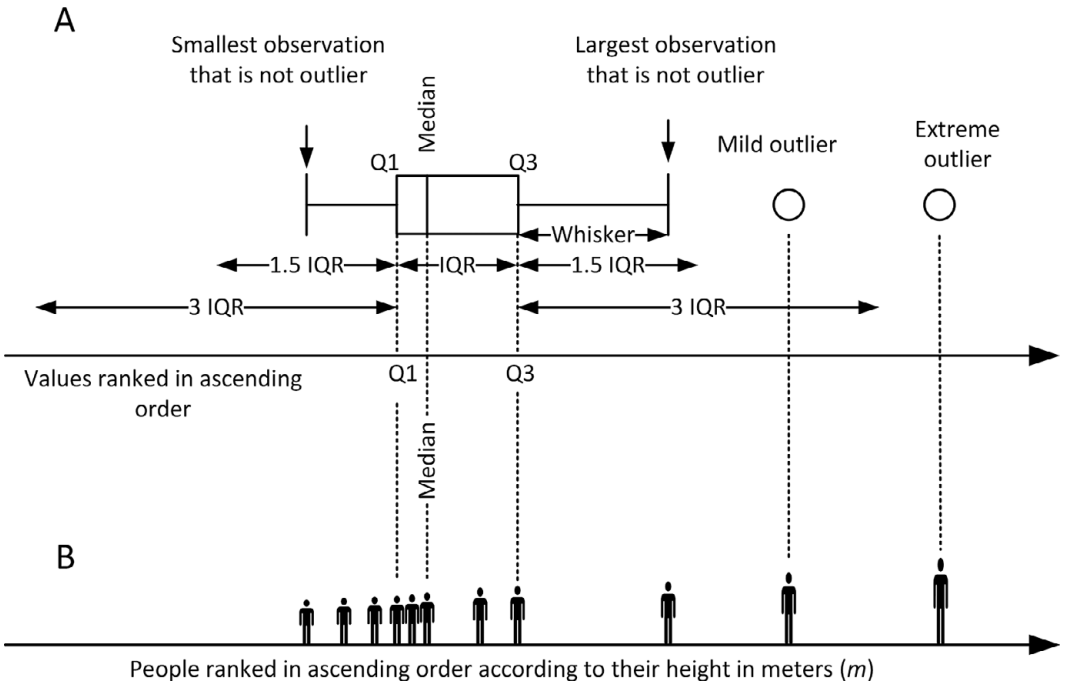

Figure 2.4 (A) Basic boxplot characteristics. In this graph, outliers exist only in the right part. Generally, outliers might exist in both parts concurrently. Whiskers stop at the largest or smallest observation that is not an outlier. In the left part, the minimum value of the variable lies less than 1.5 $\times$ *IQR* away from the lower quartile (Q1), so there is no outlier. Whisker lengths are not necessarily the same in the two parts. (B) Eleven people ranked in ascending order according to their height. The far-right person is a basketball player, and he is considerably taller than the rest (outlier).

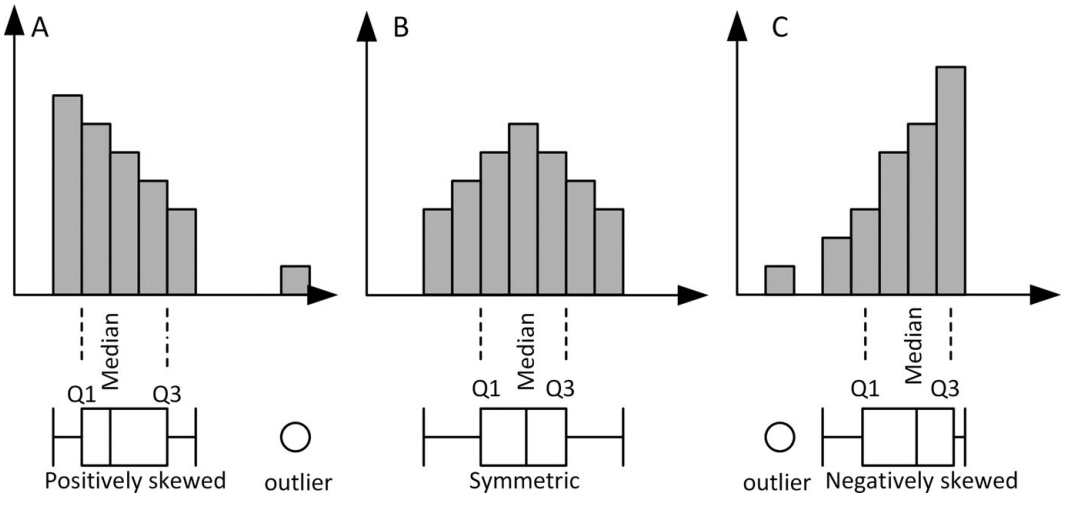

Figure 2.5 Boxplot general look for different type of distributions.

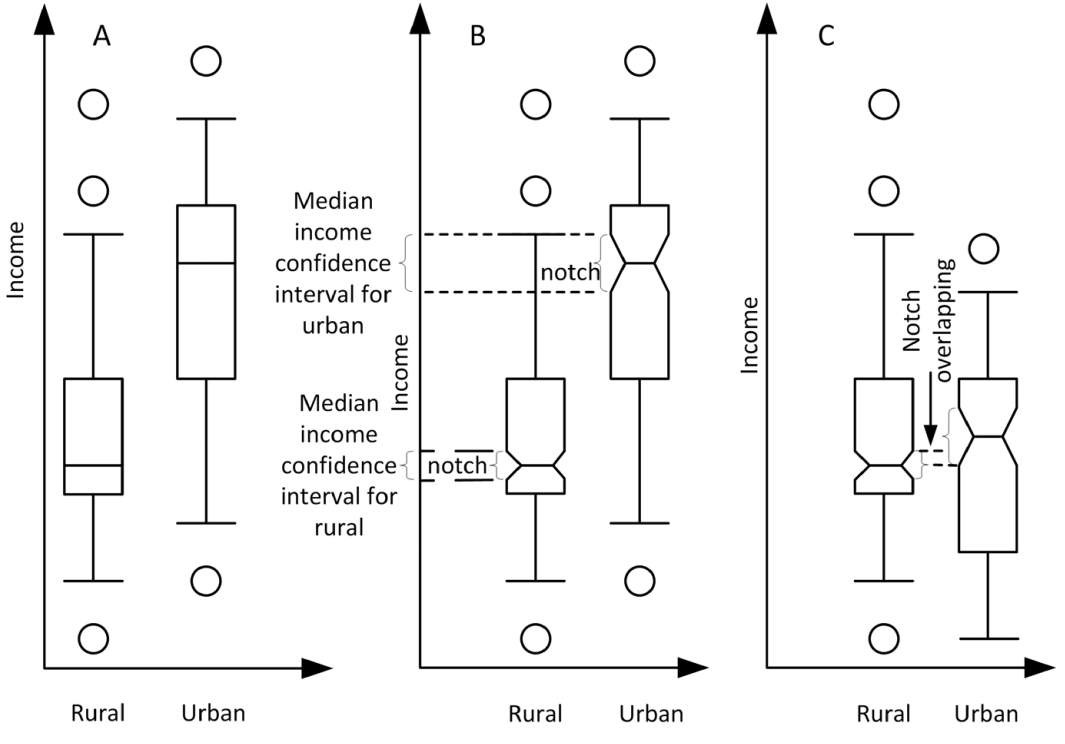

Figure 2.6 Boxplots plotted side by side to compare distributions and median. (A) A simple graphical inspection shows that the median income in urban areas is larger than the median income in rural areas. (B) Notched boxplot. (C) Notch overlapping indicates not statistical difference between the median values.

plotted side by side in a vertical representation, which is more common than the horizontal representation. We plot each group on the *x*-axis and the values of the common variable on the *y*-axis. A graphical examination is the first step in comparing these distributions – for example, to see if their medians are different. In Figure 2.6, we use parallel boxplots to describe two groups, urban and rural populations (groups on the *x*-axis), in relation to annual income (variable on the *y*-axis). Statistical tests such as a Mann–Whitney U-test should be used to check if any of the observed difference between the medians are statistically significant.

A particular type of boxplot, the notched boxplot, is used to provide a more accurate graphical representation for comparison purposes (Chambers et al. 1983; see Figure 2.6B). For each group, it provides the 95% confidence interval of the median (see Section 2.5.3 for confidence intervals). If these intervals do not overlap when we compare the two distributions (we inspect the *y*-axis), there is strong evidence at the 95% confidence level that the medians differ (see Figure 2.6B).

The confidence interval is calculated using the following formula (2.11):

$$median \pm 1.57 \frac{IQR}{\sqrt{n}} \tag{2.11}$$

In other words, the height of the notches equals 3.14 times the height of the main box divided by the square root of the sample size. This interval depicts the values in the same units as those of the variable studied. If we compare the intervals of two parallel boxplots and find that there are no common values, we may conclude, at a 95% confidence level, that the true medians of the group do differ. In Figure 2.6C, although the medians look different, the notches overlap, and we cannot conclude that there is a statistically significant difference in their medians. To better assess if there is indeed a statistically significant difference in the median between two groups, we should use statistical tests (e.g., a Mann–Whitney U test, as mentioned).

2.2.9 Normal QQ Plot

Definition

The **normal QQ plot** is a graphical technique that plots data against a theoretical normal distribution that forms a straight line.

Why Use

A normal QQ plot is used to identify if the data are normally distributed.

Interpretation

If data points deviate from the theoretical straight line, this is an indication of non-normality (see Figure 2.7). The line represents a normal distribution at a 45° slope. If the distribution of the variable is normal, then points will lie on this reference line. If data points deviate from the straight line and curves appear (especially in the beginning or at the end of the line), the normality assumption is violated. For instance, the plot in Figure 2.7 reveals non-normally distributed data.

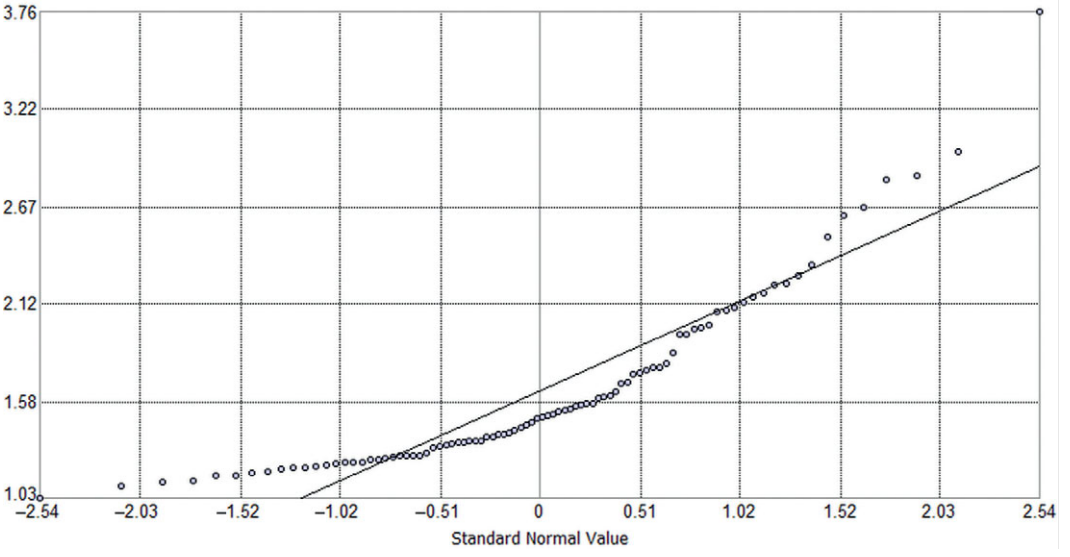

Figure 2.7 Normal QQ plot.

2.3 ESDA Tools and Descriptive Statistics for Analyzing Two or More Variables (Bivariate Analysis)

Spatial analysis often focuses on two different variables simultaneously. This type of analysis is called "bivariate," and the dataset used is called a "bivariate dataset." The study of more than two variables, as well as the dataset used, is called "multivariate." Multivariate methods will be presented in Chapter 5.

The most common ESDA techniques and descriptive statistics for analyzing bivariate data include

- Scatter plot
- Scatter plot matrix
- Covariance and variance–covariance matrix
- Correlation coefficient
- Pairwise correlation
- General QQ plot

2.3.1 Scatter Plot

Definition

A **scatter plot** displays the values of two variables as a set of point coordinates (see Figure 2.8).

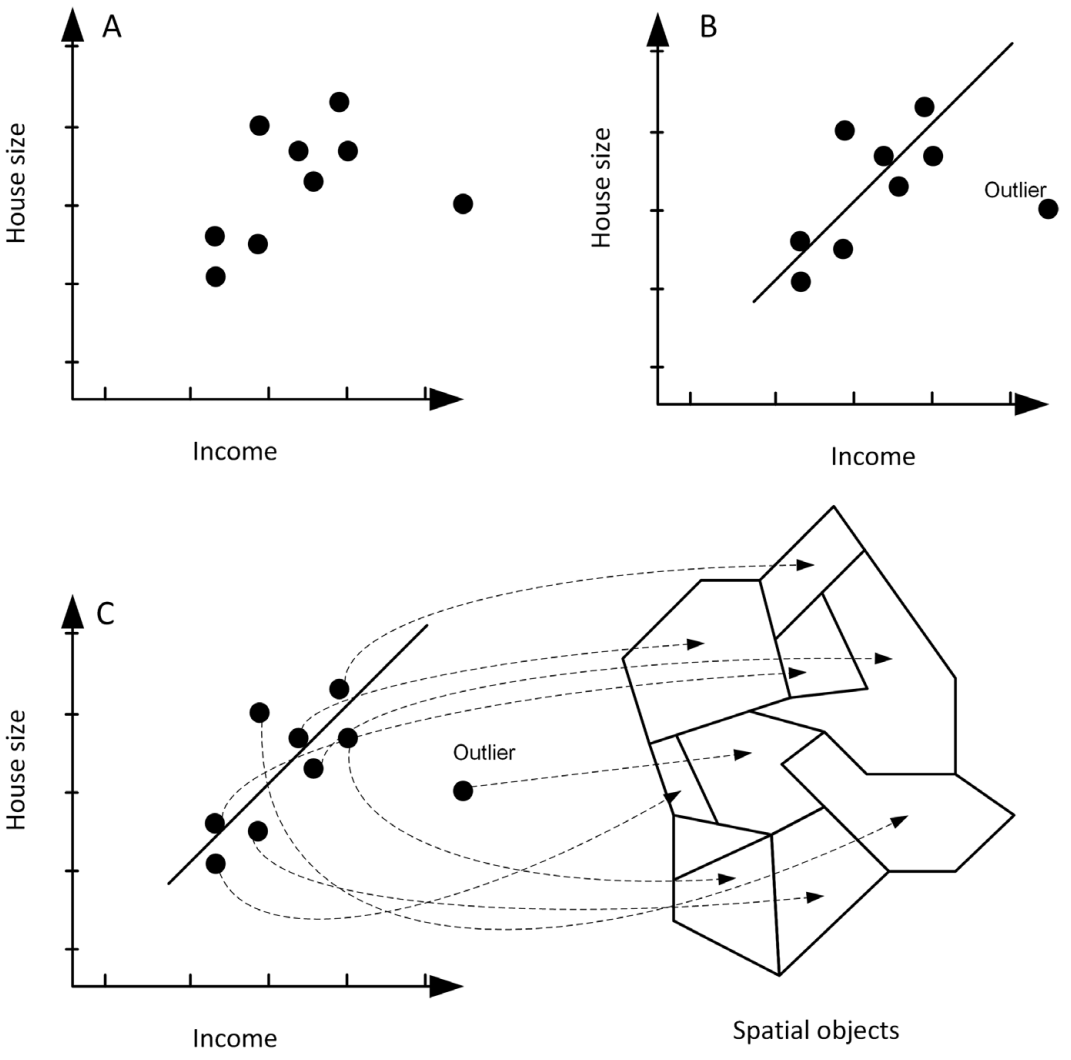

Figure 2.8 (A) Scatter plot for variables *Income* and *House size*. (B) A linear trend superimposed (with positive slope) reveals a positive linear association between the two variables. As income increases, so does house size. A value on the far right of the graph indicates a potential outlier. (C) In exploratory spatial data analysis, a one-to-one linkage exists, whereby each dot in the scatter plot stands for single spatial unit depicted in a single location on a map. Notice also that the outlier in the scatter plot is not a locational outlier, as the spatial entity does not lie far away from the rest of the entities (for locational outliers, see Section 3.1.6).

Why Use

A scatter plot is used to identify the relations between two variables and trace potential outliers.

Interpretation

Inspecting a scatter plot allows one to identify linear or other types of associations (see Figure 2.8A). If points tend to form a linear pattern, a linear relationship between variables is evident. If data points are scattered, the linear correlation is close to zero, and no association is observed between the two variables. Data points that lie further away on the *x* or *y* direction (or both) are potential outliers (see Figure 2.8B).

Discussion and Practical Guidelines

The first thing to inspect in any bivariate analysis is a scatter plot, which displays data as a collection of points (*X, Y*). In spatial data analysis, a scatter plot is a map with as many points as the spatial objects (rows in the dataset table; Figure 2.8C). For each data row in the database table, we create a set of coordinates. For example, for a single object, the value of variable *A* (*Income*) is the *X* coordinate, and the value of variable *B* (*House size*) is the *Y* coordinate (see Figure 2.8C). The *X, Y* coordinates can be switched (*A* to *Y* and *B* to *X*) with no significant change in the analysis. Each point in the scatter plot stands for a single spatial object in the map (polygons in Figure 2.8C). As the scatter plot belongs to the ESDA toolset, it offers the ability to highlight the spatial unit to which a point in the plot is linked in the map while brushing it. For instance, if we brush the outlier point, we directly locate which polygon corresponds to this value. Likewise, we can select one or more polygons in the map and identify their values in the scatter plot. We can also test if neighboring polygons in the map cluster in the scatter plot and identify if the object clustering in space creates attribute clusters as well.

2.3.2 Scatter plot matrix

Definition

A **scatter plot matrix** depicts the combinations of all possible pairs of scatter plots when more than two variables are available (see Figure 2.9).

Why Use

The visual inspection of all pair combinations facilitates (a) the locating of variables with high or no association, (b) the identification of relationship type (i.e., linear nonlinear) and (c) outlying points.

Interpretation

The closer the data points are to a linear pattern, the higher their linear correlation is to be. On the other hand, the more scattered a pattern is, the weaker the linear relationship between the two studied variables. The further away a data point lies from the main point cloud, the more likely it is to be an outlier.

Figure 2.9 All combinations of scatter plot pairs for four census variables: Income, Expenses, Bachelor degree (education), Operators (occupation).

Discussion and Practical Guidelines

By inspecting a scatter plot matrix, one can quickly identify a linear or other type of association for multiple combinations of variables in a single graph. By identifying which variables exhibit high associations, we can proceed to further analysis, such as mapping them using choropleth maps or examining for potential bivariate spatial autocorrelation (see Chapter 4). A scatter plot matrix is a quick and efficient ESDA technique for identifying the association of all variable combinations, and is usually an efficient way to begin an analysis.

2.3.3 Covariance and Variance–Covariance Matrix

Definition

Covariance is a measure of the extent to which two variables vary together (i.e., change in the same linear direction). Covariance *Cov*(*X*, *Y*) is calculated as (2.12) (Rogerson 2001 p. 87):

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{2.12}$$

where

x_i is the score of variable X of the i-th object
y_i is the score of variable Y of the i-th object
$\bar{x}$ is the mean value of variable X
$\bar{y}$ is the mean value of variable Y

This formula is for sample covariance. For population covariance, we divide by n instead of $n - 1$.

Why Use

Covariance measures the extent to which two variables of a dataset change in the same or opposite linear direction.

Interpretation

For positive covariance, if variable X increases, then variable Y increases as well. If the covariance is negative, then the variables change in the opposite way (one increases, the other decreases). Zero covariance indicates no correlation between the variables.

Covariance can also be presented along with the variance of each variable in a variance–covariance matrix (2.13). In this matrix, the diagonal elements contain the variance of each variable (calculated based on the dataset matrix A [2.14]), and the off-diagonal elements contain the covariance of all pairs of combinations of these variables.

$$\begin{matrix} & \textit{Variance} & \textit{Covariance} & & & \\ = & \begin{bmatrix} s^2_{1,1} & Cov(X_1X_2) & Cov(X_1X_3) & \cdots & Cov(X_1X_p) \\ Cov(X_2X_1) & s^2_{2,2} & \cdots & \cdots & \vdots \\ Cov(X_3X_1) & \cdots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ Cov(X_pX_1) & \cdots & \cdots & \cdots & s^2_{p,p} \end{bmatrix} \end{matrix} \tag{2.13}$$

$$A = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ a_{1,1} & \cdots & \cdots & a_{1,p} \\ \vdots & \cdots & \cdots & \cdots \\ a_{n,1} & \cdots & \cdots & a_{n,p} \end{bmatrix} \tag{2.14}$$

Where p is the total number of variables (X) and n is the total number of observations (α) of the dataset A (2.14).

Discussion and Practical Guidelines

The variance–covariance matrix is applied in many statistical procedures to produce estimator parameters in a statistical model, such as the eigenvectors and eigenvalues used in principal component analysis (see Chapter 5). It is also used in the calculation of correlation coefficients. Covariance and variance–covariance are descriptive statistics and are widely used in many spatial statistical approaches.

2.3.4 Correlation Coefficient

Definition

Correlation coefficient $r_{(x,y)}$ analyzes how two variables (X, Y) are linearly related. Among the correlation coefficient metrics available, the most widely used is the Pearson's correlation coefficient (also called Pearson product-moment correlation), given by (2.15) (Rogerson 2001 p. 87):

$$r_{(x,y)} = \frac{Cov(X,Y)}{s_x s_y} \tag{2.15}$$

where

$Cov(X, Y)$ is the sample covariance between the variables
s_x is the sample standard deviation of variable X
s_y is the sample standard deviation of variable Y

The population correlation coefficient is calculated using the population covariance and the population standard deviations.

Why Use

Correlation not only reveals if two variables are positively or negatively linearly related, but it also defines the degree (strength) of this relation on a scale of −1 to +1 by standardizing the covariance.

Interpretation

A positive correlation indicates that both variables either increase or decrease. A negative correlation indicates that a variable increases when the other decreases and vice versa. There are six main classes of correlation (see Figure 2.10). A strong positive correlation (for values larger than 0.8) indicates a strong linear relationship between the two variables; when variable X

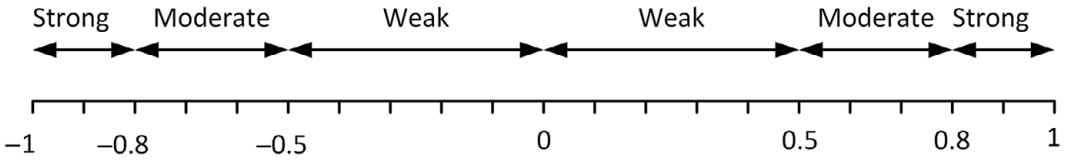

Figure 2.10 Labeling correlation.

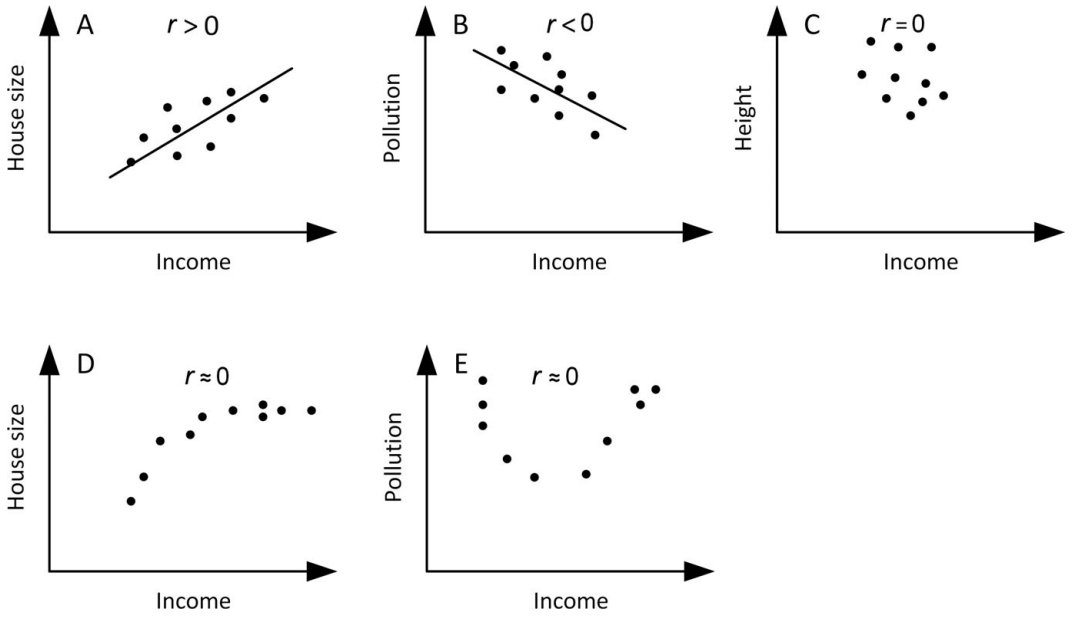

Figure 2.11 Linear correlation examples. (A) Strong positive correlation. A linear regression fit has been superimposed to the data to highlight the linear relationship. (B) Strong negative correlation. (C) No correlation (independent variables). (D) No linear correlation, but a curve pattern appears in the data. We can either use a nonlinear model or transform the data. (E) No linear correlation, but a pattern is observed in the data.

increases (or decreases), then variable *Y* also increases (or decreases) to a similar extent. A moderate positive correlation for values between 0.5 and 0.8 indicates that correlation exists but is not as intense as in a strong correlation. Observing a weak positive or weak negative correlation does not allow for reliable conclusions regarding correlation, especially when the values tend to zero. However, when the values lie between 0.3 and 0.5 (or between −0.5 and −0.3), and according to the problem studied, we may label correlation as "substantial." A moderate negative correlation (−0.8 to −0.5) means that correlation exists but is not very strong. Finally, a strong negative correlation (–1 to –0.8) indicates a strong linear relationship between the two variables (but with different directions: one decreasing and the other increasing or vice versa).

A Pearson's correlation close to zero indicates that there is no linear correlation, but this does not preclude the existence of other types of relation, as in plots D and E in Figure 2.11. Only if we use the scatter plot can we assess the potential for other types of relation.

Discussion and Practical Guidelines

Correlation coefficient is a statistical test, and its results have to be checked for statistical significance based on the null hypothesis that there is no

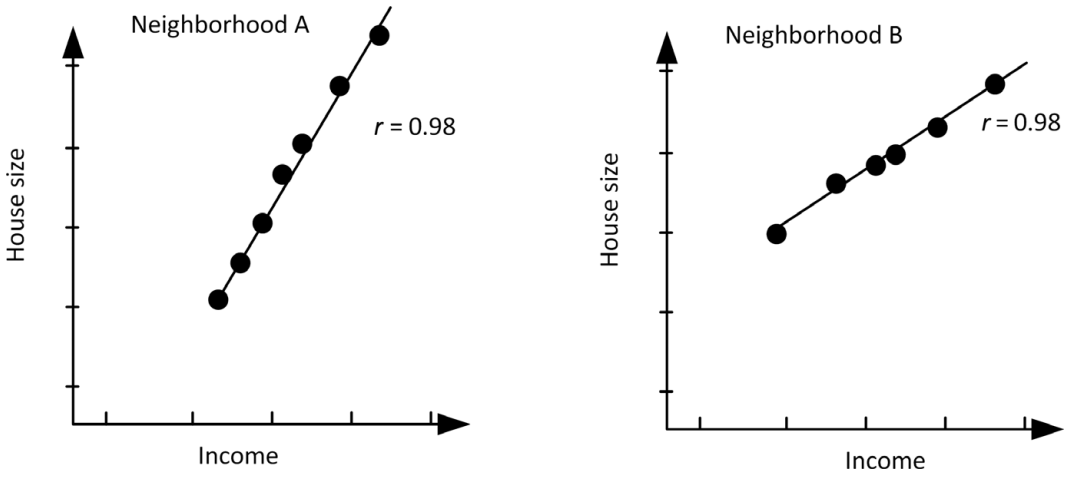

Figure 2.12 Correlation coefficient of income and house size for two different neighborhoods.

correlation between the two variables. A *p*-value is calculated expressing the probability of finding the observed value (correlation coefficient) if the null hypothesis is true (see Section 2.5.5 for a detailed analysis). A significance level should be set in advance (e.g., 0.05). If the *p*-value is smaller than the significance level (e.g., 0.001), then we reject the null hypothesis and conclude that the correlation observed is statistically significant at the 0.05 significance level. When reporting correlation values, the results should always be accompanied by their *p*-values and a related statement regarding their significance.

The slope of a regression line (the superimposed line in the data) is not the same as the correlation coefficient value unless the two variables used are in the same scale (e.g., through standardization; see Figure 2.12). The correlation provides us with a bounded measure $[-1, 1]$ of the association between the two variables. The closer to 1 or -1 it is, the closer it is to a perfect linear relationship. The slope in a regression line is not bounded by any limit and shows the estimated change in the expected value of Y for a one-unit change of X. This cannot be produced from the correlation itself. Although a positive slope is an indication of association and thus of positive correlation (i.e., the slope and correlation have the same sign), it cannot provide us with the measure of this association, as the correlation coefficient does. Nevertheless, when the variables are standardized, the slope equals the correlation coefficient.

Data in A and B depict the relation of income and house size for a set of households in two different neighborhoods (see Figure 2.12). Although the correlation is the same in both neighborhoods, the slope is different. Identical correlation means that in both neighborhoods, income is almost perfectly linearly related to housing size (all dots lie on a line), and the points have

similar deviations from the line. A difference in the slopes would indicate that, in neighborhood A, the increase in house size for one additional unit of income is far larger than that in neighborhood B. Why this happens should be determined through additional data analysis. It might be because neighborhood A lies in suburbs, where more space is available, and that neighborhood B lies close to the city center, where houses in the same price range are smaller.

Finally, correlation is a measure of association **and not** of causation. Correlation is often used to prove that one action is the result of another. This assumption is wrong if made with no further analysis. Correlation establishes only that something is related to something else. Causation and relationship/association are different. High correlation reveals a strong relation but not necessarily causation. Imagine a study on daily sales of ice cream along with the daily sales of cold drinks during the summer. If we calculate their correlation, we will probably identify a strong correlation since sales of ice cream and cold drinks are at their peak in the summer. Although this suggests an association or mathematical relationship (strong correlation), no functional relationship of causation is observed. It is not that high sales of ice cream drive (cause) the high sales of cold drinks (effect), nor is it the other way around. Another factor drives the relationship: temperature. High temperatures during summer drive (cause) the consumption (effect) of these products. Thus, the sales of ice cream and cold drinks have a functional relationship with temperature but do not have a relationship between them; this type of correlation is called "spurious." However, if we study the link between personal income and educational attainment, we will probably find a strong correlation, as people with higher income tend to have obtained at least a bachelor's degree. This is a sign of causation, as it is widely accepted that people with higher educational attainment are likely to get well paid jobs. This is merely an indication of causation and is not a definite cause-and-effect relationship. Determining whether such a link between these two variables exists would require additional scientific analysis and meticulously designed statistical experiments through explanatory analysis.

2.3.5 Pairwise Correlation

Definition

Pairwise correlation is the calculation of the correlation coefficients for all pairs of variables.

Why Use

When dealing with a large dataset, we can simultaneously calculate the correlations between all pairs of variables to identify potential linear relationships quickly.

Table 2.3 Pairwise correlation matrix for five variables. Example solved can be found in Chapter 6.

	Var1	Var2	Var3	Var4	Var5
Var1	1.0000	0.7630	0.3655	0.3560	0.4371
Var2	0.7630	1.0000	0.3281	0.2563	0.2609
Var3	0.3655	0.3281	1.0000	0.9372	0.7151
Var4	0.3560	0.2563	0.9372	1.0000	0.7399
Var5	0.4371	0.2609	0.7151	0.7399	1.0000

Interpretation

For *n* variables, the result is a square *n*-by-*n* matrix with the coefficient correlation values stored in the off-diagonal cells (see Table 2.3). Diagonal cells have a value of 1 to indicate the correlation of a variable with itself. The correlation is then interpreted as explained in Section 2.3.4.

Discussion and Practical Guidelines

We can also create a pairwise matrix plot, which displays on the off-diagonal cells (a) the scatter plots of variable pairs, (b) the correlation coefficients for each set of variables,and (c) the trend line. In the diagonal cells, the histogram of each variable is presented (see Figure 2.13). This matrix is more informative than the scatter plot matrix, which includes only the scatter plots (see Section 2.3.2).

2.3.6 General QQ plot

Definition

A **general QQ plot** depicts the quantiles of a variable against the quantiles of another variable.

Why Use

This plot can be used to assess similarities in the distributions of two variables (see Figure 2.14). The variables are ordered, and cumulative distributions are calculated.

Interpretation

If the two variables have identical distributions, then the points lie on the reference line at 45°; if they do not, then their distributions differ.

2.4 Rescaling Data

Definition

Rescaling is the mathematical process of changing the values of a variable to a new range.

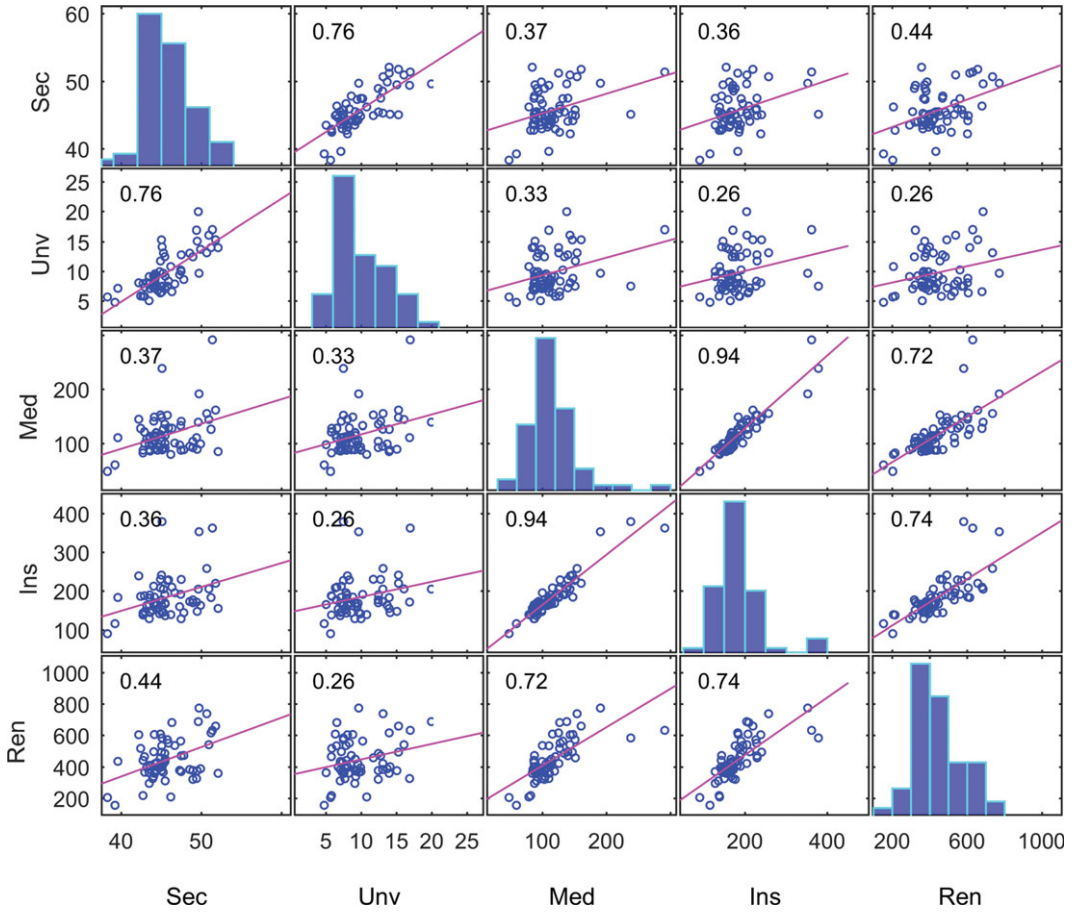

Figure 2.13 Correlation pairwise matrix plot.

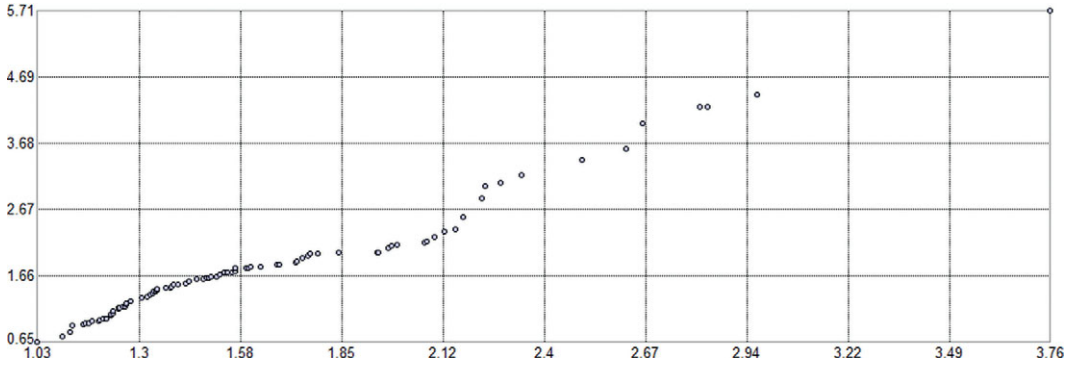

Figure 2.14 General QQ plot of *Income* (*x*-axis) and *Expenses* (*y*-axis). Points deviate from a reference line at 45°, and their distributions cannot be regarded identical.

Why Use

When variables have very different scales, it is very hard to compare them directly. By rescaling data, the spread and the values of the data change, but the shape of the distribution and relative attributes of the curve remain unchanged. Large differences in value ranges and scales, infer mainly two problems (de Vaus 2002 p. 107):

- First, comparing their descriptive statistics – such as mean, median and standard deviation – is hard, and interpretation is not straightforward.
- Second, we cannot combine data with widely differing upper and lower scores and create an index or ratio as one variable might have a larger impact on the index formula, due solely to its scale. The resulting values are likely to be too large or too small, which will also be hard to interpret.

Rescaling data is also widely used in multivariate data analysis for data reduction and data clustering (see Chapter 5). For example, as clustering methods are based in the calculation of a dissimilarity matrix which contains the statistical distance (e.g., Euclidean distance) among data points, failing to rescale data leads to assigning disproportionally more importance in variables with significantly larger values with respect to the other ones.

Methods

To avoid such problems, three rescaling methods can be used:

- **Normalize:** The following formula is a typical method of creating common boundaries (2.16):

$$Xrescale = \left(\frac{X - Xmin}{Xmax - Xmin}\right)n \tag{2.16}$$

where *Xrescale* is the rescaled value, *X* is the original value, *Xmin* is the minimum value, *Xmax* is the maximum value and *n* is the upper limit defined by user of the final variable. This rescaling method is prone to outliers, as they will scale data to a very small interval (the Xmax–Xmin range will be large).

For $n = 1$, the rescaled variable ranges from 0 to 1. This is called "normalization" and scales all numeric variables in the range [0, 1]. We can also normalize data to the $[-1, 1]$ range using (2.17):

$$Xrescale = -1 + 2\left(\frac{X - Xmin}{Xmax - Xmin}\right) \tag{2.17}$$

- **Adjust:** Another method of rescaling data is to divide a variable (or multiply it by assigning weights) by a specific value. For example, instead of using dollars to describe a person's income, we could use "income"

divided by "average income of the country." This new value has several advantages. Due to inflation, income is not directly comparable across years. This ratio provides average income as an analogy of personal income, which is comparable across years. This is called "adjustment": we adjusted personal income to average income. Adjustment is the process of removing an effect (e.g., inflation, seasonality) to obtain more comparable data. Adjustments are also needed to compare incomes in different places, such as countries. We cannot directly compare the income of someone in a Western country to that of someone in a developing country because the income variable has wide differences within its range. Adjustments could be expressed in many other ways depending on the problem studied and the research question/hypothesis tested.

- **Standardize:** Calculate z-scores. A z-score is the number of standard deviations a score lies from the mean (see Section 2.5.5, Eq. 2.21). Put simply, a standardized variable expresses the original variable values in standard deviation units (de Vaus 2002 p. 108). A standardized variable always has a mean of 0 and a variance of 1 (de Smith 2018 p. 201).

Discussion and Practical Guidelines (Normalization vs. Standardization)
Normalization and standardization are widely used in statistics and subsequently in spatial analysis. But which method is more appropriate? There is a not a straightforward answer, but let us see some basic differences:

- As a standardized variable always has a zero mean and a unit variance, it provides little information when we want to compare the means between two distributions. When the mean values are important, normalization is more descriptive.
- With standardization, the new values are not bounded. On the contrary, with normalization, we bound our data between 0 and 1 (or -1 to 1). Having comparable upper and lower bounds might be preferable and more meaningful for some studies (e.g., marketing analysis) especially when the mean values are important.
- In the presence of outliers, normalizing the non-outlying values will scale them to a very small interval. This does not happen with standardization.
- Standardization is most useful for comparing subgroups of the same variable, such as comparing between rural and urban incomes (Grekousis et al. 2015a).
- Standardization is also used in multivariate data analysis (see Section 5.1) so that variables do not depend on the measurement scale and are comparable with each other (Wang 2014 p. 144). It is occasionally preferred to normalization as it better retains the importance of each variable due to the non-bounding limitation. For example, in case of outliers,

normalized data are squeezed at a small range, and as such, when dissimilarities (through statistical distances) are calculated, they contribute less to the final values.

- Many algorithms (especially artificial neural networks' learning algorithms) assume that data are centered at 0. In this case, standardization seems a more rational choice than normalization.

Keep in mind that rescaling is not always desirable. In case we have data of similar scales or proportions (e.g., percentages) or we want to assign weights to the variables with larger values, we might not consider normalizing, adjusting or standardizing. It depends on the problem in question and the available dataset to decide if and which rescaling type to apply.

2.5 Inferential Statistics and Their Importance in Spatial Statistics

In the previous sections, we discussed the use of descriptive statistics and related ESDA tools in summarizing the key characteristics of the distribution of an attribute. However, to delve deeper into the statistical analysis of a problem, we should apply more advanced methods. For example, though we can summarize a sample, we cannot make any inference related to the statistical population it refers to. Descriptive analysis is accurate only for the specific sample we are analyzing, and the results and conclusions cannot be expanded to the statistical population it was drawn from. Suppose we query 30 households in a neighborhood about income, the size of the house and the number of cars owned by each family. Using descriptive statistics, we calculate the average family income, the number of cars owned per family and the frequency distribution of their house size. This is a very good start, but it does not tell us much about the wider area. For example, what is the average income of the census tract this neighborhood belongs to? What is the average house size in the city? What is the relation between income and the number of cars in the city? These questions attempt to generate results at a larger scale but cannot be directly answered using the descriptive statistics calculated from the 30-household sample. Making inferences from a sample to a population requires inferential statistics.

Definition

Inferential statistics is the branch of statistics that analyzes samples to draw conclusions for the entire population. In other words, through inferential statistics, we infer from the sample data what are the characteristics of the population.

Why Use

Inferential statistics are used when we need to describe the entire population through samples. With inferential statistics, we analyze a sample, and the findings

are generalized to a larger population. Descriptive statistics, by contrast, hold true only for the specific sample they were calculated for; this should be acknowledged in any analysis. Scholars often generalize and use descriptive statistics to summarize populations. A sample can describe a population if specific procedures (through inferential statistics) have been followed, but this should be clearly stated.

Importance to Spatial Statistics

Spatial statistics use inferential statistics to make inferences about a statistical population. For example, spatial autocorrelation tests, spatial regression models and spatial econometrics use inferential statistics methods. Being able to evaluate the results of spatial statistics and draw correct conclusions requires understanding of inferential statistics. Incorrect interpretations of advanced spatial statistics are common and usually stem from ignorance of inferential statistics theory. A firm knowledge of inferential statistics is required for anyone undertaking spatial analysis through spatial statistics, and the next sections cover the rudiments of the following inferential statistics topics:

- What are parametric and nonparametric methods and tests?
- What is a test of significance?
- What is the null hypothesis?
- What is a *p*-value?
- What is a z-score?
- What is the confidence interval?
- What is the standard error of the mean?
- What is so important about normal distribution?
- How can we identify if a distribution is normal?

2.5.1 Parametric Methods

Definitions

Parametric methods and tests are statistical methods using parameter estimates for statistical inferences (see Table 2.4; Alpaydin 2009 p. 61). They assume that the sample is drawn from some known distribution (not necessarily normal) that obeys some specific rules. They belong to inferential statistics.

Population parameters are values calculated from all objects in the population and describe the characteristics of the population as a whole. Population parameters are fixed values. Each population parameter has a corresponding sample statistic.

Sample statistics are characteristics of a sample. They can be used to provide estimates of the population parameters. Sample statistics do not have fixed values and are associated with a probability distribution called a sampling distribution. In practice, any value calculated from a given sample is called a statistic (Alpaydin 2009, p. 61).

Table 2.4 Parametric and nonparametric statistics according to the scope of analysis and measurement level. Parametric and nonparametric statistics can be found in many textbooks, studies and papers. A scientist should have a rough idea about these tests in order to comprehend why they are selected among so many different test options and what they intend to identify. The table focuses on statistics most commonly used to determine if differences or correlations among variables exist.

Identify	Parametric statistics (normal)	Nonparametric statistic (non- normal)	Level of measurement (for nonparametric)
Difference between two independent groups	t test (interval)	Mann–Whitney U-test Kolmogorov–Smirnov Z test Chi square	Ordinal/Interval Ordinal/Interval Nominal/Ordinal/Interval
Difference between more than two independent groups	Analysis of variance and F test (interval)	Kruskall Wallis analysis of ranks Median test Chi square	Ordinal/Interval Interval Nominal/Ordinal/Interval
Difference between two related groups	t test for dependent samples (interval)	Sign test Wilcoxon's test McNemar	Ordinal/Interval Ordinal/Interval Nominal/Ordinal/Interval
Correlation between variables	Pearson's r (interval)	Spearman's Rho Kendall's tau Gamma	Ordinal Ordinal Ordinal

Parameter estimates are estimates of the values of the population parameters. Parameter estimates are estimated from the sample using sample statistics (e.g., sample mean, sample variance). As soon as parameters are estimated, they are plugged into the assumed distribution, and the final size and shape of the distribution for the specific dataset are determined (Alpaydin 2009 p. 61). The most commonly used methods of estimating population parameters are the maximum likelihood estimation and the Bayesian estimation.

How Parametric Methods Work

To better understand how inferential statistics and parametric methods work, we should take a look at the following basic flowchart (see Figure 2.15):
The parametric statistical approach is composed of four basic steps:

A) **Population:** Define the population that the study refers to.
B) **Sampling:** Using a sampling procedure, extract data from the entire population. We use samples because we cannot practically measure each single object of a population.
C) **Sample:** Make assumptions about the distribution of the entire population – for example, that it follows the normal distribution. Next, estimate the population parameters using sample statistics.

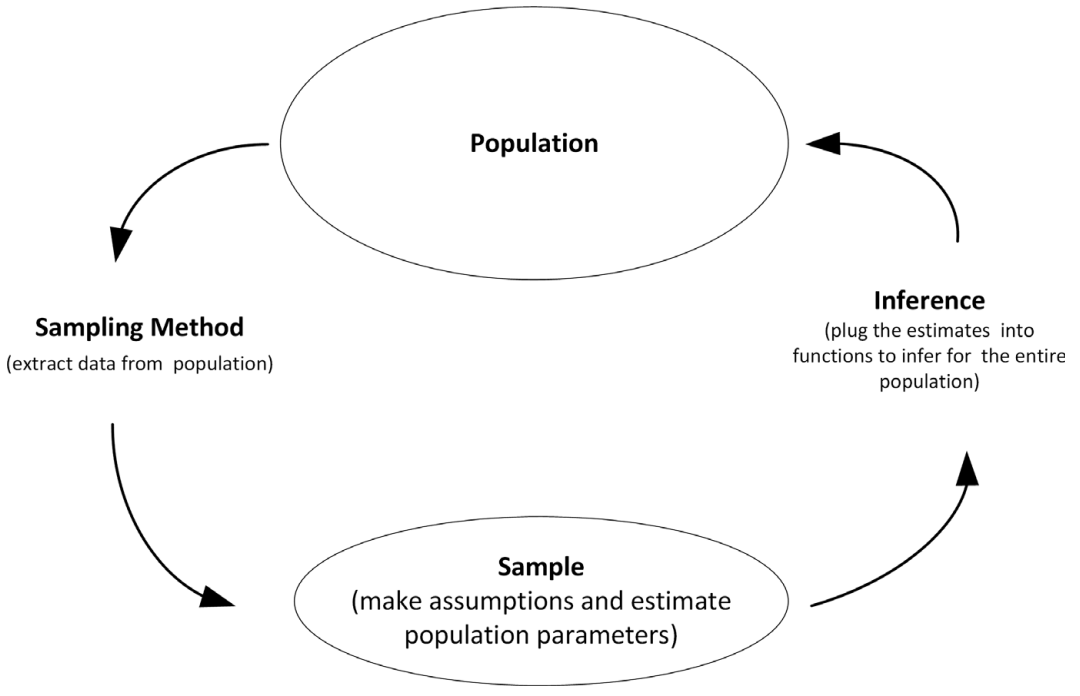

Figure 2.15 Parametric statistical approach.

D) **Inference:** Plug these parameters into functions describing the assumed distribution (e.g., probability density function), to obtain an estimation of the entire population.

Let us consider an example. Suppose we want to analyze the sales of a product according to the customers' age structure. The four basic steps are:

A) **Population:** Customers
B) **Sampling:** We first randomly select our sample (random sampling is not the only sampling method) from the total population. The sample might be *n* people who filled in questionnaires.
C) **Sample:** We then assume that the variable "sales of product" is normally distributed over the age of a potential customer (the age intervals on the *x*-axis and relative frequency of sales on the *y*-axis follow a normally shaped bell). We estimate two sample statistics: sample mean (as an estimate of the population mean) and the sample standard deviation (as an estimate of the population standard deviation) using the respective formulas.
D) **Inference:** Applying these two estimates in the normal probability density function (Eq. 2.1) allows for further analysis. For example, what is the probability that this product will be selected by customers aged 20 to 25? If the probability obtained from the probability distribution function is

not desirable (e.g., lower than the company's target), we might decide to invest more in advertisements targeting this age group. We can also calculate confidence intervals, which is the range of values within which the above estimates are likely to lie at a 95% certainty. By conducting simple parametric statistical analysis, we achieved decision making.

Discussion and Practical Guidelines

Inferential statistics use sample statistics, so termed because they refer to a sample and not to the entire population. In general, the characteristics of samples are denoted by Latin letters and the characteristics of populations by Greek letters (see Table 2.2). Formulas for the same measure (e.g., standard deviation) might change when populations or samples are calculated. The term "**statistic" is used only for samples**. The term "**parameter" refers only to a population**. (Tip to remember: "population" starts with "P," as does "parameter"; "statistics" starts with "S," as does "sample.") **Parameters** are descriptive measures of the entire population and are fixed values. Parameter estimates are calculated based on sample statistics and are not predictable. They are associated with probability distributions and margins of errors.

To sum up, the main goal of inferential statistics is to **estimate the original population parameters** and measure the **amount of error** of these estimates by depending on the sample data available. In other words, as we cannot directly measure these parameters in the entire population, we use the sample to estimate the true parameters.

For example, the normal distribution needs only two parameters to be defined, the mean and the standard deviation. Once we estimate these parameters (by using the sample data), we can plug them into a probability distribution function (see Section 2.2.2) to generate the distribution curve. The objective of describing the population is now achieved. In inferential statistics, each distribution is defined entirely by only a small number of parameters (usually one to three).

The parameters' values are called "estimates," as we do not calculate a value but produce an estimation with an associated error. The correct expression is "parameter estimate," not "parameter calculation." There are two basic approaches to evaluating a parameter estimation:

(a) Using a confidence interval (see Section 2.5.3)
(b) Hypothesis testing (see Section 2.5.5)

Parametric methods and tests are more accurate and have higher statistical power if the assumption of the distribution adopted (e.g., normal) is true relative to nonparametric methods (see next section). Another advantage is that the problem is reduced to the estimation of a small number of parameters (e.g., mean and variance). Inferences are valid only if the assumptions made in the parametric approach hold true. When the assumptions (e.g., randomly

selected and independent samples) fail, they have a greater chance of producing inaccurate results. In the spatial context, we have to assess if the assumptions hold before we use a parametric test. Randomness and thus complete spatial randomness are rare in space due to spatial autocorrelation (as a result of spatial dependence in most of the geographical problems – see Chapter 4). Spatial statistics that overcome these problems should be created. However, the new tests share similar terminology with classic statistics and are also based on significance tests, hypothesis testing and confidence intervals.

2.5.2 Nonparametric Methods

Definition

Statistical methods used when normal distribution or other types of probability distributions are not assumed are called "**nonparametric"** (Alpaydin 2009 p. 163)**.** In nonparametric methods, we do not make assumptions about the distribution of the data and the parameters of the population we study. The distribution and the number of parameters are no longer fixed.

Why Use

If assumptions are violated or if we cannot be certain whether they hold true, we may turn to nonparametric statistics/methods/models. Nonparametric models are not based on assumptions regarding either the sample (data) or the population drawn from (e.g., linear relationship or normal distribution). Nonparametric methods are based on the data, and their complexity depends on the size of the training dataset. The only assumption made is that similar inputs lead to similar outputs (Alpaydin 2009 p. 163). In nonparametric methods, the parameters' set is not fixed and can increase or decrease according to the available data.

Discussion and Practical Guidelines

Nonparametric methods do not imply the absence of parameters. They imply the non-predefined nature of the analysis and suggest that parameters are flexible and adapt according to the data (for example, the nonparametric kernel density estimation method [Section 3.2.4] has the smoothing parameter *h*). Nonparametric tests are widely used when populations can be ordered in a ranked form. Ordinal data are thus well structured for nonparametric tests. One can use an index to rank interval data and nonparametric tests if needed. Nonparametric statistics make fewer assumptions and are simpler in structure; for this reason, their applicability is wide. Nonparametric methods include Mann–Whitney U test (also known as Wilcoxon test), Kolmogorov–Smirnov test, Kruskall–Wallis one-way analysis of ranks, Kendall's tau, Spearman's rank correlation coefficient, kernel density estimation and nonparametric regression (de Vaus 2002 p. 77; Table 2.4).

2.5.3 Confidence Interval

Definition

Confidence interval is an interval estimate of a population parameter. In other words, a confidence interval is a range of values that is likely to contain the true population parameter value. A confidence interval is calculated once a confidence level is defined.

Confidence level for a confidence interval reflects the probability that the confidence interval contains the true parameter value. It is usually set to 95% or 99%. It should not be confused with the significance level (see Section 2.5.5). A confidence level of 95% reflects a significance level of 5%.

Why Use

Confidence interval is used to estimate a range of values in which a population parameter lies for a certain confidence level (probability).

Interpretation

How accurately a statistic estimates the true population parameter is always an issue. The confidence interval of a statistic (e.g., the mean) estimates the interval within which the population parameter (e.g., mean) ranges. The confidence interval is expressed in the same unit used for the variable.

Confidence intervals are constructed based on a confidence level X% defined by the user, such as 95% or 99%. Confidence levels indicate that, if we conducted the sampling procedure several times, the confidence interval would include the estimated population parameter X out of 100 times. Have in mind that the confidence level, (for example, 95%), does not indicate that for a given interval there is a 95% probability that the parameter lies within this interval. It indicates that 95% of the experiments will include the true mean, but 5% will not. Based on the definition of confidence interval by Neyman (Neyman 1937), once an interval is defined, the parameter either is included or not. As such, the probability does not refer to whether the population lies inside the interval but on the reliability of the estimation process of getting an interval that includes the true population parameter.

Discussion and Practical Guidelines

We will not detail how confidence intervals are calculated, as computing them directly is rare. It is more important to understand their use. Suppose we have selected a sample of households from a city and we want to estimate the mean household income of the households for the entire city (population). If the sample's mean income is 15,000 US dollars and the margin of error for the 95% confidence level is $\pm$500 US dollars, this typically means that we can be 95% confident that the mean income of the households in this city ranges between 14,500 and 15,500 (confidence interval). There is still a 5% chance that the mean income lies in another range. To reduce the range of the interval, we can

use a larger sample size. This will reduce the standard error and thus the interval produced.

Confidence intervals are different from significant tests (explained in Section 2.5.5). A confidence interval provides a more complete view of a variable. Instead of deciding whether or not to reject the sample estimate, a confidence interval estimates the margin of error of the sample estimate (de Vaus 2002 p. 187). The margin of error is the value to be added or subtracted from the statistic – e.g., sample mean – which reflects the interval length. This reminds us that there is no absolute precision in any estimate.

Confidence intervals (and standard error of the mean discussed in the next section) are common in reports and papers related to geographical analysis, and one should be able to interpret these statistics in the context provided.

2.5.4 Standard Error, Standard Error of the Mean, Standard Error of Proportion and Sampling Distribution

Definitions

The **standard error** of a statistic is the standard deviation of its sampling distribution (Linneman 2011 p. 540). The standard error reveals how far the sample statistic deviates from the actual population statistic.

Standard error of the mean is the standard deviation of the sampling distribution of the mean.

A **sampling distribution** is the distribution of a sample statistic for every possible sample of a given size drawn from a population.

The standard error of the mean refers to the change in mean in each different sample. This procedure is more straightforward than it seems. The standard error of the mean is calculated by the following formula (2.18): (O'Sullivan & Unwin 2003 p. 403):

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{2.18}$$

s is the sample standard deviation of the distribution studied.
n is the sample (number of objects).

In case that attribute values (scores) are expressed as percentages (P), the standard error is calculated by the following formula (2.19) (Linneman 2011 p. 177):

$$\sigma_{\bar{x}} = \sqrt{\frac{P(1-P)}{n}} \tag{2.19}$$

For example, if for n = 30 (sample size of students), 6 out of 30 (20%) smoke, then the standard error of this value would be (2.20) (de Vaus 2002 p. 193):

$$\sigma_{\bar{x}} = \sqrt{\frac{0,20(1-0,20)}{30}} = 7.3\% \tag{2.20}$$

Why Use

The standard error of a statistic shows how accurately this statistic estimates the true population parameter. The standard error of the mean, for example, is used to estimate how precisely the mean of a sample has been calculated. It measures how close the sample mean is to the real population mean. The standard error is also used to calculate the confidence interval based on the z-score and the confidence level (de Vaus 2002 p. 188).

Interpretation

Low values of the standard error of the mean indicate more precise estimates of the population mean. The larger the sample is, the smaller the standard error calculated. This is rational, as the more objects we have, the closer to the real values our approximation will be. According to two rules from probability theory:

- There is 68% probability that the population parameter is included in a confidence interval of ± 1 standard error from the sample estimate.
- There is 95% probability that the population parameter is included in a confidence interval of ± 1.96 standard errors from the sample estimate (de Vaus 2002 p. 188).

Discussion and Practical Guidelines

When we estimate a population parameter (e.g., the mean), we obtain a single value of the statistic, also called the "point estimate." As this point estimate has been calculated based on a sample, it is subject to a sampling error. It is crucial to identify how much error this point estimate is likely to contain. In other words, we cannot be confident that one sample represents the population accurately. We have to select more than one sample, calculate the statistic for each one and then analyze the formed distribution of the statistic produced. As such, estimating population parameters usually requires the use of a sampling distribution and the calculation of the standard error, which reflects how well the statistic estimates the true population parameter.

For example, to calculate the standard error of the mean, we should take many samples and then calculate the mean for each one. The resulting distribution is a sampling distribution of the mean. According to central limit theorem for a sample size larger than 30 objects/observations (O'Sullivan & Unwin 2003 p. 403):

- The distribution of the sample means will be approximately normal (as the sample size increases), regardless of the population distribution shape.
- The mean of the sampling distribution of the mean is approximating the true population mean.

As it is unfeasible to draw hundreds of samples to create a sampling distribution, we use a single sample drawn from a population that hypothetically belongs to a sampling distribution and then estimate the value of the standard error of this sampling distribution (for example, for the standard error of the mean, we use Equation [2.16] [Linneman 2011 p. 166]).

2.5.5 Significance Tests, Hypothesis, p-Value and z-Score

Definition

A **test of significance** is the process of rejecting or not rejecting a hypothesis based on sample data. A test of significance indicates the probability that the results of the test are either due to sampling error or reflect a real pattern in the population the sample was drawn from (de Vaus 2002 p. 166). A test of significance is used to determine the probability that a given hypothesis is true.

The **p-value** is the probability of finding the observed (or more extreme) results of a sample statistic (test statistic) if we assume that the null hypothesis is true. It is a measure of how unlikely the observed value or pattern is to be the result of the process described by the null hypothesis. It is calculated based on the z-score.

The **z-score** (also called z-value) expresses distance as the number of standard deviations between an observation (for hypothesis testing calculated by a specific formula for a statistical test) and the mean. It is calculated (for samples) by the following formula (2.21):

$$z\ score = \frac{x - \bar{x}}{s} \tag{2.21}$$

where

x_i is the score of the ith object
$\bar{x}$ is the sample mean value
s is the sample standard deviation

The z-score is widely used in standardization (see Section 2.4), in determining confidence intervals and in statistical significance assessments (O'Sullivan & Unwin 2003 p. 389).

Significance level α is a cutoff value used to reject or not reject the null hypothesis. Significance level α is a probability and is user-defined, usually taking values such as α = 0.05, 0.01 or 0.001, which stand for 5%, 1% and

0.1% probability levels. The smaller the p-value the more statistically significant the results. A significance level of 5% reflects a confidence level of 95%.

Interpretation

In general, significance tests use samples to decide between two opposite statements (the null hypothesis and the alternative hypothesis). For example, a null hypothesis (H_0) can be that the sample observations result purely from a random process. The alternative hypothesis (H_1) states the opposite: that the sample observations are influenced by a nonrandom cause. This type of hypothesis is the most common in statistical testing. Another null hypothesis could be that there are no differences between two samples drawn from different distributions. The alternative hypothesis states that there are differences between the samples. The statement of the null hypothesis can be set according to the problem, but the alternative is the opposite. Statistical tests reject or do not reject the null hypothesis. In this respect, they should be designed carefully to reflect the problem at hand.

- *Rejecting the Null Hypothesis* ($p \leq \alpha$)
  A low p-value means that the probability that the observed results are the outcome of the null hypothesis under consideration is low. If the p-value is smaller than α, then we reject the null hypothesis.

  > *Rejecting the null hypothesis means that there is a* (100% $-$ α) *probability that the alternative hypothesis* H_1 *is correct.*

  More analytically, we accept H_1 as true but, in relation to a probability, not a certainty. This means that there is always a chance that H_1 will be accepted as true when it is not (Type I error; de Vaus 2002 p. 172). For example, we might reject the null hypothesis with a probability of 95%, but there still is a 5% (significance level) chance that it is true.
- *Not Rejecting the Null Hypothesis* ($p > \alpha$)
  When the p-value is larger than significance value α, then we cannot reject the null hypothesis. In such a case, we have to be very careful in interpreting the results. We do not use the word "accept" for the null hypothesis. When we do not reject the null hypothesis, this does not mean that we accept it; it simply means that **we fail to reject** it, as there is insufficient evidence to do so.

  > *Not rejecting the null hypothesis means that we do not have enough evidence to reject it but we cannot accept it without further analysis.*

  For example, in case we test whether a distribution is normal or not (Ho: the distribution is normal) and $p > \alpha$, we cannot straightforwardly accept the distribution is normal. We have to examine other characteristics, such

as plots or descriptive statistics, to conclude if we will ultimately accept the null hypothesis. In correlation analysis using Pearson's r statistic, failing to reject the null hypothesis (Ho: there is no linear correlation between two variables) does not necessarily mean that there is no correlation between the variables. If we create a scatter plot, we might trace a nonlinear correlation. An efficient way to avoid non-rejecting null hypothesis vagueness is to switch hypotheses. When deciding which alternative hypothesis to use, we have to consider our research objective. Some texts may say that not rejecting the null hypothesis means that you can accept it, but this should be done with caution.

Two types of error can result from a significance test (de Vaus 2002 p. 172):

- Type I error: when we reject the null hypothesis when it is true. This is determined by the significance level, which reflects the probability of Type I error.
- Type II error: when we do not reject the null hypothesis when we should.

In the spatial context, these errors reflect the multiple comparison problem and spatial dependence (see Section 4.6 on how to deal with these problems).

Discussion and Practical Guidelines

A typical workflow for significance testing is as follows (provided that a sample is collected):

1. Make the statement for the null hypothesis (e.g., that the sample is drawn from a population that follows a normal distribution)
2. Make the opposite statement for the H_1 hypothesis (e.g., that the sample is drawn from a population that is not normally distributed)
3. Specify the significance level α.
4. Select a test statistic (some formula) to calculate the observed value (e.g., sample mean).
5. Compute the p-value, which is the probability of finding the observed (or more extreme) results of our sample if we assume that the null hypothesis is true. Put otherwise, the p-value is the probability of obtaining a result equal to (or more extreme than) the one observed in our sample if the null hypothesis is true.
6. Compare the p-value to the significance value α. If $p \leq \alpha$, then we can reject the null hypothesis and state that the alternative is true and that the observed pattern, value, effect or state is statistically significant. If $p > \alpha$, then we cannot reject the null hypothesis, but we cannot accept it either.

Significance tests are used to assess if the probability that the null hypothesis is true (e.g., correlation or differences among samples) is due to sampling error or the existence of patterns in the population. One of the problems of using tests of significance is that they produce a binary yes/no (rejected/not rejected) answer, which does not entirely apply to the real world. In addition,

significance tests do not provide a sense of the magnitude of any effect that the hypothesis is testing. On the contrary, the confidence interval approach, presented in the previous section, is more sufficient on this aspect (see Box 2.2).

Box 2.2 Think of significance and hypothesis tests as a trial. If you are accused of tax fraud, the tax bureau charge you. The null hypothesis is "guilty," and the alternative hypothesis is "not guilty." Your layer will bring forward evidence to prove that you are not guilty. If the evidence is sufficient, then the null hypothesis will be rejected and you will be declared innocent. If the evidence is not sufficient, then you cannot prove that you are not guilty, and you will most likely pay a fine. Still, this does not necessarily mean that you are guilty. Maybe the lawyer's evidence was not strong enough. In a parallel statistical world, the null hypothesis (guilty) is rejected if there is strong evidence (sample data with small errors). If the evidence is not sufficient, you cannot reject the null hypothesis, but this does not mean that the null hypothesis is true either.

Example in a Geographical Context

Suppose we study the following geographical research question: do people in cities earn more money than people in rural areas? This is a typical comparison question and is very interesting from the social and geographical perspectives. The key element here is location. Our **research question** is whether the rural–urban distinction affects income. It is a binary question, so hypothesis testing can offer valuable insights. Suppose we collected our samples and found that people living in cities have a 50% higher mean income than people living in rural areas. The real question is then as follows:

> *Is the difference in mean income between rural and urban areas real, or it is just a random result due to sampling error (caused by not asking the right people about their income)?*

In other words, can we state in a statistically sound way that this reflects the entire population? Remember that, even if we have the perception that people usually tend to earn more money in cities, we have to prove it or give a statistical context that supports our belief. Otherwise, it is just an opinion.

We could use the following hypotheses (statements) for our research question:

- H_0 = *People in cities do not have a mean income different from that of people in rural areas (null hypothesis).*
- H_1 = *People in cities have a mean income different (50% higher) from that of people in rural areas (alternative hypothesis).*

The two samples here are "people living in cities" (e.g., 1,000 people asked about their income during the last year) and "people living in rural areas" (e.g., 1,000 people asked about their income during the last year). The population characteristic to be analyzed is "income."

To answer this question (or, more correctly, to attempt to come to a conclusion), we have to run a significance test. If the entire population could be measured, we would not need such tests. We could decide which statement is correct based on the entire population. As this is usually impossible, significance tests are necessary.

- Rejecting the null hypothesis
  If the null hypothesis is rejected, we reject the hypothesis that the two samples and their distributions are the same (*not different*). In other words, there is a "good chance" that there is a difference between the distributions, and this difference is not just a result of randomness. This means that there are some patterns (reasons) by which these distributions are different. The "good chance" is estimated with a probability value using the significance level.

  If we select $\alpha = 0.05$ as the significance level and the resulting p-value calculated by the test is $p = 0.0015$, then we can reject the null hypothesis because $p < \alpha$. This means that H_1 is true and that there are differences between urban and rural incomes. The research hypothesis that income in cities is 50% higher than that in villages (as calculated earlier) is now statistically backed up. This result is typically expressed in statistical language as follows:

  > *People living in urban areas have an income statistically different (50% higher) than that of people living in rural areas at the 5% significance level.*

  The chance of finding the observed differences (e.g., in mean income) if the null hypothesis is true (no differences) is only $p = 0.0015$, or 0.15%. This can be stated using the significance level as follows:

  > *There is a less than 5% chance that this difference is the result of sampling error.*

  In other words, we have a 95% ($100\% - \alpha$) probability (confidence) that our distributions are different.

  Again, the results of all statistical significance tests are based on probability distribution functions. They produce probabilities, not certainties. In the preceding example, the conclusion is not that the "distributions are different" but that the "distributions are likely to be different with a probability of 95%." There is always a 5% chance that the distributions are similar. As a result, the smaller the significance level, the higher the chance that the results are close to the real values. The 95% probability in our example means that, if we randomly selected 100 different samples

from a population, people living in cities would have incomes 50% higher than people living in rural area in 95 of the cases. More generally, 95 of the cases would reject the null hypothesis. Still, there is always the chance that five samples would not display the difference hypothesized.

- Not rejecting the null hypothesis
  Suppose that, for the same $\alpha = 0.05$ significance level, the resulting p-value calculated by the test were $p = 0.065$. In this case, we cannot reject the null hypothesis because $p > \alpha$. If we cannot reject the null hypothesis, we have to state the following:

  We have insufficient evidence to reject the null hypothesis that people in cities do not have a mean income different from that of people in rural areas.

  Not rejecting the null hypothesis does not mean that we accept it, nor that the observed difference among the distributions is wrong. As we cannot reach a solid conclusion, we have to carry out other experiments and use other methods to decide whether incomes differ between rural and urban areas.

2.6 Normal Distribution Use in Geographical Analysis

Importance to Spatial Analysis

Spatial analysis is commonly used to study either the distribution of the locations of events/polygons or the spatial arrangement of their attributes (i.e., socioeconomic variables). As geographical analysis is an interdisciplinary field, many reports and research papers use purely statistical methods to supplement the core spatial analysis. For example, suppose we want to analyze the spatial distribution of income in relation to educational attainment. It is reasonable to begin with classic statistical analysis like the calculation of the Pearson's correlation coefficient between income and educational attainment. Spatial statistics can then be applied such as bivariate spatial autocorrelation to determine if these two variables tend to spatially cluster together. In geographical analysis, spatial statistics go along with classical statistics.

Statistics often deal with a normal distribution because many well-defined statistical tests (e.g., Pearson's correlation coefficient, analysis of variance, t-test, regression, factor analysis) are based on the assumption that the examined distribution is normal. If the observed distribution does not resemble a normal distribution (e.g., is skewed), then many statistical procedures are not accurate and cannot be used or should be used with caution. For example, if a distribution is skewed, the probability of having small values (compared to large values) differs. If data are collected through random sampling, different proportions of values are highly likely to fall into specific intervals. This will lead to an overrepresentation or underrepresentation of specific values in the sample that should be taken into account. We should highlight here that most

statistics are based on a normal underlying distribution for attribute values and on a Poisson probability distribution for point patterns (Anselin 1989 p. 4). Poisson probability distribution is used to assess the degree of randomness in point patterns (Oyana & Margai 2015 p. 75, Illian et al. 2008 p. 57).

How to Identify a Normal Distribution

There are three simple methods of determining if a distribution is normal or not:

1. Create a histogram and superimpose a normal curve. Plot inspection can enable a rough estimation of whether the distribution approximates the normal curve.
2. Calculate the skewness and kurtosis for the distribution. If the distribution is skewed and/or kurtosis is high/low, we have a clear indication that the distribution is not normal (see Section 2.2.4).
3. Create a normal QQ plot (see Section 2.2.9).

What to Do When Distribution Is Not Normal

If the distribution is not normal, we have three options (we assume that outliers have been removed):

Option 1. Use nonparametric statistics (see Table 2.4).

Option 2. Apply variable transformation. An efficient way to avoid a non-normal distribution is to transform it (if possible) to a normal distribution. Table 2.5 presents transformations that can be used to transform a variable according to its skewness value. This transformation will not necessarily lead to a normal distribution, but there is a good chance that it will.

Option 3. Check the sample size. According to the central limit theorem, if the sample is larger than 30–40, parametric statistics may be used without affecting the results' credibility. This theorem states that, given certain conditions, as the size of a random sample increases, its distribution approaches a normal distribution. In other words, even if our distribution is not normal, we can use parametric statistics if we have a large sample (de Vaus 2002 p. 78). Such a violation of the normality assumption does not cause major problems (Pallant 2013). It is not easy to define the ideal value by which a sample can be regarded as large. According to the literature, values of 30 to 40 are regarded as sufficient for a sample to be considered large and follow the central limit theory. In spatial analysis, this means that we need a sample of more than 30 spatial entities (e.g., postcodes, cities, countries) to use parametric statistics. When fewer spatial entities are involved, it is essential to check the variables for normality if we want to make inferences for a larger population.

Table 2.5 Transformations to reduce skewness and restore normality.

Skew	Transformation	Formula	Used When
High positive skew	Reciprocals or Negative reciprocal. Reciprocal of a ratio may often be interpreted as easily as the ratio. For example: population density (people per unit area) becomes area per persons.	$Y_n = 1/Y$ $Y_n = -1/Y$ Add 1 in values less than 1 up to 0	Reciprocals are useful when all values are positive. The negative formula is used to transform negative values. Reciprocal transformation reverses order among positive values: largest becomes smallest.
Moderate to low positive skew	Square root Logarithmic transformation	Y_n = Y^0.5 Y_n = logy or Y_n=lny	May have many zero's or very small values. May have a physical exponent (e.g., area). May have only positive values.
Low to moderate negative skew	Power square	Y_n = Y^2	May have a logarithmic trend (decay, survival, etc.).
High negative skew	Power cube	Y_n = Y^3	May have a logarithmic trend (decay, survival, etc.).

2.7 Chapter Concluding Remarks

- Exploratory spatial data analysis and related tools offer a comprehensive visual representation of statistics by linking graphs, scatter plots or histograms with maps.
- Data are just numbers stored in tables in a database management system. By analyzing data, we add value, creating information and then knowledge.
- Spatial statistics employ statistical methods to analyze spatial data, quantify a spatial process, discover hidden patterns or unexpected trends and model these data in a geographic context.
- Choropleth maps are typically the first thing created after a geodatabase is built.
- Designing and rendering choropleth maps is an art form. As professionals, we have to create accurate maps and graphs, always use solid statistics to back up our findings and avoid misleading messages.
- Inspecting the basic characteristics of a variable is essential prior to any sophisticated statistical or spatial analysis. Calculating the mean value, maximum value, minimum value and standard deviation of a variable provides an initial description of its distribution.
- Creating a frequency distribution histogram and inspecting for potential normality are also necessary.

- In general, calculating the common measures of center, shape and spread gives quick insights into the dataset and should be conducted prior to any other analysis.
- Boxplots are very helpful descriptive plots, as they depict measures of center, shape and spread and allow for comparison of two or more distributions.
- Scatter plot matrices and pairwise correlation matrices give a snapshot of the relationships among pairs of variables in a dataset. It is advisable to build such plots right from the beginning to gain quick insights into the data.
- Locating outliers is necessary, as statistical results might be distorted if outliers are not removed or properly handled.
- Rescaling variables with large differences in their scale and range, through normalization, adjustment or standardization is necessary when want to compare them or to use them in the same formula.
- While observations independence should exist in classical statistics, spatial dependence usually exists in spatial statistics, and classical statistics should be modified accordingly.
- A test of significance is the process of rejecting or not rejecting a hypothesis based on sample data. It is used to determine the probability that a given hypothesis is true.
- A p-value is the probability of finding the observed (or more extreme) results of a sample statistic (test statistic) if we assume that the null hypothesis is true.
- Rejecting the null hypothesis means that there is a probability (calculated as the difference: $100\% - \alpha$) that alternative hypothesis H_1 is correct (α is the significance level).
- Not rejecting the null hypothesis means that there is not sufficient evidence to reject the null hypothesis, but we cannot accept it either without further analysis.
- Statistics often deal with normal distribution because many well-defined statistical tests are based on the assumption that the examined distribution is normal.