4

Statistical measures (or quantities)

Chapter 4 concentrates on the basic descriptive measures or quantities that can be derived to compare and contrast sets of numbers in their role as measuring quantifiable differences between attributes and variables. Attention is directed towards summary measures referring to numbers, such as the median, mean, range, and standard deviation, and to the location of geographical phenomena, such as mean centre and standard distance. This chapter establishes the characteristics of the basic measures and quantities that may be subject to statistical testing.

Learning outcomes

This chapter will enable readers to:

- explain the purpose of spatial and nonspatial descriptive statistics for comparing and contrasting numerical data;
- identify the appropriate summary measures to use with variables recorded on different measurement scales;
- start to plan how to use descriptive statistics in an independent research investigation in Geography, Earth Science and related disciplines.

4.1 Descriptive statistics

Descriptive statistics are one of two main groups of statistical techniques, the other is known as inferential statistics or hypothesis testing. Descriptive statistics are about obtaining a numerical measure (or quantity), or a frequency count to describe the

Practical Statistics for Geographers and Earth Scientists Nigel Walford

characteristics of attributes and variables of observations in a dataset. We will look at frequency distributions, the principles of hypothesis testing and inferential statistics in Chapter 5, but for the present the latter can be defined as a group of analytical techniques that are used to find out if the descriptive statistics for the variables and attributes in a dataset can be considered as important or significant. Descriptive statistics are useful in their own right and in most geographical investigations they provide an initial entry point for exploring the complex mass of numbers and text that makes up a dataset. Some research questions can satisfactorily be answered using descriptive statistics on their own. However, in many cases this proves insufficient because of an unfortunate little problem known as **sampling error**, to which we will return a number of times in this text.

Descriptive statistics can be produced for both population and sample data, when they relate to a population they are known as **parameters**, whereas in the case of a sample they are simply, but perhaps slightly confusingly, called **statistics**. The purpose of descriptive statistics is to describe individual attributes and variables in numerical terms. We are all familiar with descriptions of people, places and events in books, Box 4.1a reproduces the first paragraph from Jane Austen's novel *Pride and Prejudice*, but try to follow the narrative by reading the words when they have been jumbled (Box 4.1b). It is worse than looking at a foreign language, since the individual words are familiar, but their order makes little sense. Even if the words were reorganized into a sensible order, this may not convey the meaning intended by the original author. We can make a similar comparison with a set of numbers. In Box 4.1c the numbers generated by the UK's National Lottery™ over 26 Saturdays in 2003 have been listed in the order in which they appeared. The range of possible numbers for each draw runs from 1 to 49. Were some numbers selected more often than others over this period? Were some numbers omitted? Were numbers above 25 picked more or less often than those below this midway point between 1 and 49? It is difficult to answer these questions until the numbers are presented in numerical order and Box 4.1d enables us easily to see that 10 and 12 appeared more often than any other number (7 times each), 40 is the only number not to appear in the draw during this 26-week period and that 86 of the numbers were below 25 and 90 were above. This example illustrates that the simple process of reorganizing the numbers into the sequence from lowest to highest allows us to derive some potentially useful information from the raw data. Nevertheless, the sorted sequence still includes 182 numbers (26 weeks $\times$ 7 numbers) and there is scope for reducing this detail further to obtain more summary information from the raw data. Indeed, an alternative name for the process of producing descriptive statistics is **data reduction**. Sorting the words of Jane Austen's first paragraph alphabetically is less helpful in conveying a sensible meaning (see lower part of Box 4.1b), although it does allow us to see that the word 'A' was used four times and both 'in' and 'of' twice each.

Which combination of six main numbers would seem to offer the best chance of winning the jackpot prize? Why would using this set of numbers fail to guarantee winning the jackpot?

Box 4.1: Comparison of data reduction in respect of textual and numerical data.

(a) Ordinary paragraph

'It is a truth universally acknowledged that a single man in possession of a good fortune must be in want of a wife'.

Pride and Prejudice by Jane Austen

(b) Jumbled paragraphs

Of man a in single a fortune be good universally wife that acknowledged want it truth is of possession a in that a.

A a a a acknowledged be fortune good in in is it man must of of possession single that truth universally want wife

(c) 182 Lottery numbers

08/02/03: 43, 32, 23, 37, 27, 41, **35,**
22/02/03: 30, 19, 42, 33, 38, 44, **31,**
01/03/03: 29, 31, 45, 44, 22, 24, **35,**
15/03/03: 47, 28, 04, 25, 38, 24, **11,**
22/03/03: 45, 10, 47, 49, 08, 02, **09,**
29/03/03: 33, 42, 05, 15, 35, 21, **26,**
12/04/03: 45, 25, 03, 16, 05, 43, **23,**
19/04/03: 48, 08, 38, 31, 13, 10, **14,**
26/04/03: 35, 27, 21, 09, 48, 33, **18,**
03/05/03: 07, 03, 49, 46, 06, 29, **37,**
10/05/03: 26, 32, 18, 08, 38, 24, **31,**
17/05/03: 48, 15, 36, 12, 08, 22, **37,**
31/05/03: 22, 25, 02, 09, 26, 12, **21,**
14/06/03: 46, 43, 37, 03, 12, 29, **11,**
21/06/03: 36, 20, 45, 12, 39, 44, **49,**
28/06/03: 23, 17, 35, 10, 01, 29, **36,**
05/07/03: 10, 18, 06, 19, 22, 43, **08,**
12/07/03: 15, 13, 21, 12, 09, 33, **43,**
19/07/03: 09, 34, 03, 17, 10, 48, **36,**
26/07/03: 27, 49, 16, 01, 12, 26, **23,**
02/08/03: 45, 28, 49, 47, 05, 26, **08,**
09/08/03: 41, 25, 14, 30, 19, 11, **38,**
16/08/03: 01, 09, 42, 45, 12, 10, **27,**
23/08/03: 42, 25, 15, 10, 18, 48, **11,**
30/08/03: 18, 38, 43, 44, 26, 33, **02,**
06/09/03: 21, 25, 39, 01, 27, 42, **17**

(d) Sorted lottery numbers

01	01	01	01	**02**	02	02	03
03	03	03	04	05	05	05	06
06	07	**08**	**08**	08	08	08	08
09	**09**	09	09	09	09	10	10
10	10	10	10	10	**11**	**11**	**11**
11	12	12	12	12	12	12	12
13	13	**14**	14	15	15	15	15
16	16	**17**	17	17	**18**	18	18
18	18	19	19	19	20	**21**	21
21	21	21	22	22	22	22	**23**
23	23	23	24	24	24	25	25
25	25	25	25	**26**	26	26	26
26	26	**27**	27	27	27	27	28
28	29	29	29	29	30	30	**31**
31	31	31	32	32	33	33	33
33	33	34	**35**	**35**	35	35	**35**
36	**36**	36	36	**37**	**37**	37	37
38	38	38	38	38	38	39	39
41	41	42	42	42	42	42	**43**
43	43	43	43	43	44	44	44
44	45	45	45	45	45	45	46
46	47	47	47	48	48	48	48
48	49	49	**49**	49	49		

Note: 'Bonus Ball' numbers are shown in bold.

There are two main, commonly used groups of summary measures or quantities known as **measures of central tendency** and **measures of dispersion**. These are examined in detail later in this chapter. The first group includes the mode, median and mean (commonly called the average). The main measures of dispersion are the range, interquartile range, variance and standard deviation. The two sets of measures provide complementary information about numerical measurements, in the first case

a central or typical value, and in the second how 'spread out' (dispersed) the data values are. The choice of which measure to use from each group depends largely upon the measurement scale of the attribute or variable in question. Most analyses will involve using two or three of the measures from each group according to the nature of the data. In addition to summarizing the central tendency and dispersion of a numerical distribution, it may be important to examine two further characteristics, namely its **skewness** and its **kurtosis** ('peakedness'). Measures of skewness describe whether the individual values in a distribution are symmetrical or asymmetrical with respect to its mean. Measures of kurtosis indicate the extent to which the overall distribution of data values is relatively flat or peaked.

4.2 Spatial descriptive statistics

One thing setting geography and other geo (earth) sciences apart from many other subjects is an interest in investigating variations in the **spatial location** of geographically distributed phenomena as well as their attributes and variables. In some respects the spatial location of geographical phenomena can be considered as 'just another attribute or variable', but in practice where phenomena are located in respect of each other and the patterns thus produced are often of interest in their own right. When measuring differences between nonspatial phenomena we concentrate on the values of one or more attributes and variables, but usually ignore the position of each observation in the set. Thus, a nonspatial analysis of sampled households is likely to have little interest in distinguishing between, for example, whether any of the households live next to each other or how far apart any pair of households lives. Spatially distributed phenomena have the extra dynamic that their position can also be quantified. Of course, most measurable entities, even if infinitely small or large, are spatially located, what distinguishes 'spatially aware' disciplines is that this spatial distribution itself is not dismissed as a chaotic, random occurrence, but as being capable and deserving of interpretation.

Suppose a researcher is concerned about variations in the quality of households' accommodation and wellbeing. Respondents in each sampled household might be interviewed and information obtained about the number of rooms and their size, the availability of various amenities (e.g. central heating, air conditioning, mains gas and electricity), the entrance floor level, the number of people in the household, the presence/absence of a swimming pool, etc. Analysis of these data might reveal certain nonspatial associations or connections, for instance a large living area per person and the presence of a swimming pool may both indicate relative financial and social advantage. However, if analysis of the survey data reveals households possessing these and other similar combinations of characteristics live close together, then perhaps we can also conclude that this is a relatively wealthy suburban area. In other words, the physical distance between phenomena and the patterns they display may add to the analysis of their socioeconomic characteristics. In this situation, we can start to say

something about the characteristics of the places people occupy as well as the people themselves.

The underlying reason for geographers investigating spatial distributions is that they are indicative of the outcome of a process of competition for space amongst the phenomena of interest. In some plant communities, for example, there may be a mix of species that flower at different times of the year and so it does not matter too much if there is some variation in height, provided that tall, medium and short plants do not flower at the same time. Similarly, if a gardener tries to grow too many plants in a limited space, some will thrive, whereas others will decline and possibly die. In a commercial context, a farmer producing apples will try to ensure that the fruit trees are planted an equal distance apart in order to maximize each tree's exposure to sunlight, rainfall, soil nutrients and to facilitate harvesting the fruit. However, suppose the farmer decides to grub up the apple trees and sow the field to grass in order to graze cattle. The free movement of these animals within the field is normally unconstrained and their competing demand for fresh herbage is controlled not by locating the animals at different locations in the field, but by limiting the overall number within this bounded portion of space and allowing the animals to roam freely in search of fresh pasture. Competition for space is therefore rarely, if ever, an unrestricted process, since there is normally some form of controlling or managing function in place, which may be either 'natural' (e.g. competing demand for exposure the sun's radiation), or 'human' (e.g. restricting demand for access to fresh grass).

Figure 4.1 follows on from these examples by illustrating the three main ways in which phenomena, in this case cars parked in the long-stay car park at Stansted airport, can be distributed spatially. In some parts of the car park the cars are parked in a regular fashion either in pairs next to each other or singly with one or two spaces in between. In other parts there are cars parked in small clusters with between 4 and 8 cars in each. However, taking an overall view the cars would seem to be parked in a fairly random fashion with some close together and others further apart, some forming small groups and others on their own. Figure 4.1 presents a snapshot of the spatial pattern of cars in this car park, but thinking about the process whereby cars arrive and depart an airport car park helps in understanding how the pattern may have arisen. It provides an interesting example of the dynamic interrelationships between space and time.

Where might drivers arriving at this section of the car park their cars? How will length of stay in the car park influence the spatial pattern? How might the way the operator of the car park influence the spatial pattern by controlling car arrivals and departures?

This interest in geographical patterns has led to the development of a whole series of spatial statistics, sometimes called **geostatistics** that parallel the standard measures used by researchers in other fields. Descriptive spatial statistics aim to describe the distribution and geometrical properties of geographical phenomena. For example, lines can be straight or sinuous; areas may be large or small, regular (e.g. a square) or

Figure 4.1 Spatial distribution patterns. Source: GeoInformation Group (adapted).

irregular (e.g. a polygon). Although such spatial statistics are not relevant in all geographical investigations, it is appropriate to complement explanations of the standard measures with descriptions of their spatial equivalents.

The division of statistical procedures into descriptive and inferential categories also applies to spatial statistics. Numerical measures and frequency counts that can be derived to describe the patterns produced by spatial phenomena either relate to one occasion, as in the snapshot of car parking shown in Figure 4.1, or as they change between different times, which could be illustrated by continuous filming of cars arriving and departing the car park. Geographical phenomena are usually divided into three basic types when examining spatial patterns, namely points, lines and areas. However as we saw in Chapter 2, to the question of scale can sometimes confuse this simple classification, especially between points and areas. Figure 4.2 shows the complete aerial photograph from which the extract in Figure 4.1 was taken. Zoomed out this distance the individual cars in certain parts of the car park can still be identified, but in the large central area (highlighted) they have virtually merged into one

Figure 4.2 The blurring of individual points into areas. Source: GeoInformation Group (adapted).

mass and just about transformed into an area, although the circulation routes between the parking spaces enable lines of cars to be identified. Clearly few geographers or Earth scientists spend their time analysing the patterns produced by cars in parking lots, but nevertheless this example illustrates that the scale at which spatial phenomena are considered not only affects how they are visualized but also how their properties can be analysed.

Measures of central tendency and dispersion also form the two main groups of descriptive spatial statistics and include equivalent descriptors to their nonspatial counterparts. The details of these are examined later in this chapter. These measures focus in the first case on identifying a central location within a given collection of spatial phenomena and in the second on whether the features are tightly packed together or dispersed. They are sometimes referred to as **centrographic techniques** and they provide a numerical expression of the types of spatial distribution illustrated in Figure 4.1.

4.3 Central tendency

4.3.1 Measures for nonspatial data

We have already seen that a certain amount of descriptive information about the numerical measurements for a population or sample can be obtained simply by sorting them into ascending (or descending) order. However, calculating a single quantity representing the central tendency of the numerical distribution can reveal more. There are three main measures of central tendency known as the mode, median and mean, and each provides a number that typifies the values for a nonspatial attribute or variable. The procedures for calculating these measures are outlined in Box 4.2 with respect to a variable measured on the ratio/interval scale, water temperature in a fluvioglacial stream recorded at half-hourly intervals during daytime. The lowest of the central tendency measures is the mean (9.08); the mode is highest (11.40) with the median (10.15) in between. Thinking about the variable recorded in this example and the time period over which measurements were taken, this particular sequence is not surprising. Undoubtedly it takes some time for a rise in air temperature to warm up the water melting from the glacier, which can be sustained by the warmth from the midday and early afternoon sunshine, when the peak temperature occurred. Recording of water temperature stopped sooner after this peak than it had started before it in the morning. It took five hours to reach the peak (09.00–14.00 hrs), but recording stopped three and a half hours afterwards (17.30 hrs). In other words, the length of the period over which the data measurements were captured has possibly had an impact on the statistical analysis.

Why are there three measures of central tendency?

Each has its relative strengths and weaknesses and should be used with attributes or variables recorded according to the different scales of measurement. If a large number of values are possible, then the mode is unlikely to provide much useful information. At one extreme it might simply record that each value occurs only once or, slightly more promisingly, that one out of 50 values is repeated twice or perhaps even three times. In other words, the mode is only really helpful when dealing with nominal attributes that have a limited range of values representing the labels attached to the raw data. For example, respondents in a survey may be asked to answer an attitude battery question with 'Strongly Agree', 'Agree', 'Neutral', 'Disagree' or 'Strongly Disagree' as the possible predefined responses. Suppose further that these responses have been assigned numerical codes from 1 to 5 respectively. Although the number of potential values has been limited, it is entirely feasible that the mode will be one of the extreme values (1 or 5). Hence, the notion of central tendency with respect to the mode relates the most common nominal category, since in this example the values 1 to 5 do not imply any order of magnitude in measurement and there is no requirement that 'Strongly Agree'

Box 4.2a: Central tendency measures for nonspatial data: Sample site for the measurement of water temperature in a fluvioglacial stream from Les Bossons Glacier, France.

Mean – population symbol: μ; sample symbol: $\bar{x}$

Box 4.2b: Calculation of the mode, median and mean.

The various measures of central tendency are perhaps some of the most intuitive statistics (or parameters) available. Their purpose is to convey something about the typical value in an attribute or variable and allow you to quantify whether the central values in two or more sets of numbers are similar to or different from each other. Each of the main measures (mode, median and mean) can be determined for variables measured on the interval or ratio scale, ordinal variables can yield their mode and median, but only the mode is appropriate for nominal attributes. So, if you have some data for two or more samples of the same category of observations, for instance samples of downtown, suburban and rural households, then you could compare the amounts of time spent travelling to work.

The mode is the most frequently occurring value in a set of numbers and is obtained by counting. If two or more adjacent values appear the same number of times, the question arises as to whether there are two modes (i.e. both values) or one, midway between them. If no value occurs more than once, then a mode cannot be determined. The median is the value that lies at the midpoint of an ordered set of numbers. One way of determining its value is simply to sort all the data values into either ascending or descending order and then identify the middle value. This works perfectly well if there is an odd, as opposed to even, number of values, since the median will necessarily be one of the recorded values. However, if there is an even number, the median lies halfway between the two observations in the middle, and will not be one of the recorded values when the two middle observations are different. Calculation of the arithmetic mean involves adding up or summing all the values for a variable and then dividing by the total number of values (i.e. the number of observations).

The methods used to calculate the mode, median and mean are illustrated below using half-hourly measurements of water temperature in the fluvioglacial stream from Les Bossons Glacier near Chamonix in the European Alps (see Box 4.2a).

	Mode	Median		Mean	$\bar{x} = \frac{\sum x}{n}$
	x	x	Sorted x		x
09.00 hrs	4.30	4.30	4.30		4.30
09.30 hrs	4.70	4.70	4.70		4.70
10.00 hrs	4.80	4.80	4.80		4.80
10.30 hrs	5.20	5.20	5.20		5.20
11.00 hrs	6.70	6.70	6.70		6.70
11.30 hrs	10.10	10.10	7.80		10.10
12.00 hrs	10.50	10.50	8.80		10.50
12.30 hrs	11.20	11.20	9.30		11.20
13.00 hrs	**11.40**	11.40	10.10		11.40
13.30 hrs	11.80	11.80	10.20		11.80
14.00 hrs	12.30	12.30	10.50		12.30
14.30 hrs	11.90	11.90	11.10		11.90
15.00 hrs	**11.40**	11.40	11.20		11.40
15.30 hrs	11.10	11.10	11.40		11.10
16.00 hrs	10.20	10.20	11.40		10.20
16.30 hrs	9.30	9.30	11.80		9.30
17.00 hrs	8.80	8.80	11.90		8.80
17.30 hrs	7.80	7.80	12.30		7.80
					$\sum x = 163.50$
			$\frac{10.10 + 10.20}{2}$		$\frac{163.50}{18}$
	Mode = 11.40		Median = 10.15		$\bar{x} = 9.08$

should have been labelled 1, 'Agree' as 2, etc. It is simply a matter of arbitrary convenience to allocate the code numbers in this way – they could just as easily have been completely mixed up ('Strongly Agree' as 3, 'Agree' as 5, 'Neutral' as 4, 'Disagree' as 1 and 'Strongly Disagree' as 2). The median is, by definition, more likely to provide a central measure within a given set of numbers. The main drawback with the median is its limited focus on either a single central value or on the two values either side of the midpoint. It says nothing about the values at either extreme.

The arithmetic mean is probably the most useful measure of central tendency, although its main feature is recognized as both a strength and weakness. All values in a set are taken into account and given equal importance when calculating the mean, not just those at the extremes or those in the middle. However, if the values are asymmetrical about the mean, for example there are a few outlying values at either the lower or upper extreme, then these may 'pull' the mean in that direction away from the main group of values. Trimming values at both extremes, for example by ignoring the lowest and highest 5% or 10% of values may reduce this effect, but the decision to do so represents a subjective judgement on the part of the investigator, which may or may not be supported from either a theoretical or statistical perspective. Another advantage of the mean over the median is that it can easily be calculated from totals (i.e. total count of phenomena and total sum of data values). For example, the 2001 UK Population Census reports that there were a total of 3 862 891 persons living in 1 798 864 privately rented furnished housing units in England, from which it can readily be determined that the mean number of persons per dwelling unit under this form of tenure was 0.47 (1 798 864/3 862 891). The equivalent median value cannot be determined in this fashion. The arithmetic mean possesses two further important properties. First, if the mean is subtracted from each number in the set, then these differences will add up to zero; and, secondly, if these positive and negative differences are squared and then summed, the result will be a minimum (i.e. the sum of the squared differences of any number apart from the mean would be larger). Therefore, unless the median and mean are equal, the sum of the squared differences from the median would be greater than those from the mean. These properties of the mean might at this point seem somewhat esoteric, although intuitively they suggest that it really does encapsulate the central tendency of a set of data values.

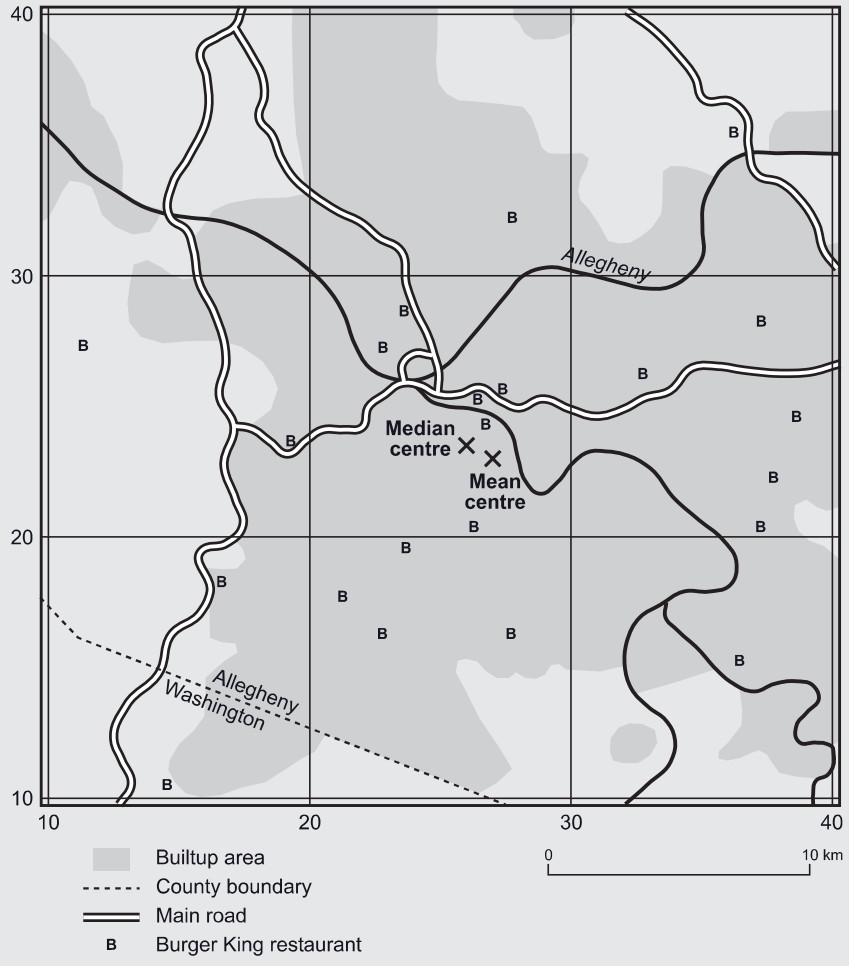
4.3.2 Measures for spatial data

The equivalent measures of central tendency with respect to spatial phenomena represented graphically as points are known as the **median centre** and the **mean centre**, both of which can be weighted according to the numerical values associated with attributes or variables quantified with respect to the phenomena occurring at each point. The procedures involved with determining the median and mean centres, and the latter's weighted equivalent, are given in Box 4.3 using the example of the location of Burger King food outlets in Pittsburgh. Box 4.3a indicates that the distribution of

the companies' outlets is concentrated downtown, but that there are also some outlying enterprises. The question is whether the overall centre of the corporation's presence in the city is in the urban core or elsewhere. The mean and median centres are relatively close together in this example and can be regarded as providing a reasonable summary of the centre of the spatial distribution and as indicating that the Burger King has good coverage across the city as a whole.

Box 4.3a: Location of Burger King restaurants as points in Pittsburgh.

Mean centre – population symbol: $\bar{X},\bar{Y}$; sample symbol: $\bar{x},\bar{y}$



Box 4.3b: Calculation of the mean and median centres of Burger King Restaurants in Pittsburgh.

Measures of central tendency for spatially distributed phenomena serve much the same purpose as their nonspatial counterparts – they capture the typical spatial location rather than a typical value. In the case of spatial phenomena this is a physically central location in the distribution. Calculation of the mean and/or median centre for the distributions of two or more spatial phenomena within a given area enables you to determine if their centres are relatively close or distant. Commercial enterprises competing with each other for customers may seek to open outlets in locations where they can attract each other's clientele. In this case you might expect the mean centres of each corporation's outlets to be reasonably close to each other. In contrast, if enterprises obtained a competitive advantage by operating at a greater distance from rival companies, then their mean centres would be further apart.

The median and mean centres are normally calculated from the X and Y coordinate grid references that locate the phenomena in Euclidean space. The median centre is determined by interpolating coordinate values midway between the middle pairs of X and Y data coordinates. The median centre lies at the intersection of the two lines drawn at right angles to each other (orthogonal lines) through these points and parallel with their corresponding grid lines. This procedure partitions the point phenomena into 4 quadrants (NW, NE, SW and SE). However, this procedure will not necessarily assign equal numbers of points to each quadrant. An alternative method involves dividing the points into four equal-sized groups in each of the quadrants and then determining the median centre as where two orthogonal lines between these groups intersect. The problem is that many such pairs of lines could be drawn and hence the median centre is ambiguous.

The mean centre is far more straightforward to determine and simply involves calculating the arithmetic means of the X and Y coordinates of the data points. The mean centre is located where the two orthogonal lines drawn through these 'average' coordinates intersect. The mean centre is thus unambiguously defined as the grid reference given by the mean of the X and Y coordinates.

The methods used to calculate the mean and median centres are illustrated below using points locating Burger King outlets in Pittsburgh, denoted by subscript B for X and Y coordinates (see Box 4.3a).

	$\bar{x} = \frac{\sum x_B}{n}$	$\bar{y} = \frac{\sum y_B}{n}$	Median	
	x_B	y_B	Sorted x_B	Sorted y_B
1	11	27	11	10
2	14	10	14	15
3	16	18	16	16
4	19	23	19	16
5	21	17	21	17
6	22	16	22	18
7	22	27	22	19
8	23	19	23	20
9	23	28	23	20

	$\bar{x} = \frac{\sum x_B}{n}$	$\bar{y} = \frac{\sum y_B}{n}$	Median	
	x_B	y_B	Sorted x_B	Sorted y_B
10	26	20	26	22
11	26	24	26	23
12	26	25	26	24
13	27	16	27	24
14	27	25	27	25
15	27	32	27	25
16	32	26	32	26
17	36	15	36	27
18	36	35	36	27
19	37	20	37	28
20	37	22	37	28
21	37	28	37	32
22	38	24	38	35
	$\sum x_B = 583$	$\sum y_B = 497$		
	$\frac{583}{22}$	$\frac{497}{22}$		
	$\bar{x}_B = 27$	$\bar{y}_B = 23$	Median $x_B = 26.0$	Median $y_B = 23.5$
	Mean centre = 27.0, 23.0			

The median centre shares the same problem as the (nonspatial) median, since it focuses on the centre of the spatial distribution and hence ignores possibly important outlying occurrences. Figure 4.3 illustrates how the median centre can occur with a range of *X* and *Y* coordinates in the shaded area at the intersection of the two sets of parallel dashed lines. Nevertheless, it is a useful, intuitively simple location for the centre of a spatial distribution. The mean centre also has some issues. A few observations located at the extremes of a distribution can lead to the mean centre being drawn in their direction. An alternative to the Median centre, especially in American texts is the Centre of Minimum Travel, which is located at the point of minimum total distance between all of the points within a given set. The mean centre lies within a rectangular area with its dimensions defined by boundary lines drawn north–south and west–east through most northerly, southerly, westerly and easterly points in the distribution. This is known as the Minimum Bounding Rectangle and defines the extreme extent of the distribution. Consequently, if a few or even one point lies at some distance from the remainder, its extreme *X* and/or *Y* coordinates will drag the mean centre away from where the majority of phenomena are located. Another seemingly anomalous result can occur if the distribution of points includes two or possibly more distinct clusters each of a similar size. The mean centre could fall between the groups. Similarly,

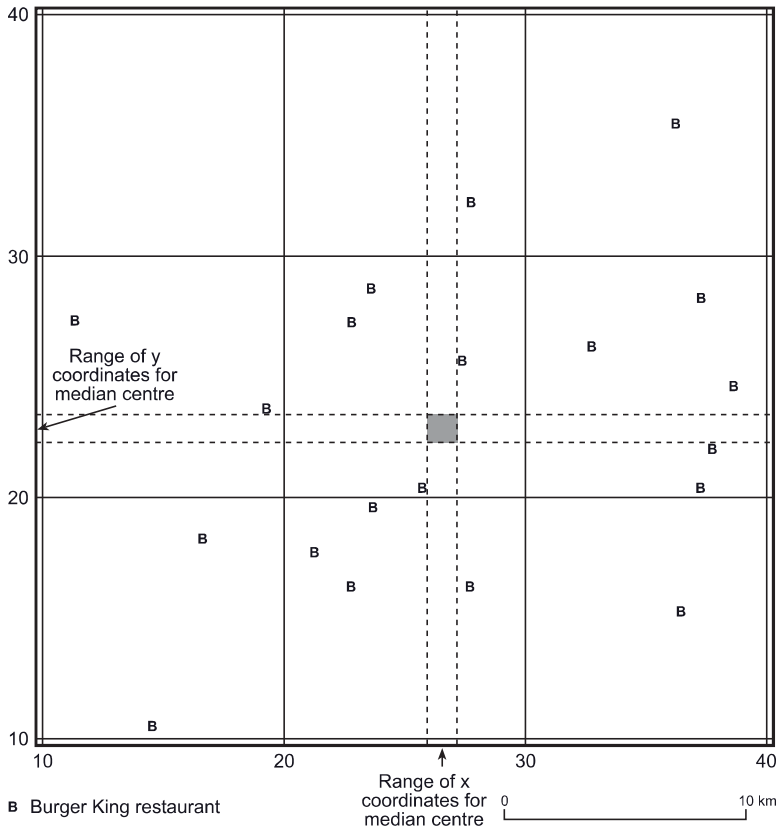


Figure 4.3 Square area containing alternative locations for the median centre.

if the points form a crescent-shaped group, the mean centre is likely to fall in space between the tips of the crescent where the phenomena themselves are not located. These problems do not entirely negate the value of the mean centre as a descriptive measure of the central tendency of spatial distributions, they simply re-emphasize the issue relating to the application of all statistical techniques – *caveat investigator*, let the researcher beware!

Weighting the location of the mean centre according to the values of a measured attribute or variable can enhance its usefulness. The distribution of spatial phenomena may be of interest in its own right, nevertheless, the underlying processes associated with producing the pattern is often equally important. Consequently, we are usually not only interested in the distribution of the phenomena, but also the values with respect to particular variables and attributes associated with each point, line or area. For example, if we were interested in finding out the mean centre of Burger King's customer base in Pittsburgh, a team of researchers could be positioned at each restaurant and count the number of people entering the premises on one day, assuming the corporation were not willing to release these data for reasons of their commercial

Box 4.4: Calculation of the weighted mean centre of Burger King Restaurants in Pittsburgh.

Weighted mean centre – population symbol: $\bar{X}_W$, $\bar{Y}_W$; sample symbol: $\bar{x}_w$, $\bar{y}_w$

The weighted mean centre takes into account the 'centre of gravity' in a spatial distribution, since it is a measure that combines the location of spatial phenomena with the values of one or more attributes or variables associated with them. If all of the spatial entities had the same data values, then the weighted mean centre would occur at exactly the same location as the 'basic' mean centre. However, if, as an extreme example, one entity in a set dominates the data values, then the weighted mean centre will be drawn towards that single case.

Calculation of the weighted mean centre involves multiplying the *X* and *Y* coordinates for each point by the corresponding data value(s) or possibly in more complex cases by the results of a mathematical function. The outcome of these calculations is a weighted *X* coordinate and a weighted *Y* coordinate, which are each totalled and divided by the sum of the weights to produce the weighted mean centre. This point is located where two lines at right angles to each other (orthogonal lines) drawn through these 'average' weighted coordinates intersect.

The method used to calculate the weighted mean centre is illustrated below using points locating Burger King outlets in Pittsburgh, denoted by subscript B for *X* and *Y* coordinates (see Box 4.3a). Note that the figures for daily customers (the weighting variable *w*) are hypothetical.

	$\bar{x} = \sum \frac{x_B}{n}$	$\bar{y} = \sum \frac{y_B}{n}$		$\frac{\sum x_B w}{\sum w}$	$\frac{\sum y_B w}{\sum w}$
	x_B	y_B	w	$x_B w$	$y_B w$
1	11	27	1200	13 200	32 400
2	14	10	7040	98 560	70 400
3	16	18	1550	24 800	27 900
4	19	23	4670	88 730	107 410
5	21	17	2340	49 140	39 780
6	22	16	8755	192 610	140 080
7	22	27	9430	207 460	254 610
8	23	19	3465	79 695	65 835
9	23	28	8660	199 180	242 480
10	26	20	7255	188 630	145 100
11	26	24	7430	193 180	178 320
12	26	25	5555	144 430	138 875
13	27	16	4330	116 910	69 280
14	27	25	6755	182 385	168 875
15	27	32	1005	27 135	32 160
16	32	26	4760	152 320	123 760
17	36	15	1675	60 300	25 125
18	36	35	1090	39 240	38 150

	$\bar{x} = \sum \frac{x_B}{n}$	$\bar{y} = \sum \frac{y_B}{n}$		$\frac{\sum x_B w}{\sum w}$	$\frac{\sum y_B w}{\sum w}$
	x_B	y_B	w	$x_B w$	$y_B w$
19	37	20	1450	53 650	29 000
20	37	22	2500	92 500	55 000
21	37	28	1070	39 590	29 960
22	38	24	1040	39 520	24 960
	$\sum x_B = 583$	$\sum y_B = 497$	93 025	2 283 165	2 039 460
	583	497		2 283 165	2 039 460
	22	22		93 025	93 025
	$\bar{x}_B = 27$	$\bar{y}_B = 23$		$\bar{x}_w = 24.5$	$\bar{y}_w = 21.9$
	Mean centre	27.0, 23.0		Weighted mean centre	24.5, 21.9

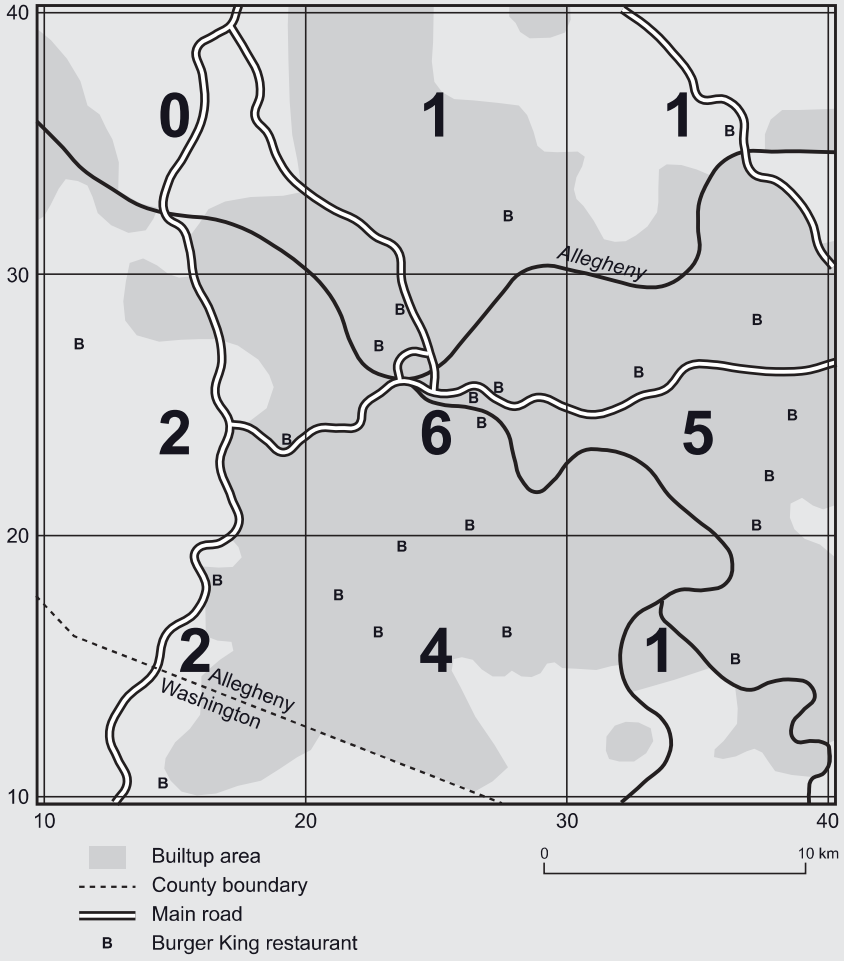
sensitivity. Each point in the distribution would then be weighted by multiplying its *X* and *Y* coordinates by the corresponding number of customers, then totalling the results and dividing them by the overall total (see Box 4.4). This process takes into account the size of some activity or characteristic occurring at the point locations, which can be thought of as a height measurement. Weighting the mean centre can also be achieved by applying a mathematical function that connects several variables together, for example the customers entering the Burger King premises during one day expressed as a percentage of the total number of people living or working within a 50 m radius.

One of the strengths of the arithmetic mean over the median was that it could be calculated from data where the individual values were unknown and the only information available was the total number of observations and the sum total of a variable. In a similar fashion the mean centre and its weighted equivalent can be estimated even if the coordinates for the individual points are missing. This may be useful if the phenomena can be aggregated into counts per regular grid square. Suppose that the exact *X* and *Y* coordinates of the Burger King restaurants in Pittsburgh were unknown, but there was a count of the number per 10×10 km grid square (see Box 4.5a). Conventionally, all the points within a square are assigned grid references representing one of the corners, typically the lower left, or the centre of the square, as in this case. In comparison with Box 4.3a, the detail of exactly where each outlet is located has been lost with their positions generalized to within a series of 10×10 km squares. Both the mean centre and the weighted version relating to such aggregate distributions of data points are clearly less accurate than those pertaining to phenomena whose exact grid coordinates are available. However, they represent useful alternative measures when analysing data obtained from secondary sources.

Box 4.5a: Allocation of Burger King restaurants to an arbitrary grid of squares in Pittsburgh.

Mean centre – population symbol: $\bar{X}$, $\bar{Y}$; sample symbol: $\bar{x}$, $\bar{y}$

Weighted mean centre – population symbol: $\bar{X}_W$, $\bar{Y}_W$; sample symbol: $\bar{x}_w$,$\bar{y}_w$



Box 4.5b: Calculation of the aggregate mean centre and the aggregate weighted mean centre of Burger King Restaurants in Pittsburgh.

The aggregate mean centre treats the spatial units, usually grid squares, in which entities are located as the primary unit of analysis. This results in some loss of detail, but may be useful where either the exact georeferenced locations of entities cannot be determined or there are so many occurrences that it is not necessary to take each one into account individually. Since

the aggregate mean centre generalizes the spatial distribution of entities, local concentrations of points can be obscured and its location will depend on which coordinates (e.g. bottom left corner or midpoint) of the superimposed grid are used. If the number of entities per grid square is fairly similar across the whole study area, then the grid unit representing the aggregate mean centre is likely to contain the 'true' mean centre of the individual point distribution. However, if, as an extreme example, one grid square includes the majority of cases with only a few in some of the others, then the aggregate mean centre will probably be in a different grid square to the one containing the 'true' mean centre.

Calculation of the aggregate mean centre involves using the X and Y grid references for the grid squares in which the individual entities occur. Multiply the chosen grid references by the number of entities within each square, then sum these values and divide the totals by the number of squares covered by the study area. This results in X and Y coordinates to a grid square that constitutes the aggregate mean centre. A weighted version of the aggregate mean centre can be calculated, which involves multiplying the count of points per square by the values of an attribute or variable.

The method used to calculate the aggregate mean centre is illustrated below in Box 4.5c using the number of Burger King restaurants in the six 10×10 km grid squares in Pittsburgh (see Box 4.5a) and grid references for the midpoint of the squares (e.g. 10.5, 20.5). The procedure for calculating the aggregate weighted mean centre is shown in Box 4.5d.

Box 4.5c: Calculation of the aggregate mean centre of Burger King restaurants in Pittsburgh.

			$\bar{x} = \frac{\sum x_B n}{\sum n}$		$\bar{y} = \frac{\sum y_B n}{\sum n}$	
	Square	Pts/sq n	x_B	$x_B n$	y_B	$y_B n$
1	10,10	2	10.5	21.0	10.5	21
2	10,20	2	20.5	21.0	20.5	41
3	10,30	0	30.5	0.0	30.5	0
4	20,10	4	10.5	82.0	10.5	42
5	20,20	6	20.5	123.0	20.5	123
6	20,30	1	30.5	20.5	30.5	30.5
7	30,10	1	10.5	30.5	10.5	10.5
8	30,20	5	20.5	152.5	20.5	102.5
9	30,30	1	30.5	30.5	30.5	30.5
		$\sum n = 22$		$\sum x_B n = 481$		$\sum y_B n = 401$
				$\frac{481}{22}$		$\frac{401}{22}$
				$\bar{x} = 21.9$		$\bar{y} = 18.2$
	Aggregate mean centre					

Box 4.5d: Calculation of the aggregate weighted mean centre of Burger King restaurants in Pittsburgh.

	Sq	Pts/sq n	Weight/sq w	$\bar{x} = \frac{\sum x_B w}{\sum w}$ x_B	$x_B w$	$\bar{y} = \frac{\sum y_B w}{\sum w}$ y_B	$y_B w$
1	10,10	2	8590	10.5	90 195.0	10.5	90 195.0
2	10,20	2	5870	20.5	61 635.0	20.5	120 335.0
3	10,30	0	0	30.5	0	30.5	0
4	20,10	4	18 890	10.5	387 245.0	10.5	198 345.0
5	20,20	6	45 085	20.5	924 242.5	20.5	924 242.5
6	20,30	1	1005	30.5	20 602.5	30.5	30 652.5
7	30,10	1	1675	10.5	51 087.5	10.5	17 587.5
8	30,20	5	10 820	20.5	330 010.0	20.5	221 810.0
9	30,30	1	1090	30.5	33 245.0	30.5	33 245.0
			93 025	$\sum x_B w = 1\,898\,263$		$\sum y_B w = 1\,636\,413$	
					$\frac{1\,898\,263}{93\,025}$		$\frac{1\,636\,413}{93\,025}$
				$\bar{x} = 20.4$		$\bar{y} = 17.6$	
Aggregate weighted mean centre					20.4, 17.6		

Unfortunately, mean centres based on aggregate counts of spatially distributed phenomena within areal units such as grid squares suffer from the problem that such units are arbitrarily defined. This arbitrariness arises from three decisions that have been taken about the nature of the grid squares: the size of the units; the position of the grid's origin; and its orientation. Figure 4.4 illustrates the impact of these decisions with respect to placing a regular square grid over the location of Burger King fast-food outlets in Pittsburgh. Three alternative grid square sizes – 500 m, 1000 m and 1500 m – and three different origins – x_1, y_1, x_2, y_2, and x_3, y_3 – are shown. Although calculations of the mean centres in each of these cases is not included, it will be apparent that decisions about the size and origin of regular grids can have a profound impact on aggregate spatial counts and hence on the location of the mean centre. There is, of course, nothing to dictate that a regular square, hexagonal or triangular grid has to be used, and once irregular, variable sized units are admitted, the arbitrariness becomes even more obvious. This example illustrates the **modifiable areal unit**

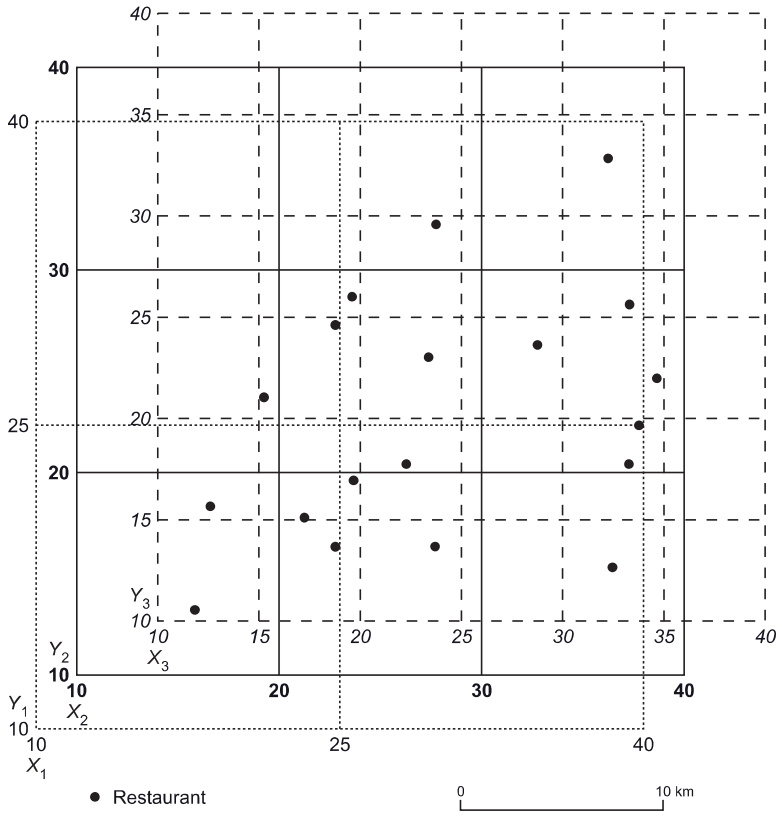


Figure 4.4 Alternative origins and sizes of regular square grid.

problem (MAUP), which is a common difficulty encountered when attempting to analyse spatial patterns.

4.3.3 Distance measures for spatial data

An alternative approach to analysing spatial patterns is to summarize distributions by reference to the **mean distance** between the set of spatial entities. Once this matrix of distances between each pair of points in the set has been determined, these measurements can be treated like any other variable that is quantified on the ratio scale and the overall mean or the **nearest-neighbour mean distance** can be calculated by summation and division. In general terms, a small mean distance implies that the collection of phenomena is more tightly packed together in comparison with a larger

mean distance. But, how small is small and how large is large? The answer partly depends on the size of the study area and how it is defined, which again raises issues associated with the MAUP. What might seem a short mean distance within a relatively large area may appear as a long distance in a small area. Box 4.6 illustrates the procedure for calculating the mean distance between the Burger King outlets in Pittsburgh. The overall mean distance is 13.0 km and the nearest-neighbour mean is 4.4 km, with modal and median values of 11.2 km and 12.0 km, respectively. This variation partly reflects the presence of some outlets outside the central downtown area. For comparison, the equivalent measures for the central downtown 10×10 km grid square are a mean of 4.3 km with the median and mode both 5.0 km, which suggests the outlets are much closer together downtown than in the study area overall.

Box 4.6a: Mean distance for Burger King Restaurants in Pittsburgh.

Mean distance – population symbol: $\bar{D}$; sample symbol: $\bar{d}$

The physical distance that separates each pair of entities can be treated as a variable much like 'standard' thematic attributes and variables in order to calculate measures of central tendency. The mean, and for that matter the modal and median, distance between phenomena provide measures of how far apart, or close together, things are. Each of these measures potentially suffers from the same problems as their nonspatial counterparts. Insofar as particular entities lying at the extremes of spatial distributions can distort the results, by suggesting a larger distance between entities is the norm rather than the exception.

The overall mean distance is obtained by determining the distance measurements between every pair of entities in the set, adding them up or summing the distances and then dividing by the total number of distance measurements (not the total number of points). The nearest-neighbour mean distance is calculated by summing the distance between each point and its nearest neighbour and dividing by the total number of points. The distance measurements can be calculated using Pythagoras' theorem, although standard statistical and spreadsheet software is not well suited to this task. Spatial analysis software is more amenable to the task.

The upper part of the table below illustrates how to calculate the distance between a selected subset of the pairs of points locating Burger King restaurants in Pittsburgh (see Box 4.3a) and indicates the complexity of the computational problem. In this example, there are 22 points giving rise to 231 pairs ($1 + 2 + 3 + 4 + \ \ldots \ + 19 + 20 + 21$, etc.) and, the calculations are shown for the distance between points 1 and 2, 2 and 3, 3 and 4, to 22 and 21. The calculations between 1 and 3, 1 and 4, 1 and 5, etc., or 2 and 4, and 2 and 5, 2 and 6, etc., and so on are not considered. The calculations shown produce the diagonal in the matrix in Box 4.6c, which includes the 231 distance measurements resulting from pairing each Burger King restaurant with all others in the set. The overall modal, median and mean distance between the Burger King restaurants in Pittsburgh can be determined only when all the distance measurements have been calculated and sorted into ascending (or descending) order (see Box 4.6d).

Box 4.6b: Calculation of distances between selected pairs (1 and 2, 2 and 3, 3 and 4, etc.) of Burger King restaurants in Pittsburgh, USA.

Calculation of point-to-point distances using Pythagoras' theorem

Pt	Pt	Start	End	Start	End			Distance =
m	n	x_m	x_n	y_m	y_n	$(x_m - x_n)^2$	$(y_m - y_n)^2$	$\sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}$
1	2	11	14	27	10	9.0	289.0	17.3
2	3	14	16	10	18	4.0	64.0	8.2
3	4	16	19	18	23	9.0	25.0	5.8
4	5	19	21	23	17	4.0	36.0	6.3
5	6	21	22	17	16	1.0	1.0	1.4
6	7	22	22	16	27	0.0	121.0	11.0
7	8	22	23	27	19	1.0	64.0	8.1
8	9	23	23	19	28	0.0	81.0	9.0
9	10	23	26	28	20	9.0	64.0	8.5
10	11	26	26	20	24	0.0	16.0	4.0
11	12	26	26	24	25	0.0	1.0	1.0
12	13	26	27	25	16	1.0	81.0	9.1
13	14	27	27	16	25	0.0	81.0	9.0
14	15	27	27	25	32	0.0	49.0	7.0
15	16	27	32	32	26	25.0	36.0	7.8
16	17	32	36	26	15	16.0	121.0	11.7
17	18	36	36	15	35	0.0	400.0	20.0
18	19	36	37	35	20	1.0	225.0	15.0
19	20	37	37	20	22	0.0	4.0	2.0
20	21	37	37	22	28	0.0	36.0	6.0
21	22	37	38	28	24	1.0	16.0	4.1
22		38		24				

Box 4.6c: Distance matrix for pairs of Burger King restaurants in Pittsburgh, USA.

Pts		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Coords	27	10	18	23	17	16	27	19	28	20	24	25	16	25	32	26	15	35	20	22	28	24
1	11																						
2	14	17.3																					
3	16	10.3	8.2																				
4	19	8.9	13.9	5.8																			
5	21	14.1	9.9	5.1	6.3																		
6	22	15.6	10.0	6.3	7.6	1.4																	
7	22	11.0	18.8	10.8	5.0	10.0	11.0																
8	23	14.4	12.7	7.1	5.7	2.8	3.2	8.1															
9	23	12.0	20.1	12.2	6.4	11.2	12.0	1.4	9.0														
10	26	16.6	15.6	10.2	7.6	5.8	5.7	8.1	3.2	8.5													
11	26	15.3	18.4	11.7	7.1	8.6	8.9	5.0	5.8	5.0	4.0												
12	26	15.1	19.2	12.2	7.3	9.4	9.8	4.5	6.7	4.2	5.0	1.0											
13	27	19.4	14.3	11.2	10.6	6.1	5.0	12.1	5.0	12.6	4.1	8.1	9.1										
14	27	16.1	19.8	13.0	8.2	10.0	10.3	5.4	7.2	5.0	5.1	1.4	1.0	9.0									
15	27	16.8	25.6	17.8	12.0	16.2	16.8	7.1	13.6	5.7	12.0	8.1	7.1	16.0	7.0								
16	32	21.0	24.1	17.9	13.3	14.2	14.1	10.0	11.4	9.2	8.5	6.3	6.1	11.2	5.1	7.8							
17	36	27.7	22.6	20.2	18.8	15.1	14.0	18.4	13.6	18.4	11.2	13.5	14.1	9.1	13.5	19.2	11.7						
18	36	26.2	33.3	26.2	20.8	23.4	23.6	16.1	20.6	14.8	18.0	14.9	14.1	21.0	13.5	9.5	9.8	20.0					
29	37	26.9	25.1	21.1	18.2	16.3	15.5	16.6	14.0	16.1	11.0	11.7	12.1	10.8	11.2	15.6	7.8	5.1	15.0				
20	37	26.5	25.9	21.4	18.0	16.8	16.2	15.8	14.3	15.2	11.2	11.2	11.4	11.7	10.4	14.1	6.4	7.1	13.0	2.0			
21	37	26.0	29.2	23.3	18.7	19.4	19.2	15.0	16.6	14.0	13.6	11.7	11.4	15.6	10.4	10.8	5.4	13.0	7.1	8.0	6.0		
22	38	70.7	62.1	22.8	19.0	18.4	17.9	16.3	15.8	15.5	12.6	12.0	12.0	13.6	11.0	13.6	6.3	9.2	11.2	4.1	2.2	4.1	

Box 4.6d: Sorting and summation of all distances to determine mean, modal and median distance measures between Burger King restaurants in Pittsburgh, USA.

Mean distance $\bar{d} = \sum \frac{d_B}{n}$

Sorted distances (d)	d	d	d	d	d	d	d	d	d
1.0	5.1	7.1	9.2	11.2	12.6	14.3	16.3	19.4	26.9
1.0	5.1	7.1	9.4	11.2	12.7	14.4	16.6	19.4	27.7
1.4	5.4	7.2	9.5	11.2	13.0	14.8	16.6	19.8	29.2
1.4	5.4	7.3	9.8	11.2	13.0	14.9	16.6	20.0	33.3
1.4	5.7	7.6	9.8	11.2	13.0	15.0	16.8	20.1	62.1
2.0	5.7	7.6	9.9	11.2	13.3	15.0	16.8	20.2	70.7
2.2	5.7	7.8	10.0	11.4	13.5	15.1	16.8	20.6	
2.8	5.8	7.8	10.0	11.4	13.5	15.1	17.3	20.8	
3.2	5.8	8.0	10.0	11.4	13.5	15.2	17.8	21.0	
3.2	5.8	8.1	10.0	11.7	13.6	15.3	17.9	21.0	
4.0	6.0	8.1	10.2	11.7	13.6	15.5	17.9	21.1	
4.1	6.1	8.1	10.3	11.7	13.6	15.5	18.0	21.4	
4.1	6.1	8.1	10.3	11.7	13.6	15.6	18.0	22.6	
4.1	6.3	8.2	10.4	11.7	13.6	15.6	18.2	22.8	
4.2	6.3	8.2	10.4	**12.0**	13.9	15.6	18.4	23.3	
4.5	6.3	8.5	10.6	12.0	14.0	15.6	18.4	23.4	
5.0	6.3	8.5	10.8	12.0	14.0	15.8	18.4	23.6	
5.0	6.4	8.6	10.8	12.0	14.0	15.8	18.4	24.1	
5.0	6.4	8.9	10.8	12.0	14.1	16.0	18.7	25.1	
5.0	6.7	8.9	11.0	12.0	14.1	16.1	18.8	25.6	
5.0	7.0	9.0	11.0	12.1	14.1	16.1	18.8	25.9	
5.0	7.1	9.0	11.0	12.1	14.1	16.1	19.0	26.0	
5.0	7.1	9.1	11.0	12.2	14.1	16.2	19.2	26.2	
5.1	7.1	9.1	11.2	12.2	14.2	16.2	19.2	26.2	
5.1	7.1	9.2	11.2	12.6	14.3	16.3	19.2	26.5	

$\sum d = 2993.8$

$\frac{2993.8}{231}$

Mode = 11.2 Median = 12.0

$\bar{d} = 12.96$

What can we conclude about the distribution of Burger King restaurants in Pittsburgh from this information?

4.4 Dispersion

4.4.1 Measures for nonspatial data

We have seen that there are various ways of quantifying the central tendency of values in a set of data. It is also useful to know the spread of data values. This information can be obtained by calculating a quantity representing the dispersion of the numerical distribution. Imagine sitting in a stadium and looking down at a football match, the 22 players (and the referee) are all located somewhere within the 90 m by 45 m rectangle that defines the pitch (minimum dimensions). At kick off, when the match starts, the players will be fairly evenly distributed around the pitch in their allotted positions. However, if one side is awarded a free kick, the large majority of the players will rush to cluster together in the goal area with those on one team attempting to score and the others to defend. In other words, sometimes the players form a compact distribution and other times they will be more dispersed.

There are three main measures of dispersion known as the range, variance and standard deviation, all of which are numerical quantities indicating the spread of values of a nonspatial attribute or variable. These measures are interpreted in a relative fashion, in the sense that a small value indicates less dispersion than a large one, and vice versa. The range is simply calculated as the difference between the smallest and largest values. Alternatively, various percentile ranges may be determined that omit an equal percentage of values at the upper and lower ends of the distribution. For example, excluding both the largest and smallest 25% of data points and calculating the difference between the values at either end of the remainder produces the interquartile range. The procedures for obtaining the variance and standard deviation measures involve rather more complicated calculations than simply sorting the data values into ascending order and are presented in Box 4.7 with a variable measured on the ratio/interval scale. Nevertheless, the basic purpose in each case is the same, namely to determine whether the data values are close together or spread out.

The reason for there being several measures of dispersion is essentially the same as why there are different measures of central tendency. Each has advantages and disadvantages, and should be used for attributes and variables recorded according to the different scales of measurement. Intuitively, the range seems the most sensible measure, since it reports the size of the numerical gap between the smallest and largest value. Unfortunately, this may be misleading, since extreme outlier values may exaggerate the situation. Discarding the top and bottom 10% or 25% helps to overcome this problem, but the choice of exclusion percentage is entirely arbitrary. A further problem with using the absolute or percentile range is that it fails to take into account

the spread in the values between the upper and lower values. We are none the wiser about whether the data values are clustered, regularly or randomly dispersed along the range.

The variance and the standard deviation help to overcome this problem by focusing on the differences between the individual data values and their mean. However, this implies that these dispersion measures are not suitable for use with attributes recorded on the nominal or ordinal scales. If most of the data values are tightly clustered around the mean, then the variance and standard deviation will be relatively

Box 4.7a: Variance and standard deviation of water temperature measurements in a fluvioglacial stream from Les Bossons Glacier, France.

Variance – population symbol: σ^2; sample symbol: s^2

Standard deviation – population symbol: σ; sample symbol: s

The variance and its close relation the standard deviation are two of the most useful measures of dispersion when your data are recorded on either the interval or ratio scales of measurement. They measure the spread of a set of numbers around their mean and so allow you to quantify whether two or more sets of numbers are relatively compact or spread out, irrespective of whether they possess similar or contrasting means. The concept of difference (deviation) is important when trying to understand what the variance and standard deviation convey. A small variance or standard deviation denotes that the numbers are tightly packed around their mean, whereas a large value indicates that they are spread out. So, if you have data for two or more samples of the same category of observations, for instance samples of soil from the A, B and C horizons, and then you can see whether or not the spreads (dispersions) of particle sizes are similar.

The standard deviation is closely related to the variance, but one of its main advantages is that its value is measured in the same units as the original set of measurements. So, for example, if these measurements are of water temperature in degrees Celsius or of journey time in minutes, their standard deviation is expressed in these units. The standard deviation is obtained by taking the square root of the variance. There are two main ways of calculating the variance and the standard deviation. The method on the left in the calculations table below involves simpler computation, since it is based on the sum of the X values and the sum of the squared X values. The second method uses the sum of the squared differences (deviations) of the X values about their mean and, although this entails slightly more complicated calculations, it provides a clearer illustration of what the standard deviation measures. These methods are illustrated using sample data that represent measurements of water temperature (X) in the fluvioglacial stream flowing from Les Bossons Glacier, France where the stream leaves the sandur plain taken at half-hourly intervals between 09.00–17.30 hrs (see Box 4.2).

The calculations show that the sample with a mean of 9.08 °C has a variance of 7.78 and a standard deviation of 2.79 °C – thankfully the two methods for deriving the standard deviation produce the same result.

Box 4.7b: Calculation of the variance and standard deviation.

	$s=\sqrt{\frac{n(\sum x^2)-(\sum x)^2}{n(n-1)}}$		$s=\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$		
	x	x^2	x	$(x-\bar{x})$	$(x-\bar{x})^2$
09.00 hrs	4.30	18.49	4.30	–4.78	22.88
09.30 hrs	4.70	22.09	4.70	–4.38	19.21
10.00 hrs	4.80	23.04	4.80	–4.28	18.35
10.30 hrs	5.20	27.04	5.20	–3.88	15.08
11.00 hrs	6.70	44.89	6.70	–2.38	5.68
11.30 hrs	10.10	102.01	10.10	1.02	1.03
12.00 hrs	10.50	110.25	10.50	1.42	2.01
12.30 hrs	11.20	125.44	11.20	2.12	4.48
13.00 hrs	11.40	129.96	11.40	2.32	5.37
13.30 hrs	11.80	139.24	11.80	2.72	7.38
14.00 hrs	12.30	151.29	12.30	3.22	10.35
14.30 hrs	11.90	141.61	11.90	2.82	7.93
15.00 hrs	11.40	129.96	11.40	2.32	5.37
15.30 hrs	11.10	123.21	11.10	2.02	4.07
16.00 hrs	10.20	104.04	10.20	1.12	1.25
16.30 hrs	9.30	86.49	9.30	0.22	0.05
17.00 hrs	8.80	77.44	8.80	–0.28	0.08
17.30 hrs	7.80	60.84	7.80	–1.28	1.65
n = 18	$\sum x=163.50$	$\sum x^2=1617.33$	$\sum x=163.50$	$\sum(x-\bar{x})^2=132.21$	

Variance

$$s^2=\frac{(18(1617.33)-(163.50)^2}{18(18-1)} \qquad s^2=\frac{132.21}{18-1}$$

$$s^2=7.78 \qquad s^2=7.78$$

Standard deviation

$$\sqrt{\frac{18(1617.33)-(163.50)^2}{18(18-1)}}$$

$$s=\sqrt{\frac{18(1617.33)-(163.50)^2}{18(18-1)}} \qquad s=\sqrt{\frac{132.21}{18-1}}$$

$$s=2.79 \qquad s=2.79$$

Sample Means

$$\frac{\sum x}{n} \quad \frac{163.50}{18}=9.08 \qquad \frac{\sum x}{n} \quad \frac{163.50}{18}=9.08$$

small even if there are a small number of extremely low and high values as well. A similar result will occur if the values are bimodally distributed with most lying at **each** end of the range and comparatively few in the middle around the mean. Conversely, the variance and standard deviation will be moderately large, if the data values are regularly spaced across the range. In the highly unlikely event that all data values are identical, then the range, variance and standard deviation measures will all equal zero and hence denote an absence of variation of the variable. In practical terms the main difference between the variance and the standard deviation is that the latter is measured in the same units as the original data values. Thus, Box 4.7 the standard deviation of 2.79 for melt water stream temperature is in °C, whereas the variance of 7.78 is in °C squared, which does make much sense.

4.4.2 Measures for spatial data

When analysing the patterns of spatial phenomena, it is often useful to indicate their dispersion, which may be achieved by calculating the **standard distance**. This is essentially the spatial equivalent of the standard deviation. It provides a measure of the extent to which the individual data points are clustered around the mean centre in two-dimensional space. Calculation of standard distance focuses on the X and Y coordinates of the set of data points and is a measure of how spread out or dispersed they are. Because a spatial distribution of points is located in two dimensions, X and Y, it is feasible for one set of coordinates to be clustered tightly around their mean and for the other to be more spread out. Figure 4.5a illustrates how a distribution of points depicting the positions of former coal pits within a relatively confined approximately north-south oriented valley in South Wales can lead to a narrower spread of X coordinates in comparison with Y coordinates. Figure 4.5b shows how points representing the location of volcanoes associated with the predominantly west–east aligned tectonic plates in the eastern Mediterranean region results in a wide spread of X coordinate values and the Y coordinates in a narrower range. These patterns contrast within a distribution in which the features are more random or tightly clustered.

Box 4.8 shows the calculation of the standard distance for the points representing Burger King Restaurants in Pittsburgh and produces the value 9.73 km. It is difficult to tell just by considering this value whether these restaurants are dispersed or compact, you would want to compare this result with another fast-food chain in Pittsburgh, with Burger King's outlets in other cities, or possibly with some theoretical figure perhaps denoting a perfectly regular distribution. Unfortunately, if comparing the standard distance values for study areas that are substantially different in size, then an adjustment should be introduced to compensate. Neft (1966), for example used the radius of his study area countries to produce a relative measure of population dispersion. Standard distance can be used as the radius of a circle drawn around the mean centre of a distribution of points and thus provide a graphic representation of their dispersion.

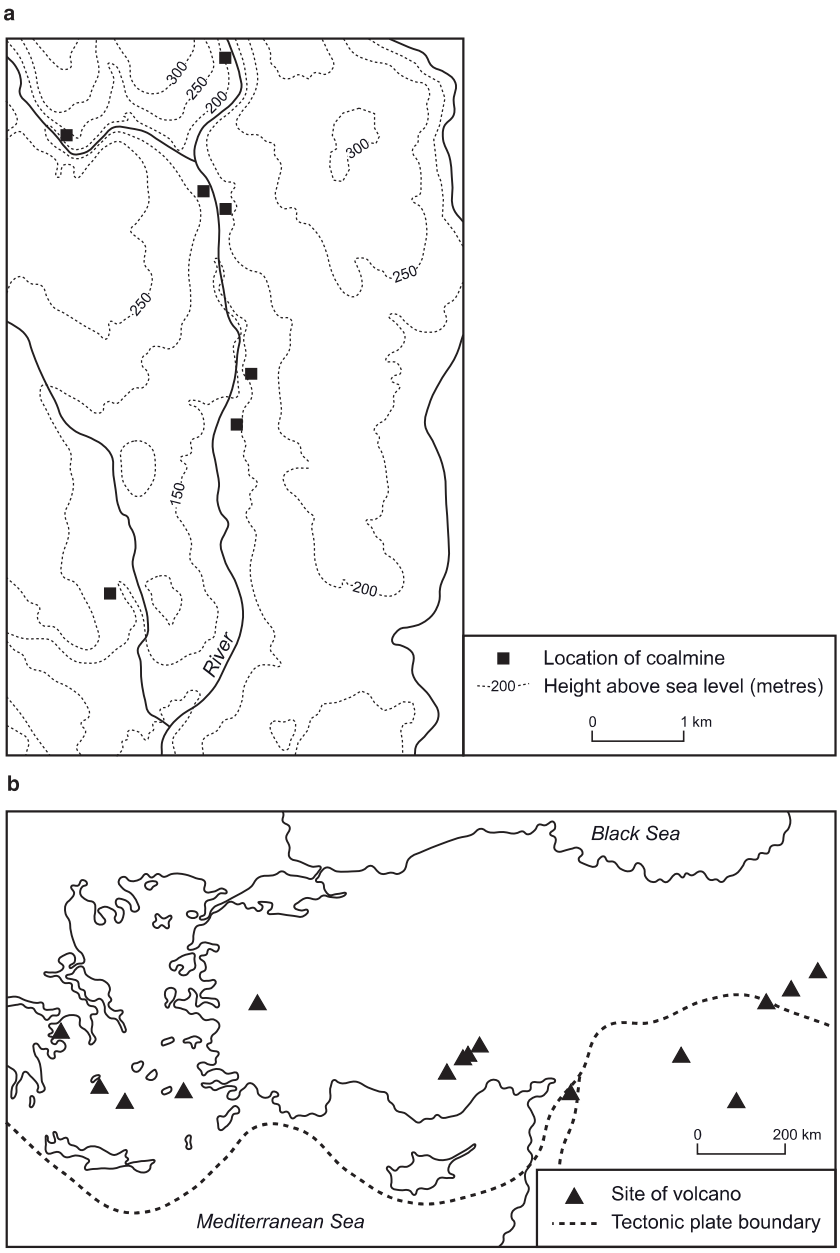
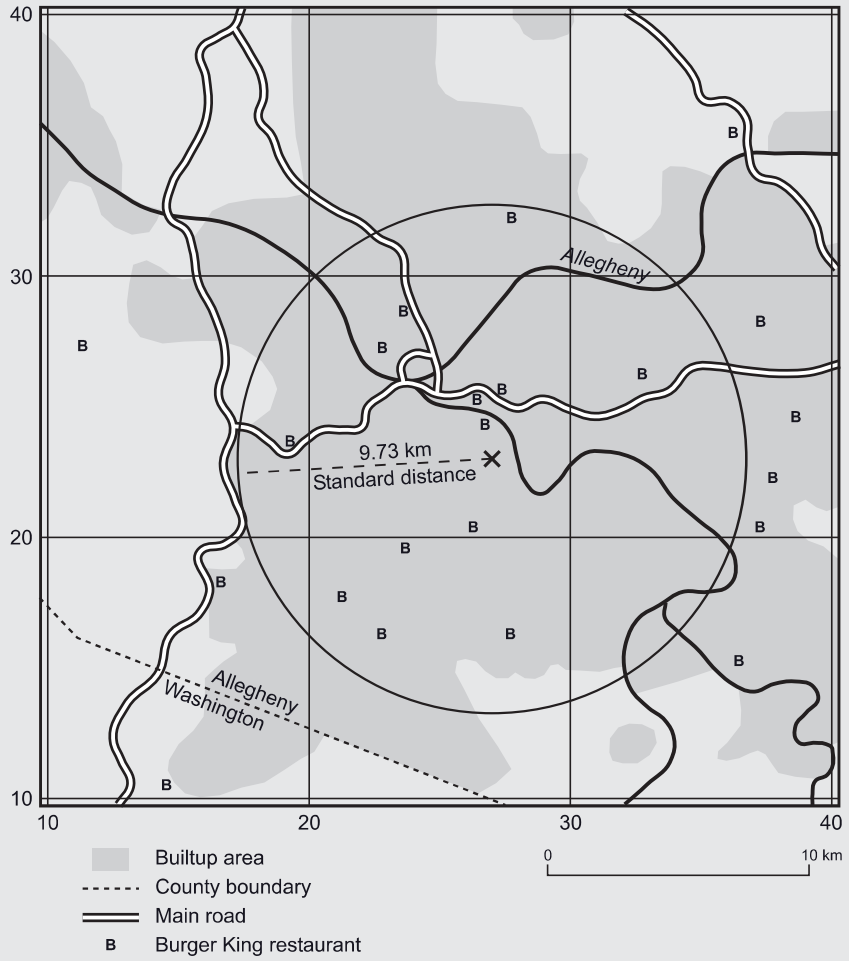


Figure 4.5 *X* or *Y* coordinates constrained within relatively narrow ranges.

Box 4.8a: Standard distance for Burger King restaurants in Pittsburgh.

Standard distance – population symbol: S_d; sample symbol: s_d



The notion that all points representing the location of a set of geographical phenomena will occur exactly on top of each other is just as unrealistic as to imagine that the values of any nonspatial variable for a given collection observations will be the same. The standard distance measure enables us to quantify how spread out the points are, just as the standard deviation measures the dispersion of a set of numerical values. The larger the standard distance then the greater the dispersion amongst the points.

Standard distance is calculated from the X and Y coordinates of the points and does not require you to calculate the distances between all the pairs of points as described in Box 4.7. The procedure given below shows calculation of the variances for the X and Y coordinates, these are summed and the square root of the result produces the standard distance. Examination of the X and Y variances indicates that the X coordinates are more dispersed that the Y coordinates.

Box 4.8b: Calculations for standard distance.

$$s_d = \sqrt{\frac{\sum (x-\bar{x})^2}{(n-1)} + \frac{\sum (y-\bar{y})^2}{(n-1)}}$$

Point	X	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
1	11	−15.5	240.25	27	4.4	19.36
2	14	−12.5	156.25	10	−12.6	158.76
3	16	−10.5	110.25	18	−4.6	21.16
4	19	−7.5	56.25	23	0.4	0.16
5	21	−5.5	30.25	17	−5.6	31.36
6	22	−4.5	20.25	16	−6.6	43.56
7	22	−4.5	20.25	27	4.4	19.36
8	23	−3.5	12.25	19	−3.6	12.96
9	23	−3.5	12.25	28	5.4	29.16
10	26	−0.5	0.25	20	−2.6	6.76
11	26	−0.5	0.25	24	1.4	1.96
12	26	−0.5	0.25	25	2.4	5.76
13	27	0.5	0.25	16	−6.6	43.56
14	27	0.5	0.25	25	2.4	5.76
15	27	0.5	0.25	32	9.4	88.36
16	32	5.5	30.25	26	3.4	11.56
17	36	9.5	90.25	15	−7.6	57.76
18	36	9.5	90.25	35	12.4	153.76
	$\sum x = 583$		$\sum (x-\bar{x})^2 = 1335.50$	$\sum y = 497$		$\sum (y-\bar{y})^2 = 749.32$
			$\frac{1335.50}{22}$			$\frac{749.32}{22}$
			60.61			34.06

$$s_d = \sqrt{60.61 + 34.06}$$

Standard distance $s_d = 9.73$

4.5 Measures of skewness and kurtosis for nonspatial data

Skewness and kurtosis are concerned with the shape of the distribution of data values and so in some respects belong with discussion of frequency distributions in Chapter 5. However, since each is a quantity capable of being calculated for data not expressed in the form of a frequency distribution and they are known, respectively, as the third and fourth moments of descriptive statistics, following on from central tendency, the first, and dispersion, the second, they are discussed here. The principle underlying skewness in respect of statistics is closely allied to the everyday word skewed, which

indicates that something is slanted in one direction rather than being balanced. For example, a political party seeking re-election may be selective in the information it chooses relating to its current period in office in order to put a 'spin' or more favourable interpretation on its record. If information or statistics are skewed, this implies that, for whatever reason, there is a lack of symmetry or balance in the set of data values.

Kurtosis may most simply be translated into everyday language as 'peakedness', and therefore provides a method of quantifying whether one set of numbers is more or less peaked than another. Just as a mountainous landscape may be described as undulating if the height of the mountains is low relative to their surface area, in contrast with a deeply dissected upland region with steeper slopes. From the numerical point of view it is a question of there being a difference in the ratio of the maximum height to radius of the mountain, assuming each peak to be perfectly circular when viewed from above (plan view).

Following on from this analogy, Box 4.9a shows two typical shapes for volcanic cones that are approximately circular in plan view, but have rather different cross sections. In one case, Mauna Loa on Hawaii, the cone has been formed from relatively free-flowing lava that was able to spread extensively over the existing surface during successive eruptions and build up what is sometimes called a 'shield cone'. In contrast, Mount Teide on Tenerife in the Canary Islands was created during a more violent series of eruptions of more viscous lava that constructed a conical volcano. Lines have been superimposed along the profiles of the volcanoes to illustrate that one is relatively flat and the other more peaked (see Box 4.9a). The calculations involved with obtaining the skewness and kurtosis measures are illustrated with elevation

Box 4.9a: Calculation of skewness and kurtosis using examples of contrasting volcanic cones in Hawaii and Tenerife.

Skewness – population symbol: β_1; sample symbol: b_1.

Kurtosis – population symbol: γ_2; sample symbol: g_2.

Mauna Loa (shield volcano)

Teide (composite cone volcano)

Skewness is most useful when you are dealing with data recorded on the interval or ratio scales of measurement. It can be viewed as measuring the degree of symmetry and the magnitude of the differences of the values of a variable around its mean. A skewness statistic of zero indicates perfect symmetry and differences of equal magnitude either side of the mean, whereas a negative value denotes that more numbers are greater than the mean (negative skewness) and a positive quantity the reverse. A large skewness statistic signifies that the values at one extreme exert a disproportionate influence. Both symmetrical and severely skewed distributions can display varying degrees of kurtosis depending on whether the values are tightly packed together or are more spread out. A kurtosis statistic of 3 indicates that the distribution is moderately peaked, whereas a set of data values possessing a relatively flat shape will produce a statistic less than 3 and those with a pronounced peak will be greater than 3. As with the other descriptive quantities skewness and kurtosis are commonly used to help with comparing different sets of numbers to see if they are characteristically similar or different. Calculating the skewness and kurtosis statistics for two or more samples of data enables you to see whether or not the distributions of their values have a similar shape.

The sum of the differences between the individual values and their mean always equals zero and the square of the differences is used in calculating the variance (see Box 4.7), so the cube of the deviations provides the basis of the skewness measure. These cubed differences are summed and then divided by the cubed variance of the values multiplied by the number of data points. Following a similar line of argument, calculation of kurtosis involves raising the differences to the fourth power and then dividing the sum of these by the variance raised to the power of four multiplied by the number of observations. These calculations are shown below in Boxes 4.8b and 4.8c with respect to samples of elevation measurements for two contrasting volcanic cones, Mauna Loa in Hawaii and Mount Teide on Tenerife. The classical broad, reasonably symmetrical profile of the Mauna Loa shield volcano results in a skewness and kurtosis statistics, respectively, of −0.578 and 2.009. The profile of Mount Teide is more conical (peaked in statistical terms) and slightly less symmetrical than Mauna Loa, which produces skewness and kurtosis values of −0.029 and 1.673.

Box 4.9b: Calculation of skewness for samples of spot heights on Mauna Loa, Hawaii (x_L) and Mount Teide, Tenerife (x_T).

(Note: standard deviation for Mauna Loa elevations (s) = 73.984; and standard deviation for Mount Teide elevations (s) = 363.354).

	Mauna Loa				Mount Teide		
	$b_1 = \frac{\sum(x-\bar{x})^3}{(n-1)s^3}$				$b_1 = \frac{\sum(x-\bar{x})^3}{(n-1)s^3}$		
N	x_L	$(x_L - \bar{x}_L)$	$(x_L - \bar{x}_L)^3$	N	x_T	$(x_T - \bar{x}_T)$	$(x_T - \bar{x}_T)^3$
1	3840	−146.44	−3 140 358.0	1	2600	−572.7	−187 817 497.0
2	3852	−134.44	−2 429 883.8	2	2700	−472.7	−105 609 182.0
3	3864	−122.44	−1 835 565.8	3	2800	−372.7	−51 761 668.0
4	3889	−97.44	−925 149.3	4	2900	−272.7	−20 274 953.0
5	3901	−85.44	−623 711.5	5	3000	−172.7	−5 149 038.3
6	3913	−73.44	−396 093.8	6	3100	−72.7	−383 923.6
7	3938	−48.44	−113 661.2	7	3200	27.3	20 391.2
8	3962	−24.44	−14 598.3	8	3300	127.3	2 063 905.9
9	3974	−12.44	−1925.1	9	3400	227.3	11 746 621.0
10	3986	−0.44	−0.1	10	3500	327.3	35 068 535.0
11	4011	24.56	14 814.4	11	3600	427.3	78 029 650.0
12	4023	36.56	48 867.3	12	3700	527.3	146 629 965.0
13	4035	48.56	114 508.1	13	3717	544.3	161 273 450.0
14	4047	60.56	222 104.6	14	3700	527.3	146 629 965.0
15	4059	72.56	382 025.0	15	3600	427.3	78 029 650.0
16	4072	85.56	626 343.1	16	3500	327.3	35 068 535.0
17	4088	101.56	1 047 533.9	17	3400	227.3	11 746 621.0
18	4072	85.56	626 343.1	18	3300	127.3	2 063 905.9
19	4059	72.56	382 025.0	19	3200	27.3	20 391.2
20	4047	60.56	222 104.6	20	3100	−72.7	−383 923.6
21	4035	48.56	114 508.1	21	3000	−172.7	−5 149 038.3
22	4023	36.56	48 867.3	22	2900	−272.7	−20 274 953.0
23	4011	24.56	14 814.4	23	2800	−372.7	−51 761 668.0
24	3986	−0.44	−0.1	24	2700	−472.7	−105 609 182.0
25	3974	−12.44	−1925.1	25	2600	−572.7	−187 817 497.0
$\sum x_L = 113356$			$\sum(x_L - \bar{x}_L)^3 = -5618013.1$	$\sum x_T = 79317$			$\sum(x_T - \bar{x}_T)^3 = -33600939.0$

$$b_1 = \frac{5\,618\,013.1}{24(404\,965.8)} \qquad b_1 = \frac{33\,600\,939.0}{24(47\,972\,275)}$$

$$b_1 = -0.578 \qquad b_1 = -0.029$$

Box 4.9c: Calculation of kurtosis for samples of spot heights on Mauna Loa, Hawaii (x_L) and Mount Teide, Tenerife (x_T).

(Note: standard deviation for Mauna Loa elevations (s) = 73.984; and standard deviation for Mount Teide elevations (s) = 363.354).

	Mauna Loa				Mount Teide		
	$g_2 = \frac{\sum(x-\bar{x})^4}{ns^4}$				$g_2 = \frac{\sum(x-\bar{x})^4}{ns^4}$		
N	x_L	$(x_L-\bar{x}_L)$	$(x_L-\bar{x}_L)^4$		x_T	$(x_T-\bar{x}_T)$	$(x_T-\bar{x}_T)^4$
1	3840	−146.44	459 874 025.8	1	2600	−572.7	107 559 324 269.5
2	3852	−134.44	326 673 582.4	2	2700	−472.7	49 919 348 352.4
3	3864	−122.44	224 746 679.3	3	2800	−372.7	19 290 538 323.2
4	3889	−97.44	90 146 548.1	4	2900	−272.7	5 528 574 182.1
5	3901	−85.44	53 289 906.6	5	3000	−172.7	889 135 929.0
6	3913	−73.44	29 089 126.0	6	3100	−72.7	27 903 563.8
7	3938	−48.44	5 505 750.6	7	3200	27.3	557 086.7
8	3962	−24.44	356 783.5	8	3300	127.3	262 776 497.6
9	3974	−12.44	23 948.7	9	3400	227.3	2 670 241 796.4
10	3986	−0.44	0.0	10	3500	327.3	11 478 632 983.3
11	4011	24.56	363 842.5	11	3600	427.3	33 343 630 058.2
12	4023	36.56	1 786 589.4	12	3700	527.3	77 320 913 021.0
13	4035	48.56	5 560 511.1	13	3717	544.3	87 784 364 145.9
14	4047	60.56	13 450 656.0	14	3700	527.3	77 320 913 021.0
15	4059	72.56	27 719 736.4	15	3600	427.3	33 343 630 058.2
16	4072	85.56	53 589 919.4	16	3500	327.3	11 478 632 983.3
17	4088	101.56	106 387 540.5	17	3400	227.3	2 670 241 796.4
18	4072	85.56	53 589 919.4	18	3300	127.3	262 776 497.6
19	4059	72.56	27 719 736.4	19	3200	27.3	557 086.7
20	4047	60.56	13 450 656.0	20	3100	−72.7	27 903 563.8
21	4035	48.56	5 560 511.1	21	3000	−172.7	889 135 929.0
22	4023	36.56	1 786 589.4	22	2900	−272.7	5 528 574 182.1
23	4011	24.56	363 842.5	23	2800	−372.7	19 290 538 323.2
24	3986	−0.44	0.0	24	2700	−472.7	49 919 348 352.4
25	3974	−12.44	23 948.7	25	2600	−572.7	107 559 324 269.5
$\sum x_L = 99661$			$\sum(x_L-\bar{x}_L)^4 = 1501060350.0$	$\sum x_T = 79317$			$\sum(x_T-\bar{x}_T)^4 = 704367516272.0$

$$g_2 = \frac{1\,501\,060\,350.1}{24(29\,961\,099.8)} \qquad g_2 = \frac{704\,367\,516\,272.0}{24(17\,430\,924\,527.8)}$$

$$g_2 = 2.009 \qquad g_2 = 1.683$$

measurements for these differing shapes in respect of 5 km long cross sections whose centres correspond with the peaks of the cones. Both cones are reasonably symmetrical across this distance, although the measures of kurtosis reflects the difference in profile.

From a computational perspective, there is no reason why measures of skewness and kurtosis should not be calculated for spatial as much for nonspatial data. For example, the skewness statistic could be calculated for all the pairs of distance measurements between Burger King restaurants in Pittsburgh and some commentary supplied about whether there are more or less smaller than the mean distance, and whether the magnitude of the differences are larger in the case of the latter or former. However, it is less common for analysis to consider the skewness or kurtosis of spatial distributions.

4.6 Closing comments

In some respects this chapter has been all about 'playing with numbers', seeing how we can describe the essential characteristics of a set of numbers. The four main elements of this statistical description are central tendency, dispersion, skewness and kurtosis. Most of these elements can be quantified in several different ways, for example by taking the most typical or frequently occurring (mode), the middle (median) or mean (average) value in the case of central tendency. In the case of measuring dispersion, the different types of range (absolute and percentile) provide an intuitive impression of how spread out the values of a variable are, but in practice the variance and standard deviation are more useful quantities. In most cases there are equivalent spatial statistics to these more 'standard' versions. Although this chapter has tended to focus on spatial phenomena as points, similar measures exist for lines and areas and in some instances linear and areal feature may be located spatially as a point by means of a pair of X and Y coordinates. Such points may relate to the midpoint of a line or the spatial or weighted centroids of an area (polygon). In these cases then the measures outlined in this chapter can reasonably be used to describe, for example the mean centre of a collection of areas.

Of course when carrying out quantitative data analysis the intention is not just to observe how different statistical quantities vary between one set of numbers and another. There needs to be a transition from regarding the numerical values as not just a set of digits but as having importance as far as the variable or attribute in question is concerned. So, for example, when looking at the values for the temperature of Les Bossons Glacier melt water stream, the purpose is not simply to regard them as 'any old set of numbers' that happen to have a mean of 9.08 and a standard deviation of 2.79, there are many other sets of numbers that can produce these results. The important point is that the numbers are measurements or values of a variable relating to a particular geographical and temporal context. In these circumstances the statistical measures are not simply the outcome of performing certain mathematical com-

putations, but are tools enabling us to say something potentially of interest about the diurnal pattern of temperature change associated with the melting of ice from this specific glacier. When planning research investigations and specifying the attributes variables that will be relevant, it is as well to think ahead to the types of analysis that will be carried out, and without knowing what the values will be in advance, thinking about how to interpret the means, variances and standard distances that will emerge from the data.