

SPECIAL SKILLS SERIES – SAMPLING AND STATISTICAL METHODS: MEASURES OF CENTRAL TENDENCY AND VARIATION

Introduction

The aim of this Geofile is to explain the nature and use of a range of statistical tests within the context of the geographical investigation process.

Background

In the past, many geographical theories and models have relied upon qualitative (non-numerical) data. With greater amounts of quantitative data being generated and available than ever before, especially in the fields of remote sensing and through the Internet, there is an increasing need to be able to summarise effectively and extrapolate information accurately and quickly. The final stage in the analysis is to attempt to prove or disprove hypotheses; many of the methods which are employed to do this are statistical.

Numerous statistical tests are available to the geographer, varying in both the information they provide about the data set and complexity of use. It is therefore essential to understand how, individually, their value can be maximised and in addition to this, the inherent limitations when using each of them.

The type of statistical test conducted during an investigation depends primarily upon the hypothesis being tested and more specifically upon whether you are attempting to summarise single data sets or establish whether there is a relationship between sets of data.

It would be both time-consuming and bad practice to employ statistical tests when it is unnecessary to do so and hence we must fully comprehend the applications of statistical tests before being able to select and effectively use them as part of Geography.

Measures of central tendency

The group of statistical tests known as measures of central tendency share a common theme: they seek to provide a single numerical value which is representative of the entire data set.

This is not as simple as first suggested, as the term central point can be interpreted in many different ways.

The central points of data sets

There are three different statistical measures of the central point of a set of data:

1. the mean
2. the mode
3. the median.

The mean

The arithmetic mean or average (\bar{x}) is perhaps the best known and commonly utilised measure of central tendency. It has many applications stretching across a range of subject areas and is calculated by summing the data within a given set together and dividing by the total number of pieces of data utilising the following equation.

Equation box 1

$$(\bar{x}) = \frac{\sum x}{n}$$

$\sum x$ = the sum of all data values
 n = the number of data values in the set
 (\bar{x}) = the mean value

Example

The quantitative data in Figure 1 is a set of midday temperature values recorded in the centre of Cheltenham throughout the month of June. Calculate the mean value.

$$\begin{aligned} \sum x &= 613 \\ n &= 30 \end{aligned}$$

Figure 1

Midday temperatures recorded in Cheltenham over a monthly period in °C			
14	19	22	23
17	22	21	22
18	21	24	22
18	22	23	21
15	24	26	20
20	21	23	20
20	21	16	
21	20	17	

Therefore the mean (\bar{x}) is equal to 20.43°C.

The mean is a relatively straightforward statistic to generate but does have several limitations which must be considered when using it to summarise a data set:

- It is heavily influenced by any extreme/outlying points within the data set, and when calculated incorporating these points the mean value could be misleading with reference to the rest of the data set.
- It gives no information as to how the data within the set is spread around this middle point, hence two data sets with similar mean values may represent widely differing distributions of data.

Means of two geographical data sets can be compared to show whether or not they are actually different, for example size of river sediment collected in two parts of the channel.

Mode

The second measure of central tendency to be considered is the mode. The mode of a set of data, referred to as the modal value, is the most frequently occurring value within the data set.

Example

Using the data within Figure 1 there are six individual days when the temperature was 21°C, which is more than any other temperature within this month. We therefore say

that the modal temperature was 21°C.

Points to consider in relation to the mode

- The mode of any given data set is not always a single value. For instance if a data set has two values that occur the same number of times then it would be termed bimodal.
- If the data are recorded within numerically defined categories the modal category may be obvious but the exact value of the mode becomes more complicated to calculate.
- The mode has little value in relation to the more complicated statistical tests discussed later.

Median

The median or mid-point value (m) is the numerical value falling within the data set at which half of data are above it and half are below it. It is again relatively simple to calculate the median:

1. First put the data into arithmetic order (either ascending or descending) based on the data values themselves or ranks previously assigned to the data.
2. Count the number of items of data (termed n).

Finally put the numerical value of (n) into the equation in Equation Box 2 to find out the location of the median within the data set.

Equation box 2: The equation for finding the median of a data set

$$m = \frac{n + 1}{2}$$

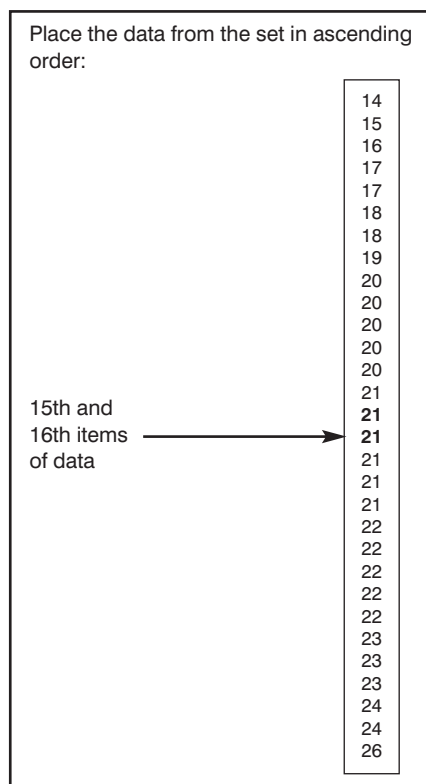
m = the median of the set of data
 n = number of pieces of data in set

Key points

If the number of pieces of data in the set (n) is **odd** then the median value will appear as an integer (whole number) and it is simply a case of finding the piece of data this refers to by counting through the data set from either end.

If the number of pieces of data is **even** then the median value will lie between two pieces of data within

Figure 2: Worked example of finding the median using the data from Figure 1



the set. To establish the median number based upon this simply add the data values on either side of the median point together and average them. The averaged value will be the median of the data set.

We must now consider whether there is an odd or even number of items of data in this set.

Calculate the median value utilising the equation in Box 2:

- Add 1 to the total number of pieces of data in the set ($30 + 1$)
- Divide the resultant number (31) by 2 . $31/2 = 15.5$
- This indicates where the median value lies (the 15.5th value). Since we only have a 15th and 16th value we must average these in order to find the median value. So $(21 + 21)/2 = 21$. The median value and temperature for June is 21°C.

Points to consider when using the median value

- It gives no information as to how the data within the set is spread around its median value, hence two data sets with similar medians may have a widely differing distribution of data.
- With large numbers of

observations within a data set it can be a tedious statistic to calculate especially when the data is being manipulated by hand as opposed to by machine where programs such as Microsoft Excel will be able to calculate this with a simple click of a button.

As explained previously there are inherent limitations related to the use of the mean, median and mode as measures of central tendency not least the fact that they are single values being used to describe what can be large sets of observations. The mean or average has been suggested as the most statistically significant of these statistics however it also has inherent flaws, as discussed above.

Measures of deviation, dispersion and variability

The terms **deviation**, **dispersion** and **variability** used in this context all refer to analysing a set of data in terms of its spread around the mean or median value.

The simplest measure of variation is termed the **range**, and this is the most obvious way to describe the scattering of the data within a set. The range defines the region within which all the data values lie. In order to calculate the range the lowest value from the data set is subtracted from the highest to provide a single numerical value, again describing the data set. There are a number of points to be considered when using the range to describe a data set:

- It is only calculated using two pieces of data from the entire data set.
- It gives no indication of the spread of data in the remaining data set within the two extremes used in the calculation of the range.
- Whenever an outlier/anomalous result is present and represents the highest or lowest value the range statistic will utilise this figure and as a result a misleading impression of the true limits/spread of data set will be given.

It would be easy to simply dismiss the range as a statistical measure of dispersion, having highlighted its

flaws, but there are ways of improving modifying the concept of the range to provide more statistically valuable information relating to the data set. The first of these is the **interquartile range**.

The interquartile range

The interquartile range is a statistical value that describes where the middle 50% of the data within any given set lies. It takes into account the median value but in addition to this gives an indication of how the data within the set are spread out around it. It uses a calculation similar to that used in finding the range. Using this technique the possibility of outliers having a significant impact upon the median value or range of the data set is reduced.

The interquartile range is found as followed:

1. First find the median value of the data set as explained utilising the equation in Box 2. This represents the point half-way or 50% through the data set.
2. Next find the 25% and 75% quartiles; these are points which represent the outer limits of the middle half of the data set. To calculate these count the number of individual pieces of data on either side of the median. Take this value to be "n" and then calculate the median of each half of the data set using the same process as described in Box 2. The value in the upper half of the data set is described as the upper quartile and in the lower set as the lower quartile.
3. The numerical difference between the upper and lower quartiles is referred to as the **interquartile range** and has one key benefit range as calculated earlier in that by considering this value the possibility of outliers giving a misleading impression of the spread of the data (as occurs with the range) is minimised.

Points to consider in relation to the interquartile range

- It can be a laborious process to calculate the location of the quartiles, especially when there is a large number of data within the set.
- The interquartile statistic, in a similar way to the range, does

Figure 3: A box and whisker plot using the data in Figure 1 highlighting the median and interquartile ranges

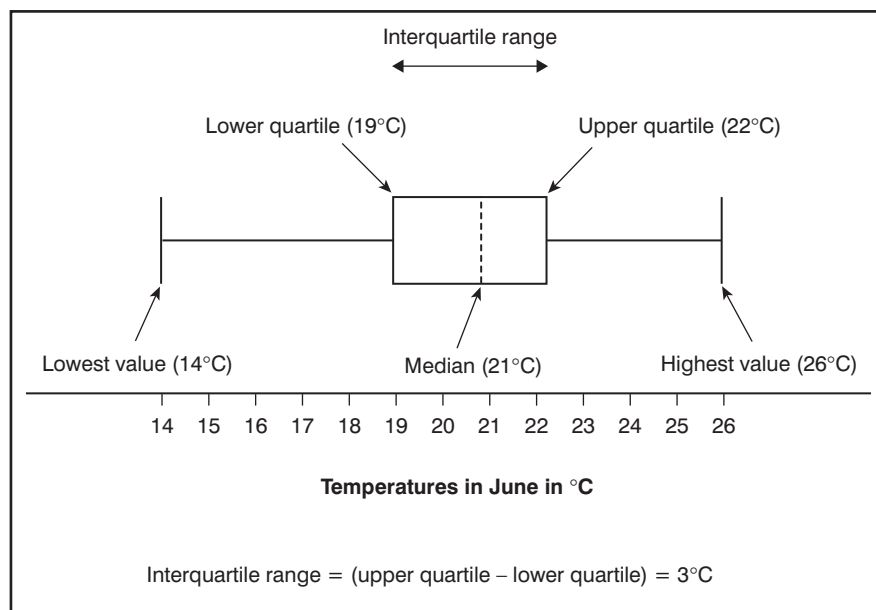
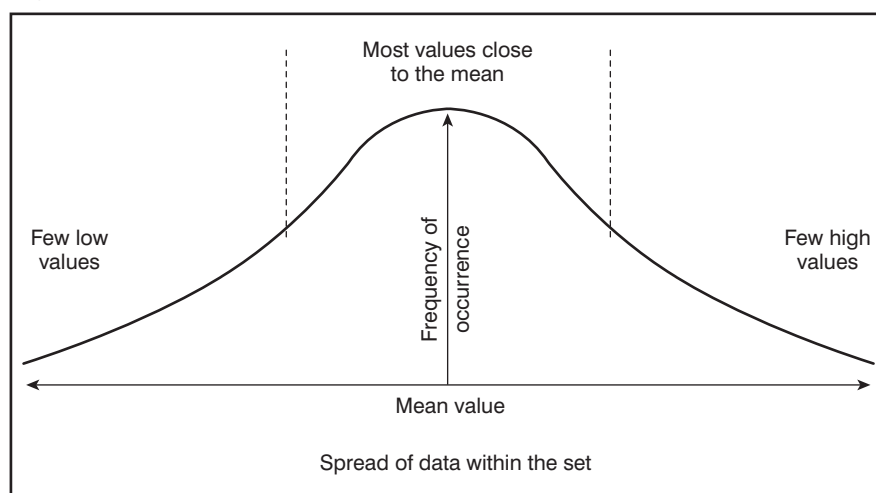


Figure 4: The normal distribution curve



not give any further indication of how the entire set of data is distributed, just the limits of the middle 50% of the data.

- Not all values are considered and hence a false impression may be given of the data set being analysed.

Despite many of the drawbacks, the interquartile range is frequently used in association with graphical techniques to represent data. One such technique is the box and whisker plot (shown in Figure 3) which, once the quartiles are found is an effective method of displaying information pertaining to the central tendency of a data set.

The box and whisker plot not only shows the interquartile range and median but also the range, and hence is a valuable method of data

presentation allowing for simple interpretation from what can be a large data set.

The interquartile range is most useful when comparing one or more data sets which appear to have similar means, medians or ranges. It effectively indicates how dispersed the data are around the median; however, without the box and whisker plot there remains no indication of the spread of the data above or below the median value!

Standard deviation

The final and perhaps the most complex of all the measures of central tendency, dispersion, deviation and variation discussed here is the standard deviation. In essence the standard deviation (represented by the greek letter sigma σ) is again a numerical value

calculated directly from the data set. It represents the average difference of the data above and below the mean value of the data set.

In order to apply the standard deviation to describe a given data set you must first make the assumption that the data within the set are normally distributed. This means:

- Most values are close to the average.
- There are only a small number of very high and very low values.
- There are equal numbers of values above and below the mean.

These assumptions when considered graphically can be represented by the following curve known as the **normal distribution curve** (Figure 4).

In order to calculate the standard deviation value (σ) of a data set it is necessary to use a formula. This formula, commonly known as the standard deviation equation, is displayed in equation box 3 below.

Equation box 3: The standard deviation equation

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

σ = the median of the set of data
 x = represents each value in the data set
 \bar{x} = mean of data set
 n = the number of pieces of data
 Σ = summation symbol

- Calculate the mean of the data set (\bar{x}).
- Calculate the difference between each piece of data and the mean value ($x - \bar{x}$).
- Square each of the differences recorded above. ($x - \bar{x}$)².
- Sum the squares. $\Sigma (x - \bar{x})^2$.
- Put the data into the standard deviation equation (Box 3) to calculate the standard deviation (σ) of the data set.

Mathematically speaking, the standard deviation is the most statistically sound technique for describing the central tendency of a data set. This is for a number of reasons:

1. In its calculation it includes **all** the data values within the set and hence there is no selective bias in choosing which pieces of data to use or ignore as in other methods.
2. It is capable of showing the variation between two data sets even if the mean values of the data sets, interquartile range or range are similar.
3. Through the application of confidence intervals it is possible to assert the likelihood of future measurements taken in line with the data set will fall within a designated range.

Standard deviation is useful to geographers in the analysis of data collected from physical measurements. Comparing rainfall figures for one location at one time of year (e.g. for January) over a period of 10 or more years would be one example. Looking for variation in samples of river water from one site tested for a particular pollutant concentration, for oxygen levels, or for numbers of a species such as water fleas over a number of years, would be another.

In addition to summarising the distribution of the data within a set around its mean value the standard deviation (σ) value can be of further use. By knowing the number of items of data within the set, the standard deviation can be compared to tables of significance to give an indication of the likelihood of the result being due to chance. These tables can also be used to infer the probability that future data collected under similar conditions will lie within one or more standard deviations of the mean value. For more information on this see the bibliography below.

Summary

In this **Geofile** a range of statistical techniques useful to the geographer in the analysis of quantitative data have been summarised. Although the limitations of each of the techniques have been illustrated it should be remembered that in the absence of these measures of central tendency and deviation, data analysis would be a laborious process and results obtained more difficult to prove.

Bibliography

Gregory, *Statistical Methods and the Geographer*, Longman.

Nagle and Witherick (1998) *Skills and Techniques for Geography A Level*, Stanley Thornes.

<http://bmj.com/collections/statsbk/1.shtml>

FOCUS QUESTIONS

1. Imagine you have recorded building height in storeys on a transect across a town. which measure of central tendency would give the most accurate indication of the most common height of the buildings within the town and what are the advantages of using this measure compared to other measures available?
2. Calculate the standard deviation of the data in Figure 1. What does the standard deviation value you have calculated imply about the variation in the temperature in June?