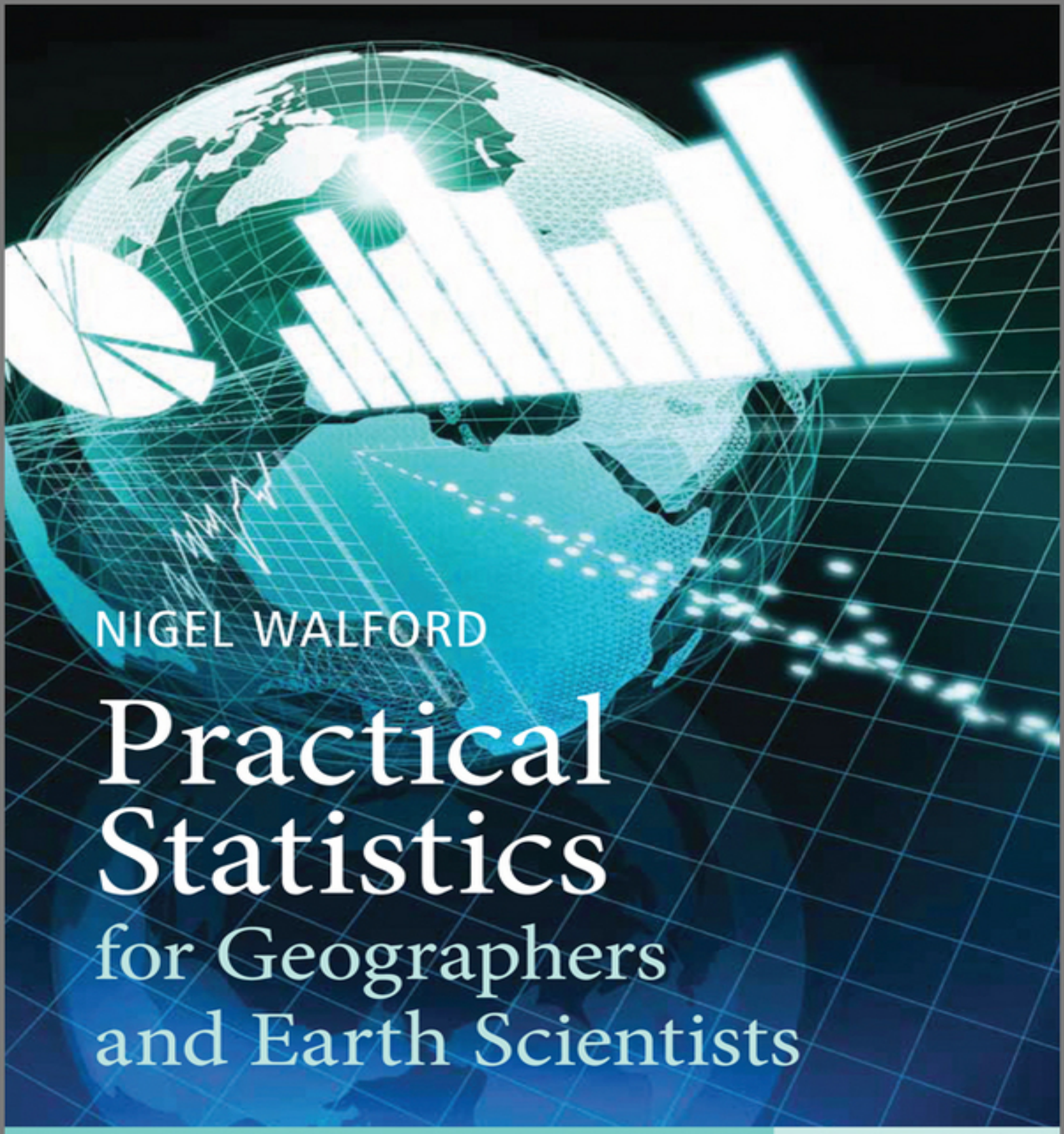

NIGEL WALFORD

Practical Statistics

for Geographers and Earth Scientists

WILEY-BLACKWELL

Contents

Practical Statistics for Geographers and Earth Scientists

Nigel Walford

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2011, © 2011 John Wiley & Sons Ltd

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices:

9600 Garsington Road, Oxford, OX4 2DQ, UK

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

111 River Street, Hoboken, NJ 07030–5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell

Library of Congress Cataloguing-in-Publication Data

Walford, Nigel.

Practical statistics for geographers and earth scientists / Nigel Walford.

p. cm.

Includes index.

ISBN 978-0-470-84914-9 (cloth) – ISBN 978-0-470-84915-6 (pbk.)

1. Geography-Statistical methods. I. Title.

G70.3.W36 2011

519.5-dc22

2010028020

A catalogue record for this book is available from the British Library.

This book is published in the following electronic formats:
ePDF: 978-0-470-67001-9

To Ann

Preface

The quantitative revolution has a lot to answer for, not least in establishing a need for students to learn about statistical techniques and how they can be applied. The majority of undergraduate students in Geography and the Earth Sciences are expected to demonstrate this understanding through carrying out independent projects as part of their degree programmes. Although it is now some 50 years since these disciplines were struck by the quantitative revolution, there remains a continuing need to reinterpret statistical techniques and to demonstrate their application in relation to contemporary issues for successive cohorts of students. Information technology in the form of powerful computers and sophisticated software has, to a large extent removed the drudgery of hand calculation that dogged the early years of quantification. In those days datasets for students were necessarily limited to small samples of observations because the time required to calculate a mean, standard deviation, correlation coefficient, regression equation or some other statistical quantity would otherwise seem interminable.

Information technology has seemingly made the application of statistics to help with answering research questions a relatively straightforward task. The entry of data into computer software is often still a time-consuming task requiring care and attention to detail in order to minimize or preferably remove any error. However, once the data have been captured in this way it is now comparatively straightforward to select from the range of statistical techniques available and to produce some results: so the need for students to undertake calculations in order to apply statistical techniques has now been removed. However, looking at how they are derived helps to understand their purpose and use. This text does not cover the details of how to use different software, how to select from the alternative

ways of presenting results or how to discuss results in a piece of written work. These topics are addressed in other places. Instead, it focuses on the practicalities of choosing statistical techniques for different situations when researching a topic in Geography and the Earth Sciences. There is a wide range of techniques available and even the comprehensive introduction given here cannot hope to include everything.

This book has been written in recognition that three features are common to many undergraduate degrees in Geography and the Earth Sciences. Students are commonly expected:

1. to develop a level of understanding and competence in statistical analysis;
2. to undertake a well-defined, limited-scale independent research investigation involving the qualitative and/or quantitative analysis;
3. to be able to interpret the results of statistical analyses presented in journal papers and other published material.

This book helps students to meet these expectations by explaining statistical techniques using examples drawn from teaching first-degree level students for over 20 years and datasets that in some cases relate directly to student projects on field-work. The chapters in the book progress from consideration of issues related to formulating research questions, collecting data and summarizing information through the application of statistical tests of hypotheses and the analysis of relationships between variables. Geography and Earth Science students are faced with not only coming to terms with classical, nonspatial statistics, but also techniques specially intended for dealing with spatial data. A range of such spatial statistics are included from simple measures describing spatial patterns to ways of investigation spatial autcorrelation and fitting surfaces. The format of the chapters includes boxed sections where the

spatial or nonspatial technique is explained and applied to one of a series of project datasets. These data sets are available on the book's website. The chapters also include some questions for students to consider either on their own or in conjunction with their tutors. The Glossary defines various statistical terms used throughout the book and acts as a constant reference point for clarifying topics. The Selected Reading assembles a range of sources into one location to enable readers to delve more deeply in the topics and techniques in the preceding chapters.

Nigel Walford

April 2010

Kingston upon Thames

Acknowledgements

I would like to thank Fiona Woods at Wiley for her patience and gentle reminders while the chapters slowly drifted in over the last few years and to others in the production and editorial offices for steering the manuscript through to publication.

Most of the figures and diagrams have been produced by Claire Ivison in the School of Geography, Geology and the Environment at Kingston University. I have been constantly mystified and astounded at how she has taken my rough drafts on scrappy bits of paper and turned them into illustrations of consistent and complete clarity.

I also acknowledge the support and assistance from colleagues at Kingston University and thank the students whose questions over the years prompted me to attempt to write this book as they attempt to understand the complexities of statistical analysis.

Living with someone trying to write a book about statistical analysis must be very difficult most of the time and I would like to offer a very sincere and heartfelt thank you to Ann Hockey, who has now had to endure this three times over the years.

Glossary

***A Posteriori* Probability** A probability that takes into account additional 'after the event' information.

***A Priori* Probability** Probability determined from past evidence or theory.

Acceptance Region The section of the range of probability associated with acceptance of the null hypothesis in hypothesis testing (e.g. 1.00 to >0.05).

Alternative Hypothesis The opposite of the Null Hypothesis.

Analysis of Variance (ANOVA) A statistical procedure used to examine whether the difference between three or more sample means indicates whether they have come from the same or different populations.

Attribute A unit of categorical data relating to an observation. Also refers to information about features or elements of a spatial database.

Autocorrelation Correlation between observations that are separated by a fixed time interval or unit of distance.

Bias Systematic deviation from a true value.

Binomial Distribution A probability distribution relating to a series of random events or trials each of which has two outcomes with known probabilities.

Bivariate Analysis where there are two variables of interest.

Bivariate Normal Distribution A joint probability distribution between two variables that follow the normal distribution and are completely uncorrelated.

Canonical Correlation Analysis A method of correlation analysis that summarizes the relationship between two groups of variables.

Categorical Data Data relating to the classification of observations into categories or classes.

Categorical Data Analysis A collection of statistical techniques used to analyse categorical data.

Census The collection of data about all members of a (statistical) population, which should not contain random or systematic errors.

Central Limit Theorem Theorem underpinning certain statistical tests (e.g. Z and t tests) that states the sampling distribution of the mean approaches normality as the sample size increases, irrespective of the probability distribution of the sampled population.

Central Tendency (Measures of) A summary measure describing the typical value of a variable or attribute.

Centroid (with spatial data) The centre of an area, region or polygon and is effectively the 'centre of gravity' in the case of irregular polygons.

Chi-square Distribution Probability distribution relating to continuous variable that is used in statistical testing with frequency count data.

Chi-square Statistic Statistic measuring the difference between observed and expected frequency counts.

Chi-square Test Statistical test used with discrete frequency count data used to test goodness-of-fit.

Cluster Sampling The entire population is divided into groups (clusters) before randomly selecting observations from the groups or a random set of the groups.

Coefficient of Variation Defined as the standard deviation divided by the mean of a variable.

Confidence Intervals (Limits) A pair of values or two sets of values that define a zone around an estimated statistic obtained from a sample within which the corresponding population parameter can be expected to lie with a specified level of probability.

Coordinate System A set of coordinate axes with a known metric.

Coordinates One, two or three ordinates defined as numeric values that are mutually independent and equal the number of dimensions in the coordinate space.

Correlation Coefficient A measure of the strength and direction of the relationship between two attributes or variables that lies within the range −1.0 to +1.0.

Correlation Matrix Correlation coefficients between a group of variables that is symmetrical either side of the diagonal where the correlation of the variables with themselves equal 1.0.

Correlogram A plot of spatially or temporally linked correlation coefficients.

Covariance The expected value of the result of multiplying the deviations (differences) of two variables from their respective means that indicates the amount of their joint variance.

Cross-sectional Data Data relating to observations at a given point in time.

Crosstabulation Joint frequency distribution of two or more discrete variables.

Data (datum) Recorded counts, measurements or quantities made on observations (entities).

Data Mining Discovery of hidden patterns in large datasets.

Datum The origin, orientation and scale of a coordinate system tying it to the Earth.

Degrees of Freedom The smallest number of data values in a particular situation needed to determine all the data values (e.g. n – 1 for a mean).

Dependent Variable A set of one or more variables that are functionally depend on another set of one or more independent variables.

Descriptive Statistics A group of statistical techniques used to summarize or describe a data set.

Dichotomous An outcome or attribute with only two possible values.

Discrete Distribution A probability distribution for data values that are discrete or integer as opposed to continuous.

Discrete Random Variable A random variable with a finite (fixed) number of possible values.

Dispersion (measures of) A measure that quantitatively describes the spread of a set of data values.

Distance Matrix A matrix of measurements that quantifies the dissimilarity or physical distance between all pairs of observations in a population or sample.

Error Difference between the estimated and true value.

Estimation Use of information from a sample to guess the value of a population parameter.

Expected Value The value of a statistical measure (e.g. mean or count) expected according the appropriate random probability distribution.

Explanatory Variable An alternative name for an independent variable.

F Distribution A group of probability distributions with two parameters relating to degrees of freedom of the numerator and the denominator that is used to test for the significance of differences in means and variances between samples.

Feature An abstract representation of a real-world phenomenon or entity.

Geocoding Allocation of coordinates or alphanumeric codes to reference data to geographical locations.

Geographically Weighted Regression A form of regression analysis that fits different regressions at different points across a study area thus weighting by spatial location.

Georeferencing Assignment of coordinates to spatial features tying them to an Earth-based coordinate system.

Geospatial Data Data relating to any real world feature or phenomenon concerning its location or in relation to other features.

Hypothesis An assertion about one or more attributes or variables.

Hypothesis Testing Statistical procedure for deciding between null and alternative hypotheses in significance tests.

Independent Events (Values) The situation in which the occurrence or nonoccurrence of one event, or the measurement of a specific data value for one observation is totally unaffected (independent of) the outcome for any other event or observation.

Independent Variable A set of one or more variables that functionally control another set of one or more dependent variables.

Inferential Statistics A group of statistical techniques including confidence intervals and hypothesis tests that seek to discover the reliability of an estimated value or conclusion produced from sample data.

Interquartile Range A measure of dispersion that is the difference between 1st and 3rd quartile values.

Interval Scale A measurement scale where 0 does not indicate the variable being measured is absent (e.g. temperature in degrees Celsius), which contrasts with the ratio scale.

Inverse Distance Weighting A form spatial data smoothing that adjusts the value of each point in an inverse relationship to its distance from the point being estimated.

Join Count Statistics Counts of the number of joins or shared boundaries between area features that have been assigned nominal categorical values.

Judgemental Sampling A nonrandom method for selecting observations to be included in a sample based on the investigators' judgement that generates data less amenable than random methods to statistical analysis and hypothesis testing.

Kolmogorov–Smirnov Test (one sample) A statistical test that examines whether the cumulative frequency distribution obtained from sample data is significantly different from that expected according to the theoretical probability distribution.

Kolmogorov–Smirnov Test (two samples) A statistical test used to determine whether two independent samples are probably drawn from the same or different populations by reference to the maximal difference between their cumulative frequency distributions.

Kriging Kriging uses inverse distance weighting and the local spatial structure to predict the values and points and to map short-range variations.

Kruskal–Wallis Test The nonparametric equivalent of ANOVA that tests if three or more independent samples could have come from the same population.

Kurtosis A measure of the height of the tails of a distribution: positive and negative kurtosis values, respectively, indicate relatively more and less observations in comparison with the normal distribution.

Lag (Spatial or Temporal) A unit of space or time between observations or objects that is used in the analysis of autocorrelation.

Level of Significance Threshold probability that is used to help decide whether to accept or reject the null hypothesis (e.g. 0.05 or 5%).

Linear Regression Form of statistical analysis that seeks to find the best fit linear relationship between the dependent variable and independent variable(s).

Local Indicator of Spatial Association A quantity or indicator, such as Local Moran's I, measuring local pockets of positive and negative spatial autocorrelation.

Mann–Whitney U Test A nonparametric statistical test that examines the difference between medians of an ordinal variable to determine whether two samples or two groups of observations come from one population.

Mean The arithmetic average of a variable for a population or sample is a measure of central tendency.

Mean Centre The central location in a set of point features that is at the intersection of the means of their *X* and *Y* coordinates.

Mean Deviation Average of the absolute deviations from a mean or median.

Measurement Error Difference between observed and true value in the measurement of data usually divided into random and systematic errors.

Median Middle value that divides an order set of data values into two equal halves and is the 50th percentile.

Median Centre A central location in a set of point features that occurs at the intersection of the median values of their *X* and *Y* coordinates or that divides the points into four groups of equal size.

Mode Most frequently occurring value in a population or sample.

Moran's *I* A quantitative measure of global or local spatial autocorrelation.

Multiple Regression Form of regression analysis where there are two or more independent variables used to explain the dependent variable.

Multivariate Analysis where there are more than two variables of interest.

Nearest-Neighbour Index An index used as a descriptive measure to compare the patterns of different categories of phenomena within the same study area.

Nominal Scale Categories used to distinguish between observations where there is no difference in magnitude between one category and another.

Nonparametric Tests Form of inferential statistics that makes only limited assumptions about the population parameters.

Normal Distribution Bell-shaped, symmetrical and singl-peaked probability density curve with tails extending to plus and minus infinity.

Normality Property of a random variable that conforms to the normal distribution.

Null Hypothesis Deliberately cautious hypothesis that in general terms asserts that a difference between a sample and population in respect of a particular statistic (e.g. mean) has arisen through chance.

Observed Value The value of a statistical measure (e.g. mean or count) obtained from sample data.

On-Tailed Test Statistical test in which there is good reason to believe that the difference between the sample and population will be in a specific direction (i.e. more or less) in which case the probability is halved.

Ordinal Scale Data values that occur in an ordered sequence where the difference in magnitude between one observation and another relates to its rank position in the sequence.

Ordinary Least Squares Regression Most common form of ordinary linear regression that uses least squares to determine the regression line.

Outlier An observation with an unusually high or low data value.

P Value Probability that random variation could have produced a difference from the population parameter as large or larger than the one obtained from the observed sample.

Paired-Sample Data Measurement of observations on two occasions in respect of the same variable or single observations that can be split into two parts.

Parameter A numerical value that describes a characteristic of a probability distribution or population.

Parametric Tests Type of inferential statistics that make stringent assumptions about the population parameters.

Pearson Correlation Coefficient A form of correlation analysis where parametric assumptions apply.

Percentile The percentage of data values that are less than or equal to a particular point in the percentage scale from 1 to 100 (e.g. the data value at which 40% of all values are less than this one is 40th percentile).

Poisson Distribution A probability distribution where there is a discrete number of outcomes for each event or measurement.

Polygon A representation of area features.

Polynomial A function, for example a regression equation containing *N* terms for the independent variable raised to the power of *N*.

Population A complete set of objects or entities of the same nature (e.g. rivers, human beings, etc.).

Precision The degree of exactness in the measurement of data.

Quadrat A framework of usually regular, contiguous, square units superimposed on a study area and used as a means of counting and testing the randomness of spatial patterns.

Quartile The 1st, 2nd and 3rd quartiles correspond with the 25th, 50th and 75th percentiles.

Random Error The part of the overall error that varies randomly from the measurement of one observation to another.

Random Sampling Sometimes called simple random sampling, this involves selecting objects or entities from a

population such that each has an equal chance of being chosen.

Range A measure of dispersion that is the difference between the maximum and minimum data values.

Rank Correlation Coefficient A method of correlation analysis that is less demanding than Pearson's and involves ranking the two variables of interested and then calculating the coefficient from the rank scores.

Raster Representation of spatial data as values in a matrix of numbered rows and columns that can be related to a coordinate system in the case of geospatial data.

Ratio Scale A scale of measurement with an absolute zero where any two pairs of values that are a certain distance apart are separated by the same degree of magnitude.

Regression Analysis A type of statistical analysis that seeks the best fit a mathematical equation between dependent and independent variable(s).

Regression Line Line that best fits a scatter of data points drawn using the intercept and slope parameters in the regression equation.

Relative Frequency Distribution Summary frequency distribution showing relative percentage or proportion of observations in discrete categories.

Residuals The differences between observed and predicted values commonly used in regression analysis.

Root Mean Square (RMS) The square root of the mean of the squares of a set of data values.

R-squared The coefficient of determination acts as a measure of the goodness of fit in regression analysis and is thought of as the proportion or percentage of the total variance accounted for by the independent variable(s).

Sample A subset of objects or entities selected by some means from a population and intended to encapsulate the latter's characteristics.

Sample Space The set of all possible outcomes from an experiment.

Sample Survey A survey of a sample of objects or entities from a population.

Sampling The process of selecting a sample of objects or entities from a population.

Sampling Distribution Frequency distribution of a series of summary statistics (e.g. means) calculated from samples selected from one population.

Sampling Frame The set of objects or entities in a population that are sampled.

Scatter Plots (Graphs) A visual representation of the joint distribution of observed data values of two (occasionally three) variables for a population or sample that uses symbols to show the location of the data points on X, Y and possibly Z planes.

Simple Linear Regression Simplest form of least squares regression with two parameters (intercept and slope).

Skewness A measure of the symmetry (or lack of it) of a probability distribution.

Smoothing A way of reducing noise in spatial or temporal datasets.

Spline The polynomial regression lines obtained for discrete groups of points that are tied together to produce a smooth curve following the overall surface in an alternative method of fitting a surface.

Standard Deviation A measure of dispersion that is the square root of the variance and used with interval and ratio scale measurements.

Standard Distance A measure of the dispersion of data points around their mean centre in two-dimensional space.

Standard Error A measure of the variability of a sample statistic as it varies from one sample to another.

Standard Normal Distribution A version of the normal distribution where the mean is zero and the standard deviation is 1.0.

Standard (Z) Score The number of units of the standard deviation that an observation is below or above the mean.

Statistic Either a number used as a measurement or a quantity calculated from sample data.

Statistical Test A procedure used in hypothesis testing to evaluate the null and alternative hypotheses.

Stratified Random Sampling A method of selecting objects or entities for inclusion in a sample that involves separating the population into homogenous strata on the basis of some criterion (e.g. mode of travel to work) before randomly sampling from each stratum either proportionately or disproportionately in relation to the proportion of the total entities in each stratum.

Systematic Error A regular repeated error that occurs when measuring data about observations.

Systematic Sampling A method of selecting objects or entities from a population for inclusion in some regular sequence (e.g. every 5th person).

t-Distribution A bell-shaped, symmetrical and single-peaked continuous probability distribution that approaches the normal distribution as the degrees of freedom increase.

Time Series Data that includes measurements of the same variable at regular intervals over time.

Time Series Analysis Group of statistical techniques used to analyse time series data.

Trend Surface Analysis The 'best fitting' of an equation to a set of data points in order to produce a surface, for example by means of a polynomial regression equation.

t-statistic Test statistic whose probabilities are given by the t-distribution.

t-test A group of statistical tests that use the t-statistic and t-distribution that are in practice rather less demanding in

their assumptions than those using the normal distribution.

Two-Tailed Test Most practical applications of statistical testing are two-tailed, since there is no good reason to argue that the difference being tested should be in one direction or the other (i.e. positive or negative) and the probability is divided equally between both tails of the distribution.

Type-I Error A Type-I error is made when rejecting a null hypothesis that is true and so concluding that an outcome is statistically significant when it is not.

Type-II Error A Type-II error is committed when a null hypothesis is accepted that is false and so inadvertently failing to conclude that a difference is statistically significant.

Univariate Analysis where there is one variable of interest.

Variance A measure of dispersion that is the average squared difference between the mean and the value of each observation in a population or sample with respect to a particular variable measured on the interval or ratio scale.

Variance/Mean Ratio A statistic used to describe the distribution of events in time or space that can be examined using the Poisson distribution.

Variate An alternative term for variable or attribute, although sometimes reserved for things that are numerical measurements (i.e. not attribute categories).

Variogram (Semivariogram) Quantification of spatial correlation by means of a function.

Vector Representation of the extent of geographic features in a coordinate system by means of geometric primitives (e.g. point, curve and surface).

Weighted Mean Centre The mean centre of a set of spatial points that is weighted according to the value or amount of a characteristic at each location.

Weights Matrix A matrix of numbers between all spatial features in a set where the numbers, for example 0 and 1,

indicate the contiguity, adjacency or neighbourliness of each pair of features.

Wilcoxon Signed Ranks Test A nonparametric equivalent of the paired sample t-test that tests a null hypothesis that difference between the signed ranks is due to chance.

Wilk's Lambda A multivariate test of difference in mean for three or more samples (groups).

Z distribution Probability distribution that is equivalent to the standardized normal distribution providing the probabilities used in the Z test.

Z Statistic Test statistic whose probabilities are given by the Z distribution.

Z Test A parametric statistical test that uses the Z statistic and Z distribution that makes relatively demanding assumptions about the normality of the variable of interest.

Z-score A Z-score standardizes the value of a variable for an observation in terms of the number of standard deviations that it is either above or below the sample or population mean.

Section 1

First principles

1

What's in a number?

Chapter 1 provides a brief review of the development of quantitative analysis in Geography, Earth and Environmental Science and related disciplines. It also discusses the relative merits of using numerical data and how numbers can be used to represent qualitative characteristics. A brief introduction to mathematical notation and calculation is provided to a level that will help readers to understand subsequent chapters. Overall this introductory chapter is intended to define terms and to provide a structure for the remainder of the book.

Learning outcomes

This chapter will enable readers to:

- outline the difference between quantitative and qualitative approaches, and their relationship to statistical techniques;
- describe the characteristics of numerical data and scales of measurement;
- recognize forms of mathematical notation and calculation that underlie analytical procedures covered in subsequent chapters;
- plan their reading of this text in relation to undertaking an independent research investigation in Geography and related disciplines.

1.1 Introduction to quantitative analysis

Quantitative analysis comprises one of two main approaches to researching and understanding the world around us. In simple terms quantitative analysis can be viewed as the processing and interpretation of data about things, sometimes called phenomena, which are held in a numerical form. In other words, from the perspective of Geography and other Earth Sciences, it is about

investigating the differences and similarities between people and places that can be expressed in terms of numerical quantities rather than words. In contrast, qualitative analysis recognizes the uniqueness of all phenomena and the important contribution towards understanding that is provided by unusual, idiosyncratic cases as much as by those conforming to some numerical pattern. Using the two approaches together should enable researchers to develop a more thorough understanding of how processes work that lead to variations in the distribution of phenomena over the Earth's surface than by employing either methodology on its own.

If you are reading this book as a student on a university or college course, there will be differences and similarities between you and the other students taking the same course in terms of such things as your age, height, school level qualifications, home town, genetic make-up, parental annual income and so on. You will also be different because each human being, and for that matter each place on the Earth, is unique. There is no one else exactly like you, even if you have an identical twin, nor is there any place exactly the same as where you are reading this book. You are different from other people because your own attitudes, values and feelings have been moulded by your upbringing, cultural background and physical characteristics. In some ways, it is the old argument of nature versus nurture, but in essence we are unique combinations of both sets of factors. You may be reading this book in your room in a university hall of residence, and there are many such places in the various countries of the world and those in the same institution often seem identical, but the one where you are now is unique. Just as the uniqueness of individuals does not prevent analysis of people as members of various different groups, so the individuality of places does not inhibit investigation of their distinctive and shared characteristics.