Fourth Edition

Statistical methods and the geographer

S Gregory

Statistical methods and the geographer

S Gregory
Professor of Geography, University of Sheffield

Fourth Edition

Longman
London and New York

Longman Group Limited London

Associated companies, branches and representatives throughout the world

Published in the United States of America
by Longman Inc., New York

First published 1963
Second edition 1968
Third edition 1973
Fourth edition 1978

ISBN 0 582 48186 4 paper

Library of Congress Cataloging in Publication Data

Gregory, Stanley.
Statistical methods and the geographer.

(Geographies for advanced study)
Bibliography: p.
Includes index.
1. Geography–Statistical methods. I. Title.
G70.3.G74 1973 519.5'02'491 77-7025
ISBN 0-582-48186-4

Printed in Great Britain by
Richard Clay (The Chaucer Press) Ltd,
Bungay, Suffolk.

Acknowledgements

We are grateful to the Syndics of the University Press, Cambridge, for permission to include adaptations of tables from *Cambridge Elementary Statistical Tables* by D. V. Lindley and J. C. P. Miller. Table 23 was reproduced from Table L which appeared in *Non-parametric Statistics for the Behavioural Sciences* by S. Siegel (McGraw-Hill Book Company), and was abridged from the sources: *Annals of Mathematic Statistics* and *Psychological Bulletin*, and we are indebted to the McGraw-Hill Book Company, the Treasurer of the Institute of Mathematical Statistics and the managing Editor of American Psychological Association for their respective permissions for the inclusion of this copyright material.

Preface

The origins of this book lie in the author's experiences, as student, research worker and lecturer, over the past 15 years. The intricacies and essential characteristics of statistical methods were first introduced to him as a student by Professor P. R. Crowe, when the latter was a Reader in the University of London. The value of such methods of analysis has been increasingly appreciated as research, especially in the field of climatology, has been pursued during succeeding years.

For the non-mathematician, however, even the simpler introductory books on statistics often raise considerable problems. These are accentuated, moreover, by the fact that the methods are applied to fields of study which are, in large measure at least, unfamiliar to the geographer—industrial or business control, sociology or economic theory, the biological sciences or medicine, or simply as a study in applied mathematics. Moreover, most geographical studies that have employed statistical techniques have equally tended simply to assume that the reader would understand the methods despite the normal lack of formal statistical training.

In an attempt to counteract these tendencies, training in statistical methods for geography students was expanded at Liverpool University in 1957. This training aimed at providing a grounding in a variety of basic methods, all of which were developed and applied in terms of geographical problems. From the course has evolved the present book, which it is hoped will provide a similar basic grounding for all geographers. Throughout the evolution of this course, and especially in encouraging me to expand it in the present form, I have had every support from Professor R. W. Steel. It is my former colleague, Dr A. T. A. Learmonth, however (now Professor of Geography in the School of General Studies, Australian National University, Canberra), to whom the greatest debt is owed, for his unfailing willingness to discuss and constructively criticize my efforts, for his persistence in exhorting me to proceed with the work, and for invaluable advice and assistance.

There are many others who, in their various ways, have provided help and guidance. Amongst these are Professor S. H. Beaver, who read and commented on the text; Dr D. J. Bartholomew, formerly lecturer in Statistics at the University of Keele, whose advice at an early stage helped to set the pattern of this book; Mr P. K. Mitchell, the Geography Department, the University College of Sierra Leone, a colleague during my year in Sierra Leone (1960–1961), when the bulk of this book was written;

Miss P. J. Treasure of the Geography Department, the University of Liverpool, who drew most of the diagrams; Miss E. M. Shaw, the University of Keele, for help at proof stage; and all my colleagues at Liverpool who willingly allowed me to try my ideas upon them.

To them, and to many others, my thanks are due—I trust that they approve of the final product.

S. G.
Liverpool, 1962

Second edition

The continually growing interest amongst geographers, at research, undergraduate and now sixth-form levels, in the relevance of statistical techniques to the subject, has made it desirable that this book should undergo some revision, both to meet wider demands and to satisfy the need for more effective presentation. The author is extremely grateful to those many friends and colleagues who wrote to him concerning the first edition, whether this was to commend sections which seemed satisfactory or to raise questions concerning others which did not completely fall into this category. Their comments and advice have all been seriously considered when the text was being revised, even when action has nevertheless not been taken.

Apart from innumerable minor textual changes which, it is hoped, will make for more effective reading and understanding, the major changes are fivefold. First, some attention has been given to the problem of the transformation of data in order to reinforce the appreciation of the need for normally-distributed data for the use of so many techniques. Secondly, the use of probability paper, at least in simple terms, has been introduced to illustrate the ways in which the labour of probability assessments can be circumvented. Thirdly, radical changes have been made, plus considerable expansion added, to the theme of non-parametric testing, to provide a more systematic approach to what is a most important group of possible techniques for geographers. Fourthly, change and expansion are also reflected in the sections on correlation and regression, including some simple consideration of curvilinear relationships and the presentation of computational techniques more geared to the use of desk calculators rather than long-hand methods. Fifthly, the bibliography has also been expanded, to incorporate a wider range of books on techniques and a selection of research papers using such techniques in a geographical (or near-geographical) context.

Nevertheless, the overall structure and framework of the book remains basically unaltered; the intellectual approach is still one of presenting the techniques as simply as possible, leaving those requiring more advanced techniques to move on to the appropriate advanced texts, and of allowing geographers to be introduced to these potentially valuable methods through examples which, it is hoped, seem relevant to them, i.e. in terms of geographical problems. The growth of the use of such methods, both of the simple and the complex varieties, over the past few years not only further stresses the need for all geographers to be conversant with such techniques and their implications, but also encourages the author to hope that these improvements in this introductory text will be appreciated and approved of by all its users.

S. G.
Liverpool, 1967

Third edition

The need to modify a book of this sort is always apparent and this opportunity to do so has been gratefully appreciated. Fundamental changes are few, however, and the express purpose of providing a simple, largely non-technical introduction, for non-numerate geographers, has been retained.

In the four years since the second edition was prepared, the number of books on more advanced techniques, and of articles and books using such techniques in geography, has expanded considerably, and many of these are included in the bibliography. The types of problems studied by geographers have also become more varied, but changes to incorporate some of these newer fields have deliberately not been made, for the methods presented are independent of specific fields and approaches. Moreover, the more traditional and simpler problems used here are perhaps more suitable as basic teaching examples—other uses and applications can be found in the texts listed.

This continued growth of more advanced work, itself a most welcome development, implies that a growing number of practising geographers are now operating at levels far beyond those in this book. It is still necessary, however, for the initial stages of statistical manipulation and thinking to be learned and appreciated by all those entering the field. It is hoped that, for these, this book will continue to prove of value and help as they take their first steps along this rocky but rewarding path.

S. G.
Sheffield, 1972

Fourth edition

The decision to redesign the lay-out and format of this volume has also presented the opportunity to restructure and update its contents. As a result, the relationships between various sections of the book have been modified in an attempt to improve the coherence and logical pattern of the ideas and methods that are included, whilst at the same time certain deletions have been more than counterbalanced by a number of new introductions. Amongst the latter are included additional comments on data characteristics and on sampling procedures, an expansion in the number of non-parametric testing techniques that have been outlined, illustrations of further uses of the analysis of variance, and a number of items concerned with correlation and regression. Putting all these changes together, it is hoped that the reader will find this modified format and new content both easier to use and even more helpful as a guide to action than were the previous editions.

The need for a text that is essentially introductory and specifically intended as an operational aid is even more marked now than when the first edition of this volume appeared fifteen years ago. Then, most geographers were still non-quantitative in their approaches to the subject, and the main purpose of the volume was to persuade as many of them as possible that there was some real value in attempting to acquire an elementary competence in statistical methods—as well as indicating that this did not pose insuperable problems. Nowadays, the quantitative methodology employed at the research level is both advanced and sophisticated, so that if undergraduates do not obtain a sound foundation, both in terms of basic concepts and simple methods, then the step needed to carry out research—*and even to read and understand research publications fully*—is far greater than it was in the early 1960s. The underlying objective of this book therefore remains the same as was indicated in the preface to the third edition, i.e. to assist in the learning and appreciation of the initial stages of statistical manipulation, as geographers commence their training in this field. It is this teaching objective that conditions the selection of therr es, the uncomplicated nature of the geographical problems used, and the mixed verbal/mathematical presentation, by all of which it is intended to communicate, at a very simple level, with those approaching the use of statistical methods in geography for the first time. It must always be remembered that, although for the subject at large the relevance and application of quantitative methods to geography is now commonplace (if far from universal!), for each new intending geographer it is still as confusing and difficult as it was two decades ago. Statistical methods and the geographer still need an effective interface, and it is hoped that this book can continue to satisfy this need.

S. G.
Sheffield, 1977

Introduction

The type of geography which admits the importance of quantification and the appropriateness of statistical methodology, but always as servants and not as masters, would appear to be the best answer the profession can furnish to the embarrassing questions which have arisen during the current debate in academic circles regarding geography's right to be included in the curricula of institutions of higher learning.
William Warntz

In the third quarter of the twentieth century the raw material with which the geographer deals became progressively more of a quantitative nature and less merely qualitative. This gradual but steady change in emphasis of necessity engendered a modification of the intellectual approach to the subject. As in any other worthwhile field of study, so in geography each generation attempts to absorb, and then advance beyond, the accumulated work of previous generations; this is no more than the outward sign of healthy development. These advances may at times be in terms of factual knowledge. At other times, however, they reflect a changing approach to the subject at large, such as this present conscious and deliberate attempt to provide a more quantitative approach to the geographer's problems.

In all branches of the subject this tendency has developed. Climatological investigations have traditionally and necessarily been concerned with numerical data. Economic geography, too, has for long utilized quantitative data as a prime source of information, although explanatory studies have tended to rest more heavily on subjective judgments than would in many cases seem desirable. Geomorphology, population studies and various other aspects of human geography, amongst many branches of the subject, have also increasingly turned to more precise numerical data over the recent past, all in the attempt to render a more accurate and objective assessment of the geography of particular areas or problems. Moreover, as geographers increasingly co-operate with scientists from other disciplines, or engage in the practical fields of planning, the need to present both data and conclusions in sound quantitative terms becomes even more pressing.

Once such an attitude is accepted, however, a necessary corollary follows, that these numerical data should be analysed by sound statistical methods so that maximum value is obtained from them. Too often a considerable body of valuable quantitative data is presented either in a raw state or after a minimal amount of processing. Sometimes, of course, this may be quite legitimate as it is all that the problem requires. In other cases, however, more fundamental, and possibly more valid, conclusions could be reached, or varied aspects of a problem investigated, by means of a more comprehensive and subtle use of existing statistical methods. Moreover, it is not simply that such methods are not always used, but that at times false interpretations are made either because of the failure to apply such methods or because they are misunderstood. The latter may unfortunately arise when a geographer quite properly consults a professional statistician without at the same time fully understanding the implications of the results which are obtained by the methods with which he is provided.

The aim of this book is therefore to present standard statistical techniques in a simple manner and to apply them to problems typical of those which geographers consider. In this way a twofold purpose is served. On the one hand the requirements of practising geographers engaged in research are at least partially met by the presentation of methods and techniques, at a relatively simple level, which should enable many geographical problems to be analysed more soundly. This is not intended to be a comprehensive work covering the full field of statistics, but rather a selective presentation of elementary methods, which are particularly applicable to geographical problems. For the investigation of more complex problems the standard statistical texts, of which a selected list is incorporated in the bibliography, must be consulted. On the other hand, the introduction to relevant elementary statistics which this book will provide will enable all students of geography more readily to interpret and understand studies based on statistical analyses. Many of the misinterpretations which occur at present result from the *reader's* failure to be conversant with either the advantages or the limitations implicit in any writer's statistical methods—this renders difficult the full and accurate assessment of the value and implication of what is written. From both viewpoints, therefore—from that of the geographer trying to analyse and present his material more effectively, and that of the student of geography trying to interpret and understand existing studies—it is hoped that this excursion into statistical methods and their uses to geographers will prove of value.

A fundamental difficulty arises here, however, and it is one which is inherent in the whole training which most potential geographers receive from their childhood onwards. Most aspiring geographers in Britain have indeed studied mathematics for G.C.E. or C.S.E. examinations. In far too many schools, however, it is either administratively impossible, or academically not permissible, to study both geography and mathematics together up to Advanced level. This lack of sixth-form training, or perhaps the actual training received prior to that date, tends to leave many prospective geographers with a built-in resistance to anything which vaguely suggests mathematics. Directly $(a + b)$ is written on the blackboard, or a square root is required, a mental barrier is irrationally erected.

This quite needless refusal to attempt to tackle such problems tends to nullify attempts to put geography on a sounder footing in its handling of quantitative data.

Throughout this book, therefore, the deliberate design is to lead the reader by the hand through these apparently difficult by-ways. Save where it is absolutely necessary, there is no attempt to delve into the mathematical theories behind the methods, but rather the concepts involved are presented in plain English instead of, or as well as, in symbols. The computational problems involved should not unduly strain the capabilities of any normally intelligent fifteen year old. What is required, on the other hand, is a conscious willingness to follow a statistical argument through to its logical conclusion, to breach this mental barrier of which I have written and in that way to discover an invaluable tool which was neglected by geographers for far too long. A short selection of publications which have used these techniques is included in the bibliography.

Thus this book is not designed for statisticians; nor does it claim to make statisticians of those who work their way through it. Many possible methods, or applications of methods, which could have been included have instead been deliberately omitted. Rather, a selection of useful methods that can be applied in the field of geography are presented, and illustrated in terms of problems which the geographer can understand. The methods themselves are in common use in so many other disciplines already, and explained—in greater or lesser complexity and clarity—in numerous other books. It is to this wide range of statistical texts at a more advanced level, such as those included in the bibliography, that the enthusiast or the specialist must turn, if the series of simple illustrations in this book stimulates further enquiry. It is not primarily as an introduction to these more advanced statistical studies that this book is designed, however. If, instead, it succeeds in enabling geographical students to handle and interpret quantitative data more effectively, then the author will feel that it has more than fulfilled its purpose.

Contents

Chapter 1

Characteristics of data

The methods and techniques used in the analysis of statistical data are in large measure controlled by the very character of the statistical data themselves. It is therefore necessary to begin with a very brief consideration of some of these characteristics so that the varied themes that will be introduced later will be more readily understood.

When any collection of data, representing some quantitative value of any given phenomenon, is to be processed it will be found that although such data all represent the same phenomenon they are not all of exactly the same value. Thus if a study were being made of the distance inland from the coast that vessels of a given draught could sail it would be found that these distances vary markedly between one river and another, or between one part of the world and another. Again, if the number of vessels sailing along these rivers were examined a very wide range in values would be found between the different rivers. This highly variable nature of the numerical data is common, to a greater or lesser extent, to all sets of data, and this quantity which varies (mileage, or numbers of vessels, in the two cases given above) is known as the *variate*, or sometimes as the *variable*.

Three broad sets of distinctions concerning such variates need to be borne in mind. Firstly, there are a range of possible types of units in terms of which data are expressed—nominal or classificatory; ordinal or ranking; interval; ratio.

(a) *Nominal*. This is a group of data which all too often in the past has been assumed by geographers to preclude quantitative description or testing. Moreover, it is a frequently occurring category of data in geography—the distinction of settlements into Celtic, Anglo-Saxon and Scandinavian origins; the classifying of soils into podzols, brown earths and rendzinas; the distinction of forest, grassland and heath vegetation complexes; the recognition of various tribal, racial or cultural groups; the functional divisions of towns or the land-use division of rural areas. None of these carries implications of quantity, nor even of relative order of magnitude; they simply refer to categories that are different from one another. Nevertheless, under sampling the various categories may occur with differing degrees of frequency, and these provide data in a form that can be analysed statistically.

(b) *Ordinal*. This is also a very common group of data in geography, in that the *relative* importance (or order of magnitude) of data may be known, even though their absolute values are not. In other words, the data can be ranked or put in order, either individually or in classes. Sometimes this reflects constraints that exist upon data collection, such that only rankings are known; in other cases, the use of data in ordinal form is a deliberate choice, even though other data forms could have been used.

(c) *Interval.* When not only is the order of magnitude known, but also the actual degree of magnitude as well, then an interval scale exists. This is characteristic of rainfall data, production values, population returns, and many other types of data of geographical relevance. In all these and similar cases, either exact measurements are made in some standard unit, or the occurrences of the phenomenon are counted.

(d) *Ratio.* In this fourth category, interval data have been converted into another form. For example, the number of persons in a given socio-economic group may be expressed as a proportion of the total population, or the number of persons voting for a particular party expressed as a percentage of the total electorate. Again, measured values may have been converted into an index, such as a pH value or an index of production. Such ratio values are often, but not invariably, characterized by finite upper and lower limits.

Secondly, a distinction must be made between *continuous* and *discrete* variates. For example, in the case of the navigable mileage of rivers outlined above, it is possible for *any* mileage value to be recorded and for fractions of a mile to be included. In other words, it is a continuous variate such that there are no clear-cut or sharp breaks between the values that are possible. Such continuous variates occur with *measured* interval data, or with ratio data. On the other hand, the number of vessels actually sailing these rivers can only be in terms of whole numbers or integers, for fractions of vessels cannot be recorded. Such a variate is known as discrete, and special care must be taken when interpreting the results of the analysis of such discrete variates. Interval data based on the counting of occurrences fall into this category.

A third distinction that must be made is between data for *individual* items and data that are *grouped* into classes or cells. The listing of each item separately is possible for all types of data units except the nominal category, which by definition implies the number of occurrences in a given class. The grouping of data can be effected for all types of data, whether this be because of the form in which data are made available, because of doubts concerning the precise accuracy of interval or ratio measurements, or for convenience in calculations or testing procedures. For example, economic or social data may be obtained from official bodies such as employment exchanges or government departments, which are often precluded by law from making individual values available. Thus the numbers of people employed by individual firms may vary from one to some high value, but data may be available only in a series of classes (1-50, 51-100, etc.). Again, the profitability or costs of certain operations may be defined by firms or farmers as high, medium and low, because they are unwilling to make actual values available. At other times, difficulties of measurement or recording may make the ordinal form of data more convenient than the interval form, as when classifying river-bed load as coarse, medium and fine, or slopes as steep, moderate and gentle, or soils as acid, neutral and alkaline. In all these cases, however, there exists some implicit underlying continuum in terms of magnitude, the discrete categories being merely a convenient division.

The variable nature of geographical data can best be understood and appreciated if the data are plotted graphically to show the frequency of occurrence of values of different given amounts. The data are first grouped into 'classes', so that it is known how many occurrences fall into each of a series of quantitatively different sets of conditions. Then the number of occurrences are plotted against the appropriate 'class', and a diagram drawn in the form of 'building blocks'. Such a diagram is known as a histogram and the pattern which it presents is called the frequency distribution for that set of data. From such a diagram a smoothed curve can be interpolated, this being known as the 'frequency curve' of that set of data. Thus in Fig. 1 can be seen the frequency distribution for population densities of the European nation-states. The values for individual states are grouped into various classes depending on their order of magnitude (e.g. 0-49·9 persons per sq. km.; 50-99·9 persons per sq. km.), and the variable character of these population densities is readily apparent. The way in which these population densities vary is shown by both the 'blocks' and by the smoothed curve. A similar frequency distribution curve can be constructed for any and all sets of data. Figure 2, for example, shows the distribution of hill summit heights in North Wales based on summit ring-contours taken from the provisional edition of the O.S. 1:25,000 maps. As with the population densities, these summit heights are a continuous variate. Moreover, both Fig. 1 and Fig. 2 also display another feature of

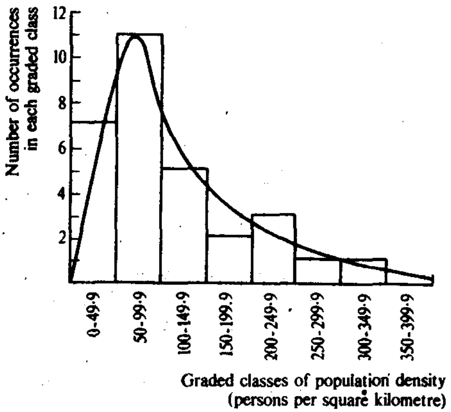

Fig. 1 Histogram and frequency distribution curve for population densities of the European nation-states

many distribution curves. It can be seen clearly that these curves are not symmetrical, having their peak markedly to one side. Such a distribution is known as *skew*, and the problems which this introduces, together with various methods by which these problems may be largely solved, will be considered later.

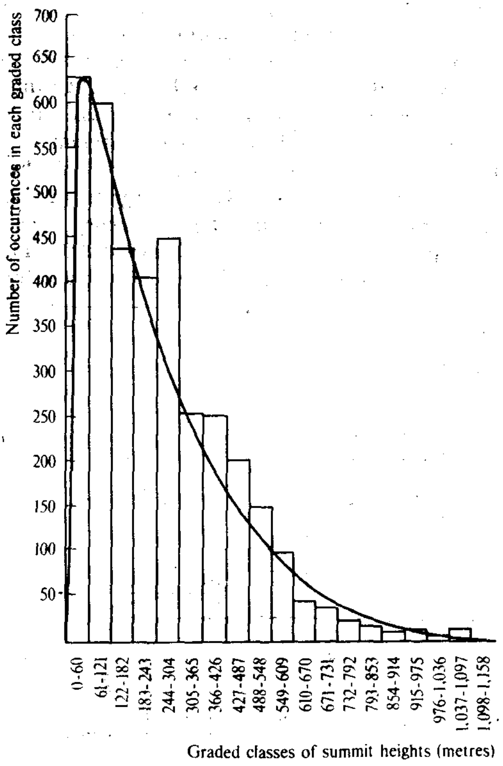

Fig. 2 Histogram and frequency distribution curve for hill-summit heights in North Wales

An alternative way of depicting a data set graphically is by means of a *cumulative frequency curve* or *ogive*. For example, in Table 1 are the grouped data for annual rainfall at Rivelin Reservoir, Sheffield, for the period 1901-1950. These are presented as a histogram in Fig. 3*a*.

If the occurrences are accumulated from the least to the largest (Table 1), it will be seen that, for example, 8 values out of 50 are equal to or less than 799 mm, and 44 values out of 50 are equal to or less than 1,099 mm. Such frequencies can be converted to percentages, and plotted on a graph against the upper limit of the appropriate class. Conventionally, this conversion is made as follows: