

STATISTICAL CONCEPTS IN
GEOGRAPHY

JOHN SILK

Statistical Concepts in Geography

JOHN SILK

University of Reading

London

GEORGE ALLEN & UNWIN

Boston

Sydney

© John Silk, 1979

This book is copyright under the Berne Convention. No reproduction without permission. All rights reserved.

**George Allen & Unwin (Publishers) Ltd,
40 Museum Street, London WC1A 1LU, UK**

George Allen & Unwin (Publishers) Ltd,
Park Lane, Hemel Hempstead, Herts HP2 4TE, UK

Allen & Unwin Inc.,
Fifty Cross Street, Winchester, Mass. 01890, USA

George Allen & Unwin Australia Pty Ltd,
8 Napier Street, North Sydney, NSW 2060, Australia

First published in 1979

Third impression, 1985

British Library Cataloguing in Publication Data

Silk, John

Statistical concepts in geography.

1. Geography – Statistical methods

I. Title

519.5'02'491 G70.3 78-40957

ISBN 0-04-910065-3

ISBN 0-04-910066-1 Pbk

Typeset in 10 on 12 point Press Roman by Preface Ltd, Salisbury, Wilts.
and printed in Great Britain by Mackays of Chatham Ltd

Statistical Concepts in Geography

In memory of Philip

Preface

This book is intended primarily for first-year undergraduate students in universities, polytechnics and colleges of further and higher education, and as such it enters what is an increasingly crowded market. However, I believe there is room for a text which tackles a number of particularly important problems at the introductory level. First, basic statistical concepts, such as those of probability, independence, randomness and sampling distribution, are emphasised, and illustrated by in-class experiments supplemented by computer-generated results. In this respect, the overall objective is to introduce greater rigour with little or no increase in mathematical formality. Secondly, the student may work through the text at her or his own pace, checking on the grasp of concepts and techniques by doing the in-text boxed exercises before reaching the more testing territory in the exercises at the end of each chapter. Thirdly, I have tried to show the relevance of statistical techniques to geographic problems through the use of numerous real-world data sets – some of which are taken from student project and thesis work – in the text and in exercises, backed by carefully selected references. Finally, no mathematical competence beyond sound Ordinary Level or high-school graduation standard is assumed, and every effort has been made to minimise the initial feeling of ‘symbol shock’ that many students encounter. Although complex derivations are avoided, references to more advanced technical treatments are also provided for those interested. Most of the ideas in the book have arisen while teaching a course at Reading for the past nine years and, needless to say, result from numerous trial-and-error exercises which students have not always borne stoically.

Many people helped, directly and indirectly, with this book. Dick Chorley and Peter Haggett inspired me to become an academic geographer in the first place, and all my colleagues at the Department of Geography in Reading University have provided a very pleasant atmosphere in which to work. In particular, Geoff Lucas commented on various parts of an early draft of the manuscript, and Sophie Bowlby, Ian Fenwick, Dave Foot, John Hardy, Trevor Meadows and John Townshend kindly made available sets of data. Ronnie Savigear provided many papers giving examples of statistical applications in physical geography, and I am particularly grateful to all those physical geography colleagues who patiently explained the background to some of the physical geography exercises to a ‘mere human’! A number of colleagues in the Department of Applied Statistics at Reading University, particularly Roger Mead, very kindly discussed a number of statistical issues with me at some length, although of course all responsibility for the particular interpretations presented, and any errors, are mine alone. I am also extremely grateful to Angela

Gurnell, Dave Unwin and Neil Wrigley for providing detailed comments on the 'penultimate draft' – they caused me an awful lot of hard work which I hope has resulted in a marked improvement in the organisation and quality of the text. Thanks also to Joyce Gillo, Patricia Hobson and Linda Tarrant for typing the earlier drafts, and to Debbie Lewis for producing an excellent final manuscript. My wife Cathy said there was no reason why I should not be capable of producing this book, and she was right.

John Silk
Reading, January 1979

Contents

PREFACE	<i>page</i>	ix
I INTRODUCTION TO STATISTICS IN GEOGRAPHY		1
1. INTRODUCTION		3
References and Further Reading		5
2. DESCRIPTIVE STATISTICS		
Introduction		6
Measurement		6
Variables and observations		9
Summarising data		9
Statistical measures		15
Exercises I		21
Descriptive spatial statistics		22
Exercises II		26
References and Further Reading		26
3. NATURAL AND ARTIFICIAL SAMPLING CONTEXTS		
Introduction		29
Natural sampling contexts		29
Artificial sampling contexts		30
Stochastic and deterministic processes		30
References and Further Reading		31
II NATURAL SAMPLING CONTEXTS		33
4. PROBABILITY		
The relative frequency definition of probability		35
Stability of probabilities		36
Predictability of individual events or cases		38
Some rules of probability		39
Exercises		41
References and Further Reading		42
5. RELATIONSHIPS INVOLVING CHARACTERISTICS OR ATTRIBUTES		
Introduction		43
Independence of characteristics		43
Randomness		45
The chi-square (χ^2) probability distribution		46

Sampling distribution and expected value	48
Looking up critical values in χ^2 tables	49
The χ^2 test	50
Type 1 and Type 2 errors	52
One- and two-tailed tests	53
Assumptions underlying use of the χ^2 test	53
Example	53
χ^2 as a test of association	54
The Multiplication Law for independent events	57
Example	59
Limitations of the technique	61
Exercises	62
References and Further Reading	64
6. RELATIONSHIPS OVER TIME	
Introduction	65
Random and systematic patterns over time	65
Random number tables	66
The runs test for randomness	68
Theory	69
Example	72
Probabilities of independent events over time	73
The binomial theorem and the binomial probability distribution	75
Mean and standard error of the binomial distribution	78
Exercises	79
References and Further Reading	79
7. THE NORMAL DISTRIBUTION	
Introduction	81
Variables showing an approximately normal distribution	81
Properties of the normal distribution	84
The normal curve	84
Areas under the normal curve	84
The standard normal distribution	86
Statistical inference	88
The normal distribution as an approximation to the binomial	90
The normal distribution as an approximation to the probability distribution of the number of runs	91
Exercises	92
References and Further Reading	93
8. RELATIONSHIPS OVER SPACE I: POINT PATTERNS	
Introduction	95
Random and systematic point patterns	95
The Poisson probability distribution and quadrat analysis	98

A test for spatial randomness based on the variance–mean ratio (χ^2 test)	101
Example	102
Advantages and limitations of quadrat analysis	105
Nearest-neighbour analysis	106
The nearest-neighbour index, R	108
Advantages and limitations of nearest-neighbour analysis	109
Exercises	110
References and Further Reading	112
9. RELATIONSHIPS OVER SPACE II: AREA PATTERNS	
Introduction	114
<i>Preliminary definitions and concepts</i>	114
<i>Spatial autocorrelation on regular lattices</i>	117
Contiguity test for spatial autocorrelation	117
Statistical hypotheses and test statistics	122
Examples	122
Searching for spatial autocorrelation in particular directions	125
One-tailed tests	126
Spatial autocorrelation on irregular lattices	127
Advantages and limitations of the contiguity test	128
Exercises	129
References and Further Reading	130
III ARTIFICIAL SAMPLING CONTEXTS	131
10. RANDOM SAMPLING PROCEDURES	
Introduction	133
Populations and samples	133
Objectives of sampling	134
Advantages and limitations of sampling	135
Sampling procedures	136
Independent random sampling and simple random sampling	137
Systematic sampling	140
Stratified sampling	141
Sampling under experimental conditions	143
Other design considerations	144
Scope of inferences based on sample observations	144
Exercises	144
References and Further Reading	146
11. PARAMETRIC METHODS I: THEORY AND ESTIMATION	
Introduction	148
Distribution theory for a sample proportion	148
The Central Limit Theorem	153

Estimation and confidence intervals	154
Confidence interval for Π	154
Notation	157
Confidence interval for μ	158
Sample size and sampling error	160
Working backwards – estimating required sample size	162
Exercises	165
References and Further Reading	165
12. PARAMETRIC METHODS II: HYPOTHESIS TESTING	
Introduction	167
Comparison of two sample means	167
Proportions	167
Means	169
Checking assumptions	172
Comparison of three or more sample means – the Analysis of Variance (ANOVA)	173
Assumptions and theory	174
Example	177
Transformations	179
Exercises	180
References and Further Reading	182
13. NON-PARAMETRIC METHODS	
Introduction	184
Comparing two distributions – the Mann–Whitney U test	185
Sampling distribution of U	186
Examples	188
Correction for ties	191
Comparing three or more distributions – the Kruskal–Wallis	
Analysis of Variance by ranks	192
Comparison of distributions based on nominal or categorical data	196
Exercises	196
References and Further Reading	198
14. SIMPLE CORRELATION	
Introduction	200
Non-parametric techniques	200
Parametric techniques	203
Interpreting the results of a correlation analysis	206
Testing the significance of r	209
Transformations	213
Correlation and regression	213
Exercises	214
References and Further Reading	216

15. SIMPLE REGRESSION ANALYSIS

Introduction	218
Regression analysis in the context of prediction and forecasting	218
Estimation	222
Goodness of fit and variation explained	223
Choice and judgement in regression analysis	226
Exercises I	227
Regression analysis and statistical inference	229
The regression model and its underlying assumptions	229
The sample regression line	232
Precision of estimates from the sample regression equation	233
Significance tests with respect to the sample regression equation	235
Non-linear relationships	237
Examination of residuals	242
Exercises II	248
References and Further Reading	251

APPENDICES	255
------------	-----

STATISTICAL TABLES

A.1	Critical values of the chi-square distribution.	257
A.2	Critical values of u in the runs test for n_1 or n_2 from 2 to 20. Any calculated value of u equal to or smaller than the values given in (a), or equal to or larger than those given in (b), is significant at the 0.05 level.	
	(a) Values of u_L .	
	(b) Values of u_U .	259
A.3	Areas under the normal curve.	260
A.4	The Student t distribution.	261
A.5	The F distribution.	262
A.6	Critical values of U for the Mann-Whitney test.	
	(a) Critical values for one-tailed test at $\alpha = 0.025$, or for two-tailed test at $\alpha = 0.05$.	
	(b) Critical values for one-tailed test at $\alpha = 0.01$, or for two-tailed test at $\alpha = 0.02$.	265
A.7	Random numbers.	266

GLOSSARY OF SYMBOLS	268
---------------------	-----

INDEX	273
-------	-----

INTRODUCTION TO STATISTICS IN GEOGRAPHY

1 Introduction

Statistical methods are used by geographers because they help us to come to conclusions based upon **empirical** data, these being measurements derived either from observation or experiment. Geography was a strongly empirical subject well before the ‘quantitative revolution’ of the 1960s, and so statistical methods can be regarded as essential aids to geographic enquiry.

The way in which empirical evidence and conclusions are related statistically can be illustrated if we consider briefly the difference between mathematics, on the one hand, and statistics, on the other. Mathematics is chiefly **deductive** in nature. Consider the following statements:

- (1) A is greater than B ($A > B$), and
- (2) B is greater than C ($B > C$).

Provided both statements or premises are true, it follows logically that:

- (3) A is greater than C ($A > C$).

We can be quite certain that the conclusion expressed in statement (3) is correct, given the premises expressed in statements (1) and (2), just as we can be quite certain that the theorems of Euclidean geometry are correct, and such logical necessity is a hallmark of deductive reasoning. No matter how complex the argument, therefore, no additional information is required to reach a firm conclusion. In this sense, deductive arguments cannot go beyond the information given.

The position is different in statistics, which is primarily **inductive** in nature. Inductive arguments provide conclusions which in some sense exceed the content of the premises upon which they are based. Suppose we polled a representative sample of adults living in a suburb of a large city, and found that 20% of them commuted to work in the city centre each day. This may be stated as a premise:

- (a) 20% of the adults in the sample from the suburb commute to the city centre.

From this, it might be concluded that:

- (b) 20% of *all* adults in the suburb commute to the city centre.

However, we should feel bound to add a qualification to the effect that the conclusion is ‘probably true’ or ‘approximately correct’, simply because not all adults were questioned. Because of the gap in our knowledge, it is necessary to infer that approximately 20% of all the adults are commuters. This belief or opinion is the result of an **inductive inference**. A **statistical inference** is a form of inductive inference which allows the investigator to be relatively precise about her or his degree of uncertainty, stating, for example, that there is a 95% chance or probability that the true percentage of adults commuting to the city centre lies between the limits set by 14% and 26%. The techniques of **inferential statistics** provide formal procedures for calculating such limits and probabilities, for testing statistical hypotheses, and thereby drawing statistical inferences.

The overall plan of the book is shown in Figure 1.1. Although Chapter 2, on descriptive statistics, is primarily concerned with methods for summarising characteristics of large bodies of data, it should be clear that informal inferences must be made leading to interpretations based on the investigator’s own knowledge and judgement. This also holds true for all ensuing chapters. Following this (Ch. 3), we describe two contexts in which inferential statistics are generally

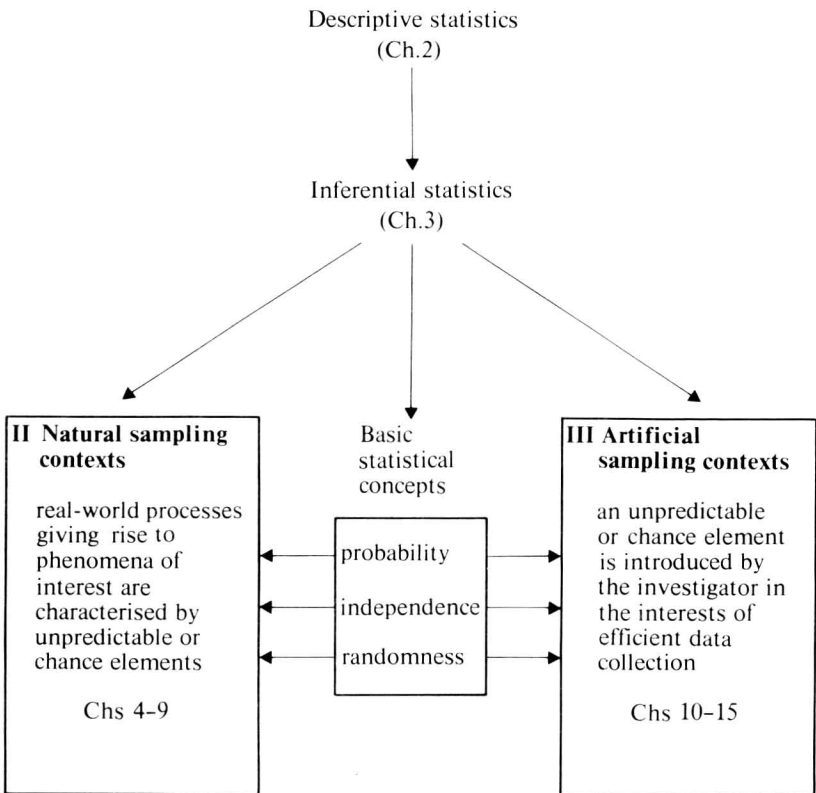


Figure 1.1 Overall plan of the book.

employed: **natural sampling contexts**, involving study of real-world processes giving rise to phenomena characterised by unpredictable or chance elements, and **artificial sampling contexts**, in which an unpredictable or chance element is introduced by the investigator in the interests of efficient data collection. Sections II and III are devoted to techniques employed in natural and artificial sampling contexts respectively, although many techniques can be used in either context, as is made clear where appropriate in the text. The basic statistical concepts of probability, independence and randomness are discussed in detail in Chapters 4 and 5, and reappear throughout the book, particularly in Section II and early in Section III.

Brief comments should be made on the quality of information or data upon which any statistical analysis is based. The data should be both valid and reliable. Measurements are said to be **valid** if they really measure what we think they are measuring – if we are trying to ascertain an individual's knowledge of places in her or his home town, then the number of correct identifications may be regarded as a more valid measure if photographs of places are presented than if place names or maps are used. This is simply because use of names or maps is liable to provide a better measure of an individual's ability to remember names or road maps than of their spatial knowledge. Measurements are said to be **reliable** if they are free from substantial bias – poorly maintained equipment or careless handling of soil or vegetation samples may lead to consistent over- or underestimates of cation exchange capacities or species counts – and liable to relatively small errors. Official statistics should never be regarded as sacrosanct in this respect – errors of up to 22% have been reported as a result of checks carried out on the 1966 Sample Census of England and Wales, and errors exceeding 5% were quite common (Gray & Gee 1972). Careful choice and correct use of a statistical technique counts for little if the measurements are of poor quality – as the computer experts say 'garbage in means garbage out'!

REFERENCES AND FURTHER READING

Comments on census errors may be found in:

Gray, P. and F. A. Gee 1972. *A quality check on the 1966 ten per cent sample census of England and Wales*. London: H.M.S.O., Office of Population Censuses and Surveys, Social Survey Division.

A discussion of data reliability and validity may be found in:

Nachmias, D. and C. Nachmias 1976. *Research methods in the social sciences*. London: Edward Arnold.

An introductory treatment of the topics of inference, deduction and induction is given in:

Salmon, W. C. 1963. *Logic*. Englewood Cliffs: Prentice-Hall. (Chs 1–3).